UNIVERSITY OF
BIRMINGHAM

Molecular approaches to understand the effect of acetic acid on *Escherichia coli*

By

FATEMAH A A H ALATAR

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

Institute of Microbiology & Infection

School of Biosciences

College of Life and Environmental Sciences

University of Birmingham

January 2022

Abstract

Acetic acid has long been known for its antibacterial activity which can be used to treat infected burn wounds. However, there is a lack of a detailed mechanism on acetic acid's effect on *E. coli* at a molecular level. To learn more about this, we have used Transposon Directed Insertion Sequencing (TraDIS) to investigate the molecular mechanisms by which acetic acid acts as an antibacterial agent, by identifying non-essential genes whose loss alters the fitness of different strains of *E. coli*. We grew transposon libraries in three different strains of *E. coli* (uropathogenic *E. coli* EO499 (serotype 131), uropathogenic UTI89, and the lab strain MG1655) in M9 media + 0.2% casamino acids and 0.2% glucose at neutral pH 7 and mildly acidic pH 5.5, with or without acetic acid. Libraries were sequenced pre- and post-growth, using a transposon-specific primer to generate positions and frequencies for each transposon. RPKMS and insertion indices were generated by a TraDIS pipeline. To determine the impact of acetic acid on gene fitness value, the numbers of reads before and after the stress were compared for each gene in each strain. This enables us to identify genes where transposon inserts lead to a decrease or increase of fitness under acetic acid stress. Comparing the results between the strains will enable the identification of both strain-specific genes and genes shared between strains that have a role in fitness under acetic acid stress.

This project consists of two parts. In the first part, we evaluated the roles of candidate genes identified in a previously generated EO499 TraDIS library under acetic acid. Eight of these were depleted under acetic acid stress and they were chosen for further study: *nuoM, nuoG, sucA, sthA, pitA, apaH, rssB* and *ytfP.* Because of the difficulties of constructing gene deletions in the uropathogenic strain for validating the TraDIS results, we tested the relative fitness of the corresponding gene deletion mutants from the Keio library (in lab strain BW25113), with the growth conditions used for EO499. Interestingly, only a few knockouts showed a reduction in relative fitness in competitions at pH 5.5 with acetic acid. This may occur due to the differences between strains used in TraDIS and competition. To overcome this issue, we have also isolated transposon mutants from *E. coli* EO499 transposon library to determine relative fitness.

In the second part, we have optimized and constructed a UTI89 transposon library. The three *E. coli* (EO499, MG1655 and UTI89) transposon libraries were subjected to acetic acid stress as described, for a passaged for a time course of five days. Only day one and day five were sequenced. For the analysis, several bioinformatic pipelines were used for *E. coli* genome annotation and sequencing analysis. TraDIS allowed us to identify essential genes in three *E. coli* strains. The results presented here show which genes tend to be enriched or depleted under acetic acid.

Dedication

To my mother, my father, my brothers

And auntie Ann.

Thank you for always believing in me.

Acknowledgment

I would like to express my special thanks to my supervisor Dr. Peter Lund, for the valuable guidance and advice during the last four years. Dr. Peter helped me enormously with my thesis and scientific writing skills. I am extremely thankful for having your door open to me anytime when I needed guidance and reassurance, you are truly an understanding supervisor. I would like to also express my deepest thanks and appreciation to my colleague Dr. Mathew Milner, who helped me with all the informatics aspects of this project.

I am certainly thankful to all my colleagues for the insightful conversations and the collaborations we shared during this project: Dr. Swaine Chen, Dr. John Herbert, Dr. Emily Goodall, Dr. Keith Turner, Dr. Francesco Falciani, and Dr. Thippesh Sannasiddappa.

I want to extend a huge thanks to Dr. Maria Masoura and Dr. Francesca Bushell, you were indeed excellent work partners and friends. To my T101 lab members: Dr. Mathew Milner, Dana Alfawaz, Jessica Gray, Max Alexander, Jack Bryant, Emily Goodall, Georgia Isom, Samantha Mckeand, Kara Staunton, Rachel Chandler, Emma Sheehan, Christopher Icke, Deema Alodaini, Gabriela Boelter, and Shahida Butt, thanks for the friendly relationships we had that were spent by chatting over coffee and lunch breaks. Thank you all for making T101 a fun and lively place to work, you honestly made T101 worth it.

Most importantly, a special thanks goes to the Kuwaiti government for funding this project and giving me this amazing opportunity. I owe a huge thanks to my academic advisor Ms. Mahasin El-Gaddal in the Embassy of Kuwait – Cultural office London, for the scholarship assistance during the last four years, I am very much appreciative of all your support.

I would like to also offer my special thanks to my supportive and kind friends throughout my studies: Dr. Eiman Ali, Dr. Dana Alsaqer, Dr. Ghayda AlHashem, Lujain Aljahdali, Balqees Al Haddabi, Ragia Mohammed, Shaima Ahmed and Eman Alateeqi, you were always there when I needed you. A special thanks to my cousin Mariam Buzobar for her moral support and encouragement throughout my studies.

To the most important members of my success story, my family; my brothers, my mom and my dad, and my sister in law (Shahad Alshatti), I am entirely thankful for the continuous support, and the regular phone calls or visits that honestly gave me joy, thank you for always spoiling me.

Thank you, auntie Ann, for taking care of me, surprising me with all the wonderful gifts these past four years. Thank you for checking up on me every day and for treating me as one of your daughters. You are such a kindhearted, loving, gorgeous, and an elegant human being and there aren't enough words that can come close to describing your loving nature. May God bless you and your wonderful family.

I would like to thank everyone who made this thesis possible.

Table of contents

List of figures

List of tables:

Abbreviations

AA	Acetic acid
ADP	Adenosine diphosphate
BEDtools	Browser extensible data
bp	Base pair
CDS	Coding sequence
Cm	Chloramphenicol
EB	Elution Buffer
EdgeR	A Bioconductor package for differential expression analysis of digital gene expression data
ENA	European Nucleotide Archive
ESBL	Extended-spectrum β-lactamases
FDR	False discovery rate
GFF	General Feature Format
HYB	Hybridization buffer
ITL	Initial transposon library
Kan	Kanamycin
Kb	Kilobase pairs
LB	Lysogeny broth medium
logCPM	log2 Count Per Million
logFC	log2 Fold Change
MLST	Multi-locus sequence types
mM	Millimolar
NPWT	Negative pressure wound therapy
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PEG	Polyethylene glycol
PGAP	Genome Annotation Pipeline
pH	Power of hydrogen
pK_a	Acid dissociation constant
Prokka	Rapid Prokaryotic Genome Annotation
qPCR	Quantitative PCR
Roary	High speed *stand* alone pan genome pipeline
RPKM	Reads Per Kilobase Million
SAMtools	Sequence alignment/map
SCFA	Short chain fatty acids
SPRI	Solid phase reversible immobilization
ST	Sequence type
TCA	Tricarboxylic acid
Tn	Transposon
Tn-seq	Transposon sequencing
TraDIS	Transposon directed Insertion Site Sequencing
UPEC	Uropathogenic *E.* coli
UTI	Urinary tract infection
w	Relative competitive index

1 Introduction

1.1 The problem of antimicrobial resistance

Antimicrobial resistance is associated with increased morbidity and mortality. Antimicrobial resistance reduces or eliminates the effectiveness of antibiotics and thus compromises human health (Decousser et al., 2003; Diekema et al., 2000, Coelho et al., 2006). Antimicrobial resistance develops when the pathogens causing infection to become less susceptible to antibiotic therapy, and so become difficult to treat. Once the antibiotic fails, the doctors may have to start using undesirable antibiotics with bad side effects, such as colistin, which can cause kidney failure (Boucher et al., 2009).

Examples of some major factors that promote the spread of antimicrobial resistance are poverty, suboptimal control of the sale and quality of antimicrobials, extensive use, poor sewage and water systems. Antibiotic resistance can be spread among humans from antibiotic use in animals, through the food chain or contact with animals (Witte, 1998). The yearly death toll in Europe as a result of multidrug-resistant bacterial infections is estimated to be 25,000 people and this costs the European Union economy €1.5 billion each year (Davies et al., 2013). It was reported in 2014, in total 700,000 people die yearly from bacterial infections, HIV (human immunodeficiency virus), TB (Tuberculosis) and malaria (O'Neill, 2015). In US, more than 2 million cases of bacterial infection found to have resistance to at least the first antibiotic used, this cost the health system around 20 billion USD each year (Smith and Coast, 2013). By 2050, based on current trends, it has been estimated death due to antimicrobial resistance could cost 10 million lives each year, if no measures are taken (O'Neill, 2015). One of the main reasons for antibiotic resistance is the spread of resistance genes via horizontal gene transfer (Hawkey, 2003; Hooper,

2001) and overuse of antimicrobials with lack of new drugs to combat antibiotic resistant (Shankar, 2016).

Drug resistance can occur naturally where the resistance genes are already present in the bacterial chromosome or it can be acquired via mobile genetic elements such as plasmids, transposons, bacteriophages or naked DNA (e.g. PCR fragment). Resistance can also occur spontaneously through mutation (Palmer and Kishony, 2014; Levy and Miller, 1989).

1.2 The envelop of Gram-negative bacteria and antibiotic resistance

Multidrug resistance has been identified in both Gram-negative and Gram-positive bacteria. Gram-negative bacteria generally have a higher level of antibiotic resistance than Gram-positive due to their different cell membranes and cell wall structures (Nikaido, 1996). The cell envelop of Gram-negative bacteria consist of the outer membrane, the peptidoglycan, and the cytoplasmic or inner membrane. The outer membrane is unique to Gram-negative group and it is more permeable than the inner membrane. It contains integral pore proteins and porins, which generally allow the passage of water and small molecules e.g. glucose and monosaccharides. One important function of the outer membrane is to act as protective barrier. This helps slow and even prevent the entry of bile salts, antibiotics, and other toxic substances present in the environment that might kill or injure the bacterium. The space between the outer membrane and the inner membrane is the periplasm (periplasmic space). This space contains the peptidoglycan layer. The inner membrane retains the cytoplasm, separating it from the surroundings. It is very

important site for many metabolic processes such as respiration, the synthesis of lipids and cell wall constituents. In contrast, Gram-positive bacteria lack the outer-membrane and have a thicker peptidoglycan layer in compare to the Gram-negative bacteria.

There are several structures that strengthen the Gram-negative wall, for example the outer membrane and peptidoglycan are firmly linked by lipoproteins. Lipoproteins are membrane proteins that can bind the underlying peptidoglycan and are embedded in the outer membrane by their hydrophobic end of the phospholipid bilayer. Also, the presence of lipopolysaccharides the outer membrane is important. The lipopolysaccharides consist of three parts (a) lipid A (b) the core polysaccharide (c) the O-antigen chain. The host antibodies can recognize O-antigen chains, but Gram-negative bacteria can evade the host defense by changing the nature of O-antigen to escape detection. So, the antibodies interact with the lipopolysaccharides before getting to the outer membrane which protect the cell wall from host attack.

There are several mechanisms that can cause drug resistance in Gram-negative bacteria such as the presence of a double membrane surrounding the bacterial cell wall which excludes large antibiotics. Also, the presence of porin which are the outer membrane protein found in the outer membrane, which regulate the transport of hydrophilic molecules across the outer membrane. These features from Gram-negative bacteria are the major obstacles in antibiotic finding against clinically important Gram-negative pathogens, as shown in figure 1 which compares the structures of cell envelop in Gram positive and Gram-negative bacteria (Prescott, 2005). For instance, glycopeptides, vancomycin and teicoplanin are bactericidal antimicrobials used only against to Gram-positive bacteria. These are antibiotics that act by inhibiting cell-wall

synthesis in Gram-positive bacteria. These antibiotics have a large molecular weight size and are not able to penetrate the outer membrane of Gram-negative bacteria (Greenwood, 1988; Choi, 2019).

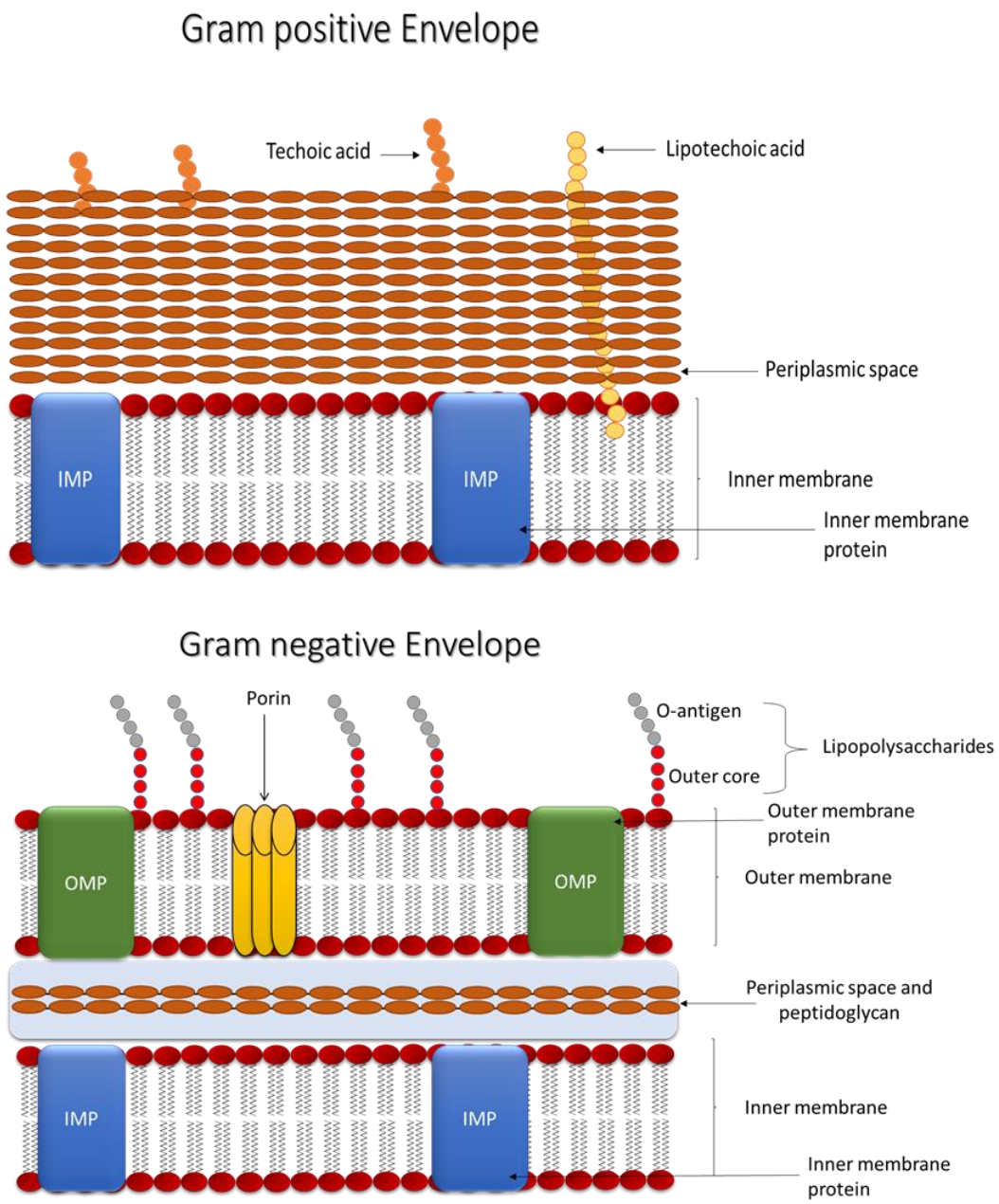


Figure 1. Cell wall structures of Gram-positive and Gram-negative bacteria.

Antibiotic resistance can be also acquired either (a) by mutation of an existing gene or (b) by either of a plasmid carrying antibiotic resistance genes. For example, fluoroquinolone resistance has emerged in *E. coli* and other Enterobacteriaceae as a result of mutations in the genes encoding the target enzymes (DNA topoisomerases), or mutations that cause over-expression of the membrane protein will pump out the incoming drugs (Wang et al., 2001; Levy, 1985). DNA topoisomerases are enzymes that are responsible for overwinding or underwinding of DNA during replication. Another example of antibiotic resistance is the spread of plasmid encoded genes for extended-spectrum β-lactamases (ESBL), which is common among Gram-negative species. Infection caused by ESBLs-expressing bacteria can cause serious illness. These infections are difficult to treat and they became common in the community and healthcare settings (Woodford et al., 2011).

The medically most common Gram-negative pathogens found in Germany, Austria, India and U.S are *Acinetobacter, Campylobacter, Escherichia coli, Salmonella, Shigella, Klebsiella, Vibrio cholerae, Pseudomonas aeruginosa* and *Helicobacter pylori* (Exner et al., 2017). In England and Wales, a study done on Enterobacteriaceae of *E. coli*, *Klebsiella* and *Enterobacter* bacteraemias shows a resistance level rising up to 14 % ciprofloxacin resistance between 1990 – 2002 (Livermore, 2004) (Ciprofloxacin belongs to the group Fluoroquinolones which inhibit DNA synthesis as described above).

1.3 New therapeutic approaches

There are more than 14 classes of antibiotics acting as bacteriostatic and bactericidal agents whose targets are generally involved in essential cellular and physiological processes of bacterial cells. None of these has escaped the evolution of bacterial resistance mechanisms (Levy, 2002). So alternative approaches are required to deal with infectious diseases caused by multidrug resistant bacteria (Barrow and Soothill, 1997). One of these could be combination therapies of antibiotics with natural antibacterial agents, which may also help to overcome the problem of antimicrobial resistance (Castro et al., 2019). There is therefore interest in investigating antibiotic alternatives and natural antimicrobial agents. It remains crucial that we find alternative to conventional antibiotics (Zhitnitsky et al., 2017). Organic acids are an example of environmentally friendly antibacterial alternatives. Many studies showed the antimicrobial activities of organic acids, and these are discussed in the following section.

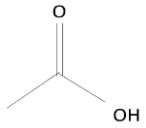
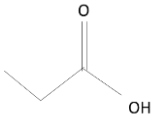
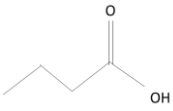
1.4 Organic Acids

Organic acids are an organic compound with acidic properties. They are known as weak acids because they don't fully dissociate in water. They are typically naturally occurring compounds found in nature from different sources (i.e., animals, plants, and microbial organisms). The most common type of organic acids are the carboxylic acids, in which the carbon is connected by four covalent bonds, two bonds with oxygen (C=O) and one bond to hydroxyl group (-OH). The last fourth bond connects the carbon to hydrogen. The carboxyl (-COOH) group

is named as the result of having both a carbonyl (C=O) group and a hydroxyl group (-OH) (March, 1968). The other type of organic acids has a (-OH) attached to a carbon atom that is a part of an aromatic ring structure (i.e., phenol).

The relative stability of the conjugate base of the acid determines its acidity, as shown with further details in section 1.6. Organic acids can have long aliphatic tails (chain of carbon and hydrogen) with varying numbers of carbon atoms. They can vary from short chain fatty acids (SCFA) which has a fewer than six carbon atoms to a very long chain fatty acid which has 22 or more carbons. As the focus of this project is the short chain fatty acids, the three most common short chain fatty acids are acetic acid, propionic acid and butyric acid. Table 1 showed the nomenclature, molecular formula and diagram of these three short chain fatty acids. These SCFAs are produced in millimolar quantities in the gastrointestinal tract and occur in high concentrations in those areas where strictly anaerobic microorganisms are predominant.

Table 1. Examples of short chain fatty acids.

Common name	Nomenclature	Formula	Diagram
Acetic acid	Ethanoic acid	CH_3COOH	
Propionic acid	Propanoic acid	CH_3CH_2COOH	
Butyric acid	Butanoic acid	$CH_3(CH_2)_2COOH$	

1.5 Organic Acids as Antimicrobial Agents

Organic acids have been used for very long time in both the human and animal food industries, and in human infection treatment. In the food industry, organic acids are used because of their preservative properties as they help to prevent food spoilage by reducing bacterial and fungal growth, which extends the shelf life of perishable foods (Frank, 1994). Acetic, sorbic and propionic acids are examples of major food preservatives. Additionally, organic acids sprays are used as sanitizers during meat processing (Acuff, 1987; Hamby, 1987).

In addition to studies on food, many studies have investigated the effect of acetic acid against infection of burn wounds. Patients suffering from burn wounds that become infected often develop subsequent sepsis (blood poisoning), which is difficult to treat and can be fatal, especially with the growing issue of antibiotic resistance. Burn wound infections often form

biofilms which are also hard to treat. It has been shown that low concentration of acetic acid is effective against common wound-infecting pathogens in the laboratory (Halstead et al., 2015). On top of that, a nationwide telephone survey of UK burns units was conducted to determine the antimicrobial agent used against *Pseudomonas* burn wound infections. In 19 burns units, seven types of antimicrobials dressing were frequently used, acetic acid-soaked dressing was the most common agent used for treating patients with burn wounds infected by *Pseudomonas* in UK (Nour et al., 2021). Moreover, in clinical trails involving 100 patients with infected wounds suffering from diabetic, trauma, burns, venous ulcers and infection developed at the graft donor site. A number of Gram-negative and Gram-positive pathogens from the wounds were isolated. The patients were treated with 1% acetic acid dressing material. The 1% acetic acid were diluted with normal saline. Acetic acid found to be effective and faster heal against the isolated wound pathogens, in which these pathogens were found to be resistance to the commonly used antibiotics. The study have direct implications for efforts aimed reducing health care cost (Agrawal et al., 2017)

A clinical study over approximately five years and with more than 200 patients, showed that acetic acid with combination of negative pressure wound therapy (NPWT) was effective against biofilms in chronic infections. NPWT is a vacuum system that helps to draw out fluid from the infected wound. The study was done on patients with ulcers and postoperative infections after osteosynthesis and articular infections (Bjarnsholt et al., 2015). Similarly, seven hospitalized patients with *P. aeruginosa* wound infections were successfully treated with 3% of acetic acid for 15 minutes over 12 applications (Nagoba et al., 2013). The use of various organic acids like boric

acid, ascorbic acid, citric acid, salicyclic acid and acetic acid on *P. aeruginosa* of skin and soft tissue infection sites has been reported to be effective in various studies (Nagoba et al., 2013; Kundukad et al., 2020).

Recently, an in vitro study on non-typhoidal multidrug resistant *Salmonella* strains found that a combination of 2% acetic acid and vancomycin were significantly effective at killing bacteria. *Salmonella* is resistant to vancomycin, but the acetic acid enhanced the permeability of the bacterial outer membrane which would allow that substance access inside the cell. Then, the acetic acid enhanced the activity of vancomycin which inhibits peptidoglycan biosynthesis (Castro et al., 2019). Another recent study examined the antimicrobial activity of different types of weak organic acids such as citric acid, malic acid, propionic acid, lactic acid, benzoic acid, mandelic acid, pyruvic acid and hippuric acid against urinary tract infections (UTIs) pathogens caused by urinary catheters. Between 15-25% of enrolled hospital patients received catheterization. These urinary catheter devices are very susceptible to infection due to their position in the bladder which often developed bacterial biofilm growth. About 70% of the hospital acquired UTIs were catheters associated infections. The study showed that all the examined weak organic acids have bactericidal activities against uropathogens during both planktonic and biofilm modes of growth. The study also used a combination of the weak organic acids against the UTIs pathogens which tends to greatly reduce the concentrations needed to inhibit bacterial growth in compare when used alone (Burns et al., 2021).

1.6 The General mode of action of organic acids at low pH

To understand the properties of weak organic acids there is need to understand the factors which affect their relative strength. First, acids are defined simply as substances that are able to donate hydrogen ions (protons, H^+). In water strong acids such as hydrochloric acid (HCl) dissociate almost 100%: in other words, the equilibrium position in the equation below is very far to the right. This results in hydroxonium (H_3O^+) and aqueous anions:

$$HA + H_2O \rightleftharpoons A^- + H_3O^+ \qquad (1)$$

where, (A^-) is a conjugate base and (H_3O^+) is a conjugate acid. In general, the organic acids are weak in the sense that this ionization is very incomplete. So, only a small percentage of organic acids dissociate into organic acids anions and hydroxonium ions, and the majority remain as uncharged organic acid molecules. In this case, the equilibrium position lies further to the left.

The strength of acid can be measured by the acid dissociation constant (K_a) value, which is the equilibrium constant for the above reaction, shown as:

$$K_a = \frac{\left[[H_3\,O^+][A^-]\right]}{[HA]}$$

A better measure of the strength of an acid is the pK_a value, which is defined as :

$$pK_a = -log_{10}\,K_a$$

The smaller the value of pK_a the stronger the acid is, as a small pK_a indicates the acid dissociates more readily in water. Weak acids have a small K_a values and a higher pK_a values compared to strong acids, which have large K_a values and negative pK_a values. For example, the

K_a of acetic acid is 1.8×10^{-5}, but the pK_a constant is 4.76. while the a strong acid such as hydrochloric acid HCl the K_a is 1.0×10^{6}, but the pK_a constant is -6.

There is inverse relationship between the relative acid strength and the relative base strength. This means the conjugate base of a strong acid is a weak base, and the conjugate base of weak acid is strong base. Figure 2 shows the relative strengths of some common conjugate acid–base pairs. For example, HCl is a strong acid, and strong acids are very good proton donors. The HCl conjugate base is (Cl^-), the Cl^- is weak proton acceptor. Therefore, the Cl^- ion is a weak base. But the acetic acid is a weak acid, when it donates a proton to water, its conjugate base (acetate ions, CH_3COO^-), are stronger than the conjugate base of HCl (Cl^-).

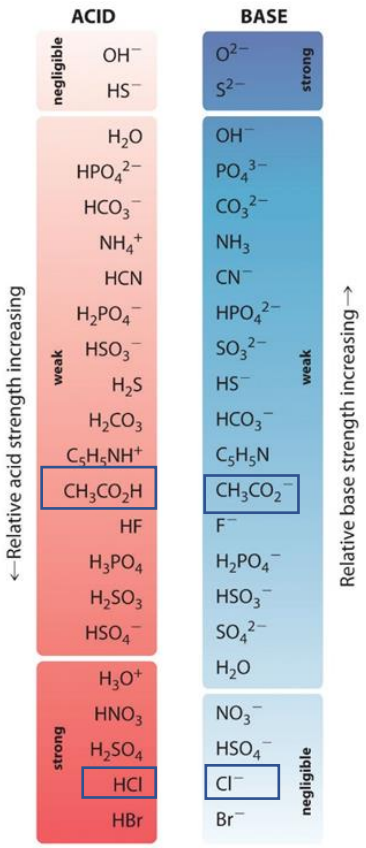


Figure 2. The relative strengths of some common conjugate acid–base pairs.

The squares show acetic acid (CH_3COOH) and hydrochloric acid (HCl). Taken from: (https://chem.libretexts.org/Bookshelves/General_Chemistry, 2012)

Acetic acid (CH_3COOH) reacts with water to form acetate anion and hydroxonium (hydrated proton).

$$CH_3COOH + H_2O \rightleftharpoons CH_3COO^- + H_3O^+ \qquad (2)$$

The K_a equilibrium expression for this reaction (2) is:

$$K_a = \frac{[CH_3COO^-]\,[H_3O^+]}{[CH_3COOH]} = 1.8 \times 10^{-5}\ \text{M}$$

$$pK_a = -log_{10}(1.8 \times 10^{-5}) = 4.75$$

From the known pK_a value of any acid, we can calculate the amounts of protonated and deprotonated forms at any pH. This can be calculated by Henderson-Hasselbalch equation which connects the pH of a solution with the acid dissociation constant (pK_a):

$$pH = pK_A + log_{10} \frac{[conjugate\ base\ (A^-)]}{[weak\ acid\ (HA)]}$$

Or: $$log_{10} \frac{[conjugate\ base\ (A^-)]}{[weak\ acid\ (HA)]} = pH - pK_A$$

where the [A-] is molar concentration of a conjugate base and [HA] is the molar concentration of undissociated weak acid. Table 2 shows the calculated values of percentage dissociated and undissociated at pH 7 and pH 5.5 of acetic acid and hydrochloric acid. These parameters were used in this project.

Table 2. The values of acetic acid and hydrochloric acid at pH 5.5 and 7 dissociated and the undissociated by Henderson-Hasselbalch equation.

Compound	pH	pK_a	H^+= (10^{-pH})	K_a= (10^{-pKa})	HA= $[H^+] \times [A^-] / K_a$	% Dissociation = $[H^+]/[A^-]+[HA] \times 100$	% Undissociation = (100 - % dissociation)
Acetic acid	7	4.75	1×10^{-7}	1.7×10^{-5}	5.6×10^{-10}	99.44	0.56
Acetic acid	5.5	4.75	3.1×10^{-6}	1.7×10^{-5}	5.6×10^{-7}	84.90	15.10
HCl	7	-7	1×10^{-7}	10000000	1×10^{-21}	100.00	0.00
HCl	5.5	-7	3.1×10^{-6}	10000000	1×10^{-21}	100.00	0.00

It is important to note based on the Henderson-Hasselbalch equation and table 2, for acetic acid or organic acids in general, as the pH drops the amounts of undissociated (i.e. unionized) form increase. Undissociated organic acids are typically lipid soluble can cross the hydrophobic inner membrane. Once the organic acids are in the cells, as the intracellular pH is often neutral, this will cause the organic acid to dissociate based on the Henderson-Hasselbalch equation in table 2. Then, the organic acids will liberate their proton (H^+) and organic acid anions. The protons lead to pH reduction in the cytoplasm (Booth, 1985). Subsequently, the cells will start pumping the excess protons across the membrane using F_1F_o-ATPase to restore the cytoplasmic pH to normal. The F_1F_o-ATPase is the prime producer of ATP, using the proton gradient generated by oxidative phosphorylation, in the inner membrane (Russell, 2007). The accumulation of organic acid ions may be toxic or cause osmotic stress, and the effect of the anions depends on the type of organic acid (Mira et al., 2010), as shown in figure 3. Consistent with the mode of action, the effect of the weak acids depends on the external pH. The effect of these compounds increases as the extracellular pH becomes more acidic. However, this may reduce cell growth, and this might occur for a variety of reasons, such as interaction with the membrane or build-up of toxic anions in the cytoplasm or reduction of the cytoplasmic pH and subsequent partial collapse of the transmembrane proton gradient due to pH difference (Kitko et al., 2009).

Furthermore, the response to stress of lowered intracellular pH leads to attempts to restore pH homeostasis and results in the reduction of available energy pools for growth and other essential metabolic functions. Under acid stress, pH homeostatic is the process of increasing expression of genes and the activity of proteins or pathways that result in proton pumping out of

the cell (Brul and Coote, 1999; Kroll and Booth, 1983). In particular, the first response to cope with the stress is the ability to adjust membrane properties, such as lipid composition in the cytoplasmic membrane, which effectively changes the proton permeability and channel size of the membrane (Booth, 1999; Sohlenkamp, 2017). Some microbes modify the distribution of fatty acids in the bilayer structure to regulate membrane fluidity (Lindberg et al., 2013). The most common mechanism used by bacteria is altering the ratios of unsaturated to saturated and branched to unbranched fatty acids of phospholipids to control membrane fluidity (Denich et al., 2003). Saturated or unsaturated refers to whether or not double bonds exist between the carbons in the fatty acid tails. Saturated fatty acids exist without double bonds which results in straight tails, while unsaturated fatty acids have a double bond which result in crooked tails. The saturated fatty acids are arranged in a way to increase the interactions between the tails, to decrease fluidity of the bacterial membrane (Eden, 2017). Many studies reported that higher level of unsaturated level of membrane fatty acids contributes to stress resistance to low pH (Wu et al., 2012; Tan et al., 2016). Moreover, altering membrane fluidity can be by modifying the proportion or the type of branching of fatty acids (Kaiser et al., 2016; Sen et al., 2015). In some bacteria, cyclopropane fatty acid (three carbon atoms linked together to form a ring in the fatty acids chain) formation modifies the lipid bilayer, and this can be a critical factor in acid stress by decreasing in membrane proton permeability (Chang and Cronan, 1999). Also, some bacterial strains are found to increase the length of the fatty acid chain to increase survival under acidic stress (Wu et al., 2012).

Previous studies suggested that some bacteria modulate the size of membrane channels as one of the adaptive mechanisms to maintain pH homeostasis. Increase the expression of peptide loop (L3) across the porins of the outer membrane in response to acidic stress. L3 is a large external loop part of OmpC outer membrane porin forming a folding model Omp40. This loop narrows the size of the porin gateway, which restrict the diffusion of proton form the outside of the cell to the inside and vice versa (Amaro et al., 1991; Guiliani and Jerez, 2000; Guan and Liu, 2020).

Additionally, urease is known in some bacteria to neutralize acidic pH by producing ammonia. Urease converts urea (CH_4N_2O) into ammonia (NH_3) and CO_2. The ammonia can be produced endogenously from amine-containing amino acids such as arginine or glutamine, to maintain pH homeostasis which raises the cellular pH. This action depends on urea influx through urea channel located in the inner membrane, and ammonia efflux, transporting the ammonium into periplasm. In the periplasmic space, the ammonia binds with protons leading to neutralization of the acidic pH ($H^+ + NH_3 \rightarrow NH_4^+$). Finally, NH_4^+ exits the bacterial cell (Ansari and Yamaoka, 2017; Vollan et al., 2017).

Another source of stress with organic acid is the accumulation of organic acid anions, which affects the cellular growth in various ways. In general, increased anion concentrations have been shown to be associated with increased transport of potassium ions into the cell, which raises turgor pressure (Kroll and Booth, 1983). To regulate the increased turgor pressure glutamate is transported out of the cells (McLaggan et al., 1994). The glutamate result from deamination of

glutamine, via acid-glutaminase (YbaS/GlsA) (Pennacchietti et al., 2018). The transport of glutamate effects the cellular osmolarity cell growth and viability. Enzymes involved in protein synthesis become sensitive because of the cytoplasmic acidification or the anion pool or both (Roe et al., 2002).

The cause of growth inhibition by organic acids is affected by the intracellular pH and the cellular energetic status. However, with all the available data the basis of growth inhibition by weak organic acids remains uncertain. This is one of the key points this project will focus on.

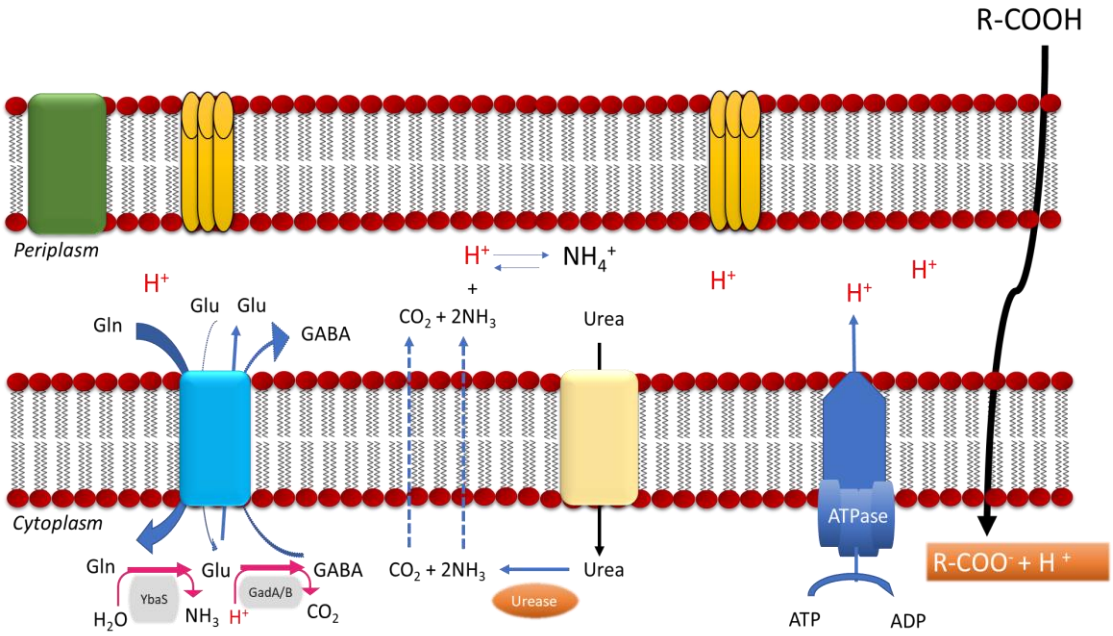


Figure 3. General mode of action of organic acids against gram-negative bacteria.

1. The uncharged carboxylic acids penetrate the bacterial cell membrane and dissociate to organic acid anions and protons (H^+) due to the intracellular pH. 2. Accumulation of organic acids anions in the cell may be toxic and/or cause osmotic stress and protons cause decrease in the intracellular pH. The details are discussed in the text. Gln, glutamine; Glu, glutamate; GABA, ϒ-aminobutyric acid; RCOOH, general form of undissociated organic acid; $RCOO^-$, dissociated form of organic acid. Figure adapted from (Lund et al., 2020).

1.7 Acetate in *E. coli*

With reference to previous section 1.6 general mode of action of organic acids, the accumulation of acetate increases the internal osmotic stress. *E. coli* can use acetate as a source of both carbon and energy, thus removing the toxin from its environment by consuming it. The

process of using acetate as a carbon source rather than secreting it from the cell is known as the "acetate switch". This section will discuss acetate metabolism in *E. coli* briefly.

In *E. coli*, acetate is excreted as an end product during aerobic condition and anaerobic fermentation (Wolfe, 2005; Bernal et al., 2016). The process of formation and secretion of acetate is known as acetogenesis. During aerobic growth conditions in glucose supplemented medium, acetate production referred as overflow metabolism. Acetogenesis takes place when the carbon or glucose flux to the tricarboxylic acid (TCA) is high and the TCA cannot operate fast enough to metabolize all the input carbon (Chang et al., 1999; Farmer and Liao, 1997). In general acetate is produced directly from pyruvate by decarboxylation using by the pyruvate oxidase (PoxB). The second pathway involved in acetate production starts with the conversion of acetyl-coenzyme A (acetyl-CoA) to acetyl~P by phosphate acetyltransferase (*pta*) then to acetate by acetate kinase (*ackA*). These two steps of the conversion of acetyl-CoA to acetate generates ATP by consuming ADP and inorganic phosphate Pi, figure 4 (Brown et al., 1977). The same process can be used in the opposite direction for assimilation of acetate at medium or high external concentrations, figure 4 (Brown et al., 1977). This assimilation process functions primarily by AMP-forming acetyl-CoA synthetase (AMP-ACS). This enzyme catalyzes the production of acetyl-CoA from acetate. When the nutrients needed for growth are depleted in the medium, metabolism can alter to allow continued cell survival and growth. This shift in the bacterial metabolic pathway involves importing and using the acetate that was previously excreted. Thus, *E. coli* grown in glucose medium not only produces acetate but also consumes acetate (Wolfe, 2005). The acetate switch takes a place when the environment of acetate-producing carbon source such as glucose is

depleted, then the cells start to use acetate as a carbon source (Kumari et al., 1995). A study recent study found cells are not able to consume acetate at low pH (pH 6) during growth. Also, they found PTA-ACKA is the essential pathway for acetate production (Orr et al., 2019).

During mixed acid fermentation under anaerobic culture of *E. coli* various products are produced including acetate, figure 4 (Sawers and Clark, 2004). Fermentation is an anaerobic reaction in bacteria, in which it converts six-carbon sugar (e.g. glucose) to a variable mixture of acids. This type of anaerobic cell growth happens when electron acceptors for cellular respiration are absent (e.g. nitrate (NO_3^-) or nitrite (NO_2^-)). The mixed acid fermentation starts with the production of pyruvate produced by the glycolytic pathway (the process in which glucose is converted into pyruvate, ATP and NADH). Then the pyruvate will be converted into two products acetyl-CoA and formate facilitated by the pyruvate formate lyase enzyme. In the same metabolic pathway acetyl-CoA can be transferred to acetate by phosphotransacetylase (PTA) and acetate kinase (ACKA), producing ATP. Also, the acetyl-CoA can be converted to ethanol by alcohol dehydrogenase (ADH) utilizing NADH (Sawers and Clark, 2004).

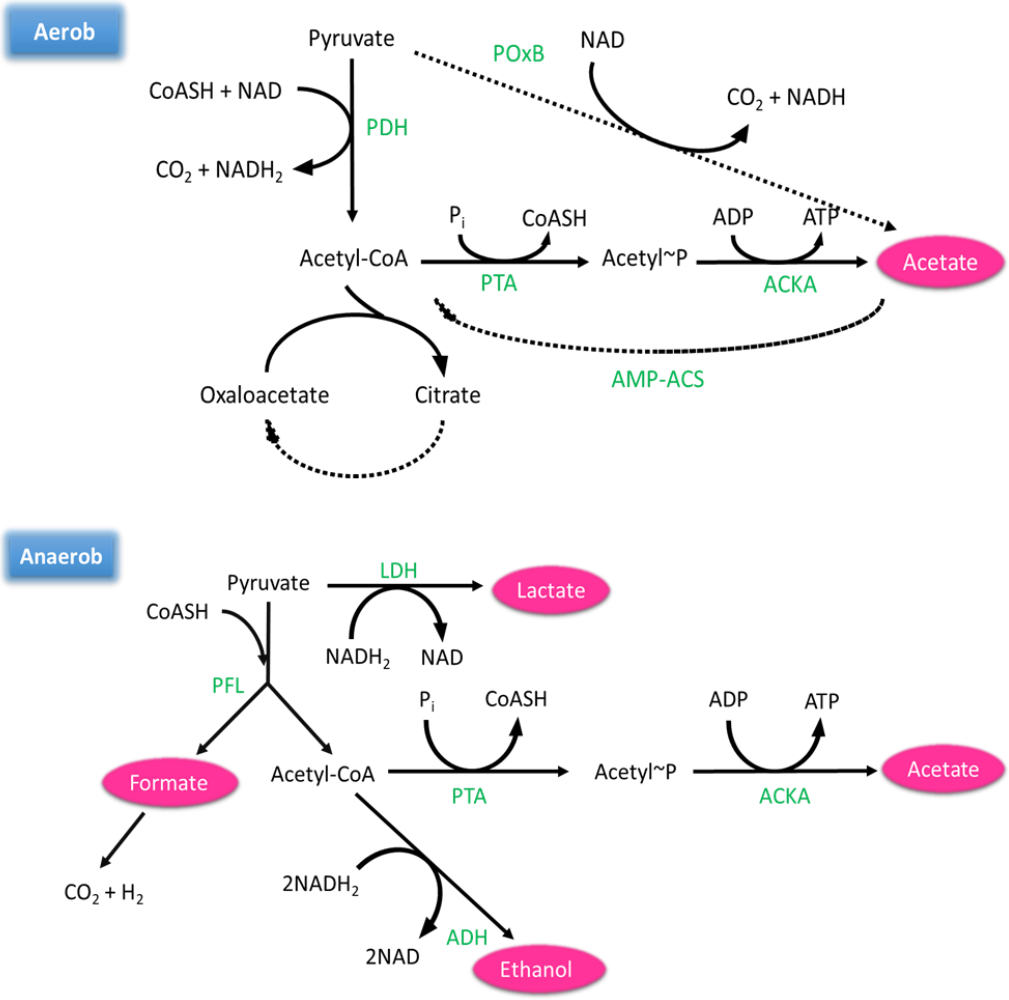


Figure 4. Acetate metabolism pathway in *E.coli* under aerobic and anaerobic conditions. The green text indicates: enzymes PDH, pyruvate dehydrogenase complex; ACKA, acetate kinase; PTA, phosphotransacetylase; POxB, ACS, AMP-forming acetyl-CoA synthetase, Pyruvate Oxidase; LDH, lactate dehydrogenase; PFL, pyruvate-formate lyase; ADH, alcohol dehydrogenase. Figure adapted from (Schütze, 2020; Wolfe, 2005).

1.8 Strain used in this study

Escherichia coli is a Gram-negative facultative anaerobic rod-shaped bacterium. “Facultative anaerobic” means an organism that does not require oxygen for growth but grows better in the presence of oxygen. So, they can change their process depending on the presence of oxygen, in the presence of oxygen the cells obtain energy from aerobic respiration and in the absence of oxygen the cells use anaerobic respiration or fermentation. *E. coli* is common in the normal intestinal flora in humans and animals (Meng et al., 2007). *E. coli* is a chemoheterotroph and hence requires a chemical source of energy for growth. “Chemoheterotroph” refer to the use of chemicals (rather than light) as an energy source. *E. coli* grows in a variety of defined laboratory media, an example of typical defined medium for *E. coli* contains glucose ($C_6H_{12}O_6$), ammonium phosphate monobasic ($(NH_4)H_2PO_4$), sodium chloride (NaCl), magnesium sulfate ($MgSO_4$), potassium phosphate dibasic (K_2HPO_4) and water. Several studies have firmly established *E. coli* as a choice to study molecular biology and it was used for classical studies into the genetic code, and the processes of transcription, translation, and DNA replication (Crick et al., 1961; Nirenberg, 1965; Idalia and Bernardo, 2017).

E. coli is a well-known bacterium that lives as a harmless inhabitant of the guts of human and many animals, but some strains of *E. coli* are dangerous pathogens. In humans, the pathogenic *E.* coli stains can cause urinary tract infection UTI, gastroenteritis, meningitis, and wound infections. In this study, three *E. coli* strains were used, two uropathogenic *E.* coli strains (UPEC) and a non-pathogenic lab strain. The UPEC strains were EO499 and UTI89 and the lab strain is *E. coli* K-12 MG1655.

Phylogenetic studies show that *E. coli* strains are distributed among six phylogroups: A, B1, B2, D, E and F. Phylogenetic trees are diagrams that depicts the origin and evolution of groups of any species from their common ancestor. Lineages often split when populations or groups become genetically isolated from the common ancestor. As a consequence of this genetic isolation, the lineages will evolve separately over time. A recent phylogenetic analysis on 20 different *E. coli* genomes constructed a phylogenetic tree showing different *E. coli* groups (Forde et al., 2014). In figure 5, the two UPEC strains UTI89 and EO499 in this study belong to phylogroup B2 as shown. Group B2 is primarily dominated by uropathogenic strains. EO499 is not specifically mentioned in figure 5 in this paper, but it is a member of the ST131 clade which also belongs to phylogroup B2. In addition, MG1655 can be seen to be a member of phylogroup A. Group A contains the commensal strains and their derivatives (Sims and Kim, 2011). We would expect to find more overlap in genes between the two phylogroup B2 strains than with MG1655 in phylogroup A. It could be further expected that there may be essential gene homologues in EO499 and UTI89 that however not found in MG1655. These predictions were tested here.

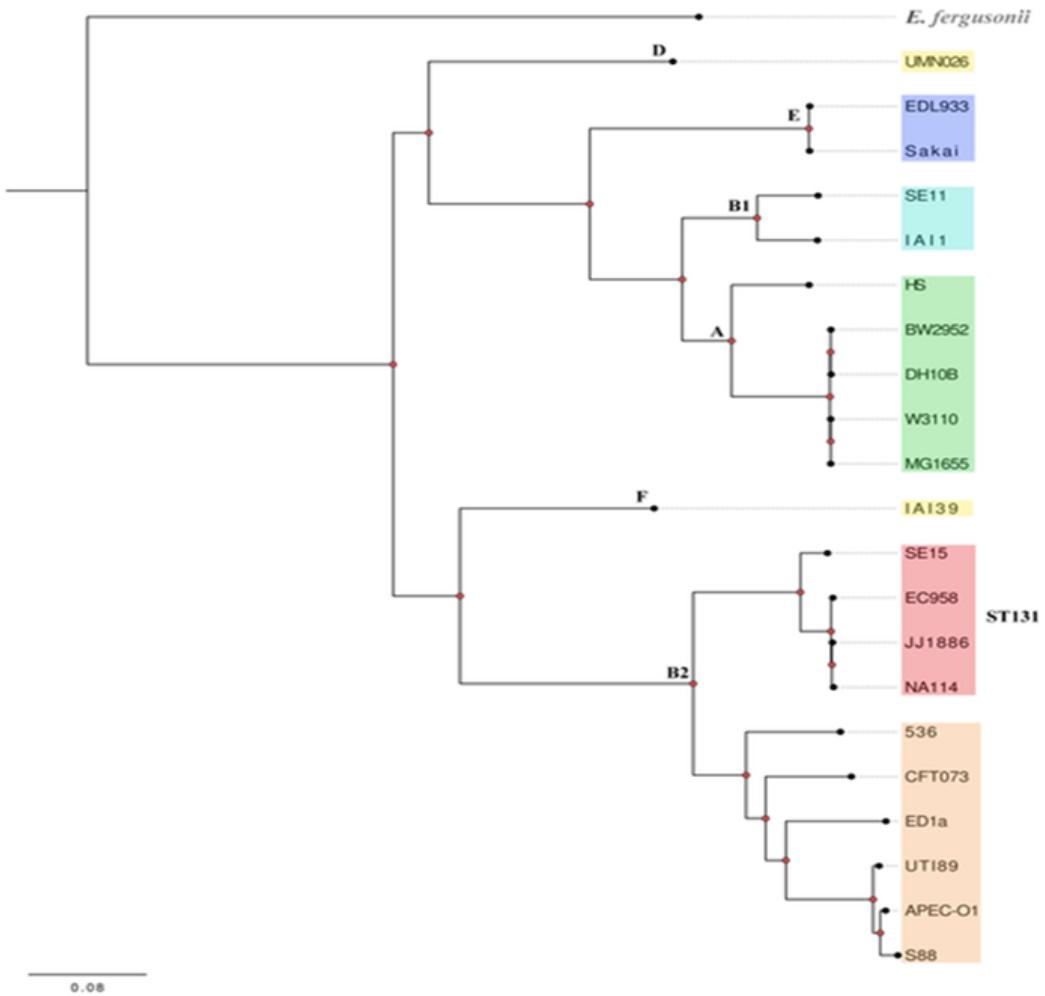


Figure 5. Phylogenetic tree of representative *E. coli* isolates.

The figure was taken from (Forde et al., 2014).

1.9 Uropathogenic *E. coli*

1.9.1 *Escherichia coli* EO499

EO499 was isolated in 2003 in UK from a patient attending general practice (Woodford et al., 2004). This project will investigate strain EO499 which belongs to sequence type ST131.

"Sequence type" refers to Multi-locus sequence types (MLST) which is a molecular technique using DNA sequencing of six to seven well-conserved housekeeping genes within the bacterial genome such as (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*). For each housekeeping gene, each unique fragment sequence represents a distinct allele, for each isolate, and by assigning each allele a number, and considering the combinations of numbers obtained across several genes, then each one is assigned a sequence type (ST) number. ST131 strain belongs to phylogenetic group B2. It is serotype O25:H4 and the national collection of type cultures (NCTC) number is NCTC is 13441. The whole genome sequence of EO499 is available in European Nucleotide Archive (ENA) and was determined by the Sanger institute. Serotypes (serological typing) are separated groups within a single species of microorganisms. The serotype classification is based on the observation of mainly two structures on their surface: the H (flagella) antigen based on the protein content of flagella and O (outermost) antigen which is part of a specific lipopolysaccharide. EO499 contains the pEK499 plasmid, the complete nucleotide sequence of the plasmid is available in GenBank with accession number EU935739 (Woodford et al., 2009). The pEK499 plasmid produces CTX-M enzymes that are a group of extended- spectrum β-lactamases (ESBL). The β-lactamase enzymes give bacteria resistance to antibiotics such as such as cefotaxime, ceftriaxone, ceftazidime, or cefepime and aztreonam, and represent a major public health concern. β-lactamase provides antibiotic resistance by hydrolyzing the β-lactam ring of these antibiotics (Coque et al., 2008, Pitout, 2008). In total the pEK499 contains around 10 antibiotic resistance genes $bla_{\text{CTX-M-15}}$, $bla_{\text{OXA-1}}$, $bla_{\text{TEM-1}}$ (Beta-lactamase), *tetA* (resistance to tetracycline), *aac6'-Ib-cr (*provide *resistance to* aminoglycosides and ciprofloxacin*)*, *mph* (provide

resistance to macrolides class antibiotic containing: erythromycin, roxithromycin, azithromycin and clarithromycin), *catB4* (resistance to chloramphenicol*), dfrA7* and *aadA5* (trimethoprim and streptomycin resistances) and *sulI*, (sulfonamide resistance). This is associated with limited treatment options (Woodford et al., 2009).

UPEC is the most common cause of urinary tract infection responsible of ~ 80% of all cases in United states (Flores-Mireles et al., 2015). UPECs encode several virulence pathogenicity factors which enable the bacteria to colonize the urinary tract such as adhesins, toxin secretion, iron acquisition factors, lipopolysaccharides, polysaccharide surface structures, flagella and plasmids (Nicolas-Chanoine et al., 2008; Petty, 2014, Kakkanat, 2017; Totsika, 2011). Colicins have also been proposed to be virulence factors in UPEC strains (Sharp et al., 2019).

1.9.2 *Escherichia coli* UTI89

The strain was kindly provided by Dr. Swaine L. Chen. The complete genome sequences were generated using multiple sequencing platforms and recently have been published on GenBank (Fenlon et al., 2020).

UTI89 is a clinical isolate from urine of a patient suffering from an acute bladder infection. The strain has serotype of O18:K1:H7 and it contains a pUTI89 plasmid (Li et al., 2012) which encodes many UPEC virulence pathogenicity factors. Furthermore, the pUTI89 plasmid can be divided to two major sections: One section contains genes involved in replication and conjugative DNA transfer, while the other contains genes present on the EIEC virulence plasmid and other

ORFs present on the chromosomes of a number of pathogenic bacteria, as shown in figure 6. The addition of these genes to the plasmid is likely to play a role in UTI89 pathogenesis. These genes encode enterotoxin, type III secretion systems, or/and adhesive structures (pili) (Cusumano et al., 2010). The type III secretion systems are inner and outer membrane protein channels, found in many pathogens, that deliver unique virulence factors to host cells.

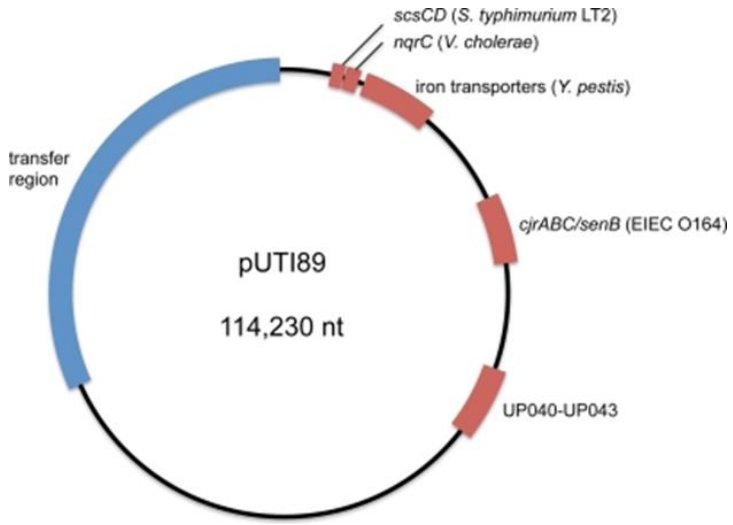


Figure 6. Diagram of pUTI89.

The blue highlighted genes encode proteins that are involved in conjugative DNA transfer. The red highlighted parts represent genes homologous to those found on virulence plasmids of other pathogenic *E. coli* (EIEC O164), and genes homologous to chromosomal genes in pathogenic bacteria such as *Salmonella enterica* serovar Typhimurium LT2, *V. cholerae*, *Y. pestis*. The figure is taken from (Cusumano et al., 2010).

1.9.3 *Escherichia coli* MG1655

The laboratory strain used in this study is a "wild-type" *E. coli* MG1655, serotype: OR:H48:K-, which has a genotype listed as F -, lambda- , *rph-1*. The strain was cured of the temperate bacteriophage lambda (by ultraviolet light), then the F plasmid (by acridine orange). These treatments resulted in a frameshift mutation at the end of *rph-1* (1 bp deletion). *rph* = enzyme truncated RNase PH, involve in phosphorolytic exoribonuclease in *E. coli* that participates in tRNA maturation by removing nucleotide residues following the -CCA terminus of tRNA. This frameshift caused a low expression of downstream gene *pyrE,* caused pyrimidine starvation phenotype.

The strain is descended from a strain that was isolated in 1922, originally isolated from stool of a convalescent diphtheria patient. The genome sequence of this strain was published in 1997, and it is very widely used as laboratory model (Blattner et al., 1997).

1.10 Transposable elements Tn5

Bacteria have served as a model to study transposition. Transposition is a powerful genetic tool to investigate and understand many features of the bacterial lifecycle (Goryshin and Reznikoff, 1998). Transposons are DNA fragments that can move from one location to another in the genome by a process called transposition. If the transposon insert in a gene, it will usually inactivate the gene. The transposon Tn5 and its derivatives consists of two key elements (a) an antibiotic resistance marker (b) flanked by 19 bp inverted repeat which is recognized by the

transposase. Most derivatives of Tn5 used experimentally are “mini-Tn5” with a smaller size than the complete transposon. Mini-Tn5 derivatives are commercially available bound to a modified mini-Tn5 transposase with high transposition efficiency. The mini-Tn5-transposase complex is stable and can be electroporated into cells where it is activated by the intracellular Mg^{++2} resulting in random insertion of transposon in the host genome (Lucigen, 2016). Transposon insertions are stable once made as the transposase cannot be passed on to daughter cells as they grow and divide. The mechanism of transposition is known as “cut and paste”, in which the transposase recognizes and binds the 19 bp terminal inverted repeat in the mini-Tn5. The transposon insertion starts with the transposase making a 9 bp staggered cut in the DNA target site making two sticky ends of the double DNA strand. Then the transposon is ligated by DNA ligase from the 3’ hydroxyl ends to the 5’ phosphate end into the sticky ends. The DNA polymerase fills the gaps at the sticky ends. This results in short direct 9 bp repeat of the target DNA, shown in figure 7 (Reznikoff, 2008; Goryshin and Reznikoff, 1998).

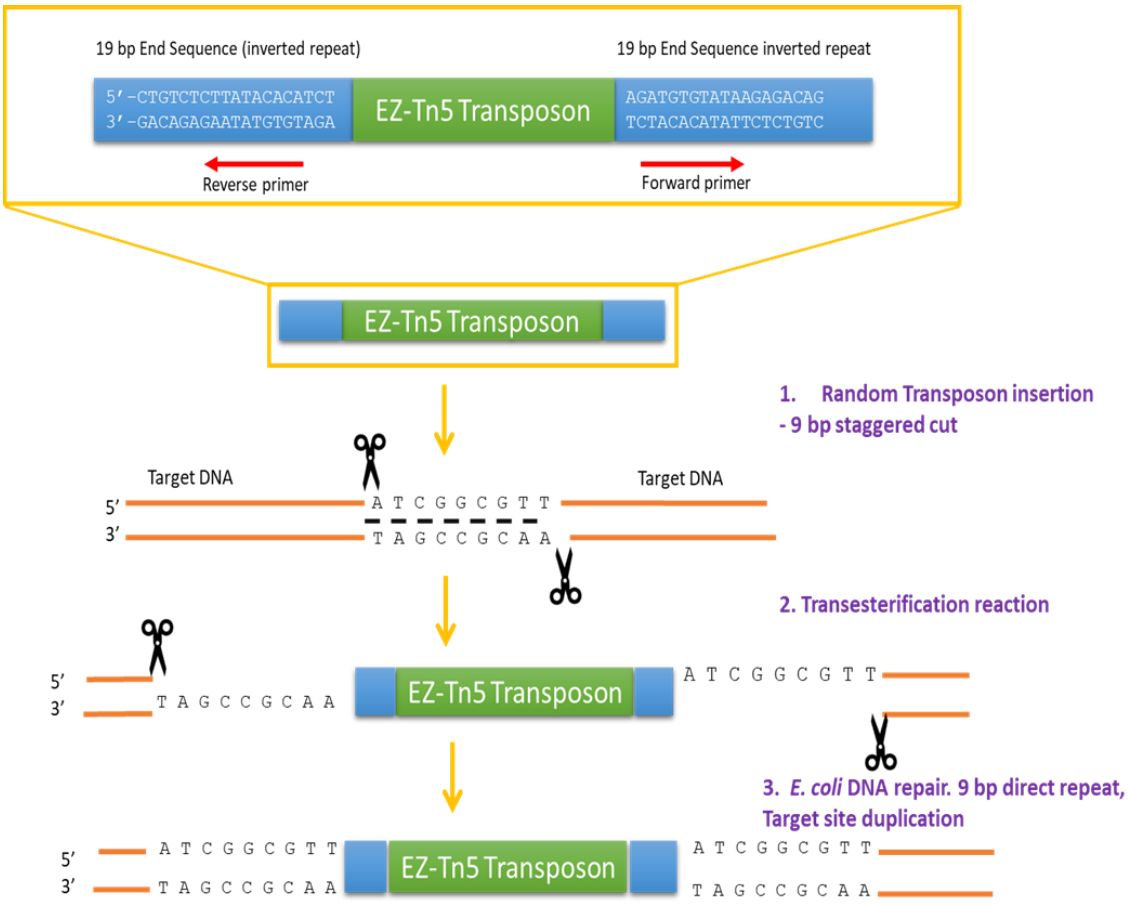


Figure 7. Tn5 Transposition mechanism.

The transposon structure shown in the yellow box which contains, the DNA sequence of resistence marker and 19 bp inverted repeat on each end. The orange strands represent the genomic DNA target. The scissors represent the site of the staggered cut.

First, the transposase makes a 9 bp staggered in the target DNA. The 5’ end of the transposon is ligated to the 3’ of the cut DNA sequence and the 3’ end of the transposon is ligated to the 5’ end of the DNA sequence. Third, the gaps in the ssDNA are filled by DNA replication, which results in the formation of flanking short direct repeats of the target site.

1.10.1 Transposon directed Insertion Site Sequencing (TraDIS)

Recently, transposon insertion sequencing methods have been developed, which allows genome-wide analyses to identify of the bacterial fitness contribution and the essentiality of the genetic components. Several studies have used transposon insertion sequencing methods to determine genes required for viability and optimal fitness under selective conditions; these are referred to as essential genes (genes are absolutely required for organism to growth or survive in standard lab conditions) or conditionally essential genes (genes required for growth under one growth condition but not another).

The method has been used in hundreds of different bacterial species which has led to new discoveries including gene function and regulation, functional non-coding RNAs, understand molecular pathways of pathogenesis, characterizing antibiotic mechanisms of action and identifying drug and vaccine targets (van Opijnen and Levin, 2020). Figure 7 shows the workflow typical of transposon-insertion sequencing methodology (Van Opijnen and Camilli, 2013; Barquist et al., 2013).

For example, studies on transposon insertion sequencing have identified virulence genes in different bacteria such as *Salmonella enterica* (Hensel et al., 1995), *Staphylococcus aureus* (Mei et al., 1997), *Vibrio cholerae* (Chiang and Mekalanos, 1998) and *Streptococcus pneumoniae* (Lau et al., 2001). Moreover, a number of bacterial strains were screened for essential genes such *Mycobacterium tuberculosis* (Griffin et al., 2011), *S. pneumoniae* (van Opijnen and Camilli, 2012), *E. coli* K12 (Goodall et al., 2018) and *Yersinia pseudotuberculosis* (Willcocks et al., 2018).

Additionally, Cyanobacterium strain *Synechococcus elongatus library* was used to investigate essential genes, essential intergenic regions and regulatory elements (Rubin et al., 2015).

TraDIS is a new technology which combines transposon insertional mutagenesis with massively parallel sequencing of the transposon insertion sites. As described above, one use of TraDIS is to identify essential genes in a genome. In the essential genes the transposon insertion is not tolerated, so essential genes are defined by the absence of transposon insertion or a very few insertions (Goodall et al., 2018). Also, TraDIS can be used to identify genes and networks which are involved in cellular survival and fitness in certain growth conditions. This experiment is done using a pooled transposon library containing millions of mutants, and allows linkage of genotype to phenotype in a high-throughput manner. This method allowed us to measure the fitness of many mutants by quantifying all the transposon insertions of each gene in each growth condition. Bacterial fitness is a measure of how well a particular cell replicates compared to other cells with different genotype in any given environment (van Opijnen and Levin, 2020).

The basic workflow for Transposon sequencing (Tn-seq) and transposon directed insertion site sequencing (TraDIS) is as follows:

(1) T**ransposon mutagenesis:** a bacterial transposon (derived from Tn5) that encodes Kanamycin resistance and has been adapted for laboratory use is used to create a library of random transposon mutants with a high frequency of insertions.

(2) Pool construction: after the transposon has been introduced into competent cells, they are cultured on suitable selective media for the chosen transposon then the colonies are picked and pooled.

(3) Determination and quantification of transposon-insertion junctions: A sample is taken from the 'input pool' for genomic DNA extraction from the pooled mutant bacterial cells. A sample is also taken after growth of the library under a specific condition of interest, this is the 'output pool'. DNA from the input and the output pools are then cleaved either by a specific enzyme or via physical shearing, followed by ligation of sequencing adapters. Then PCR amplification is used to selectively enrich for the end of the transposon and the sequence into which it has inserted. This is achieved by a transposon specific primer and a sequencing adaptor-specific primer. The amplified transposons insertion junctions are then sequenced to determine the location and number of insertion junctions, and changes in frequency between the input and output DNA pool can then be determined. A requirement of the method is that the transposon inserts are evenly distributed with limited sequence bias (Shevchenko et al., 2002). Also, another key point is that libraries should contain a high number of insertions, so that smaller genes are not missed in the analysis (van Opijnen and Levin, 2020).

Analysis of the sequencing output: Transposon insertion frequencies are determined in both the input and the output pool. The ratio of these is taken to be a measure of the contribution of each gene to bacterial fitness in the particular condition under which the output pool was generated. The calculation of fitness value is discussed in the relevant chapters 3 and 6. In general, the fitness effects can be captured, including neutral, negative, positive or conditionally

essential. Under the tested condition, if the total number of reads of all transposons in a given gene does not change significantly in comparison to the input DNA, this indicate that loss of the gene has little effect on the fitness (neutral) of the strain under these conditions (fitness = 1). This means the specific gene doesn't contribute significantly to fitness in the examined condition. If the ratio of reads for a given gene after the stress is < 1 (negative), this shows the organism that carries the insertion in that gene is less fit under the specific condition, and so that the gene contributes to that fitness in that condition. And if the ratio of the reads for a particular gene increase after the stress > 1 (positive), this shows the organism that carries the insertion in this gene is more fit under the specific stress condition. The conditionally essential genes can be identified when the transposon insertions disappear from the population in specific environment.

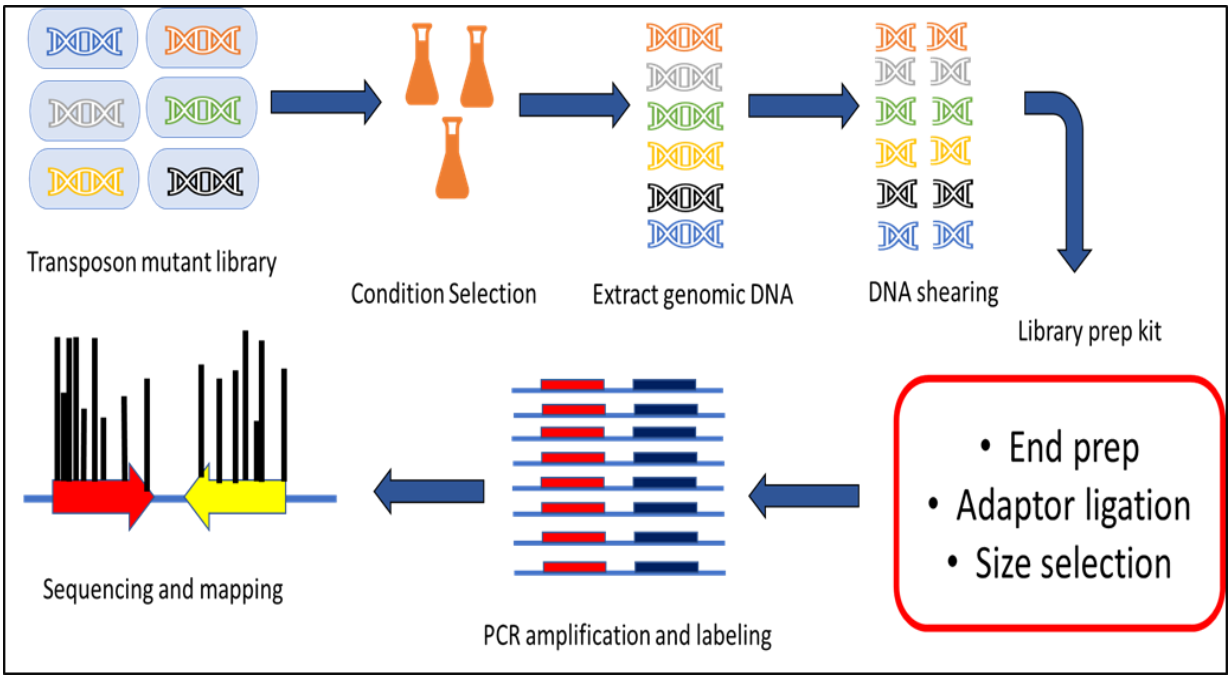


Figure 8. The workflow of transposon-insertion sequencing procedures.

The arrows indicate the procedure flow. The method starts with transposon library construction by transforming in a bacterial transposon that encodes for an antibiotic resistance marker to generate a library of random transposon mutants with high frequency of insertions. The library is then subjected to a selective condition. Cells are then collected for genomic DNA extraction, followed by DNA shearing where DNA is broken into shorter fragments by sonication. Subsequently, the library prep is done using a commercial kit start with end prep, adaptor ligation and size selection. The prepped DNA fragments are PCR amplified with transposon specific primer and a sequencing adaptor specific primer. Finally, the amplified fragments are sequenced, and the sequencing reads are mapped to the bacterial genome showing the frequency and the location of insertions.

1.11 Work leading to this project

This research project originated from a previous BBSRC-funded project run in collaboration with Liverpool university. Dr. Francesca Bushell and Dr. Thippesh Sannassidappa used using a transposon mutant library (Tn5) in *Escherichia coli* UPEC EO499 which belongs to serotype ST131 and which was gifted from Dr. Keith Turner, Discuva.

Basically, this TraDIS library in *E. coli* EO499 was grown under 35 different conditions including pH 7, pH 5.5, different types of weak organic acids (e.g. propionic acid, butyric acid, acetic acid), urea, ethanol, and bile salts, with all experiments being done both aerobically and anaerobically. The cultures were grown in 250 ml flasks containing 50 ml of M9 medium supplemented with 0.2% casamino acids, 0.2% glucose, MOPS and MES. MOPS and MES are a buffering agent in biological assays. MOPS buffers over the range pH 6.9 – 8.3, while MES buffers over the range pH 5.2 – 7.1.

In this project I focus strictly on acetic acid stress. In acetic acid stress, the EO499 library was grown either at pH 7 in the presence of 40 mM acetic acid or at pH 5.5 in the presence of 4 mM acetic acid. These conditions were chosen based on a number of factors. First, the cells were mildly affected so that the final optical density of both the test and control condition (with and without acetic acid added) was not too different: in other words, the addition of acetic acid slowed bacterial growth by about 10% but did not stop it. Second, Dr. Francesca performed RNAseq experiments under the same conditions and these experiments were repeated under anaerobic conditions. The anaerobic condition added a further stress on the growth, which made it difficult for bacteria to grow and reach the desired optical density. These parameters were

chosen to obtain good level of overall growth in both experiments and maintain the same stress in both TraDIS and RNAseq.

Each medium was inoculated with 10 µl of EO499 initial sequencing library which is approximately an OD_{600} of 0.01 (8 x 10^6 cells/ml). The cultures were grown at 37 °C and rotated at 180 rpm for 24 hours. Each experiment was carried out in biological triplicate. Next day, 10 ml of the culture was centrifuged for 5 min at 4000 rpm and the pellet was frozen at – 20 C. The genomic DNA was extracted with Qiagen DNeasy Blood and Tissue kit according to the guidelines provided (Bushell, 2019).

TraDIS data sequencing was carried out at Liverpool University Centre for Genome Research and data analysis was performed by Dr. John Herbert. The sequencing data was processed using the ESSENTIALS software for rapid analysis of high throughput transposon insertion sequencing data. The software was used to predict genes that showed significantly altered fitness profiles when mutated in the presence of acetic acid by calculating the false discovery rate (FDR) and the adjusted p- value significant values of each gene (Zomer et al., 2012).

The analysis suggested a list of non-essential genes that have an impact on fitness in the presence of acetic acid, that will be describe later in the results chapter 3.

1.12 Aim and objectives

The purpose of this project is to use TraDIS to identify genes which, when mutated, cause a decrease in fitness for cells in the presence of acetic acid. First, an attempt was made to validate

the results of the EO499 TraDIS data under acetic acid that have been generated as described above. The data obtained from Dr. John Herbert were analyzed to give ratios of Reads Per Kilobase Million (RPKMs) for each gene before and after acetic acid treatment. The application of TraDIS assists the classification of genes to three categories: essential genes, non-essential genes and unclear genes under the examined condition. The non-essential genes with lower relative fitness index were selected for validation to see whether they did contribute to fitness under organic acid stress. Eight mutants with fitness defect according to TraDIS were chosen to re-expose the mutant to the conditions, to measure the survival of the mutant relative the wild type strain. This was attempted using two different methods:

a) Strains containing knockouts of the genes in the list were examined under the conditions used to generate the TraDIS data. Ideally these knockouts would be in strain EO499, but as making knockouts in EO499 strains is difficult and attempts to do this in the laboratory were not successful, I have used knockouts derived from the Keio gene knockout library of *E. coli* BW25113 (Baba et al., 2006). This enabled us to determine whether TraDIS in one strain of *E. coli* can be validated by using knockouts in a different strain. Also, this is to determine the fitness of these knockouts relative to the wild type under the different stress conditions using competition experiments. I have expected to see loss of relative fitness index in the selected genes in the presence of acetic acid.

b) Secondly, I have attempted to develop a PCR-based method to isolate individual transposon mutants from the EO499 initial transposon sequencing library and to validate the

phenotypes of these mutants in the presence of acetic acid. Nested PCR method was used with two transposon specific primers and two flanking primers. After the long time spent on trying to validate TraDIS genes list I have decided to replicate the EO499 TraDIS data under acetic acid stress over 5 days time course.

I have also aimed to compare three transposon mutant libraries at pH 5.5 with 4 mM acetic acid by TraDIS over a time series; UPECs strains (EO499 and UTI89) and a lab strain MG1655. Thus, this will provide us with clearer view of the genes involve in acetic acid at pH 5.5 and needed for the optimum growth. Also, this will indicate if different *E. coli* strains will have the same organic acids targets. Additionally, I am trying to examine if the lab strain can behave differently from pathogenic variants of the same species. Moreover, reconstruction of EO499 TraDIS will allow us to learn if I can achieve the result by replicating the experiment by two different labs. This was done by:

a) Construction of UTI89 transposon mutant library (ITL – initial transposon library). This started with optimization and troubleshooting for high transposon transformation efficiency. EO499 (constructed by Keith Turner) and MG1655 (constructed by Dr. Mathew Milner) ITLs were already available in the lab. The libraries were sequenced by Miseq and analysed by TraDIS data processing pipeline. The three genomes were annotated by Prokka annotation pipeline. And Rory pipeline was used to determine if gene is present or absent among the three strains. Then I have performed a comparison between the three strains and grouped the genes to determine core genome, accessory genome and unique genome. Then, analysing and comparison of the three

transposon mutant libraries to determine essential genes, non-essential and unclear genes. The essential genes list were used to exclude them from TraDIS under acetic acid analysis. I have expected to find large correlation between UPECs libraries than MG1655.

b) UPEC strains and MG1655 were subjected to pH 5.5 with and without acetic acid stress over a time course of five days and each condition was carried in duplicate. I have sequenced two-time points day one and day three. TraDIS libraries were sequenced by Miseq and analysed by TraDIS data processing pipeline to identify genes required for fitness under acetic acid stress. Next, the $\log_2$ fold changes for the pH 5.5 with acetic acid to pH 5.5 with the FDR and P-value were measured by EdgeR pipeline. The results shown in the chapter with the different approaches taken to the analysis.

2 Materials & Methods

2.1 Bacterial strains and growth conditions

2.1.1 List of strains

Table 3. List of the strains used in this study.

Strain	Genotype	Source
E. coli K-12 MG1655	*F- lambda- ilvG- rfb-50 rph-1*	Lab stock
E.coli EO499 ST131	EO499 *E. coli* associated with CTX-M-15 extended-spectrum β-lactamase resistance gene. Clinical isolate harbouring sequenced plasmid pEK499.	Keith Turner, Discuva
E.coli EO499 *ytfP*::Tn5		Dr.Francesca Bushell Thesis
E.coli EO499 *rssB*::Tn5		This study
E. coli K-12 BW25113	*F- λ- rrnB3 ΔlacZ4787 hsdR514 Δ(araBAD)567 Δ(rhaBAD)568 rph-1*	(Baba et al., 2006)
BW25113 Δ *nuoM*::kan^R	*E. coli* K-12 BW25113 with *nuoM* deletion	(Baba et al., 2006)
BW25113 Δ *sucA*::kan^R	*E. coli* K-12 BW25113 with *sucA* deletion	(Baba et al., 2006)
BW25113 Δ *nuoG*::kan^R	*E. coli* K-12 BW25113 with *nuoG* deletion	(Baba et al., 2006)
BW25113 Δ *sthA*::kan^R	*E. coli* K-12 BW25113 with *sthA* deletion	(Baba et al., 2006)
BW25113 Δ *pitA*::kan^R	*E. coli* K-12 BW25113 with *pitA* deletion	(Baba et al., 2006)
BW25113 Δ *apaH*::kan^R	*E. coli* K-12 BW25113 with *apaH* deletion	(Baba et al., 2006)
BW25113 Δ *ytfP*::kan^R	*E. coli* K-12 BW25113 with *ytfP* deletion	(Baba et al., 2006)
BW25113 Δ *rssB*::kan^R	*E. coli* K-12 BW25113 with *rssB* deletion	(Baba et al., 2006)
BW25113 Lac$^+$	*E. coli* K-12 BW25113 with *lacZ* insertion	This study
Uropathogenic *Escherichia coli* (UPEC) UTI89 wild type		Swaine L. Chen, Singapore

2.1.2 Growth media, supplements and buffers

The strains were grown in lysogeny broth medium (LB) containing 10 g/l tryptone, 10 g/l NaCl and 5 g/l yeast extract. The overnight cultures were always inoculated in 5 ml broth in a 20 ml universal bottle. Selected liquid cultures were grown in 250 ml Erlenmeyer flasks with 1:20 ratio culture to flask volume. LB agar + 1.5 % (w/v) bacteriological agar was used for plating, if required. Cultures were incubated at 37 ˚C with continuous shaking at 180 rpm. The cultures were supplemented with 30 μg/ml Chloramphenicol (Cam), or 50 μg/ml Kanamycin (Kan), 100 μg/ml Ampicillin (Amp.

M9 minimal medium was used for TraDIS, and competition experiments. The medium contains 42.3 mM di-sodium hydrogen orthophosphate, 22.1 mM potassium dihydrogen orthophosphate, 8.56 mM sodium chloride, 18.7 mM ammonium chloride and supplemented with 22.2 mM D-glucose, 0.2% w/v cas-amino acids, 100 mM MOPS, 100 mM MES hydrate, 2 mM magnesium sulphate 7-hydrate and 0.1 mM calcium chloride. The pH was adjusted with NaOH or HCl when required.

P1 Luria-Bertani Medium Broth (P1 LB) was used for P1 transduction, containing 10 g/l tryptone, 5 g/l yeast extract, 10 g/l NaCl, 0.2% glucose and 5 mM $CaCl_2$.

2.1.2.1 1 M Tris-HCl Buffer Stock Solution

For 1 L of Tris-HCl, 121.14 g of Tris (American Bioanalytical #AB14042) was added into 800 ml dH_2O. The pH was adjusted pH to 8.0. The final volume was brought to 1 liter with deionized water. Then, the buffer was autoclaved and store at room temperature.

2.2 Growth conditions

2.2.1 Competition experiments and measurements of fitness

The wild type strain was competed against the appropriate mutant strain in the same condition. The two competitors were distinguished based on their lactose fermentation phenotypes; Lac^+ and Lac^- cells produce pink and red colonies, respectively, on MacConkey agar plates supplemented with 1% w/v lactose. Cultures of wild type and mutant strains were grown separately at 37° C and 180 rpm overnight. These cultures were mixed at OD_{600} of 0.025 to give a final OD_{600} of 0.05 in 5 ml of supplemented M9 minimal media. This was done at pH 7.0 or 5.5 and in the presence or absence of acetic acid. In addition, the mixed cultures were serially diluted and 50 µl of 10^{-2} and 10^{-3} dilutions were spread on lactose MacConkey agar, in order to quantify the colonies by counting after the incubation and to differentiates the lactose ferments form the non lactose ferments based on their colors red and white, respectively. The plates were incubated overnight at 37 ˚C. The competition mixtures were then incubated at 37 °C, 180 rpm for 24 h. Most experiments were done for 24h, but some additional experiments were done over three days with fresh medium dilutions every 24h. 1:10 dilutions of 10^{-5}, 10^{-6}, 10^{-7} were then plated

onto lactose MacConkey agar. The numbers of red and white colonies were counted after overnight incubation of the plates. Assays were performed 3 to 9 times.

Richard Lenski's (Wiser and Lenski, 2015) formula for relative competitive index is given as

$$w = \frac{K_A}{K_B} = \frac{\ln\left(\frac{A_f}{A_i}\right)}{\ln\left(\frac{B_f}{B_i}\right)}$$

where A and B are the two strains, f is the final population size and i is the initial population size.

Means and standard error of mean for all experiments were determined. For Statistical analysis, Statpages website (http://statpages.info/anova1sm.html) was used to determine whether there were significant differences between the tested means by one-way ANOVA. The test required means and standard deviations (SD) for each examined condition. A P-value < 0.05 showed a significant difference between the tested groups. This was followed by Tukey post-hoc test to show which groups were significant from others.

2.3 P1 transduction

P1 transduction was used to transfer the *lacZ* gene between different *E. coli* strains using a modified standard protocol (Thomason et al., 2007). To make P1 lysate donor strain *E. coli* MG1655 was grown in 5 ml of P1 Luria-Bertani Medium Broth. The overnight culture was diluted

1: 100 into 5 ml P1 LB and incubated further with shaking for 45 min at 37 °C. Then, the culture was infected with 100 µl of recently prepared P1 lysate and continued the incubation with shaking for additional 3 hours. Few drops $CHCl_3$ were added and continued shaking for another few minutes to ensure cell lysis and killed bacteria. The supernatant was transferred into 50 ml of centrifuge tube and centrifuged 10 min at∼9200 ×g, 4 ˚C. The supernatant was transferred to 5 ml screw-cap tube labelled with donor strain and date and stored at 4 °C.

For P1 transduction, 1.5 ml of an overnight *E. coli* BW25113 recipient culture was centrifuged in microcentrifuge tube for 2 mins at maximum speed at room temperature. The supernatant was discarded, and the pellet was resuspended with 750 µl of P1 LB medium. Then, 100 µl of the cells was mixed with different volume of the P1 lysate (1, 10 and 100 µl) in sterile microfuge tube. For a control, 100 µl of recipient cells was used without P1 lysate. The tubes were incubated at room temperature for 30 min, to allow phage to adsorb the cells. After that, 1 ml of LB broth and 1 M sodium citrate was added to each tube, the sodium citrate removes the calcium essential for phage adsorption to the bacteria. Then, the tubes were incubated further for ~ 1 hr at 37 °C with shaking. Each culture was centrifuged for 2 min at maximum speed, RT. Pellet was removed and the cells were resuspended in 100 µl of 1x PBS + 5 mM sodium citrate. The resuspended culture was diluted and 100 µl was plated into the appropriate media, MacConkey agar and incubated at 37 °C for 24 hrs.

Next day, The BW25113 lactose fermenter, red colonies were selected as a positive transduction. The Lac^+ colonies were confirmed by using colony PCR with primers using LacZ-1-F, LacZ-1-R, LacZ-2-F, LacZ-2-R and LacZ-3-F, LacZ-3-R (refer to in primers supplementary table S1.)

2.4 Molecular methods

2.4.1 Preparation of DNA for a colony PCR

For colony PCR, a single colony was cultured in 5 ml of LB overnight. The next day, 10 µl of the culture was boiled at 98 ˚C for 10 min and then cooled for 10 mins at 4 ˚C in a PCR machine. Then, 90 µl of sterile distilled water was added to the sample. 2 µl was then used as a DNA template for the PCR reaction.

2.4.2 Amplification of isolated DNA

Primers used in this study are listed in table S1. Polymerase chain reactions (PCR) for amplifications were carried in 25 µl reaction mixture volume containing 2µl of template DNA, 2 µl (20 µM) of each primer, 12.5 µl of 2x MyTaq Red Mix (Bioline), 0.5 µl of 50 mM $MgCl_2$ (BioLabs) and water was used to make up the final reaction volume. The general thermocycle condition that were used are shown in table 4.

2.4.3 Agarose Gel Electrophoresis

The amplified products were separated on 2 % agarose gel (Bioline) gels in 1x TAE buffer containing 50x TAE buffer = 2 M Tris, 1 M acetic acid, 0.05 M EDTA in water, and stained with Midori Green (Nippon Genetics). Hyperladder 1 kb or 100 bp (Bioline) was used as a marker to

confirm the DNA fragment size. The agarose gels were run at 120 V for 60 minutes. Bands were visualized using Syngene Gene Snap system.

2.4.4 Primers list

Primers, inline barcodes and transposon amplifying primers used in this study are listed in the supplementary data, Table S1.

Table 4. General thermocycling conditions used for PCR.

Cycles	Step	Temperature	Time
1	Initial denaturation	95˚C	2 minutes
30	Denaturation	95˚C	30 seconds
	Annealing	Range 52˚C to 66˚C refer to table S1, (supplementary Data)	2 minutes
	Extension	72˚C	2 minutes
1	Final extension	72˚C	1 minutes
∞	Refrigerator	4˚C	∞

2.4.5 DNA extraction by boiling method

20 µl – 1 ml of overnight bacterial culture (OD_{600} between 3.0 and 4.0) were transferred into a fresh 1.5 ml Eppendorf tube and mixed with 1 ml of 1x phosphate buffered saline (BR0014, Oxoid Ltd., UK). The cells were centrifuged for 5 min at 12,000 rpm. The supernatant was discarded, and the pellet was resuspended with 200 µl sterile water. The tube was placed in a pre-

heated block (98°C) and held for 10 min and then shifted to ice-bath for 5 min. The sample was centrifuged for 10 mins at 12,000 rpm and the supernatant was transferred to fresh Eppendorf tube. Around 2 – 4 µl of the supernatant was used as a DNA template for PCR screening.

2.4.6 PCR purification protocol

PCR products were purified by Qiagen QIAquick PCR purification kit (cat.# 28104), the procedures were followed as manufacturer recommendations.

2.4.7 Sample verification by sequencing

Amplified DNA samples were sent to Source Bioscience, Nottingham, United Kingdom to ensure correct sequences and orientation.

2.5 Isolating mutants from the TraDIS library

The desired transposon insertion mutants were isolated from a TraDIS library using a PCR screening method. The library was diluted to 2500 cells per ml. In each well of a 96 well microtiter plate, 200 µl of cells were added. Column 12 and row H were left empty. The plate was then incubated overnight with the lid on at 37°C without shaking. After the incubation, the rows and columns were pooled vertically and horizontally into the empty wells by taking 4 µl from each well (total volume 44 µl in column 12 and 28 µl in row H) as illustrated in figure 9. Next, the genomic DNA was extracted from each pooled sample by the boiling method.

PCR screening was done with the genomic DNA of the pooled cells using a transposon specific primer and primer binding to the flanking region of the target gene. For pools where amplification was positive, the location of the well containing the mutant in the 96-well plate was determined as depicted in figure 10. Culture from the positive well then was diluted down in another 96 well plate to 50 cells per well leaving column 12 and row H empty. The same process of pooling and PCR was performed again. Then, culture from the positive well was spread on LB agar plates and the resulting colonies were selected for individual colony PCR to identify the colony with desired mutant. The positive colony was cultured in LB and stored with 15% glycerol in -80 °C for further analysis. This method will be explained in detail with the optimization in chapter 4.

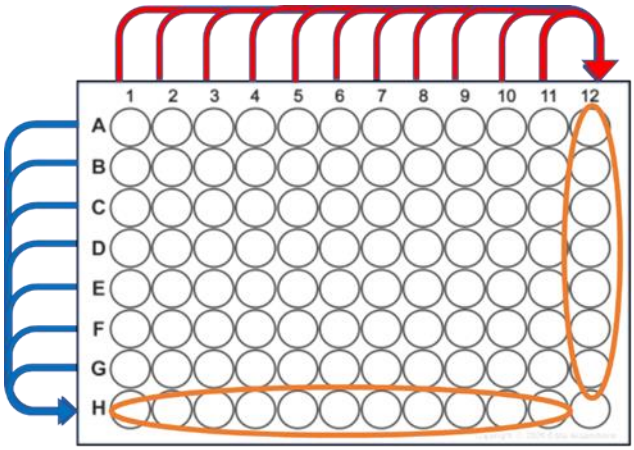


Figure 9. The pooling pattern of the EO499 library in the 96 well plate.
The red arrows showed cultures in rows pooled horizontally to column 12. The blue arrows showed cultures in columns pooled vertically in row H.

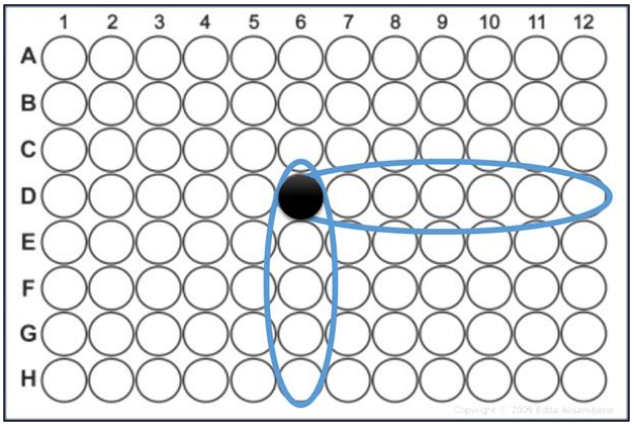


Figure 10. An example of the way to coordinate the location of the desired mutant after PCR. If a positive band appeared in D12 and H6, then the mutant is located in well D6.

2.6 UTI89 TraDIS library construction

Constructing a UTI89 transposon library involves preparing competent cells, constructing a mutagenesis transposon library, library preparation for sequencing and library sequencing. The workflow of the procedures is presented simplified in figure 11. These steps will be discussed in detail in the following sections:

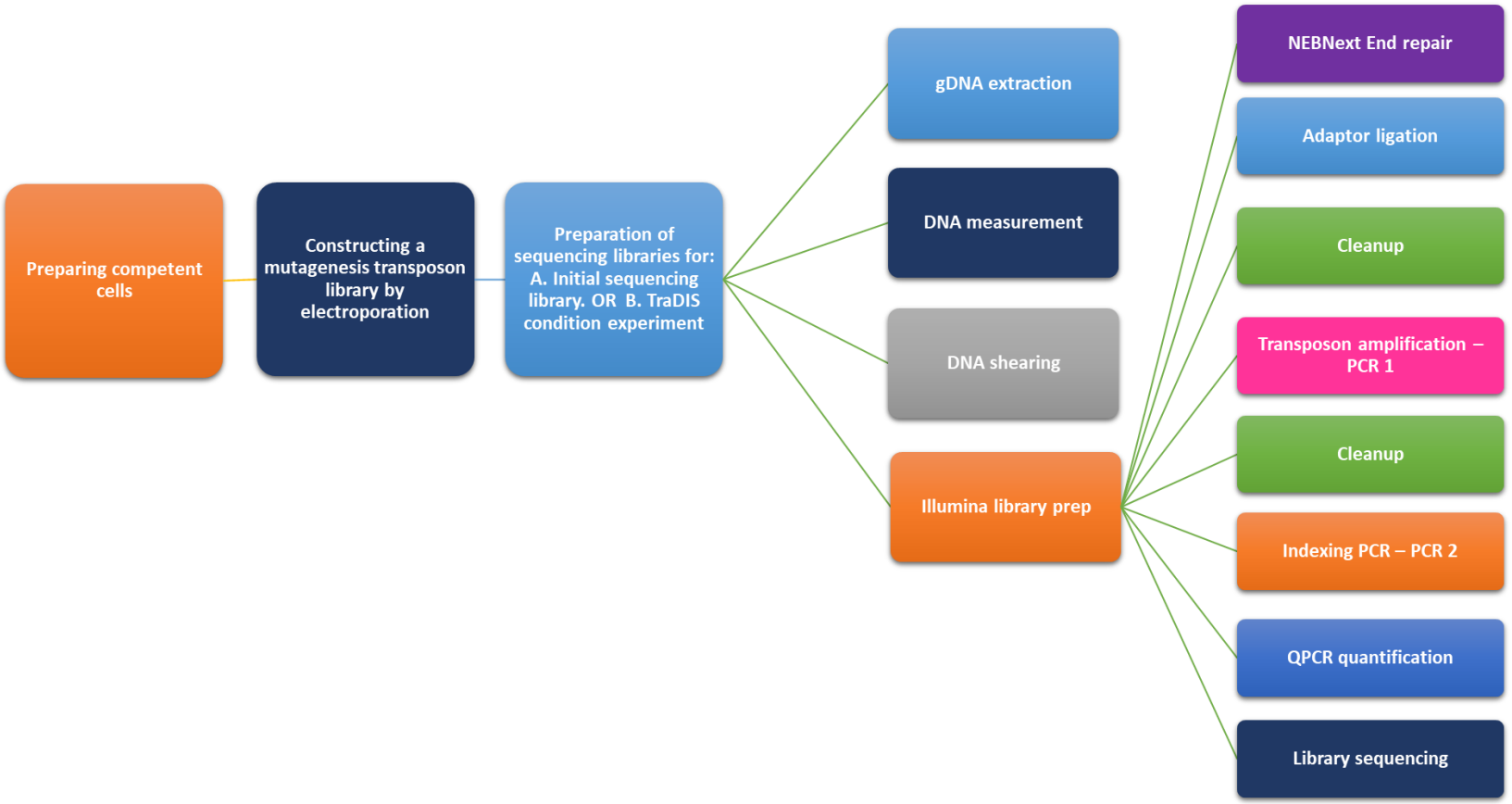


Figure 11. The workflow of constructing a transposon library and sequencing.

Each box represents a step and the grey arrows represent the direction of the procedures.

2.6.1 Preparing competent cells

The library construction approach used was essentially the same as in (Langridge et al., 2009). All the glassware was washed thoroughly in advance with soap and extensively rinsed with distilled water to ensure there was no residual detergent. Then the glassware was autoclaved for sterilization. The Lysogeny broth or agar medium used in this protocol were prepared by me rather central services. A single colony was inoculated in 250 ml shaking flask containing 20 ml of LB for overnight culture. On the next day, 7 ml of the overnight culture was inoculated into 2 flasks of 2 liters containing 800 ml of prewarmed 2x YT broth. 2x YT medium contains 15 g/l tryptone, 10 g/l yeast extract and 5 g/l sodium chloride with a final pH 6.8. The medium pH was adjusted with few drops of 1mM of sodium hydroxide. The cells were grown at 37 °C, 180 rpm until an OD_{600} of ~ 0.2 – 0.3. Once the cells reached the desired OD, the cultures were decanted into pre-chilled 50 ml Falcon tubes placed in ice. 600 ml of the culture was aliquoted into 24 of 50 ml Falcon tubes. The tubes were incubated in ice for 30 mins. The cultures were centrifuged at 5000 *g* for 10 mins in a chilled centrifuged at 4 °C. The supernatants were discarded, and the pellets were resuspended gently in 25 ml of 10% ice cold glycerol. Cultures were combined resulting in 12 Falcon tubes. The cells were centrifuged, and pellets were washed with 25 ml of 10% ice cold glycerol. The cultures were combined again resulting in 6 Falcon tubes. The tubes were centrifuged, and pellets were washed with 15 ml of 10% ice cold glycerol. Each two cultures were combined into two flacon tubes resulting in 3 tubes. The tubes were centrifuged, and pellets were washed with 1 ml of 10% ice cold glycerol and combined in 1 Falcon. The cells were

centrifuged and resuspended in 1 ml of 10% ice cold glycerol. Finally, 60 µl of the competent cells were aliquoted in ice cold 1.5 ml Eppendorfs. The competent cells were stored in -80 °C.

2.6.2 Transposon mutagenesis library

For transformation, 0.2 µl of transposon (cat. No. TSM99K2, EZ-Tn5™ <KAN-2> Tnp Transposome™ Kits, Lucigen) was mixed with 60 µl aliquots of the competent cells and incubated on ice for 30 minutes. The cells were transferred into 2 mm electroporation cuvettes for electroporation using an Eppendorf Eporator set to 2.2 kV. Then immediately, 1 ml of pre-warmed SOC medium was added to the electroporated cells in the cuvette (cat No. SLBW4174, Sigma Aldrich). Twelve electroporation transformations were done. The samples were incubated in 37 °C at 180 rpm for 1 to 1:30 hours. The samples were diluted with 2.5 ml of LB broth and 250 µl was spread on square LB agar plates. Initially the first batch was done in round plate, subsequently changed using square plate for higher yield. The plates were supplemented with 25 µg/ml kanamycin. The plates were incubated at 37 °C overnight. The next day, around 500 colonies were counted on each plate. 500 µl of LB was added to each plate and colonies were resuspended with a plastic spreader. The resuspension was transferred by 1000 µl pipette to 50 ml Falcon tube placed on ice. After all the colonies had been resuspended and transferred to the Falcon tube, the tube was vortexed to ensure an even mixture. Sterile 50 % glycerol was added to the library to the concentration of 15 % (w/v). Finally, the library was divided in to 500 µl aliquots in 1.5 ml Eppendorf and stored in – 80 °C for further use and analysis.

2.6.3 TraDIS growth condition for sequencing

Three different transposon libraries of MG1655, EO499 and UTI89 were each exposed to pH 5.5 with and without acetic acid over five days. An OD_{600} of 0.01 of the three initial transposon libraries (MG1655, EO499 and UTI89) was added into a 250 ml flask containing 50 ml supplemented M9 at pH 5.5 or at pH 5.5 with 4 mM acetic acid. The medium was incubated at 37°C, 180 rpm for 22-24 hrs. OD_{600} of 0.1 was used for passaging the cultures. All experiments were done in triplicates for 5 days. Genomic DNA from day 1 and day 5 was extracted using RTP® Bacteria DNA Mini Kit. Two replicates of each condition were used for sequencing.

For archiving, 10 ml of the overnight culture was centrifuged at 4000 rpm for 5 mins at 4 °C and the pellet was stored at – 20 for further studies. Also, 1 ml of the culture was frozen with 15% glycerol and stored at – 80 °C for further studies.

2.6.4 Preparation DNA screening and isolation for sequencing libraries

2.6.4.1 gDNA extraction

Genomic DNA extraction was done using RTP® Bacteria DNA Mini Kit (1 x 10^9 bacteria cells). The kit, used according to the manufacturer's instructions, was designed to purify DNA for Gram negative bacteria. The filtrate was used as a DNA template for the sequencing library preparation.

2.6.4.2 DNA measurement

To quantify the DNA, a Qubit® Fluorometer was used. Qubit® dsDNA HS Assay Kit (cat. No. Q32851, Q32854) was used as the manufacturer's instructions.

2.6.4.3 DNA shearing

For DNA shearing, 500 µl of nuclease free water were added to a 15 ml Falcon tube (cat. # 11507411). An amount of DNA, calculated to reach a final concentration of 1 µg, was added to the tube. Genomic DNA was sheared to generate 200 bp fragments using a Bioruptor probe. The sonication probe was decontaminated with absolute ethanol and immersed in the sample. A blank containing 500 µl of water was used if needed. The samples were placed in the sonicator in balance with each other. The sonicator was set with the following parameters: 30 seconds on and 90 seconds off, low intensity, for 13 cycles. This method allowed 20 mins time intervals between multiple runs. The fragmented DNA was transferred into 1.5 ml Eppendorf tube and placed into a condenser at RT until the volume reached 50 µl. The final sample volume was made up to 55.5 µl with sterile water.

2.6.5 Sequencing library preparation

2.6.5.1 NEBNext End repair

The next step was done using a NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (cat. # E7370) with starting materials from 1 µg of fragmented DNA. In a PCR tube placed in an ice

block the following materials were mixed to reach a total volume 65 µl: 55.5 µl of fragmented DNA, 3 µl End prep enzyme mix and 6.5 µl End repair reaction buffer (10x). The PCR tube was placed in the PCR thermocycler at 20°C for 30 mins, then 65°C for 30 mins.

2.6.5.2 Adaptor ligation

For adaptor ligation, 15 µl blunt/TA ligase master mix, 2.5 µl NEBNext adaptor for illumina and 1 µl of ligation enhancer were added to the End prep reaction mixture from the previous section to a final volume of 83.5 µl. The ligation mixture was placed in the thermal cycler at 20 °C for 15 minutes.

Next, a 3 µl of USER™ enzyme was added to the ligation mixture and was placed in the thermal cycler at 37 °C for 15 minutes.

2.6.5.3 Size selection or cleanup of Adaptor-ligated DNA

The library was constructed with an average insertion size of 200 bp fragment. Magnetic bead was used for selecting the DNA fragments with specific size and removing the unwanted DNA from aqueous solution. The DNA fragment size selection can be controlled by varying the ratio of the bead added to the DNA solution. SPRI (Solid Phase Reversible Immobilization) - AMPure XP were made in the laboratory as described in DeAngelis, 1995 (DeAngelis et al., 1995). The AMPure XP beads were made of carboxylated paramagnetic combined with a buffer containing a polyethylene glycol (PEG) and salt. Paramagnetic meaning that they become

magnetized when an external magnetic field applied to prevent beads from falling out of the solution or clumping. The beads were first allowed to come to room temperature for half an hour. The AMPure XP beads were then vortexed to resuspend them thoroughly. In new Eppendorf, 13.5 µl of sterile water was added to 86.5 µl ligation reaction and 44 µl AMPure XP beads. The tube was incubated at room temperature for 5 minutes and placed on a magnetic stand for 5 minutes to sperate the beads from the supernatant. After the sample had cleared, the supernatant containing the DNA was transferred to a new Eppendorf tube and the remaining material was discarded. This process was repeated again with 20 µl of AMPure XP beads were added to the supernatant, but this time the supernatants containing unwanted DNA were carefully removed and discarded. 200 µl of 80 % of freshly prepared ethanol was added to the tube while in the magnetic stand and incubated for 30 seconds. After incubation, the ethanol was removed and discarded. The washing step with ethanol was repeated twice and residual ethanol was removed from the bottom of the tube without disturbing the beads, while the tube was still in place. The tube was removed from the magnetic stand and allowed to air dry for 5 minutes with the lid open. Care was taken to avoid over-drying, as this can result in a lower recovery of DNA. The target DNA was eluted with 17 µl buffer EB Qiagen (Cat No. # 19086) and incubated for 2 minutes at room temperature. The eluted DNA was still mixed with the beads, so it was placed on magnetic stand for 5 minutes to separate the DNA from the beads. After the solution was cleared, 15 µl of the supernatant was transferred into a new PCR tube for amplification. Less amount was aliquoted to avoid pipetting the beads.

2.6.6 PCR amplification of the transposon junction

To amplify the transposon junction for sequencing. , the following were added to 15 µl of the adapter-ligated DNA fragments from step b: 25 µl of NEBNext Q5 Hot Start HiFi PCR Master Mix (KAPPA), 2.5 µl TTC-SLXP1-F1 primer (labmade primer)-TKK R, 2.5 µl TTC-SLXP1-TKK F (kanamycin primer) and 5 µl water to reach a final volume of 50 µl. The PCR tube was placed in the thermocycler, with the following condition: Initial denaturation at 98 °C for 48 seconds and 10 cycles denaturation at 98 °C for 15 seconds, annealing at 65 °C for 30 seconds and extension at 72 °C for 30 seconds. The final extension was for 1 minutes at 72 °C and the sample was then held at 4 °C.

2.6.7 Cleanup following PCR Amplification

AMPure XP beads were vortexed and allowed to come to room temperature. The previous PCR reaction was transferred to a 1.5 ml Eppendorf, and 36 µl of AMPure XP beads were added to the PCR reaction and mixed well. The tube was incubated at room temperature for 5 minutes. Then the tube was spun for 30 sec and placed in the magnetic stand to separate the beads from supernatant for about 5 minutes. After the solution cleared the supernatant was carefully discarded. While the tube on the magnetic stand, the beads were washed with 200 µl of 80 % ethanol and incubated at room temperature for 30 seconds. Washing with ethanol was repeated, and the ethanol was discarded after. The sample was removed from the magnetic stand and the beads were allowed to air day with the lid open for 5 minutes. The DNA was eluted with 17 µl

buffer EB and incubated for 2 minutes at room temperature. Then the tube containing the beads and the DNA was placed in the magnetic stand and for 5 minutes. When the solution was cleared from the beads, 15 µl of the of the superannuant was transferred to a PCR tube.

2.6.8 The second PCR amplification- library preparation

The second PCR reaction contained 15 µl of adaptor ligated DNA fragments, 25 µl of NEBNext Q5 Hot Start HiFi PCR Master Mix, 2.5 µl inline index custom forward primer in table S2 supplementary data, 2.5 µl illumina index (cat. # E7335S) and 5 µl water for a total volume of 50 µl.

The PCR tube was placed in the thermocycler for: initial denaturation at 98 °C for 48 seconds and 20 cycles denaturation at 98 °C for 15 seconds, annealing at 65 °C for 30 seconds and extension at 72 °C for 30 seconds. The final extension was followed for 1 minutes at 72 °C and held at 4 °C. After the PCR reaction, the reaction was cleaned up as described in section 2.6.7 with a final elution volume of 33 µl EB and 32 µl was transferred to labelled Eppendorf. The prepared library was stored – 20 °C. Summarizing all the above steps for preparing the DNA fragment for sequencing is outlined in figure 12.

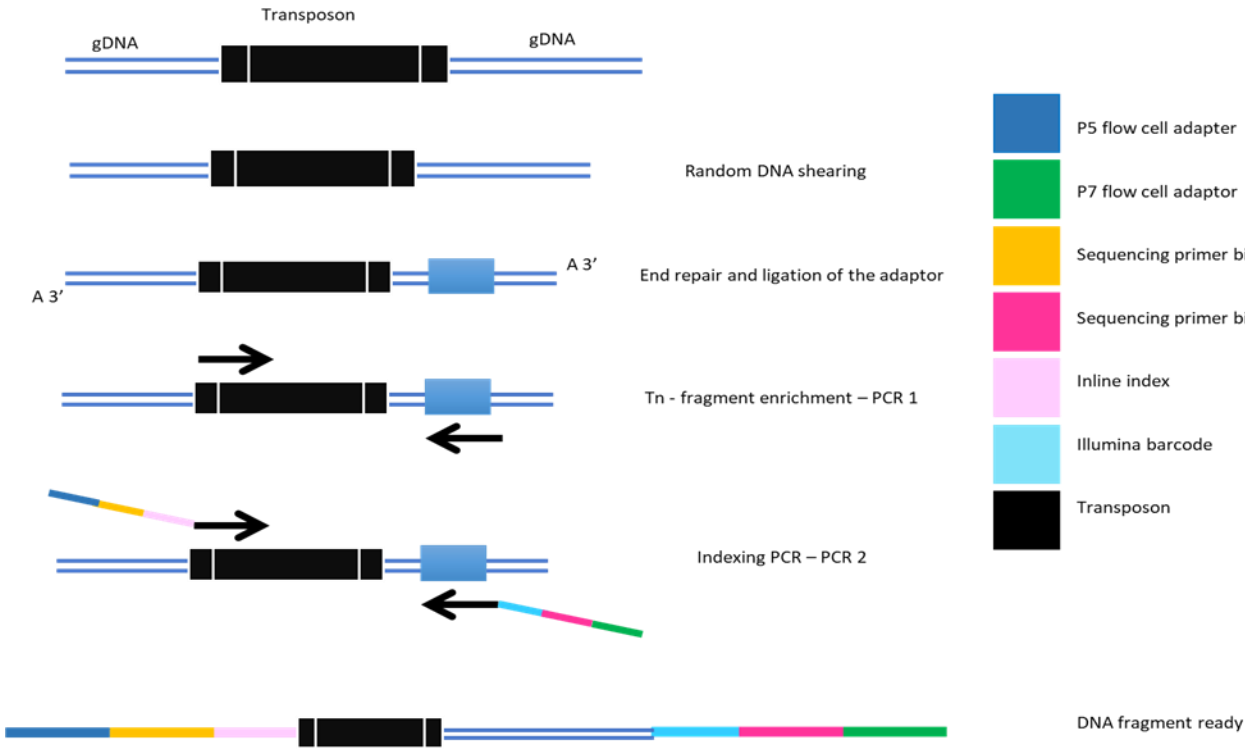


Figure 12. Preparing DNA fragment for sequencing.

General approach of Illumina transposon preparation for sequencing involves fragmentation, End repair and adaptor ligation, transposon amplification (PCR 1) and indexing in PCR 2. PCR 2 step introduces sequencing specific adaptors (to allow the sample to bind to Miseq flow cell), and barcodes.

2.6.9 Quantification of sequencing libraries by qPCR

The genomic libraries were quantified using the KAPA Library Quantification Kits from Roche (cat. Number KK4824). The kit consists of the qPCR Master Mix, a platform-specific library quantification primer premix, and a pre-diluted set of 6 DNA standards. The DNA Standards represent a 10-fold dilution series (20 pM to 0.0002 pM). The kit was used according to the manufacturer's protocol.

The library prep was vortexed well and diluted with 10mM Tris-HCl at pH 8 in 3 independent replicates. The dilution was done as follows: 1 µl of library prep into 999 µl (1000) → 20 µl into 980 µl (50,000) → lastly 100 µl into 900 µl (500,000). Each dilution was done in triplicate. The three replicates of dilutions 50,000 and 500,000 were quantified and water was used as control. The qPCR preparation was done on ice using a 96 well plate, semi-skirted. The following table illustrates the well composition for the qPCR:

Components	**Volume per well**
SYBR + primers	6 µl
Nuclease free dH_2O	2 µl
The standard or the diluted library prep	2 µl
Total	10 µl

In the PCR plate the six standards were loaded twice with at least two negative controls (sterile water). The PCR plate were spun down and sealed with PCR optical film and placed in an Agilent Aria MX (W131) qPCR machine with the brown mat on top and holes lined up with caps. The thermal profile was set for: 95 °C for 5 minutes followed by 35 cycles for 95 °C for 30 seconds and 60 °C for 30 seconds, melting the curve were obtained after PCR from 65 to 95 °C to assess sample quality.

2.6.10 qPCR result analysis

The results were analyzed with Aligent AriaMX software. In the experiment area analysis criteria were selected for all the cells in the 96 well plate and in the graphical display the threshold of the fluorescence (ΔRn) was adjusted. The results were exported with Cq (ΔR) values and saved as an Excel file.

2.6.11 The QPCR data calculation

An excel template from ROACH was used to analyse the Cq value (KAPA library quantification Kit Illumina® platforms, Data analysis template). The sheet provides a summary of the analyzed libraries simply done by entering the Cq values and it calculates the concentration of the undiluted library. Then the library was diluted down to 8nM stock concentration for sequencing on the Illumina Miseq.

2.6.12 MiSeq sequencing of the transposon library

The sequencing kit used was for the Illumina sequencing platform**,** MiSeq Reagent Kit v3 (150-cycle) MS-102-3001. The diluted libraries from previous step were pooled with 1.5 µl of each into one Eppendorf called PAL (pooled amplified library). In a 1.5 ml Eppendorf tube: 3 µl of TE at pH 8, 2 µl of PAL and 5 µl of fresh 0.2 M NaOH were added and incubated at RT for 5 minutes. Followed by addition of 990 µl of HYB buffer and transferred to a 98 °C heat block for 2 mins. At the same time, 30 µl of 20 pM bacteriophage PhiX DNA were denatured at 98 °C for 2 mins. After

denaturation, libraries were mixed with illumine generated control libraries PhiX 30 µl (5%) and 600 µl was into the cartridge well.

To start the run, the instructions in Miseq were followed to load PR2 bottle, waste bottle, flow cell and cartridge. The run aimed for an optimal cluster density of 1,000 clusters per mm2 using 150 cycle v3 cartridges.

2.6.13 Downloading the results:

The result was in the form of zip fastq files found in BaseSpace Sequence hub https://basespace.illumina.com/dashboard . The sequencing output file was downloaded into the computer and uploaded into the ubuntu (operating system branch of Linux) server using Putty (remote access). Details on data processing were in bioinformatics section 2.7.

2.6.14 Genomic DNA Sequencing

Bacterial culture was sequenced by MicrobesNG for standard whole genome sequencing at University of Birmingham. Once the sequencing request was accepted, the sample was prepared according to the provided method from MicrobesNG.

2.7 Bioinformatics

This section explains the bioinformatics workflow, I have followed for transposon sequencing data handling of *E. coli* genome. The workflow is outlined in figure 13. The figure shows the pipeline used to analyse sequencing data with the required input files and the obtained output files. Each script used will be explained in detail in the following sections.

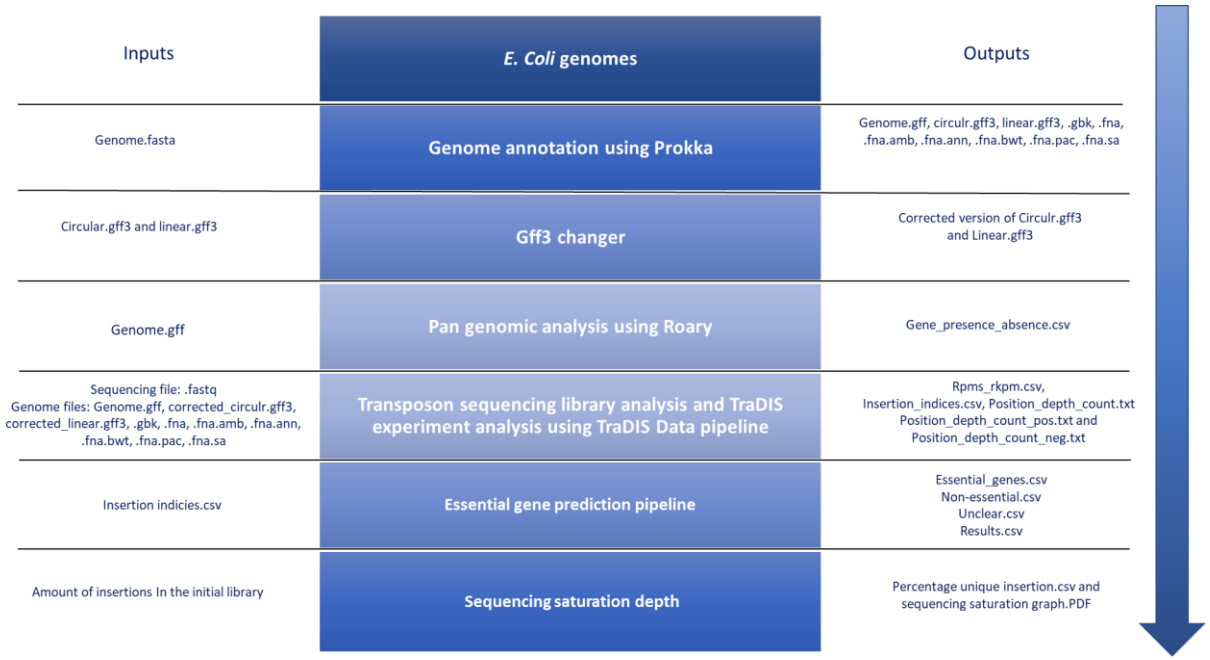


Figure 13. The data handling process workflow outlined of the scripts used in the sequencing library analysis and TraDIS experiment analysis.

The inputs column shows the required files for the corresponding scripts. The outputs column shows the output files obtained and the blue arrow indicates the order in which the scripts were used.

2.7.1 PROKKA: genome annotation

PROKKA is an automated genome annotation package used to annotate bacterial genome sequences with useful information (Seemann, 2014). The script was installed from https://github.com/tseemann/prokkaUbun and updated to version 1.14.0. The command line was followed from the help menu for the annotation options. The input genomic DNA sequence were in FASTA format. The protein annotation was performed against MG1655 geneBank file (NCBI Reference Sequence: NC_000913.3).

2.7.2 Gff3 Changer

General Feature Format (gff3) files were used to alignment of the sequencing file .fastq to the genome .gff3 by TraDIS pipeline. GFF3 linear and circular files from PROKKA outputs were used as an input files in gff3 changer script. This is to re-identify the ID attributes in column 9, figure 14. This was done because TraDIS pipeline can only identify the attribute in a specific format. The script was written by Dr. Mathew Milner.

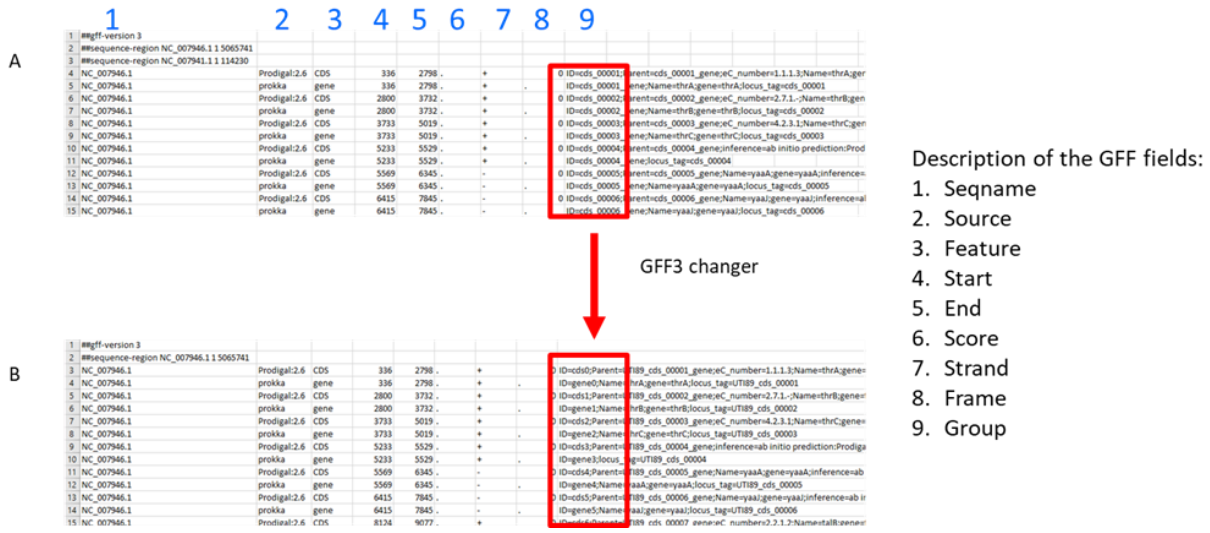


Figure 14. Screenshot of genome annotation Gff3 files.

A- Input gff3 file format to gff3 changer script. The script re-identify the attribute in gff3 file in column number 9. B- The gff3 file corrected version output used for TraDIS pipeline. The script was written by Dr. Mathew Milner.

2.7.3 Roary

Roary is a pan-genome pipeline used to determine gene presence or absence in the tested strains. The pipeline was installed from https://sanger-pathogens.github.io/Roary/ into ubuntu, version 3.13.0 (Page et al., 2015). The PROKKA annotation file .gff was used as input for Roary to produce a gene presence/absence matrix. The default usage of Roary was used.

2.7.4 Sequencing analysis (Sequencing libraries analysis)

The sequencing output .fastq files were downloaded from BaseSpace of illumina and uploaded into ubuntu server. These files were processed with a customized version of the scripts (TraDIS pipeline) written by Dr. Ashley Robinson and some minor editing done by Dr. Mathew Milner. A flow chart outlining the data processing steps shown in figure 15. In brief, sequencing reads were first trimmed to remove the Illumina adaptor sequences, then further separated according to the inline barcode used for each replicate, using the Fastx barcode splitter and trimmer tools (Pearson et al., 1997). The sequence reads were checked and trimmed in two steps: at first 25 bp of the transposon sequence was used allowing 3 bp mismatch. The resulting match transposon sequence was checked for the last 10 bp of the transposon allowing 1 base mismatch. Subsequently, the identified transposon sequence was discarded using Trimmomatic (Bolger et al., 2014). Sequence reads with low quality or < 20 bp were discarded. The trimmed reads were then mapped to the bacterial reference genome using Burrows-Wheeler alignment tool (BWA) (Li and Durbin, 2009). Mapped reads were sorted and indexed using SAMtools (sequence alignment/map) and converted to BED format using BEDtools (browser extensible data) and then intersected against the annotated protein coding sequences in bacterial genome or the plasmid in the format .gff of the PROKKA output.

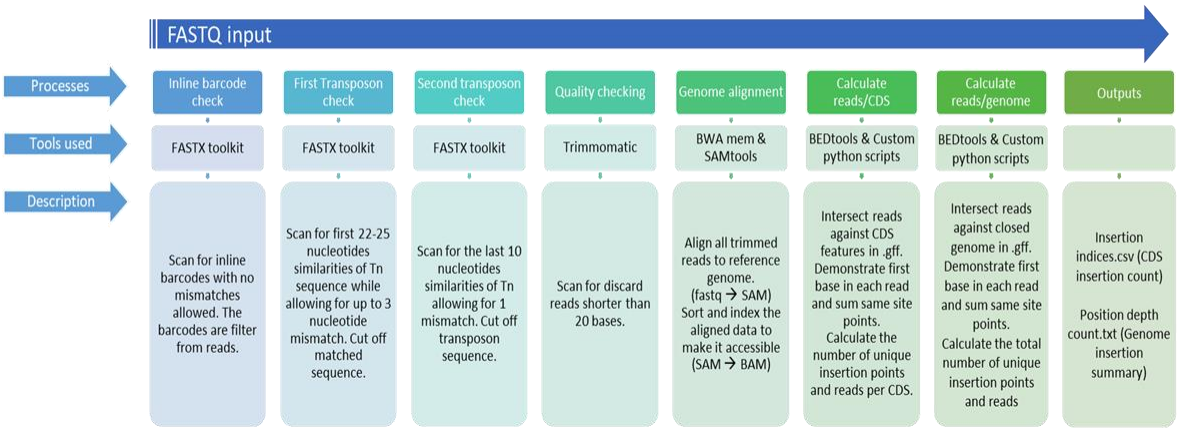


Figure 15. Workflow of TraDIS data analysis.

The raw data (fastq files) were processed using a series of tools combined in one pipeline. The arrow represents the direction of the data handling. Tools are described in details in the box below.

2.7.5 Essential gene prediction

Essential gene prediction for the sequencing libraries data was done by Dr. Sara Jabbari using edited version of the scripts from (Langridge et al., 2009). The final essential gene prediction was processed by me. The Freedman–Diaconis approach was used to produce a histogram with data-informed bin widths. Using the R Project for Statistical Computing (http://www.r-project.org), an exponential distribution was found on the left for the essential genes and the gamma distribution found on the right for the non-essential genes. For each data set (essential or non-essential), the probability was estimated, then the resulted probability ratio was used to calculate a log-likelihood value. The log-likelihood values were used with 12-fold likelihood threshold. Thus, by calculating what genes are likely to be in the left side of the graph (essential

genes) or right side of the graph (non-essential genes). Essential genes were 12 times were more likely to be in the left side and the non-essential were 12 times were more likely to be in the right side. The unclear genes were classified when they have a value that falls in between the maximum and the minimum of the $\log_2(12)$ threshold from (3.6 to -3.6). Data were then exported to Microsoft Excel for further analysis (Goodall, 2019).

2.7.6 Sequencing saturation depth script

The script is written by Dr. Mathew Milner. This script is designed to identify the total number of sequence reads required for TraDIS outgrowth samples to reach saturation depth. The formula depends on the insertion site number of sequencing libraries. The equation used $I = S - S\left(\frac{s-1}{s}\right)^n$, where '$I$' is the new number of insertions identified, 'S' the sample size, 'n' is the number of sample intervals which equates to the sequencing reads. 10,000 iterations were used as a read count to 4 million. Use the formula is described in result section 6.2.

2.7.7 EDGER

Once the raw sequencing fastq files were processed in the TraDIS pipeline, the pipeline aligns the reads to a reference genome. The alignment is to separate locations on the genome and result with final insertion count.txt file. This file contains the read counts correspond to each gene. These read counts were then analysed in the software package, "empirical analysis of digital

gene expression" (Robinson et al., 2010). EDGER requires two input .csv files: the first file contains the read counts corresponding to each gene in the libraries and the second file contains the total read counts for each library. For producing a table of read counts for EDGER, figure 16, the excel .csv should be in this order: gene name, read counts of initial sequencing library and followed by read counts of outgrowth TraDIS conditions. Read counts should be at least in two biological replicates to estimate biological variability.

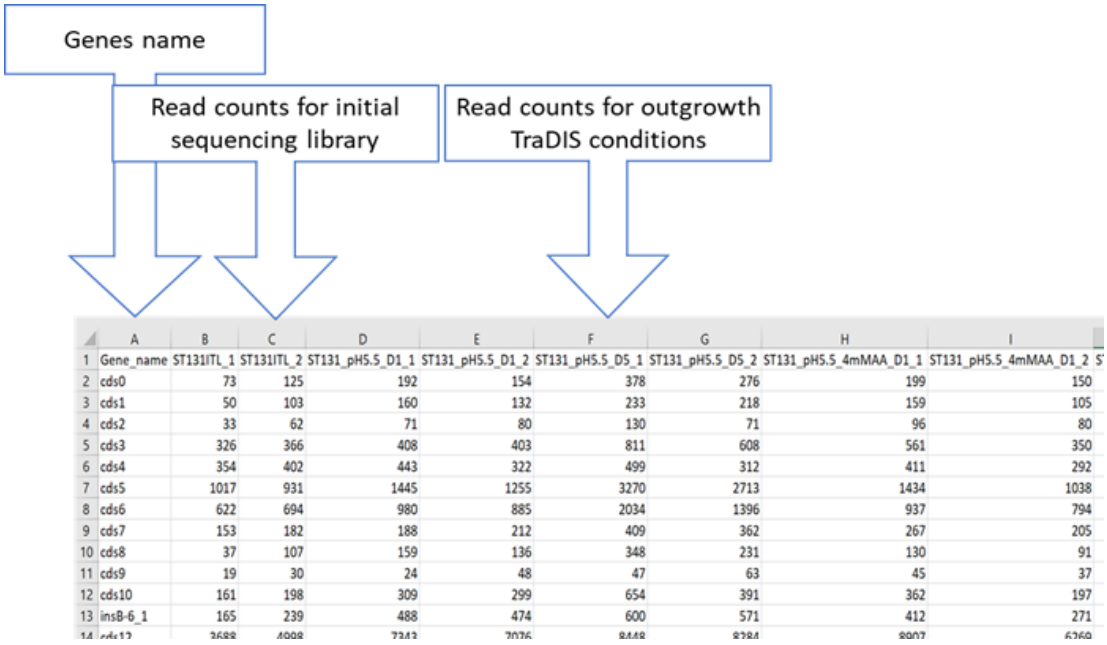
The EDGER package was used to identify statistically significant differences in read counts and mutant frequencies between experimental conditions. It models the read counts data as negative binomial (NB) distributed to measure gene fitness (gene expression).

$$Y_{gi} \sim NB\ (M_{iPgj}, \emptyset_g)$$

The number of reads i for a gene g is denoted as Y_{gi}. M_i is the total number of reads for the library size, $\emptyset_g$ refer for dispersion and Pgj is the relative abundance of gene g in experimental group j to which sample i belongs. NB parameterization where the mean is $\mu_{gi} = M_{iPgj}$ and variance is $\mu_{gi}(1 + \mu_{gi\emptyset gj})$. The $\emptyset_g$ the dispersion captures the library's variation between replicates.

EDGER estimates the genewise dispersion by conditional maximum likelihood, conditional on the total read count for that gene. It implements an empirical Bayesian strategy for squeezing the estimated dispersions towards the common dispersion. Finally, differential expression was computed using an exact test analogous to Fisher' exact test.

EDGER produces for each gene the logFC, log2 Fold Change: difference between experimental conditions, logCPM (log2 Count Per Million: normalised average between all samples), P-Value (exact test for the negative binomial distribution) and FDR (False discovery rate) (Robinson et al., 2010).



	A	B	C	D	E	F	G	H	I
1	Gene_name	ST131ITL_1	ST131ITL_2	ST131_pH5.5_D1_1	ST131_pH5.5_D1_2	ST131_pH5.5_D5_1	ST131_pH5.5_D5_2	ST131_pH5.5_4mMAA_D1_1	ST131_pH5.5_4mMAA_D1_2
2	cds0	73	125	192	154	378	276	199	150
3	cds1	50	103	160	132	233	218	159	105
4	cds2	33	62	71	80	130	71	96	80
5	cds3	326	366	408	403	811	608	561	350
6	cds4	354	402	443	322	499	312	411	292
7	cds5	1017	931	1445	1255	3270	2713	1434	1038
8	cds6	622	694	980	885	2034	1396	937	794
9	cds7	153	182	188	212	409	362	267	205
10	cds8	37	107	159	136	348	231	130	91
11	cds9	19	30	24	48	47	63	45	37
12	cds10	161	198	309	299	654	391	362	197
13	insB-6_1	165	239	488	474	600	571	412	271
14	cds12	3688	4998	7343	7076	8448	8284	8907	6269

Figure 16. Producing a table of reads count .csv file as input for EDGER.

The gene names were obtained from gene length.txt file. The read counts were obtained from final insertion counts.txt by processing the raw data in TraDIS pipeline. The .csv file should be in the same order; gene name, read counts in initial library and read counts in outgrowth TraDIS conditions. The rows correspond to genes and columns to independent libraries. The read counts for initial library or TraDIS should be at least in two biological replicates.

2.7.8 Data visualization

The transposon insertion read counts where visualized using the Artemis genome browser (Rutherford et al., 2000). Also Artemis was used to create DNAploter, genome transposon insertion sites mapped to the genome.

3 Validation of *E. coli* EO499 TraDIS Results Obtained Under Acetic Acid Stress

Declaration:

The focus of this chapter is to try to validate TraDIS data from EO499 generated by previous student, Francesca Bushell, and continuation of work on her thesis. In this chapter, the initial transposon sequencing library of EO499 was constructed byDr.Keith Turner. The TraDIS growth experiment was carried out by Dr. Francesca Bushell and Dr. Thippesh Sannassidappa. TraDIS libraries were sequenced in Liverpool University and processed using bioinformatic pipeline by Dr. John Herbert. The output files from Dr. John Herbert were preliminary analyzed by my supervisor Dr. Peter Lund. Based on the preliminary analysis on TraDIS EO499 at pH 5.5 with and without acetic acid, lists of candidate mutant genes were selected for TraDIS validation. I have selected the mutants from the lists to validate TraDIS data using Keio library knockouts BW25113 using competition experiments. All the validation experiments in this chapter were completed by me including: the construction of Lac^+ in *E. coli* BW25113, measuring the relative fitness of BW25113 Lac^+, confirmation of kanamycin insertion in the Keio collection, fitness of the candidate gene knockouts from Keio collection and the time course competition experiments.

3.1 Overview

The aim of this study was to try to validate results obtained through a series of TraDIS experiments performed by previous lab members Dr. Thippesh Sannassidappa and Dr. Francesca Bushell, as it is explained in the introduction section 1.11. In this chapter I am only going to focus on TraDIS data generated at pH 7 with and without 40 mM acetic acid and at pH 5.5 with and without 4 mM acetic acid, aerobically.

One simple way to measure bacterial fitness of an individual strain is to measure the growth rate. The growth rate was measured as the optical density at 600 nm over time for 8 hours using a spectrophotometer. However, this provides the growth rate which is only one component of fitness. As a fitter strain is expected to grow more rapidly than the other, therefore direct measurement of growth rate could be enough to determine relative fitness. However, differences in growth rates can be small and hard to observe and may only be seen when the strains are in direct competition. Moreover, differences in fitness may not result only in different growth rates but also different behavior elsewhere in the culture cycle, which may include different times taken to come out of log phase and stationary phase dynamics.

In this study I will use competition experiment to attempt to validate TraDIS data. The competition experiments are used to determine the bacterial relative fitness changes between ancestral and evolved populations (Lenski et al., 1991). The relative fitness of the two competitors were determined as the competitors diverge in number and the less fit population becomes smaller in proportion. The genotype with higher relative fitness will go through more generations and therefore increase in frequency over the considered time period in relation to the less-fit

competitors. In this method, the relative fitness is determined by measuring the relative net growth of the two competitors in batch culture, integrated across the full growth cycle: lag phase, log phase, stationary phase. The fitness was measured by the ratio of the growth rate of the two populations grown in the same sized flask and same environmental conditions as used in the TraDIS experiments. As shown in the materials and methods section 2.2.1, the relative fitness (*w*) as calculated:

$$w = \frac{In\left(\frac{A_f}{A_i}\right)}{In\left(\frac{B_f}{B_i}\right)}$$

Where N_f and N_i are the final and initial population sizes based on the colony counts of each competitor. The relative fitness formula is derived from the exponential growth equation:

$$x\,(t) = x_0\,.\,e^{kt}$$

where

$$e^{kt} = \frac{x_t}{x_0}$$

For example, for a population growing exponentially, the population at time t is the population at time zero multiplied by e^{kt}, where k the growth constant. The natural logarithm $In\,(x)$ is the inverse of the exponential function. Then, taking the natural log of both sides (Note: $In\,e = 1$, because $e^1 = e$, while the $In\,(1) = 0$, as $e^0 = 1$.

$$In\left(\frac{x_t}{x_0}\right) = kt$$

To compare two strains A and B k is replaced by the ratio of $\frac{K_A}{K_B}$, the same as w in the above formula.

$$w = \frac{K_A}{K_B} = \frac{\ln\left(\frac{A_f}{A_i}\right)}{\ln\left(\frac{B_f}{B_i}\right)}$$

In another words, the w is defined simply as the ratio of the growth constants K of the two populations A and B during competition. This is the same as the ratio of the number of doubling times of the two competitors (Wiser and Lenski, 2015). It should be noted that a limitation of this measure of fitness is that the formula measures the fitness based on the assumption that two competing strains are growing exponentially throughout the experiment. This would be correct if the two competitors are growing exponentially all the time, as the relative fitness w would remain constant and not change over time during the experiment, so the length of time doesn't matter. This formula ignores the fact that bacteria don't grow exponentially in a typical growth experiment and exponential phase is followed by stationary phase which has different dynamic behavior. I did the competition by growing the cells for 24 hrs, at this point we know our cells will not be in exponential phase. So, if the cells stop growing this will not affect the fitness value because once the cells reached a certain density the value w will not change, as the formula does not consider time. Despite the limitation, competition measured using this formula is widely used to validate transposon sequencing data (van Opijnen and Camilli, 2012; Jayeola et al., 2020; de Moraes, 2017), therefore I decide to use it.

The goal of this experiment was to identify genes that are important for fitness under these stresses, as this may give useful insights into the pathways which are inhibited or disrupted by acetic acid treatment. Using the relative frequency of insertions in non-essential genes before and after growth from TraDIS experiments with and without a particular stress, the relative fitness

of each mutant can be calculated. However, it is always best to validate such results by an alternative method. So ideally an individual knockout of each gene needs to be constructed and competed against the wild type under the different stress conditions to obtain a direct measurement of relative fitness for each gene. If the TraDIS method is reliable, I would expect to observe the relative fitness calculated from a TraDIS experiment to correlate with that from a competition experiment using single gene knockouts.

In the following section, I will discuss the nature of the TraDIS data and how it was analyzed for each gene and how the genes were ranked. Then I will discuss a variety of approaches that were taken to obtain validated mutants of the genes of interest, in EO499 and in BW25113. Lastly, I will present the competition data using these mutants and some mutants and discussed their relationship with the TraDIS data.

3.2 Analysis of TraDIS data

In this section I will discuss the nature of the sequencing data, and how I have generated a list of potential candidate genes affected by the presence of acetic acid at pH 7 and pH 5.5. The aim of this analysis was to find genes required for optimum growth of EO499 in the presence of acetic acid by identifying genes that when mutated, caused loss of bacterial fitness.

The TraDIS libraries from experiments where the EO499 library was grown under different conditions were generated in Birmingham and sequenced in Liverpool University and analyzed by Dr. John Herbert. The sequence reads were displayed in a custom-built browser which combines

the annotated EO499 genome with TraDIS short sequence reads, all mapped to the genome to show precise insertion sites and numbers of read for each growth condition, refer figure 17. The EO499 genome annotation was carried out using Prokka pipeline as described in the method and material section 2.7.1.

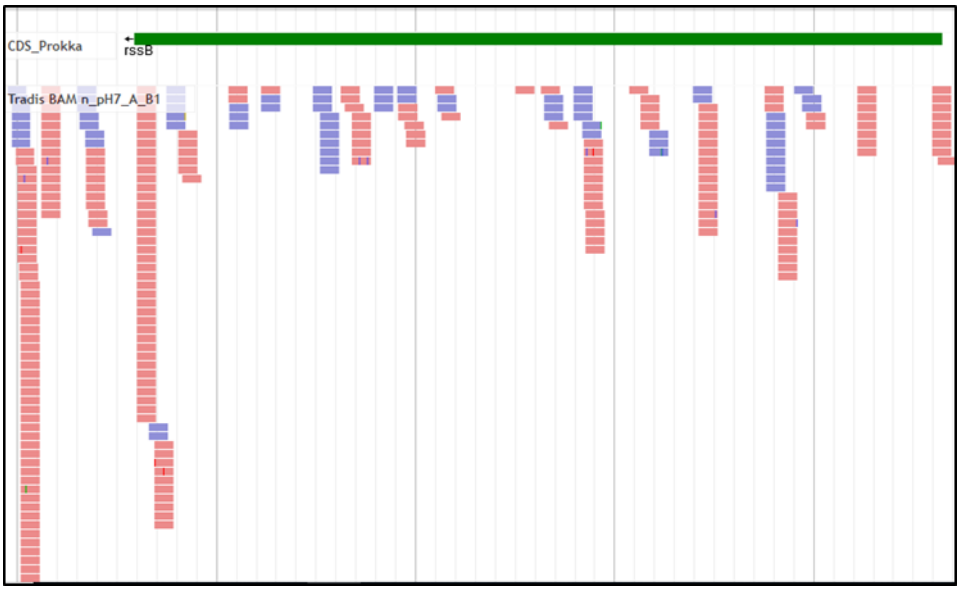


Figure 17. A screen capture of the browser built by Dr. Herbert, Liverpool University to show TraDIS data of EO499 under different conditions.

Data could be displayed for all the different growth conditions. The green line shows the *rssB* gene (size of 1,013 bp). The bands represent read alignments generated at pH 7 with pink and blue representing the orientation of the Tn.

Moreover, Dr. Herbert also provided us with Excel files containing the summaries of the TraDIS data as analysed using the ESSENTIALS pipeline (Zomer et al., 2012). These spreadsheets contained the gene list of EO499, values for raw fitness, values for fitness expressed as $\log_2$ Reads Per Kilobase Million (RPKM), the raw count of gene insertions, and the mean insertion indices for each gene. The raw fitness expression is the counts of total reads mapping to insertion sites of genes. The given RPKM values were normalized for each replicate and generated using the formula: RPKM = (total number reads mapped to a gene x 10^6) / (gene length in kilobases). The fitness expression $\log_2$ RPKM is the $\log_2$ total reads mapping to insertion sites of genes. The insertion indices are the non-redundant number of insertion sites in a gene divided by the gene length. For example, to calculate the insertion index for gene **A** which has a length of 1000 bp and 19 sequencing reads within 6 unique insertion sites. So, 6 unique insertion sites /of the gene length 1000 bp = 0.006 insertion index. Calculating the insertion indices for all the genes is important to determine gene essentiality (Langridge et al., 2009), discussed in chapter 5.

Dr. Herbert also provided us with pairwise comparisons between pH 7 versus pH 5.5, pH 7 acetic versus pH 5.5 acetic, pH 7 versus pH 7 acetic and pH 5.5 versus pH 5.5 acetic giving a value for log fold change logFC ($\log_2$) and a measure of the false discovery rate (FDR). The FDR (corrected p-value) is the expected proportion of tests which are incorrectly called significant out of the total significant tests. Using a suitable FDR reduces the likelihood of incorrectly rejecting the true null hypothesis (there are no differences between the conditions) and reduces Type I errors (false positive).

When I started my project, a detailed statistical analysis of the TraDIS data had not been completed by our collaborators in Liverpool. So I have started the analysis with a relatively simple approach (described below) to rank gene lists for validation, in spite of the fact that the initial data did not have any P-value connected to it. Later analysis showed that our generated gene lists were in fact highly similar to those generated by a much more detailed analysis, which justifies the approach I have took. Dr. John Herbert's analysis will be provided later in this section.

This preliminary analysis of TraDIS sequencing data was done simply by Excel rather than using a computational pipeline. The aim of this analysis was to find genes required for optimum growth of EO499 in the presence of acetic acid and mutants that showed loss of fitness. This enables us to find genes affected by the presence of acetic acid and to begin to determine the molecular pathways involved.

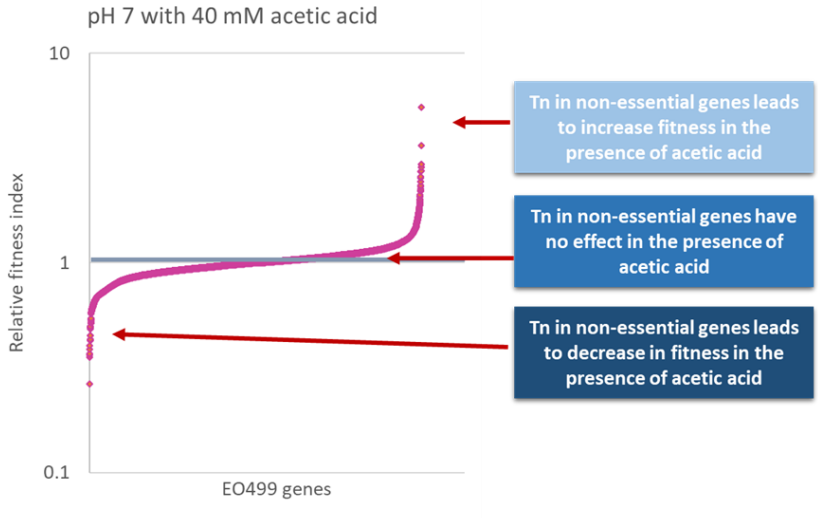
First, from TraDIS data I have tried to determine if there were genes normally non-essential for growth that become essential in the presence of acetic acid, in both cases (pH 7 and pH 5.5). I have examined this by looking for genes that had a plenty of inserts in the ITL (Initial transposon library) or when grown without acetic acid, but where there were no reads for the genes after the stress was applied. In our analysis, no genes were found to be completely essential in the presence of acetic acid.

Therefore, I have looked for other genes which may contribute to fitness when acetic acid is present. To do this, I have divided the RPKM value for each gene after growth with and without acetic acid, at both pH values. Generally, most non-essential genes are not expected to provide a fitness advantage under any specific growth condition. So, if I calculate the relative fitness based

on the ratio of RPKM values for most of the genes for strains grown with or without acetic acid I would expect the ratio will be close to one. But, a small number of individual genes would be expected to have values below one: i.e., inactivation of these genes causes reduced fitness under the environmental condition. These are the genes I am interested in. Note also that some genes may when mutated give an increase in fitness under acetic acid, these will have a relative fitness value larger than one, discuss later on this section.

I predicted that by using the ratio of RPKM values for most of the genes before and after growth, most of these values will be close to one. To evaluate all the genes grown in the presence of acetic acid at both pH values I did the calculation of relative RPKM values for all genes, and the fold change of the RPKM values is shown in figure 18. In figure 18, as predicted, a small number of genes showed significantly reduced relative fitness values when mutated by transposon insertion, while some of the inactivated genes gave a relative fitness value above one. Genes with low relative fitness were the genes of interest; these data will be discussed in the discussion section. Table 5 shows the genes that have a larger reduction in fitness, at both pH 7 and pH 5.5, in the presence of acetic acid, relative to growth at the same pH but without organic acid. These have been ranked in order of increasing relative fitness. Notably, in table 7 (generated from the subsequent analysis done by Dr. Herbert) although the ranking changes slightly among the two calculated relative fitness groups, the same genes appear to be equally present.

A


pH 7 with 40 mM acetic acid
10
1
0.1
Relative fitness index
EO499 genes
Tn in non-essential genes leads to increase fitness in the presence of acetic acid
Tn in non-essential genes have no effect in the presence of acetic acid
Tn in non-essential genes leads to decrease in fitness in the presence of acetic acid

B

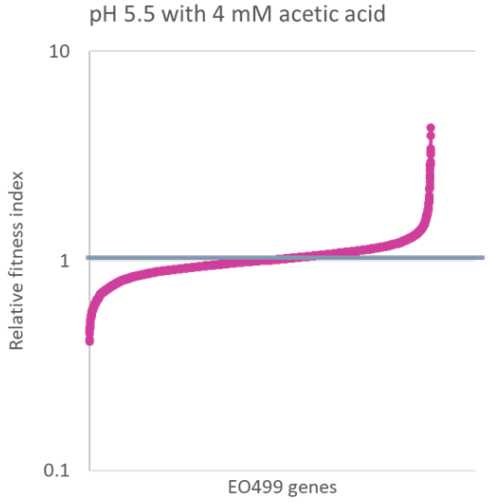

pH 5.5 with 4 mM acetic acid
10
1
0.1
Relative fitness index
EO499 genes

Figure 18. The relative fitness index ranked of all non-essential genes, plus/minus acetic acid. At (A) pH 7 with 40 mM acetic acid and (B) pH 5.5 with 4 mM acetic acid. The relative fitness values of all non-essential genes were calculated under these two conditions; all genes were ranked by fitness value and plotted. Genes where inserts lead to a gain fitness have a relative fitness value of more than one. The opposite applies to genes where inserts lead to a decrease fitness: these have a relative fitness values less than one. As made according to the instructions from Dr. Peter Lund.

The analysis of Dr. John Herbert of pH 5.5 with acetic acid, that was provided later on, with log fold change and FDR, is shown in table 7. These data were ranked based the log fold change from highest to lowest fitness value, but in this analysis the log fold change was calculated as pH 5.5 + 4mM acetic acid/ pH 5.5. Although both analyse were generated differently, the genes selected were similarly in both lists but with different ranking order. For example, our analysis in table 5 showed a strong fitness reduction for mutations in the *sthA, ytfP, nuo* genes, *apaH, rssB, pitA, suc* genes and *sdh* genes at pH 7 and pH 5.5 with acetic acid. These genes were not affected in the absence of acetic acid as it shown in table 5 column A and B.Dr.John Herbert's analysis identify similar genes under at pH 5.5 with acetic acid (although with a different ranking) and significant FDR value, table 7.

Table 5. The top 26 genes due to the decrease on their relative fitness effect in TraDIS in EO499.

(1) columns A and B: pH 5.5 vs pH 7; (2) columns C and D: pH 7 with acetic acid vs pH 7; (3) pH 5.5 with acetic acid vs pH 5.5. The genes were ranked based on their relative fitness values, calculated by using the fold change of RPKMs with and without the stress. I have selected eight genes for further validation; *nuoM, nuoG, sucA, sthA, pitA, apaH, rssB* and *ytfP*.

A	B	C	D	E	F
Genes	Relative fitness of pH 5.5 / pH 7	Genes	Relative fitness of pH 7 with acetic acid/ pH 7	Genes	Relative fitness of pH 5.5 with acetic acid/ pH 5.5
dgkA	0.18	*sthA*	0.27	*ytfP*	0.41
cpxA	0.44	*carB*	0.35	*apaH*	0.42
finO	0.49	*sdhA*	0.36	*ST131v2_05400*	0.45
trkA	0.49	*lon_1*	0.37	*sthA*	0.46
ST131v2_05445	0.49	*nuoG*	0.39	*traM*	0.46
ST131v2_05453	0.51	*aspC*	0.40	*ST131v2_05394*	0.48
ST131v2_05402	0.53	*pitA*	0.43	*bla_3*	0.48
ST131v2_05452	0.53	*nuoM*	0.43	*sdhB*	0.50
ST131v2_05403	0.53	*sixA*	0.43	*ST131v2_05412*	0.50
ST131v2_05435	0.54	*cpxA*	0.43	*nuoB*	0.50
sapA	0.54	*sucD_1*	0.45	*pitA*	0.50
vapB	0.54	*ytfP*	0.45	*ST131v2_05397*	0.51
ST131v2_05299	0.54	*sucA_1*	0.48	*folA_3*	0.51
yfjX_3	0.55	*rcsA*	0.49	*ST131v2_05428*	0.51
ST131v2_05311	0.55	*apaH*	0.49	*rssB*	0.52
ST131v2_05406	0.56	*sucC_1*	0.50	*folA_2*	0.52
yfjQ_5	0.56	*rssB*	0.50	*chpB_2*	0.52
ST131v2_05296	0.56	*phoR*	0.51	*nuoM*	0.53
ST131v2_05283	0.56	*ST131v2_01743*	0.52	*ST131v2_05382*	0.53
sok_4	0.56	*yejM*	0.52	*ybiI_2*	0.54
ST131v2_05369	0.56	*ST131v2_05448*	0.53	*clpP*	0.54
ST131v2_05400	0.56	*sucB_1*	0.54	*pbl*	0.54
yhcR	0.57	*nuoL*	0.54	*agp_2*	0.54
ST131v2_05397	0.57	*nuoE*	0.54	*traY*	0.54
pinE_4	0.57	*nuoH*	0.54	*ST131v2_05287*	0.55
vapC_2	0.57	*sdhC*	0.55	*sucA_1*	0.55

Genes where mutations caused increased in fitness under acetic acid are shown in table 6. Although the relative fitness ranking of genes in column C and E in the presence of acetic acid is different, many genes were enriched in both lists in the presence of acetic acid. For example, there are some genes where mutations caused increase in fitness under acetic acid such as *pta, aroBE, nadBC, panCB, dnaJ,* and *bioAH*. Moreover, there are some genes where mutations cause increased fitness both with and without acetic acid such as *aroAD*, *bioCF*, *ppc*, *pgi*, *metJ* and *ackA.* Looking to the gene function of some of these genes they can be seen to have key roles in normal acetic acid metabolism; for example, the enzyme phosphotransacetylase (*pta*) is involved in the pathway for production of aetate and catalyzes the reversible interconversion of acetyl-CoA and acetyl phosphate. The gene *ackA* encodes acetate kinase which is also involved in the acetate metabolism and catalyzes the formation of acetyl phosphate from acetate and ATP (Klein et al., 2007; Ren et al., 2017).

There was not much time in this project to look at these gene. Experimentally the approach would be the same as I have been following, either gene knockout or isolation of the mutant from the library for validation. The isolation of mutants from the library is discussed in the next chapter. Subsequently, I would expect increase of fitness in competition but due to time limitation this was not investigated further. Biologically, it is not yet clear what is the reason gain of fitness caused by mutations in these genes. However, they are of interest because mutations in these genes could lead to strains acquiring acetic acid resistance if it was widely used for treatment.

Table 6. The top 26 genes due to the increase in their relative fitness effect in TraDIS EO499. (1) columns A and B: pH 5.5 vs pH 7; (2) columns C and D: pH 7 with acetic acid vs pH 7; (3) pH 5.5 with acetic acid vs pH 5.5. The genes were ranked based on their relative fitness values, calculated by using the fold change of RPKMs with and without the stress.

A	B	C	D	E	F
Genes	Relative fitness of pH 5.5 / pH 7	Genes	Relative fitness of pH 7 with acetic acid/ pH 7	Genes	Relative fitness of pH 5.5 with acetic acid/ pH 5.5
pabA	2.01	*ackA*	5.51	*aroE*	4.33
serA_1	1.95	*pta*	3.61	*ackA*	3.95
pabB	1.91	*nadB*	2.95	*panB*	3.42
moaE	1.90	*pgi*	2.94	*bioF*	3.33
Ppc	1.86	*bioF*	2.85	*dnaJ*	3.24
aroA	1.86	*bioC*	2.85	*bioH*	3.21
mobA	1.80	*bioH*	2.75	*panC*	2.96
mntS	1.79	*panB*	2.72	*pta*	2.88
Pgi	1.69	*aroA*	2.58	*bioC*	2.87
bioC	1.68	*aroB*	2.53	*bioA*	2.83
metJ	1.67	*sspA*	2.44	*aroH*	2.72
yifE	1.65	*nadC_2*	2.44	*ppk*	2.71
yeaC	1.64	*panD*	2.34	*aroC*	2.57
panE	1.64	*panC*	2.30	*aroB*	2.52
serB	1.63	*ppc*	2.26	*trpE_2*	2.47
aroD	1.63	*surA*	2.22	*bioB*	2.38
bioF	1.59	*metJ*	2.21	*aroA*	2.28
RyhB	1.59	*ytfK*	2.19	*nadB*	2.26
pmrD	1.58	*aroE*	2.19	*nadC_2*	2.23
znuB_2	1.58	*fis*	2.12	*yijF*	2.21
ychF	1.57	*bioA*	2.08	*rlmE*	2.20
yacL	1.57	*relA*	2.06	*aroK*	2.05
ackA	1.57	*aroD*	2.02	*rsmA*	2.02
sthA	1.56	*xerC*	2.01	*cspE*	2.01
tolA	1.56	*dnaJ*	2.00	*carA*	1.98
yejM	1.55	*nadA*	1.96	*ST131v2_00839*	1.97

Table 7. Top 25 genes due to the decrease in their relative fitness effect at pH 5.5 with acetic acid vs pH 5.5 in TraDIS EO499 ranked by the FDR.

The genes are ranked based on their log fold change of pH 5.5 / pH 5.5 with acetic acid, along the FDR and gene annotation. The analysis was conducted by Dr. John Herbert, (Bushell, 2019).

Gene	logFC change	FDR	Product
rssB	0.74	7.97E-17	PcnB-degradosome interaction factor; response regulator
sdhB	0.64	2.56E-11	succinate dehydrogenase, FeS subunit
sthA	0.83	2.12E-09	pyridine nucleotide transhydrogenase, soluble
nuoG	0.80	2.36E-08	NADH:ubiquinone oxidoreductase, chain G
apaH	0.44	4.23E-08	diadenosine tetraphosphatase
ytfP	0.66	7.97E-17	GGCT-like protein
lon_1	0.66	1.54E-07	DNA-binding ATP-dependent protease La
nuoB	0.53	6.45E-06	NADH:ubiquinone oxidoreductase, chain B
nuoL	0.73	6.78E-06	NADH:ubiquinone oxidoreductase, membrane subunit L
nuoM	0.67	1.76E-05	NADH:ubiquinone oxidoreductase, membrane subunit M
pitA	0.76	2.96E-05	phosphate transporter, low-affinity; tellurite importer
ST131v2_01428	0.51	0.000157	-
clpP	0.58	0.000212	proteolytic subunit of ClpA-ClpP and ClpX-ClpP ATP-dependent serine proteases
yccJ	0.64	0.000334	uncharacterized protein
nuoF	0.77	0.000357	NADH:ubiquinone oxidoreductase, chain F
ST131v2_01182	0.57	0.001832	-
ST131v2_01427	0.70	0.012085	tRNA-Arg
ST131v2_02655	0.68	0.024297	Bacterial protein of unknown function (DUF977)
sdhC	0.88	0.024297	succinate dehydrogenase, membrane subunit, binds cytochrome b556
hokD_1	0.77	0.029758	Qin prophage; small toxic polypeptide
pspD	0.61	0.032873	peripheral inner membrane phage-shock protein
pdxB	0.79	0.03708	erythronate-4-phosphate dehydrogenase
sdhD	0.66	0.049023	succinate dehydrogenase, membrane subunit, binds cytochrome b556
sucA_1	0.55	0.13514	2-oxoglutarate decarboxylase, thiamine triphosphate-binding
atpC	0.86	0.30884	F1 sector of membrane-bound ATP synthase, epsilon subunit

I have selected the eight genes where inserts had the largest negative effect on fitness at both pH 7 and pH 5.5 in the presence of acetic acid from table 8 for further validation studies. These were *nuoM*, *nuoG*, *sucA*, *sthA*, *pitA*, *apaH*, *rssB* and *ytfP*. The biological function and cellular location of each of the candidate genes is shown in table 8. Further details of these genes and the proteins they encode will be covered in the discussion, linked to pre-existing data of the general mode of action of organic acids.

Table 8. Biological function and cellular location of the candidate genes from acetic acid stress.

Gene	Gene Description	Biological Process	Cellular Location
nuoM	Polypeptide: component of NADH quinone oxidoreductase	Proton translocation; involved in proton pumping	Inner membrane
nuoG	Polypeptide: NADH: quinone oxidoreductase, represents the electron input part of the enzyme	NuoG is part of the soluble fragment of NADH dehydrogenase I, which represents the electron input part of the enzyme.	Inner membrane
sucA	Enzyme: 2-oxoglutarate decarboxylase, thiamine-requiring.	Catalyzes the conversion of 2-oxoglutarate to succinyl-CoA and CO_2.	Cytosol
sthA	Enzyme: Soluble pyridine nucleotide transhydrogenase	Involved in the reoxidation of NADPH. Membrane-bound proton-translocating transhydrogenase.	Inner membrane
pitA	Transporter: metal phosphate: H+	The PitA phosphate transporter is a member of the Inorganic Phosphate Transporter (PiT) family.	inner membrane
apaH	Enzyme: diadenosine tetraphosphatase	An *apaH* mutation leads to elevated abundance of diadenosine tetraphosphate (Ap4A).	Cytosol
rssB	Polypeptide: regulator of RpoS	Regulates the turnover of the sigma S factor (RpoS) by promoting its proteolysis in exponentially growing cells. Acts by binding and delivering RpoS to the ClpXP protease. RssB is an adaptor protein that facilitates degradation of σS by the protease ClpXP.	Cytosol
ytfP	Polypeptide: γ-glutamylamine cyclotransferase family protein YtfP	May play a role in antibiotic biosynthesis.	Cytosol

One way to study the link between genotype and phenotype on the tested conditions, is to make deletion ("knock out") mutations of the eight candidate genes and measure the fitness of the resultant mutants relative to the wild type parent. The bacterial fitness of these mutants can be tested by competition experiment by growing the wild type-strain in competition with the

mutant at pH 7 and pH 5.5 with and without acetic acid. I have chosen competition experiments for validating the TraDIS observations because TraDIS itself involves microbial competition between all the mutants in the transposon library. Next, I will explain below the three different ways I can use for validating the relative fitness for these genes, using knockouts in EO499 and using the Keio collection of BW25113.

3.3 Attempts to isolate mutations in EO499

The first objective was to construct knockout mutants of the individual genes that have been identified using TraDIS as contributing to fitness under acetic acid stress. Ideally these knockouts should be in strain EO499, but making knockouts in EO499 strains is known to be difficult. Previously, two lab members Dr. Thippesh Sannassidappan and Dr. Hadi Mohammad attempted to generate knockouts of the target non-essential genes in EO499. Dr. Thippesh tried to replace *lacZ* gene with chloramphenicol resistant marker using lambda red recombinase system under the arabinose inducible promoter pKD46 (Datsenko and Wanner, 2000). Hadi tried construct a knockout in *nuoG* using gene doctoring method (Lee et al., 2009). Both methods were unsuccessful despite multiple attempts.

Therefore, I have tried to solve this problem by directly isolating strains containing transposons in the genes of interest from the transposon library of EO499. Dr. Francesca Bushell and I have successfully isolated two transposon mutants of *ytfP* and *rssB* respectively from EO499

transposon library. These two mutants were competed with the parent EO499 to examine and validate TraDIS data. Unfortunately, this method is time-consuming. The details of this method and results will be discussed in the next chapter.

3.4 Competition experiments in BW25113

Because of the difficulties in making knockouts in the EO499 strain, I instead used Keio collection of knockouts in *E. coli* K12 BW25113 (where each non-essential gene is replaced with a Kanamycin cassette (Baba et al., 2006) to determine the relative fitness of candidate genes by competition under TraDIS conditions. This enabled us to determine whether TraDIS in one strain of *E. coli* can be validated by using knockouts in a different *E. coli* strain. Moreover, it allows us to identify the fitness of these knockouts relative to the wild type under different stress conditions using competition experiments.

In this study, I have tested whither the phenotype of BW25113 knockouts was as predicted based on the EO499 data in the presence of acetic acid at pH 7 and pH 5.5. I have chosen to work with *E. coli* BW25113 for few reasons. First, I had easy access to the Keio collection library, the library was available in our lab. This reduces the time used to construct the mutants in any other strains. Second, because they are both *E. coli* and I might expect the data will correspond quite well. Also, it is well known that core genome (genes shared by all strains of *E. coli*) is highly similar and conserved among *E. coli* strains (Abram et al., 2020; Van Elsas et al., 2011). These genomes sustain key cellular functions that are likely to be essential. However, EO499 is a pathogenic strain

and BW25113 is a lab strain, so it might not correspond very well. Therefore, I have decided to use different strain of a single species and to evaluate EO499 TraDIS candidate genes under acetic acid stress using knockouts in BW25113 and to determine the fitness of candidate genes among another *E. coli.* Furthermore, the predictions sometimes disagree with experimental data, this might leads us to have different bacterial expression under a certain condition.

In the next section, I will explain in detail how competitions were done using strains from the Keio collection. Then, I will describe how BW25113 was constructed with the appropriate genetic marker, followed by testing the relative fitness of the marked strain (as it should be neutral to be used in these experiments). I will show how I have confirmed the Kanamycin insertion in strains from the Keio collection, and finally, I show how I have determined the relative fitness results obtained from competing the knockout strains in the TraDIS-identified candidate genes and the parent BW25113.

3.5 Construction of Lac$^+$ in *E. coli* K-12 BW25113

Since the Keio collection knockout strains are marked with kanamycin resistance cassette, a competition experiment could be performed using kanamycin agar and LB agar to distinguish between the wild type BW25113 and mutants. This can be done in the same way as shown in the methods and materials (section 2.2.1), but with the two competitors plated on LB and kanamycin agar plates. In this case the wild type and the mutant will grow on the LB agar and only the mutant will grow in on kanamycin plate. To determine the relative fitness, the mutant colonies number

on kanamycin would be subtracted from the number of colonies on LB agar to determine the wild type colonies number.

Another option is to use Lactose-fermenting and non-lactose fermenting BW25113 that can be distinguished by their color on lactose MacConkey plates, allowing measurement of relative fitness on the same plate of competed pairwise strains before and after competition under different conditions. In this study, I have chosen to use lactose MacConkey to perform competition. Using this method reduces the number of the plates to half and it takes less time to count the plates and to identify the wild type *lacZ*$^{+}$ and *lacZ*$^{-}$ mutants.

BW25113 wild type is Lac^{-} because of deletion *ΔlacZ4787* as are the selected mutants. To be able to distinguish between the two, I have needed to construct a BW25113 Lac^{+} and then to compete the wild type BW25113 *lacZ*$^{+}$ against the mutant *lacZ*$^{-}$ strains. To do this I have used P1 transduction to transfer *lacZ*$^{+}$ from *E. coli* MG1655 to the wild type BW25113, shown in material and methods section 3.2. The transductant colonies were screened on MacConkey agar with lactose to distinguish the positive pink color colonies from other transductants colonies which are white. The transduced colonies were checked by PCR using primers found in supplementary table S1, and the thermocycler condition illustrated in table 4.

To confirm that Lac^{+} colonies are BW25113 and not MG1655 that had contaminated the P1 lystate, I have confirmed that controls of P1 phage alone showed no growth on MacConkey plates. Second, I have also used PCR to confirm the transduced colonies were recipient strain BW25113 and not donor strain MG1655, by checking genes present in MG1655 but not BW25113.

To do this, primers araA-F, araA-R, rhaA-F and rhaA-R were designed to amplify *araA* gene and *rhaA,* respectively. These are present in the parent strain MG1655 but not BW25113, because of the deletion in rhamnose Δ*(rhaBAD)568* and arabinose Δ*(araBAD)567* operons in BW25113. These genes were selected from a (genome variation between *E. coli* MG1655 and *E. coli* BW25113) (http://bioinfo.ccs.usherbrooke.ca/BW25113.html).

So transduced colonies would not show positive bands for *araA* and *rhaA* genes, but they would be present in MG1655. The PCR results are shown in figure 19.

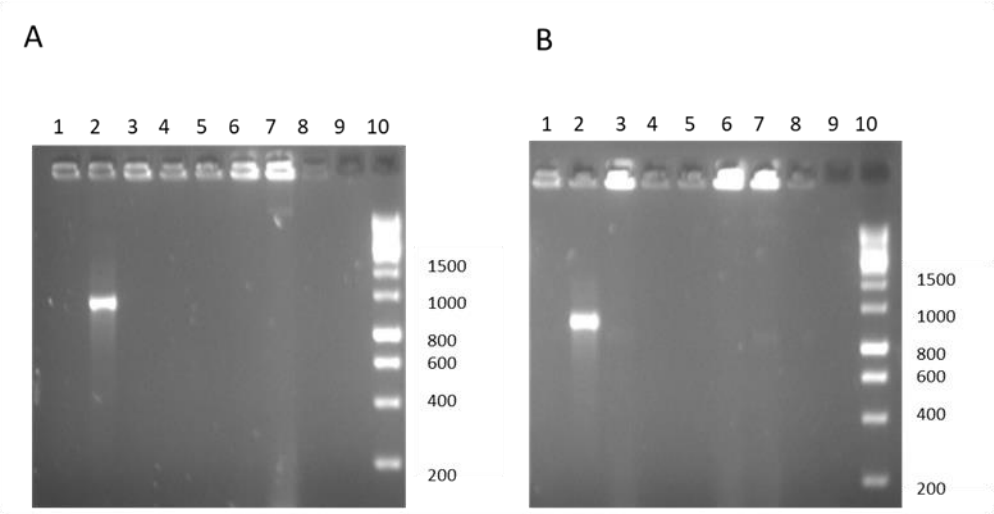


Figure 19. PCR and gel confirmation of BW25113 transduced strain.

A. Amplification of *araA* gene in MG1655 using internal primer pairs araA-F and araA-R. The expected PCR product of *araA* gene was 1.3 kb B. Amplification of *rhaA* gene in MG1655 using internal primer pairs rhaA-F and rhaA-R. The expected PCR fragment was 1.2 kb. 1: BW25113 parent strain (negative control) 2: MG1655 parent strain (positive control), 3 to 8 lanes are tested colonies 9: Negative control (water instead of template DNA) 10: Molecular weight marker (1 kb).

Another PCR was carried out to confirm the transduction of the *lacZ* genes into BW25113. The PCR amplifications of *lacAYZI* genes were carried out using LacZ-1-F, LacZ-1-R, LacZ-2-F, LacZ-2-R and LacZ-3-F, LacZ-3-R on transduced colonies. The primers designed for *lacAYZI* regions are shown in figure 20. The PCR results is shown in figure 21 and confirm the presence of *lacAYZI* gene. One of the colonies was used for the competition assays.

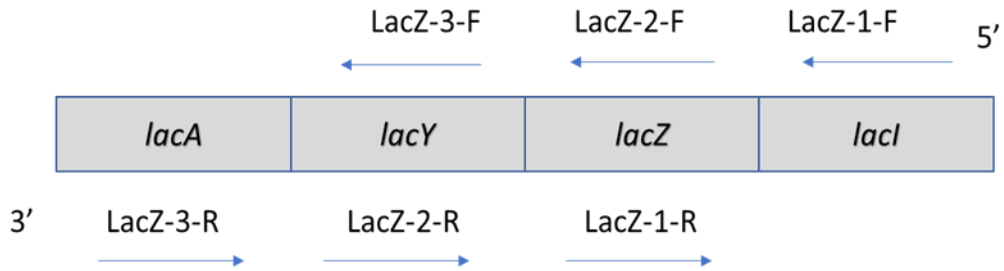


Figure 20. Primers used for PCR amplification of the *lacAYZI* genes.

Upstream primers LacZ-1-F, LacZ-2-F and LacZ-3-F were pair downstream with LacZ-1-R, LacZ-2-R and LacZ-3-R respectively. (Primers were designed byDr.Francesca Bushell)

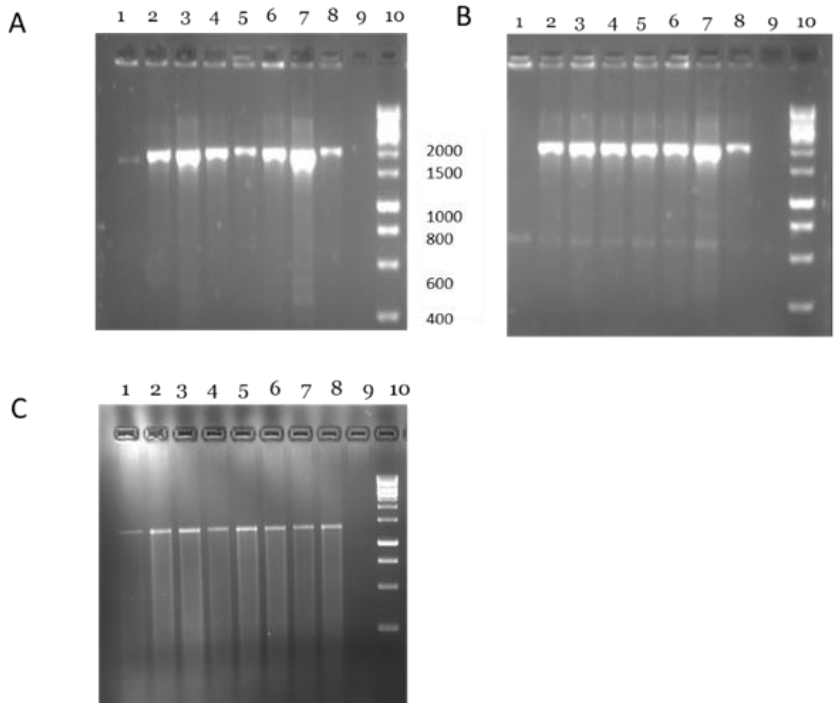


Figure 21. Confirmation of *lac* gene amplification in transduced colonies.

A. PCR amplification of *lacZ* gene using lacZ1 primer pairs, the expected product size was 2 kb. B. PCR amplification of *lacY* gene using lacZ2 primer pairs, the expected product size was 2 kb. C. PCR amplification of *lacA* gene using lacZ3 primer pairs, the expected size 1.5 kb. 1: BW25113 parent strain. 2: MG1655 parent strain. 3 to 8 lanes are tested colonies. 9: Negative control (water instead of template DNA) 10: Molecular weight marker (1 kb).

3.6 Relative fitness of BW25113 Lac$^+$

Before doing competition experiments with mutants from the Keio library, it was important to show that the fitness of wild type BW25113 is not affected by presence or absence of the *lacZ* gene at any of the pH values used. A competition experiment was therefore performed between BW25113 wild type and BW25113 Lac$^+$ in supplemented M9 medium at pH

7 and pH 5.5. The results indicate no statistically significant difference (P- value = 0.4652) between the strains at either pH, see figure 22. Since the two strains display no difference in the fitness index the BW25113 Lac$^+$ was able to be used as reference strain for competition experiment with Keio collection knockouts.

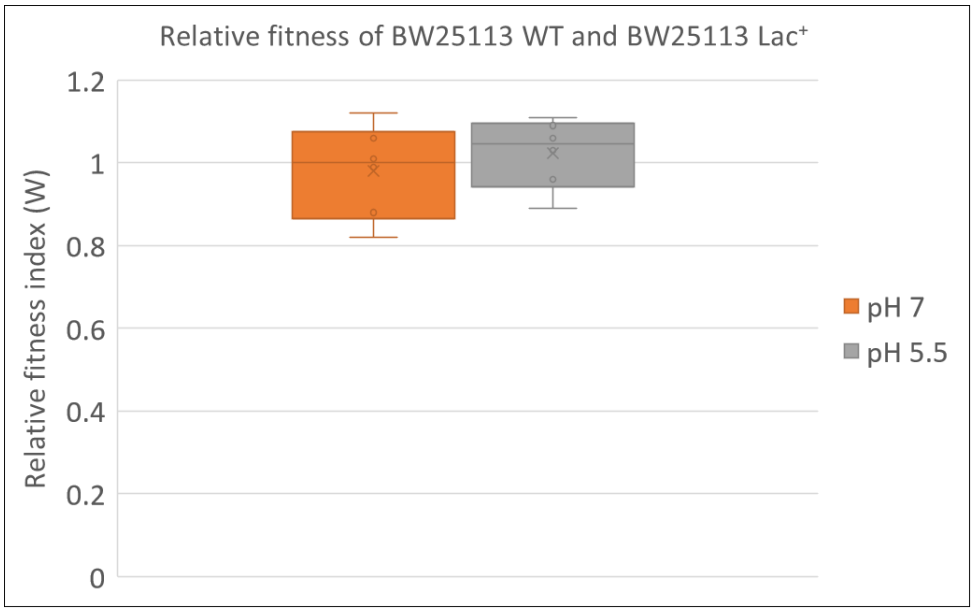


Figure 22. Relative fitness index of BW25113 wild type against BW25113 Lac$^+$ at pH 7 and pH 5.5.

The box and whisker plotted with the relative fitness values of six biological repeats. The dashed line represents a neutrality relative fitness at 1. Statistical analysis was done using an unpaired t-test (P- value = 0.4652).

3.7 Confirmation of Kanamycin insertion in the Keio collection

For doing the competition against BW25113 Lac$^+$ as the parental strain, mutants were isolated from the Keio collection. In the Keio collection, two independent mutants were made for

each gene deletion. Sometimes there was some of contamination in the collection due to poor maintenance. So, before using the knockouts from Keio collection, I have needed to ensure the Kanamycin cassette was inserted in the correct genes. To do this, four PCR confirmations of the selected genes were done. For example, kanamycin cassette insertion in *ytfP* was confirmed by four PCRs showed in figure 23; the primers used are labelled on top of each gel figure 23. The primer sequences can be found in supplementary data table S1. These PCRs were applied to all of the colonies which should contain deletions of *nuoM, pitA, rssB, sucA, nuoG, apaH* and *sthA.*

The amplification results in figure 23, showed no Kanamycin cassette insertion in Keio collection strains of *ytfP* 40C2, *rssB* 3E10, *sucA* 1E12. All the eight mutants from Keio collection showed at least one confirmed mutant, so these mutants were used for competition experiments.

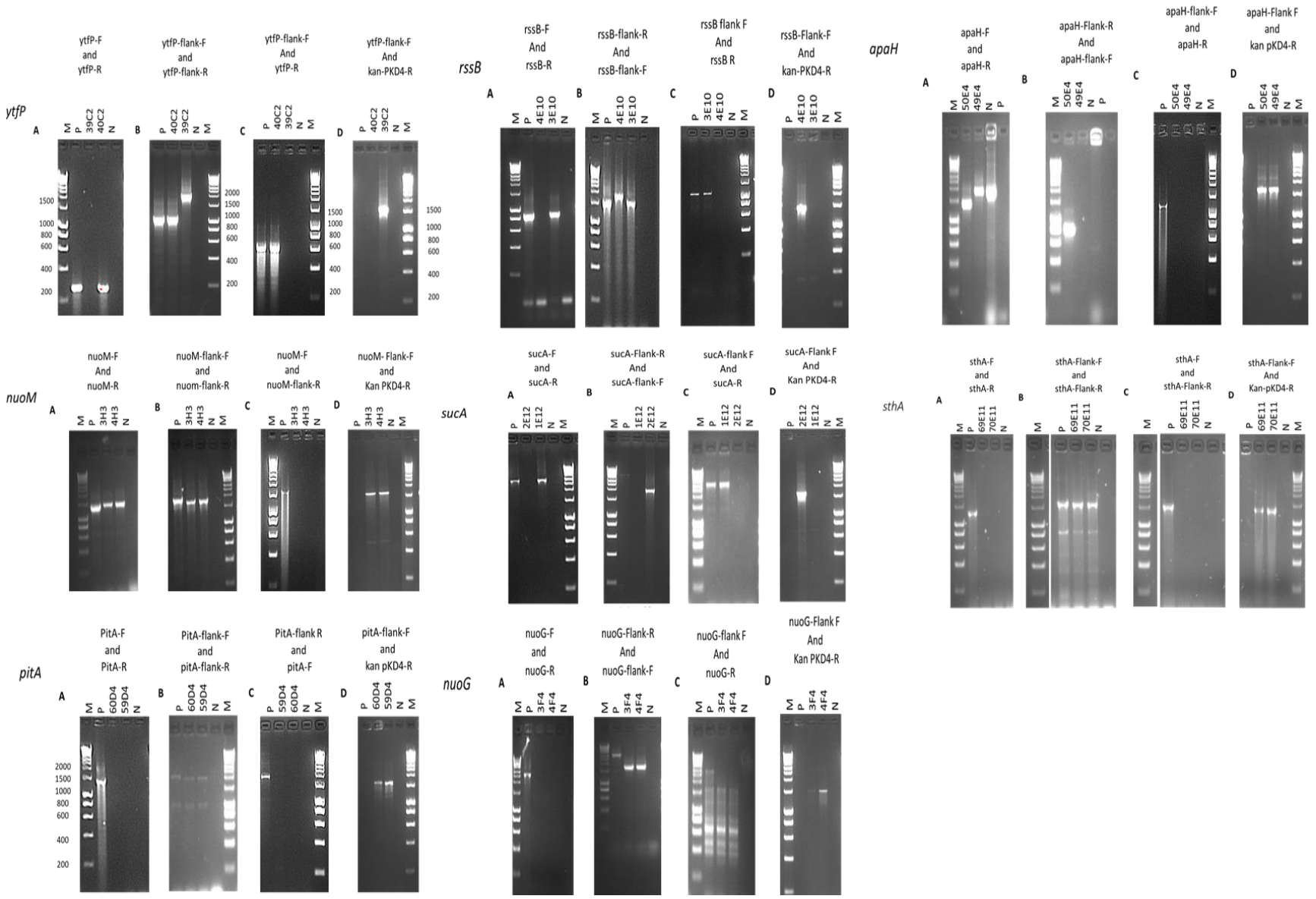

ytfP
ytfP-F and ytfP-R
ytfP-flank-F and ytfP-R
ytfP-flank-F and ytfP-flank-R
ytfP-flank-F And kan-PKD4-R
rssB
rssB-F And rssB-R
rssB-flank-R And rssB-flank-F
rssB flank F And rssB R
rssB-Flank-F And kan-PKD4-R
apaH
apaH-F and apaH-R
apaH-Flank-R And apaH-flank-F
apaH-flank-F and apaH-R
apaH-Flank F and kan pKD4-R
nuoM
nuoM-F And nuoM-R
nuoM-flank-F and nuom-flank-R
nuoM-F and nuoM-flank-R
nuoM- Flank-F and Kan PKD4-R
sucA
sucA-F and sucA-R
sucA-Flank-R And sucA-flank-F
sucA-flank F And sucA-R
sucA-Flank F And Kan PKD4-R
sthA
sthA-F and sthA-R
sthA-Flank-F and sthA-Flank-R
sthA-F and sthA-Flank-R
sthA-Flank-F And Kan-pKD4-R
pitA
PitA-F and PitA-R
PitA-flank-F and pitA-flank-R
PitA-flank R and pitA-F
pitA-flank-F and kan pKD4-R
nuoG
nuoG-F and nuoG-R
nuoG-Flank-R And nuoG-flank-F
nuoG-flank F And nuoG-R
nuoG-Flank F And Kan PKD4-R

Figure 23. PCR confirmation of knockout and kanamycin cassette insertion in *ytfP*, *nuoM*, *PitA*, *rssB*, *sucA*, *nuoG*, *apaH* *and* *sthA* from the Keio collection.

The mutant pairs from the Keio collection were tested. A. Amplification of the genes using gene specific primers B. Flanking primers of the tested gene C. Primers of gene and flanking region. D. Primer of Kanamycin cassette and flanking site. The primers used areshown on top of each gel image. M: Molecular marker 1 Kb ladder, P: Positive control BW25113 wild type, N: negative control, water instead of DNA template. The results showed no Kanamycin cassette insertion in Keio collection strain in *ytfP* 40C2, *rssB* 3E10, *sucA* 1E12. Confirmed mutants were used for competition. The sizes of bands scored in all the gels between 271 bp – 2.7 kb.

3.8 Fitness of the candidate gene knockouts from Keio collection

The aim in these experiments was to determine the phenotype of the eight candidate knockout mutants using competition experiments, and to compare the data with that from the TraDIS analysis of EO499. The competitions were done using derivative BW25113 Lac^+ vers the selected mutants from Keio collection. I have expected to see a significant decrease in the relative fitness of the mutants in the presence of acetic acid.

Competition experiments were done on eight individual knockouts: *rssB, apaH, sucA, pitA, nuoM, ytfP, sthA* and *nuoG* from Keio collection against derivative Lac^+ BW25113 on MacConkey medium, as shown in figure 24, under each of the stress conditions that had been used to generate TraDIS data (pH 7 and pH 5.5 with acetic acid concentrations of 40 mM and 4 mM, respectively, with or without acetic acid). Each assay was performed 3 - 9 times. The relative fitness index was calculated for each knockout under each condition as described in materials and methods, section 2.2.1.

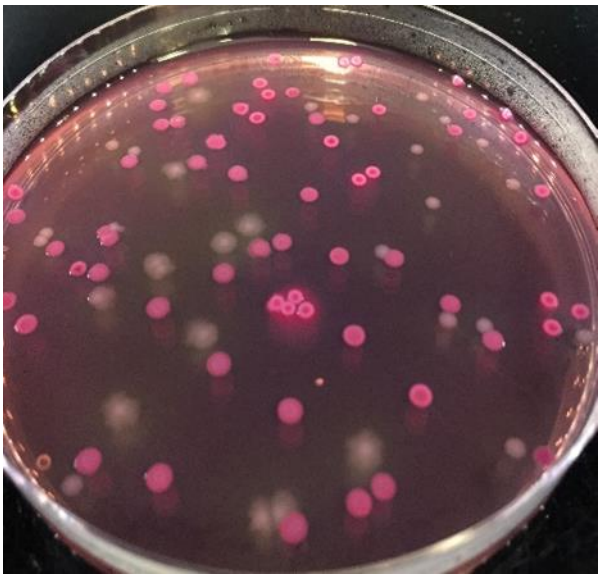


Figure 24. Results of competition between BW25113 Lac$^+$ derivative and Lac$^-$ mutants from Keio collection, as identified on MacConkey plate.

The results are shown in figure 25. Although some of mutants showed a reduction in the relative fitness index (w) such as *apaH* and *sthA* in pH 7 with 40 mM acetic acid and *rssB, apaH* and *sthA* at pH 5.5 with 4 mM acetic acid, the overall correlation was surprisingly poor with TraDIS values of relative fitness under the tested conditions.

Because of this result, I have investigated whether increasing the acetic acid stress (by increasing the concentration in both conditions to 100 mM at pH 7 and 10 mM at pH 5.5), would show a better correlation with TraDIS fitness index. Fitness reductions were expected when increasing the acetic acid concentration for the mutants. The results showed loss of fitness in *rssB, apaH* and *pitA* at pH 7 with 100 mM acetic acid, while *apaH* showed reduction in fitness at pH 5.5

with 4 mM and 10 mM acetic acid (see the results in figure 25. These findings therefore partially support our predictions.

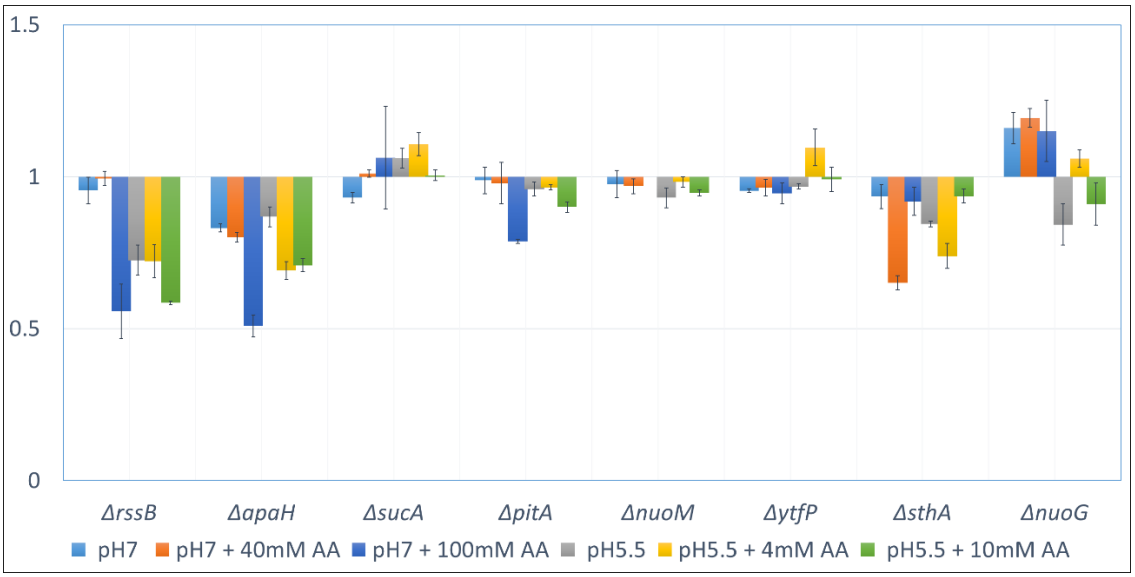


Figure 25. The relative fitness (w) index of candidate gene knockout strains under each assay condition.

A relative fitness index of 1 would show no effect; values <1 would show the mutation was deleterious relative to the wildtype. The bars show the mean of three to nine biological replicates; error bars show the standard error of the mean.

To determine the significance of the difference between the acetic acid condition and controls (e.g. pH 7 vs pH 7 + 40 mM acetic acid) for each of the mutants, Tukey's p-test of significance was applied to the relative fitness of the mutants, with the results shown in table 9. This analysis showed there was a significant effect of *sthA* at pH 7 with 40 mM acetic acid in comparison to pH 7, and of *apaH* at pH 5.5 with 4 mM acetic acid in comparison to pH 5.5. With

increasing acetic acid concentration, only *rssB*, *apaH* and *pitA* were significant at pH 7 with 100 mM acetic acid to pH 7.

Table 9. The Tukey post-hoc test indicated the significant difference between different conditions.

(Significance is shown if P < 0.05, cells colored in red). Note: acetic acid =AA.

Conditions	*rssB*	*apaH*	*sucA*	*pitA*	*nuoM*	*ytfP*	*sthA*	*nuoG*
pH 7 vs pH 7 + 40 mM AA	0.997	0.996	0.858	1	0.999	0.999	0.0006	0.995
pH 7 vs pH 7 +100 mM AA	0.003	0.0006	0.559	0.047	NA	0.999	0.999	1
pH 5.5 vs pH 5.5 + 4 mM AA	1	0.001	0.960	1	0.644	0.176	0.277	0.074
pH 5.5 vs pH 5.5 + 10 mM AA	0.741	0.060	0.948	0.854	0.463	0.996	0.411	0.964

TraDIS calculates the relative fitness based on the ratio of the mutants abundance before and after selection (e.g. pH 5.5 + acetic acid / pH 5.5), while competition experiment measures the relative fitness of a particular mutant directly in one condition (e.g. pH 5.5 or pH 5.5 + acetic acid), calculated using the formula for fitness shown in section 2.2.1. Therefore, to make a direct comparison of both sets of data easier I have divided the competition relative fitness of one mutant in stress condition over the competition relative fitness of the mutant in the control condition. Table 10 shows the relative fitness values as estimated from TraDIS in EO499 and the relative fitness as estimated from competition experiments in BW25113, for each gene. The conditions were as shown. A clear fitness reduction of mutations in *sthA* in pH 7 with acetic acid

was seen in both TraDIS and competition data, while mutations in *apaH* reduces fitness at pH 5.5 with acetic acid and mutations in *nuoG* reduces fitness at pH 5.5 in comparison to pH 7.

Table 10. The relative fitness effects of gene mutation as measured by competition experiments and TraDIS for the indicated mutants at pH 7 and pH 5.5 with 40 mM or 4 mM acetic acid , respectively.

Highlighted cell in red show cases where both TraDIS and competition data show drop in relative fitness.

Genes	TraDIS relative fitness index at pH 7 with 40 mM acetic acid	Competition relative fitness index at pH 7 with 40 mM acetic acid	TraDIS relative fitness index at pH 5.5 with 4 mM acetic acid	Competition relative fitness index at pH 5.5 with 4mM acetic acid	TraDIS relative fitness relative index of pH 5.5 to pH 7	Competition relative fitness index of pH 5.5 to pH 7
rssB	0.50	1.04	0.51	0.99	**0.93**	**0.76**
apaH	0.49	0.96	**0.42**	**0.79**	0.95	1.04
sucA_1	0.48	1.08	0.54	1.04	0.73	1.13
pitA	0.42	0.99	0.50	1.00	0.91	0.97
nuoM	0.42	0.99	0.53	1.05	0.69	0.95
ytfP	0.45	1.01	0.41	1.13	0.98	1.01
sthA	**0.26**	**0.69**	**0.45**	**0.87**	1.56	0.90
nuoG	0.38	1.02	0.59	1.25	**0.73**	**0.72**

In order for correlations to be identified, these values in table 10 were plotted in figure 26. It showed higher correlation in the presence acetic acid only (R^2=0.49), in figure 26A in compare to the 26B where the all the conditions plotted, (R^2=0.005). That could be due the reason I have selected genes generated from TraDIS data in the presence of acetic acid only. It is more

likely I would observe consistency between TraDIS data and competition data in the presence of acetic acid. In the absence of acetic acid these genes are not in top of the list for TraDIS, so this explains the low and poor effect on fitness correlation. The point is, when looking to individual genes the picture is not clear, but when looking to a number of genes there is actually quite reasonable amount of positive correlation between TraDIS data and competition measurements, with acetic acid, such as under conditions where the genes were chosen as the first genes in the list. The correlation disappears in the absence of acetic acid, which does support the hypothesis that these genes are important in giving *E. coli* the ability to survive acetic acid stress, not only low pH stress.

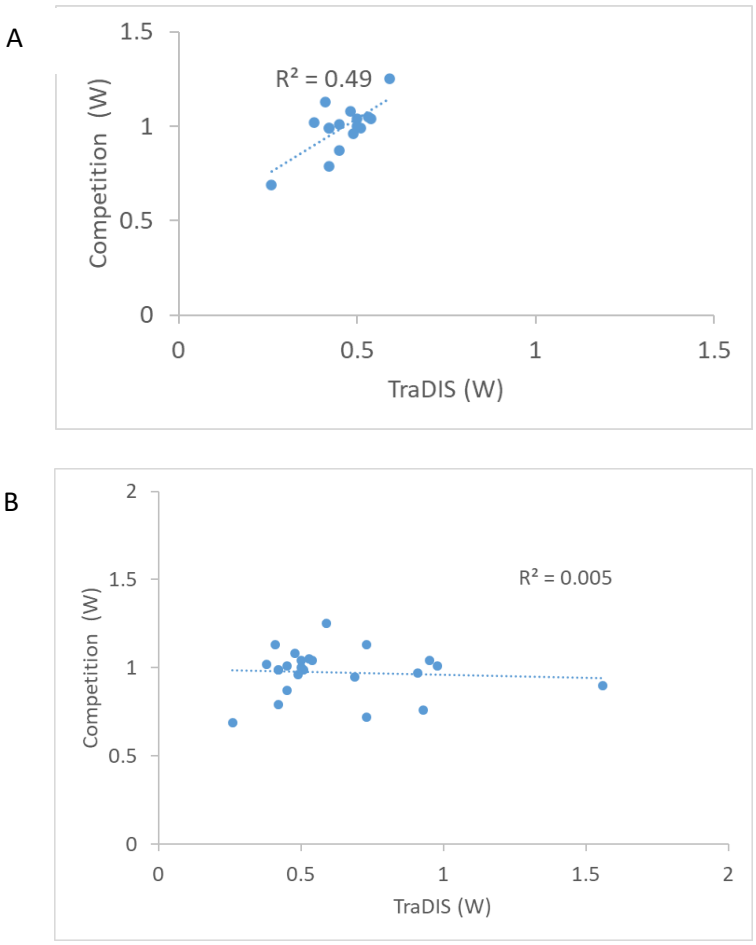


Figure 26. The TraDIS relative fitness in EO499 with competition relative fitness in BW25113, where single gene deletion from Keio collection is competed against wildtype.

A. The correlation between genes at pH 7 and pH 5.5 in the presence of acetic acid. B. The correlation between genes at pH 7 and pH 5.5 with and without acetic acid.

There was no identifiable consistent pattern of the different genes contribution to fitness in the presence of acetic acid in comparison to TraDIS. Some agreements were however seen: *sthA* was significant at pH 7 with 40 mM acetic acid and *apaH* at pH 5.5 with 4mM acetic acid, as

was seen in the TraDIS experiments. *rssB* and *pitA* also showed significant effects on fitness at higher acetic acid concentration at pH 7 with 100 mM acetic acid.

I have tried one further experimental approach to look for correlations with TraDIS data. Researchers have used time point competition experiment to study the evolution of bacterial populations under many stress conditions in changing environments to demonstrate evolutionary mechanisms (Kram et al., 2017; Pletnev, 2015). Therefore, I have decided to examine selected mutants using time course competition experiments at pH 5.5 with and without acetic acid, in which the mutants were exposed for longer to the stress. These experiments are discussed, in the next section.

3.9 Time course competition experiments:

In these experiments I have performed competition experiment where the cells transitioned repeatedly through three phases of growth during the long-term stationary phase of growth. Cells enter the stationary phase when nutrients are depleted, and this state remains relatively steady for several hours. This phase is characterized by equilibrium between dividing and dying cells. The stationary phase is followed by death phase, in which the death of cells exceeds the formation of new cells. In the stationary phase, it is possible that the wild type would cope with the stress and starvation better than the mutants under acetic acid stress. As I have repeated passages of the cells into fresh medium, I might have a better chance of seeing phenotypic changes associated with the differences in fitness (Pletnev et al., 2015; Kram, 2017).

In order to do these experiments, competitions were performed as described in the methods section using BW25113 Lac^+ and Keio collection mutants at pH 5.5 with 4 mM acetic acid or without. The cells with OD_{600} of 0.05 were grown to stationary phase and serially passaged into fresh M9 medium (diluting each time to OD_{600} 0.05) every 24 hrs for three days. During the passage cells were plated on MacConkey agar for direct measurement of relative fitness changes over time, with the results shown in figure 27. I have found that the *apaH* showed a large drop in numbers over time in the presence of acetic acid at pH 5.5. Surprisingly, *pitA* showed a greater fitness reduction under acetic acid over time at pH 5.5. This reduction was not demonstrated in the first type of competition experiment (section 3.8) or under higher acetic acid concentration at pH 5.5. *nuoM* showed growth reduction at pH 5.5 without acetic acid after 48 hrs. Note that *rssB* and *sthA* should show fitness reduction at 24 hrs time point as it is shown in figure 27, but due to the large error bars it was difficult to compare these results, and this was not investigated further. The data shows that although some mutations cause a clear additional reduction in fitness using this method, correlation with the TraDIS data is still quite poor.

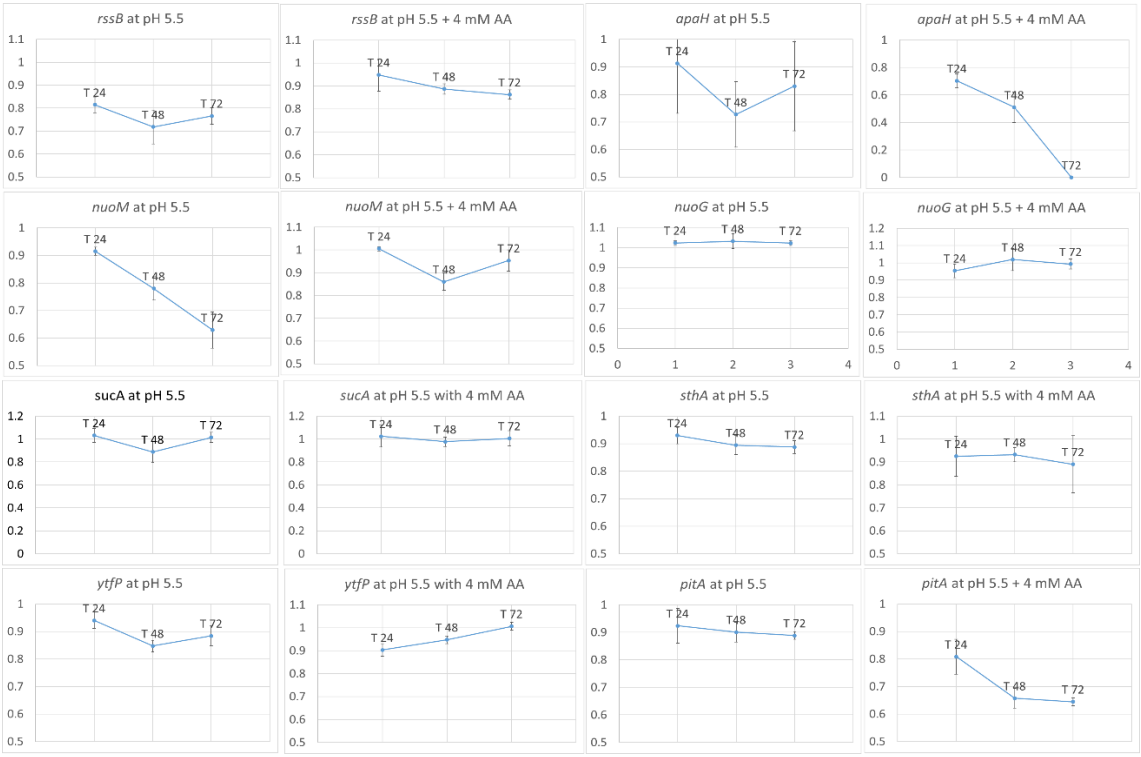


Figure 27. The relative fitness (w) of candidate gene knockout strains under each assay condition over a time course of three days.

The time point show the mean of three to nine biological replicates; error bars show the standard error of the mean.

3.10 Discussion

TraDIS analysis in strain EO499 data identified a set of genes which appear to be important for fitness in the presence of acetic acid. This chapter focused on the analysis of this previously generated TraDIS data and selection of candidate genes that may be important for fitness in the presence of acetic acid stress at pH 5.5 and pH 7. It also explored the phenotype - genotype

relationship between TraDIS relative fitness by use of competition to measure relative fitness index. The best way to test these predictions is to make knockouts in those genes to see whether the mutant strain do show such fitness defects under acetic acid stress. Since making knockouts in EO499 was extremely challenging, therefore I have attempted to validate TraDIS data in *E. coli* BW25113. In this chapter I have tried to answer the question to what extent EO499 data can be informative in other *E. coli* strains.

Also, TraDIS showed genes where mutation caused an increase in fitness of EO499 under pH 5.5 and acetic acid stress. Due to time limitations, it was not possible to validate these genes. I predict that mutations in these genes might get selected to cause some acetic acid resistance, if it was used for treatment.

The genes chosen following analysis of the TraDIS data for further validation are all ones that ranked highly in the gene lists, but also had connections with the expected effects of organic acid, as explained in the introduction. For example, *nuoG*, *nuoM, sthA* were selected for TraDIS analysis because their function is related to forming the proton gradient, which can be completely or partially collapsed by organic acid. *nuoM* is involved in proton pumping, *nuoG* is involved in proton translocation and *sthA* in reoxidation of NADPH under metabolic conditions. In particular, *nuoM* and *nuoG* (ubiquinone oxidoreductase NADH) are significant as they are part of the *nuo* operon which codes for proteins involve in energy converting: NADH ubiquinone oxidoreductase which is involved in transfering two electrons from NADH to quinone. The NADH ubiquinone oxidoreductase is the main entry point for electrons from NADH in both aerobic and anaerobic respiratory chains. The *nuo* complex structure and the electron transfer and proton translocation

flow are shown in the figure 28 adapted from (Friedrich et al., 2016). The complex contains three modules. First, the NADH dehydrogenase module in the cytoplasm assists the oxidation of NADH. This module consisting of flavin mononucleotide (FMN) protein facilitates the interaction with protein and binds a cofactor, and iron-sulfer Fe/S clusters. Then the NADH dehydrogenase module passes the electron from NADH to the hydrogenase module. The hydrogenase module is made up of six subunits NuoBCDHIN. The third module is the transporter module which consists of three subunits NuoLMK, this module transports protons across the membrane. The genes *nuoG* and *nuoM* were chosen to see if two genes under the control of the same promoter will behave the same.

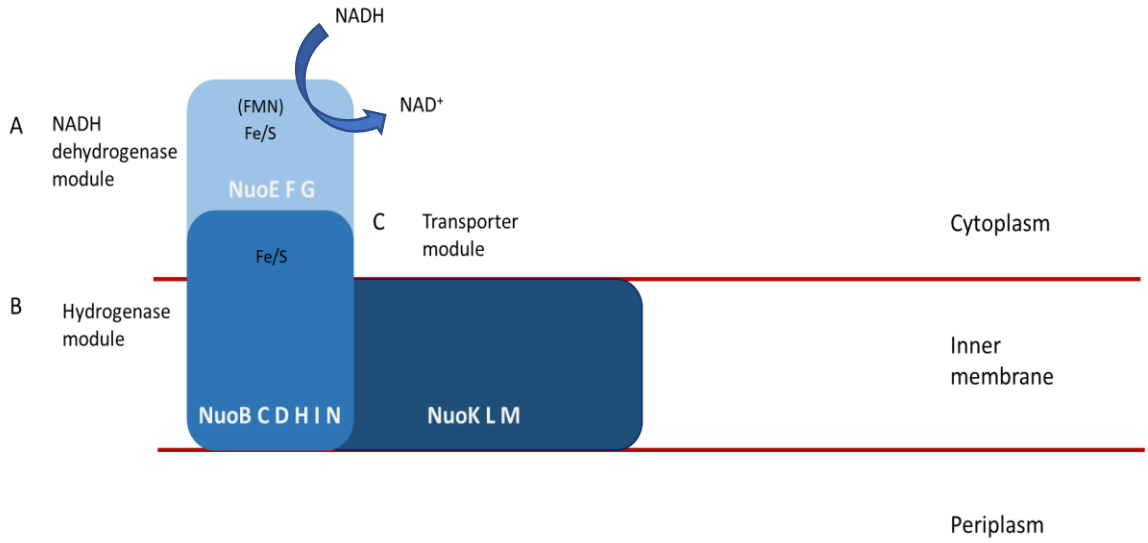


Figure 28. *nuo* **complex structure.**

A. In this complex, NADH dehydrogenase module accepts two electrons from NADH in NuoEFG subunit. B. Then, the NADH dehydrogenase module delivers the electrons from NADH to the amphipathic hydrogenase module. The module consists of NouBCDHIN subunits. C. The final transport module transports protons across the membrane and contains multi-subunits NuoKLM. Figure adapted from (Friedrich et al., 2016)

The gene *sthA* encodes a soluble pyridine nucleotide transhydrogenase in the cytoplasm of *E. coli*, which converts NADPH produced in various catabolic pathways, to NADH, which enters the respiratory chain (Zhao et al., 2008; VOORDOUW et al., 1983). The reaction catalysed by SthA is:

$$NADPH + NAD^+ \xrightarrow{sthA} NADP^+ + NADH$$

It was reasonable for mutants in *nuoM*, *nuoG* and *sthA* genes to be found in the TraDIS acetic acid stress experiments since the general mode of action of organic acid affects the proton gradient across the inner membrane, which is generated by NADH formation and oxidation.

ytfP was selected as an example of a gene of unknown function, because it provided a test for the ability of TraDIS to identify the roles of such genes.

apaH encodes diadenosine tetraphosphatase enzyme, which hydrolyzes the P^4-diadenosine tetraphosphate (AP_4A) to two adenosine diphosphate (ADP), figure 29 shows its enzymatic activity (Guranowski et al., 1983). It is known that the proton motive force generated the ATP as a result of oxidative phosphorylation, when the ADP and phosphate added together. It could be that the absence of *apaH* in the cell effect the amount of ATP synthesis which leads to lower growth rate. *apaH* was chosen as this mutation is known to have general defect in stress resistance (Lévûe et al., 1990). In another study, deletion of *apaH* in MG1655 was shown not to be able to grow at low pH 4.5 (Vivijs et al., 2016).

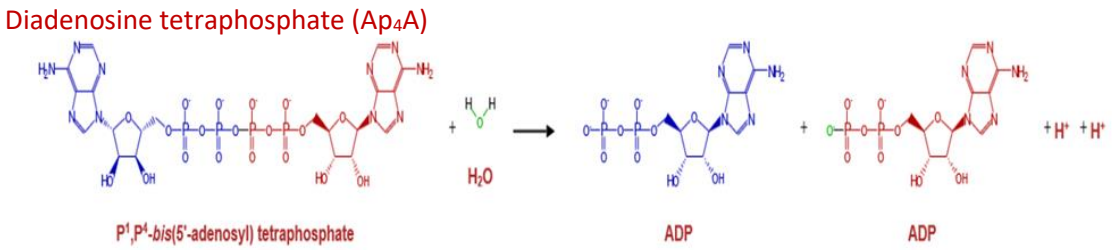


Figure 29. The reaction of hydrolysis of diadenosine tetraphosphatase to two molecules of adenosine diphosphate (ADP), catalysed by ApaH.

The figure is taken from (https://biocyc.org/gene?orgid=ECOLI&id=EG10048#)

The gene *pitA* codes for a low affinity inorganic phosphate transporter which depends on the proton motive force. The inorganic phosphate molecules can react with ADP in ATP synthesis (Harris et al., 2001). It is thus possible that in the presence of organic acids which cause loss of proton motive force, strains are unable to transport enough phosphate for their requirements and this is why mutations in this gene show reduced fitness.

The *sucA* gene encodes the 2-oxoglutarate decarboxylase enzyme which is involved in the TCA cycle (tricarboxylic acid cycle). *sucA* catalyzes the following reaction:

$$2-oxoglurate + coenzyme\ A + \ NAD^+ \ \rightarrow \ succinyl-coA + \ CO_2 + NADH$$

A mutation in *sucA* was previously showed to increase the sensitivity to acetate in comparison to the wild type in *E. coli* (Green and Emerich, 1997).

The *rssB* gene codes for a protein whose function is known to control the stability of the RpoS (σ^S), a sigma subunit of RNA polymerase which is the master regulator of stationary phase gene expression and is important under a variety of stressful conditions. The concentration of σ^S

in the cell is important to determine gene expression in stationary phase and stress resistant cells, and its stability is regulated in order to maintain the appropriate level of σ^S. The degradation of σ^S is facilitated by ClpXP protease and the recognition of σ^S by ClpXP is helped by the RssB adaptor protein to promote rapid proteolysis of σ^S. In *E. coli* K-12, three anti-adaptors (IraD, IraM, and IraP) interact with RssB in a unique way to inhibit the RssB activity in responce to different environmental stress conditions. For instance, IraD is involved in DNA damage, IraM is involved in magnesium starvation and IraP is involved in phosphate starvation. Under standard lab conditions, in *E. coli* K- 12 RssB levels are lower than the σ^S and the RssB levels increase upon entry into stationary phase. As the transcription of *rssB* itself depends on σ^S, consequently, RssB levels depend on the proportion the σ^S in the cell, and the proportion of σ^S depends on the level of RssB present to promote degradation (Becker et al., 2000; Muffler et al., 1996; Micevski et al., 2015). To clarify and summarize the molecular mechanism of σ^S and *rssB* role, see figure 30 which is adapted from (Cavaliere and Norel, 2016; Hengge, 2011).

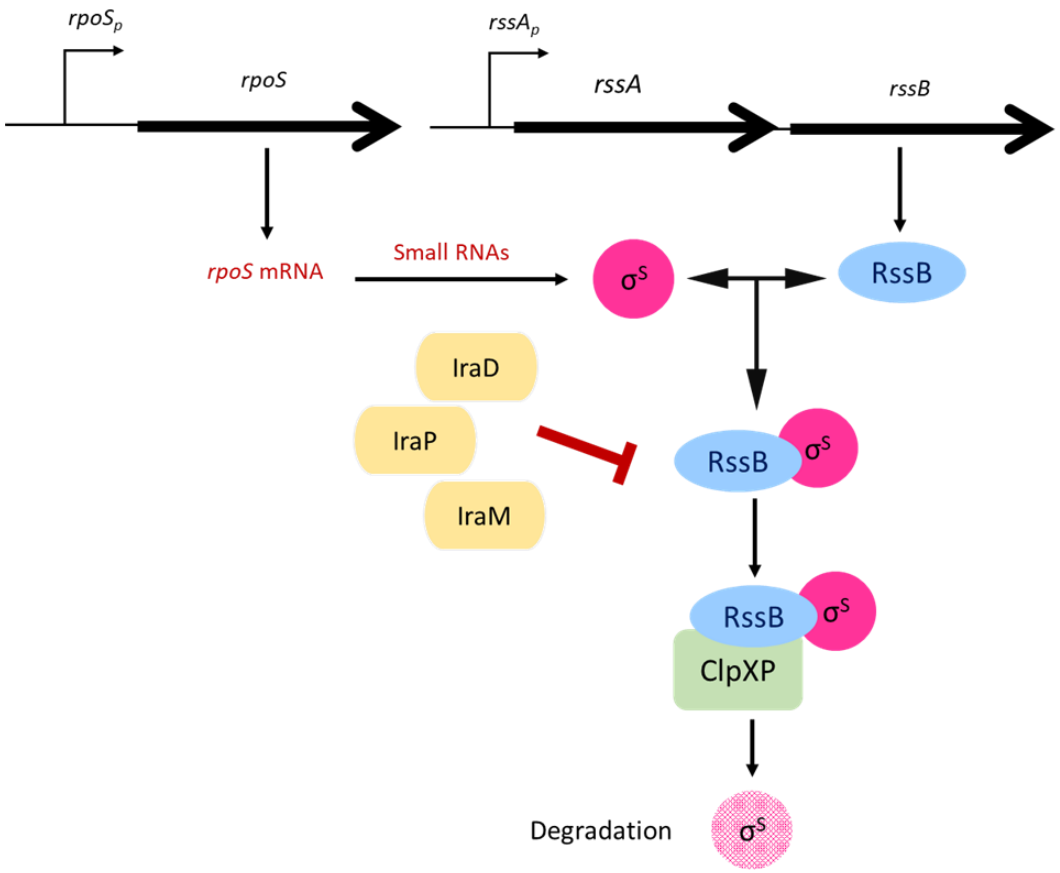


Figure 30. Molecular mechanism in the regulation of σ^S.

The translation of *rpoS* mRNA is regulated by small regulatory RNAs. The RssB protein binds to σ^S, then RssB will deliver it to ClpXP protease for degradation. This can occur unless the anti-adaptors protein (IraD, IraP, IraM) are produced and interfere with RssB protein and block the degradation process. Figure adapted from (Cavaliere and Norel, 2016, Hengge, 2011).

In an *rssB* knockout, σ^S becomes more stable (Muffler et al., 1996). In *rssB*, I therefore expected to have higher stress resistance since the cells have increased level of σ^S which no longer

will be degraded by ClpXp protease. This is the opposite of what I have found. However, one study reported that two uropathogenic *E. coli* (CFT073 and GR12) have a similar level of σ^S during log phase and stationary phase growth (Culham et al., 2001). This suggest there are differences in the timing and magnitude of the expression of σ^S between pathogenic and nonpathogenic bacteria. Also, interestingly, a fitness browser (Price et al., 2018), which has information on two hundred different BW25113 genes based on TraDIS, showed mutation of *rssB* in BW25113 showed defects under several different stress conditions including potassium acetate using transposon sequencing. Also of interest, this browser also showed defects of *nuoM*, *nuoG*, *apaH* and *sucA* mutants in the presence of 20mM potassium acetate, consistent with our data, but not *ytfP*, *pitA* or *sthA*. Another study showed that *rssB* mutants have a highly pleiotropic phenotype, which means this gene exhibits multiple effects on phenotype (Zhou et al., 2003). Therefore, our data for a role *rssB* under acetic acid stress are consistent with these other studies.

The results of the competitive relative index were unexpected in comparison to measures of relative fitness from TraDIS data. Our results indicate low correlation between the mutants' relative fitness in competition experiment in comparison to TraDIS. Among the eight selected candidate genes from TraDIS analysis, only two genes showed a reduction of fitness when deleted under the tested conditions. These were *sthA* and *apaH* at pH 7 with 40 mM acetic acid and pH 5.5 with 4 mM, respectively. At higher stress level in competition *rssB* and *pitA* were significant at pH 7 with 100 mM acetic acid. While in time course competition assay *pitA* showed a fitness drop after 48 hrs at pH 5.5 with acetic acid. In general, four mutants (*sthA*, *apaH*, *rssB* and *pitA*) out of the eight showed loss of fitness under acetic acid in comparison to the TraDIS results.

There is a one possible explanation for the unexpected poor correlation between competition in BW25113 and TraDIS in EO499. The reason is that EO499 data is not completely relevant or informative in BW25113. It has been suggested that although *E.coli* K-12 is biology's favorite model strain to study bacterial genetics, biochemistry and physiology, it unwise to generalize K-12 data to any other strains. This is because subtle differences in the promoters or coding sequences could have a huge effect on the cells (Hobman et al., 2007; Fux et al., 2005). It was not possible to determine the exact reason of this results, therefore I have attempted to isolate mutants key genes in EO499 directly from the TraDIS library for a direct comparison to TraDIS, in the coming chapter.

4 Isolation of *rssB* mutant from EO499 library

4.1 Overview

As stated in the previous chapter (3), analysis of TraDIS results identified a list of candidate genes that may affect the fitness of EO499 in the presence of acetic acid when mutated. To verify this result achieved from TraDIS analysis, ideally gene knockouts would be made of these genes. If the identification using TraDIS is correct, the knockout mutant should show altered fitness in competition with the wild type under the conditions used to generate the TraDIS data.

Two Post-doctoral researchers, Dr. Thippesh Sannasiddappa and Dr. Hadi Mohammed in the lab, had previously tried to create knockouts in EO499 but were unsuccessful. This may be due to the presence of endonucleases in the mega-plasmid that is present in this strain, that cleave the introduced plasmids or PCR products used to generate a gene knockout. Another choice is therefore to isolate the desired mutants directly from the EO499 transposon library. Considering the library contains around 1 million independent mutants, this approach is also difficult to accomplish. To establish this experiment a lot of optimizations were required to achieve the results we are seeking.

It is also the case that these mutants will not be equivalent to a gene deletion, because the transposon is introduced in the gene and the gene is still present. The position of the transposon insertion site within the gene is crucial because transposon insertions closer to the 3' end are not likely to have such a strong phenotype. Therefore, finding the mutant might not show a phenotype in comparison to a complete knockout of the gene. Despite this potential problem, it was decided to attempt to isolate specific mutants directly from the transposon library. The next

section will explain optimization of the steps needed to isolate a mutant from the library, and the isolation of the *rssB* mutant from the library.

4.2 Choosing the right master mix

Finding a polymerase or PCR master mix with the right sensitivity is critical to identify the target mutant in any large bacterial library. Therefore, different commercially available master mixes or *Taq* polymerases were tested for their sensitivity. These master mixes were provided as free samples from different suppliers upon our request. The master mixes used were A: 2x My taq red mix (Bioline, Cat. # BIO-25043), B: 2x PCRBIO HS Taq Mix Red (PCR biosystems, Cat. # PB10.22-02), C: Type it mutation detect PCR kit (Qiagen, Cat. # 206343), D: Taq DNA polymerase (Qiagen, Cat. # 201203), E: 5x FIREPol master mix (Solis biodyne, Cat. # 04-12-00115) and F: MyTaq HS mix (Bioline ,Cat. # BIO-25045).

To test the sensitivity of these master mixes the following experiment was carried out, using a strain of *E. coli* EO499 with a transposon insertion in the *ytfP* gene that had been isolated from the library earlier in the lab by Dr. Francesca Bushell. This strain was used as a control in this experiment. *E. coli* ST131. *ytfP*::Tn5 was mixed with EO499 library in 1:1 – $1:10^8$ ratios, in tenfold dilution steps, and the DNA from these mixtures was used as templates. In some cases the dilutions used were reduced to the $1:10^6$ samples, because the volume of the master mix or the *Taq* polymerases were not enough to do more tests. In order to amplify the transposon insertion within the gene, primers were designed to anneal to *ytfP* flanking region and to the

chloramphenicol transposon within the *ytfP* gene. The primers used were TnR-cc and ytfP-flanking-F listed in supplementary table S1. The 25 µl PCR contained 4 µl of template, 1 µl of 20 µM of each primer, the amount of each PCR master mix or *Taq* polymerase and PCR buffer were used according to the manufacturers' recommendations.

The PCR annealing temperature was 60°C for 30 - 35 cycles, the other cycling conditions were according to the manufacturers' recommendations. To detect whether a PCR reaction had produced a visible product of the correct size, the amplified products were resolved on 2% agarose gel. Figure 31 shows the presence of a band corresponding to the *ytfP* mutant and different ratios of the EO499 library, after using different master mixes. No amplifications were noticed with the in D: Taq DNA polymerase this might be to poor sample handiling during shipment. These results suggest that sample F (MyTaq HS mix) shows the highest sensitivity. MyTaq HS was therefore used to isolate mutants from EO499 library.

,

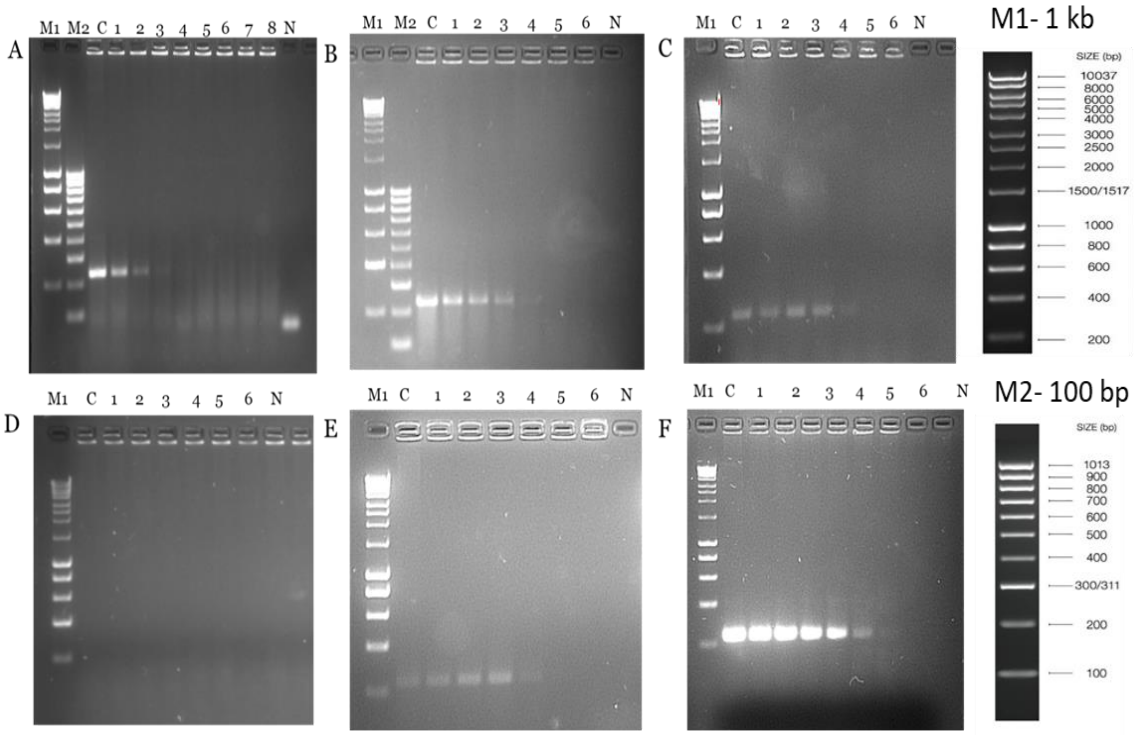


Figure 31. PCR amplification of *ytfP* gene in E. coli EO499 *ytfP*::miniTn5 using different Master mixes and polymerases.

The fragment of approximately 250 bp was amplified using different Taq polymerase or master mixes. EO499 *ytfP*::Tn5 was mixed with EO499 library in 1:1 – $1:10^6$ or $1:10^8$ ratio. M1 and M2: molecular marker Hyperladder 1 kb (Bioline) and Hyperladder 100 bp (Bioline), respectively, while C and N denote EO499 *ytfP*::Tn5 and negative controls, respectively. Lanes 1-8 denote the mutant-to-library ratio of 1:10 to $1:10^6$ or $1:10^8$.

4.3 Optimization of isolation of mutants from the transposon mutant library

In order to find the mutant of interest in the transposon library, the pooling strategy described above was used with PCR amplification. The total library (composed of 1 million independent mutants) was distributed at ca. 500 cells / well across 77 wells. The cells were distributed in the 96 well plate leaving column 12, and row H empty for pooling resulted in 77 wells of cultures. Thus, each 96-well plate would carry 38,500 mutants. Assuming that the strain contains around 5000 genes, if the transposons are evenly distributed there would be approximately 7 mutants on the plate.

10 µl of EO499 mutant library was resuspended in 20 ml of supplemented M9 medium (Materials and methods, section 2.1.1) at pH 7 which result in 1.11 x 10^8 mutant cells. Then, this culture was diluted down to 2500 cells per ml (500 cells /200 µl) in fresh supplemented M9. 200 µl was pipetted into each of 77 wells of the 96 wells leaving the margins empty (column 12 and row H). Then, the plate was incubated with the lid on at 37 °C without shaking. The next day, 4 µl from each well was pooled vertically and horizontally to the empty corresponding wells. A sterile pipette tip was used to transfer cells every time to prevent sample mix up between the wells. DNA extractions were done from the pooled cultures using the boiling method. In total 18 DNA extractions were performed on the pooled wells. Glycerol was added to 15% and 30 µg/ml of chloramphenicol to final concentration, and the 96 well plate was preserved at -80 °C for further study.

The extracted DNA from the pooled wells was used for PCR amplification to find the mutant of interest. The first attempt to isolate the mutant from the library used nested PCR

method as shown in figure 32, with the primers listed in supplementary table S1. Nested PCR strategy involved the use of two primers sets are used in two successive PCR-runs. The first primers set were designed to anneal to the region upstream of the second set PCR. The amplicon from the first PCR product was used as template for the second amplification step. The sensitivity and specify may be enhanced this technique. The idea of this method was to use nested PCR to amplify a larger band first then to use the first PCR product with the closer pair of primers to amplify the shorter band.

In this case, *rssB* flanking primers rssB-R and rssB-F were used to amplify the *rssB* gene, which was expected to be observed for all the wells with different bands sizes between 1.1 to 2 kb. Then, in the second step PCRs were carried out to amplify the transposon insertion in *rssB.* so that bands can be detected in few wells. The expectation was to amplify only *rssB* genes with a transposon insertion. This PCR reaction was done on the extracted DNA template from 500 cells per well plate and F-flanking- rssB and R-flanking- rssB in PCR 1, F- makeTnCmlong and R-makeTnCmlong were used in the nested PCR. Transposon chloramphenicol amplifying primers were obtained from Dr. Emma Sheehan.

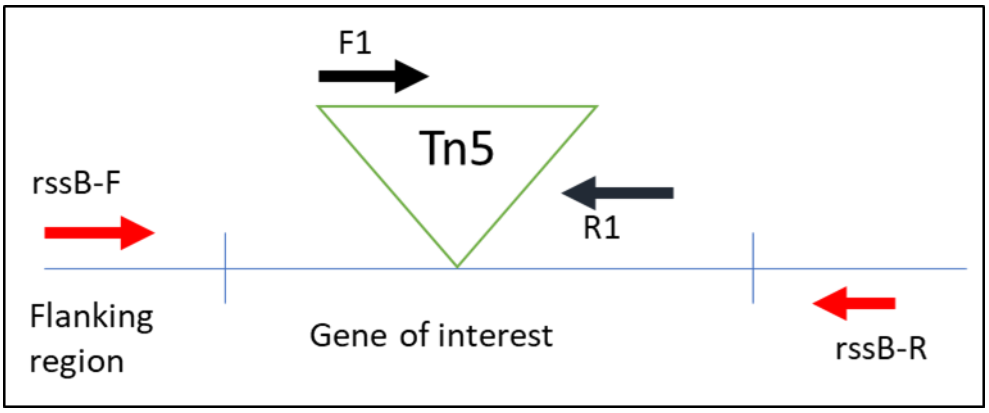


Figure 32. Nested PCR primers designed to find *rssB* mutant.

RssB-F/ R used to amplify the flanking region, F1 and R1 to amplify the transposon. F1: F-makeTnCmlong and R1: R-makeTnCmlong.

Figure 33 A shows the result of this nested PCR, with amplification of *rssB* gene product size of 1.14 kb with F-flanking-rssB and R-flanking-rssB primers, while figure 33B shows amplification of figure 33A PCR product with primers F-makeTnCmlong and R-makeTnCmlong which would give a fragment size of 924 bp if a transposon insert were present. As can be seen, in the first PCR 31-A all the wells showed *rssB* amplification, as expected. However, in figure 33B all the wells also showed a positive transposon amplification of the same size. This result didn't support the hypothesis, in which amplifying the *rssB* gene first followed by amplifying the transposon insert in *rssB* only.

The expectation was to amplify different positive sizes of *rssB* between 1.1 to 2 kb depending on the location of the transposon in *rssB*, figure 33A, as different sizes are a sign of transposon insertion within the gene. There are a few possible explanations for what happen in the nested PCR, figure 33B; Firstly, it might be there is no available *rssB* mutant within the plate, because I

expected variable band sizes between 1.1 to 2 kb if the mutant was present in the plate. But instead, the band size amplified in 31-A was 900 bp which is less than I have expected. Secondly, the mutant might be there. The presence of insertion in the gene will increase the flanking band size. In this case the bands were amplified for *rssB* without insertion. This could be because the bands corresponding to the *rssB* gene are preferentially amplified because (a) they are shorter and (b) cells containing no *rssB*::Tn5 inserts are much more common in the samples, figure 33A. I therefore came up with a different experiment designed as described in the next section.

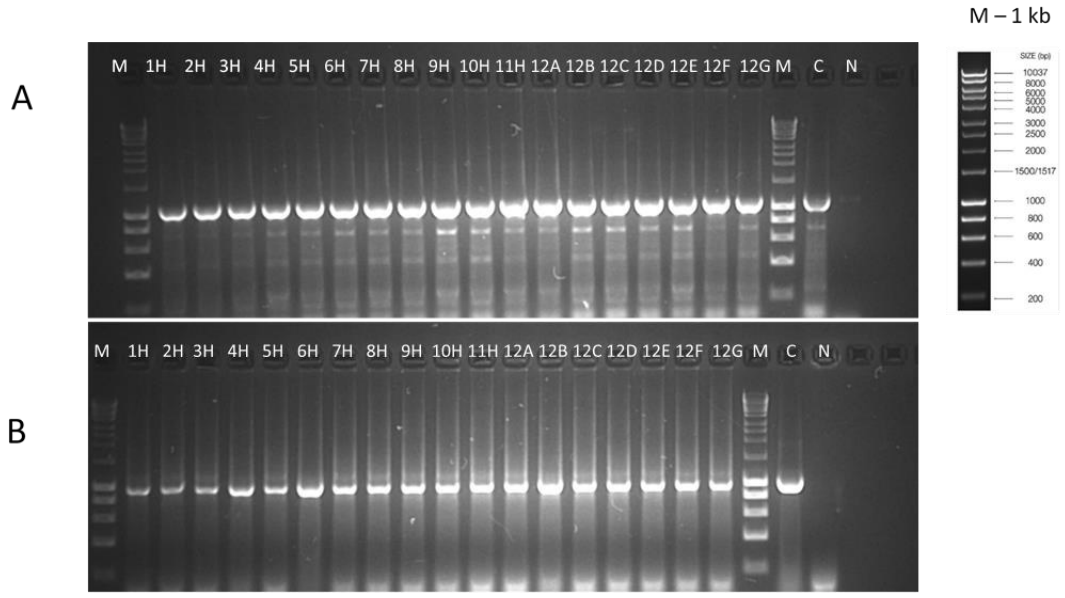


Figure 33. Nested PCR amplification of *rssB.*

M: molecular marker Hyperladder 1 kb (Bioline), 1H - 11H lanes represent the location of wells pooled vertically, 12A – 12G represents the wells pooled horizontally, C: EO499 *ytfP*: Tn and N: negative control. A. Shows the first PCR of amplification of rssB gene with F-flanking-rssB and R-flanking- rssB primers with a size of 1.1 kb. B. Transposon amplification fragment ~ 900 bp with F- makeTnCmlong and R- makeTnCmlong primers.

4.4 Alternative method for mutant isolation from the library

As the first attempt of isolating the mutant from the library wasn't successful, I have designed a second strategy. To avoid random amplification in the PCR like the first attempt, this experiment design allows more specificity and sensitivity. Different primer sets were designed for the nested PCR. These sets will anneal specifically to *rssB* mutant by using transposon specific primer and *rssB* flanking primer. The primers were designed as shown in figure 34, the first primer set were designed to amplify the larger fragment from F1 in the transposon and the flanking region of *rssB* gene where R1 anneals, while the second nested set were designed to amplify the shorter fragment F2 to R2. This way of primers design were made to make sure only *rssB* mutant will be amplified.

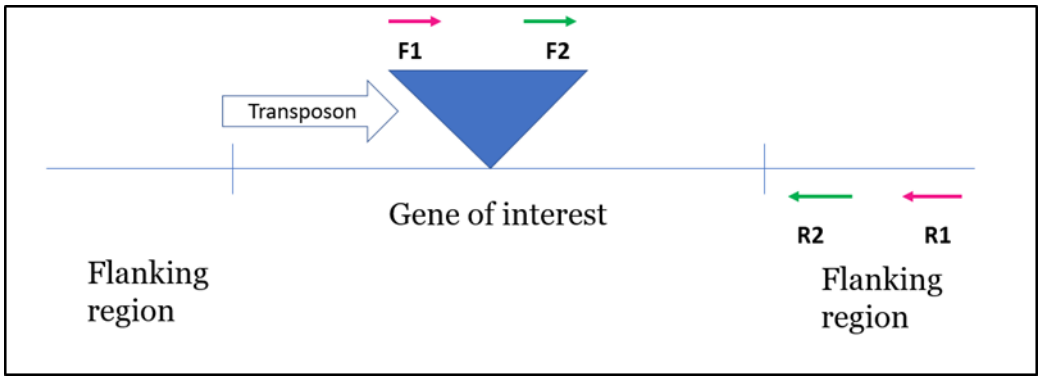


Figure 34. Primers design for nested PCR to amplify transposon insertion located in the gene of interest.

F1 and F2 refer to the transposon primers and R1 and R2 refer to rssB flanking region. F1 and R1 (pink arrows) primers are used in the first PCR to amplify the larger fragment, while F2 and R2 (green arrows) amplify the shorter fragment.

The nested PCR reaction was performed using the primers listed in table S1 and the pooled DNA prepared as described in the previous section, 500 cells per well. Tn-F2 and RssB-Flank-R2 were nested in Tn-F1, RssB-Flank-R1. The primers used were diluted to 20 µM. The nested PCR reaction used is presented in table 11 and the PCR conditions in table 12:

Table 11. The nested PCR reaction.

- Empty cells refer to no PCR components were added.

PCR components	PCR 1 - µl	PCR 2 nested - µl
5x MyTaq reaction buffer	4	4
MyTaq HS DNA polymerase	0.4	0.4
20 µM Tn-F1	1	-
20 µM RssB-Flank-R1	1	-
Tn-F2	-	1
RssB-Flank-R2	-	1
Sterile water	11.6	11.6
DNA	2	2
Total volume	20	20

Table 12. Nested PCR condition.

Process	PCR 1	PCR 2
Initial denaturation	1 min; 95°C	1 min; 95°C
Denaturation	15 sec; 95°C	15 sec; 95°C
Annealing	1 min; 57°C	1 min; 57°C
Extension	2:30 min; 72°C	2:30 min; 72°C
Cycles	30	20
Final elongation	5 min; 72°C	5 min; 72°C

PCRs were analysed by running the products on 2% agarose gels. Figure 35 shows the gel of this PCR reaction. Figure 35A shows the results from the first PCR and Figure 35B shows the results of the nested PCR. In both PCR gels a positive band appears in well 6H, with a size of 1.2 kb which corresponds to column 6 in the 96 well plate. However, no positive band was found in the pooled row from row 6. This might be because the pooled rows contained less cells than the pooled columns in the 96 well plate, which made the master mix less able to detect the positive mutant, discussed in section 4.2. In order to find the positive well with the transposon insertion in *rssB* gene, DNA was therefore extracted from each of the wells in column 6. From each well in column 6, 100 µl was used for DNA extraction. A single PCR was done with the extracted DNA using PCR1 only in table 11 and the PCR condition in table 12. As shown in figure 36, a single strong band was observed in DNA extracted from cells in well 6B with 1.2 kb size, suggesting that well 6B contains a transposon insertion mutant.

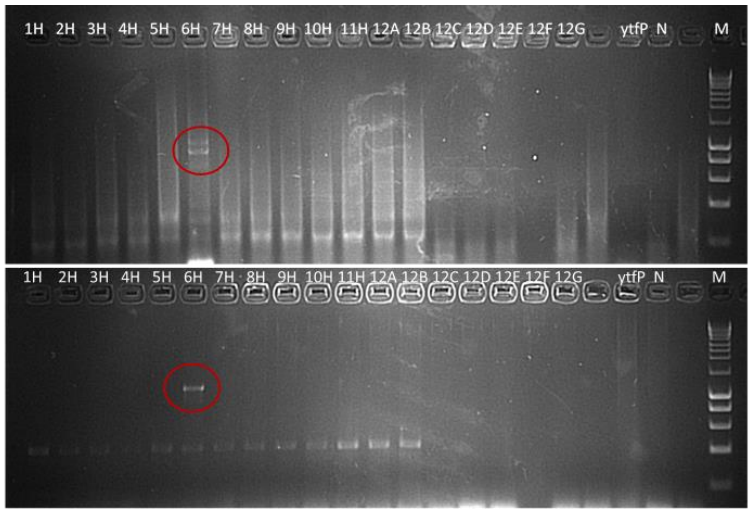


Figure 35. Nested PCR amplification of *rssB* with transposon specific primers and rssB-flanking primers.

(A): results of first PCR (B) results of nested PCR. M: molecular ladder 1 kb, 1H to 11H represent pooled column wells, 12A to 12G represent pooled row wells, EO499 wild-type, *ytfP*: *ytfP*: Tn, N; negative control (water template).

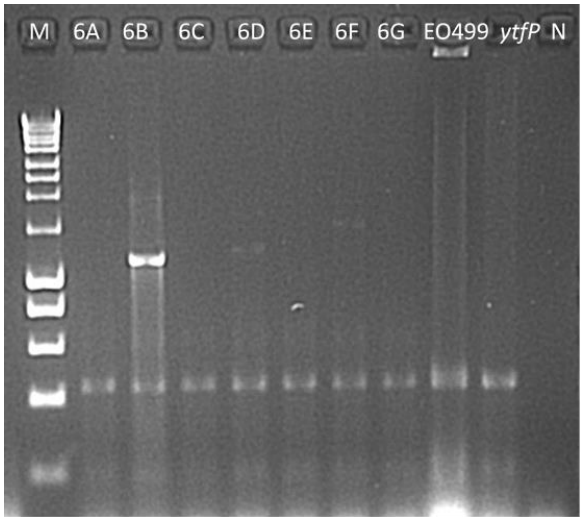


Figure 36. PCR of individual wells in column 6.

M: molecular ladder 1 kb, 6A to 6G PCR from wells A to G in row 6; EO499: DNA from wild-type; *ytfP*: DNA from EO499 *ytfP*::Tn, N: negative control (water template).

Having identified a positive well, the next step was to isolate the mutant from the mixed culture in this well. As each well had initially been seeded with 500 cells, a further step to narrow down the number of cells was needed. This was done by diluting the culture from well 6B to seed a 96 well plate with 200 µl of culture at approximately 50 cells per well. The final volume of the dilution is 20 ml, this to cover the plate, showed in figure 37. Column 12 and row H were left empty. The plate was incubated overnight at 37 °C without shaking. Next day, 4 µl of each well was pooled vertically and horizontally. PCR was done on the extracted DNA of the pooled wells as described above in section 2.5.

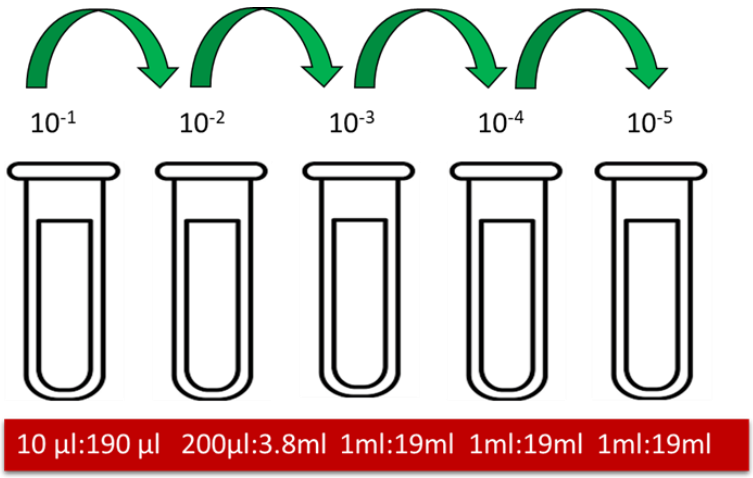


Figure 37. The serial dilution bacterial culture of well 6B.
This to dilute 500 cells per well to 50 cells per well.

The primers used were rssB-Flank-R1 and rssB-Tn-F1 using PCR1 from table 11 and the condition in table 12. I have expected a few positive wells with *rssB* mutant within this PCR reaction, because the cells were diluted to 50 cells per well. The results are shown in figure 38,

and as can be seen a positive band was seen in most of the wells with size 1.2 kb. The strongest two bands were seen in wells G12 and H4, corresponding to well G4 in the 96 well plate. The strongest bands probably results from a large amount of amplified mutants in the well. These bands were chosen to yield high concentration of DNA by purifying the PCR product required for sequencing confirmation.

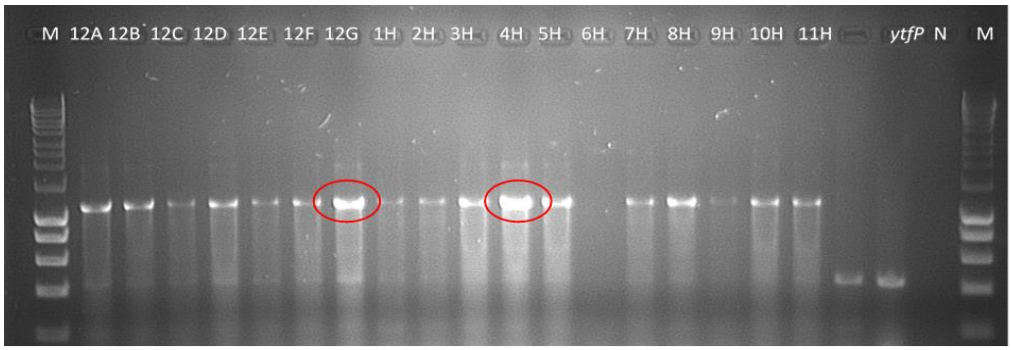


Figure 38. PCR amplification of *rssB*::mini-Tn with rssB-Flank-R1 and rssB-Tn-F1 primers. The size of the amplifies bands were 1.1 kb. M: molecular ladder 1 kb, 12A to 12G represent pooled row wells, 1H to 11H represent pooled column wells, EO499 wild-type, *ytfP*: *ytfP*: Tn, N; negative control (water template). The circled wells which correspond to well G4 were chosen for further investigation.

To verify that the positive band does result from the presence of cells containing a transposon insertion in *rssB*, the PCR product from G12 was purified and send for sequencing to Source Bioscience. The sequencing results confirmed the transposon insertion was in the reverse orientation five bases from the 5' start of *rssB* gene, as shown in figure 39.

GCAATATCTACGCTTGATTTCCATCGCGCGCACGCTGCCATTGCGGCC
GGACAGCTGGCAGTGGAAAAGAAAATGGACGAACTTTTGCCGTTGGTA
CGCACCAACATTTGACCAGAATTTTTATCTACACTTAAGTTAATTCTG
ACAGGCGTAGGTGGCAATAGCATGCCACTATTGAGTAAAGCCAGTCAG
GGGAGAGAACATGACCTGTCTCTTATACACATCTTTGGCGAAAATGAG
ACGTTGATCGGCACGTAAGAGGTTCCAACTTTCACCATAATGAAATAA
GATCACTACCGGGCGTATTTTTTGAGTTATCGAGATTTTCAGGAGCTA
AGGAAGCTAAAATGGAGAAAAAAATCACTGGATATACCACCGTTGATA
TATCCCAATGGCATCGTAAAGAACATTTTGAGGCATTTCAG

- The Flanking region of *rssB*
- Strat of transposon IR sequence
- Transposon sequence
- *rssB* gene from 5'

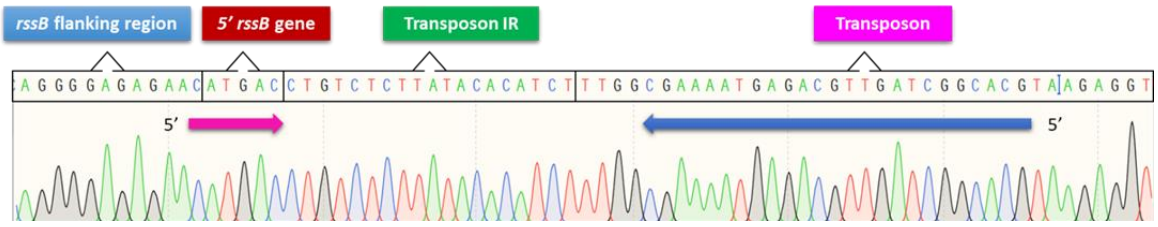


Figure 39. Analysis of PCR product from well G12.

The transposon insertion location within the *rssB* based was determined based on the sequencing results form (Source Bioscience). The transposon is located 5 bases from the 5' end.

After confirmation of the presence of an *rssB*::mini-Tn5 in well G4, the next step was to screen for a single positive colony of *rssB*::mini-Tn5 . To do this, 10 µl of well G4 was diluted to give single colonies, each colony was transferred to a single well in a 96 well plate continuing 200 µl of LBB in every well. The plate was incubated overnight at 37 °C without shaking. A colony PCR was done on each of the wells using PCR 1 from table 11 and PCR condition in table 12.

The PCR result showed a positive colony in well 9E with the expected size ~ 1.2 kb in figure 40. The PCR product was purified and sent for sequencing to confirm the transposon insertion in *rssB*::mini-Tn. The rest of the culture in well 9E was frozen in -80°C with 15% glycerol for further investigation.

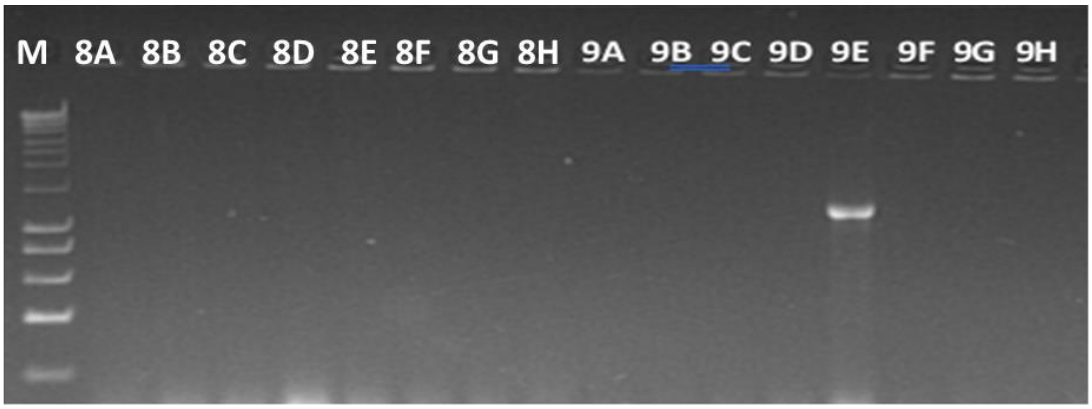


Figure 40. Identification of *rssB*::mini-Tn5.

Lanes from 8A : 9H refer to the location of the single colony in the 96-well plate. The band size 1.2 kb. M: 1 kb molecular ladder

4.5 Analysis of the *rssB*::mini-Tn mutant phenotype

The purpose of isolating the mutant from the library was to examine if it was less fit than the wild-type EO499 in the presence of acetic acid, in order to test the original hypothesis derived from the TraDIS data that this gene has a significant role in fitness under acetic acid stress. A competition experiment was therefore done between Lac$^-$ EO499 and the *rssB*::mini-Tn mutant . The Lac$^-$ mutant was isolated from the Tn library by my colleague Dr. Francesca Bushell, a former PhD student in the lab, who also showed it had the same fitness as the wild-type ancestor under all the conditions used for analyzing the *rssB*::mini-Tn mutant (Bushell, 2019). As the Lac$^-$ mutant and wild-type showed no difference in relative fitness index, the Lac$^-$ mutant was used for competition with *rssB* mutant. The results, shown in figure 41, showed no significant difference in fitness between the *rssB*::mini-Tn mutant and the Lac$^-$ EO499. The p-value was calculated using one sample t-test showed a statistically significant p-value of 0.0117 at pH 7 shown in figure 41.

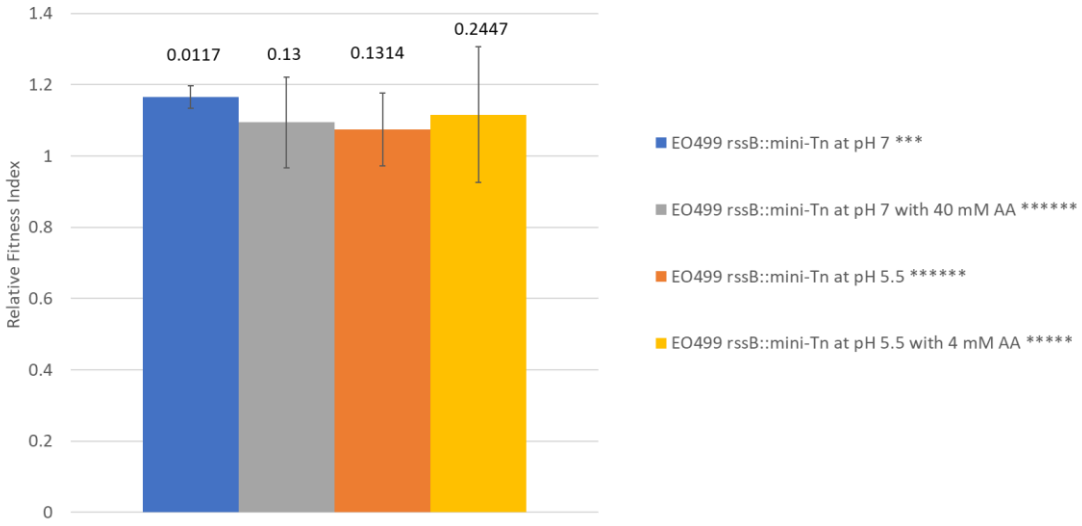


Figure 41. Relative fitness index determination for the *rssB* mutant.

The *rssB*::mini-Tn mutant of EO499 was competed against Lac$^-$ EO499 in the conditions shown and relative fitness index of (*rssB*/wild type). The number of stars in the legend showed the biological repetitions of each condition. The error bars are standard deviation of the mean. The p-value, shown on top of each column, was calculated by conducting one sample t-test compared to 1.

The results reveal that the transposon mutant of *rssB* did not show any significant in the presence of acetic acid. The significant differences were indicated by p-values < 0.05.At pH 7 the relative fitness index of *rssB* mutant was fitter and showed a significant p-value of 0.0117, but at this point the relative fitness TraDIS data in the presence of acetic acid was what was of interest. Studies found using multiple applications of t-test, may result in type I error rates (false positives), when observing a difference when the truth there is none.

So, this result is not consistent with the hypothesis derived from the results of the TraDIS experiment. It also does not correspond well with the analysis of the Keio library knockout *rssB*, as shown in chapter 3, figure 25. In this case, the results showed a decrease in fitness of *rssB* at pH 5.5 with and without acetic acid. A possible reason for this discrepancy is that EO499 is a pathogenic strain and BW25113 is a non-pathogenic lab strain; thus EO499 may have a higher tolerance to organic and low pH due to it is normally an inhabitant of the intestinal tract unlike the BW25113.

4.6 Isolating *apaH* mutant from EO499 library

As the method for isolating a transposon insert in *rssB* from the TraDIS library was successful, the same method was applied to attempt to isolate an *apaH*::miniTn5 mutant from the library. The same DNA samples that had been prepared from the pooled cells of 500 cells per well were used. The primers used in this study are listed in table S1 and the PCR reaction used followed in table 13 and PCR condition in table 14. Primers set 1 was used for the nested PCR.

Table 13. PCR reaction used to isolate *apaH* mutant.

Empty cells refer to no PCR components were added.

PCR components	PCR 1 - µl	PCR 2 nested - µl
5x MyTaq reaction buffer	4	4
MyTaq HS DNA polymerase	0.4	0.4
20 µM Tn-F1	1	-
20 µM apaH-Flank-R1	1	-
20 µM Tn-F2	-	1
20 µM apaH-Flank-R2	-	1
Sterile water	11.6	11.6
DNA	2	2
Total volume	20	20

Table 14. PCR condition used to isolate *apaH* mutant.

Process	PCR 1	PCR 2
Initial denaturation	1 min; 95°C	1 min; 95°C
Denaturation	15 sec; 95°C	15 sec; 95°C
Annealing	1 min; 57°C	1 min; 59°C
Extension	1 min; 72°C	1 min; 72°C
Cycles	30	30
Final elongation	5 min; 72°C	5 min; 72°C

The PCR results in figure 42 shows two positive bands appearing in lane 9H and 12G with a size of 1.4 kb which correspond to well 9G in the 96 well plate. Next, 50 µl of 9G was DNA extracted to ensure the *apaH* mutant is in the well.

Figure 42. PCR of pooled samples from EO499 library using *apaH* flanking primers and transposon specific primers.

M: molecular ladder 1 kb, 1H to 11H represent pooled column, 12A to 12G represent pooled row, EO499 wild-type type, *ytfP*: *ytfP*::Tn, N; negative control (water template). Bands size 1.4 kb.

Nested PCR with primers Tn-F2 and apaH-flank-R2 to detect the putative *apaH::miniTn5* mutants were performed using the extracted DNA from well 9G. The result figure 43, showed in 9G lane no amplification, this result does not correspond to the bands in pooled wells in figure 42. No further work was done on isolating mutant from EO499 library. There are a few possible reasons why such a method cannot be generalized. First, failure to amplify the right bands at the

predicted annealing temperature (T_m). The temperature was calculated by online T_m calculator (https://tmcalculator.neb.com), but the annealing temperature can't be optimized in the absence of positive control. The positive control is the strain EO499 with the desirable Tn insertion. A high temperature could cause the primers to be unable to bind to the template. Second, the absence of positive control in this experiment made it hard to optimize other PCR parameters, such as cycling times and PCR components. Third, another reason could be that there is no *apaH* mutant in the plate. I have assumed that mutants were evenly covered in the entire genome which makes 7 mutants in the 96-well plate, which is relatively few. *apaH* had an insertion index of score 0.047 and *rssB* had a score of 0.053. Given that *rssB* had a slightly higher insertion index score than *apaH* the chance of finding the *apaH* mutant in the library is slightly more challenging. The procedures could be repeated with different parameters including more plates if there were enough time and resources.

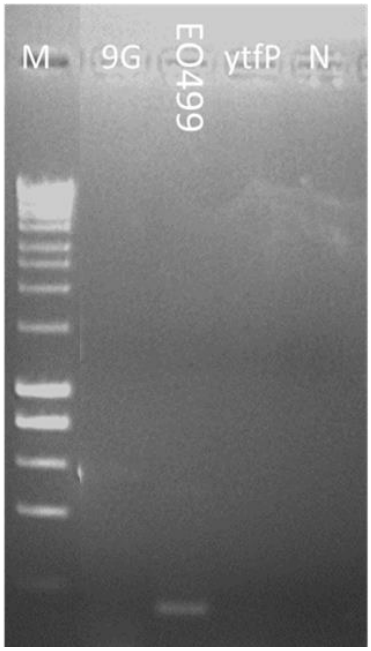


Figure 43. PCR result on well 9G.

M: 1 kb Molecular ladder, 9G: DNA from well 9G in the 96 wells plate of 500 cells/ well. EO499 wild-type, *ytfP*: *ytfP*::Tn, N: negative control.

4.7 Discussion

In this study I developed and attempted to optimize a pooling method to find TraDIS mutants. This enabled us to identify cells containing an insert in one of the genes in our list for candidates which play a part in acetic acid resistance. A similar method has been used for isolation of mutants from a transposon library with a gene-specific primer and a transposon specific primer (Holeva et al., 2004; Mesarich et al., 2017). Isolating mutants from the transposon library is

however quite a challenging technique. A few optimization variables were considered in this experiment, but the major factor was selecting the right master mix with high amplification efficiency. With this method I have managed to successfully isolate an *rssB* mutant from the library. But attempts to isolate an *apaH* mutant with the same method were unsuccessful. For the reasons mentioned in section 4.6, this method could not be generalized, considering the time and the resources spent on this method.

However, I have managed to validate EO499 *rssb::*mini-Tn5 mutant which by TraDIS was a candidate for having a role in acetic acid resistance. The biological function of *rssB* is involved in regulating sigma factor RpoS, which it delivers it to protease ClpXP for degradation in exponential phase of growing cells (Pruteanu and Hengge-Aronis, 2002). RpoS is the master regulator of general stress response and is active as cells move into stationary phase. It is responsible for the induction of at least 100 genes (Melamed et al., 2016). This known to protect the cells from nutrient starvation, hyperosmotic stress, acid and alkaline stress, heat and cold shock. Mutation in *rssB* was found to lead to RpoS stabilization (Zhou and Gottesman, 1998). On this basis it could be predicted that cells lacking *rssB* function would show greater resistance to stress, as the elevated RpoS would turn on the stress resistance pathways. This is the opposite of what TraDIS is showing, i.e. that *rssB* loss of function mutations caused reduction in fitness under the specific stress I have investigated. However, *rssB* interaction with RpoS and ClpXP are not fully understood. Therefore, I have isolated and tested an *rssB* mutant to attempt to validate the TraDIS results.

However, the results of the competition experiments with wild type strain showed it had no competitive advantage compared to the *rssB::*mini-Tn5 mutant isolated from the TraDIS library, under acetic acid stress. The *rssB::*mini-Tn5 mutant wasn't tested under extreme conditions such as high concentration of acetic acid. In the previous chapter, competing *rssB* knockout *rssB*::kan^R from Keio library against wild type BW25113 showed different relative fitness index. The results showed significant reduction in the relative fitness index of *rssB*::kan^R at pH 5.5 and pH 5.5 with 4 mM acetic acid. Moreover, significant reduction in fitness values were obtained when *rssB* knockout was tested under higher acetic acid concentration stress. These findings in BW25113 supported the TraDIS data which appeared to confirm that *rssB* has a role, particularly under acetic acid stress.

The difference between the two competition results of *rssB::*mini-Tn5 EO499 and *rssB*::kan^R BW25113 in chapter 3 might be due to the different *E. coli* strains used, as each of the strain belong to a different phylogeny group, discussed in the introduction 1.8. It might be also that (a) TraDIS data on EO499 is not applicable to BW25113 or the (b) the initial TraDIS analysis I have followed were incomplete. As validating the candidate genes list from EO499 in lab strain BW25113 was not the best option. In brief, some mutants showed fitness defect but general correlation with TraDIS data was weak.

EO499 *rssB::*mini-Tn5 mutant exhibited different behaviors in both competition experiment and TraDIS under acetic acid stress. To explain this, I have thought about the location of the transposon insertion within *rssB*. The location of the insertion was five bases downstream of the 5' start point of the gene. Insertions at the 3' end of genes sometimes have a reduced

phenotype (Goodall et al., 2018), but this is very unlikely to apply to a mutant where the transposon location was at the 5' end.

A possible explanation for this finding can be that TraDIS is a massive scale competition experiment between millions of different single transposon mutants in the library. The nature of the experiment allowed comparison between large pool of mutants providing a numerical measure of which mutants were positively or negatively affected during growth. The results of this experiment may be strongly influenced by the nutrient availability, where fitter mutants accelerate nutrient consumption. However, the proportion of *rssB* mutants is very small in TraDIS in relationship to the library size. This is different to the competition experiment done to validate the mutant, where I measure fitness in a mixture of only two strains in equal proportion of the mutant and the wild type strain, and the competition over nutrients might be less intense in this case. But studies, are able to validate transposon insertion sequencing with 1 X 1 traditional competition over a range of tested conditions (van Opijnen and Camilli, 2012). The differences between TraDIS and traditional competition is unlikely to be the reason that explain our results in *rssB* mutant in the presence of acetic acid.

Another possibility is that TraDIS carried out at 50 ml volume while the competition was done at 5 ml. Dr. Mathew Milner has compared competition of certain mutants under the same condition at 5 ml and at 50 ml volume cultures. He found some differences in results between the competition carried out in 5 ml and 50 ml cultures, but this was not investigated further. It would be an option to repeat the *rssB* competition exactly the same conditions as TraDIS experiment was performed.

ytfP mutant was another candidate mutant showed a defect under acetic acid by TraDIS. Dr. Francesca Bushell successfully isolated a *ytfP* mutant from the EO499 library but she also failed to validate TraDIS data in this matter as well. In the chapter 6, repeating EO499 TraDIS, *ytfP* mutant also showed reduction in fitness with a logFC of 0.66. As Dr. Francesca did *ytfP* competition in 5 ml culture, which support the above point.

Considering I was unable to create a knockout in EO499 due to the presence of endonucleases in the mega-plasmid that is present in this strain, one option was to validate EO499 strain cured of the plasmid to create a gene knockout of the detected genes to validate our TraDIS data. The cured strain was available in (T102 lab – School of Biosciences - University of Birmingham). But Due to the long time was spend on validating the EO499 library, I have decided to repeat the exposure of EO499 library to acetic acid to investigate the data reproducibility at different time point. Also, I wanted to investigate the reproducibility of different analysis method used. The details are discussed in chapter 5 and 6.

5 Transposon Sequencing Libraries

Declaration:

The focus of this chapter is the analysis of the generated sequencing libraries data. UTI89 library transformation optimization was carried out by me. MG1655 transposon library was constructed by Dr. Mathew Milner, EO499 was constructed by Dr. Keith Turner and UTI89 was constructed by me with help of (Dr. Mathew Milner, Shahida Butt, Dr. Maria Masoura, Dr. Santosh Kumar, Bakul Piplani). Sequencing of the initial transposon libraries: MG1655, EO499 and UTI89 were sequenced by me. The sequencing libraries files were processed with TraDIS pipeline by me with a script written by Dr. Ashley Robinson. The essential gene predictions for the sequencing libraries were processed by me with script written by Dr. Sara Jabbari. The analytical scripts were used with the help of Mathew Milner. The genomes were annotated by me using Prokka pipeline. The gene presence and absence were carried by Dr. Mathew Milner using Roary. In this chapter, scripts written by Dr. Mathew Milner were the GFF3 changer and Sequencing depth saturation. Further analysis and comparison between the three Tn libraries were completed by me. The trouble shooting in UTI89 libraries were done by me except the correction of UTI89 was done with help of Dr. Mathew Milner.

5.1 Overview

Tn-Seq is a tool that combines transposon mutagenesis with Illumina next generation sequencing. Transposon sequences are aligned to the reference genome to identify the location of the regions flanking the insertions and hence the position of the insertion. Transposon mutant libraries are constructed with the aim of identifying genotypic and phenotypic features of bacteria under certain specific conditions. In another word, this approach allows investigation of all non-essential genes to assess the effect of a loss of function of a gene under a tested condition. In the initial transposon library, genes that contains few or no transposon insertions are likely to be essential. Essential genes are those required for cell growth and viability, or nearly so, in the growth conditions that were used in the study. The term conditionally essential refers to genes required for growth under one growth condition but not another.

Essential genes can be grouped into two categories: core essential genes and accessory essential genes. The core essential genome can be defined as a set of genes present in all strains of the species and universally required in all the strains. While the accessory essential genome is genes present only in a subset of the species but are indispensable for survival of these strains. Identifying the accessory essential genes could lead to promising drug or vaccine targets in bacteria, as targeting these genes would not affect related species where these genes are absent or non-essential. Inactivating essential genes can cause inhibition of the metabolic or regulatory pathways which ultimately leads to cell death. Some of these pathways may include crucial transport and catalytic proteins which could also be promising new targets for antibacterial drug

discovery. Knowing the absolute essential genes is also important in order to exclude them from the conditionally essential analysis, which will be discussed in depth in the next chapter.

There are several techniques to identify essential genes. These methods can be classified as non-transposon-based methods or transposon-based methods. Non-transposon-based methods include comparative genomics approaches which involve targeting a gene-knockout of specific genes after bioinformatic analysis. The premise behind this approach is that genes of unknown function that are highly conserved between strains and species are more likely to be found to be essential than non-conserved genes. An alternative approach would be to predict gene essentiality based on genomic information from the sequence of the gene itself (Arigoni et al., 1998). Another approach used is genome-wide knockout by attempted construction of all single gene knockouts e.g. by homologous recombination with PCR products. This type of single gene deletion of all genes was done in BW25113 and genes which were unable to be disrupted were classified as essential (Baba et al., 2006). Transposon-based methods such as saturation transposon mutagenesis have been used over the past few years for identification of essential genes. There are many different methods of transposon mutagenesis which are described in depth in here (Van Opijnen and Camilli, 2013). For example, many studies have identified essential genes in *E. coli*, the most recent one identified essential genes in *E. coli* K-12 strain BW25113 using TraDIS using visual analysis of insertion index scores (Goodall et al., 2018).

In general, usually genome-wide knockout has been relatively rarely done because of limitations in funding resources, time and labor. In contrast, random transposon mutagenesis libraries and

Transposon sequencing can be simply constructed in a new organism to with minimal effort and much less labor.

It was difficult to come to firm conclusions from the results that have been presented in chapter 3 and 4. In chapter 3, TraDIS analysis of EO499 grown in M9 + 0.2% casamino acids and 0.2% glucose at pH 7 or pH 5.5 with acetic acid at 40 mM and 4 mM, respectively, or without added acetic acid, was done. Several candidate genes in EO499 were identified as being important for fitness under the examined conditions. As it is challenging to make knockouts in EO499 strain, I initially used Keio collection knockouts in *E. coli* K12 BW25113 to determine the relative fitness of candidate genes by competition assay under the same conditions that had been used for TraDIS. The competition experiments were done under a range of different conditions: pH 7 and pH 5.5 with and without acetic acid, 40 mM and 4 mM, respectively. Because of the initial difficulty in identifying fitness difference, competitions were also done over a longer time period (3 days). In addition, they were also performed under higher concentration of acetic acid 100 mM at pH 7 and 10 mM at pH 5.5. However, the fitness values determined from all the examined competition experiments with knockouts in BW25113, showed a partial positive correlation in the presence of acetic acid with fitness values as identified in TraDIS analysis of EO499. The use of BW25113 as a proxy strain can be partially justified, but a better approach would be to construct the gene knockouts in the EO499 where the TraDIS experiment were carried out. As a consequence, I have to re-generate EO499 TraDIS library. In chapter 4, where a specific transposon mutant was isolated from the EO499 library for the purpose of validating TraDIS, the isolated mutant also showed no phenotypic effect under acetic acid stress.

Given the results in chapter 3 and 4, I have decided to generate new data from EO499 under acetic acid stress, with longer exposure to stress than 24 hours. This can be done using a time course experiment where the transposon library is grown and diluted repeatedly over several days. The aim here is to study the effect of acetic acid on EO499 over a longer period of time. Additionally, I have wished to compare the newly generated data to the previous TraDIS results generated by Dr. Francesca Bushell in her PhD study, to see the extent to which the data are reproducible when similar experiments are done by different people. Additionally, I have wanted to compare the candidate genes identified from TraDIS in EO499 under acetic stress with those detected in different *E. coli* strains using the same method, to see how much they are conserved between different *E. coli* strains. In order to do these experiments, transposon mutant libraries were constructed and sequenced, and then the sequencing results were analysed to determine the essential and non-essential genes within each tested strain. Then, the new mutant libraries were subjected to acetic acid stress as had been done previously with EO499. This chapter presents results analysing gene essentiality from the initial sequencing libraries among three *E. coli* strains (EO499, UTI89 and MG655). The next chapter presents the results obtained from TraDIS experiments done under acetic acid stress over time on all three libraries.

The focus of this chapter is to analyse and compare essential genes in three suggests genetic selection *E. coli* strains: lab strain MG1655, uropathogenic EO499 and uropathogenic UTI89. To do this, I have constructed a new transposon library in UTI89. As far as I am aware, this is the first direct comparison of two independently made TraDIS libraries in the same EO499 UPEC strain of *E. coli*, and it gives important evidence about experiment reproducibility. This represent

two independent experiments on the same library by two different people. Using the same library same and same conditions but different worker. And the interesting question is how this do these two results compare.

The genome analysis classification was applied using the same method as described previously in (Goodall et al., 2018). In this chapter, I will view and discuss the results of genome annotation of MG1655, EO499 and UTI89, the transposon library construction sequencing libraries, the analysis method applied, and the result from the comparison of three independent UTI89 libraries.

5.2 Library transformation and optimization of UTI89

The transposon mutant library of UTI89 was constructed in this study, while the EO499 library was constructed and kindly donated to us by Keith Turner and MG1655 library was constructed and sequenced by Dr. Mathew Milner. UTI89 and MG1655 libraries were constructed using a mini-Tn5 transposon with a kanamycin resistance cassette, and the EO499 library was constructed with a mini-Tn5 transposon with chloramphenicol resistance cassette. A study has been done in the lab (T101 lab – School of Biosciences - University of Birmingham) by Dr. Emily Goodall to determine whether the use of different antibiotic markers has any effect on the output, but the study has not been published yet. In theory, no differences are expected.

The UTI89 strain used for library construction was provided byDr.Swaine Chen from National University of Singapore. Many studies have been performed on this UTI89 strain (Wright

et al., 2007, Justice et al., 2006, Anderson et al., 2003). A modified method was used to construct a transposon mutant library in this strain (Langridge et al., 2009). This modified method is illustrated in method section 2.6.

Constructing the UTI89 library required highly efficient electrocompetent cells to achieve a high-density transposon library. Several factors have been identified that cause an impact on the efficiency of electroporation transformation. Some of these factors include: the cells' optical density, the electrical field strength (voltage), preparation temperature, cell or plasmid type, type of buffer and the amount of nucleic acid to be transformed (Dower, 1988; Tu, 2016). To test UTI89 transformation efficiency before constructing the transposon library, UTI89 and MG1655 competent cells were prepared exactly as described (Goodall, 2019). These competent cells were transformed using the Kanamycin Mini-Tn5 transposon-transposase complex. Colony count showed that the efficiency of UTI89 was about half of MG1655. This showed the transformation efficiency for UTI89 is lower than MG1655 using mini-Tn5. This result suggested that further optimization is required to yield the best results of transformed colonies.

Further transformation optimization experiments were carried out using different parameters to increase the transformation efficiency of UTI89.Dr.Keith Tuner advised us that for obtaining high transformation efficiency, cells should be harvested with an OD_{600} of 0.2 – 0.3, early log phase. Moreover, the transformation voltage and Kanamycin concentration in the culture plates were optimized. Tn5 was used to transform UTI89 cells at different transformation voltages, with the results shown in table 15. The optimization range of voltages applied were

between 1.4 kV to 2.5 kV, because the Eppendorf Eporator recommended using 1.7 kV to 2.5 kV in the product supporting information. As in (Goodall, 2019) the transformation was carried out at 2.2 kV, so I have decided to test the transformation voltage to 2.5 kV. At 2.5kV, the highest number of transformed colonies were obtained, so this voltage was selected for library construction. A study showed that preparing competent cell at room temperature increased the transformation efficiency (Tu et al., 2016). To test this, competent cells were prepared at room temperature and transformed at 2.2 kV. This showed higher number of transformed colonies in comparison to the 2.2 kV ice-cold prepared competent cells. In a small experiment of Kanamycin concentration in the culture plates of UTI89, the untransformed UTI89 cells were plated at a range of kanamycin concentrations (10 – 30 µg/ml), small colonies appeared up to kanamycin contraction of 19 µg/ml, so 25 µg/ml were chosen for selection. This experiments was done to determine the lowest Kanamycin concentration that would inhibit all growth of the non-transformed cells.

Table 15. The parameters used for transformation optimization in UTI89.

Average number of colonies and standard deviation obtained by transforming Tn5 to UTI89. Transformation were carried out at different voltages from 1.4 kV to 2.5 kV. The amount loaded in the plate (µl) = 300 µl from the recovery transformed cells contains: 60 µl of the competent cells, 0.2 µl of transposome, 1 ml of SOC medium and with 2.5 ml of LB broth, as shown in methods and materials 2.6.1 and 2.6.2. Each experiment was performed in biological triplicate.

Condition	Measures	Average number of colonies	Standard deviation	Amount loaded in the plate (µl)
Transformation voltages	1.4 kV	29	4.1	300
	1.6 kV	22	3.7	300
	1.8 kV	32	2.5	300
	2.2 kV	296	28.3	300
	2.2 V at 28 °C	382	50.3	300
	2.5 kV	475	11.2	300

5.3 Genome annotation

A genome sequence .FASTA file of EO499 was obtained from Dr. Thippesh Sannassidappa. This strain had been sequenced using PACBIO when Dr. Sannassidappa was a member of our lab. MG1655 and UTI89 were sequenced by the MicrobesNG facility in University of Birmingham.

As I have mentioned earlier TraDIS relies on mapping the sequenced reads to the reference genome to identify the unique position of the transposon and relative number of insertions. For this reason, to get full information the reference genome needs to be well

annotated. (Genome annotation is the process of identifying the locations of coding regions sequence (CDS) in the genome and assigning functions to them where possible). CDS is DNA region that code for protein, this is indicated from the start point of initiation codon to the end point termination codon location.

The available genome annotation of UTI89 was done using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP), While the available genome annotation of EO499 was annotated by Rapid Prokaryotic Genome Annotation PROKKA, provided by Dr. John Herbert. Different annotation pipelines can produce different results, an example of this can be found in (Seemann, 2014). To overcome this issue, I have re-annotated the genomes to obtained unified genomic features (Seemann, 2014).

Genome annotations of MG1655, EO499 and UTI89 were done using PROKKA (Seemann, 2014). MG1655 gene bank file .gbk (NC_000913.3) was used as reference genome to guide the annotation in PROKKA for MG1655, EO499 and UTI89, because MG1655 is a well-studied model bacterium for complete genome annotation (Riley et al., 2006).

For a better understanding of the conserved genome and the accessory genome of the three annotated strains, I have used Roary, a rapid pan-genome pipeline that uses PROKKA output .GFF3 files to generate gene presence/absence matrix in .csv format. The. GFF3 of EO499, MG1655 and UTI89 were used as input for Roary to obtain a gene presence/absence list. Processing via this pipeline is illustrated in Materials and Methods 2.7.3. Details of the gene

presence/ absence matrix can be found in (http://sanger-pathogens.github.io/Roary/). The gene presence/absence matrix lists each gene and which samples it is present in (Page et al., 2015).

Roary produced a total of 6,606 protein-coding gene sequence clusters with 51.7 % of the genes considered as a core gene presented in all three strains. The core genes are set of genes shared in all the strains of a species. Most of these genes are related to basic cellular functions and typical phenotypic properties of the organism (Meng et al., 2017). Genes which were present in at least two strains or more genome but not all were labelled as accessory genes. In fact, it would be possible to find an accessory gene that is essential for viability in one strain but not the other. This is part of this chapter's investigation. The accessory genes are acquired by horizontal gene transfer or by mutation of pre-existing genes. If the genes arise from mutation (i.e., single nucleotide polymorphisms (SNPs) of pre-existing genes, it would still be very homologous to them, in which they will be identified as core genes. Normally, they are involved in adaptive functions such as specific metabolism, virulence, and antibiotic resistance mechanisms. While genes found in only one strain were labelled as unique genes (Raskin et al., 2006, Meng, 2017; Martínez-Carranza, 2018). The non-core genomes with 3187 genes were divided into two categories: 749 accessory genes which present in two strains and 2,438 unique genes present in one particular strain, figure 44. The unique gene lists for EO499, MG1655 and UTI89 identified by Roary found in table S6, table S7 and table S8, respectively.

A

	Roary output analysis	
	Protein-coding gene sequence clusters	6,606
Core genome	Genes present in all three strains, core genome	3419
Accessory genome	Genes present in MG1655 and EO499	116
	Genes present in MG1655 and UTI89	78
	Genes present in EO499 and UTI89	555
Unique genome	Genes unique in MG1655	705
	Genes unique in EO499	945
	Genes unique in UTI89	788

B

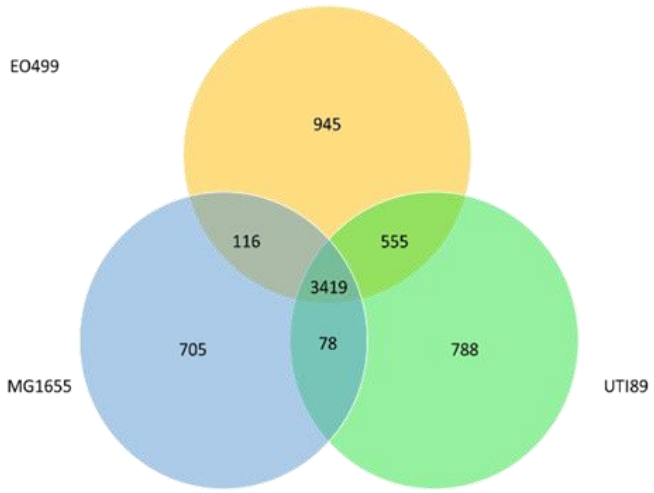


Figure 44. Genomic comparison of EO499, MG1655 and UTI89 by Roary.

Core genome analysis of the three strains. The numbers indicate the core genome, accessory genome and unique genome among the indicated genomes. A. table, and B. Venn diagram.

5.4 Transposon library mutant construction

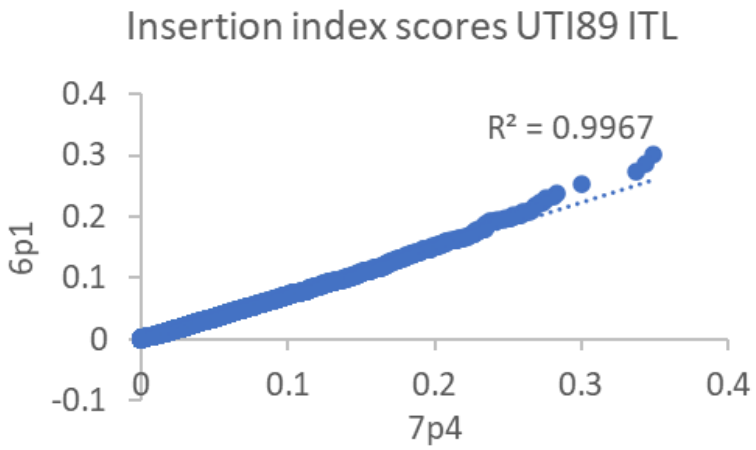
The UTI89 transposon sequencing library was constructed with help from Dr. Mathew Milner, Dr. Maria Massoura, Shahida Butt, Dr. Santosh Kumar, and Bakul Piplani. Transposon library prep started with transformation of a mini-Tn5 transposon with a kanamycin resistance marker into electrocompetent cells, followed by plating onto selective medium and overnight growth. After incubation, individual colonies were pooled to constitute the library. Some plates showed large numbers of individual colonies and some showed a lawn of growth, which made estimation of total numbers difficult, but I have estimated approximately 180,000 to 200,000 mutants were pooled in this way.

For sequencing and comparison of the different strains, three technical replicates DNA extractions of the UTI89 transposon libraries were prepared. Two independent library preps were also made from the MG1655 and EO499 transposon libraries.

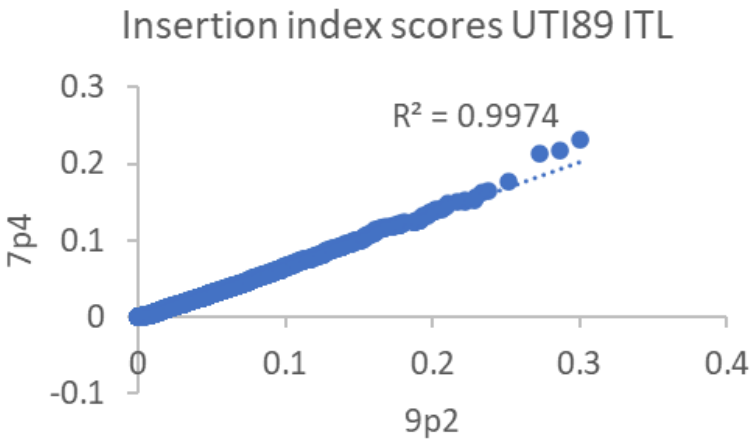
In each case, the genomic DNA was fragmented and labeled with inline index barcodes to enable identification of each sample. Illumina MiSeq was used for sequencing the libraries with PhiX control as a quality standard during sequencing, which helps to determine whether an error is related to library preparation. PhiX serves as a calibration control for examining the performance of the sequencing run. Also, PhiX can help in improving the run quality for low nucleotide diversity libraries, because this DNA (derived from bacteriophage PhiX174) has balanced fluorescent signals for all four bases. High nucleotide diversity or a well-balanced library is when the proportion of nucleotides are equally distributed among the inline barcodes in the run.

The samples were sequenced to obtain approximately 3 - 4 million reads. The three technical replicates of UTI89 ITL were compared using the gene insertion index scores. A high correlation coefficient of 0.99 was found between the replicates. Because of the time limitation in this project, three replicates were prepared in order to boost the number of reads mapped to the genome. In case of EO499 and MG1655 ITLs the correlation coefficients were 0.93 and 0.96, respectively, figure 45.

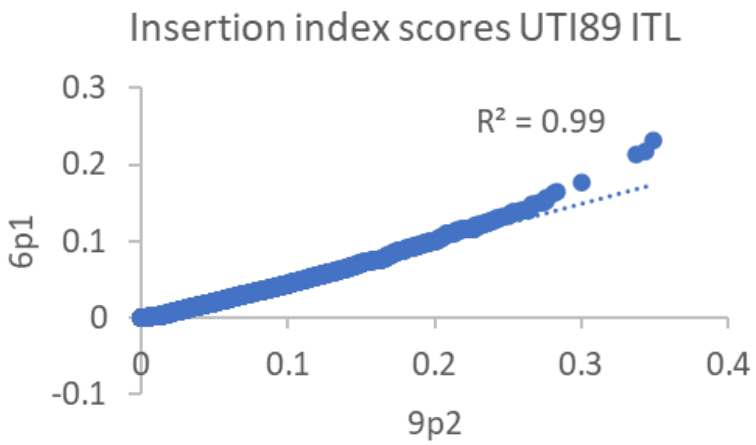
A


Insertion index scores UTI89 ITL
R² = 0.9967
6p1
7p4
0.4
0.3
0.2
0.1
0
-0.1
0
0.1
0.2
0.3
0.4

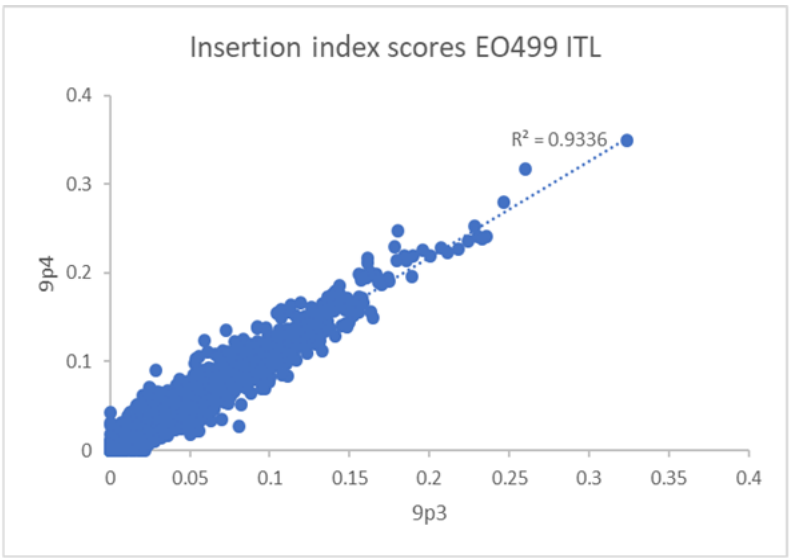
B


Insertion index scores UTI89 ITL
R² = 0.9974
7p4
9p2
0.3
0.2
0.1
0
-0.1
0
0.1
0.2
0.3
0.4

C


Insertion index scores UTI89 ITL
R² = 0.99
6p1
9p2
0.3
0.2
0.1
0
-0.1
0
0.1
0.2
0.3
0.4

D


Insertion index scores EO499 ITL
R² = 0.9336
9p4
9p3
0.4
0.3
0.2
0.1
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4

E

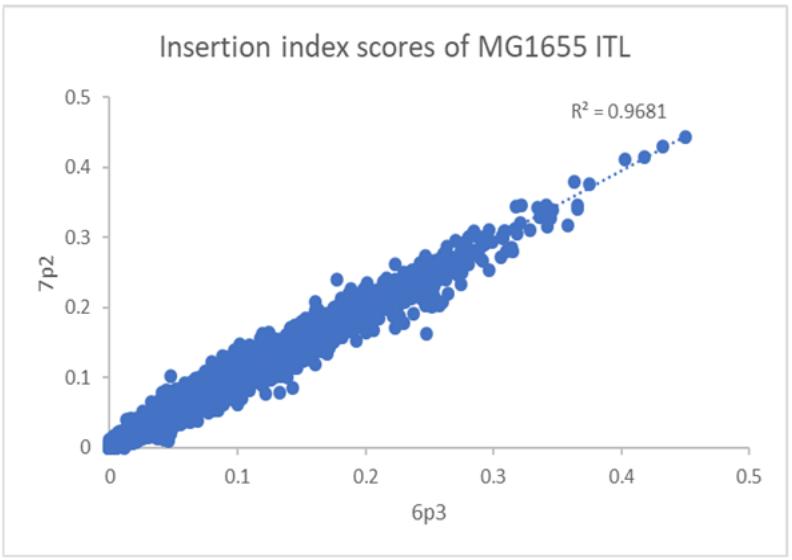

Insertion index scores of MG1655 ITL
R² = 0.9681
7p2
6p3
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5

Figure 45. The correlation coefficients of gene insertion index scores for the sequenced technical replicates in UTI89, EO499 and MG1655.

UTI89 ITL with three sequenced technical replicates. Each blue dot represent a gene. UTI89 with inline barcodes: A- 6p1 and 7p4. B- 7p4 and 9p2 C- 6p1 and 9p2. D- EO499 ITL with two sequenced technical replicates 9p3 and 9p4. E- MG165 ITL with two sequenced technical replicates 7p2 and 6p3.

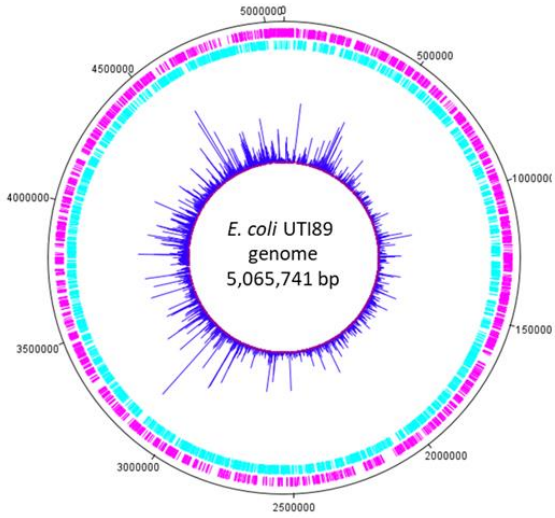
Therefore, the output sequencing files were combined and analyzed by TraDIS data analysis pipeline against the reference genome. The sequencing output files of EO499 and UTI89 were also aligned against the reference plasmid sequence separately (as both strains contain a megaplasmid in addition to the circular chromosome), as shown in table 16. The alignment of the sequencing output files were done twice; once against the genome, and secondly against the plasmid, as these two files were independent. This is in order to calculate the total read count and to locate the insertion count sites. Details on the TraDIS pipeline are given in the Methods and Materials section 2.6.12. The outputs of the TraDIS pipeline on UTI89, EO499 and MG1655 libraries are summarized in table 16.

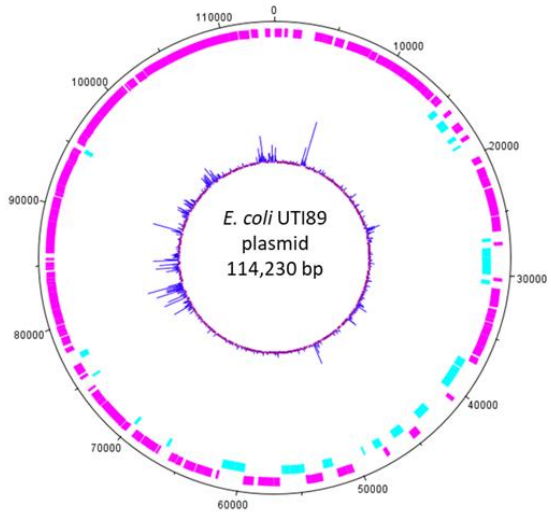
The MG1655 library shows the highest number of unique insertion sites, which were distributed along the genome with an average of one insertion every 8 bp. The UTI89 library shows a slightly less dense insertion with an average of one insertion every 15 bp for the genome and every 9 bp for the plasmid. This is probably, because fewer colonies were pooled in UTI89 than were pooled for the MG1655 library, when the MG1655 library was made. The initial transposon libraries were sequenced to a depth of 3 – 4 million reads. But in case of EO499, some sequencing files (.Fastq) had to be removed from the analysis because these files showed a sequencing noise background. Noise background in fastq will be explained later on (see Section 5.7.3). DNA plotter in Artemis was then used to construct the circular genome plot, see figure 46.

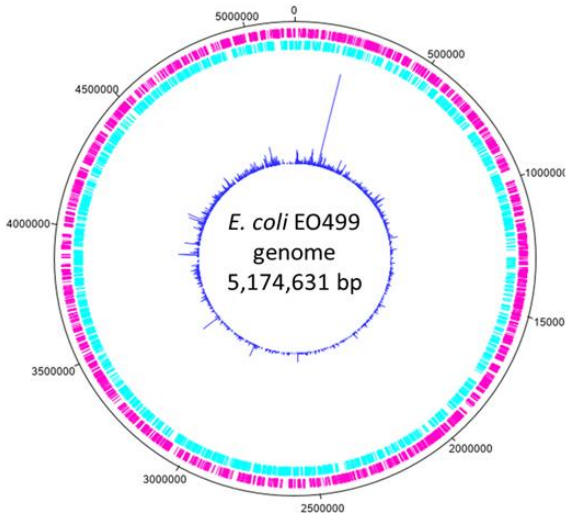
Table 16. Summary of sequencing libraries and mapping data.

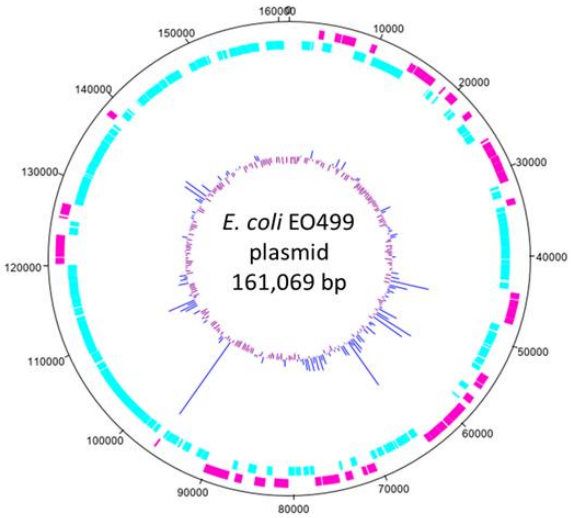
This shows the total number of reads mapped to each genome and mega-plasmid (when present), and the number of unique insertion sites determined from the TraDIS pipeline, for the three strains. The average distances between inserts were calculated as the genome length divided by the unique insertion sites.

Sequencing library	Genome or plasmid size in (bp)	No. of reads mapped to the genome	No. of unique insertion sites	Average distance between inserts (bp)
E. coli MG1655	4,641,652	6,012,402	574,734	8
E. coli UTI89 genome	5,065,741	5,918,004	240,029	15
E. coli UTI89 plasmid	114,230	299,961	9,439	9
E. coli EO499 genome	5,174,631	1,921,727	270,968	19
E. coli EO499 plasmid	161,069	372,906	27,860	5


E. coli UTI89
genome
5,065,741 bp


E. coli UTI89
plasmid
114,230 bp


E. coli EO499
genome
5,174,631 bp


E. coli EO499
plasmid
161,069 bp

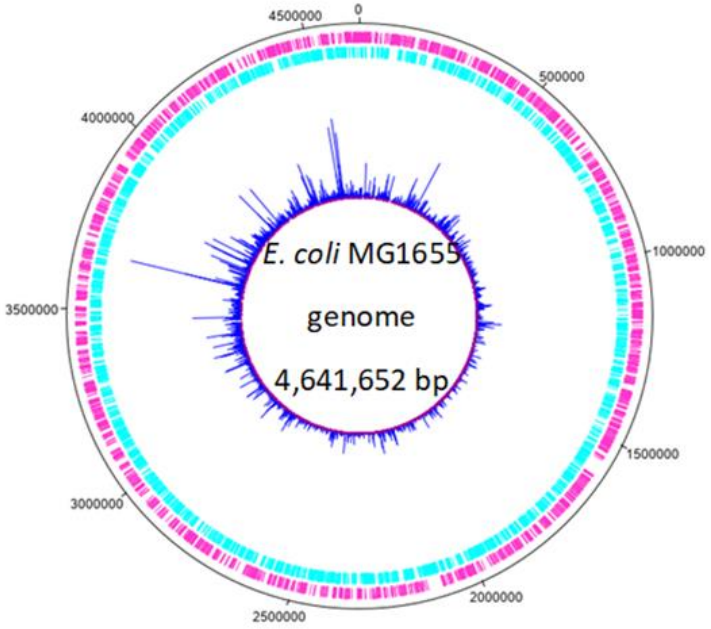


Figure 46. Genetic maps showing frequency and location of transposon insertions, mapped to the chromosome or the plasmid, assigned in the inner track.

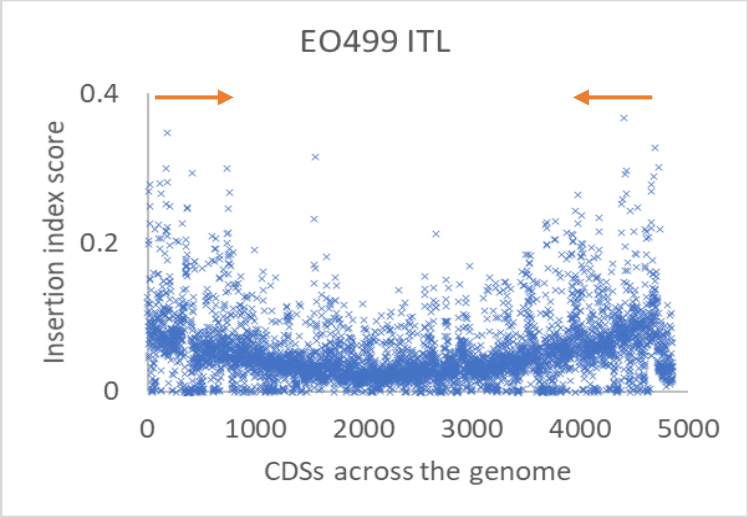
The outermost track marks the chromosome or the plasmid in base pairs starting at the annotation origin. The next two inner tracks correspond to forward and reverse CDS, respectively pink and turquoise. The innermost circle in blue corresponds to the position and the frequency of transposon insertion sequences. The figures were generated using DNAPlotter tool of Artemis.

5.5 The effect of the genomic position on the transposon insertion density

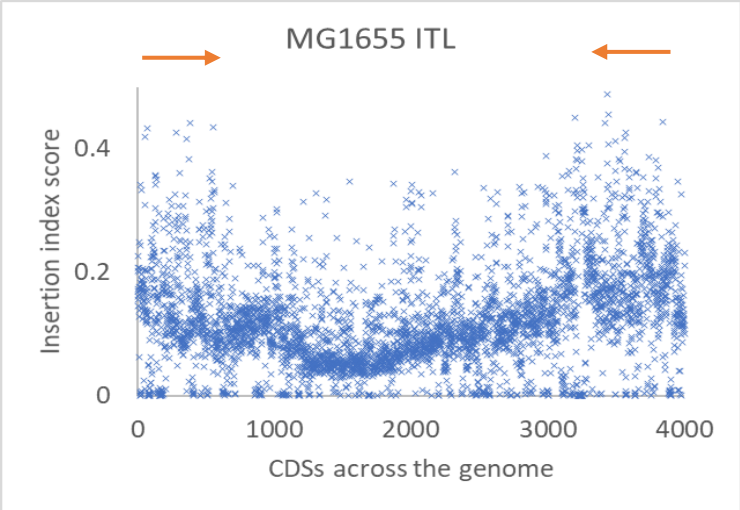
During the exponential phase, replication-associated gene dosage is seen (Cooper and Helmstetter, 1968). DNA replication in *E. coli* is bidirectionally, being initiated at a single origin, *oriC*, and proceeding in opposite directions toward the terminus region, (Wang et al., 2011). In

(Chao et al., 2016), transposon insertion density was shown to have a positional bias toward the origin of replication and a density decrease toward the terminus. This what is due to the relative abundance in the of origin-proximal chromosomal regions during cellular growth. This positional bias affects the comparison of the relative abundance of each transposon mutant in the library (Chao et al., 2013). Therefore, our data were plotted in figure 47 which shows the insertion index score verses the CDSs across the genome for the positional bias for the three ITLs. The three graphs show slightly a bias toward the origin of replication compared to the terminus site. EO499 and MG1655 ITLs showed a slight drop near the midpoint (terminus site) around CDS 2000 and CDS 1600, respectively. While it is noticeable in UTI89 showed from one side a slight bias toward the origin of replication and not very clear drop of the midpoint. This might be due to lower density library in compare to the others. The lower density of inserts around the terminus most likely to increase the of false negative essential genes. It is possible to correct the data for positional bias mathematically, but it was not considered (Chao et al. 2013).

A


EO499 ITL
Insertion index score
0.4
0.2
0
0
1000
2000
3000
4000
5000
CDSs across the genome

B


MG1655 ITL
Insertion index score
0.4
0.2
0
0
1000
2000
3000
4000
CDSs across the genome

C

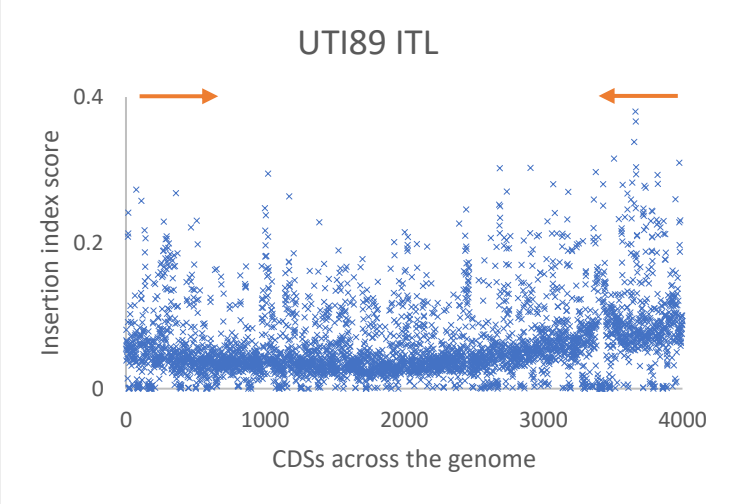

UTI89 ITL
Insertion index score
0.4
0.2
0
0
1000
2000
3000
4000
CDSs across the genome

Figure 47. The distribution of insertion index score across the CDSs.

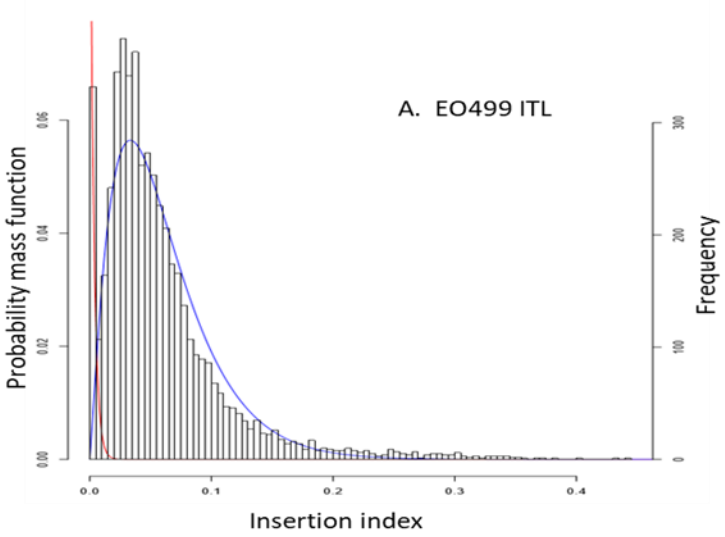
The Insertion index scores plotted in the annotation order of the genome for each library, A- EO499 ITL (1-4861) B- MG1655 ITL (1-4318) C- UTI89 ITL (1-4707). The insertion index position varies according to the relative position on genome. The orange arrows showed direction of chromosome replication.

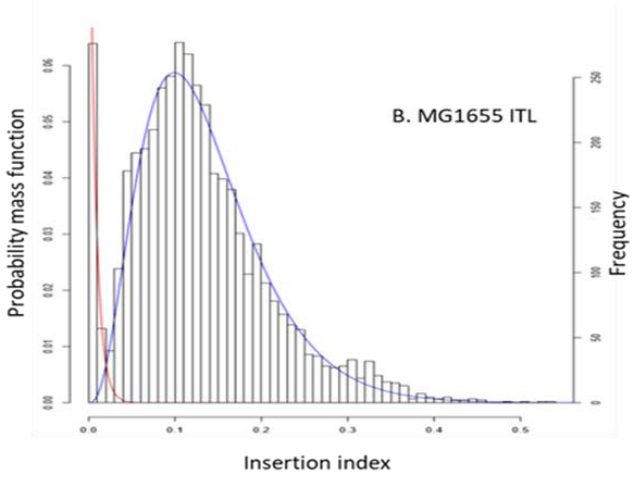
5.6 Identification of putative essential genes across transposon libraries

In this study *E. coli* genes were classified as essential, non-essential, or unclear, based on the number of insertions per CDS. The CDS is the gene defined as the region of DNA that will be transcribed and translated into protein, inclusive of the start and stop codons. The essential genes and partially essential genes were identified using the ESSENTIAL GENE PREDICTION pipeline based on insertion index value (Langridge et al., 2009; Goodall et al., 2018). This approach has been shown to work successfully with high density mutant libraries (Solaimanpour et al., 2015).

The frequency distribution of the insertion index score of the three initial transposon libraries (ITLs) was bimodal, figure 48, as previously shown by (Langridge et al., 2009) and others. The graphs show genes with low transposon insertion number in the left peak. These genes are classified as essential for cell viability. Genes with a higher number of insertions make up the right peak, and are classified as non-essential. Unclear genes are genes located in between the two peaks, where the software is unable to classify them as essential or non-essential. The number of genes in each category for these three ITLs, is shown in table 17, in the ESSENTIAL GENE PREDICTION outputs column. In MG1655, 333 genes were classified as essential because they had

a log-likelihood score less than a $\log_2(12)$. Conversely, 3764 genes were classified as non-essential because they had a log-likelihood score more than a $\log_2(12)$. And 221 genes were classified as unclear, because they had a log-likelihood score between these two cutoff values $\log_2(12)$ and $-\log_2(12)$, so it is not possible to assign the genes to essential or non-essential with an accurate degree of certainty.





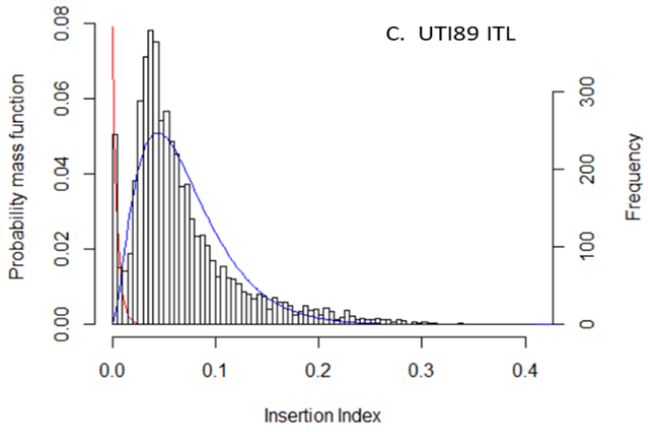


Figure 48. Distribution of insertion index scores for the three ITLs.

The normalized value insertion index for each coding sequence was calculated as the number of insertions per CDS/CDS in bp. The frequency of insertion index score was plotted for A. EO499 ITL, B. MG1655 ITL, C- UTI89 ITL. All the three display a bimodal distribution. The distributions of the data were fitted to two models: an exponential distribution (red) fitted to the left mode which covers the essential genes and a gamma distribution mode (blue) being right-skewed which includes non-essential genes. The log-likelihood-ratio test for insertion index score was calculated for the two modes. If the ratio was less than $\log_2$ (12), the gene was classified as essential, therefore 12 times more likely to be in the red mode than blue mode for non-essential.

The ESSENTIAL GENE PREDICTION pipeline classified the genes based on the insertion index score to essential, non-essential and unclear genes. The outputs of pipeline were compared using Excel. The results in table 17, show overall comparisons between the three strains. The table shows a combination of Roary output data (described in section 5.3), and ESSENTIAL GENE PREDICTION pipeline output data.

Table 17. Summary of results obtained by ESSENTIAL GENE PREDICTION pipeline in relationship to genome annotation (Roary) across the three ITLs.

The Roary output row shows the number of genes based on gene presence or absence among the strains. The essential pipeline column shows the number of genes in each ITL classified based on the insertion index score. The Essential pipeline grouped the genes into three categories: essential, non-essential and unclear.

	Tn libraries	ESSENTIAL GENE PREDICTION output (Raw Data)	Genes present in all the three strains genome (core genome)	Genes present in MG1655 and EO499 genome	Genes present in MG1655 and UTI89 genome	Genes present in EO499 and UTI89 genome	Genes unique to MG1655 genome	Genes unique to EO499 genome	Genes unique to UTI89 genome
Roary output (Raw data)			3419	116	78	555	705	945	788
Essential	MG1655	333	312	0	1		20		
	EO499	355	327	3		1		24	
	UTI89	295	286		1	0			8
Non-essential	MG1655	3764	2937	113	76		638		
	EO499	4310	2798	110		537		862	
	UTI89	4139	2789		75	531			739
Un-clear	MG1655	221	170	3	1		47		
	EO499	373	294	3		17		59	
	UTI89	411	344		2	23			41

Gene essentiality of EO499, UTI89 and MG1655 were classified based on genes present or not or unclear. Based on Roary output genes were classified to core, accessory genes and the remaining were unique genes as it is shown in section 5.3. To focus on the essential genes, in MG1655 312 genes (93.7 %), in EO499 327 genes (92.1 %) and in UTI89 286 genes (97.0 %) were shown to be core essential genes. As expected, the majority of the essential genes were found in the core genome. The remaining of essential genes were accessory or unique: that is, they were not found in all three strains. As described earlier in this chapter, accessory essential genes could be potential drug targets especially for UPEC strains, UTI89 and EO499. The list of accessory genes common in two strains were found in supplementary data table S9, S10 and S11. While the list of genes found to be unique to each strain were listed in the supplementary data table S12, S13 and S14. In the unique essential gene lists for the three strain MG1655, UTI89 and EO499, a large number of these genes were identified as hypothetical proteins, putative protein and prophage. Further studies were required to confirm these results, but no further validation was conducted.

In order to validate the unique essential gene lists obtained by TraDIS in S12, S13 and S14, these lists were compered to other data. A comparison between three data sets were conducted by (Goodall et al., 2018) using TraDIS (BW25113), Keio collection (BW25113) and PEC (W3110) to determine the overlapped essential genes. The fact that the unique essential genes found in UTI89 were mostly hypothetical proteins and putative transcriptional regulator, they were excluded from the analysis. MG1655 and UTI89 were compared to Goodall, 2018 data set, the analysis found in table S15. Although the analysis showed these genes were unique essential in MG1655 or EO499 in this study, some of these genes were presented in BW25113 and W3110.

Mostly these unique essential genes overlapped with the essential genes found in BW25113 TraDIS by Goodall. The result of this analysis in MG1655 confirm the reproducibility of this data in TraDIS BW25113. While half of the unique essential genes in EO499 were found to overlapped with Keio collection or TraDIS BW25113. This analysis confirms the reproducibly of the essential genes in this study to other data. It would be interesting to manually inspect the genes involved in this analysis by Artemis. This is because in (Goodall et al., 2018) it was found that essential genes which contain non-essential regions were classified as non-essential by the essential gene prediction based in the insertion index score. No further analysis was done on these genes here.

As expected, a higher number of genes are non-essential in comparison to the essential genes. However, there were about 531- 537 non-essential genes present in both EO499 and UTI89. This was expected and discussed in section 5.8. UPEC strains are expected to share a higher proportion of genes since they belong to the same phylogroup B2. The small discrepancies were due to Roary identifying genes across the strains, identifying these discrepancies required a manual inspection. Also, these results showed that the major differences among strains were due to unique non-essential genes. Among these, 14.7% genes were unique to MG1655, 17% genes were unique to EO499 and 15.2% were unique to UTI89.

5.7 Unexpected results and troubleshooting

In this coming section, I will investigate the unpredicted results and some troubleshooting I had during the analysis. This is such as the number of unclear genes in EO499 and UTI89 were

high, in table 17. Additionally, illustrate a gene identified by Prokka and not Roary in UTI89 ITL and the applied solution. Lastly, sequencing noise background in UTI89 and how this was identified.

5.7.1 Unclear Genes

Unclear genes (those which could not be classified as essential or non-essential) were more frequent in the EO499 and UTI89 libraries. This could be due to these being lower density transposon libraries in comparison to MG1655. If this is correct, I would predict that unclear genes would be shorter in length than genes overall. To test this, I have calculated the average genes size in the three strains and the average size of the unclear genes. This is shown in figure 49. This showed significant difference among the size of unclear genes in MG1655, UTI89 and EO499 in compare to the average gene length of the genome. All the statistic with t-test, p-values were < 0.05 were considered statistically significant. These results support the hypothesis that most of the unclear genes in MG1655, UTI89 and EO499 have a shorter gene length. As smaller genes are statistically more likely to have few insertions by chance, they are therefore more likely to be classified as unclear genes.

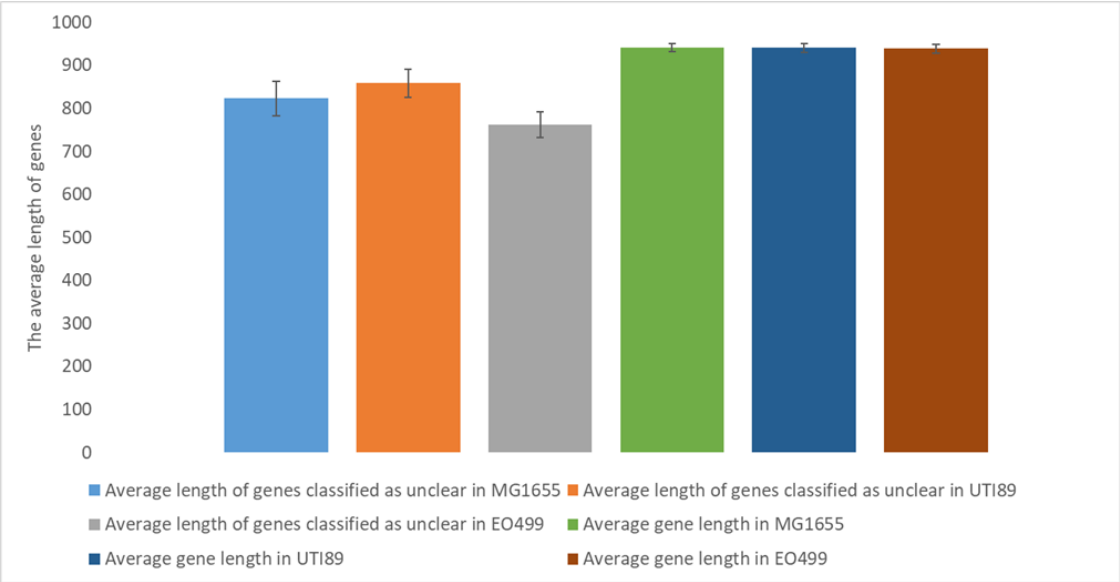


Figure 49. Average length of genes in MG1655, EO499 and UTI89.

The columns showing the average length of genes for MG1655, EO499 and UTI89 and the average gene length of the unclear genes. The error bars are the standard error of the mean.

5.7.2 Correction of UTI89 annotation

As I was examining different gene list between the three sequencing libraries, not convincing gene list was spotted by Dr. Mathew Milner. Which made us to cross-check the gene presence/absence matrix from Roary, this was checked manually. A gene (cds2169) was found in the UTI89 genome annotated by Prokka classified as hypothetical protein with a length of 90 bp. It was not identified by Roary gene presence/absence matrix in compare to other two database MG1655 and EO499, figure 50. It was classified as non-essential in the ESSENTIAL pipeline. So, to obtain a uniform genome annotation lists among the three strains, cds1269 was deleted from UTI89 list. The rest of the analyses were carried out with the correct list.

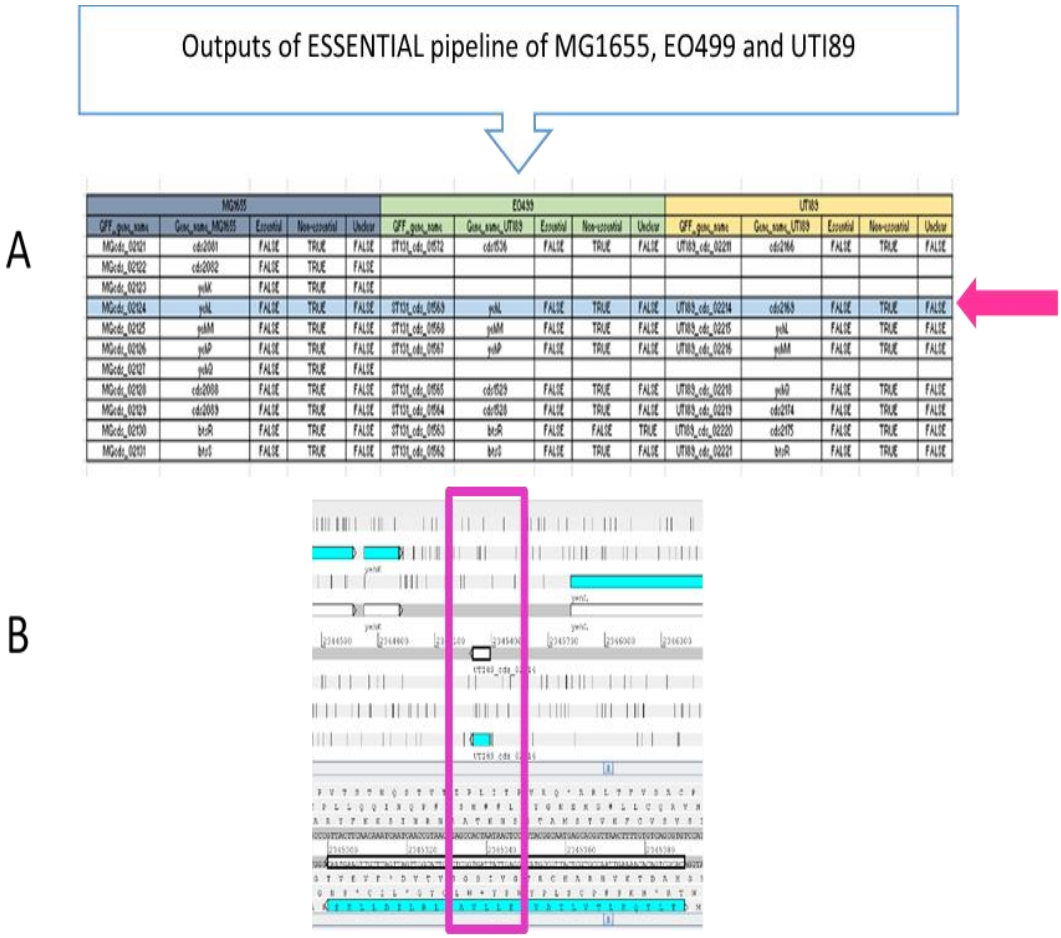


Figure 50. Correcting UTI89 annotation.

A- Screen shot of the outputs of ESSENTIAL pipeline for MG1655, EO499 and UTI89, respectively as a color-coded header. The three lists were ranked based on the (GFF_gene_name) General Feature Format column of UTI89. The arrow points to the blue highlighted row, and shows an extra gene (cds2169) found in UTI89. B- cds2169 found in UTI89 genome in a size of 90 bp viewed by Artemis.

5.7.3 Noise in the UTI89 sequencing library

Another issue that arose was caused by a high noise background in the sequencing data analysis. The issue was first spotted during the analysis, when the output of ESSENTIAL GENE PREDICTION pipeline showed a much higher number of unclear genes in UTI89 ITL in comparison to MG1655 and EO499 ITLs, table 18.

Table 18. The ESSENTIAL GENE PREDICTION output in UTI89 sequencing library.

First attempt for sequencing UTI89 ITL.

ESSENTIAL GENE PREDICTION output in UTI89 ITL	
Essential	289
Non-essential	4046
Un-clear	510

To look more carefully at this issue, a few genes known to be essential were selected to be viewed by Artemis such *groS, groL, rpoD* and *rpoA,* figure 51. These genes showed a clear noise background of inserts in these genes. Transposon insertion in essential genes would result in a lethal phenotype, so essential genes should be free of insertions, as is explained in the introduction 1.10.1. This noise background expected to be from poor quality of some sequencing reads. To check this, in the sequencing analysis pipeline explained in materials and methods

2.6.12, the pipeline programmed to check the sequencing reads in two levels. At first 25 bp of the transposon sequence was check allowing 3 bp mismatch. Then, the matched transposon followed by checking for the last 10 bp allowing 1 base mismatch. I have assumed if I have ran the pipeline with UTI89 ITL allowing stringent mismatches for transposon check 1 and transposon check 2. As this will remove and filter parts of the reads containing low quality bases, to minimize any possible artifacts. I have processed the .fastq files few times in the pipeline testing different mismatches values, 2 and 0, 1 and 0, and no mismatches allowed. The output results were viewed by Artemis and no significant changes were observed in these genes. So this hypothesis could not be confirmed.

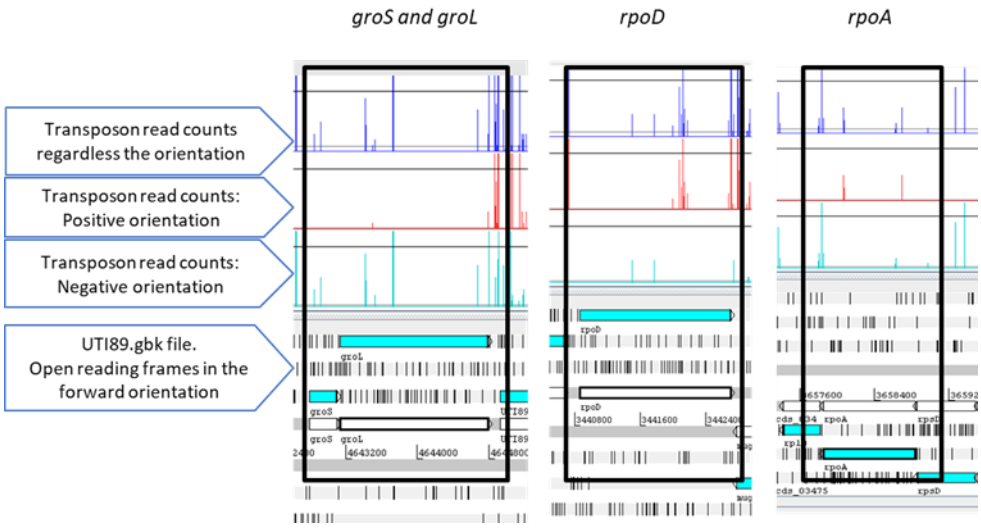


Figure 51. Transposon read count in UTI89 ITL viewed in Artemis.

Essential genes showing a noise background processed through stringent TraDIS pipeline. The blue vertical lines are the total transposon read counts, the red vertical lines are read counts in the positive orientation and the turquoise are read counts in negative orientation.

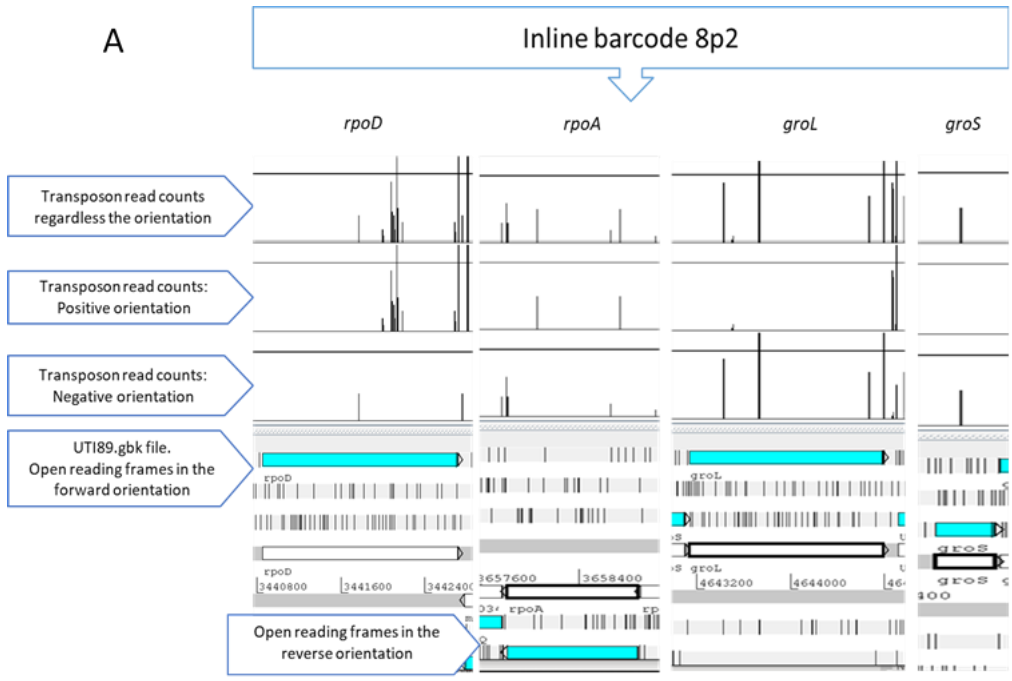
Secondly, I have tested the hypothesis that the background noise might be from the quality of overall of the .fastq files. As the quality of every sequence base is categorized by symbols and letters, arranged from lowest quality ! to highest quality ~, figure 52. I have expected to observe a quality values toward the low. To do this, the .fastq sequence files were manually inspected for the read quality for *groS, groL, rpoD* and *rpoA*. The quality of these reads were an average quality characters, the quality were shifted toward the right (High) in figure 52. I have expected to find a low quality of reads in which TraDIS pipeline might wrongly classified these reads as real reads. This test failed to locate the source of the background noise.

Figure 52. the quality character values for sequencing reads in .fastq

Finally, I have investigated the different inline barcodes that had been used to prepare the UTI89 ITL. Initially, the UTI89 library was prepared independently with two inline barcodes 8p2 and 9p2. Both libraries were processed in TraDIS pipeline separately with each inline barcode. Then, the read counts of both ITLs were compared by Artimes. The comparison was done on few random essential genes selected from MG1655 and EO499. The genes were *rpoD, rpoA, groL* and *groS.* The results showed the noise background was much higher in the library prepared using the inline barcode 8p2 than in the library prepared using 9p2, as shown in figure 53A. This showed that one of the library preparations was the source of the noise background found in the initial UTI89 ITL. To overcome this problem, I have removed the sequencing files of UTI89 library prepared with inline barcode 8p2. This reduced the number of reads, so two more independent libraries were prepared with different inline barcodes 6p1 and 7p4. Time limitation prevented the preparation of any additional samples. The final library prep was inspected using Artemis as shown in figure 53B. The overall noise background was reduced in the final library prep in compare to the first UTI89 sequencing library.

Then, I examine if the level of noise is acceptable in the final UTI89 library and it doesn't conflict with the further analysis. As essential genes should be free of insertion reads, it is

important to know the noise will still allow us to classify *gorL* and *groS* essential. Because both of these genes showed insertion reads by Artemis in figure 54. Successfully, these two genes were classified as essential genes. Since the level of background noise in the library did not interfere with the analysis, this was not further investigated.



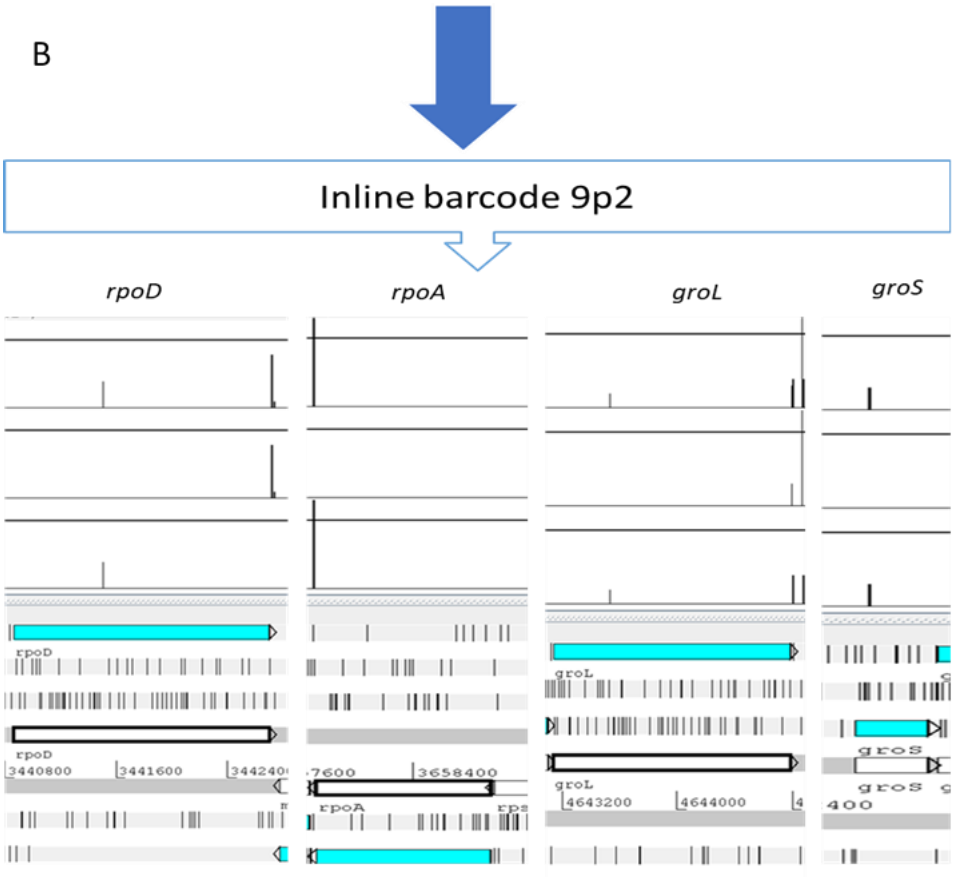


Figure 53. Comparison between two independent inline barcodes 8p2 and 9p2 of UTI89 ITL by Artemis.

The comparison was done on the essential genes *rpoD*, *rpoA*, *groL* and *groS*. UTI89 ITL prepared with A. 8p2 B. 9p2 inline barcodes.

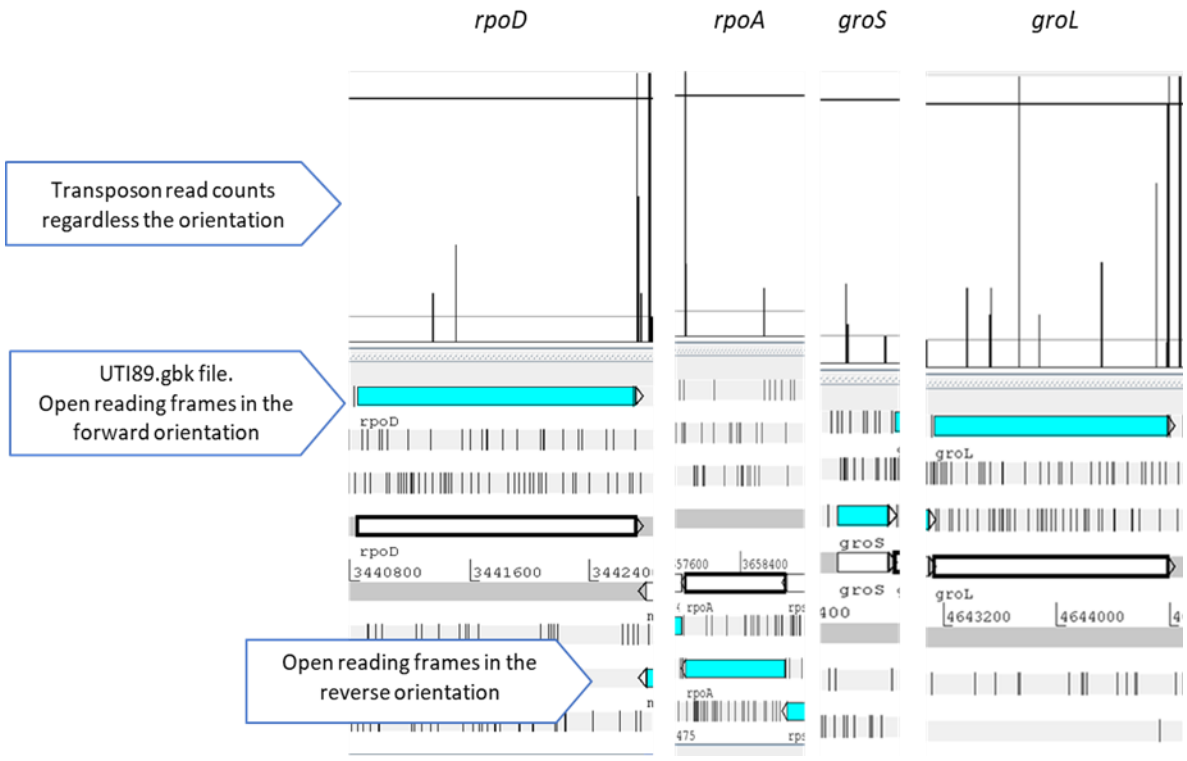


Figure 54. Manual inspection of the final UTI89 ITL sequencing by Artemis.

5.8 Discussion

Work described in this chapter involved the construction and analysis of the UTI89 library. Comparative analysis was used to identify the essential genes in MG1655, EO499 and UTI89 genomes using transposon sequencing method. These libraries will be used to perform TraIDS experiment under acetic acid stress condition. Identifying the essential genes will be used in the coming chapter to exclude them from TraDIS under acetic acid analysis in the next chapter.

Identifying essential genes is important for two main reasons: a- Highlight the biological process in bacteria, b- May offer new antimicrobial therapeutics for the treatment.

I have successfully constructed UTI89 transposon sequencing library in this study. The library construction method was modified for better transformation efficiency for UTI89. The method was modified for the growth phase of the bacterial cells when they were harvested for competent cells, transformation voltage and the antibiotic concentration in the recovery plates of the mutants.

The three sequencing libraries were analyzed by mapping the sequencing reads to the annotated reference genome. For genome annotation I have used two main pipelines Prokka and Roary. Prokka was applied first for rapid labelling and identify features of the genomic DNA sequence. Then, gff3 files produced by Prokka were used for Roary as an input file to obtain a matrix providing lists of genes and which strains it is present in.

Two further pipelines were used to identify essential genes in each genome and plasmid if present. First, TraDIS pipeline was used to obtain the number of reads mapped to the genome or the plasmid besides the number of the unique insertion sites. MG1655 sequencing library showed the highest unique insertion sites followed by EO499 then UTI89. Second, the ESSENTIAL GENE PREDICTION pipeline was used to classify the genes to essential, non-essential, or unclear based on insertion index value of the library. I have compared the ESSENTIAL GENE PREDICTION output among the three strains by grouping them to core genome, accessory genome and unique genome. Our analysis reveals most of the essential genes belong to the core gene set of the three

strains. I have expected to observe a large overlap of genes between UTI89 and EO499 as both belong to the same phylogenetic group B. This was successfully shown in study, a large proportion of the non-essential genes overlap between in EO499 and UTI89 in compare to MG1655.

However, a high number of un-clear genes in EO499 and UTI89 were obtained in this study. The average length of these unclear genes shown to be smaller in comparison to the average length of all genes in the genome. Manual inspection of the unclear genes in EO499 and UTI89 and comparing them to MG1655 and other *E. coli* data bases such as BW25113 (Goodall et al., 2018) would be an option. This will allow us to sort the unclear genes manually to essential or non-essential for better data quality. This this was not investigated further.

To perform reliable analysis and high quality of the sequencing library it is important to manually inspect the data in Artemis to identify noise background in the sequencing. I had successfully identified the source the noise in one of our sequencing libraries in UTI89. A new library prep of UTI89 sample overcome this issue. It was difficult to track further the cause of the noise background. It was reported that error and noise background in the sequencing library could be for several reasons: sequencing run itself, technical errors during sample preparation, library preparation or PCR enrichment (Park et al., 2017).

6 Identification of genes required for growth under acetic acid by TraDIS

6.1 Overview

Due to the limited time remaining in this project, this chapter will present an analysis of TraDIS results obtained at pH 5.5 with and without acetic acid, without further validation. In the previous chapter I have successfully classified the genome of the three examined *E. coli* strains (MG1655, EO499 and UTI89) into essential, non-essential and unclear genes. To fulfil the aims as stated in the previous chapter in section 5.1, this chapter focus on identifying genes required for growth in the presence of acetic acid by TraDIS in these three *E. coli* strains. In here, I attempt to demonstrate the effect of acetic acid over longer period of time. Moreover, I compare Dr. Francesca Bushell TraDIS data to the newly generated TraDIS data, to examine the extent to which TraDIS data are reproducible when similar experiments were repeated by others. Additionally, Iattempt to compare the candidate genes identified from TraDIS in a lab strain and pathogenic strains under acetic acid stress, to see how much they are conserved between various *E. coli* strains.

To do this I have grown each of the transposon mutant libraries (MG1655, EO499 and UTI89) in M9 medium at pH 5.5 with and without 4 mM acetic acid, diluted repeatedly over a period of five days, as described in Methods and Materials section 2.6.3. Experiments were carried out for a period of five days but only populations from day 1 and day 5 were sequenced. The workflow of TraDIS library preparation is shown in figure 11. All experiments were performed in duplicate, and consisted of DNA extraction, DNA shearing and sequencing library preparation, as shown in Methods and Materials section 2.6.3 to 2.6.14. The inline barcodes used in library preparation are showed in supplementary data, table S3, S4 and S5.

By applying a time-course TraDIS approach, Iexpect to observe divergence over time under the examined stress conditions, with data showing genes where insertions make the cells more fit and less fit. Transposon sequencing or TraDIS studies are often done using a single sampling time point (Chao et al., 2013; Yasir et al., 2019; Boinett et al., 2019; Knöppel et al., 2018). As shown in one TraDIS study done using a time-course approach, reliance on single time point to identify significant genes under a selective condition may fail to capture some of the mutants required under the examined condition (Yang et al., 2017). In this study, I am trying to identify genes that contribute both positively and negatively under acetic acid stress.

6.2 Sequencing saturation depth script

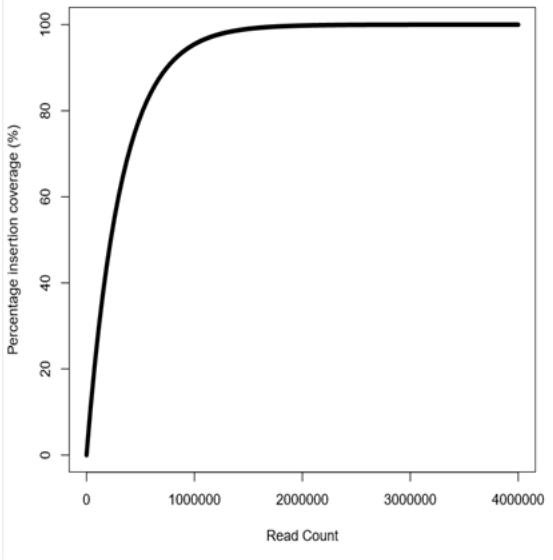
It is necessary to correctly determine the ideal total number of sequence reads for the outgrowth (pH 5.5 with and without acetic acid) sequencing samples; in other words, to determine how many reads are required to obtain saturation for the outgrowth samples. Some samples were required to be sequenced multiple times to approach saturation. As sequencing depth increases, it more likely to detect more genes, but the saturation depth of the outgrowth depends on the number of unique insertions of the initial transposon sequencing library.

To investigate this, a script was written in R by Dr. Mathew Milner based on the description in (Goodall, 2019). Using the unique insertion position of the sequencing allowed us to identify the approximate number of reads required for the outgrowth sampling. The equation used is:

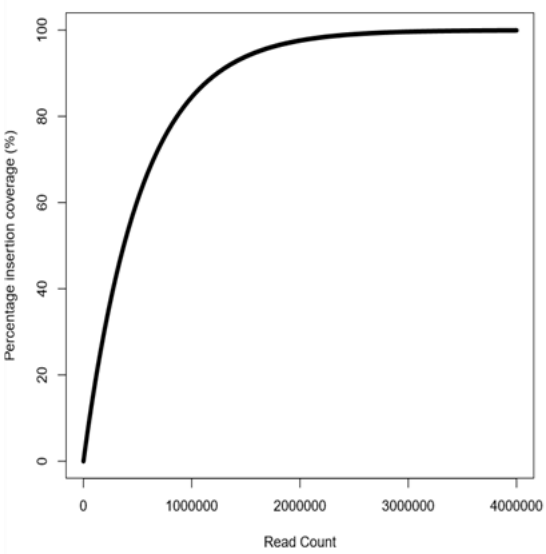
$$I = S - S\left(\frac{S-1}{S}\right)^{n}$$

where 'I' is the new number of insertions identified, 'S' the sample size (Unique insertion position of sequencing libraries), given in table 19, 'n' is the number of sampling iteration which equates to the sequencing reads. The number of sampling iteration were selected between 10,000 – 4 million iterations were used as a read count to 4 million. This means the new number of insertions 'I' was determined 400 times, in which n was calculated for every 10,000-point starting from 10,000 to 4 million. The results were plotted from 0 – 4 million read count against the percentage of insertion coverage of the total data, figure 55. The graphs were used to estimate the minimum number of reads count required for a given outgrowth TraDIS sample for each strain. These values are found in table 19. For example, in the case of UTI89 the graph shows that for a 99 % insertion coverage approximately ~ 1.49 million read count is required. Therefore, UTI89 TraDIS outgrowth samples were sequenced to ~ 1.49 million read counts before determining the fitness of each mutant.

A- UTI89



B- MG1655



C- EO499

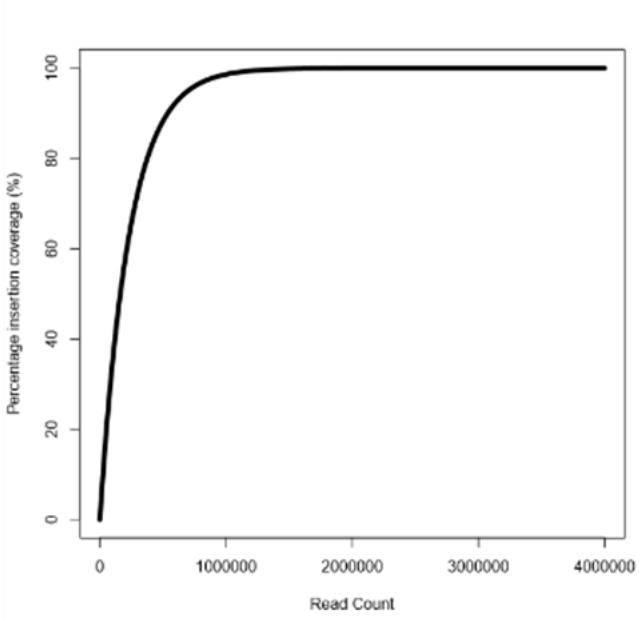


Figure 55. Estimation the number of reads required to have sufficient sample collection of TraDIS library.

The vertical axis shows the percentage insertion coverage and the horizantal shows the read counts. The graphs showed the initial transposon libraries (ITL): A- UTI89, B- MG1655 and C- EO499.

Table 19. Sample size needed to generate 99% saturation of outgrowth samples.

	Unique insertion position of sequencing libraries (S)	Percentage insertion coverage (%)	Read Count (M)
UTI89	323,214	99.00	~ 1.49
MG1655	574,734	99.00	~ 2.48
EO499	270,968	99.00	~ 1.25

6.3 Processing the outgrowth samples by sequencing analysis pipeline

As shown in the Methods section, processing the sequencing analysis pipeline required the annotated genome to map the corresponding sequencing reads to. The genome annotations were created previously in chapter section 5.3 using Prokka and Roary. The bioinformatic analysis was done as described in Materials and Methods in section 2.7.1 to section 2.7.3. In this section analysis started with processing the sequencing library (.fastq files) through the sequencing analysis pipeline. A summary of the output data from the sequencing libraries pipeline including the number of insertions on the genome (and the plasmid if present) and the number of reads

mapped to the genome for each strain (EO499, MG1655 and UTI89) is shown in the supplementary data table S3, S4 and S5.

6.3.1 EO499 outgrowth replicates

The RPKMs scores of EO499 at pH 5.5 with and without acetic acid on day 1 and day 5, resulted from processing TraDIS pipeline were plotted to see the degree of correlation of the two replicates of EO499 as shown in figure 56. The correlation between the replicates were high under all conditions, (R^2 = 0.9). The graphs were presented in log scale to better display the data, as it stretches the values in the graphs together and present the variability of low values more clearly. Therefore, EO499 data can also be used for further analysis using EdgeR, as shown in next section 6.4.

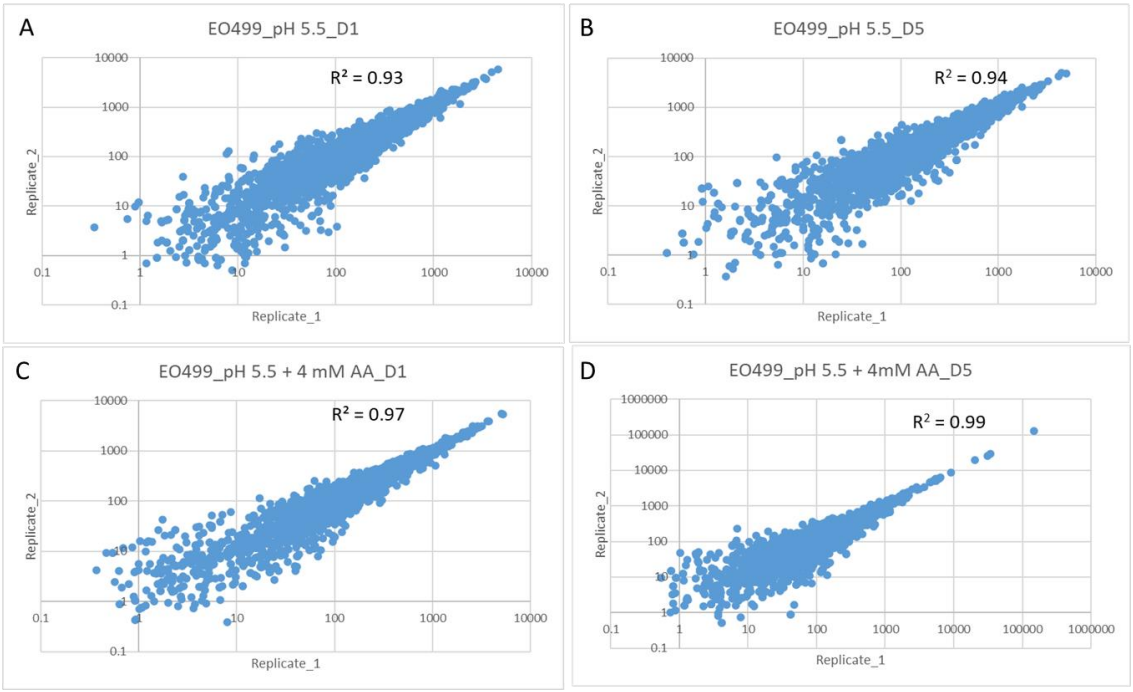


Figure 56. Correlation graphs of EO499 replicates based on RKPMs on log scale for all the examined conditions.

A. At pH 5.5 on day 1. B. At pH 5.5 on day 5. C. At pH 5.5 with acetic acid on day 1 and D. At pH 5.5 with acetic acid on day 5. AA: acetic acid. Note: The R-squared was calculated from the data without adjustments.

6.3.2 MG1655 outgrowth replicates

The replicates correlation of MG1655 based on RPKMs scores resulted from processing the acetic acid conditions by TraDIS are shown in figure 57. The correlation between the replicates was mostly good, but the replicates of MG1655 at pH 5.5 on day 5 showed lower correlation. Note that R^2 value was not presented in MG1655 at pH 5.5 on day 5, because of the presence of outliers which significantly distort the data (R^2 can only be calculated when the root mean

squared distances for the given regression line are normally distributed, which is not true in this case). The outlier point in the top right of figure 57B this refers to *rsmG* which is a methyltransferase responsible for methylation of 16S rRNA (**r**RNA **s**mall subunit **m**ethyltransferase **G**). I have tried to see if this outlier point is a result of numerous reads in *rsmG* in the ITL which get dominant through passaging or show a gene with numerous reads as a result of passaging the outgrowth samples. In MG1655 ITL, the gene with largest number of reads (RPKMs) found to be *hdeA,* is a periplasmic protein that is involve in acid resistance, with 11285 RPKMS. Ranking the RPKMS from highest to lowest, *rsmG* found in the 414 rank within the ITL, with 501 RPKMS. So the *rsmG* is a result from passaging MG1655 not from the ITL.

The most likely reason that the correlation between replicates at pH 5.5 on day 5 were poor is because the first replicate (R1) has lower sequencing read counts (722,830) in compare to the second replicate (R2) (1,925,379) read counts, as shown in the supplementary data table S4. Also, note the R1 has three technical repeats prepared for sequencing to boost the number of reads. The MG1655 data showed a less good data sets than EO499 data. The reasons of poor-quality sample in MG1655 at pH 5.5 on day 5, will be in the discussion section later on.

The large variability between these two replicates may lead to inaccurate estimate of the results when processing the data by EdgeR. However, I have decided to process the data from MG1655 by EdegR, and this is described in the next section 6.4.

It is important to know that EdgeR requires biological replicates from libraries with which to estimate the biological variability. Since I have only one good biological replicate of MG1655,

pH 5.5 on day 5, EdgeR can be used to process one repeat, but the obtained p-value and the number of significant genes will be very sensitive to error.

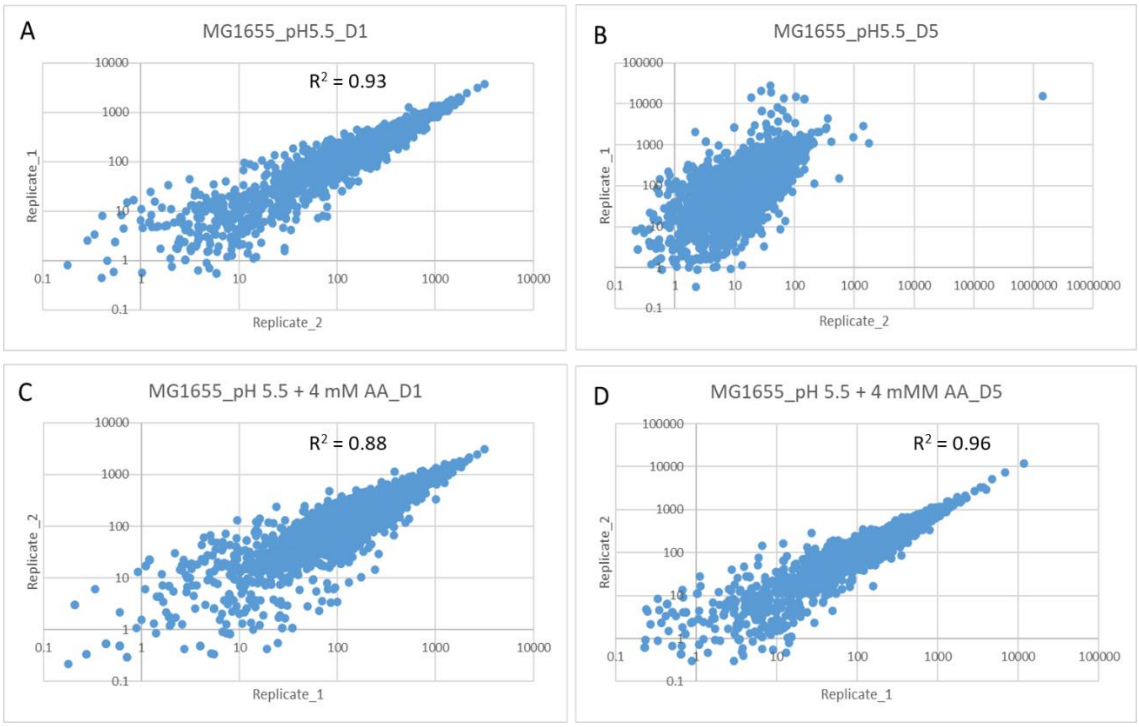


Figure 57. Correlation graphs of MG1655 replicates based on RKPMs on log scale for all the examined conditions.

A. pH 5.5 on day 1. B. pH 5.5 on day 5. C. At pH 5.5 with 4 mM acetic acid on day 1 and D. At pH 5.5 with 4 mM acetic acid on day 5. AA: acetic acid. Note: The R-squared was calculated from the data without adjustments.

6.3.3 UTI89 outgrowth replicates

Furthermore, in TraDIS of UTI89 the RPKMs for both replicates under all examined conditions (at pH 5.5 with and without acetic acid on day 1 and day 5) were plotted to determine the correlation between replicates as well. From the figure 58, the graphs showed in most of the cases correlation is extremely poor, at pH 5.5 on day 1 and day 5, and pH 5.5 with 4 mM acetic acid on day 5. In these cases, note also the R^2 value was not presented in the graphs for the same reason mentioned in the above section of MG1655 6.3.2.

As can be seen from UTI89 graphs in figure 58A, B and D, from the far-right top genes (one or two genes) with extremely high numbers of reads (RPKM scores) were seen in both populations. If these strains become dominant in the population, this would lead to the other mutants in the remaining genes becoming much lower in frequency, and this fact might explain the poor correlation between most genes.

To identify the genes that correspond to these points, the RPKM lists were ranked from larger to smaller value for both UTI89 ITL and TraDIS outgrowth conditions, see table 20 and table 21. I have investigated if the number of reads were high in the original UTI89 ITL before the selection, which could partially explain their dominance mutant after selection due to large initial input values.

As shown in the table 20 for UTI89 ITL, *mutL* has the highest RPKMs score value. In table 21, it can be seen that *mutL* and *recR* were the two genes responsible for most of the reads in these examined conditions: in other words, strains containing inserts in these genes had become very common in the populations. Only *mutL* showed a large skew UTI89 ITL with 44983 RPKMs,

which was reflected in the outgrowth conditions. While *recR* was in the 1873 in the ranking of ITL with 174 RPKMs. *recR* showed a skew as result of the outgrowth conditions. *mutL* encodes a protein involved in the repair of mismatches in DNA, and *recR* encodes a protein that is also involved in DNA repair. Altogether, these examples showed there some correlation between the genes with a skew in the ITL and the effect in the outgrowth samples, or one but not the other.

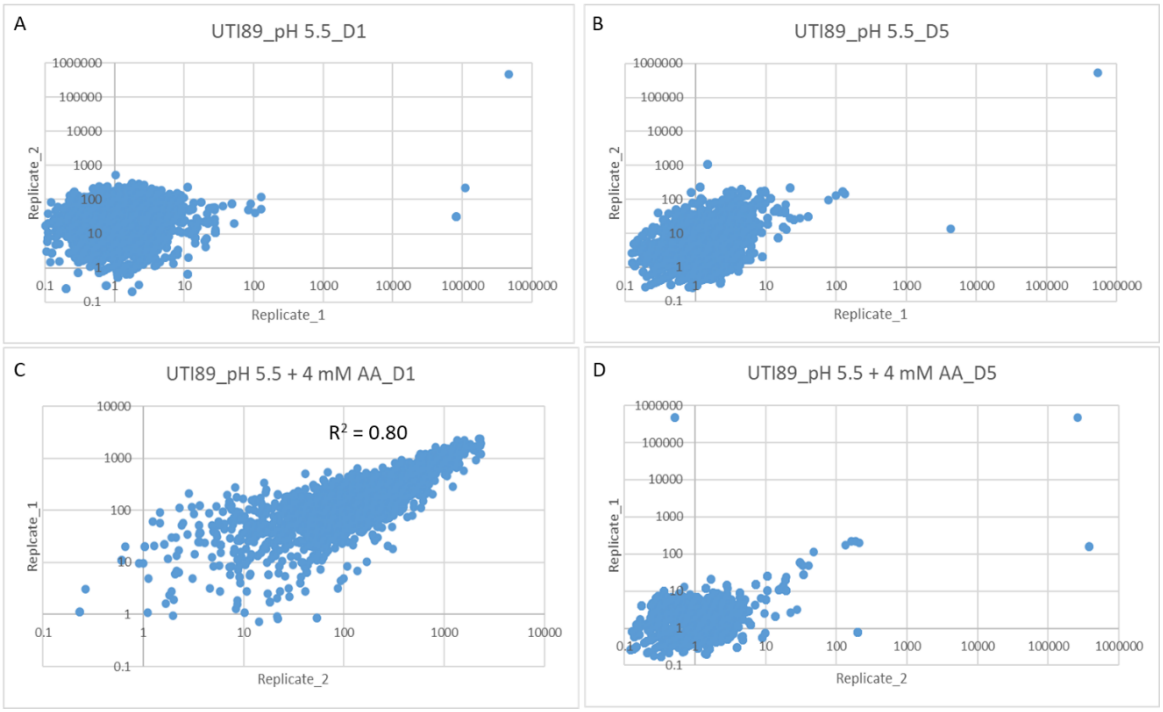


Figure 58. Correlation graphs of UTI89 replicates based on RKPMs on log scale for all the conditions.

A. At pH 5.5 on day 1. B. At pH 5.5 on day 5. C. At pH 5.5 with acetic acid on day 1 and D. At pH 5.5 with acetic acid on day 5. Note: The R-squared was calculated from the data without adjustments.

Artemis was used to visualize the frequency and the location of the sequencing reads within the genes *mutL* and *recR* for both UTI89 ITL and TraDIS outgrowth, as shown in figure 59, *mutL* and *recR* showed roughly equal distribution of inserts through the genes in both directions positive and negative. As seen in the figures the vertical lines of the read counts are off the scale, this is because of the overloaded read counts. Note that trying to zoom out the scale to see the overloaded reads would make the non-overloaded reads invisible (insertions with low read numbers). For this reason, I have chosen a scale (0-10) to view all the read counts including the small ones.

According to the table 21, I have expected to observe a high frequency of insertions by Artemis in *mutL* at pH 5.5 on day 1 and day 5, and at pH 5.5 with acetic acid on day 5. This is clearly shown in figure 60A and C. It is noticeable that the insertions are mostly toward the 3' end of the gene in these cases. While in *recR*, according to table 21, I expected to observe high frequency of insertions at pH 5.5 in replicate 1 on day 1 and pH 5.5 with acetic acid day 5 in both replicates. As shown in figures 60 B and D, a high frequency of insertions in *recR* were seen in the 5' end in these cases.

I hypothesized that the transposon insertion orientation in these genes (*mutL* and *recR*) might be biased toward one orientation and not the other. For example, it was found by Goodall et al that insertion mutants can exhibit polar effects on downstream genes, where expression of nearby genes is disrupted by the insertion (Goodall et al., 2018), because the mini-transposon used to construct the TraDIS libraries has an internal promoter that reads in one direction only. To see if this is the same case here, I have investigated the insertions of *mutL* and *recR* in depth

to look to whether there is a significant difference between negative orientation and positive orientation. For example, if these transposon insertions in the 3' end of *mutL* (where the promoter of *miaA* is located) face one orientation but not the other, this might affect the expression of the downstream gene *miaA*. *miaA* gene encodes enzyme tRNA isopentenyltransferase and mutations in *miaA* have been found to decrease RpoS expression. The *miaA* is affects expression of RpoS, as knockout of *miaA* affects translation of the *rpoS* open reading frame. This is because *miaA* facilitates the addition of 2-methylthio-N^6-(Δ^2-isopentenyl) to tRNAs for codon recognition (Thompson and Gottesman, 2014).

I show one example at pH 5.5 only, replicate 1 on day 1, to see if the insertion orientations were toward one particular direction or found in both directions, figure 61A and B. As shown, for *mutL* and *recR*, the insertions were found in both directions, i.e. not biased in one direction or the other. The same analysis was applied to the other examined conditions, and the same results were obtained (data in ITL not shown). So the insertions in *mutL* and *recR* are not likely to be having an effect due to expression from the internal promoter.

No further analysis was conducted on UTI89 TraDIS library because the number of sequencing reads were very high in a few genes, as shown in table 21, and this left the rest of the genes with a very low number of sequencing reads. This would therefore lead to errors when calculating the relative fitness for any given genes. This is further discussed in the Discussion section.

Table 20. The top three RPKMs scores for genes in UTI89 ITL.

The lists are arranged in descending order starting with larger RPKM scores. The light blue cells show common genes with very large RPKM scores.

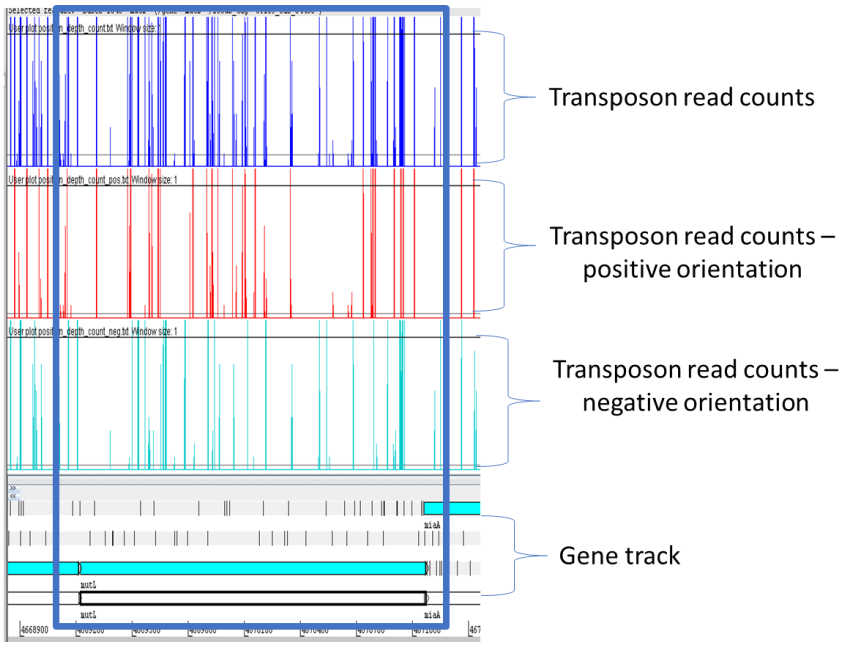
UTI89 ITL	
Gene	RPKMs
mutL	44983
hdeA	3576
cds3976	2698

Table 21. The top three RPKMs scores for genes in TraDIS UTI89, for both replicates (R1 and R2).

The lists are arranged in descending order starting with larger RPKM scores. AA: acetic acid, R: replicate. The light blue cells show common genes with very large RPKM scores.

UTI89 - pH 5.5 - R1 - D1		UTI89 - pH 5.5 + 4 mM AA – R1 - D1	
Gene	RPKMs	Gene	RPKMs
mutL	470853	cds4083	2326
recR	110530	cds1022	2317
ydiV_2	81450	*nlpA*	2303
UTI89 - pH 5.5 - R1 -D5		UTI89 - pH 5.5 + 4 mM AA – R1 - D5	
mutL	537754	*mutL*	377384
rsmG	4304	*recR*	262484
mlaD	135	*ydiV_2*	194926
UTI89 - pH 5.5- R2 -D1		UTI89 - pH 5.5 + 4 mM AA – R2 - D1	
mutL	471707	*hdeA*	2403
hdeA	527	*yhiD*	2367
rimP	369	*slp*	2226
UTI89 - pH 5.5 - R2 - D5		UTI89 - pH 5.5 + 4 mM AA – R2 - D5	
mutL	526891	*recR*	478808
cds833	1055	*flp*	468858
cds837	304	*mlaC*	221

A


Transposon read counts
Transposon read counts – positive orientation
Transposon read counts – negative orientation
Gene track
mutL

B

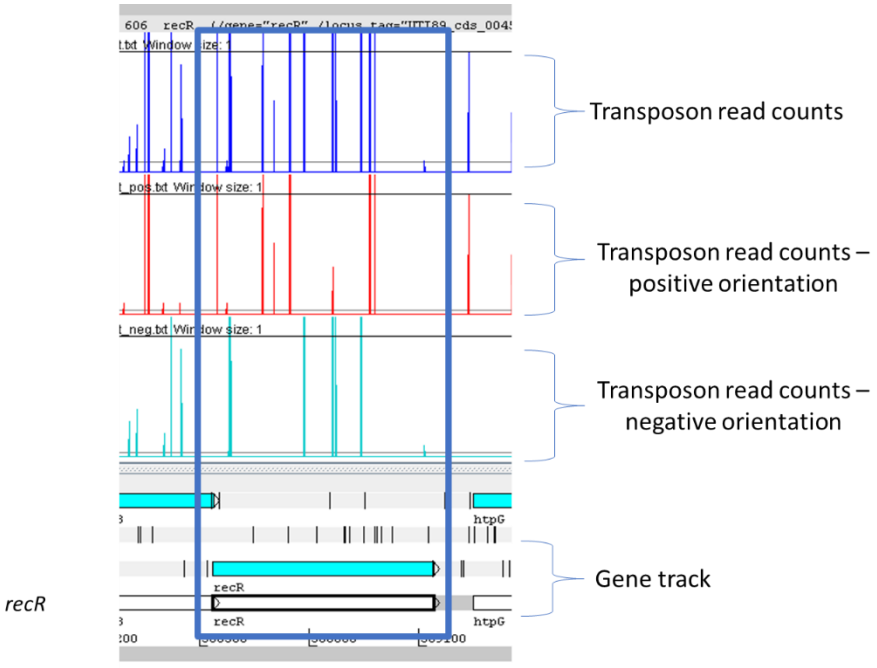
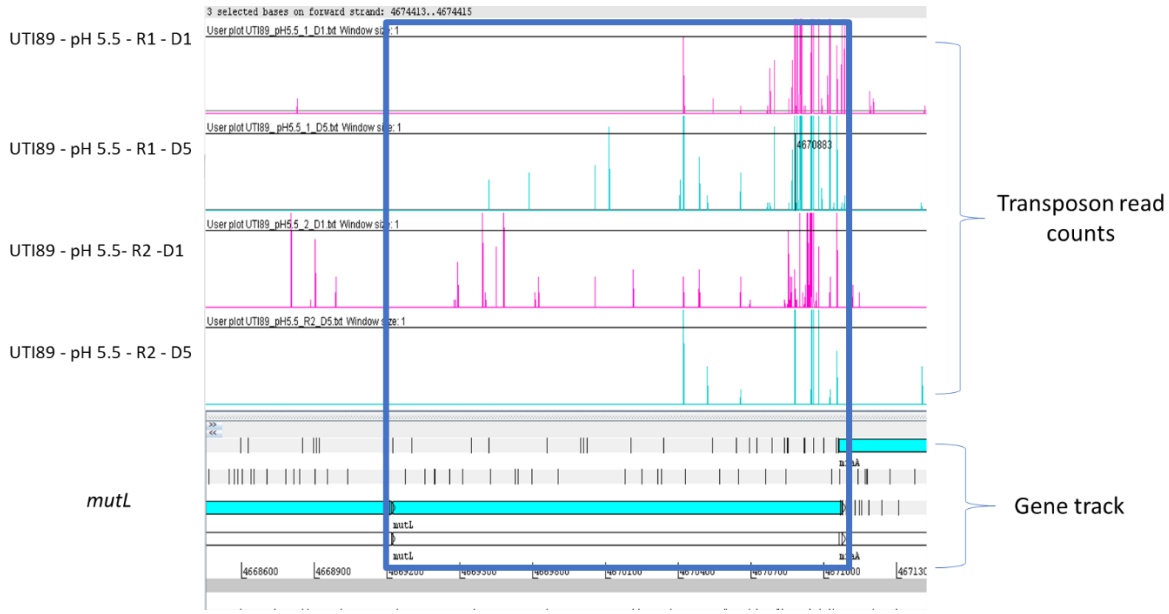

Transposon read counts
Transposon read counts – positive orientation
Transposon read counts – negative orientation
Gene track
recR

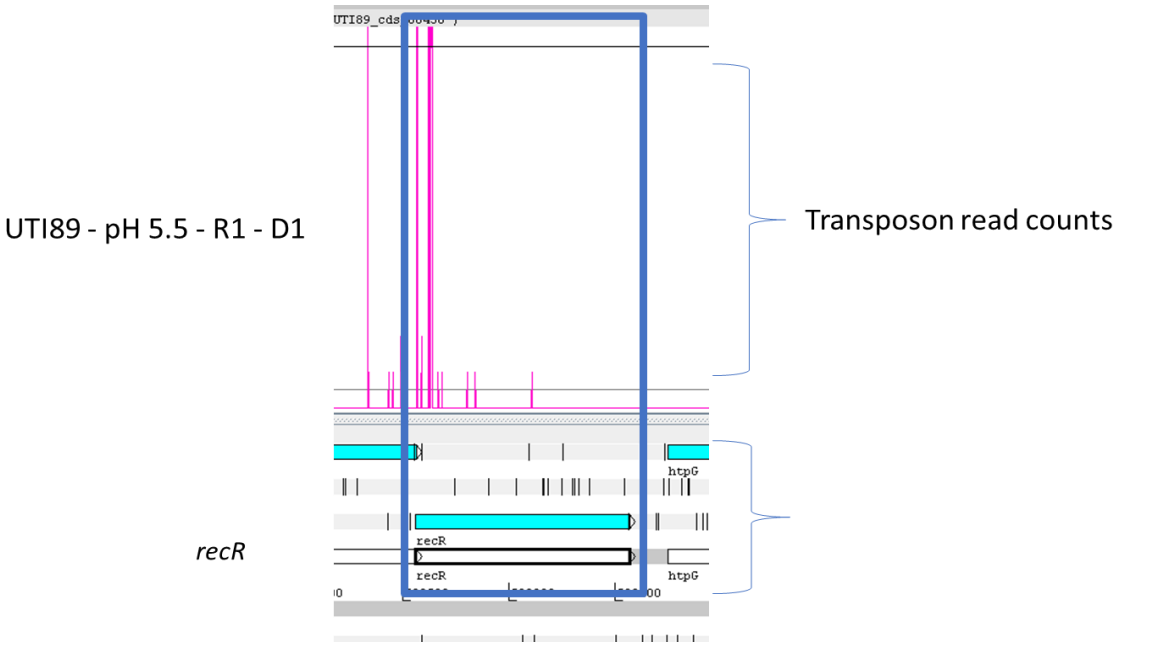
Figure 59. ITL data for *mutL* and *recR* in UTI89 with the insertion site orientation viewed by Artemis.

The insertion sites and frequency are represented by the vertical lines and mapped reads to the genome (gene track). A. *mutL*. B. *recR*. The blue vertical lines are the total transposon read counts in both orientations, the red vertical lines are read counts in the positive orientation and the turquoise are read counts in negative orientation.

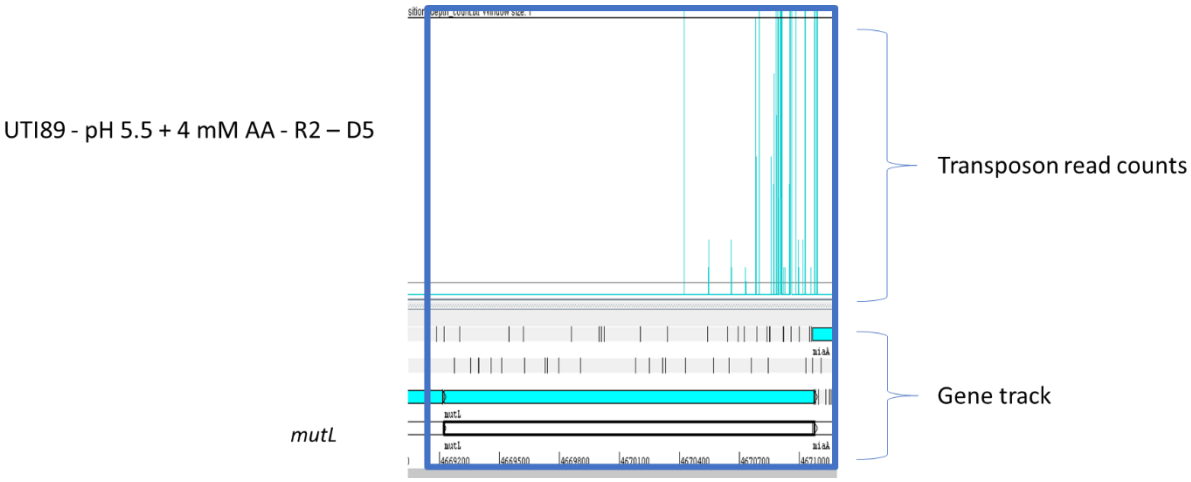
A



B


UTI89 - pH 5.5 - R1 - D1
Transposon read counts
recR
htpG
recR
recR
htpG

C


UTI89 - pH 5.5 + 4 mM AA - R2 – D5
Transposon read counts
Gene track
mutL
mutL
mutL
4668200
4668500
4668800
4669100
4669400
4669700
4670000

D

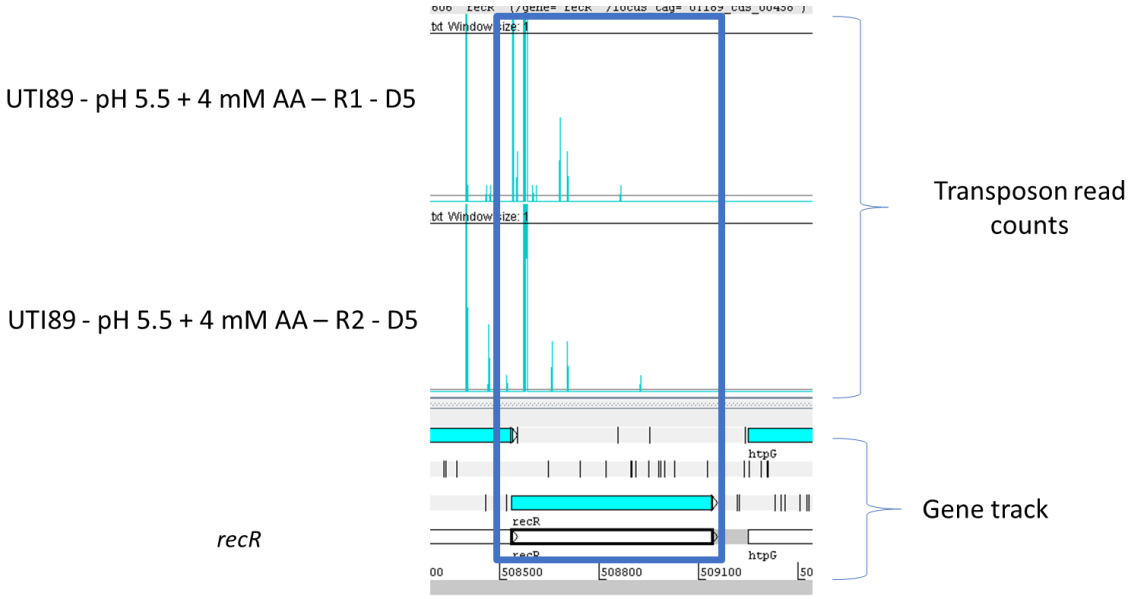
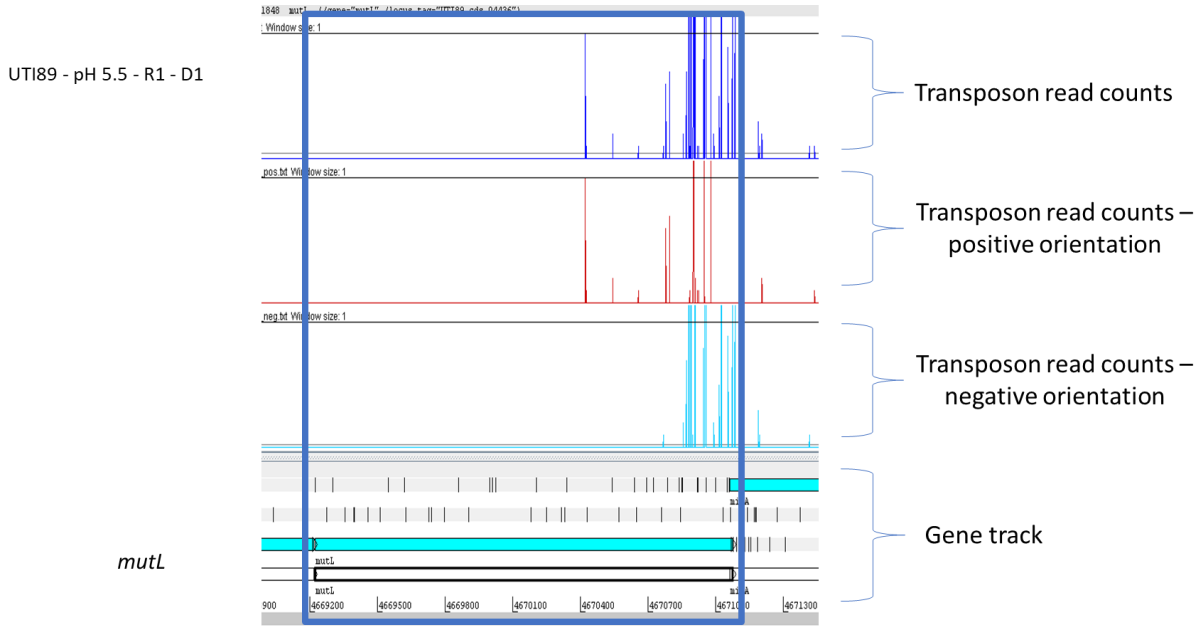


Figure 60. TraDIS data for *mutL* and *recR* in UTI89 viewed by Artemis.

The insertion sites and frequency are represented by the vertical lines and mapped reads to the genome (gene track). Pink color vertical lines represent the condition at day 1 (D1), and the light green vertical lines represent the condition at day 5 (D5). A. *mutL* at pH 5.5, both replicates (R1 and R2), on day 1 and 5 (D1 and D5). B. *recR* at pH 5.5, replicate 1 on day 1. C. *mutL* at pH 5.5 with acetic acid replicate 2 on day 5. D. *recR* at pH 5.5 with acetic acid, both replicates (R1 and R2), on day 5.

A


UTI89 - pH 5.5 - R1 - D1
Transposon read counts
Transposon read counts – positive orientation
Transposon read counts – negative orientation
Gene track
mutL
4669200
4669500
4669800
4670100
4670400
4670700

B

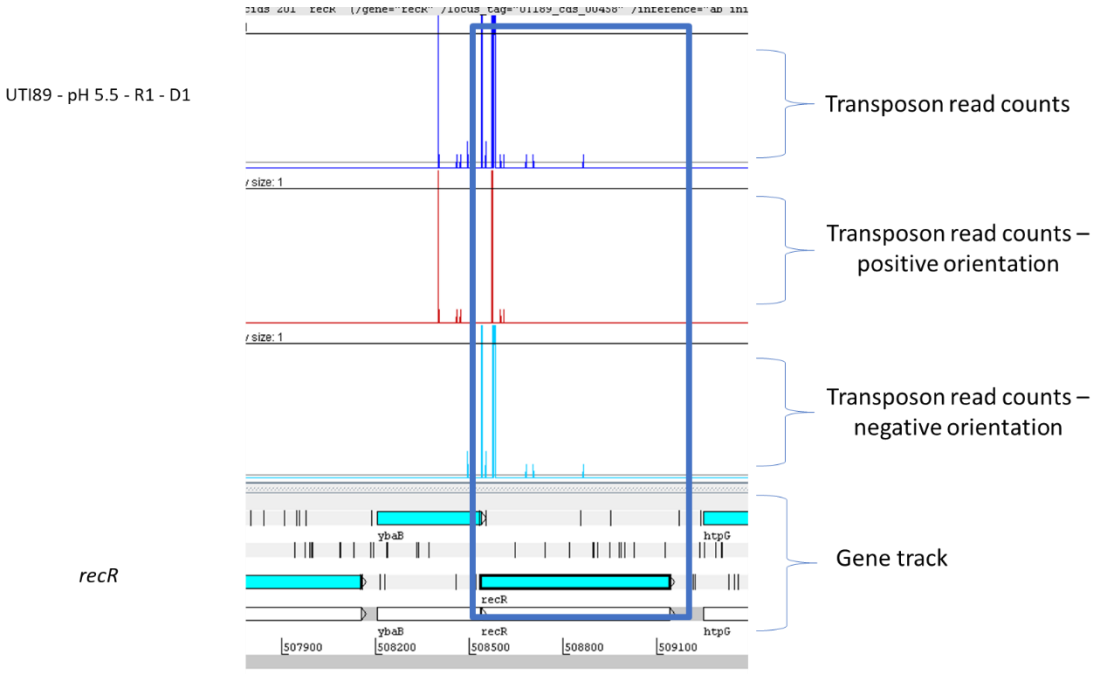

UTI89 - pH 5.5 - R1 - D1
Transposon read counts
Transposon read counts – positive orientation
Transposon read counts – negative orientation
Gene track
recR
ybaB
recR
htpG
507900
508200
508500
508800
509100

Figure 61. TraDIS data for *mutL* and *recR* in UTI89 with the insertion site orientation.
The insertion sites and frequency are represented by the vertical lines and mapped reads to the genome (gene track). Transposon read counts in UTI89 viewed in Artemis. A. *mutL* at pH 5.5, replicate 1 (R1), on day 1 (D1). B. *recR* at pH 5.5, replicate 1 on day 1. The blue vertical lines are the total transposon read counts in both orientation, the red vertical lines are read counts in the positive orientation and the turquoise are read counts in negative orientation.

With these findings in this section, I have decided to process further with the analysis for EO499 TraDIS outgrowth as the data looks very reproducible. The MG1655 outgrowth TraDIS data were less reproducible, but the analysis will be processed with the same obtained results. The data of both EO499 and MG1655 was therefore processed next by EdgeR to identify significantly changing genes.

6.4 Processing the outgrowth data by EdgeR

For reasons explained in the previous sections 6.3.3, UTI89 data was not included in this analysis. Only EO499 and MG1655 were analysed in this section. The EO499 and MG1655 data were processed by EdgeR to determine the effects on bacterial fitness of insertions in each non-essential gene, as explained in Materials and Methods section 2.7.7. The read count obtained from sequencing corresponds to number of insertion per gene in each sample. To generate the

gene list, the read counts for each gene obtained from the sequencing analysis pipeline were normalized for gene length:

$$Normilized\ Read\ Counts\ for\ gene\ X = \frac{Read\ counts}{Gene\ length}$$

The relative fitness for any X gene was calculated, using the following formula:

$$x = log_2\left(\frac{pH\ 5.5\ + AA}{pH5.5}\right)$$

The formula is used to determine the effect of transposon inserts in each non-essential gene on the relative fitness of these strains (EO499 and MG1655) under the two conditions (pH 5.5 + acetic acid and pH 5.5) on day 1 and day 5. Precisely, the formula looks for the effect of acetic acid only at pH 5.5. The reason for comparing (pH 5.5 + acetic acid / pH 5.5) but not (pH 5.5 + acetic acid / ITL), is to remove the effect caused by pH 5.5 alone.

EdgeR calculates the relative fitness for all genes including the essential genes. Some of the essential genes showed a relative fitness of zero due to absence of inserts while some other essential genes showed a relative fitness result due to low background noise or the presence of a few rare inserts mapped to essential genes. For this reason, the essential genes obtained in the previous chapter for EO499 and MG1655, section 5.6, were removed from the relative fitness gene list obtained from EdgeR.

To visualize the data from EdgeR, the log_2 fold of relative fitness of all the non-essential genes under acetic acid stress were plotted against their FDR score. The data were ranked by the

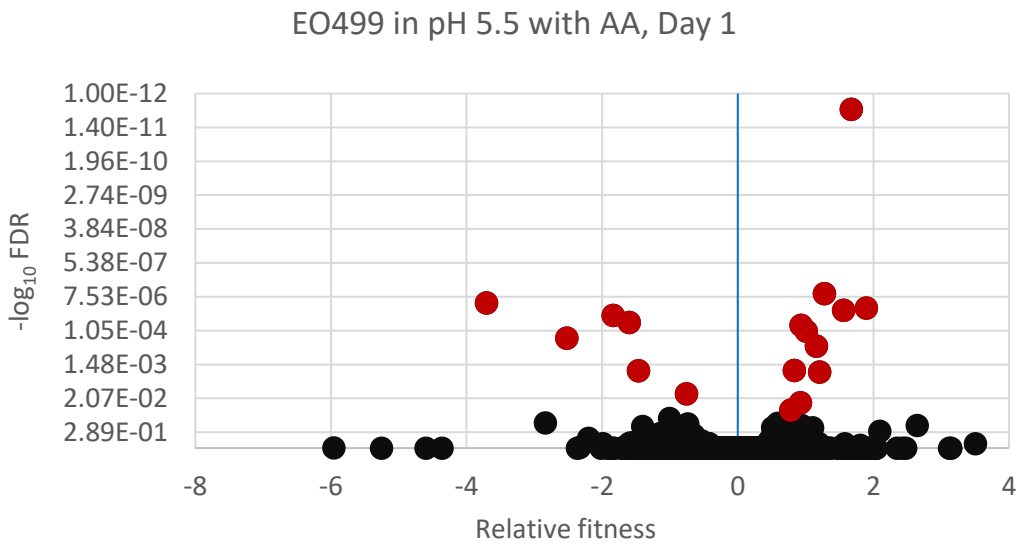
significance score (FDR) with a cutoff value of < 0.05 in both EO499 and MG1655 on day 1 and day 5, as shown in figure 62. The graphs show a scatter plot of $\log_{10}$ transformed FDR scores from EdgeR on the y-axis against the relative fitness of the $\log_2$ fold change between the two conditions (pH 5.5 with and without acetic acids) on the x-axis. The red color-coded mutants have a significant difference in fitness in the presence of acetic acid (The presence of acetic acid/ the absence of acetic acid). The graphs 62 A and 62 B showed volcano patterns for EO499 on both days 1 and 5. The data showed black color-coded mutants which refer to non-significant mutants under acetic acid stress. The red color-coded mutants refer to mutants with significant changes in fitness under acetic acid stress with a $\log_2$ fold change of relative fitness and FDR < 0.05. On the right side of the graph are cases of genes where strains containing Tn inserts are fitter in presence of acetic acid and the left side shows cases where Tn inserts in these genes cause a decrease in fitness in the presence of acetic acid. The red color-coded genes are the genes of interest, which can be further validated by experiments to observe if their deletion does indeed cause a change in fitness as measured in TraDIS experiments.

In EO499 day 1, there were fewer significant genes in comparison to day 5 after longer exposure of acetic acid was applied. This means genes which contributed to fitness under acetic acid were more statistically enriched over longer time points.

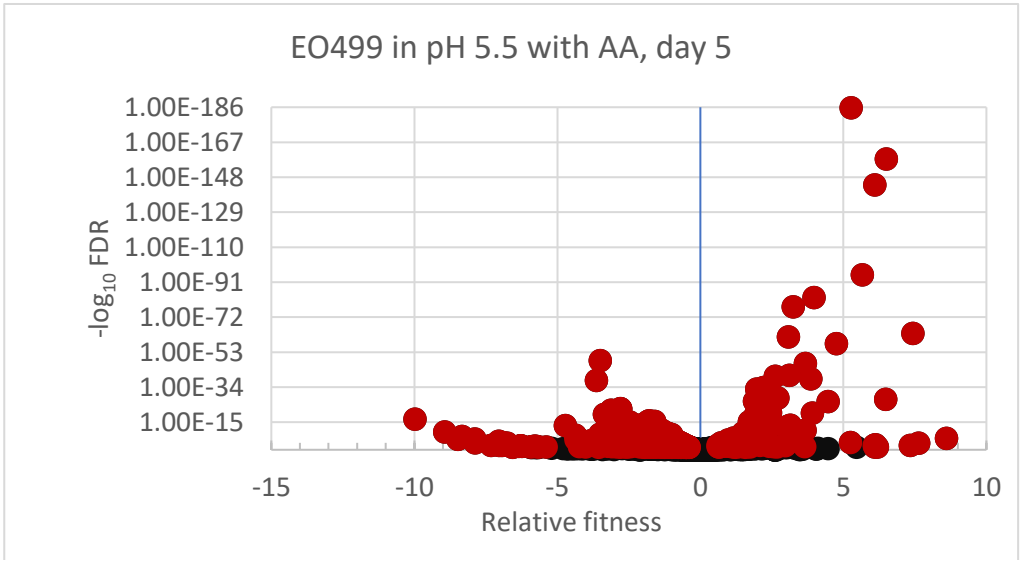
In the case of MG1655, on day 1 the EdgeR data did not show any significant mutants under acetic acid with FDR < 0.05, so the data were not plotted for day 1. Therefore, only MG1655 on day 5 were plotted, as shown in figure 62 C. The data for MG1655 also showed a volcano

pattern, as seen with EO499 62 A and 62 B. As can be seen from these graphs, MG1655 showed more significant mutants which were enriched in the presence of acetic acid than mutants which lost fitness in the presence of acetic acid. Further discussion of this result is coming later in this section.

A


EO499 in pH 5.5 with AA, Day 1
1.00E-12
1.40E-11
1.96E-10
2.74E-09
3.84E-08
5.38E-07
7.53E-06
1.05E-04
1.48E-03
2.07E-02
2.89E-01
-log10 FDR
-8
-6
-4
-2
0
2
4
Relative fitness

B


EO499 in pH 5.5 with AA, day 5
1.00E-186
1.00E-167
1.00E-148
1.00E-129
1.00E-110
1.00E-91
1.00E-72
1.00E-53
1.00E-34
1.00E-15
-log10 FDR
-15
-10
-5
0
5
10
Relative fitness

C

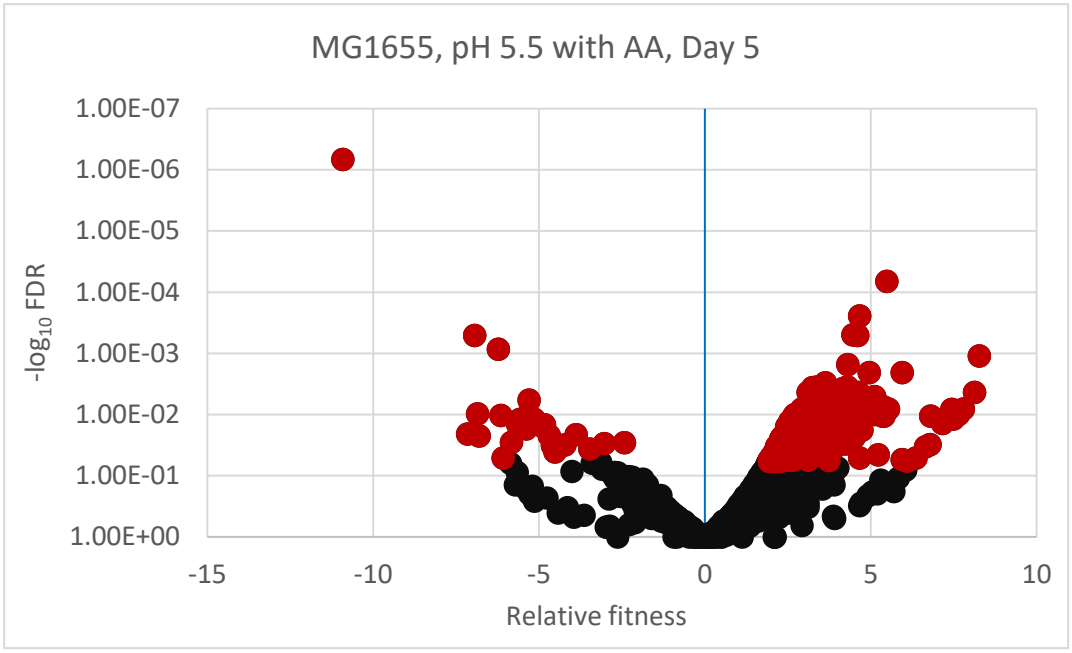


Figure 62. The relative fitness of insertions in all the non-essential genes in EO499 and MG1655 in the presence of acetic acid stress.

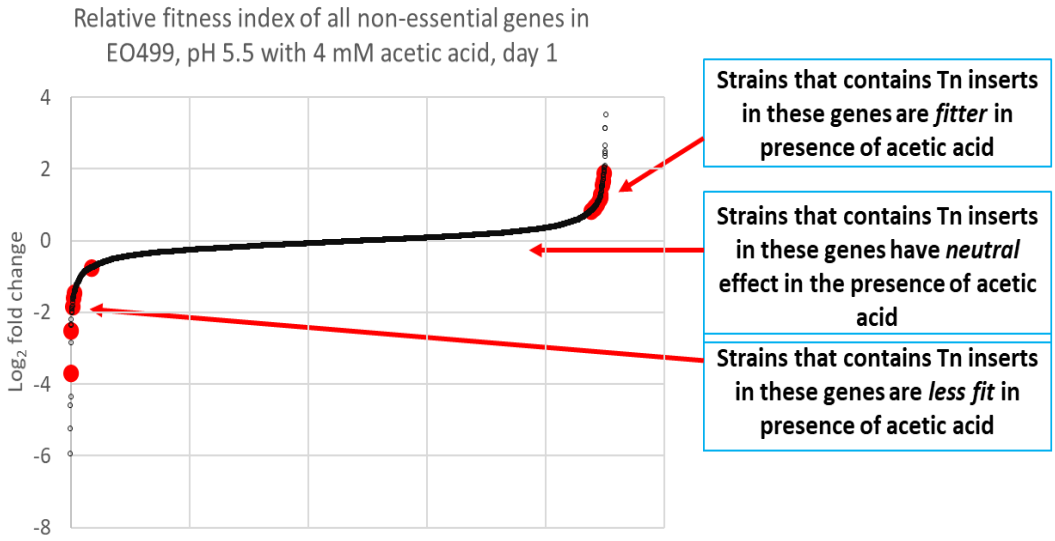
The genes were ranked according to FDR. The graphs represent the $\log_2$ fold change of each mutant vs the FDR score with a cutoff value of < 0.05 calculated by EdgeR. The black color represents non-significant mutants and red color represent genes significant difference in fitness in the presence of acetic acid with cutoff value (FDR < 0.05). A. EO499 at pH 5.5 with 4 mM AA on day 1. B. EO499 at pH 5.5 with 4 mM AA on day 5. C. MG1655 at pH 5.5 with 4 mM AA on day 5. AA: acetic acid.

Another way to plot the EdgeR data is by ranking all the non-essential genes by the $\log_2$ fold change of the relative fitness from lowest to highest, as shown in figure 63. This enables to classify genes to three different categories, in both day 1 and day 5. First, genes which, when they contained transposon inserts, led to increased fitness in the presence of acetic acid. Second, genes which, when they contained transposon inserts, led to decreased fitness under acetic acid. Third, genes which, when they contained transposon inserts caused no significant effect on fitness under acetic acid stress. The same ranking pattern was obtained previously in EO499, as shown in section 3.2. The significant mutants were defined for this analysis as being those with an FDR < 0.05, color-coded in red, figure 63. Note that this cut-off is chosen to try as much as possible to limit the false positives and false negatives. At the start with the analysis I chose FDR < 0.05 cutoff point, and if the significant gene list was too long I would choose a more stringent FDR value. This is to have a more manageable gene list. Starting with EO499 63 A, the significant genes on day 1 in red color showed a very small number of genes in the first two categories, while by day 5, figure 63B the red tail with significant genes becomes longer on both sides, showing that the longer time point reveals more genes with an impact on fitness. The tail with significant genes which are enriched was smaller than the tail with depleted genes under acetic acid. This suggested that relatively larger proportion of the mutants showed a defect in fitness on day 5 (i.e. they are important to the cell under acetic acid stress), compared to the mutants that showed enrichment under acetic acid. Considering the large number of mutants showing defect under acetic acid, this might require choosing more stringent cutoff value to FDR < 0.01 when doing further analysis on the gene list to minimize the chance of errors and to focus first on subset of these genes. Or

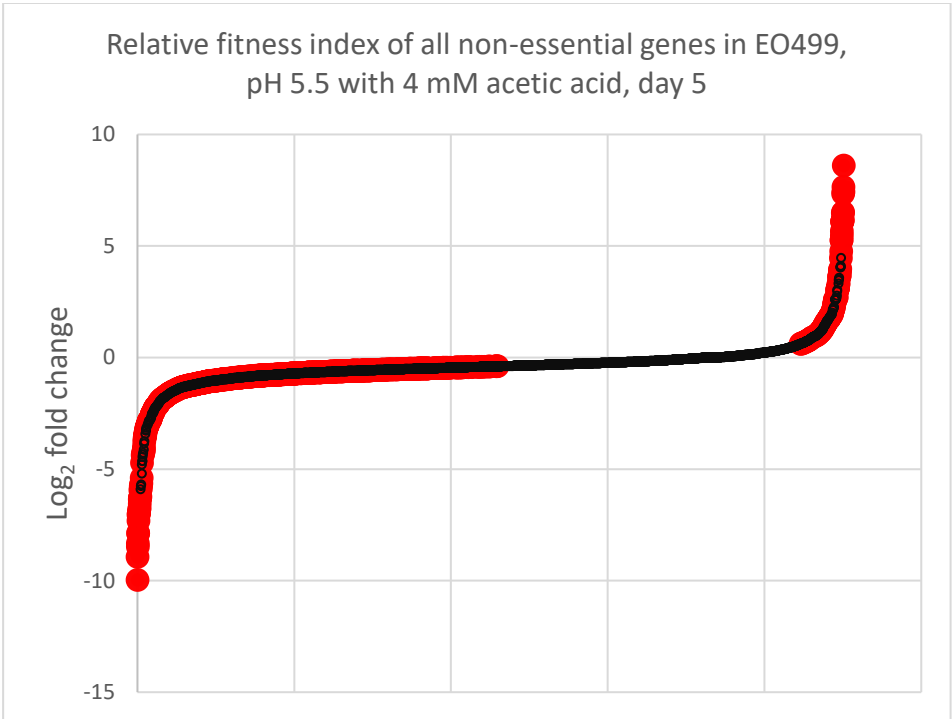
another option is to choose a cutoff point based on the graph figure 63B where there is a sudden change in the fitness. Then I generate our gene list and perform the pathways analysis. In this analysis I have chosen the first option to lower the cutoff value to FDR < 0.01. Further discussion about choosing the FDR cutoff score, coming later in the chapter.

In the case of MG1655, as mentioned earlier in this section there were no genes with FDR < 0.05 on day 1 in the presence of acetic acid, so the data were not plotted. Although there were no significant mutants under acetic acid stress the data were reproducible for both with and without acetic acid, as shown in earlier the figure 63C. MG1655 on day 5, showed fewer genes depleted in the presence of acetic acid compared to the genes enriched in the presence of acetic acid, figure 63C. Most of the FDR values were > 0.05. This pattern with MG1655 on day 5 is the opposite to that seen for EO499 on day 5. Note that this might be because MG1655 on day 5 there was low data reproducibility between replicates at pH 5.5 as shown earlier in this chapter, section 6.3.2. In the next section, significant gene lists under acetic acid for both EO499 and MG1655 will be presented.

A


Relative fitness index of all non-essential genes in EO499, pH 5.5 with 4 mM acetic acid, day 1
Log2 fold change
4
2
0
-2
-4
-6
-8
Strains that contains Tn inserts in these genes are *fitter* in presence of acetic acid
Strains that contains Tn inserts in these genes have *neutral* effect in the presence of acetic acid
Strains that contains Tn inserts in these genes are *less fit* in presence of acetic acid

B


Relative fitness index of all non-essential genes in EO499, pH 5.5 with 4 mM acetic acid, day 5
Log2 fold change
10
5
0
-5
-10
-15

C

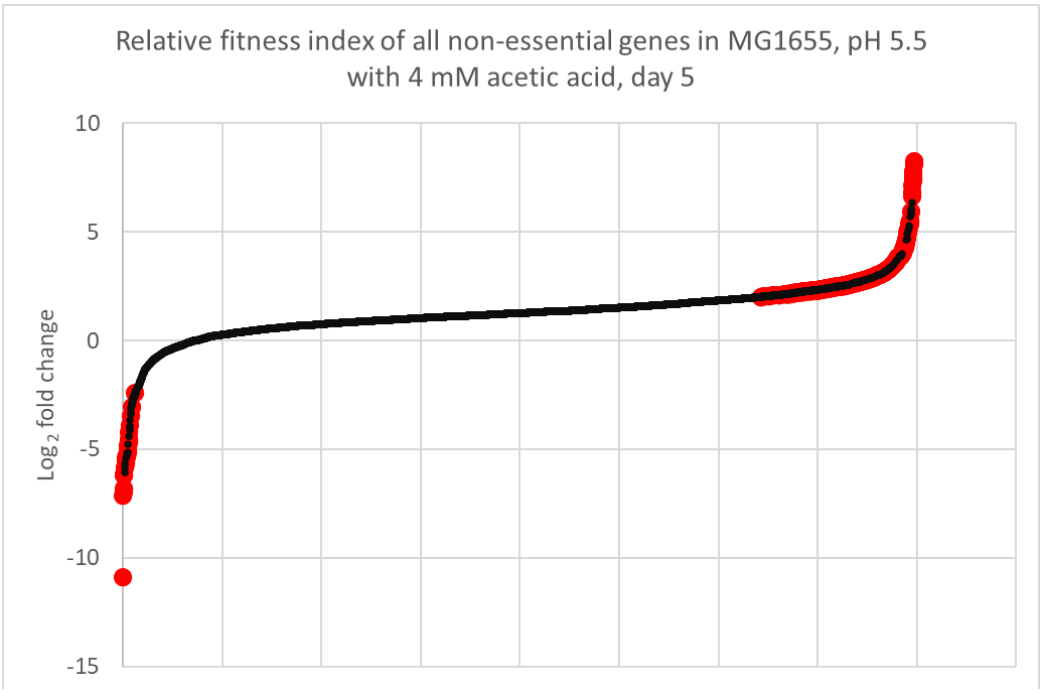


Figure 63. The impact on relative fitness of insertions in all non-essential genes in EO499 and MG1655, log_2 fold (pH 5.5 + 4 mM AA/ pH 5.5).

The red color-coded refers to log_2 fold change of mutants with a FDR < 0.05. The black color-coded referred to log_2 fold change of mutants with a FDR > 0.05. A. EO499 on day 1. B. EO499 on day 5. C. MG1655 on day 5. All genes were ranked by the log2fold change as calculated in EdegR.

6.4.1 Lists of significant genes for EO499 and MG1655

The genes where insertions have a significant relative effect on fitness in the presence of acetic acid were ranked by log_2FC on both strains (EO499 and MG1655) on day 1 and day 5, with FDR cutoff value < 0.05. The total number of significant genes of EO499 on day 1 and day 5 are

shown in table 22. As was clear in the above graphs 63 A and 63 B, the number of significant genes on day 1 were fewer compared to day 5. Also, there were more mutants causing loss of fitness in EO499 than mutants that caused enrichment under acetic acid stress.

In the case of MG1655 no significant genes were found on day 1, table 22. On day 5, the mutants that caused enrichment in the presence of acetic acid were more frequent than in EO499. In this experiment, the single time point (24 hrs.) did not fully capture the impact of acetic acid stress, as only a few genes were observed to be changing significantly in EO499 and none in MG1655. The use of a five daytime point is thus more useful to identify genes important in determining fitness in the presence of acetic acid at pH 5.5.

Table 22. A summary of the total number of significant mutants in both strains EO499 and MG1655 on day 1 and day 5.

There were no significant genes identified in MG1655, day 1. The numbers refer to mutants with $\log_2$ change and FDR < 0.05.

Strains	Total significant genes	Day 1	Day 5
EO499	Genes in which Tn inserts caused increase of fitness under acetic acid stress	10	112
	Genes in which Tn inserts cased loss of fitness under acetic acid stress	8	829
MG1655	Genes in which Tn inserts caused increase of fitness under acetic acid stress		513
	Genes in which Tn inserts cased loss of fitness under acetic acid stress		29

For significant genes identified by TraDIS, the lists of genes below are provided with the general format feature gene name (GFF), gene name, and the predicted function, alongside the log_2FC and FDR score. Starting with EO499, the 10 genes where inserts caused a significant increase of fitness under acetic acid stress on day 1, are shown in table 24, ranked by their log_2FC. The 8 genes in which Tn inserts cased loss of fitness under acetic acid stress on day 1 are listed in table 24. The genes were ranked based on log_2FC, from mutants that caused the largest decrease in fitness. For EO499 on day 5, the genes in which Tn inserts caused an increase of fitness under acetic acid are shown in table 26, ranked in the same way as in table 24 because of the large number of genes. Table 25 shows only the top 25 genes and the rest of the significant genes list is available in the supplementary data table S16. Finally, table 27, showed the top 25 significant genes in which Tn inserts caused decreased fitness under acetic acid stress in EO499 on day 5. The rest of the significant genes list are available in the supplementary data table S17. In this case, using a cutoff value of FDR < 0.05 showed a very long genes list, increasing the chance of false positive. In order to have more manageable genes list for EO499 on day 5 for genes that were depleted, an FDR < 0.01 was used to lower the number of gene list from 829 to 278 genes.

As a comparison between EO499 on day 1 and day 5, I have expected to observe that the significant mutants on day 1 will be also present in day 5. The significant enriched mutants were found both day 1 and day 5, in table 24 and table 26. While five mutants (*sucA*, *purH*, *carB*, *mnmG* and *cpxA*) were depleted in day 1 under acetic acid in EO499, were found in day 1 but not significant in day 5.

Table 23. Lists of genes in which Tn inserts caused increase of fitness under acetic acid stress in EO499 on day 1 identified by TraDIS.

The genes were ranked based on logFC score from larger to smaller, with cutoff value of FDR < 0.05.

GFF gene name	Gene name	Predicted function	log_2FC	FDR
ST131_cds_01394	*ackA*	Acetate kinase	1.89	1.82E-05
ST131_cds_01014	*rpoS*	Rna polymerase, sigma s (sigma 38) factor	1.67	3.38E-12
ST131_cds_02023	*cspC*	Stress protein, member of the cspa family	1.56	2.13E-05
ST131_cds_03908	*dnaJ*	Chaperone protein dnaj	1.27	5.84E-06
ST131_cds_04206	*hfq*	Rna-binding protein hfq	1.20	2.67E-03
ST131_cds_04457	*fabR*	Dna-binding transcriptional repressor fabr	1.15	3.54E-04
ST131_cds_00169	*gadX*	Dna-binding transcriptional dual regulator gadx	0.93	7.02E-05
ST131_cds_00175	*gadE*	Dna-binding transcriptional activator gade	0.92	2.93E-02
ST131_cds_01470	*rcsB*	Dna-binding transcriptional activator rcsb	0.83	2.31E-03
ST131_cds_03532	*iraP*	Anti-adaptor protein for sigma(s) stabilization	0.78	5.12E-02

Log_2 FC: logFC of number of reads mapped to each gene from pH 5.5 + acetic acid at day 1 / pH 5.5 at day 1.

Table 24. Lists of genes in which Tn inserts caused loss of fitness under acetic acid stress in EO499 on day 1 identified by TraDIS.

The genes were ranked based on logFC score from smaller to larger, with cutoff value of FDR < 0.05.

GFF gene name	Gene name	Predicted function	log_2FC	FDR
ST131_cds_03211	*sucA_1*	Subunit of E1(0) component of 2-oxoglutarate dehydrogenase	-7.42	1.45E-02
ST131_cds_04413	*purH*	Bifunctional AICAR transformylase/IMP cyclohydrolase	-3.71	1.22E-05
ST131_cds_03889	*carB*	Carbamoyl phosphate synthetase subunit beta	-2.53	1.89E-04
ST131_cds_04699	*mnmG*	5-carboxymethylaminomethyluridine-trna synthase subunit mnmg	-1.84	3.24E-05
ST131_cds_04511	*cpxA*	Sensory histidine kinase cpxa	-1.61	5.61E-05
ST131_cds_03515	*phoR*	Sensory histidine kinase phor	-1.47	2.40E-03
ST131_cds_04665	*gpp*	Guanosine-5'-triphosphate,3'-diphosphate phosphatase	-1.01	9.96E-02
ST131_cds_00839	*gcvP*	Glycine decarboxylase	-0.76	1.45E-02

Log_2 FC: logFC of number of reads mapped to each gene from pH 5.5 + acetic acid at day 1 / pH 5.5 at day 1.

Table 25. Top 25 genes in which Tn inserts caused increase of fitness under acetic acid stress in EO499 on day 5 identified by TraDIS.

The genes were ranked based on logFC score from larger to smaller, with cutoff value of FDR < 0.05.

GFF gene name	Gene name	Predicted function	log_2FC	FDR
ST131_cds_00519	*rimP*	ribosome maturation factor RimP	8.60	7.04E-07
ST131_cds_02770	*rpmF*	50S ribosomal subunit protein L32	7.63	2.25E-04
ST131_cds_01394	*ackA*	acetate kinase	7.42	7.01E-64
ST131_cds_00047	*rpmG*	50S ribosomal subunit protein L33	7.34	4.23E-03
ST131_cds_03908	*dnaJ*	chaperone protein DnaJ	6.50	1.20E-158
ST131_cds_01393	*pta*	phosphate acetyltransferase	6.47	3.46E-28
ST131_cds_04707	*atpG*	ATP synthase F1 complex subunit gamma	6.18	2.80E-02
ST131_cds_04425	*thiC*	phosphomethylpyrimidine synthase	6.11	2.06E-03
ST131_cds_01014	*rpoS*	RNA polymerase, sigma S (sigma 38) factor	6.10	1.67E-144
ST131_cds_00403	*def*	peptide deformylase	6.09	3.59E-02
ST131_cds_04414	*purD*	phosphoribosylamine--glycine ligase	6.07	4.95E-03
ST131_cds_00918	group_4746	hypothetical protein	5.97	4.83E-02
ST131_cds_02023	*cspC*	stress protein, member of the CspA family	5.66	9.17E-96
ST131_cds_00690	*exbB*	Ton complex subunit ExbB	5.45	4.24E-02
ST131_cds_01013	*nlpD*	murein hydrolase activator NlpD	5.26	2.19E-186
ST131_cds_00073	*secB*	SecB chaperone	5.25	1.28E-04
ST131_cds_02016	*proQ*	RNA chaperone ProQ	4.75	2.38E-58
ST131_cds_00034	*spoT*	bifunctional (p)ppGpp synthase/hydrolase SpoT	4.46	7.66E-27
ST131_cds_03909	*dnaK*	chaperone protein DnaK	4.28	5.62E-05
ST131_cds_01146	*srmB*	ATP-dependent RNA helicase SrmB	3.91	1.05E-20
ST131_cds_04154	group_4441	hypothetical protein	3.86	2.91E-39
ST131_cds_04457	*fabR*	DNA-binding transcriptional repressor FabR	3.66	1.30E-47
ST131_cds_03351	*fes_1*	enterochelin esterase	3.65	2.35E-11
ST131_cds_02430	*rnb*	RNase II	3.56	1.06E-09
ST131_cds_00430	*fis*	DNA-binding transcriptional dual regulator Fis	3.32	6.57E-09

Log2 FC: logFC of number of reads mapped to each gene from pH 5.5 + acetic acid at day 5 / pH 5.5 at day 5.

Table 26. Top 25 genes in which Tn inserts caused loss of fitness under acetic acid stress in EO499 on day 5 identified by TraDIS.

The genes were ranked based on log_2 FC score from smaller to larger, with cutoff value of FDR < 0.01.

GFF gene name	Gene name	Predicted function	log_2FC	FDR
ST131_cds_02017	*prc*	Tail-specific protease	-9.98	1.97E-17
ST131_cds_03208	*sucD_1*	Succinyl-coa synthetase subunit alpha	-8.95	2.48E-10
ST131_cds_00838	*gcvH*	Glycine cleavage system h protein	-8.49	2.33E-06
ST131_cds_02885	*ompA*	Outer membrane porin a	-8.34	5.71E-08
ST131_cds_03213	*sdhA*	Succinate:quinone oxidoreductase, fad binding protein	-7.89	1.35E-06
ST131_cds_02269	*mgtS*	Small protein mgts	-7.32	3.81E-03
ST131_cds_04222	*epmA*	Ef-p-lysine lysyltransferase	-7.05	2.78E-05
ST131_cds_03244	*pgm*	Phosphoglucomutase	-7.04	3.79E-03
ST131_cds_03217	*gltA*	Citrate synthase	-6.98	1.28E-04
ST131_cds_00600	*higA*	Antitoxin/dna-binding transcriptional repressor higa	-6.94	4.21E-03
ST131_cds_01472	*ompC*	Outer membrane porin c	-6.89	1.58E-04
ST131_cds_03215	*sdhC*	Succinate:quinone oxidoreductase, membrane protein sdhc	-6.80	3.01E-03
ST131_cds_02774	*rne*	Ribonuclease e	-6.76	2.33E-04
ST131_cds_04552	*glnA*	Glutamine synthetase	-6.57	3.84E-02
ST131_cds_03083	*ybjN*	Protein ybjn	-6.36	1.15E-02
ST131_cds_01405	*nuoC*	Nadh:quinone oxidoreductase subunit cd	-6.29	1.09E-02
ST131_cds_01284	*cysK*	O-acetylserine sulfhydrylase a	-6.28	7.27E-03
ST131_cds_03450	group_5294	Hypothetical protein	-6.25	1.57E-02
ST131_cds_03185	*nadA*	Quinolinate synthase	-5.95	2.58E-02
ST131_cds_01411	*nuoJ*	Nadh:quinone oxidoreductase subunit j	-5.89	1.85E-02
ST131_cds_03780	*panB*	3-methyl-2-oxobutanoate hydroxymethyltransferase	-5.79	8.11E-03
ST131_cds_03157	*bioD*	Dethiobiotin synthetase	-5.71	3.59E-02
ST131_cds_03870	*apaH*	Diadenosine tetraphosphatase	-5.65	2.96E-02
ST131_cds_03918	*thrC*	Threonine synthase	-5.40	3.13E-02
ST131_cds_00461	*mdh*	Malate dehydrogenase	-4.72	6.00E-14

Log_2 FC: logFC of number of reads mapped to each gene from pH 5.5 + acetic acid at day 5 / pH 5.5 at day 5.

Starting with the enriched mutants under acetic acid stress, *ackA* mutant was enriched in day 1, while *ackA* and *pta* mutants were enriched in day 5. To check whether the enrichment of *ackA* and *pta* mutants is already present in the ITL or is an effect of acetic acid. *ackA* and *pta* were investigated in the ITL, for RPKMs and read counts. First, the RPKMs in the EO499 ITL were ranked from largest to smallest, the *ackA* was found in the rank of 3949 (RPKM= 60), and *pta* was found in rank of 3925 (RPKM= 61). This means these mutants have a low level of RPKMs. Second, Artemis inspection for *ackA* and *pta* did not show any enrichment (data not shown). Therefore, I can say *ackA* and *pta* enrichment as seen after EdgeR analysis is an effect of the acetic stress.

As it is mentioned in the introduction section 1.7, *ackA* and *pta* play a major role in acetate production versus the switch to acetate consumption as a carbon source in *E. coli*. The products of these two genes catalyze the conversion of acetyl-CoA to acetate with ATP production, figure 4 (Wolfe, 2005). Acetate production occurs both under aerobic and anaerobic growth conditions as part of the mixed acid fermentation pathway. This result suggested that loss of acetate metabolism is advantageous under acetic acid stress, as strains containing either of these two mutants were found to be enriched in the presence of acetic acid at pH 5.5, compared to pH 5.5 only. It was found earlier that *ackA* and *pta* genes showed reduction of expression level at pH 6 medium supplemented with acetate. (Orr et al., 2019). This is consistent with our data here. As strains with mutations in *ackA* and *pta* cannot produce acetate, this suggests that the production of acetate might be harmful in the presence of acetate in the medium (for example, by increasing the internal acetate concentration). Therefore, when cells can't produce acetate by this pathway, they have growth advantages, as shown here.

Also, strains with mutations in *rpoS* were enriched on day 1, and strongly enriched on day 5, under acetic acid stress. *rpoS* is a sigma factor that regulates stress responses in *E. coli*, including the acid resistance systems (AR) in response to acid stress in *E. coli*. This system is activated by the σ factor RpoS and cAMP receptor protein (CRP) (Lin et al., 1995). The gene *cpdA* encodes the cAMP phosphodiesterase, and this also showed enrichment on day 5, supplementary table S6. This is an enzyme which regulates the hydrolysis of cAMP will affect the level of transcription expression of genes controlled by cAMP-CRP. The AR systems can maintain the internal pH by consuming the protons by amino acid decarboxylation reaction, using the antiporters of the substrates glutamate, arginine, lysine and ornithine (Foster, 2004; Castanie-Cornet et al., 1999). There are up to five AR systems, but the most effective is AR2, which is the most complex one, figure 64. The AR2 system uses the glutamate decarboxylases, GadA and GadB, to convert glutamate to GABA (γ-aminobutyric acid). This reaction is done by replacing the α-carboxyl group (-COOH) with a proton in glutamate which results in GABA and CO_2. This reaction lowers the cytoplasmic pH by reducing the proton concentration. Then the antiporter GadC, imports glutamate with exchange of GABA (the decarboxylation product) (Lund et al., 2014). This regulation of this system depends on a number of two component systems and regulatory proteins. The two component systems are EvgSA and PhoPQ. The regulatory protein includes RpoS, YdeO, GadE, GadX, GadW, and RcsB, all together forming a complex regulatory network. (Lund et al., 2014; Zhao and Houry, 2010). The two-component system EvgSA encodes a sensor histidine kinase that confers acid resistance by activation of AR2. For PhoPQ, PhoP is a DNA-binding transcriptional dual regulator and PhoQ is a bifunctional sensor histidine kinase. Among

other effects, PhoPQ regulates the protein IraM which is involved in the acid resistance control. Among all the regulators GadE is the central activator of the acid resistance and its expression involve regulators such as EvgA, YdeO, GadE, PhoP, GadX, and GadW (Sayed et al., 2007). When the EvgS sensor is activated it activates the transcription factor EvgA, which activates YdeO, which in turn activates GadE, which induces the expression of a large number of genes involve in acid resistance. PhoQ/PhoP is activated by SafA (sensor associating factor A), which connects both EvgS/EvgA and PhoQ/PhoP systems. Once the PhoQ/PhoP is activated it will activate the anti-adaptor IraM, that will bind to the RssB preventing RpoS to be degraded by ClpXP. Also, the PhoQ/PhoP induces the expression of *gadW* and *hdeA*. The RcsB stand for Regulator capsule synthesis B, it belongs to two components regulatory system RcsC/RcsB. It is a response regulator in which RcsB and interact with GadE to activate the expression on *gadABC* (Xu, 2020; Lund, 2014; Eguchi, 2011).

Our data showed enrichment of strains carrying mutations in the genes *rpoS*, *gadE*, *gadX* and *rcsB* on both day 1 and day 5, table 24, table 25 and supplementary table S16, in the presence of acetic acid at pH 5.5. A study showed that the expression of *gadAB* is activated by *gadE* rather than RpoS but does required *gadE* for activation (Castanie-Cornet et al., 2007). This means the cells will still be able to activate the glutamate-dependent acid resistance with the absence of RpoS, and this mutant can enhance the activity of GadE toward have a positive effect toward acetic acid stress. It was shown, RpoS mutation induce the alteration in the genes expression, in the stationary growth phase. Also, RpoS have important regulatory roles in controlling many genes and pathways, in the exponential growth phase (Rahman et al., 2006). However, it is hard

to determine the true extent of the effects of loss of RpoS function because it effect many genes and pathways that play key roles in stress response. For instance, RpoS involve in the expression of *poxB* that catalyzes the conversion of pyruvate to acetate (Suryadarma et al., 2012). Also, RpoS has a large impact on gene expression during growth in stationary phase including the expression of acid resistance system AR2. Therefore, RpoS mutants expected to be changed in many different ways, and it is not possible to tell from TraDIS data only the cause of enrichment of *rpoS* mutants from the examined conditions.

The enrichment of mutations in *gadE*, *gadX* and *rcsB*, showed that loss of function of these genes has a beneficial effect on cells under acetic acid stress at pH 5.5. It was shown earlier that RcsB is a fundamental and necessary requirement for the glutamate-dependent acid resistance, as both GadE and RcsB were needed for the activation of GadAB acid resistance system (Castanie-Cornet et al., 2007). GadX was also shown to regulate the GadAB system when GadE is absent (Seo et al., 2015). These results suggested that the absences of one protein that activate the acid resistance system can be replaced by another protein in EO499 for pH homeostasis. The much more likely explanation for this is that, in spite of the fact that AR2 system is required for optimum growth when pH is low, but it is more likely to be deleterious when acetate present in the medium. Thus, mutants that are less efficient at activating it (*gadE, rcsB, gadX*) currently do better. The reason is not known yet, but it is supported with our data.

Also, mutations in *phoQ* cause loss of fitness of the strain under acetic acid on day 5, supplementary table S7. A study found a mutant phoQ can't activate the PhoP in response to

Mg^{2+} limitation (Regelmann et al., 2002). This might limit the RpoS available is the cell for acid resistance tolerance. Thus, because the RssB to deliver the RpoS to the ClpXP protease for degradation. Then, the available RpoS will not be enough to activate GadE. In summary, the acid resistance system is a large complicated system, a number of genes were spotted in the system to cause an enrichment or reduction when mutated under acetic acid stress. This showed the acid resistance system is important under acetic acidic stress.

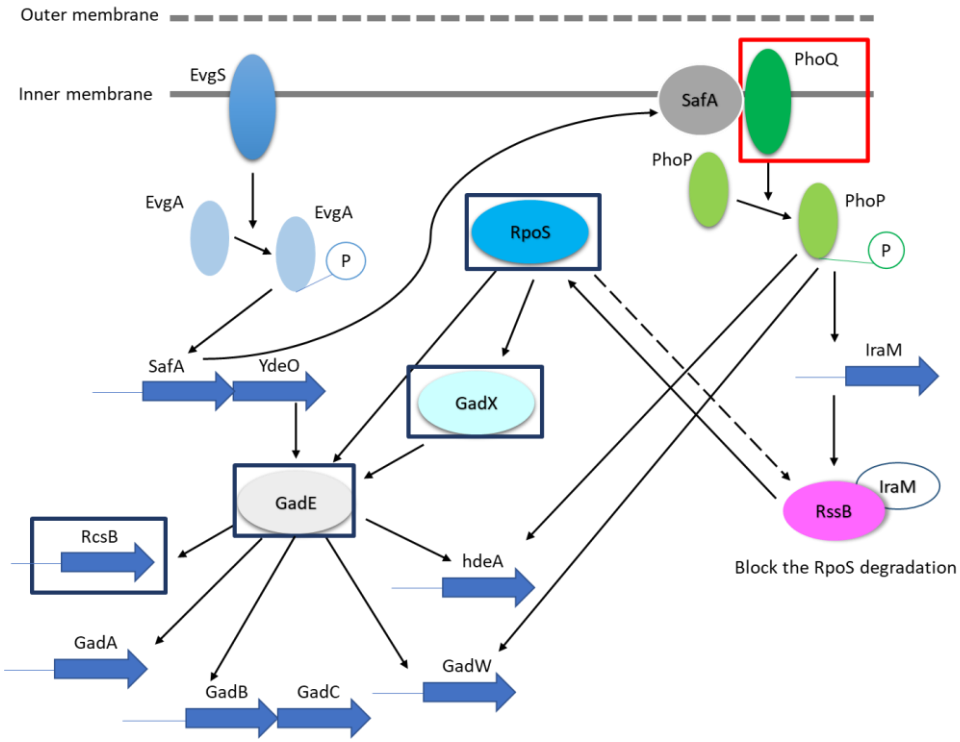


Figure 64. The acid resistance 2 (AR2) system in *E. coli.*

Activation of this acid resistance system is initiated by EvgS. This activates a series of regulators EvgA-YdeO-GadE. EvgS/EvgA also activates the PhoQ/PhoP via a small protein SafA, which enhances acid resistance by increasing the RpoS level in the cell. This occurrs via IraM anti-adaptor which blocks RssB from assisting the degradation of RpoS. Then RpoS will activate GadE which is the central regulator of AR2. P: Phosphate. The black arrows indicate activation. Blue boxes show enriched mutants and the red box shows mutants with loss of fitness under acetic acid stress as shown by TraDIS data. Figure adapted from (Burton et al., 2010; Eguchi et al., 2011; Eguchi and Utsumi, 2014).

Furthermore, our results showed enrichment of mutations in *fabR* and *fabF*. *fabR* was enriched on both day 1 and day 5, table 24 and 26 supplementary table S16, while *fabF* was enrich on day 5 only. *fabR* stand for Fatty Acid Biosynthesis Regulator, and it was reported as repressor for both *fabA* and *fabB,* which are required for synthesis of monounsaturated fatty acids (Feng and Cronan, 2011). *E. coli* in low pH tends to increase the saturated fatty acyl chains as a response to low pH (Arneborg et al., 1993). This can explain why *fabR* mutant strain showed enrichment in our TraDIS results, as the absence of *fabR* from the cells, may increase saturated fatty acid synthesis which will control the membrane fluidity under stress. *fabF* encodes the enzyme β-ketoacyl-[acyl carrier protein] synthase (KAS), which is responsible for chain elongation steps during the biosynthesis of saturated fatty acids, which is associated with stress responses (Tsay et al., 1992; Zhang et al., 2012a). This process can be carried out by FabF or FabB (Cronan and Thomas, 2009). There is a lack of studies on the fatty acid synthesis during acetic acid stress, and our results suggest this may be worth further study.

Our TraDIS data showed enrichment in *atpG* mutants on day 5, table 26. *atpG* is a subunit of ATP synthase F1Fo-ATPase complex (ATP biosynthesis), mentioned in the introduction section 1.6. The F1Fo-ATPase produces ATP from ADP and inorganic phosphate using the electro-chemical gradient (movement of protons) in the inner membrane during oxidative phosphorylation. *atpG* is involved in production of ATP using the proton gradient. In low pH, the FoF1-ATPase is also involved in proton pumping, consuming ATP in order to maintain the internal pH (Richard and Foster, 2004).

I suggest that because acetic acid at pH 5.5 will disrupt and reduce the efficiency of the proton gradient, but maybe not to zero, so cells can continue to produce ATP using the proton gradient. This means any mutation that reduce the ability needed to make ATP using the energy stored in the proton gradient will have greater effect in reducing the proton gradient in the presence of acetic acid. Including mutations in the components of TCA cycle (that makes NADH/$FADH_2$), components of the oxidative phosphorylation pathway or the ATPase.

There are several genes involved in the TCA pathway which caused loss of fitness under low pH with acetic acid when mutated. This included *sucA* on day 1, and *sdhAC, gltA, sucADC, fumAC, mdh* on day 5, found in table 25, 27 and S17. All of these genes correspond to enzymes or component of enzymes that catalysis a reaction in the TCA, figure 65. As described in the Introduction section 1.6, at low pH the organic acids including acetic acid will cause a collapse of the proton gradient, formed by the electron transport chain re-oxidising NADH and $FADH_2$ from the TCA cycle. Therefore, cells exposed to acetic acid will have already be limited in ATP production by the process of oxidative phosphorylation that requires a proton gradient. Because the TCA cycle is the central metabolic hub of the cell, which is needed to produce of NADH and $FADH_2$ for the proton gradient under the acetic acid effect, finding these mutants as more deleterious is not surprising because mutants in the mentioned genes will lose the ability to maintain the TCA cycle, which will restrict even more their ability to produce ATP. By the same argument, I have expected mutations in components of the electron transport chain would also have a negative effect on fitness of the cell in the presence of acetic acid. Starting with NADH dehydrogenase complex I in the electron transport chain of the oxidative phosphorylation, *nuoC*

and *nuoJ* are part of this first protein complex in the electron transport chain. And *sdhA* and *sdhC,* both are a part of the succinate dehydrogenase, the second complex in the electron transport chain. These genes are the key component of the oxidative phosphorylation and they also show loss of fitness when they are mutated under acetic acid, as predicted.

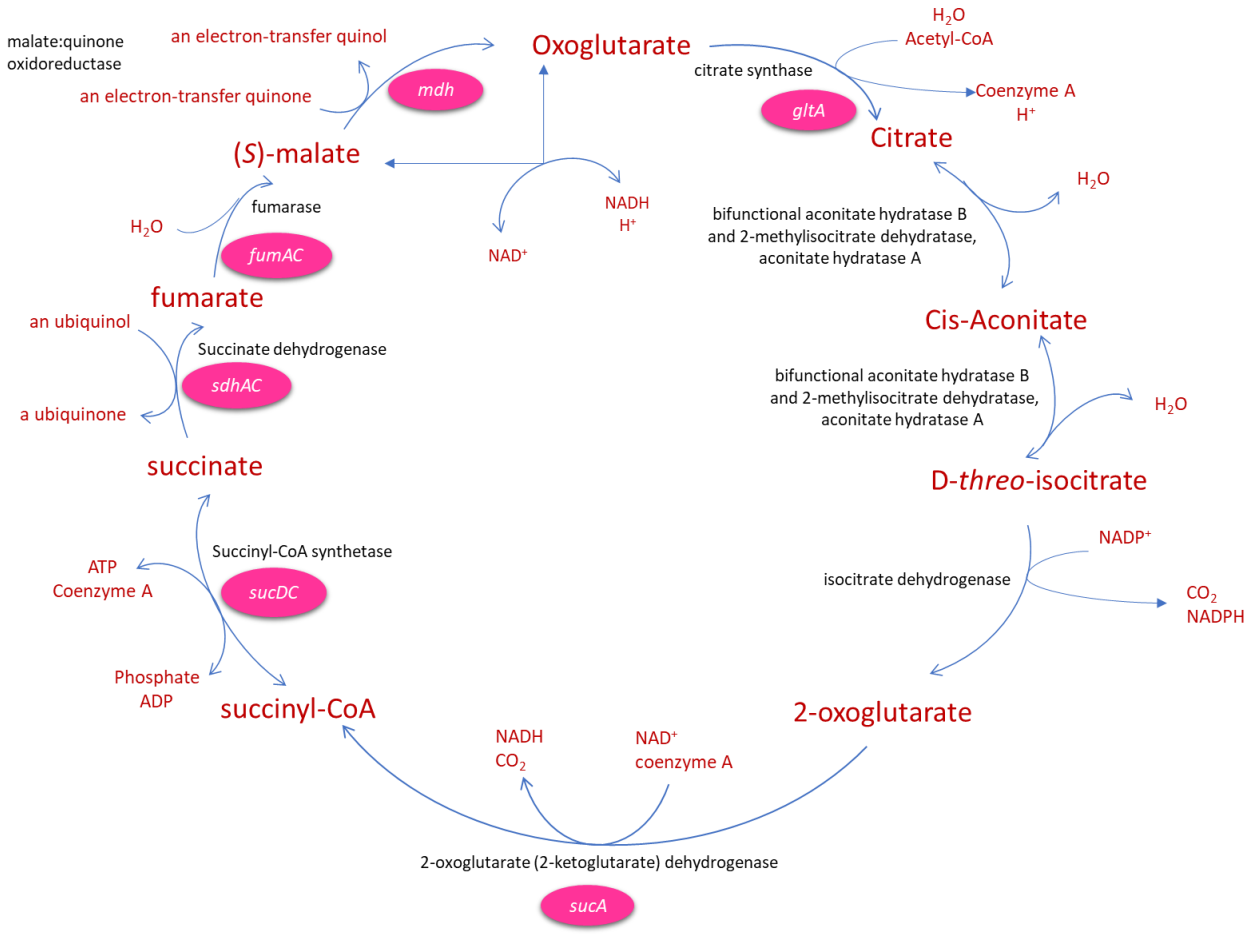


Figure 65. TCA pathway of *E. coli.*

The pink circles represent genes which showed loss of fitness when they are mutated in TraDIS under acetic acid stress.

In MG1655, EdgeR showed the list of significant mutants of MG1655 on day 5, table 28, showed the top 25 significant genes in which Tn inserts caused increased of fitness of MG1655 on day 5. The genes list was ranked based on log_2FC, from mutants that caused the larger increase in fitness, with a cutoff FDR < 0.05. The rest of the significant gene list available in the supplementary data in table S18. While the top genes in which Tn insets caused loss of fitness under acetic acid stress shown in table 29. The mutants were ranked based on log_2FC score from smaller to larger, with cutoff value of FDR < 0.05. The rest of the of the significant mutants list were shown in the supplementary data table S19.

MG1655, although it doesn't look much like EO499 and it is a bit less reliable in day 5, where is most of the significant changes happen. As it is stated earlier, no significant mutants were determined on day 1 under acetic acid stress. For MG1655, I am not attempting to do the analysis in this much detail as EO499, because of the low reproducibility shown in the graphs and from the EdegR output from the analysis the significant mutants causing the strains less fitness under acetic acid were much lower. However, a few examples of the significant genes under acetic acid stress.

Also, our data showed enrichment on *eutNQ* mutants*, eutN* encode a protein for ethanolamine utilization. It is suggested that is a pore with a transport function due to its structure, table 28. The information on these genes is limited so far (Forouhar et al., 2007). The EdgeR data showed enrichment of *efeU* mutants on day 5, encode inner membrane protein specific for ferrous iron induced by low pH, table 28. It was found that *efeU* was enriched in both

E. coli K-12 and O157:H7 during the induction to acetic acid (King et al., 2010). But it still not clear yet why *efeU* mutant showed a growth advantages under acetic acid stress, it could be that iron acquisition is toxic to the cell under acetic acid stress. Another gene encode inner membrane protein *ccmD*, involve in the pathway: cytochrome *c* biogenesis. cytochrome are pigments that contain iron, table 26. Bacteria can use cytochrome as electron carrier in the electron transport chain (Sanders et al., 2010). Also, another membrane protein found were *gltJL* mutants showing enrichment in day 5, the data not provided due to long list of genes. *gltJL* encode glutamate transporter complex. There are a number of genes encode a variety of different membrane proteins showing advantages under acetic acid in MG1655. I have expected to find genes encode membrane protein to be show reduction in fitness when they are mutated. Because it is generally known the effect of organic acid known to cause perturbation of membrane function. This is different to what EO499 were showing, in EO499 a number of genes encode membrane protein showed a fitness defect when they are mutated under acetic acid. However, this could be due to difference among strains or it could be the results are not reliable at this stage.

The data in MG1655 have also showed enrichment of *ackA* similar to EO499 (details in the nest section). This suggested the block of acetate production are advantageous, in which the acetate is almost certainly toxic in some way. As the knockout of *ackA* prevents the accumulation of toxic acetate inside the cell. This overlap in MG1655 and EO499 confirm the acetate production pathway, is an upregulated. Another mutant *arpA* which regulates the of acetyl CoA synthetase which is also involve in the acetate production pathway.

While there are some genes showing reduction in fitness in MG16655 when they are mutated on day 5 in acetic acid. An example, *phoR* encode enzyme sensor histidine kinas, regulate the response to the level of extracellular inorganic phosphate and it is part of the two-component system PhoRB which regulates a large number of genes. This gene *phoR* showed reduction in EO499 when mutated as well on day 1. These results suggested *phoR* is important in both MG1655 and EO499 under acid. It was found that PhoRB system helps the cell sense external acidity environment and regulate the transcription of genes that mediate the response for acid shock resistance (Marzan and Shimizu, 2011). Furthermore, the MG1655 showed fitness reduction of *atpG* when mutated, *atpG* encode ATP synthase F1FoATPase complex (ATP biosynthesis). This is the opposite of the EO499 data, where *atpG mutant* showed enrichment under acetic acid. This means mutation in *atpG* cause growth defect in MG1655 and cause growth advantages in EO499.

Table 27. Top 25 genes in which Tn inserts caused increased of fitness under acetic acid stress in MG1655 on day 5 identified by TraDIS.

The genes were ranked based on logFC score from larger to smaller, with cutoff value of FDR < 0.05.

GFF gene name	Gene name	Predicted function	log_2FC	FDR
MGcds_02466	*eutN*	Putative carboxysome structural protein	8.27	1.09E-03
MGcds_01143	*ymfT*	E14 prophage; putative dna-binding transcriptional regulator ymft	8.12	4.31E-03
MGcds_02281	*yfbN*	Putative protein yfbn	7.79	8.05E-03
MGcds_01969	group_2678	Hypothetical protein	7.65	9.83E-03
MGcds_01018	*efeU*	Ferrous iron permease efeu	7.49	1.18E-02
MGcds_04225	*rpsR*	30s ribosomal subunit protein s18	7.46	3.61E-09
MGcds_02454	*yffM*	Cpz-55 prophage; uncharacterized protein yffm	7.44	8.16E-03
MGcds_02203	*ccmD*	Heme trafficking system membrane protein ccmd	7.38	1.19E-02
MGcds_02864	*ygeG*	Tpr repeat-containing putative chaperone ygeg	7.17	1.39E-02
MGcds_02870	*ygeM*	Putative protein	6.80	1.05E-02
MGcds_01363	*ydaS*	Rac prophage; toxin ydas	6.79	3.05E-02
MGcds_00150	*erpA*	Iron-sulfur cluster insertion protein erpa	6.67	4.36E-02
MGcds_01331	*ymjC*	Hypothetical protein	6.65	3.34E-02
MGcds_01664	*ynhF*	Stress response membrane protein ynhf	6.36	5.10E-02
MGcds_00079	*ftsL*	Cell division protein ftsl	6.27	3.30E-02
MGcds_00678	group_1672	Hypothetical protein	5.95	5.42E-02
MGcds_00661	*ybeY*	Endoribonuclease ybey	5.94	2.05E-03
MGcds_03712	*dnaN*	Beta sliding clamp	5.55	5.02E-03
MGcds_01728	*arpA_1*	Regulator of acetyl coa synthetase	5.53	8.05E-03
MGcds_01256	*yciB*	Inner membrane protein	5.49	8.05E-03
MGcds_01093	*fabF*	Beta-ketoacyl-[acyl carrier protein] synthase ii	5.48	6.55E-05
MGcds_01873	*nudB*	Dihydroneopterin triphosphate diphosphatase	5.40	1.05E-02
MGcds_01692	*ydiH*	Protein ydih	5.38	7.43E-03
MGcds_00734	*cydX*	Cytochrome bd-i ubiquinol oxidase subunit cydx	5.36	1.05E-02
MGcds_01393	*paaB*	Phenylacetyl-coa 1,2-epoxidase subunit b	5.23	4.53E-02

Log_2 FC: logFC of number of reads mapped to each gene from pH 5.5 + acetic acid at day 5 / pH 5.5 at day 5.

Table 28. Top 25 genes in which Tn inserts caused loss of fitness under acetic acid stress in MG1655 on day 5 identified by TraDIS.

The genes were ranked based on logFC score from smaller to larger, with cutoff value of FDR < 0.05.

GFF gene name	Gene name	Predicted function	log_2FC	FDR
MGcds_03750	*rsmG*	16s rrna m(7)g527 methyltransferase	-10.91	6.70E-07
MGcds_01884	*argS*	Arginine--trna ligase	-7.47	4.15E-02
MGcds_02511	*purM*	Phosphoribosylformylglycinamide cyclo-ligase	-7.15	2.05E-02
MGcds_00400	*phoR*	Sensory histidine kinase phor	-6.94	5.07E-04
MGcds_00740	*tolB*	Tol-pal system periplasmic protein tolb	-6.86	9.60E-03
MGcds_00928	*aspC*	Aspartate aminotransferase	-6.81	2.21E-02
MGcds_02790	*pyrG*	Ctp synthetase	-6.28	5.35E-02
MGcds_04400	*prfC*	Peptide chain release factor rf3	-6.23	8.45E-04
MGcds_01826	*manZ*	Mannose-specific pts enzyme iid component	-6.15	1.02E-02
MGcds_03743	*atpG*	Atp synthase f1 complex subunit gamma	-6.08	5.06E-02
MGcds_03651	*pyrE*	Orotate phosphoribosyltransferase	-5.83	2.83E-02
MGcds_01825	*manY*	Mannose-specific pts enzyme iic component	-5.65	1.42E-02
MGcds_02329	*pdxB*	Erythronate-4-phosphate dehydrogenase	-5.57	1.19E-02
MGcds_04051	*lamB*	Maltose outer membrane channel/phage lambda receptor protein	-5.46	1.41E-02
MGcds_03420	*malT*	Dna-binding transcriptional activator malt	-5.39	1.71E-02
MGcds_00687	*pgm*	Phosphoglucomutase	-5.31	9.83E-03
MGcds_04006	*thiC*	Phosphomethylpyrimidine synthase	-5.30	5.77E-03
MGcds_01236	*galU*	Utp--glucose-1-phosphate uridylyltransferase	-5.28	1.19E-02
MGcds_03639	*waaP*	Lipopolysaccharide core heptose (i) kinase	-5.15	1.19E-02
MGcds_04005	*thiE*	Thiamine phosphate synthase	-4.90	1.55E-02
MGcds_03640	*waaG*	Lipopolysaccharide glucosyltransferase i	-4.84	1.45E-02
MGcds_00029	*carA*	Carbamoyl phosphate synthetase subunit alpha	-4.70	2.20E-02
MGcds_02107	*thiD*	Bifunctional hydroxymethylpyrimidine kinase/phosphomethylpyrimidine kinase	-4.60	3.19E-02
MGcds_03407	*ompR*	Dna-binding transcriptional dual regulator ompr	-4.52	4.06E-02
MGcds_01095	*mltG*	Endolytic murein transglycosylase	-4.22	3.09E-02

6.4.2 Comparison between EO499 and MG1655

One of the aims in this project where to look for differences between the three strain UTI89, EO499 and MG1655 in acetic acid. It was not possible to perform any comparison to UTI89 due to the issue with reads distribution in *mutL* and *recR*, discussed in section 6.3.3. Therefore, I am left with EO499 and MG1655 for a comparison. EO499 showed significant genes under acetic acids stress in both day 1 and day 5, but most of the significant genes appeared on day 5. While MG1655 did not show any significant mutant on day 1, because the no mutant with FDR < 0.05 were detected. Given the fact significant mutants of MG1655 only were shown on day 5. This analysis will be limited to EO499 and MG1655 on day 5. The candidate genes of EO499 on day 1 and day 5 were combined, for the fact significant genes in day 1 were expected to be in day 5, explained earlier section 6.4.1.

As a comparison between the two strains, table 30 showed 19 overlap candidate genes that caused the strains to be fitter when mutated. As an example of these genes discussed earlier section 6.4.1, *fabFR* involve in the membrane lipid synthesis, *ackA* involve in acetate production and acetate utilization. While the overlap among the two strains of the candidate genes causing the strains to be less fit when mutated under acetic acid, table 31. Looking to the $\log_2$ fold change of EO499, most of the overlap candidate genes did not show such drastic changes. The overlap between MG1655 and EO499 data sets were very low, probably this is due to the low reproducibility of MG1655 data on day 5. Or it could be the effect of acetic acid at low pH is different on the two strains. For a better analysis, further gene set enrichment analysis (GSEA) are required to understand the pathway in which these genes are involved in.

Table 29. Overlap of gene list in TraDIS EO499 and TraDIS MG1655.

The overlap is for genes showing enrichment in fitness when Tn is inserted in the gene under acetic acid stress.

Overlap gene list between EO499 and MG1655, in which Tn inserts caused increase of fitness under acetic acid stress				MG1655		EO499	
GFF gene name EO499	GFF gene name MG1655	Gene name	Predicted function	log_2FC	FDR	log_2FC	FDR
ST131_cds_02764	MGcds_01093	*fabF*	Beta-ketoacyl-[acyl carrier protein] synthase ii	5.48	6.55E-05	1.25	4.21E-03
ST131_cds_00508	MGcds_03180	*greA*	Transcription elongation factor grea	4.31	1.51E-03	2.39	2.34E-15
ST131_cds_00519	MGcds_03169	*rimP*	Ribosome maturation factor rimp	3.90	2.97E-02	8.60	7.04E-07
ST131_cds_03347	MGcds_00589	*fepC_3*	Ferric enterobactin abc transporter atp binding subunit	3.80	9.14E-03	1.79	4.87E-02
ST131_cds_03869	MGcds_00047	group_1193	Hypothetical protein	3.64	1.00E-02	2.23	1.15E-03
ST131_cds_03862	MGcds_00056	*rapA*	Rna polymerase-binding atpase and rnap recycling factor	3.29	1.00E-02	2.06	2.30E-24
ST131_cds_01394	MGcds_02305	*ackA*	Acetate kinase	3.20	2.05E-02	7.42	7.01E-64
ST131_cds_03868	MGcds_00048	*rsmA*	16s rrna m(6)2a1518,m(6)2a1519 dimethyltransferase	3.03	1.77E-02	2.55	1.79E-02
ST131_cds_00069	MGcds_03622	*envC*	Murein hydrolase activator envc	2.86	1.19E-02	1.05	6.43E-05
ST131_cds_04457	MGcds_03973	*fabR*	Dna-binding transcriptional repressor fabr	2.77	9.83E-03	3.66	1.30E-47
ST131_cds_03696	MGcds_00209	*yafS*	Putative s-adenosyl-l-methionine-dependent methyltransferase	2.74	1.10E-02	1.78	4.35E-04
ST131_cds_01676	MGcds_02010	group_2701	Hypothetical protein	2.72	1.71E-02	2.14	1.35E-05
ST131_cds_00523	MGcds_03165	*truB*	Trna pseudouridine(55) synthase	2.70	2.21E-02	0.95	4.26E-03
ST131_cds_02023	MGcds_01830	*cspC*	Stress protein, member of the cspa family	2.63	2.99E-02	5.66	9.17E-96
ST131_cds_00514	MGcds_03174	*secG*	Sec translocon subunit secg	2.50	5.56E-02	2.96	6.37E-03
ST131_cds_00431	MGcds_03258	*dusB*	Trna-dihydrouridine synthase b	2.42	2.32E-02	1.85	1.63E-16
ST131_cds_02769	MGcds_01088	*plsX*	Putative phosphate acyltransferase	2.32	3.28E-02	1.14	3.02E-02
ST131_cds_03953	MGcds_04396	*rsmC*	16s rrna m(2)g1207 methyltransferase	2.28	3.30E-02	0.65	4.74E-02
ST131_cds_04688	MGcds_03762	*rbsK_2*	Ribokinase	2.27	3.28E-02	1.89	4.47E-27

Table 30. Overlap of gene list in EO499 and MG1655.

The overlap is for genes showing reduction in fitness when Tn is inserted in the gene under acetic acid stress.

Overlap gene list between EO499 and MG1655, in which Tn inserts caused loss of fitness under acetic acid stress				MG1655		EO499	
GFF gene name EO499	GFF gene name MG1655	Gene name	Predicted function	log_2FC	FDR	log_2FC	FDR
ST131_cds_03515	MGcds_00400	*phoR*	Sensory histidine kinase phor	-6.94	5.07E-04	-1.47	2.40E-03
ST131_cds_03949	MGcds_04400	*prfC*	Peptide chain release factor rf3	-6.23	8.45E-04	-0.91	6.82E-03
ST131_cds_03244	MGcds_00687	*pgm*	Phosphoglucomutase	-5.31	9.83E-03	-7.04	3.79E-03
ST131_cds_04374	MGcds_04051	*lamB*	Maltose outer membrane channel/phage lambda receptor protein	-5.46	1.41E-02	-0.99	2.78E-05
ST131_cds_04699	MGcds_03751	*mnmG*	5-carboxymethylaminomethyluridine-trna synthase subunit mnmg	-2.43	2.84E-02	-1.84	3.24E-05

I have found a number of candidate genes were showing inverse relation under acetic acid in MG1655 and EO499, table 32. In this case, these genes when mutated in MG1655 caused reduction in fitness and in EO499 caused increase in fitness. Also, I have found the reverse occur for another number of genes, table 33. Where mutant in these genes caused MG1655 to be fitter whereas in EO499 these mutants where less fit in the presence of acetic acids. At this point I am not sure if these results are reliable for MG1655 due to the low reproducibility data in day 5. And if these genes do differ in behavior under acetic acid stress these will require a further confirmation by experiments.

Considering I have done this kind of analysis is not satisfactory because it doesn't take all the significant genes into the account, but it only takes individual genes. It would be significantly better to investigate the pathways analysis of the significant genes. In the coming section I have attempted to do gene ontology analysis.

Table 31. Candidate genes in both MG1655 and EO499 with inverse relation under acetic acid.

In MG1655 mutation in these genes leading to fitness defect and in EO499 caused increase in fitness.

Candidate genes in which Tn inserts caused oppisite realtion in fitness under acetic acid stress				MG1655		EO499	
GFF gene name	GFF gene name	Gene name	Predicted function	log2FC	FDR	log2FC	FDR
ST131_cds_04707	MGcds_03743	*atpG*	Atp synthase f1 complex subunit gamma	-6.08	0.0506249	6.18	0.0279716
ST131_cds_04700	MGcds_03750	*rsmG*	16s rrna m(7)g527 methyltransferase	-10.91	6.703E-07	1.22	5.72E-05
ST131_cds_04425	MGcds_04006	*thiC*	Phosphomethylpyrimidine synthase	-5.30	0.0057664	6.11	0.0020597

Table 32. Candidate genes in both MG1655 and EO499 with inverse relation under acetic acid.

In MG1655 mutation in these genes leading to increase in fitness and in EO499 caused in defect in fitness.

Candidate genes in both MG1655 and EO499 with opposite relation under acetic acid.				MG1655		EO499	
GFF gene name	GFF gene name	Gene name	Predicted function	log2FC	FDR	log2FC	FDR
ST131_cds_03858	MGcds_00060	*araB*	Ribulokinase	2.21	0.040553315	-0.963	0.00082264
ST131_cds_02851	MGcds_01000	*cbpM*	Chaperone modulator cbpm	2.59	0.032821194	-2.838	0.0007212
ST131_cds_00838	MGcds_02915	*gcvH*	Glycine cleavage system h protein	2.69	0.033134973	-8.487	2.327E-06
ST131_cds_03284	MGcds_00656	*gltJ*	Glutamate/aspartate abc transporter membrane subunit gltj	2.36	0.032608145	-1.376	0.00422946
ST131_cds_03770	MGcds_00138	*gluQ*	Glutamyl-q trna(asp) synthetase	2.90	0.013719626	-0.871	0.00772359
ST131_cds_01655	MGcds_02032	*gnd*	6-phosphogluconate dehydrogenase, decarboxylating	2.62	0.022510733	-1.93	3.5408E-06
ST131_cds_01572	MGcds_02121	*group_207*	Hypothetical protein	2.45	0.034293892	-1.017	5.3876E-05
ST131_cds_02184	MGcds_01661	*mepH*	Peptidoglycan dd-endopeptidase meph	2.78	0.032608145	-4.12	0.00025594
ST131_cds_03577	MGcds_00321	*yahF_1*	Putative acyl-coa synthetase yahf	2.48	0.019109045	-0.734	0.00752605
ST131_cds_01999	MGcds_01854	*yebF*	Secreted protein yebf	3.08	0.034654447	-2.896	0.00422946
ST131_cds_00127	MGcds_03560	*yiaG*	Putative dna-binding transcriptional regulator yiag	2.48	0.04856312	-1.124	0.00874647
ST131_cds_04539	MGcds_03896	*yiiD*	Putative acetyltransferase yiid	2.14	0.053025093	-0.671	0.00322015

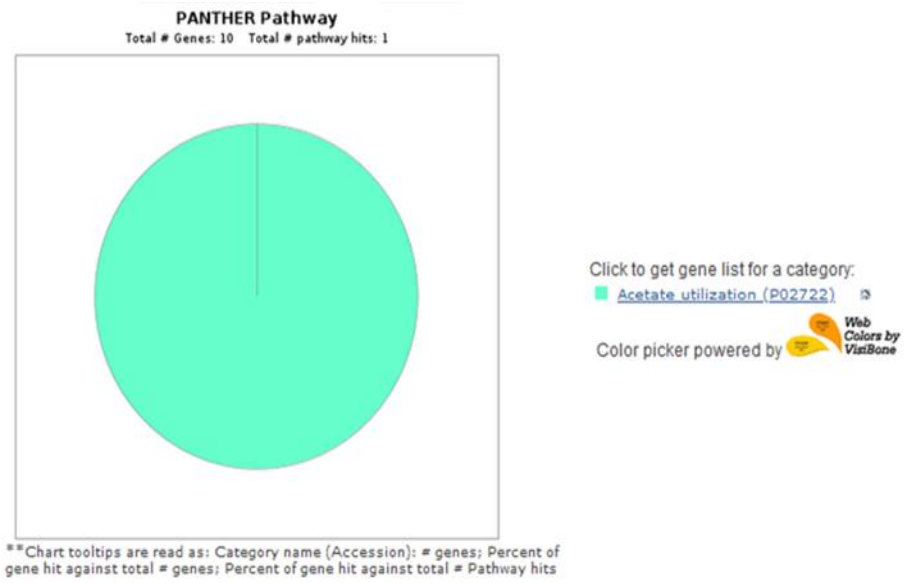
6.5 Gene ontology pathway

In order to understand the genes list obtained by TraDIS of EO499 and identify the pathways in which these genes are likely to be involved in. Or which pathway is overrepresented or under-represented in TraDIS data. I have tried to do the analysis using several different tools for gene ontology such as ShinyGO (Ge et al., 2020), DAVID (Dennis et al., 2003), and PANTHER (Mi et al., 2016). Attempts to use these tools didn't address the significant pathways involve under acetic acid stress, for one reason or another. For example, using PANTHER (http://pantherdb.org/), in which the genes list was uploaded in Panther web browser using the functional classification of PANTHER pathway, figure 66. PANTHER pathway graphs were generated for EO499 on both day 1 and day 5 for all the genes that were enriched and depleted. On day 1, in figure 66A the aetate utilization pathway was overrepresented under acetic acid stress. While underrepresented pathways were Arginine biosynthesis, purine biosynthesis, pyrimidine ribonucleotide biosynthesis and TCA cycle, figure 66B. The number of pathway hits were small these two cases, might be due to short significant gene list provided. While in day 5, in figure 66C shows the overrepresented pathways under acetic acid stress. Two of the hit pathways were Parkisnson disease and Wnt signaling pathway in which they are not involve in bacterial metabolic pathway, so they were considered as miss labelling. Also, in figure 66D, considering the long list of genes were included, probably will be lots of false positive pathways detected. The second issue was found using PANTHER, the number of genes were classified for each individual pathway was around one or two genes. Therefore, I couldn't classify these

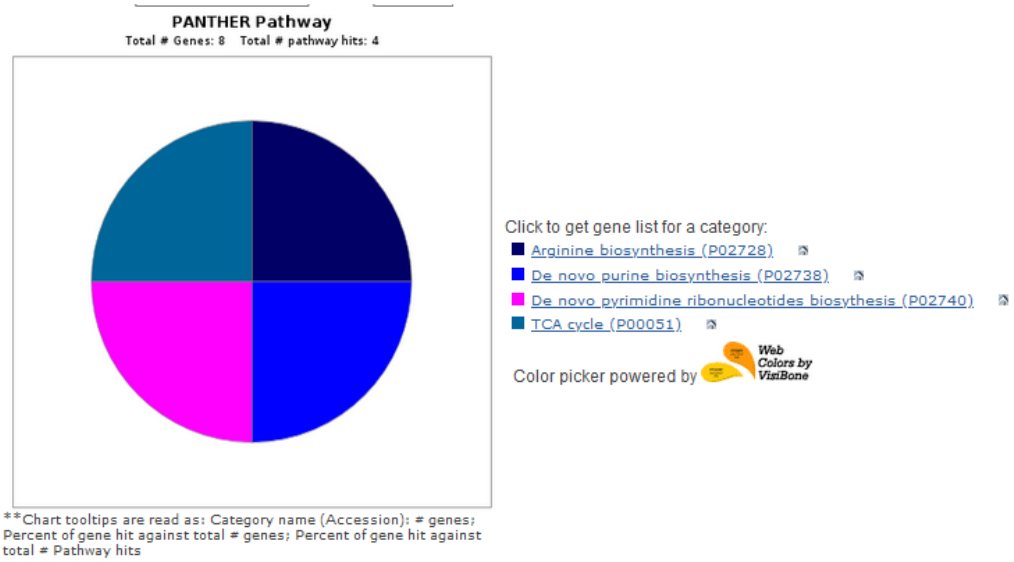
pathways as overrepresented or underrepresented under acetic acid stress, because there is not very large number of genes are involved in.

Therefore, gene ontology for MG1655 were not investigated further. To this point the analysis were not complete and further sophisticated should be carried out using gene set enrichment analysis (Bushell et al., 2021).

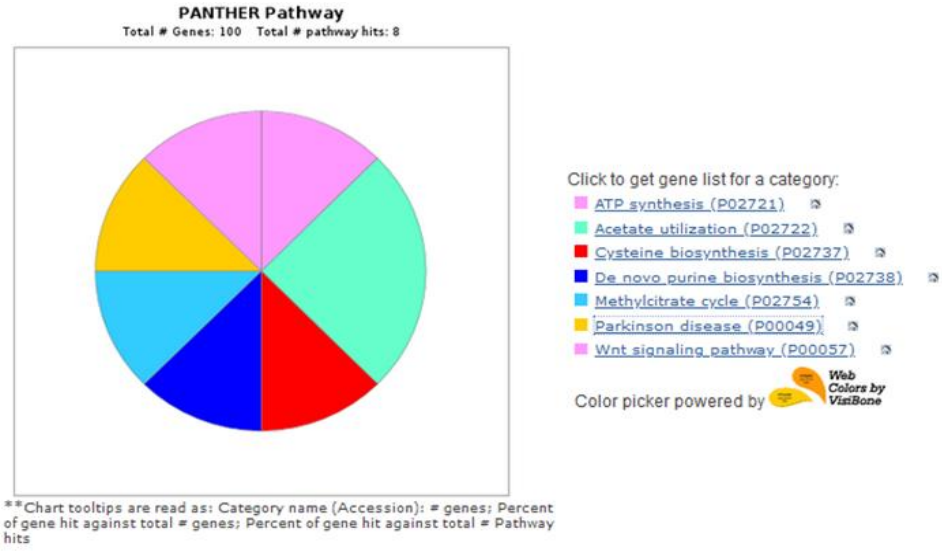
A- Upregulated pathway in EO499 on day 1, under acetic stress conducted by PANTHER pathway.



B- Downregulated pathway in EO499 on day 1, under acetic stress conducted by PANTHER pathway.



C- Upregulated pathway in EO499 on day 5, under acetic stress conducted by PANTHER pathway.



D- Downregulated pathway in EO499 on day 5, under acetic stress conducted by PANTHER pathway.

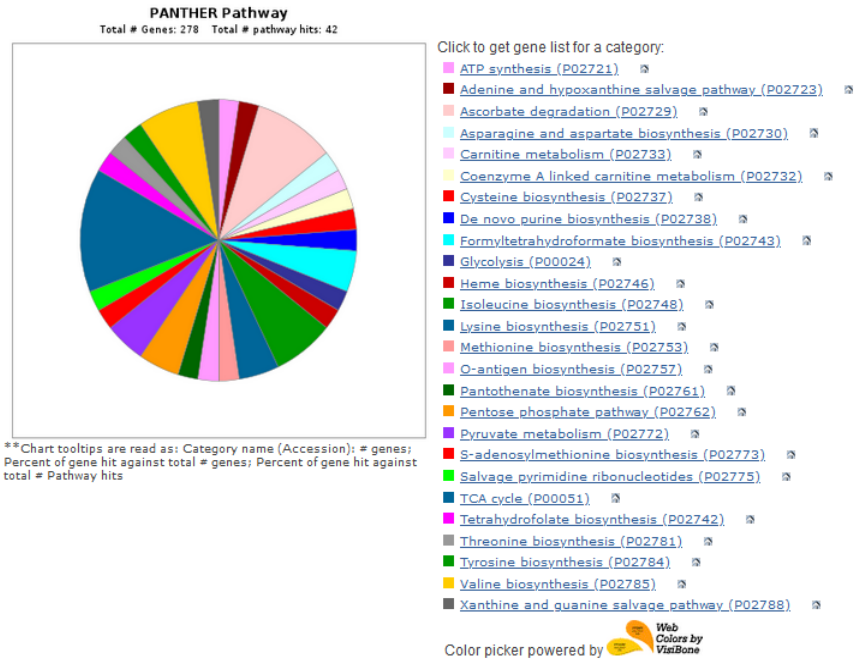


Figure 66. The pathways analysis performed by PANTHER.

Upregulated and downregulated metabolic pathways showed in EO499 for both day 1 and day 5. The charts were labeled accordingly.

6.6 Comparisons between two EO499 libraries

The final aim of the project is to evaluate previous TraDIS data in EO499 generated by Dr. Francesca (Bushell, 2019). This is done to determine the reproducibility of TraDIS (using the same

Initial library, performing the same experiment) when it is done by two different people. As a comparison between the two EO499 TraDIS libraries, one constructed in this project and second constructed by Dr. Francesca, we both have used the same initial sequencing library of EO499, as explained in chapter 3, section 3.2. Note that EO499 TraDIS in this study were done for two time points, day 1 and day 5 and the previous TraDIS library was done for 24 hrs only. A summary of the number of significant genes used in this analysis for both libraries, shown in table 34. The significant candidate genes in EO499 made in this study were identified earlier in section 6.4.1. While in the second EO499 TraDIS library created by Francesca, the significant genes were selected using a cutoff value of FDR < 0.05. The number of the overlap significant genes of both libraries were identify in figure 67. The overlap of the significant genes between the two TraDIS library were more common to day 5 rather than day 1. For some reason, EO499 in this study took longer to show the effect of acetic acid. But the two sets of data correspond moderately well. But the analysis would be limited to a discussion of effects of some examples of the single genes effect.

Both genes candidate genes list on day 1 and day 5 were combined and then compared to Dr. Francesca TraDIS data. Table 35, showed the overlap of significant genes that between the two studies that showed increase in fitness under acetic acid when mutated. The table contains the log FC and FDR for both EO499 TraDIS libraries. As an example, both libraries were enriched in acetate production and utilization (*pta* and *ackA*), and in the fatty acid biosynthesis (*fabF* and *fabR*). Table 36, showed the overlap of significant genes between the two studies that showed reduction in fitness under acetic acid when mutated. In this table there are common few genes

from the TCA cycle (*gltA*, *sdhC* and *mdh*) and *phoR* which is part of the PhoRB is the two-component system. The important of these two pathways in relation to acetic acid at low pH were explained earlier in section 6.4.1. For better interpretation of the obtained results, it would be useful to perform gene set enrichment analysis. To determine the extent of this overlap, further tests are required for these candidate genes.

Table 33. The common significant genes in EO499 done by Francesca and in this study at pH 5.5 with acetic acid.

A. The total significant genes number in both libraries. B. The common genes lists in in both EO499 libraries.

		Genes increased fitness in the presence of acetic acid	Genes reduced fitness in the presence of acetic acid
This study TraDIS EO499	Day 1	10	8
	Day 5	100	299
Francesca TraDIS EO499	Day 1	139	120

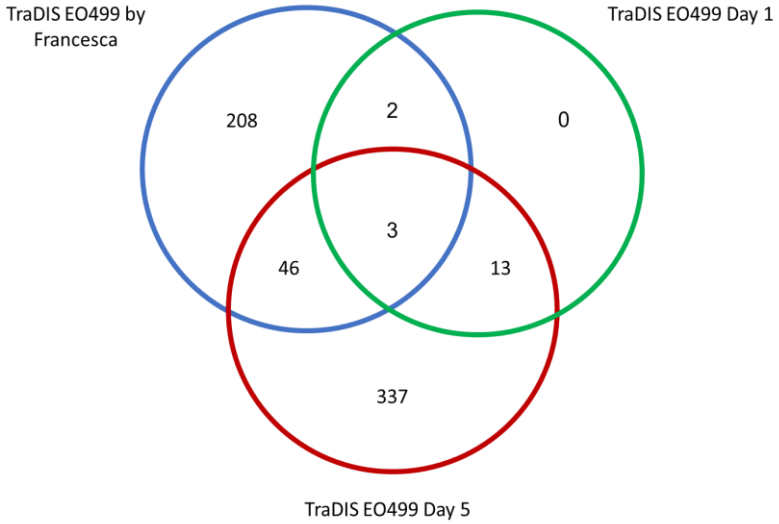


Figure 67. A comparison of two TraDIS libraries in EO499 at pH 5.5 with acetic acid.

This study TraDIS data compared to previous EO499.

Table 34. Common list of genes in which Tn inserts caused increase of fitness under acetic acid stress in two TraDIS studies.

Enriched mutants under acetic acid					
		EO499 in this study		EO499 in Francesca study (Bushell, 2019)	
Gene	**Annotation**	**LogFC**	**FDR**	**LogFC**	**FDR**
rpmF	50S ribosomal subunit protein L32	7.63	0.000224911	-3.34	2.60E-09
rpmG	50S ribosomal subunit protein L33	7.34	0.004229459	-1.39	0.031768503
dnaJ	chaperone protein DnaJ	6.5	1.20E-158	-1.58	1.46E-15
pta	phosphate acetyltransferase	6.47	3.46E-28	-1.49	1.48E-06
rnb	RNase II	3.56	1.06E-09	-0.62	0.001881012
fis	DNA-binding transcriptional dual regulator Fis	3.32	6.57E-09	-1.34	2.28E-05
rcsB	DNA-binding transcriptional activator RcsB	3.25	2.20E-78	-0.45	0.039638232
arcA	DNA-binding transcriptional dual regulator ArcA	3.09	1.70E-06	-0.87	0.057679413
sapB	putrescine ABC exporter membrane subunit SapB	2.67	0.03288073	-0.86	0.040403119
rsmA	16S rRNA m(6)2A1518,m(6)2A1519 dimethyltransferase	2.55	0.017907411	-1.01	0.000140588
rph	truncated RNase PH	2.36	9.30E-23	-0.57	0.006963342
iscX	accessory iron-sulfur cluster assembly protein IscX	2.08	0.004242371	-0.99	0.012563162
rapA	RNA polymerase-binding ATPase and RNAP recycling factor	2.06	2.30E-24	-0.39	0.025988469
rbsR	DNA-binding transcriptional dual regulator RbsR	1.96	1.25E-33	-0.48	8.41E-05
cyaY	frataxin CyaY	1.9	1.97E-06	-0.72	8.26E-05
ackA	Acetate kinase	1.89	1.82E-05	-2.56	9.10E-28
hupA	DNA-binding protein HU-alpha	1.81	9.72E-11	-0.49	0.024993373
cspC	Stress protein, member of the cspa family	1.56	2.13E-05	-0.55	0.001209891
rluB_2	23S rRNA pseudouridine(2605) synthase	1.51	0.000684734	-0.72	0.004339223
thiI	tRNA uridine 4-sulfurtransferase	1.5	0.011403569	-0.72	6.57E-05
speB	agmatinase	1.28	0.001315749	-0.57	0.002428044
dnaJ	Chaperone protein dnaj	1.27	5.84E-06	-1.58	1.46E-15
fabF	beta-ketoacyl-[acyl carrier protein] synthase II	1.25	0.004205338	-0.86	1.12E-08
fabR	Dna-binding transcriptional repressor fabr	1.15	0.000354362	-0.59	0.000145087
rhlB	ATP-dependent RNA helicase RhlB	1.15	5.03E-07	-0.55	0.00224717

Table 35. Common list of genes in which Tn inserts caused decrease in fitness under acetic acid stress in EO499 on this study in compared to Francesca TraDIS study.

Depleted mutants under acetic acid					
		EO499 in this study		EO499 in Francesca study (Bushell, 2019)	
Gene	**Annotation**	**LogFC**	**FDR**	**Log_2 FC**	**FDR**
prc	Tail-specific protease	-9.98	1.97033E-17	0.56	0.016344891
sucD_1	Succinyl-coa synthetase subunit alpha	-8.95	2.48088E-10	1.07	8.73828E-06
gltA	Citrate synthase	-6.98	0.000128387	1.06	1.31585E-05
sdhC	Succinate:quinone oxidoreductase, membrane protein sdhc	-6.80	0.003013984	1.14	0.024296705
cysK	O-acetylserine sulfhydrylase a	-6.28	0.007271417	0.77	0.009949248
apaH	Diadenosine tetraphosphatase	-5.65	0.029635492	2.25	4.23384E-08
mdh	Malate dehydrogenase	-4.72	5.99967E-14	0.79	0.000554759
sucC_1	Succinyl-coa synthetase subunit beta	-3.46	0.0049777	1.01	0.023140875
ytfP	Gamma-glutamylamine cyclotransferase family protein ytfp	-3.12	0.012688041	1.53	9.76877E-08
ybeZ	Phoh-like protein	-2.47	5.62136E-05	0.59	0.004510166
gpp	Guanosine-5'-triphosphate,3'-diphosphate phosphatase	-2.33	1.31868E-07	0.74	0.000357039
arcB	Sensory histidine kinase arcb	-2.01	1.17215E-06	0.64	0.043859692
waaL	O-antigen ligase	-1.81	1.73056E-05	0.47	0.015787254
cysU	Sulfate/thiosulfate abc transporter inner membrane subunit cysu	-1.79	0.004398189	0.57	0.046782653
phoR	Sensory histidine kinase phor	-1.69	0.000438205	0.83	3.44572E-06
cyoE	Heme o synthase	-1.60	0.03253817	0.79	0.000574434
aroM	Protein arom	-1.56	0.004347606	0.69	9.38898E-05
hyaD	Putative hydrogenase 1 maturation protease hyad	-1.52	0.010187232	0.55	0.039049968

6.7 Discussion

TraDIS is an approach can be used to identify genes significant for growth under different conditions. Due to limited time in this project the analysis was not completely done. The detailed analysis is still in progress in for the above data. The UTI89 TraDIS showed insertion bias toward 3' end in *mutL* which left the rest of the genome with very low or non-reads in most of the examined conditions. Therefore, I was not able to determine the relative fitness of the genes,

leading to errors when calculating the relative for any given genes. Also our data showed, *recR* insertions bias toward 5' end in some cases. It was notable during passaging UTI89, the optical densities were very low for all the conditions (~ 0.02). Therefore, larger amount of cultures were required for DNA extraction to increase the concentration of DNA required for library prep. The cultures passaging UTI89 were unlike passaging MG1655 and EO499. In principle using larger volume of cultures for DNA extraction should not affect the correlation, but low growth might be important in this case.

One possibility is this insertion bias (high number of insertions) could be because of DNA contamination in TraDIS library. For example, if there is any source of contamination from other TraDIS libraries that would be amplified by PCR and detected by sequencing in our libraries. The source of contamination could be from contaminated reagents (e.g. beads, PCR polymerase, sterile water or the primers) used to prepare the libraries for sequencing. In here I have ruled out the possibility of genomic DNA contamination because these reagents were used to prepare EO499, MG1655 TraDIS libraries and other samples in the LAB, and no contamination were detected.

In order to confirm the insertions in *mutL* and *recR*, TraDIS libraries under these examined conditions should be diluted and plated on LBA. Then random single colonies were selected for PCR confirmation using flanking primers of *mutL* or *recR*. If there is a transposon insertion in *mutL* or *recR*, this will show a larger fragment size in compare to the non-mutant. This will tell us wither the insertions is there or not. But this were not further investigated.

Another possibility it could be a true result for *mutL*, because it is involved in a major pathway of DNA mismatch repair (MutS and MutL), which repair of DNA replication in both prokaryotes and eukaryotes. This pathway associated with mismatches, insertions and deletions (Harfe and Jinks-Robertson, 2000). A study found MutS and MutL can clear a mismatch, by guiding the endonuclease to the target sites, no matter how far it is (Polosina et al., 2009).

RecR involve in DNA repair protein, in which function in RecA replication recovery (stabilized the RecA filament DNA complex). Also, RecA recombinase protein plays a major role in the DNA repair system of the homologous recombination, to repair the double-stranded DNA breaks. Based on studies, it is known there is overlap between the two systems (mismatch repair and homologous recombination). It is suggested that the recombination can be inhibited by the mismatch repair system. The DNA mismatch protein *MutL* with DNA binding and ATPase activity, found to enhance the inhibition of RecA of the homologous recombination pathway (Zhang et al., 2012b). There were no studies shown the relationship between *mutL* and *recR*, but since *mutL* found to inhibit *recA*, and *recR* stabilize RecA filament DNA complex, they might be linked or maybe not. At this point, it is unclear the link between the bias transposon insertions toward *mutL* and *recR*. Therefore, no further analysis were carried out for UTI89 with EdgeR due to low correlation between the examined replicates, that will result in low statistical significance.

However,Dr.Swaine Chen has constructed UTI89 TraDIS library under several growth conditions, including biofilm formation. His sequencing data is accessible to us, it would be interesting for further study to examine and compare *mutL* and *recA* insertion patterns in his

sequencing data. Also, it would be helpful to borrowDr.Swaine TraDIS library and perform the same experiment again if there were no insertion bias toward *recA* and *mutl* in the library.

TraDIS of three *E. coli* strains (EO499, UTI89 and MG1655) were used were exposed to acetic acid stress over a period 5 days. Only EO499 and MG1655 were used for further data analysis. EdgeR were used to determine genes that positively and negatively contributed to fitness under acetic acid stress in MG1655 and EO499 based on their significant FDR values. Our analysis showed that passaging the TraDIS libraries over a longer period of time capture a better data view. In order to generate the significant candidate genes from EdgeR outputs a cut-off value should be applied to the data. There are many ways to choose the significant genes list from the EdgeR output, as the one I have done in this study choosing a cutoff value based on FDR < 0.05 or FDR < 0.01 (in one case, EO499 mutants showing reduction in fitness on day 5). Another option is to define significant genes as a $\log_2$ fold change >1 or <-1, with FDR < 0.01 or < 0.05. The last option I can choose is to rank the $\log_2$ fold change values and choose the significant genes list based on the graphs where there is a sudden change in the corner curve of fitness with consider to the FDR < 0.05 or FDR < 0.01, figure 63B.

Looking to EO499 candidate genes lists, it suggests that there is come important pathways effected negatively or positively under acetic acid stress such as the TCA cycle, electron transport chain, oxidative phosphorylation, fatty acid synthesis, the acid resistance pathway, acetate production and utilization and some membrane proteins. In MG1655, some candidate genes were involved in acetate production and utilization, membrane protein and oxidative phosphorylation. The discussion of these pathways and the genes are involved in mentioned earlier in the relative

section 6.4.1. A comparison were conducted for both strains based on two genes lists, genes which showed reduction in fitness when mutated or genes which showed enhance in fitness when mutated. Our results show that, the overlap between the two strains was very low suggesting that may be different pathways were affected by acetic acid. Or it could be the low reproducibility of MG1655 TraDIS data on day 5 is causing the low overlap between the two strains. I have expected to observe different behaviour of these two strains, as both of them sits in a different phylogroup which may results in a different biological behaviour. There haven't been many studies comparing different strains of the same species, but recently a study on three strains of *Pseudomonas aeruginosa* were investigated for genes contributed for survival under biofilm growth conditions. The study showed that the overlap between the three strains were quite distinct in both cases when were genes depleted or enriched (Schinner et al., 2020). Another recent study, shows that genes identified against three antibiotics (β-lactams: ampicillin, cefradine and cefoxitin) in BW25113 overlap by less than half to genes identified in TraDIS data on *E. coli* UTI EC958, due to different cellular effect of these antibiotics on these strains (Phan et al., 2020; Liu et al., 2010). These studies may support our findings in here, which strains from the same species do differ in their cellular response to stress.

The second differences I have found between comparing EO499 and MG1655 TraDIS libraries. The significant candidate genes in acetic acid, there were several genes that showed an inverse of relation to their $\log_2$ fold change between the two libraries. This remain hypothetical until I do the validation experiments by making the knockouts of these mutants to come to conclusion. There are two possibilities to explain this, it is either this result is not valid be due to

low reproducibility of MG1655 data quality on day 5 or it could be real results. Because the same previously cited study showed that the overlapped of significant candidate genes between BW25113 and *E. coli* UTI EC958, showed some differences in their antibiotic's susceptibility when they mutated for results validation. Where in BW25113 the antibiotic susceptibility of some mutants increased to cefoxitin and did not show any changes for EC958 strain (Phan et al., 2020). In order to obtain better data set for MG1655 at pH 5.5 on day 5, I can sequence the third biological replicate which has been stored as an additional sample. Because only two samples were sequenced for each time point and each condition. Then I can process the analysis and the pairwise comparison between the two strains.

Secondly, It was shown that lab strain *E. coli* K-12 MG1655 is not always the best model in genomic view. Because MG1655 was isolated in 1922, since then it was going through repeated subculture, storage and treatments (UV for phage lambda and acridine orange for F plasmid) to alter the genome sequence. Several other genetic lesions and accumulated mutations have been identified by sequencing K-12 while it is life in the lab. Also, it was recommended to study K-12 alongside other *E. coli* strains from it is natural habitats before trying to generalize the results to the entire species (Hobman et al., 2007).

Whereas, the overlap of the gene lists between two libraries produce independently by two different people (Francesca TraDIS library and this study TraDIS library) in EO499 TraDIS under acetic acid stress demonstrate good reproducibility. Note that Francesca TraDIS library was analyzed by ESSENTIALS pipeline, while this study used EdegR to obtain the relative fitness and FDR. Different pipelines used for the analysis may introduce variation due to different statistical

method conducted, but this has not been examined yet. The second difference is that both set of TraDIS libraries were conducted at different time point. In this study EO499 library was conducted at two-time points day 1 and day 5, while Francesca TraDIS library was conducted at one-time point day 1. It would be more reasonable in future to process Francesca sequencing data through EdgeR pipeline to have a uniform data set. Inferring from these findings in section 6.6, I speculate that both data sets correspond moderately well in spite of the minor differences in both experiments. Considering, the genes contributing to each pathway is very small, it is a concern that single gene analysis wouldn't be an accurate representation of the pathways are involved in, a better method would be gene set enrichment analysis but it hasn't been time to do this. Experimental validation options for both TraDIS libraries are coming later on in this section.

At this stage of analysis, it is early to come to conclusion what molecular pathways are being affected under acetic acid stress. Because investigating the candidate significant gene lists were based on individual gene which is not the best option to come to conclusion. Furthermore, it is time consuming to investigate the molecular pathway of every single gene separately. It would be interesting to perform annotation enrichment analysis to define the shared functional properties of these gene sets (Glass and Girvan, 2014). There are few databases or pipelines do this with such as KEGG (Kyoto Encyclopedia of Genes and Genomes) or Go enrichment analysis. This will allow us to determine the significant pathways involve under acetic acid stress, the number of genes falls under each pathway and comparisons between EO499 and MG1655. Examples of KEGG analysis done on TraDIS data found in these studies (Charbonneau et al., 2017; Bushell et al., 2021). For further validation of TraDIS data under acetic acid stress, gene

inactivation should be conducted for genes less fit under acetic acid stress starting with the significant genes with the larger log fold change. Similar approach used was explained in earlier in chapter 3. For MG1655, gene knockout can be conducted by P1 transduction or using direct gene knockout (Datsenko and Wanner, 2000). But in case of EO499, for the explained reasons in chapter 3.3, I couldn't conduct a gene knockout in this strain. I have two options, either I can confirm TraDIS results based on reproducibility between this study TraDIS data and Francesca TraDIS data, or gene knockout in EO499 can be conducted in EO499 cured from it is mega-plasmid. Additionally, changes in cell morphology can be obtained with these mutants at pH 5.5 with or without acetic acid, besides the cell length using Zeiss microscopy. It was shown in a recent study, some candidate mutants form TraDIS post treated with cefotaxime increased cell length (Phan et al., 2020). It would be beneficial to conduct the cell morphology by microscopy as it was generally proposed that the acetic acid stress affect the proton gradient and may collapse the transmembrane gradient. Moreover, many of the genes found in our analysis encode a protein located in the inner membrane which may leads to effect the bacterial membrane. Additionally, considering the fact that acetate can be a by-product from the cell and it can be consumed by the cell this made the interpretation of what is going on of these tested conditions (pH 5.5 with and without acetic acid) really hard to understand. One way probably can help us to interpretate our results is to measure the internal pH of the cell such as the cytoplasmic pH and the periplasmic pH, with and without acetic acid (Wilks and Slonczewski, 2007). Besides, measuring the metabolise taken up by the cell and excreted by the cells such as the acetate ions (Pinhal et al., 2019). High concentration of acetate is proven to be toxic for the cells and inhibits the growth

rate to 50 %, because it inhibits a step in biosynthesis of methionine which caused an accumulation of the homocysteine. Figure 68, showed the methionine biosynthesis pathway in *E. coli* (Roe et al., 2002). Usually, acetate ions produced in the cell can across the membrane outside the cells through transporters (ActP or SatP) when there is no acetic acid outside the cell (Pinhal et al., 2019). But in this study it is hard for the acetate to cross outside the membrane against gradient because the concentration gradient will push it in, particularly at low pH with the presence of acetic acid. This process involves many aspects, the production of acetate as by product, movements of acetate in the charged form by transporter, also the diffusion of the uncharged acetate from outside the cells.

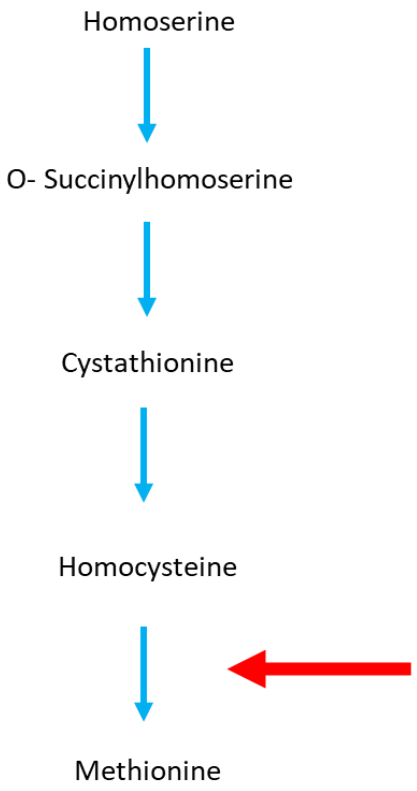


Figure 68. Methionine biosynthetic pathway of *E. coli.*

The red arrow represents acetate inhibition site in Methionine biosynthesis pathway leading to accumulation of homocysteine. Figure adapted from (Roe et al., 2002).

For further prospective study, it would be interesting to build a network of different *E. coli* strains to evaluate the cellular effects of acetic acid on those strains across a period of five days. Due to the high expenses of TraDIS experiment, it would be ideal to use two different *E. coli* strains from each of the six common phylogroups *E. coli* (A, B1, B2, D, E). In order to build a general model in *E. coli* of the pathways affected by acetic acid stress. This will help use to determine the potential effect of acetic when it is used in infected burn wound. As identifying the significant mutants causing reduction in fitness would reveal the effect the acetic acid and allow us to

determine the pathways, they are involved in. While identifying the mutant that causing the increase in fitness could lead the strain to acquire acetic acid resistance when it used for treatment.

7 Comparative analysis of TraDIS, RNA-seq and evolution

Declaration:

This chapter is written to find the overlapped important genes in *E. coli* EO499 under acetic acid stress at pH 5.5 among three different approaches: TraDIS, RNA-seq and evolution experiment.The EO499 initial transposon library was constructed by Dr. Keith Turner. The TraDIS experiment was constructed in this project by me, and the data analysis were carried out by me and with the help of Dr. Mathew Milner. The RNA-seq libraries were performed by Dr. Francesca Bushell and Dr. Thippesh Sannasiddappa. The RNA-seq libraries were sequenced in Liverpool University and the data analysis was conducted by Dr. John Herbert. The evolution experiment was done by Dr. Francesca Bushell and the sequencing was done in MicrobesNG facility in University of Birmingham. The evolution data analysis was conducted by Dr. Mathew Milner. The data comparison in this chapter among the three methods were done by me.

7.1 Overview

This chapter describes a comparison between the TraDIS candidate genes with those previously identified by RNA-seq or in long-term evolution experiments done at pH 5.5 with and without acetic acid stress in EO499. In this comparative analysis, only EO499 TraDIS was used to compare with RNA-seq and evolution experiments, because the RNA-seq and evolution experiment data were generated using EO499 only.

The RNA-seq data were generated to estimate gene expression (transcription levels), whereas TraDIS measures the bacterial fitness under pH 5.5 with and without acetic acid. The evolution experiment identifies the genetic targets and the genetic variation across the whole genome after prolonged growth under the selected environment. Here, the goal is to ask whether TraDIS data correlates at gene level to RNA-seq data and evolution experiment data. If so, candidate genes could be used for a follow up study to link them to the pathways they are involved in. In the cross-comparison between TraDIS and RNA-seq, the assumption is if a gene is required under a particular condition, this gene's expression will likely increase under the given condition. This might not always be true, but generally speaking it makes biological sense if gene X is important under a selection, then more protein would be made from gene X under the selected stress. Additionally, if gene X is important under a selection, gene X knockout will cause the cell to be less fit. It is thus a reasonable hypothesis to suggest, that if a gene is required, an increase of expression would correlate with decrease of fitness in TraDIS. But for the converse situation, it is more challenging to justify because if expression of gene X is deleterious under the examined condition, in theory making a gene knockout would cause the strain to become fitter.

The assumption is, if the gene expression is deleterious then the strain will evolve to down regulate expression or to lose the gene entirely. I won't anticipate this will affect every single gene, but in this chapter any evidence for trends in gene lists has been looked for to see if there is any correlation between the data sets. With more time, more information might be gained by looking at the overlap among pathways that these genes are involved in, rather than in simple gene to gene relationships (Bushell et al., 2021; Bushell, 2019).

In the next sections, the RNA-seq and evolution methods under acetic acid stress done by Dr. Francesca will be explained briefly. The read distribution pattern of the RNA-seq data will be shown. A cross-comparison between the data from the TraDIS, RNA-seq and long-term evolution experiments has been conducted to identify any genes in common under acetic acid stress.

7.2 RNA-seq: method overview

RNA sequencing (RNA-seq) it is a technique that uses next-generation sequencing to study the quantity of input RNA under specific growth conditions at a given moment in time. The method investigates and compare the transcriptome profile [(transcriptome is the total set of transcripts of the mRNA)] for various gene expression by measuring expression in different conditions. The most common transcriptome studied is the mRNA which is translated into proteins expressed by bacteria. RNA-seq measures which genes are turned on or off in a cell to evaluate their level of transcription which may help to identify genes or pathways associated with the condition. In theory, transcriptome can capture the transcripts of genes which have unknown

function when they are turned on in the growth condition. In general, the RNA-seq approach is done as shown in figure 69. This method was used by Dr. Francesca Bushell to generate RNA-seq data under acetic acid stress in EO499. Also, this represents the approached used to library prep the RNA-seq by Liverpool University, where the sequencing was done.

The approach involves several steps, as follows. A- the strain EO499 was grown under the control condition (pH 5.5) and selective condition (pH 5.5 with 4 mM AA). B- Then the total RNAs were purified using RNeasy kit (Qiagen) following the manufacturer's instructions. Normally, the total RNA purification approach isolates the cellular RNA including rRNA, tRNA, ncRNA and the mRNA. Also, the RIN (RNA integrity number) value for all samples were obtained using Prokaryote Total RNA Nanochip. The RIN values were arranged from (1 -10) lowest to highest quality with 10 being the least degraded. C- This is followed by removal unwanted rRNA, as the rRNA is highly abundant in the total RNA sample isolated from bacterial cell, and rRNA removal increases the efficiency of RNA-seq coverage of mRNA. This will lead to an increase in the detection of low abundance transcripts/gene in both the RNA-seq and qPCR. D- RNA fragmentation. Because many current sequencing platforms provide short reads, therefore most of the RNA libraries prep methods include an RNA fragmentation step which allows PCR amplification and sequencing. The long RNA can be shortened by any of three means: physical, enzymatic, or physical. Next, E- RNA is converted to complementary DNA (cDNA), the RNA fragments were converted to double-stranded cDNA by reverse transcription (RT). First, the single RNA fragments were converted into cDNA by reverse transcriptase, then the second strand of the cDNA were made using DNA polymerase. Followed by F- Adenylate 3' Ends, an A-base is added to 3' ends of the blunt ends

fragments to avoid self-ligation while adding the adapters. G- Adapters ligation, in this step a complementary T-base is added to the 3' end of the adapter for ligating the adapter to the cDNA. Ligation of the adapter to the cDNA allows the fragments to attach onto the flow cell by hybridization. Next, H- DNA fragment enrichment - DNA library, the cDNA products were enriched with PCR to obtain sufficient material for sequencing and purified to generate the final library. Note, only fragments with sequencing adapters will be amplified. I- Normalize and pool libraries, each library was generated with specific index adapter at each end (This step is not shown in this workflow), because the sequencing systems required an almost equal index presentation of the libraries in order to separate multiplexed samples after the sequencing has been done. In this step, the libraries were quantified by qPCR to pool equal amounts of each library. J- Paired end sequencing on an illumina 2500 HiSeq platform. In the paired end reads, both ends of the fragments were sequenced. K- Map reads onto the reference genome, before mapping the reads to the annotated genome the reads were filtered by index barcodes and the unwanted and low-quality reads were removed. Finally, L- Computational analysis was used to generate sequencing counts for all genes in the genome and calculate $\log_2$-fold changes, i.e. relative changes in the expression of each gene between the control condition and the selection condition. The P-value and the FDR values were generated as well for each gene (Bushell et al., 2021).

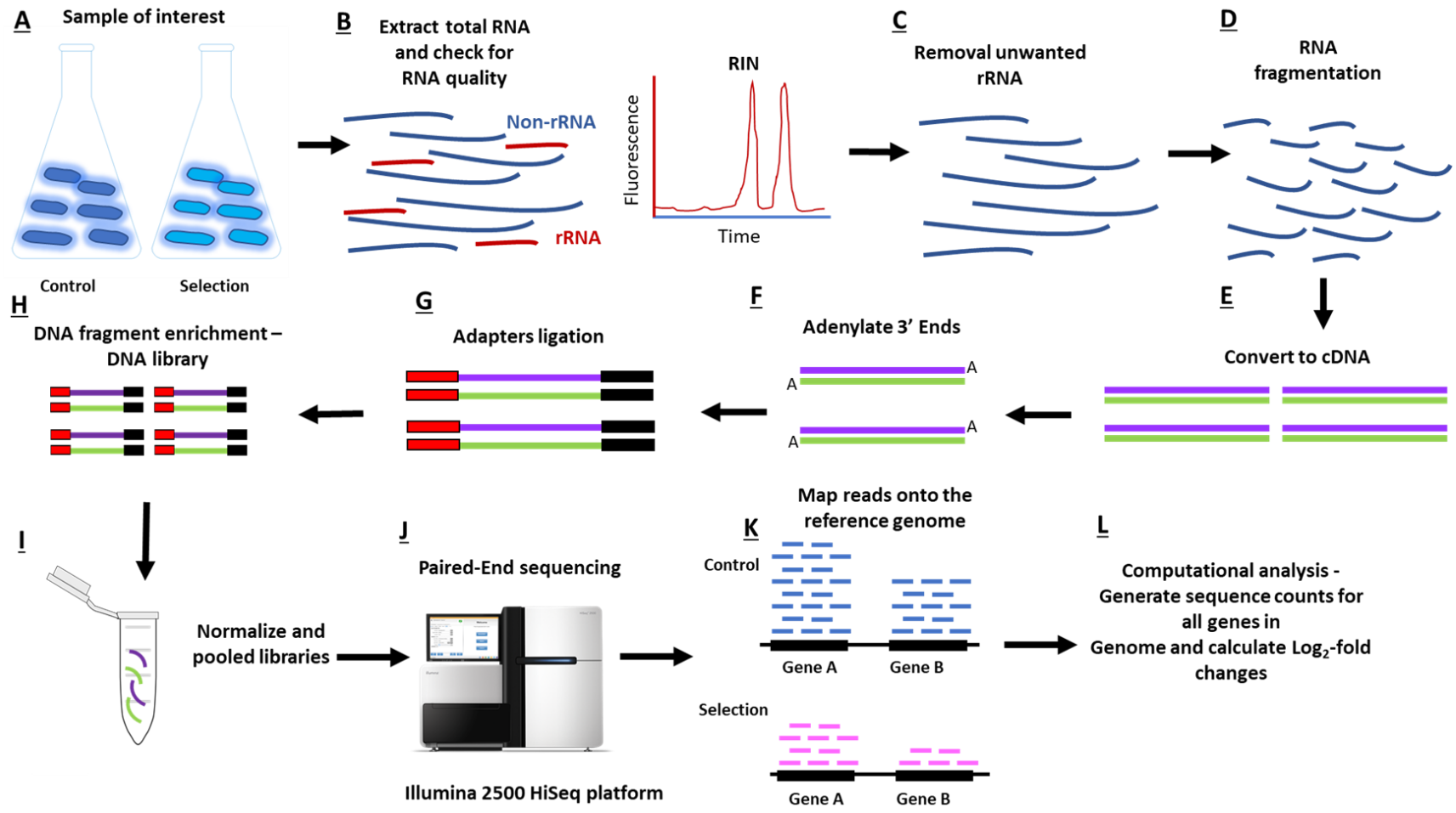


Figure 69. The workflow of RNA extraction, library preparation, and sequencing.

The details were explained in the text.

7.3 Lab-based evolution, and analysis of evolved EO499 populations

The second aim of this chapter is to compare the TraDIS results with a long-term lab-based evolution experiment. *E. coli* EO499 was evolved under pH 5.5 and pH 5.5 with acetic acid, as explained in section 2.10 (Bushell, 2019). These stresses were exactly the same as those used in TraDIS experiment in this study and the RNA-seq experiment done by (Bushell, 2019). This is to whether a EO499 tolerance phenotype toward acid or organic acid could be evolved and to identify whether any of the mutations detected in the evolution experiment overlapped with RNA-seq gene expression or TraDIS mutants (Bushell, 2019). Figure 70 showed a summary of the method used in (Bushell, 2019). Dr. Francesca Bushell did the assay by growing six populations at pH 5.5 and another six populations at pH 5.5 with 4 mM AA, with serial transfer and dilution every day for 11 months. These evolving population were only genome sequenced after 11 months of passaging. In this comparison, populations which were evolved under pH 5.5 with acetic acid only were analysed. The sequencing reads of the evolved populations were analyzed by Dr. Mathew Milner using breseq (v0 .31.1.) computational pipeline. The breseq pipeline determines the full sequence of the newly evolved genome, detecting due to point mutations, chromosomal deletions, short insertions, and other structural mutations relative to the ancestral genome (Deatherage and Barrick, 2014). In breseq, the mapped sequence reads from the evolved populations are expected to be different from the reference genome due to the mutations that occur during the evolution experiment. The breseq output file was in excel format and contains the nucleotide position in the EO499 reference genome, read alignments, the maximum frequency of the base change, amino acid substitution, locus tag and description. The evolution

data used in this analysis were obtained from a supplementary data file (tab non-synonymous mutation) for Dr. Francesca Bushell's thesis, on the University of Birmingham eThesis repository.

Further explanation of the breseq output files can be found in (Deatherage and Barrick, 2014) and in the breseq manual (https://barricklab.org/). The coverage numbers represent the relative proportions of mutation populations that have read counts mapped to the reference genome. And the amino acid substitution showed when there has been amino acid change in a gene, for example in position 264466 in the reference genome, gene *ftsX*, in evolving sample 6 under pH 5.5 with acetic acid, the amino acid substitution shown was G222W (GGG_TGG). This means the mutation caused a change in the 222nd codon of *ftsX*, causing Glycine (G) to change to Tryptophan (W) in the encoded protein. This is defined as non-synonymous mutation (i.e. when a mutation alters the amino acid sequence of the protein). Synonymous substitutions are when a mutation occurs but it doesn't change the amino acid sequence (also known as silent mutation). Mutation of nucleotide bases can also occur in the intergenic region, but the focus of this comparison will be only on changes in genes. This is because non-synonymous mutation can cause structural and functional changes in the mutated protein and can allow us to understand the changes of DNA sequence during the evolution under acetic acid. The mutant's coverage proportion scores were arranged from 0.05 – 1 (5% - 100% of the evolved population). By examining the non-synonymous mutations, the actual genetic variation (deletions, new insertions of mobile elements, and deletions by mobile elements and rearrangements) might be underestimated in this comparison. Due to the limited time in Dr. Francesca Bushell's project this genetic variation was not considered.

For this analysis, mutants which have a coverage value > 0.05 were considered. The other important thing to clarify here is that there are some mutations have a 1.00 (100%) coverage proportion in all the evolved samples. These are most likely to be differences between the strain used in the evolution experiments and the sequenced strain (which provided the reference genome). These mutations were not considered in the analysis, they were assumed to be present at the start of the experiment.

For the comparison between TraDIS candidate genes and evolution experiment mutants at pH 5.5 with acetic acid, non-synonymous mutants with a coverage threshold > 0.05 was selected from evolution experiment to be compared to TraDIS. In total there were 249 mutants found inDr.Francecsa Bushell's evolution experiment.

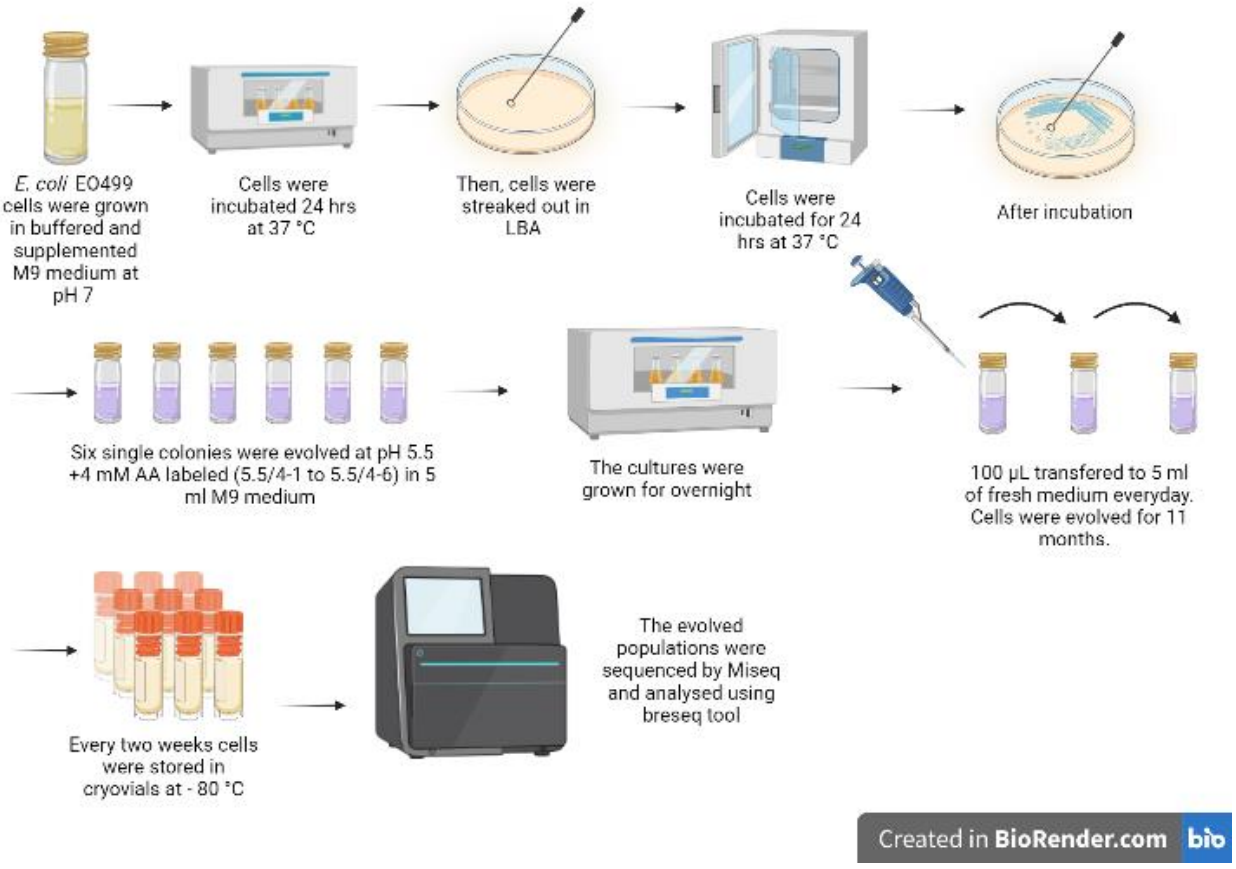


Figure 70. A summary of the procedures used to develop long-term lab-based evolution at pH 5.5 with acetic acid. The arrows express the procedures direction.

The M9 medium was supplemented with glucose and casamino acids as found in materials and methods section 2.1.2. Created with BioRender.com

7.4 Gene expression analysis (RNA-seq)

The RNA-seq analysis process and genome annotation was explained in Materials and Methods in (Bushell et al., 2021). The output analysis file of RNA-seq at pH 5.5 with acetic acid used in this analysis was obtained from the supplementary data in the University of

Birmingham e- Theses Repository (Bushell, 2019). In this output file, a gene measured under two experimental conditions (stress/control), was considered differently expressed if the difference in expression level between the conditions is statically significant. A FDR value of < 0.01 was chosen as the cut-off, regardless of whether the gene was upregulated or downregulated.

The bacterial transcriptional responses in a stress environment are expected to have different dynamics in order to adapt to the environmental changes. The transcriptional response is expected to assist the strain to increase its fitness in the stress environment. The significantly differentially expressed genes were identified by calculating the $\log_2$ fold changes (condition/control) with an adjusted FDR < 0.01. The genes were categorized to upregulated genes, downregulated genes, and change not significant. The up-regulated genes mean there is an increase in gene expression in response to the stressful environment, whereas the down-regulated genes are reverse of this. Figure 71 showed the $\log_2$ fold changes of all the genes plotted against their FDR value. In total there are 5078 genes whose expression was measured in RNA-seq. Around 37.04 % were unchanged genes, 35.60 % were upregulated genes and 27.35 % down regulated genes. The percentage of significantly up-regulated genes was greater than the down-regulated genes.

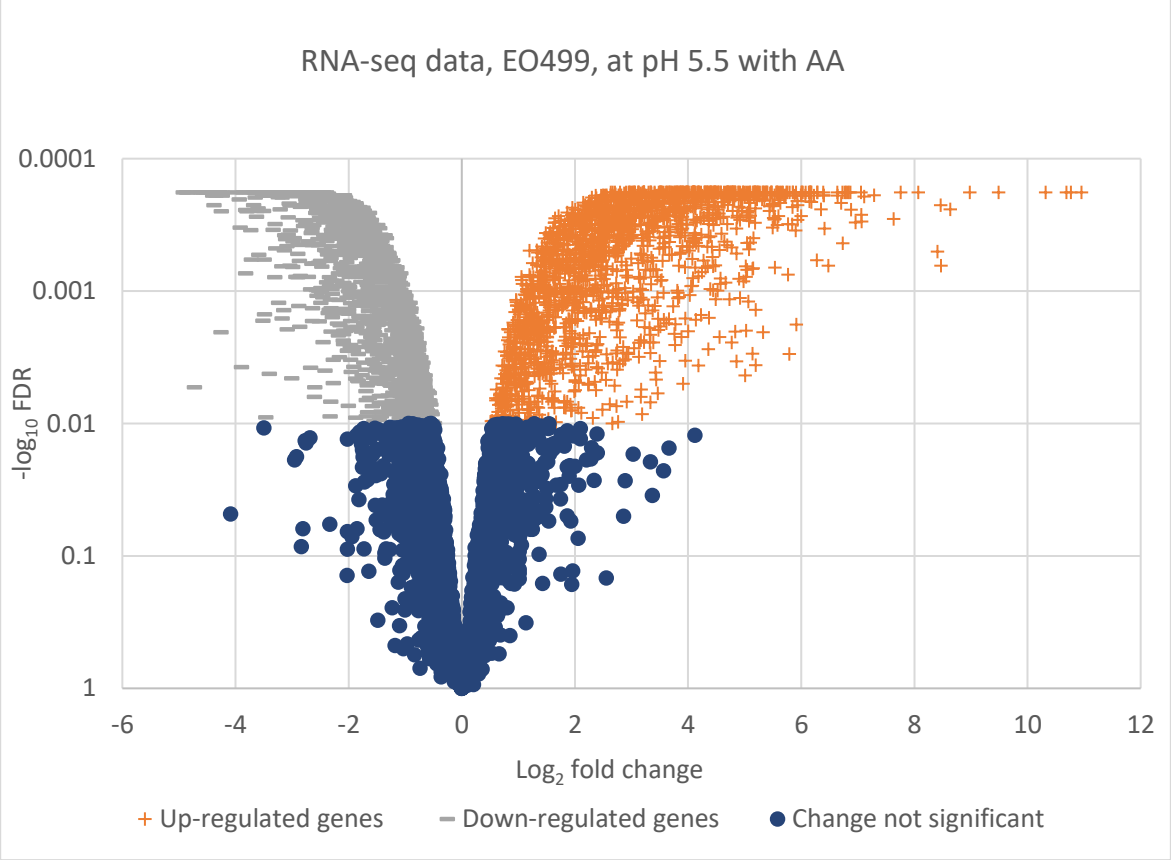


Figure 71. Volcano plot of the distribution of all differentially expressed genes for RNA-seq EO499 at pH 5.5 with acetic acid.

Grey and orange color indicate significantly expressed genes with FDR < 0.01. In total there where 1808 up-regulated genes (orange crosses), 1389 down-regulated genes (grey dashes) and 1881 where the change is not significant (blue bullets) (FDR > 0.01). The $\log_2$ fold change for each gene was calculated as $\log_2$ (Stress/Control).

7.5 A cross comparison between RNA-seq data, TraDIS and long-term evolution under acetic acid stress

The cross comparison between RNA-seq and TraDIS data sets were done at gene level. TraDIS data analysis has been shown earlier in chapter 6, section 6.2 – 6.4.1. The genome annotation for all the three data sets (TraDIS, RNA-seq and long-term evolution experiment) were done using Prokka. For the TraDIS data, EO499 genome was annotated against the MG1655 GenBank file (NCBI Reference Sequence: NC_000913.3) using Prokka. The EO499 genome annotation for RNA-seq and the evolution experiment was done using Prokka without any modification or without specifying the genome reference. This annotation file was run and obtained from Dr. John Herbert. As the EO499 genome annotation for TraDIS was generated slightly differently than EO499 annotation used in RNA-seq and evolution experiment, this will result in minor differences in the output annotation. This means the genes identified as overlapping between two different conditions in this comparison will be a slight underestimate of the true number. Ideally, the two annotation files (GFF3) should be run through Roary pipeline in order to find gene presence and absence in these genome annotations. Then the Roary genome annotation could be used to proceed the comparative analysis. But this option was not considered in this analysis.

For data analysis and comparison between TraDIS, RNA-seq and Evolution, the significant gene lists in TraDIS were used for comparison to RNA-seq data and evolution. Since the comparison was done based on gene level, genes which were found in one list but not the other were neglected from the analysis due to Prokka annotation output differences. The significant

mutants in TraDIS were determined when the FDR value < 0.05 for day 1 and day 5, in both cases when the mutants were enriched or depleted.

7.5.1 A comparison between RNA-seq data and TraDIS data

The simplest way to start this comparative analysis, is to begin with a comparison of RNA-seq and TraDIS data, and then include the evolution experiment in the comparison. EO499 TraDIS were sequenced at two-time points day 1 and day 5, while RNA-seq was done at one time point after four hours of growth. So, RNA-seq was compared to TraDIS day 1 and TraDIS day 5. In this comparison, in TraDIS day 1 and RNA-seq data, no genes were ignored from the TraDIS significant gene list on day 1, because of the different annotation used between the two methods. While in TraDIS day 5, there were 18 genes eliminated from the list when the Tn inserts caused enrichment of fitness under acetic acid. Also, there were 211 genes were eliminated from the list when the Tn inserts caused depletion of fitness under acetic acid.

The two candidate gene lists (when mutants were depleted or enriched) from TraDIS on day 1 and day 5, were compared to the RNA-seq data (when genes were up-regulated or down-regulated). This comparison results in four lists:

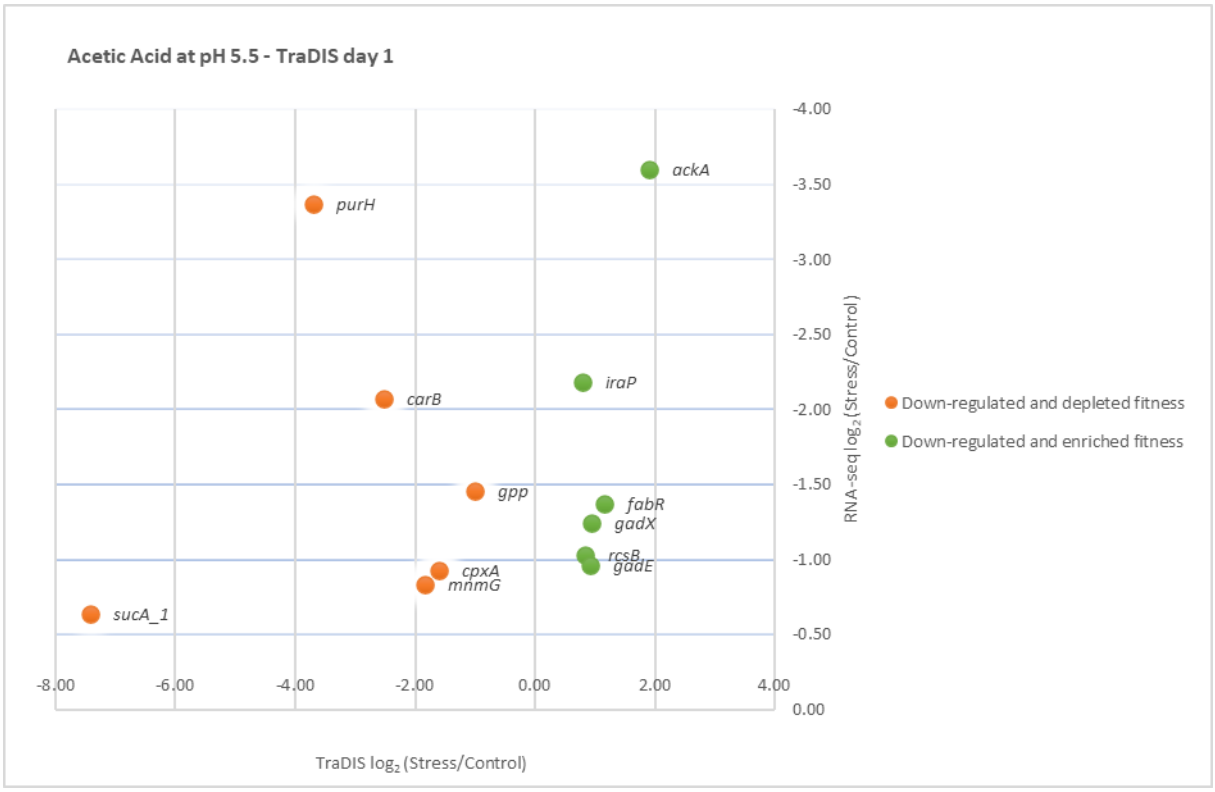
- Down-regulated genes in RNA-seq and depleted mutants in fitness in TraDIS.
- Up-regulated genes in RNA-seq and enriched mutants in fitness in TraDIS.
- Down-regulated genes in RNA-seq and enriched mutants in fitness in TraDIS.

- Up-regulated genes in RNA-seq and depleted mutants in fitness in TraDIS.

In order to determine if there are any overlap between the alteration of fitness and the level of expression under acetic acid stress, both data sets were graphed to indicate the correlation pattern. Figure 72A and 72B, shows the $\log_2$ fold changes of the significant genes for both TraDIS and RNA-seq plotted against each other at pH 5.5 with acetic acid. The $\log_2$ fold changes of the gene expression and the mutant fitness were calculated by the reads of $\log_2$ [Stress (pH 5.5 + 4 mM AA)/ Control (pH 5.5)]. Figure 72A shows the significant overlapped genes between TraDIS day 1 and RNA-seq data. There were no commonly up-regulated genes, and only 12 down-regulated genes were matched. The number of significant candidate genes in TraDIS day 1 were small which resulted in a small overlap with the RNA-seq data. Because of the low contrast in this case the data interpretation is difficult. As shown earlier in chapter 6, section 6.4.1, TraDIS day 5 is just a subset of TraDIS day 1, but there are few exceptions. For the simplicity of analysis, further analysis will be only focused TraDIS day 5, because many mutants change at day 5 but not at day 1. Figure 72B shows that approximately 67% of genes identified in TraDIS at day 5 overlapped with significant genes in RNA-seq under acetic acid, pH 5.5. Whether TraDIS mutants were depleted or enriched there where around ~ 85 % down-regulated genes and ~ 48.5 % up-regulated genes shared with RNA-seq. In spite of the large percentage of significant gene similarity between the two data sets, the correlation coefficients (R^2) were very small for each of the four comparison categories as shown in the graph. The correlation coefficient was less than 0.20 in all cases and can't be considered statistically significant. (Note, the R^2 values are not presented in figure 4B). The only cross comparison that slightly fitted with the assumption is that

the log_2 fold change of the down-regulated genes in RNA-seq correlated better with the mutants that were enriched in TraDIS with R^2= 0.2. This considered as a weak positive correlation.

A)



B)

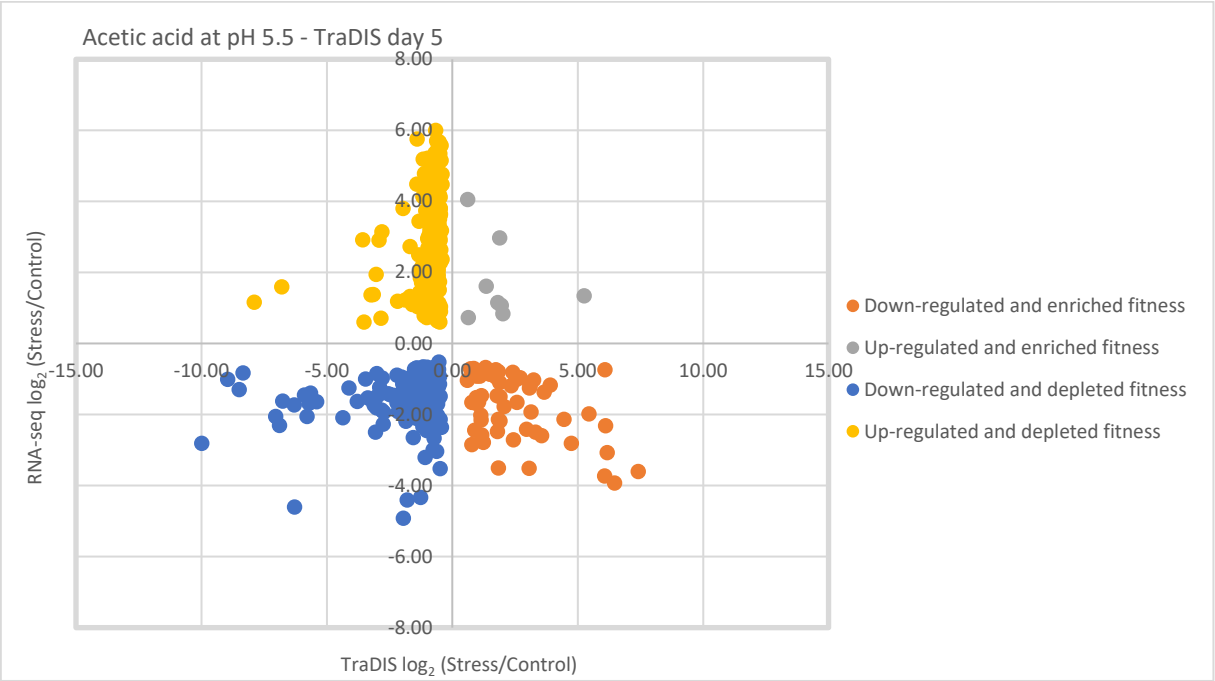


Figure 72. A comparison between RNA-seq data previously generated in the lab and TraDIS data generated in this study.

The log_2 fold changes (pH 5.5 with/without acetic acid) for RNA-seq and TraDIS were plotted against each other. RNA-seq data were collected after four hours of growth. The log_2 fold changes were calculated to determine both the expression changes in RNA-seq and the fitness changes in TraDIS as: Stress log_2 (pH 5.5 + acetic acid)/Control (pH 5.5). Genes significantly expressed in RNA-seq were with cutoff value FDR < 0.01 of the log_2 fold change. A. The comparison of RNA-seq and TraDIS data were collected on day 1. The TraDIS cutoff FDR value < 0.05 for both cases when the transposon inserts cause the strain to enrich in fitness or deplete in fitness. B. The comparison of RNA-seq and TraDIS data were collected on day 5 after serial daily passages. The significant Tn inserts that cause the strain to increase in fitness or decrease in fitness were those with a cutoff FDR < 0.05.

7.5.2 Comparison between RNA-seq, TraDIS, and experimental evolution data.

The previous sections explained the methods and analysis of both RNA-seq and evolution experiments. In this section, the significant TraDIS candidate genes are compared to significantly expressed genes from RNA-seq and genes containing non-synonymous mutants from the evolution experiment. The cut off points or the threshold used for each data set were stated earlier, for TraDIS in section 6.4.1 and for RNAs-seq in section 7.4. The complete data set of the comparison is available in the supplementary data tables S20, S21, S22, and S23. The comparison of these three data sets resulted in four tables:

1. TraDIS candidate genes in EO499 on day 1 where Tn inserts caused loss of fitness under acetic acid stress compared to RNA-seq genes and evolution mutants.
2. TraDIS Candidate genes in EO499 on day 1 where Tn inserts caused increase of fitness under acetic acid stress compared to RNA-seq genes and evolution mutants.
3. TraDIS candidate genes in EO499 on day 5 where Tn inserts caused loss of fitness under acetic acid stress compared to RNA-seq genes and evolution mutants.
4. TraDIS candidate genes in EO499 on day 5 where Tn inserts caused increase of fitness under acetic acid stress compared to RNA-seq genes and evolution mutants.

The tables show the TraDIS candidate significant genes, the predicted function with the corresponded log_2FC, and the FDR, plus the corresponding expressed genes in RNA-seq with log_2FC and the FDR. Note that in RNA-seq the genes can be upregulated or down-regulated and only the significant genes (cutoff <0.01) were highlighted in red. Moreover, the result of the

evolution experiment in six evolved cultures corresponded to TraDIS significant mutants. The evolved cultures were labelled the same as in Dr. Francesca Bushell's thesis (Bushell, 2019). The cultures were labelled form 5.5/4 -1 to 5.5/4 -6; the label stands for pH 5.5 with 4mM AA from 1 to 6 evolved cultures. Genes which have a mutation referred to as 1, and genes which have no mutation detected are shown as zero. The red highlighted cells in the table show where a mutation has occurred in the corresponded genes.

The second comparison is of overlapping genes between TraDIS and RNA-seq correspond to the evolved mutants under pH 5.5 with acetic acid, found in table S24. The table shows the non-synonymous mutants found in the evolution in correspond to TraDIS and RNA-seq, whether present or not.

Venn diagrams (figure 73) were drawn to help to visualize the relative numbers in different overlapping and non-overlapping categories between different conditions, using table S22, S23 and S24. As clarified earlier, this comparison was focused on TraDIS data from day 5. The Venn diagrams were generated using TraDIS significant gene list compared to RNA-seq, and the non-synonymous evolved genes were compared to TraDIS and RNA-seq. There were some small discrepancies in the Venn diagram and that is because of different annotation lists as mentioned earlier.

The overlap between the depleted genes in TraDIS data at day 5 and the significantly expressed genes in RNA-seq were more than the overlap between the enriched genes in TraDIS day 5 and the significantly expressed genes in RNA-seq in both cases, whether genes were up-regulated or down-regulated. While the evolution experiment overlapped very little with the

enriched genes in TraDIS day 5. Also, the Venn diagram shows the data from the evolution experiment is more similar to that from the RNA-seq experiment, particularly for genes whose expression is upregulated, than it is to the TraDIS data.

The other thing to consider is that the timepoints of each data sets were different, which may affect the result of the comparison. However, the number of overlapping genes between the three data sets was very small in all four comparisons, but there is greater between TraDIS and evolution or RNA-seq and evolution.

To find out the significance overlap between the two data sets, an online tool in http://nemates.org was used to determine the probability of this overlap happening by random chance. The test required the number of genes in set 1, the number of genes in set 2, the overlap between both sets and number of genes in the genome used. The probability results for the overlap were found in table 36. The data labelling in the Venn diagram correspond to table 36. A value of < 0.05 was used as significant threshold for P-value. As it is shown in the table, the significant overlap were found between TraDIS day 5 – Enriched and the up- or down-regulated in RNA-seq. While the rest of the overlap between two data sets were not considered statically significant.

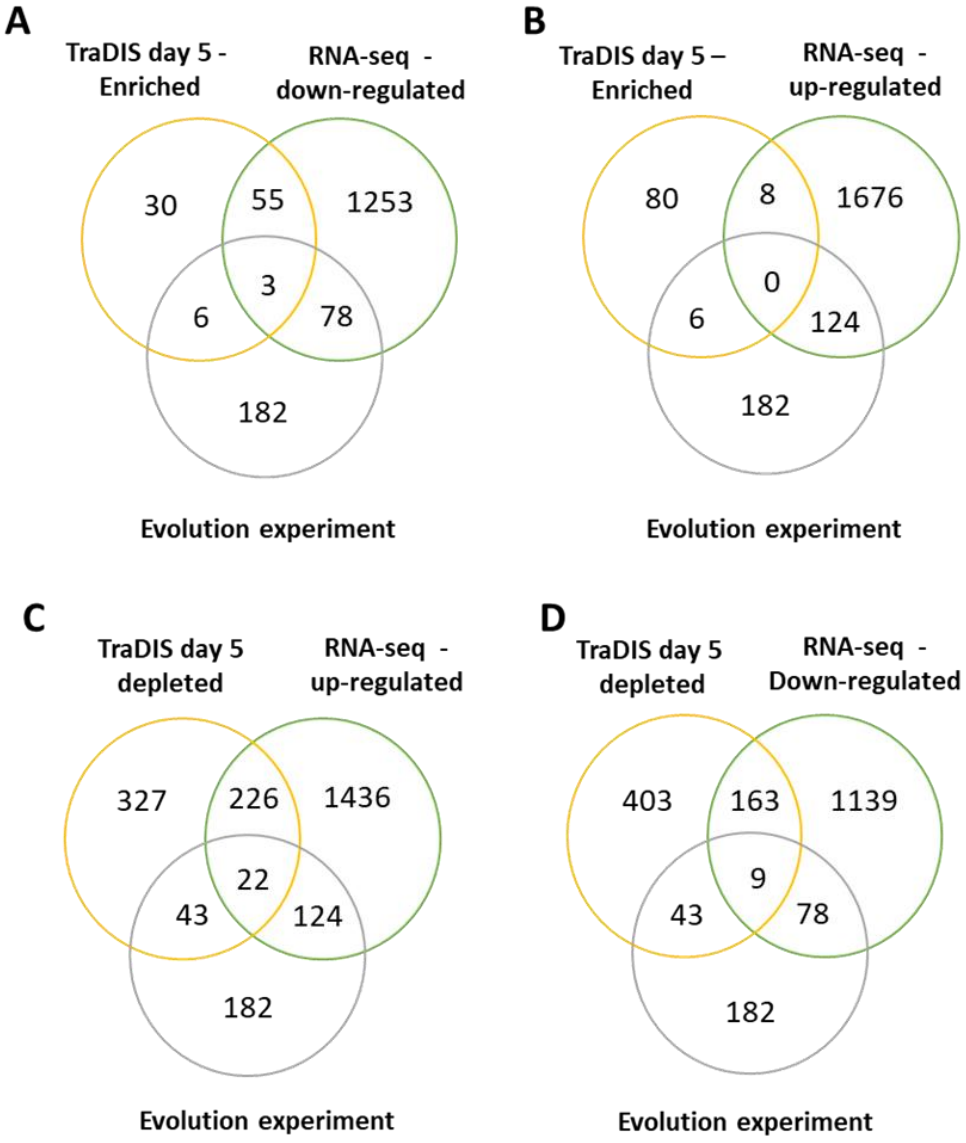


Figure 73. Comparisons of RNA-seq and TraDIS data. The Venn diagram shows the intersection of significant genes in TraDIS, RNA-seq and evolution under acetic acid stress. A more detailed description can be found in the text.

Table 36. The statistical significance of the overlap between the three sets of experiments.

The Venn diagram numbers correspond to figure 73. The cutoff value for significance overlap was P-value < 0.05. The P-values with red underline were considered significant.

Venn diagram number	The overlap between groups of genes	P-value (cutoff < 0.05)
A	TraDIS day 5 – Enriched and RNA-seq – down-regulated	**$< 9.174 \times 10^{-9}$**
	TraDIS day 5 – Enriched and evolution experiment	< 0.420
	RNA-seq – down-regulated and evolution experiment	< 0.461
B	TraDIS day 5 – Enriched and RNA-seq up-regulated	**$< 1.491 \times 10^{-10}$**
	TraDIS day 5 – Enriched and evolution experiment	< 0.399
	RNA-seq – up-regulated and evolution experiment	< 0.182
C	TraDIS day 5 – depleted and RNA-seq up-regulated	< 0.385
	TraDIS day 5 – depleted and evolution experiment	< 0.280
	RNA-seq – up-regulated and evolution experiment	< 0.065
D	TraDIS day 5 – depleted and RNA-seq down-regulated	< 0.106
	TraDIS day 5 – depleted and evolution experiment	< 0.304
	RNA-seq – down-regulated and evolution experiment	< 0.083

7.6 Discussion:

In this chapter I have tried to determine whether TraDIS data from experiments where EO499 was treated with acetic acid at pH 5.5 shows overlap with RNA-seq and evolution experiment data generated by others. This comparative analysis was done based on gene level of the significant candidate genes obtained by TraDIS in this study and the significantly expressed genes in RNA-seq and non-synonymous mutants found in evolution experiments done previously in the lab.

These three different methods measure different things, as described earlier in this chapter. For example, the RNA-seq determines gene transcription or the gene expression under a particular condition, but it can't provide information about the function or the contribution of the genes on the survival of the strain. By comparison TraDIS is able to quantify all the non-essential independent mutants in the examined condition, therefore the genotype-phenotype relationship can be detected and then the related pathways involved in the condition can be determined. Laboratory evolution experiments detect genome variation that leads to an increase in strain fitness.

The comparative analysis of $\log_2$ fold changes (figure 72) showed no correlation or a very weak positive correlation between the TraDIS day 5 candidate genes (fitness) and the significantly expressed genes in RNA-seq under acetic acid based on the $\log_2$ fold change. The correlation between the four different categories didn't fulfill predictions for the hypothesis mentioned earlier in this chapter. This may be simply because each experiment leads up to different cellular measurements. Recently, the same finding was found in (Bushell et al., 2021), where their TraDIS

data at one time point and the RNA-seq data showed no significant correlation at the gene level between the $\log_2$ fold change scores, although some significant overlaps in the enriched pathways contributing to metabolism were identified. In contrast, the correlation between two approaches such as TraDIS day 5 when the mutants were enriched vs the RNA-seq when genes were up- or down-regulated were considered statically significant. While there were no significant overlap were detected among the other data sets. The overlapped between TraDIS or RNA-seq to the evolution were not considered significant, possibly due to the fact there is more genome structural variation in evolution but the provided data were focused only in the non-synonymous mutation, which might lead to less of the overlap of TraDIS or RNA-seq to evolution.

In order to have a more complete picture and to gain more detailed insights into the cellular defects of the overlapped genes between the different experiments, gene enrichment analysis should be conducted, as these overlapping genes may regulate or be part of the same metabolic pathways, which are changing in response to acetic acid stress. This can answer the question about which method is more sensitive to detect the overlapping pathways under acetic acid stress. Mutants with significant relative fitness change, genes with significant relative expression, or the non-synonymous mutants can be further investigated by making knockouts which can then be competed against the wildtype under the examined condition, to determine the impact of individual knockout on phenotype.

8 Bibliography

ABRAM, K. Z., UDAONDO, Z., BLEKER, C., WANCHAI, V., WASSENAAR, T. M., ROBESON, M. S. & USSERY, D. W. 2020. What can we learn from over 100,000 *Escherichia coli* genomes? *bioRxiv***,** 708131.

AGRAWAL, K. S., SARDA, A. V., SHROTRIYA, R., BACHHAV, M., PURI, V. & NATARAJ, G. 2017. Acetic acid dressings: Finding the Holy Grail for infected wound management. *Indian Journal of Plastic Surgery,* 50**,** 273-280.

AMARO, A., CHAMORRO, D., SEEGER, M., ARREDONDO, R., PEIRANO, I. & JEREZ, C. 1991. Effect of external pH perturbations on in vivo protein synthesis by the acidophilic bacterium Thiobacillus ferrooxidans. *Journal of bacteriology,* 173**,** 910-915.

ANDERSON, G. G., PALERMO, J. J., SCHILLING, J. D., ROTH, R., HEUSER, J. & HULTGREN, S. J. 2003. Intracellular bacterial biofilm-like pods in urinary tract infections. *Science,* 301**,** 105-107.

ANSARI, S. & YAMAOKA, Y. 2017. Survival of Helicobacter pylori in gastric acidic territory. *Helicobacter,* 22**,** e12386.

ARIGONI, F., TALABOT, F., PEITSCH, M., EDGERTON, M. D., MELDRUM, E., ALLET, E., FISH, R., JAMOTTE, T., CURCHOD, M.-L. & LOFERER, H. 1998. A genome-based approach for the identification of essential bacterial genes. *Nature biotechnology,* 16**,** 851-856.

ARNEBORG, N., SALSKOV-IVERSEN, A. S. & MATHIASEN, T. E. 1993. The effect of growth rate and other growth conditions on the lipid composition of *Escherichia coli*. *Applied microbiology and biotechnology,* 39**,** 353-357.

BABA, T., ARA, T., HASEGAWA, M., TAKAI, Y., OKUMURA, Y., BABA, M., DATSENKO, K. A., TOMITA, M., WANNER, B. L. & MORI, H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology,* 2.

BARQUIST, L., BOINETT, C. J. & CAIN, A. K. 2013. Approaches to querying bacterial genomes with transposon-insertion sequencing. *RNA biology,* 10**,** 1161-1169.

BARROW, P. A. & SOOTHILL, J. S. 1997. Bacteriophage therapy and prophylaxis: rediscovery and renewed assessment of potential. *Trends in microbiology,* 5**,** 268-271.

BECKER, G., KLAUCK, E. & HENGGE-ARONIS, R. 2000. The response regulator RssB, a recognition factor for σS proteolysis in *Escherichia col*i, can act like an anti-σS factor. *Molecular microbiology,* 35**,** 657-666.

BERNAL, V., CASTAÑO-CEREZO, S. & CÁNOVAS, M. 2016. Acetate metabolism regulation in *Escherichia coli*: carbon overflow, pathogenicity, and beyond. *Applied microbiology and biotechnology,* 100**,** 8985-9001.

BJARNSHOLT, T., ALHEDE, M., JENSEN, P. Ø., NIELSEN, A. K., JOHANSEN, H. K., HOMØE, P., HØIBY, N., GIVSKOV, M. & KIRKETERP-MØLLER, K. 2015. Antibiofilm properties of acetic acid. *Advances in wound care,* 4**,** 363-372.

BLATTNER, F. R., PLUNKETT, G., BLOCH, C. A., PERNA, N. T., BURLAND, V., RILEY, M., COLLADO-VIDES, J., GLASNER, J. D., RODE, C. K. & MAYHEW, G. F. 1997. The complete genome sequence of *Escherichia coli* K-12. *science,* 277**,** 1453-1462.

BOINETT, C. J., CAIN, A. K., HAWKEY, J., DO HOANG, N. T., KHANH, N. N. T., THANH, D. P., DORDEL, J., CAMPBELL, J. I., LAN, N. P. H. & MAYHO, M. 2019. Clinical and laboratory-induced colistin-resistance mechanisms in Acinetobacter baumannii. *Microbial genomics,* 5.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

BOOTH, I. R. 1985. Regulation of cytoplasmic pH in bacteria. *Microbiological reviews,* 49**,** 359.

BOOTH, I. R. The regulation of intracellular pH in bacteria. Novartis Foundation Symposium, 1999. Wiley Online Library, 19-37.

BOUCHER, H. W., TALBOT, G. H., BRADLEY, J. S., EDWARDS, J. E., GILBERT, D., RICE, L. B., SCHELD, M., SPELLBERG, B. & BARTLETT, J. 2009. Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America. *Clinical infectious diseases,* 48**,** 1-12.

BROWN, T., JONES-MORTIMER, M. & KORNBERG, H. 1977. The enzymic interconversion of acetate and acetyl-coenzyme A in *Escherichia coli*. *Microbiology,* 102**,** 327-336.

BRUL, S. & COOTE, P. 1999. Preservative agents in foods: mode of action and microbial resistance mechanisms. *International journal of food microbiology,* 50**,** 1-17.

BURNS, J., MCCOY, C. & IRWIN, N. 2021. Synergistic activity of weak organic acids against uropathogens. *Journal of Hospital Infection,* 111**,** 78-88.

BURTON, N. A., JOHNSON, M. D., ANTCZAK, P., ROBINSON, A. & LUND, P. A. 2010. Novel aspects of the acid response network of *E. coli* K-12 are revealed by a study of transcriptional dynamics. *Journal of molecular biology,* 401**,** 726-742.

BUSHELL, F. M. L. 2019. *Examining the response of Escherichia coli and Pseudomonas aeruginosa to organic acid stress.* University of Birmingham.

BUSHELL, F., HERBERT, J. M., SANNASIDDAPPA, T. H., WARREN, D., TURNER, A. K., FALCIANI, F. & LUND, P. A. 2021. Mapping the Transcriptional and Fitness Landscapes of a Pathogenic *E. coli* Strain: The Effects of Organic Acid Stress under Aerobic and Anaerobic Conditions. *Genes,* 12**,** 53.

CASTANIE-CORNET, M.-P., PENFOUND, T. A., SMITH, D., ELLIOTT, J. F. & FOSTER, J. W. 1999. Control of acid resistance in *Escherichia coli*. *Journal of bacteriology,* 181**,** 3525-3535.

CASTANIE-CORNET, M.-P., TREFFANDIER, H., FRANCEZ-CHARLOT, A., GUTIERREZ, C. & CAM, K. 2007. The glutamate-dependent acid resistance system in *Escherichia coli*: essential and dual role of the His–Asp phosphorelay RcsCDB/AF. *Microbiology,* 153**,** 238-246.

CASTRO, V. S., VIEIRA, B. S., CUNHA-NETO, A., DE SOUZA FIGUEIREDO, E. E. & CONTE-JUNIOR, C. A. 2019. Acetic Acid Increased the Inactivation of Multi-drug Resistant Non-typhoidal Salmonella by Large-Scaffold Antibiotic. *Indian journal of microbiology,* 59**,** 508-513.

CAVALIERE, P. & NOREL, F. 2016. Recent advances in the characterization of Crl, the unconventional activator of the stress sigma factor σS/RpoS. *Biomolecular concepts,* 7**,** 197-204.

CHANG, D.-E., SHIN, S., RHEE, J.-S. & PAN, J.-G. 1999. Acetate metabolism in a pta mutant of *Escherichia coli* w3110: Importance of maintaining acetyl coenzyme a flux for growth and survival. *Journal of bacteriology,* 181**,** 6656-6663.

CHANG, Y. Y. & CRONAN, J. E. 1999. Membrane cyclopropane fatty acid content is a major factor in acid resistance of *Escherichia coli*. *Molecular microbiology,* 33**,** 249-259.

CHAO, M. C., ABEL, S., DAVIS, B. M. & WALDOR, M. K. 2016. The design and analysis of transposon insertion sequencing experiments. *Nature Reviews Microbiology,* 14**,** 119.

CHAO, M. C., PRITCHARD, J. R., ZHANG, Y. J., RUBIN, E. J., LIVNY, J., DAVIS, B. M. & WALDOR, M. K. 2013. High-resolution definition of the Vibrio cholerae essential gene set with hidden Markov model–based analyses of transposon-insertion sequencing data. *Nucleic acids research,* 41**,** 9033-9048.

CHARBONNEAU, A. R., FORMAN, O. P., CAIN, A. K., NEWLAND, G., ROBINSON, C., BOURSNELL, M., PARKHILL, J., LEIGH, J. A., MASKELL, D. J. & WALLER, A. S. 2017. Defining the ABC of gene essentiality in streptococci. *BMC genomics,* 18**,** 1-11.

CHIANG, S. L. & MEKALANOS, J. J. 1998. Use of signature-tagged transposon mutagenesis to identify Vibrio cholerae genes critical for colonization. *Molecular microbiology,* 27**,** 797-805.

COELHO, J. M., TURTON, J. F., KAUFMANN, M. E., GLOVER, J., WOODFORD, N., WARNER, M., PALEPOU, M.-F., PIKE, R., PITT, T. L. & PATEL, B. C. 2006. Occurrence of carbapenem-resistant Acinetobacter baumannii clones at multiple hospitals in London and Southeast England. *Journal of clinical microbiology,* 44**,** 3623-3627.

Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major *Escherichia coli* lineages from the United Kingdom, all belonging to the international O25: H4-ST131 clone. *Antimicrobial Agents and Chemotherapy,* 53**,** 4472-4482.

COOPER, S. & HELMSTETTER, C. E. 1968. Chromosome replication and the division cycle of *Escherichia coli* Br. *Journal of molecular biology,* 31**,** 519-540.

COQUE, T., BAQUERO, F. & CANTON, R. 2008. Increasing prevalence of ESBL-producing Enterobacteriaceae in Europe. *Eurosurveillance,* 13**,** 19044.

CRICK, F. H., BARNETT, L., BRENNER, S. & WATTS-TOBIN, R. J. 1961. General nature of the genetic code for proteins. *Nature,* 192**,** 1227-1232.

CRONAN, J. E. & THOMAS, J. 2009. Bacterial fatty acid synthesis and its relationships with polyketide synthetic pathways. *Methods in enzymology,* 459**,** 395-433.

CULHAM, D. E., LU, A., JISHAGE, M., KROGFELT, K. A., ISHIHAMA, A. & WOOD, J. M. 2001. The osmotic stress response and virulence in pyelonephritis isolates of *Escherichia coli*: contributions of RpoS, ProP, ProU and other systemsThe GenBank accession numbers for the DNA sequences of the rpoS loci in *E. coli* strains HU734 and CFT073 are AF275947 and AF270497, respectively. *Microbiology,* 147**,** 1657-1670.

CUSUMANO, C. K., HUNG, C. S., CHEN, S. L. & HULTGREN, S. J. 2010. Virulence plasmid harbored by uropathogenic *Escherichia coli* functions in acute stages of pathogenesis. *Infection and immunity,* 78**,** 1457-1467.

DATSENKO, K. A. & WANNER, B. L. 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proceedings of the National Academy of Sciences,* 97**,** 6640-6645.

DAVIES, S. C., FOWLER, T., WATSON, J., LIVERMORE, D. M. & WALKER, D. 2013. Annual Report of the Chief Medical Officer: infection and the rise of antimicrobial resistance. *The Lancet,* 381**,** 1606-1609.

DEANGELIS, M. M., WANG, D. G. & HAWKINS, T. L. 1995. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic acids research,* 23**,** 4742.

DeAngelis, M.M., Wang, D.G. and Hawkins, T.L., 1995. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic acids research*, *23*(22), p.4742.

DEATHERAGE, D. E. & BARRICK, J. E. 2014. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Engineering and analyzing multicellular systems.* Springer.

DEATHERAGE, D. E., TRAVERSE, C. C., WOLF, L. N. & BARRICK, J. E. 2015. Detecting rare structural variation in evolving microbial populations from new sequence junctions using breseq. *Frontiers in genetics,* 5**,** 468.

DECOUSSER, J.-W., PINA, P., PICOT, F., DELALANDE, C., PANGON, B., COURVALIN, P., ALLOUCH, P. & GROUP, C. S. 2003. Frequency of isolation and antimicrobial susceptibility of bacterial pathogens isolated from patients with bloodstream infections: a French prospective national survey. *Journal of Antimicrobial Chemotherapy,* 51**,** 1213-1222.

DENICH, T., BEAUDETTE, L., LEE, H. & TREVORS, J. 2003. Effect of selected environmental and physico-chemical factors on bacterial cytoplasmic membranes. *Journal of microbiological methods,* 52**,** 149-182.

DENNIS, G., SHERMAN, B. T., HOSACK, D. A., YANG, J., GAO, W., LANE, H. C. & LEMPICKI, R. A. 2003. DAVID: database for annotation, visualization, and integrated discovery. *Genome biology,* 4**,** 1-11.

DIEKEMA, D., PFALLER, M., JONES, R., DOERN, G., KUGLER, K., BEACH, M., SADER, H. & GROUP, T. S. P. 2000. Trends in antimicrobial susceptibility of bacterial pathogens isolated from patients with bloodstream infections in the USA, Canada and Latin America. *International journal of antimicrobial agents,* 13**,** 257-271.

DOWER, W. J., MILLER, J. F. & RAGSDALE, C. W. 1988. High efficiency transformation of *E. coli* by high voltage electroporation. *Nucleic acids research,* 16**,** 6127-6145.

EDEN. 2017. *What is the phospholipid bilayer and what determines its fluidity?* [Online]. Available: http://blog.cambridgecoaching.com/what-is-the-phospholipid-bilayer-and-what-determines-its-fluidity [Accessed 15/01/2021].

EGUCHI, Y. & UTSUMI, R. 2014. Alkali metals in addition to acidic pH activate the EvgS histidine kinase sensor in *Escherichia coli*. *Journal of bacteriology,* 196**,** 3140-3149.

EGUCHI, Y., ISHII, E., HATA, K. & UTSUMI, R. 2011. Regulation of acid resistance by connectors of two-component signal transduction systems in *Escherichia coli*. *Journal of bacteriology,* 193**,** 1222-1228.

ESHAGHI, M., MEHERSHAHI, K. S. & CHEN, S. L. 2016. Brighter fluorescent derivatives of UTI89 utilizing a monomeric vGFP. *Pathogens,* 5**,** 3.

EXNER, M., BHATTACHARYA, S., CHRISTIANSEN, B., GEBEL, J., GORONCY-BERMES, P., HARTEMANN, P., HEEG, P., ILSCHNER, C., KRAMER, A. & LARSON, E. 2017. Antibiotic resistance: What is so special about multidrug-resistant Gram-negative bacteria? *GMS hygiene and infection control,* 12.

FARMER, W. R. & LIAO, J. C. 1997. Reduction of aerobic acetate production by *Escherichia coli*. *Applied and environmental microbiology,* 63**,** 3205-3210.

FENG, Y. & CRONAN, J. E. 2011. Complex binding of the FabR repressor of bacterial unsaturated fatty acid biosynthesis to its cognate promoters. *Molecular microbiology,* 80**,** 195-218.

FENLON, S. N., CHEE, Y. C., CHEE, J. L. Y., CHOY, Y. H., KHNG, A. J., LIOW, L. T., MEHERSHAHI, K. S., RUAN, X., TURNER, S. W. & YAO, F. 2020. Sequencing of *E. coli* strain UTI89 on multiple sequencing platforms. *BMC Research Notes,* 13**,** 1-4.

FLORES-MIRELES, A. L., WALKER, J. N., CAPARON, M. & HULTGREN, S. J. 2015. Urinary tract infections: epidemiology, mechanisms of infection and treatment options. *Nature reviews microbiology,* 13**,** 269-284.

FORDE, B. M., ZAKOUR, N. L. B., STANTON-COOK, M., PHAN, M.-D., TOTSIKA, M., PETERS, K. M., CHAN, K. G., SCHEMBRI, M. A., UPTON, M. & BEATSON, S. A. 2014. The complete genome sequence of *Escherichia coli* EC958: a high quality reference sequence for the globally disseminated multidrug resistant *E. coli* O25b: H4-ST131 clone. *PloS one,* 9**,** e104400.

FOROUHAR, F., KUZIN, A., SEETHARAMAN, J., LEE, I., ZHOU, W., ABASHIDZE, M., CHEN, Y., YONG, W., JANJUA, H. & FANG, Y. 2007. Functional insights from structural genomics. *Journal of structural and functional genomics,* 8**,** 37-44.

FOSTER, J. W. 2004. *Escherichia coli* acid resistance: tales of an amateur acidophile. *Nature Reviews Microbiology,* 2**,** 898-907.

FRANK, K. 1994. Measures to preserve food and feeds from bacterial damage. *UÈ bersichten zur TierernaÈhrung,* 22**,** 149-63.

FRIEDRICH, T., DEKOVIC, D. K. & BURSCHEL, S. 2016. Assembly of the *Escherichia coli* NADH: ubiquinone oxidoreductase (respiratory complex I). *Biochimica et Biophysica Acta (BBA)-Bioenergetics,* 1857**,** 214-223.

FUX, C., SHIRTLIFF, M., STOODLEY, P. & COSTERTON, J. W. 2005. Can laboratory reference strains mirror 'real-world'pathogenesis? *Trends in microbiology,* 13**,** 58-63.

GE, S. X., JUNG, D. & YAO, R. 2020. ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics,* 36**,** 2628-2629.

GLASS, K. & GIRVAN, M. 2014. Annotation enrichment analysis: an alternative method for evaluating the functional properties of gene sets. *Scientific reports,* 4**,** 1-9.

GOODALL, E. C. 2019. *Using TraDIS to probe the model organism Escherichia coli.* University of Birmingham.

GOODALL, E. C., ROBINSON, A., JOHNSTON, I. G., JABBARI, S., TURNER, K. A., CUNNINGHAM, A. F., LUND, P. A., COLE, J. A. & HENDERSON, I. R. 2018. The essential genome of *Escherichia coli* K-12. *MBio,* 9**,** e02096-17.

GORYSHIN, I. Y. & REZNIKOFF, W. S. 1998. Tn5 in vitro transposition. *Journal of Biological Chemistry,* 273**,** 7367-7374.

GREEN, L. S. & EMERICH, D. W. 1997. Bradyrhizobium japonicum does not require alpha-ketoglutarate dehydrogenase for growth on succinate or malate. *Journal of bacteriology,* 179**,** 194-201.

GRIFFIN, J. E., GAWRONSKI, J. D., DEJESUS, M. A., IOERGER, T. R., AKERLEY, B. J. & SASSETTI, C. M. 2011. High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathog,* 7**,** e1002251.

GUAN, N. & LIU, L. 2020. Microbial response to acid stress: mechanisms and applications. *Applied microbiology and biotechnology,* 104**,** 51-65.

GUILIANI, N. & JEREZ, C. A. 2000. Molecular cloning, sequencing, and expression ofomp-40, the gene coding for the major outer membrane protein from the acidophilic bacterium Thiobacillus ferrooxidans. *Applied and environmental microbiology,* 66**,** 2318-2324.

GURANOWSKI, A., JAKUBOWSKI, H. & HOLLER, E. 1983. Catabolism of diadenosine 5', 5"'-P1, P4-tetraphosphate in procaryotes. Purification and properties of diadenosine 5', 5"'-P1, P4-tetraphosphate (symmetrical) pyrophosphohydrolase from *Escherichia coli* K12. *Journal of Biological Chemistry,* 258**,** 14784-14789.

HALSTEAD, F. D., RAUF, M., MOIEMEN, N. S., BAMFORD, A., WEARN, C. M., FRAISE, A. P., LUND, P. A., OPPENHEIM, B. A. & WEBBER, M. A. 2015. The antibacterial activity of acetic acid against biofilm-producing pathogens of relevance to burns patients. *PloS one,* 10**,** e0136190.

HARFE, B. D. & JINKS-ROBERTSON, S. 2000. DNA mismatch repair and genetic instability. *Annual review of genetics,* 34**,** 359-399.

HARRIS, R. M., WEBB, D. C., HOWITT, S. M. & COX, G. B. 2001. Characterization of PitA and PitB from *Escherichia coli*. *Journal of bacteriology,* 183**,** 5008-5014.

HAWKEY, P. M. 2003. Mechanisms of quinolone action and microbial response. *Journal of Antimicrobial Chemotherapy,* 51**,** 29-35.

HENGGE, R. 2011. Stationary-Phase Gene Regulation in *Escherichia coli* §. *EcoSal Plus,* 4.

HENSEL, M., SHEA, J. E., GLEESON, C., JONES, M. D., DALTON, E. & HOLDEN, D. W. 1995. Simultaneous identification of bacterial virulence genes by negative selection. *Science,* 269**,** 400-403.

HOBMAN, J. L., PENN, C. W. & PALLEN, M. J. 2007. Laboratory strains of *Escherichia coli*: model citizens or deceitful delinquents growing old disgracefully? *Molecular microbiology,* 64**,** 881-885.

HOLEVA, M. C., BELL, K. S., HYMAN, L. J., AVROVA, A. O., WHISSON, S. C., BIRCH, P. R. & TOTH, I. K. 2004. Use of a pooled transposon mutation grid to demonstrate roles in disease development for Erwinia carotovora subsp. atroseptica putative type III secreted effector (DspE/A) and helper (HrpN) proteins. *Molecular plant-microbe interactions,* 17**,** 943-950.

HOOPER, D. C. 2001. Mechanisms of action of antimicrobials: focus on fluoroquinolones. *Clinical Infectious Diseases,* 32**,** S9-S15.

HTTPS://CHEM.LIBRETEXTS.ORG/BOOKSHELVES/GENERAL_CHEMISTRY. 2012. *A Qualitative Description of Acid–Base Equilibriums* [Online]. Saylor Academy. Available: https://saylordotorg.github.io/text_general-chemistry-principles-patterns-and-applications-v1.0/s20-02-a-qualitative-description-of-a.html [Accessed 2021].

HTTPS://CHEM.LIBRETEXTS.ORG/BOOKSHELVES/GENERAL_CHEMISTRY. 2012. *A Qualitative Description of Acid–Base Equilibriums* [Online]. Saylor Academy. Available: https://saylordotorg.github.io/text_general-chemistry-principles-patterns-and-applications-v1.0/s20-02-a-qualitative-description-of-a.html [Accessed 2021].

IDALIA, V.-M. N. & BERNARDO, F. 2017. *Escherichia coli* as a model organism and its application in biotechnology. *Recent Advances on Physiology, Pathogenesis and Biotechnological Applications. In Tech Open, Rijeka, Croatia***,** 253-274.

JAYEOLA, V., MCCLELLAND, M., PORWOLLIK, S., CHU, W., FARBER, J. & KATHARIOU, S. 2020. Identification of novel genes mediating survival of Salmonella on low-moisture foods via transposon sequencing analysis. *Frontiers in Microbiology,* 11**,** 726.

JUSTICE, S. S., HUNSTAD, D. A., SEED, P. C. & HULTGREN, S. J. 2006. Filamentation by *Escherichia coli* subverts innate defenses during urinary tract infection. *Proceedings of the National Academy of Sciences,* 103**,** 19884-19889.

KAISER, J. C., SEN, S., SINHA, A., WILKINSON, B. J. & HEINRICHS, D. E. 2016. The role of two branched-chain amino acid transporters in S taphylococcus aureus growth, membrane fatty acid composition and virulence. *Molecular microbiology,* 102**,** 850-864.

KING, T., LUCCHINI, S., HINTON, J. C. & GOBIUS, K. 2010. Transcriptomic analysis of *Escherichia coli* O157: H7 and K-12 cultures exposed to inorganic and organic acids in stationary phase reveals acidulant-and strain-specific acid tolerance responses. *Applied and environmental microbiology,* 76**,** 6514-6528.

KITKO, R. D., CLEETON, R. L., ARMENTROUT, E. I., LEE, G. E., NOGUCHI, K., BERKMEN, M. B., JONES, B. D. & SLONCZEWSKI, J. L. 2009. Cytoplasmic acidification and the benzoate transcriptome in Bacillus subtilis. *PLoS One,* 4**,** e8255.

KLEIN, A. H., SHULLA, A., REIMANN, S. A., KEATING, D. H. & WOLFE, A. J. 2007. The intracellular concentration of acetyl phosphate in *Escherichia coli* is sufficient for direct phosphorylation of two-component response regulators. *Journal of bacteriology,* 189**,** 5574-5581.

KNÖPPEL, A., KNOPP, M., ALBRECHT, L. M., LUNDIN, E., LUSTIG, U., NÄSVALL, J. & ANDERSSON, D. I. 2018. Genetic adaptation to growth under laboratory conditions in *Escherichia coli* and Salmonella enterica. *Frontiers in microbiology,* 9**,** 756.

KRAM, K. E., GEIGER, C., ISMAIL, W. M., LEE, H., TANG, H., FOSTER, P. L. & FINKEL, S. E. 2017. Adaptation of *Escherichia coli* to long-term serial passage in complex medium: evidence of parallel evolution. *Msystems,* 2**,** e00192-16.

KROLL, R. & BOOTH, I. 1983. The relationship between intracellular pH, the pH gradient and potassium transport in *Escherichia coli*. *Biochemical Journal,* 216**,** 709-716.

KUMARI, S., TISHEL, R., EISENBACH, M. & WOLFE, A. J. 1995. Cloning, characterization, and functional expression of acs, the gene which encodes acetyl coenzyme A synthetase in *Escherichia coli*. *Journal of bacteriology,* 177**,** 2878-2886.

KUNDUKAD, B., UDAYAKUMAR, G., GRELA, E., KAUR, D., RICE, S. A., KJELLEBERG, S. & DOYLE, P. S. 2020. Weak acids as an alternative anti-microbial therapy. *Biofilm,* 2**,** 100019.

LANGRIDGE, G. C., PHAN, M.-D., TURNER, D. J., PERKINS, T. T., PARTS, L., HAASE, J., CHARLES, I., MASKELL, D. J., PETERS, S. E. & DOUGAN, G. 2009. Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome research,* 19**,** 2308-2316.

LAU, G. W., HAATAJA, S., LONETTO, M., KENSIT, S. E., MARRA, A., BRYANT, A. P., MCDEVITT, D., MORRISON, D. A. & HOLDEN, D. W. 2001. A functional genomic analysis of type 3 Streptococcus pneumoniae virulence. *Molecular microbiology,* 40**,** 555-571.

LEE, D. J., BINGLE, L. E., HEURLIER, K., PALLEN, M. J., PENN, C. W., BUSBY, S. J. & HOBMAN, J. L. 2009. Gene doctoring: a method for recombineering in laboratory and pathogenic *Escherichia coli strains*. *BMC microbiology,* 9**,** 252.

LENSKI, R. E. 2017. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *The ISME journal,* 11**,** 2181-2194.

LENSKI, R. E., ROSE, M. R., SIMPSON, S. C. & TADLER, S. C. 1991. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *The American Naturalist,* 138**,** 1315-1341.

LÉVÊQUE, F., BLANCHIN-ROLAND, S., FAYAT, G., PLATEAU, P. & BLANQUET, S. 1990. Design and characterization of *Escherichia coli* mutants devoid of Ap4N-hydrolase activity. *Journal of molecular biology,* 212**,** 319-329.

LEVY, S. 2002. The antibiotic paradox: How misuse of antibiotics destroys their curative powers (Perseus Cambridge, 2002).

LEVY, S. B. & MILLER, R. V. 1989. *Gene transfer in the environment*, McGraw-Hill Companies.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics,* 25**,** 1754-1760.

LI, Z., BOUCKAERT, J., DEBOECK, F., DE GREVE, H. & HERNALSTEENS, J.-P. 2012. Nicotinamide dependence of uropathogenic *Escherichia coli* UTI89 and application of nadB as a neutral insertion site. *Microbiology,* 158**,** 736-745.

LIN, J., LEE, I. S., FREY, J., SLONCZEWSKI, J. L. & FOSTER, J. W. 1995. Comparative analysis of extreme acid survival in Salmonella typhimurium, Shigella flexneri, and *Escherichia coli*. *Journal of bacteriology,* 177**,** 4097-4104.

LINDBERG, L., SANTOS, A. X., RIEZMAN, H., OLSSON, L. & BETTIGA, M. 2013. Lipidomic profiling of Saccharomyces cerevisiae and Zygosaccharomyces bailii reveals critical changes in lipid composition in response to acetic acid stress. *PloS one,* 8**,** e73936.

LIU, A., TRAN, L., BECKET, E., LEE, K., CHINN, L., PARK, E., TRAN, K. & MILLER, J. H. 2010. Antibiotic sensitivity profiles determined with an *Escherichia coli* gene knockout collection: generating an antibiotic bar code. *Antimicrobial agents and chemotherapy,* 54**,** 1393-1403.

LIVERMORE, D. 2004. The need for new antibiotics. *Clinical microbiology and infection,* 10**,** 1-9.

LUCIGEN 2016. EZ-Tn5™ <KAN-2> Insertion Kit, Cat. No. TSM99K2. Epicentre an illumina company.

LUND, P. A., DE BIASE, D., LIRAN, O., SCHELER, O., MIRA, N. P., CETECIOGLU, Z., FERNÁNDEZ, E. N., BOVER-CID, S., HALL, R. & SAUER, M. 2020. Understanding How Microorganisms Respond to Acid pH Is Central to Their Control and Successful Exploitation. *Frontiers in Microbiology,* 11**,** 2233.

LUND, P., TRAMONTI, A. & DE BIASE, D. 2014. Coping with low pH: molecular strategies in neutralophilic bacteria. *FEMS microbiology reviews,* 38**,** 1091-1125.

MARCH, J. 1968. Carboxylic acid. *Encyclopaedia Britannica* New York.

MARZAN, L. W. & SHIMIZU, K. 2011. Metabolic regulation of *Escherichia coli* and its phoB and phoR genes knockout mutants under phosphate and nitrogen limitations as well as at acidic condition. *Microbial cell factories,* 10**,** 1-15.

MCLAGGAN, D., NAPRSTEK, J., BUURMAN, E. T. & EPSTEIN, W. 1994. Interdependence of K+ and glutamate accumulation during osmotic adaptation of *Escherichia coli*. *Journal of Biological Chemistry,* 269**,** 1911-1917.

MEI, J. M., NOURBAKHSH, F., FORD, C. W. & HOLDEN, D. W. 1997. Identification of Staphylococcus aureus virulence genes in a murine model of bacteraemia using signature-tagged mutagenesis. *Molecular microbiology,* 26**,** 399-407.

MELAMED, S., PEER, A., FAIGENBAUM-ROMM, R., GATT, Y. E., REISS, N., BAR, A., ALTUVIA, Y., ARGAMAN, L. & MARGALIT, H. 2016. Global mapping of small RNA-target interactions in bacteria. *Molecular cell,* 63**,** 884-897.

MENG, J., DOYLE, M., ZHAO, T. & ZHAO, S. 2007. Food microbiology: fundamentals and frontiers. ASM press Washington DC.

MENG, P., LU, C., ZHANG, Q., LIN, J. & CHEN, F. 2017. Exploring the genomic diversity and cariogenic differences of Streptococcus mutans strains through pan-genome and comparative genome analysis. *Current microbiology,* 74**,** 1200-1209.

MESARICH, C. H., REES-GEORGE, J., GARDNER, P. P., GHOMI, F. A., GERTH, M. L., ANDERSEN, M. T., RIKKERINK, E. H., FINERAN, P. C. & TEMPLETON, M. D. 2017. Transposon insertion libraries for the characterization of mutants from the kiwifruit pathogen Pseudomonas syringae pv. actinidiae. *PloS one,* 12.

MI, H., POUDEL, S., MURUGANUJAN, A., CASAGRANDE, J. T. & THOMAS, P. D. 2016. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic acids research,* 44**,** D336-D342.

MICEVSKI, D., ZAMMIT, J. E., TRUSCOTT, K. N. & DOUGAN, D. A. 2015. Anti-adaptors use distinct modes of binding to inhibit the RssB-dependent turnover of RpoS (σS) by ClpXP. *Frontiers in Molecular Biosciences,* 2**,** 15.

MIRA, N. P., TEIXEIRA, M. C. & SÁ-CORREIA, I. 2010. Adaptive response and tolerance to weak acids in Saccharomyces cerevisiae: a genome-wide view. *Omics: a journal of integrative biology,* 14**,** 525-540.

MUFFLER, A., FISCHER, D., ALTUVIA, S., STORZ, G. & HENGGE-ARONIS, R. 1996. The response regulator RssB controls stability of the sigma (S) subunit of RNA polymerase in *Escherichia coli*. *The EMBO journal,* 15**,** 1333-1339.

NAGOBA, B., SELKAR, S., WADHER, B. & GANDHI, R. 2013. Acetic acid treatment of pseudomonal wound infections–a review. *Journal of infection and public health,* 6**,** 410-415.

NICOLAS-CHANOINE, M.-H., BLANCO, J., LEFLON-GUIBOUT, V., DEMARTY, R., ALONSO, M. P., CANIÇA, M. M., PARK, Y.-J., LAVIGNE, J.-P., PITOUT, J. & JOHNSON, J. R. 2008. Intercontinental emergence of *Escherichia coli* clone O25: H4-ST131 producing CTX-M-15. *Journal of Antimicrobial Chemotherapy,* 61**,** 273-281.

NIKAIDO, H. 1996. Multidrug efflux pumps of gram-negative bacteria. *Journal of bacteriology,* 178**,** 5853.

NOUR, S., REID, G., SATHANANTHAM, K. & MACKIE, I. 2021. Acetic acid dressings used to treat pseudomonas colonised burn wounds: A UK national survey. *Burns*.

O'NEILL, J. 2015. The Review on Antimicrobial Resistance Chaired by Jim O'Neill. *HM Government, Wellcome Trust*.

ORR, J. S., CHRISTENSEN, D. G., WOLFE, A. J. & RAO, C. V. 2019. Extracellular acidic pH inhibits acetate consumption by decreasing gene transcription of the tricarboxylic acid cycle and the glyoxylate shunt. *Journal of bacteriology,* 201.

PAGE, A. J., CUMMINS, C. A., HUNT, M., WONG, V. K., REUTER, S., HOLDEN, M. T., FOOKES, M., FALUSH, D., KEANE, J. A. & PARKHILL, J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics,* 31**,** 3691-3693.

PALMER, A. C. & KISHONY, R. 2014. Opposing effects of target overexpression reveal drug mechanisms. *Nature communications,* 5.

PARK, G., PARK, J. K., SHIN, S.-H., JEON, H.-J., KIM, N. K., KIM, Y. J., SHIN, H.-T., LEE, E., LEE, K. H. & SON, D.-S. 2017. Characterization of background noise in capture-based targeted sequencing data. *Genome biology,* 18**,** 1-13.

PEARSON, W. R., WOOD, T., ZHANG, Z. & MILLER, W. 1997. Comparison of DNA sequences with protein sequences. *Genomics,* 46**,** 24-36.

PENNACCHIETTI, E., D'ALONZO, C., FREDDI, L., OCCHIALINI, A. & DE BIASE, D. 2018. The glutaminase-dependent acid resistance system: qualitative and quantitative assays and analysis of its distribution in enteric bacteria. *Frontiers in microbiology,* 9**,** 2869.

PHAN, M.-D., BOTTOMLEY, A. L., PETERS, K. M., HARRY, E. J. & SCHEMBRI, M. A. 2020. Uncovering novel susceptibility targets to enhance the efficacy of third-generation cephalosporins against ESBL-producing uropathogenic *Escherichia coli*. *Journal of Antimicrobial Chemotherapy,* 75**,** 1415-1423.

PINHAL, S., ROPERS, D., GEISELMANN, J. & DE JONG, H. 2019. Acetate metabolism and the inhibition of bacterial growth by acetate. *Journal of bacteriology,* 201.

PLETNEV, P., OSTERMAN, I., SERGIEV, P., BOGDANOV, A. & DONTSOVA, O. 2015. Survival guide: *Escherichia coli* in the stationary phase. *Acta Naturae (англоязычная версия),* 7.

POLOSINA, Y. Y., MUI, J., PITSIKAS, P. & CUPPLES, C. G. 2009. The *Escherichia coli* mismatch repair protein MutL recruits the Vsr and MutH endonucleases in response to DNA damage. *Journal of bacteriology,* 191**,** 4041-4043.

PRESCOTT, H., KLEIN 2005. *Microbiology,* New York, McGraw-Hill.

PRICE, M. N., WETMORE, K. M., WATERS, R. J., CALLAGHAN, M., RAY, J., LIU, H., KUEHL, J. V., MELNYK, R. A., LAMSON, J. S. & SUH, Y. 2018. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature,* 557**,** 503-509.

PRUTEANU, M. & HENGGE-ARONIS, R. 2002. The cellular level of the recognition factor RssB is rate-limiting for σS proteolysis: implications for RssB regulation and signal transduction in σS turnover in *Escherichia coli*. *Molecular microbiology,* 45**,** 1701-1713.

RAHMAN, M., HASAN, M. R., OBA, T. & SHIMIZU, K. 2006. Effect of rpoS gene knockout on the metabolism of *Escherichia coli* during exponential growth phase and early stationary phase based on gene expressions, enzyme activities and intracellular metabolite concentrations. *Biotechnology and bioengineering,* 94**,** 585-595.

RASKIN, D. M., SESHADRI, R., PUKATZKI, S. U. & MEKALANOS, J. J. 2006. Bacterial genomics and pathogen evolution. *Cell,* 124**,** 703-714.

REGELMANN, A. G., LESLEY, J. A., MOTT, C., STOKES, L. & WALDBURGER, C. D. 2002. Mutational analysis of the *Escherichia coli* PhoQ sensor kinase: differences with the Salmonella enterica serovar Typhimurium PhoQ protein and in the mechanism of Mg2+ and Ca2+ sensing. *Journal of bacteriology,* 184**,** 5468-5478.

REN, J., SANG, Y., LU, J. & YAO, Y.-F. 2017. Protein acetylation and its role in bacterial virulence. *Trends in microbiology,* 25**,** 768-779.

REZNIKOFF, W. S. 2008. Transposon tn 5. *Annual review of genetics,* 42**,** 269-286.

RICHARD, H. & FOSTER, J. W. 2004. *Escherichia coli* glutamate-and arginine-dependent acid resistance systems increase internal pH and reverse transmembrane potential. *Journal of bacteriology,* 186**,** 6032-6041.

RILEY, M., ABE, T., ARNAUD, M. B., BERLYN, M. K., BLATTNER, F. R., CHAUDHURI, R. R., GLASNER, J. D., HORIUCHI, T., KESELER, I. M. & KOSUGE, T. 2006. *Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005. *Nucleic acids research,* 34**,** 1-9.

ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26**,** 139-140.

ROE, A. J., O'BYRNE, C., MCLAGGAN, D. & BOOTH, I. R. 2002. Inhibition of *Escherichia coli* growth by acetic acid: a problem with methionine biosynthesis and homocysteine toxicity. *Microbiology,* 148**,** 2215-2222.

RUBIN, B. E., WETMORE, K. M., PRICE, M. N., DIAMOND, S., SHULTZABERGER, R. K., LOWE, L. C., CURTIN, G., ARKIN, A. P., DEUTSCHBAUER, A. & GOLDEN, S. S. 2015. The essential gene set of a photosynthetic organism. *Proceedings of the National Academy of Sciences,* 112**,** E6634-E6643.

RUSSELL, J. B. 2007. The energy spilling reactions of bacteria and other organisms. *Journal of molecular microbiology and biotechnology,* 13**,** 1-11.

RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M.-A. & BARRELL, B. 2000. Artemis: sequence visualization and annotation. *Bioinformatics,* 16**,** 944-945.

SANDERS, C., TURKARSLAN, S., LEE, D.-W. & DALDAL, F. 2010. Cytochrome c biogenesis: the Ccm system. *Trends in microbiology,* 18**,** 266-274.

SAWERS, R. G. & CLARK, D. P. 2004. Fermentative pyruvate and acetyl-coenzyme a metabolism. *EcoSal Plus,* 1.

SAYED, A. K., ODOM, C. & FOSTER, J. W. 2007. The *Escherichia coli* AraC-family regulators GadX and GadW activate gadE, the central activator of glutamate-dependent acid resistance. *Microbiology,* 153**,** 2584-2592.

SCHINNER, S., ENGELHARDT, F., PREUSSE, M., THÖMING, J. G., TOMASCH, J. & HÄUSSLER, S. 2020. Genetic determinants of Pseudomonas aeruginosa fitness during biofilm growth. *Biofilm,* 2**,** 100023.

SEEMANN, T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics,* 30**,** 2068-2069.

SEN, S., SIROBHUSHANAM, S., HANTAK, M. P., LAWRENCE, P., BRENNA, J. T., GATTO, C. & WILKINSON, B. J. 2015. Short branched-chain C6 carboxylic acids result in increased growth, novel 'unnatural'fatty acids and increased membrane fluidity in a Listeria monocytogenes branched-chain fatty acid-deficient mutant. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids,* 1851**,** 1406-1415.

SEO, S. W., KIM, D., O'BRIEN, E. J., SZUBIN, R. & PALSSON, B. O. 2015. Decoding genome-wide GadEWX-transcriptional regulatory networks reveals multifaceted cellular responses to acid stress in *Escherichia coli*. *Nature communications,* 6**,** 1-8.

SHANKAR, P. R. 2016. Book review: tackling drug-resistant infections globally. *Archives of Pharmacy Practice,* 7**,** 110-111.

SHARP, C., BOINETT, C., CAIN, A., HOUSDEN, N. G., KUMAR, S., TURNER, K., PARKHILL, J. & KLEANTHOUS, C. 2019. O-antigen-dependent colicin insensitivity of uropathogenic *Escherichia coli*. *Journal of bacteriology,* 201.

SHEVCHENKO, Y., BOUFFARD, G. G., BUTTERFIELD, Y. S., BLAKESLEY, R. W., HARTLEY, J. L., YOUNG, A. C., MARRA, M. A., JONES, S. J., TOUCHMAN, J. W. & GREEN, E. D. 2002. Systematic sequencing of cDNA clones using the transposon Tn5. *Nucleic acids research,* 30**,** 2469-2477.

SIMS, G. E. & KIM, S.-H. 2011. Whole-genome phylogeny of *Escherichia coli/*Shigella group by feature frequency profiles (FFPs). *Proceedings of the National Academy of Sciences,* 108**,** 8329-8334.

SMITH, R. & COAST, J. 2013. The true cost of antimicrobial resistance. *Bmj,* 346**,** f1493.

SOHLENKAMP, C. 2017. Membrane homeostasis in bacteria upon pH challenge. *Biogenesis of fatty acids, lipids and membranes***,** 1-13.

SOLAIMANPOUR, S., SARMIENTO, F. & MRAZEK, J. 2015. Tn-seq explorer: a tool for analysis of high-throughput sequencing data of transposon mutant libraries. *PLoS One,* 10**,** e0126070.

SURYADARMA, P., OJIMA, Y., FUKUDA, Y., AKAMATSU, N. & TAYA, M. 2012. The rpoS deficiency suppresses acetate accumulation in glucose-enriched culture of *Escherichia coli* under an aerobic condition. *Frontiers of Chemical Science and Engineering,* 6**,** 152-157.

TAN, Z., YOON, J. M., NIELSEN, D. R., SHANKS, J. V. & JARBOE, L. R. 2016. Membrane engineering via trans unsaturated fatty acids production improves *Escherichia coli* robustness and production of biorenewables. *Metabolic engineering,* 35**,** 105-113.

THOMASON, L. C., COSTANTINO, N. & COURT, D. L. 2007. *E. coli* genome manipulation by P1 transduction. *Current protocols in molecular biology***,** 1.17. 1-1.17. 8.

THOMPSON, K. M. & GOTTESMAN, S. 2014. The MiaA tRNA modification enzyme is necessary for robust RpoS expression in *Escherichia coli*. *Journal of bacteriology,* 196**,** 754-761.

TRCHOUNIAN, A. & TRCHOUNIAN, K. 2019. Fermentation revisited: how do microorganisms survive under energy-limited conditions? *Trends in biochemical sciences,* 44**,** 391-400.

TSAY, J.-T., ROCK, C. & JACKOWSKI, S. 1992. Overproduction of beta-ketoacyl-acyl carrier protein synthase I imparts thiolactomycin resistance to *Escherichia coli* K-12. *Journal of bacteriology,* 174**,** 508-513.

TU, Q., YIN, J., FU, J., HERRMANN, J., LI, Y., YIN, Y., STEWART, A. F., MÜLLER, R. & ZHANG, Y. 2016. Room temperature electrocompetent bacterial cells improve DNA transformation and recombineering efficiency. *Scientific reports,* 6**,** 1-8.

VAN ELSAS, J. D., SEMENOV, A. V., COSTA, R. & TREVORS, J. T. 2011. Survival of *Escherichia coli* in the environment: fundamental and public health aspects. *The ISME journal,* 5**,** 173-183.

VAN OPIJNEN, T. & CAMILLI, A. 2012. A fine scale phenotype–genotype virulence map of a bacterial pathogen. *Genome research,* 22**,** 2541-2551.

VAN OPIJNEN, T. & CAMILLI, A. 2013. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nature Reviews Microbiology,* 11**,** 435-442.

VAN OPIJNEN, T. & LEVIN, H. L. 2020. Transposon Insertion Sequencing, a Global Measure of Gene Function. *Annual Review of Genetics,* 54.

VIVIJS, B., AERTSEN, A. & MICHIELS, C. W. 2016. Identification of genes required for growth of *Escherichia coli* MG1655 at moderately low pH. *Frontiers in microbiology,* 7**,** 1672.

VOLLAN, H. S., TANNÆS, T., CAUGANT, D. A., VRIEND, G. & BUKHOLM, G. 2017. Outer membrane phospholipase A's roles in Helicobacter pylori acid adaptation. *Gut pathogens,* 9**,** 36.

VOORDOUW, G., VAN DER VIES, S. M. & THEMMEN, A. P. 1983. Why are two different types of pyridine nucleotide transhydrogenase found in living organisms? *European journal of biochemistry,* 131**,** 527-533.

WANG, H., DZINK-FOX, J. L., CHEN, M. & LEVY, S. B. 2001. Genetic characterization of highly fluoroquinolone-resistant clinical *Escherichia coli* strains from China: role ofacrR mutations. *Antimicrobial agents and chemotherapy,* 45**,** 1515-1521.

WANG, X., LESTERLIN, C., REYES-LAMOTHE, R., BALL, G. & SHERRATT, D. J. 2011. Replication and segregation of an *Escherichia coli* chromosome with two replication origins. *Proceedings of the National Academy of Sciences,* 108**,** E243-E250.

WILKS, J. C. & SLONCZEWSKI, J. L. 2007. pH of the cytoplasm and periplasm of *Escherichia coli*: rapid measurement by green fluorescent protein fluorimetry. *Journal of bacteriology,* 189**,** 5601-5607.

WILLCOCKS, S. J., STABLER, R. A., ATKINS, H. S., OYSTON, P. F. & WREN, B. W. 2018. High-throughput analysis of Yersinia pseudotuberculosis gene essentiality in optimised in vitro conditions, and implications for the speciation of Yersinia pestis. *BMC microbiology,* 18**,** 1-11.

WISER, M. J. & LENSKI, R. E. 2015. A comparison of methods to measure fitness in *Escherichia coli*. *PloS one,* 10**,** e0126210.

WITTE, W. 1998. Medical consequences of antibiotic use in agriculture. *Science,* 279**,** 996-997.

WOLFE, A. J. 2005. The acetate switch. *Microbiology and molecular biology reviews,* 69**,** 12-50.

WOODFORD, N., CARATTOLI, A., KARISIK, E., UNDERWOOD, A., ELLINGTON, M. J. & LIVERMORE, D. M. 2009.

WOODFORD, N., TURTON, J. F. & LIVERMORE, D. M. 2011. Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. *FEMS microbiology reviews,* 35**,** 736-755.

WOODFORD, N., WARD, M., KAUFMANN, M., TURTON, J., FAGAN, E., JAMES, D., JOHNSON, A., PIKE, R., WARNER, M. & CHEASTY, T. 2004. Community and hospital spread of *Escherichia coli* producing CTX-M extended-spectrum β-lactamases in the UK. *Journal of antimicrobial chemotherapy,* 54**,** 735-743.

WRIGHT, K. J., SEED, P. C. & HULTGREN, S. J. 2007. Development of intracellular bacterial communities of uropathogenic *Escherichia coli* depends on type 1 pili. *Cellular microbiology,* 9**,** 2230-2241.

WU, C., ZHANG, J., WANG, M., DU, G. & CHEN, J. 2012. Lactobacillus casei combats acid stress by maintaining cell membrane functionality. *Journal of industrial microbiology & biotechnology,* 39**,** 1031-1039.

YANG, G., BILLINGS, G., HUBBARD, T. P., PARK, J. S., LEUNG, K. Y., LIU, Q., DAVIS, B. M., ZHANG, Y., WANG, Q. & WALDOR, M. K. 2017. Time-resolved transposon insertion sequencing reveals genome-wide fitness dynamics during infection. *MBio,* 8.

YASIR, M., TURNER, A. K., BASTKOWSKI, S., PAGE, A. J., TELATIN, A., PHAN, M.-D., MONAHAN, L. G., DARLING, A. E., WEBBER, M. A. & CHARLES, I. G. 2019. A new massively-parallel transposon mutagenesis approach comparing multiple datasets identifies novel mechanisms of action and resistance to triclosan. *bioRxiv***,** 596833.

ZHANG, F., OUELLET, M., BATTH, T. S., ADAMS, P. D., PETZOLD, C. J., MUKHOPADHYAY, A. & KEASLING, J. D. 2012a. Enhancing fatty acid production by the expression of the regulatory transcription factor FadR. *Metabolic engineering,* 14**,** 653-660.

ZHANG, M., ZHOU, Y., LI, T., WANG, H., CHENG, F., ZHOU, Y., BI, L. & ZHANG, X.-E. 2012b. MutL associates with *Escherichia coli* RecA and inhibits its ATPase activity. *Archives of biochemistry and biophysics,* 517**,** 98-103.

ZHAO, B. & HOURY, W. A. 2010. Acid stress response in enteropathogenic gammaproteobacteria: an aptitude for survival. *Biochemistry and cell biology,* 88**,** 301-314.

ZHAO, H., WANG, P., HUANG, E., GE, Y. & ZHU, G. 2008. Physiologic roles of soluble pyridine nucleotide transhydrogenase in *Escherichia coli* as determined by homologous recombination. *Annals of microbiology,* 58**,** 275.

ZHITNITSKY, D., ROSE, J. & LEWINSON, O. 2017. The highly synergistic, broad spectrum, antibacterial activity of organic acids and transition metals. *Scientific Reports,* 7.

ZHOU, L., LEI, X.-H., BOCHNER, B. R. & WANNER, B. L. 2003. Phenotype microarray analysis of *Escherichia coli* K-12 mutants with deletions of all two-component systems. *Journal of bacteriology,* 185**,** 4956-4972.

ZHOU, Y. & GOTTESMAN, S. 1998. Regulation of proteolysis of the stationary-phase sigma factor RpoS. *Journal of bacteriology,* 180**,** 1154-1158.

ZOMER, A., BURGHOUT, P., BOOTSMA, H. J., HERMANS, P. W. & VAN HIJUM, S. A. 2012. ESSENTIALS: software for rapid analysis of high throughput transposon insertion sequencing data. *PloS one,* 7**,** e43012.