CAN TRANSPOSON DIRECTED INSERTION-SITE SEQUENCING BE USED TO PREDICT POSSIBLE OUTCOMES OF EVOLUTION?

by

Mathew Thomas Milner

A thesis submitted to The University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

School of Biosciences
College of Life and Environmental Sciences
University of Birmingham
February 2021

Abstract

Laboratory-based evolution has become a tool that is widely used to understand an organism’s response to stressful environments through linking the genotype to the phenotype. Within laboratory evolution, the role that loss of function mutations play in adaptation is a topic of debate, with recent observations suggesting that adaptive loss of function mutations are a common adaptive strategy. One limiting factor of this technique is that the time taken to conduct a single experiment can be extensive. With these points in mind, we proposed to see if a short term selection experiment on a high density transposon library, using Transposon Directed Insertion-site Sequencing (TraDIS) to analyse the data, would produce results which correlate with those from long-term evolution experiments. Since TraDIS provides a measure of relative contributions to fitness of each gene, in principle it should be possible to use TraDIS to identify genes whose loss of function provides a fitness benefit on a significantly shorter timescale. Previously in our laboratory, five populations of *E. coli* K-12 MG1655 were evolved in a dynamic pH environment by daily passaging over five months in unbuffered LB, starting at pH 4.5. Whole genome resequencing of the final populations and clones revealed many striking similarities in the evolutionary trajectories of these populations. Therefore, to explore the hypothesis that short term selection of a high density transposon library could identify genes that were also found in the five month evolution experiment, an *E. coli* K-12 MG1655 transposon library was constructed and passaged for 10 days under similar conditions as the evolution experiment at both pH 4.5 and pH 7. TraDIS analysis showed that, within these populations, insertions in a few genes had accumulated, suggesting there was a fitness advantage for a strain carrying these insertions. These genes showed a significant overlap with the ones identified in the evolution experiment. These results highlight a possible alternative approach to laboratory evolution when attempting to understand an organism’s response to stress, providing a foundation for future work to explore different conditions.

“ Frodo: I wish the ring had never come to me. I wish none of this had happened

Gandalf: So, do we all who live to see such times, but that is not for them to decide. **All we have to decide is what to do with the time given to us**. “

J. R.R Tolkien - The Lord of the Rings

“Coffee is a way of stealing time that should by rights belong to your older self.”

Terry Pratchett - Thud!

For

Sylvia Rushton

(Nan)

Acknowledgements

I’d firstly like to thank my family, to my mum and dad and sister, for the continuing love and support , and who have helped me through writing during this time. Without them I do not think I would be writing these acknowledgments now. To my nan and grandad, thank you for being there and also supporting me with your unwavering love and encouragement.

To my supervisor Pete Lund, who stepped in as the role of proof-reader and as such this thesis copy was edited by him only for conventions of language spelling and grammar only. I would like to say huge thank you for all the help and guidance within the lab. Also, for giving me the freedom to explore new ideas and for supporting me when it was needed. I look forward to continuing the work with you !

I would also like to thank everyone in T101, to Dana, JG, Jack, Micro-Max, Sammi, Manpreet, Gabby, Dena, Rob, Bakul, Rachel and Shahida, thank you for making this PhD a joyful experience. I must especially mention Kara and Emma for the laughs and the drinks: a pint always kept the work going. To Santosh, for our random biology chats and to Maria for the meals and wanderings. Most importantly to Fatima, aka the ‘boss’ who throughout my PhD has made my time in Birmingham memorable. Finally, to Emily who taught me TraDIS and put up with my questions. Thank you to the students who have worked in the lab on this project.

Finally, to my mates, Joe, Bill, Norman, Dyl, Bobby, James, Tom, and Rosie. Thanks for the drinks, the games, and the company!

Table of contents

Contents of Figures

Contents of Tables

Chapter 1

Introduction

1.1 Prologue

Adaptation in the context of evolution can be described as a process by which a species changes to become fitter in its environment. This definition does not consider the specific 'factors' in an environment that can drive the process of selection once mutation has taken place. One way to consider these factors in the broadest sense is as a combination of **stresses**. A stress can be considered as any factor of physical, chemical, or biological origin within an environment to which a species can potentially adapt. Adaptation can be seen as a response to stress and therefore stress as a driving factor for adaptation.

One method of understanding an organism's response to stress is through laboratory-based evolution (LE). This technique involves culturing an organism under defined conditions over an extended period of time allowing adaption to a defined stress (Dragosits and Mattanovich, 2013, Van den Bergh *et al*., 2018, McDonald, 2019). Upon completion of the experiment, the genotype can then be obtained by using High Throughput Sequencing (HTS). This technique has proven effective in gaining further insight into a variety of stress responses (Table 1.1). However, within LE, when mutations are identified the precise nature that the mutation has upon the function of the gene is typically not considered. Recently, studies have begun to suggest mutations which occur in evolution experiments are more likely to cause a loss of function within the gene (Cooper *et al*., 2001, Gompel and Prud'homme, 2009, McDonald *et al*., 2009, Hottes *et al*., 2013, Lind *et al*., 2015). (See section 1.3.3).

Since 2009, multiple methods have arisen which fall under the term transposon insertion sequencing (TIS). This technique utilizes a high density transposon library coupled with HTS to obtain the location and relative frequency of each unique transposon insertion within a library (Gawronski *et al*., 2009, Goodman *et al*., 2009, Langridge *et al*., 2009, van Opijnen *et al*., 2009). Using this technique allows for genotype to phenotype relationships to be inferred by looking at the frequency and location of insertions within each gene on the genome, to simultaneously assess the effect on fitness of insertions within individual genes (See Section 1.4).

Considering the observation that loss of function mutations can be common within LE, a concept arose that similar results from an evolution experiment could be produced by the outgrowth of a transposon library, but on considerably shorter time frame, this concept is tested in the thesis. LE typically starts with a single clonal isolate, time is then required for mutations to occur and accumulate within a population. In a transposon library, coupled with TIS, a very large number of mutations have already been created and this should reduce the time required to observe the change in frequency of these mutations during an outgrowth experiment (Figure 1.1.1). Although the data gathered from these two experiments will be different, both will identify genes where mutations are present which confer a fitness advantage under stress and if they both cause loss of function, in theory the same genes will be identified (Figure 1.1.1). This would then present a method which could be used to determine, some of the outcomes of an evolution experiment, but on a considerably shorter timeframe than a standard LE experiment.

Previously, an evolution experiment was conducted in unbuffered LB at pH4.5 for several months within our laboratory (Sen, 2018). This experiment aimed to investigate *E. coli*'s acid response to growth under mild pH stress. However the results of this experiment, suggested that the adaptations observed were towards other stresses present within the conditions of the evolution experiment (see Chapter 3). However using this, the aim of this thesis was to consider whether similar results from this evolution experiment could be obtained, using a transposon library created using the same strain. This would then be put under the same conditions as the evolution experiment, but for considerably shorter time. The relative frequency of insertions could then be tracked over time, using the TIS method Transposon directed insertion-site sequencing (TraDIS), and genes associated within insertions causing an increase in fitness, identified (Figure 1.1.1). This would then allow for a comparison of these two techniques to see if similar results could be obtained.

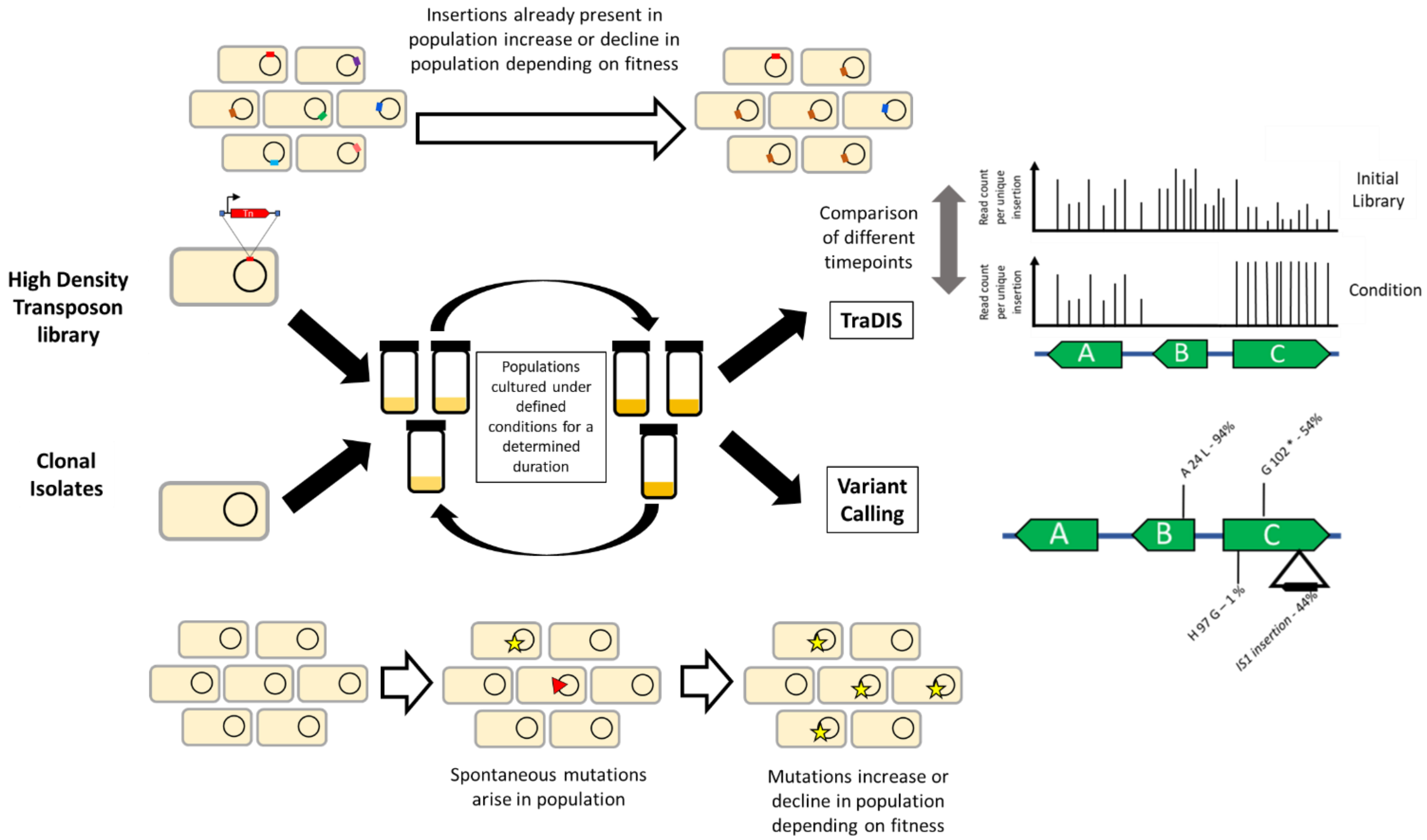

Insertions already present in population increase or decline in population depending on fitness
Tn
High Density Transposon library
Clonal Isolates
Populations cultured under defined conditions for a determined duration
Comparison of different timepoints
TraDIS
Variant Calling
Read count per unique insertion
Initial Library
Condition
A
B
C
A 24 L - 94%
G 102 * - 54%
H 97 G – 1 %
IS1 insertion - 44%
Spontaneous mutations arise in population
Mutations increase or decline in population depending on fitness

Figure 1.1.1: The rationale of this study. A laboratory based evolution experiment using a clonal isolate was performed. During the experiment, spontaneous mutation arise, which if they confer a fitness benefit are selected for and accumilate within the population. High throughput sequencing is then conducted at the end of the experiment to ascertain the genotypes. The genotypes observed indicate mutations within genes which potentially confer an advantageous fitness. In the same strain, a high density transposon library was created, and passaged under the same conditions as the evolution experiment for a considerably shorter time. Transposon Directed Insertion-site Sequencing was then employed to track the frequency of insertions within genes, before and after growth under the selective condition. The effect on fitness of insertions can then be ascertained, and genes identified where insertions have either an **Advantageous** fitness and increase in the population (Gene C) or a **Disadvantageous** fitness and decrease in the population (Gene B). While the former suggests loss of function of a gene increases fitness, the latter shows that the function of the gene is required for fitness. Genes identified in both experiments, can then be compared to see if similar results are found.

1.2 Understanding *E. coli*'s response to acid stress.

Escherichia coli is an enteric organism and can be found within the intestines of most mammalian species. *E. coli* is described as a model organism, due to its simple requirements for growth, quick growth rates, and relative ease to be genetically manipulated. It is these traits and others, which have made *E. coli* one of the most fundamentally understood organisms in biology, with a considerable amount of information coming from experimentation in *E. coli* laboratory strains. *E. coli* laboratory strains have a certain pedigree, with many strains sharing the same ancestry having originated from only a few patient isolates, as well as spending decades being cultured and selected to laboratory conditions (Hobman *et al*., 2007). However, within the postgenomic era, *E. coli* as a species, has been shown to have high diversity with different strains varying in number of genes and genome size and being present within a variety of different environmental niches with laboratory strains only representing a small proportion of the total strains identified (Rasko *et al*., 2008, Horesh *et al*., 2021). This highlights an issue that understanding gained in laboratory strains, may not represent what occurs in strains which are present in the natural environment, and presents the need consider information within the context of the genotype (Hobman *et al*., 2007).

In this study, the laboratory strain *E. coli* K-12 MG1655 was used. This strain is of K-12 origin with its ancestor originating from a diarrhoea sample of a diphtheria patient in 1922, and then subsequently was used as a model organism from the 1940's until present day (Hobman *et al*., 2007). The strain *E. coli* K-12 MG1655 has been considered as a strain of K-12 origin which has had very little genetic manipulation conducted to it with it being subjected only to UV radiation to remove its lambda phage, before being subsequently cured of its F plasmid using acridine orange (Blattner *et al*., 1997, Hobman *et al*., 2007).

In it natural habitat of the gut, *E. coli* routinely encounters acid stress within the gastrointestinal tract, particularly within the stomach which is typically between pH 1 and 4 (Giannella *et al*., 1972, Evans *et al*., 1988, Peterson *et al*., 1989, Lin *et al*., 1995). Due to this, *E. coli*'s response to acid stress has been studied extensively. An evolution experiment conducted by Sen, (2018) in unbuffered LB at pH 4.5 , was a major component of this study. Therefore in this section *E. coli*'s response to acid stress will be outlined in brief, however two reviews that provide a good overview of the organisms response to acids stress are highlighted in Lund, Tramonti and De Biase, (2014) and Guan and Liu, (2020).

1.2.1 The effect that acid has upon *E. coli* K-12

When exposed to acid stress, *E. coli* actively maintains the homeostasis of its cytoplasmic pH within a range of pH 7.2 - 7.8 (Slonczewski *et al*., 1981, Wilks and Slonczewski, 2007, Martinez *et al*., 2012). When *E. coli* is subjected to an environment of pH 5.5 in which the organism is still able to grow, there is a sudden decrease in cytoplasmic pH towards that of the external environment which recovers within the first 30 seconds and by 2 minutes pH homeostasis is maintained (Wilks and Slonczewski, 2007, Martinez *et al*., 2012). At the same time the periplasmic pH is unable to be maintained and eventually becomes acidified, equalising to the extracellular pH (Wilks and Slonczewski, 2007).

The precise mechanism of how protons cross the membrane is unknown. Considering the effect that this has upon the cell, *E. coli* is able to maintain its cytoplasmic pH, as the inner membrane acts as a barrier to the influx of protons, albeit leaky. As such *E. coli*'s response to acid stress is to actively maintain the cytoplasmic pH homeostasis.

1.2.2 Consideration of *E.coli's* acid response

E. coli's response to acid is not a single overarching response. Since acid stress in its broad sense is defined as anything < pH 7, the scale and the degree of acid stress can vary the overall response and phenotype of *E. coli*. Therefore, acid stress has been considered within two types of response: Extreme acid stress (< pH3), in which the *E. coli* acid response is to survive the stress and maintain cytoplasmic pH until a more preferred environment occurs (Gorden and Small, 1993, Lin *et al*., 1995, Cheville *et al*., 1996, Castanie-Cornet *et al*., 1999, Boot *et al*., 2002, Seputiene *et al*., 2006); and mild acid stress (pH 4-6) where *E. coli* is still able to proliferate albeit at a slower rate (Harden *et al*., 2015, He *et al*., 2017, Du *et al*., 2020, Xu *et al*., 2020). Considering these two broad categories, the majority of our understanding of acid response has come from understanding the *E. coli* response to extreme acid stress.

The method used to characterise the *E. coli* response to extreme acid stress involves a survival assay where the bacteria are shocked at a low pH typically for 2 hours, before being plated and a percentage of survival calculated (Gorden and Small, 1993, Lin *et al*., 1996, Castanie-Cornet *et al*., 1999, Boot *et al*., 2002, Seputiene *et al*., 2006). Under these conditions two acid resistance phenotype were observed: one at stationary phase (Gorden and Small, 1993, Castanie-Cornet *et al*., 1999) the other in exponential phase. In order to observe an acid resistance phenotype in exponential phase *E. coli* has to be conditioned by being grown in mild acid stress before being subjected to extreme acid stress (Boot *et al*., 2002, Seputiene *et al*., 2006). The observation of these two different phenotypes suggested that acid resistance mechanisms were potentially regulated by different mechanisms.

1.2.3 The acid resistance (AR) mechanisms

In *E. coli* four major acid resistance mechanisms have been identified towards extreme acid stress. These are termed AR1 – AR4 and have been shown to confer resistance to extreme acid stress in the presence or absence of a specific nutrient. Of the four AR mechanisms, each has been characterised through the use of survival assays, described above. However, in order to be able to distingish between each mechanism the exposure to extreme acid stress is conducted in a minimal media, supplimented or not supplimented with the specific nutrient which defines the AR mechanism. Considering each AR mechanism separately, the AR1 mechanism is the least understood although is shown to be dependent upon RpoS, and induced in the absence of glucose (Cheville *et al*., 1996, Lin *et al*., 1996, Castanie-Cornet *et al*., 1999, Aquino *et al*., 2017). However, the AR2 – AR4 share a similar mechanism in that all are dependent on an external source of amino acids to be present in the environment for an acid resistance phenotype to be observed. Each AR mechanism is dependent on a single amino acid, these are glutamate (AR2), arginine (AR3) and lysine (AR4) (Gale and Epps, 1942, Lin *et al*., 1995, Castanie-Cornet *et al*., 1999). A further system was identified using ornithine (ODAR) which is also present in *E. coli* (Kashiwagi *et al*., 1991, Aquino *et al*., 2017). Each of these AR systems are able to maintain cytoplasmic pH as each have an amino acid decarboxylase that removes a proton from the cytoplasm by the decarboxylation of its specific amino acid to form a product and CO_2 (Figure 1.2.1)(Kashiwagi *et al*., 1991, Castanie-Cornet *et al*., 1999, Foster, 2004). This product is then removed from the cytoplasm by a product/substrate antiporter in the inner membrane (Figure 1.2.1) (Kashiwagi *et al*., 1991, Castanie-Cornet *et al*., 1999, Foster, 2004, Aquino *et al*., 2017). Although the genes involved in the mechanisms have been identified the regulation of these mechanisms has yet to be fully characterised. A brief description of these mechanisms will be given below.

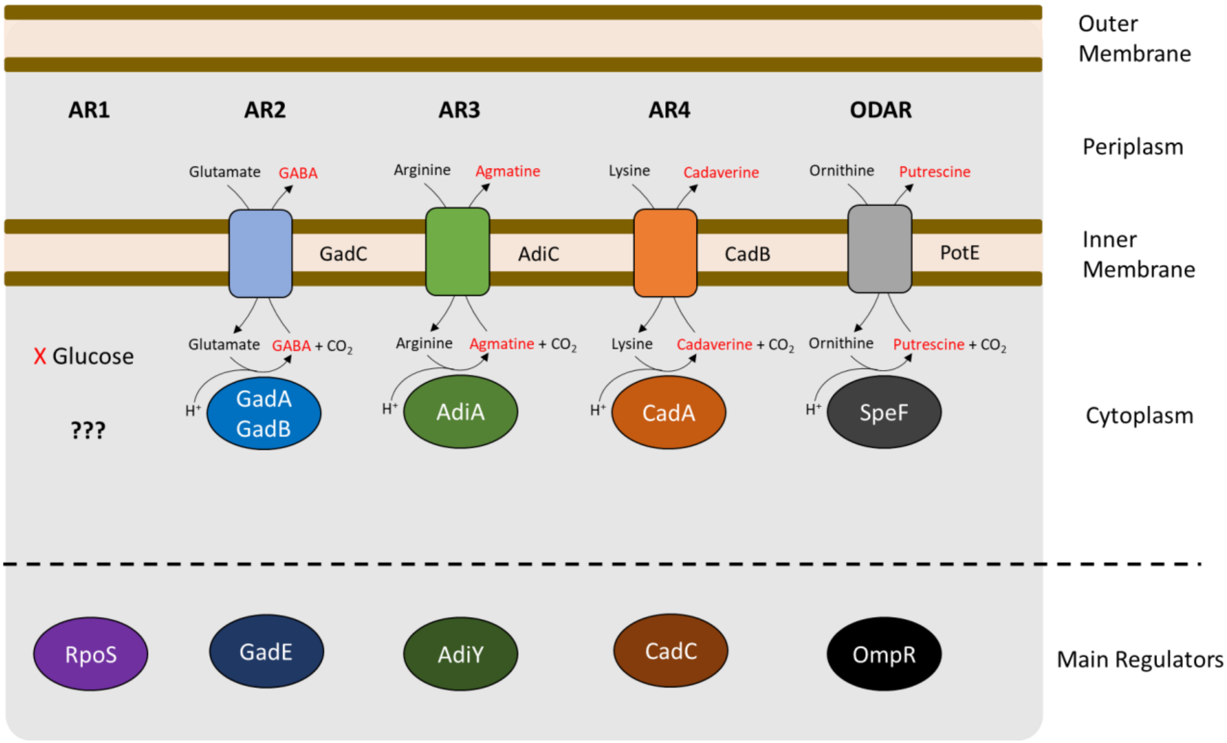


Figure 1.2.1: Overview of the Acid Resistance systems (AR1-AR5). AR1, not much is known about this mechanism, only that it occurs in the absence of glucose and is RpoS dependent. AR2-AR4 and ODAR are termed amino acid dependent acid resistant system highlighting the decarboxylase and product/substrate antiporters which constitute these AR systems. The reactions of these systems are illustrated. The main regulator of each AR mechanism is also highlighted . Adapted from (Kanjee and Houry, 2013).

1.2.3.1 AR2 – Glutamate dependent acid resistance.

The most effective acid resistance mechanism at extreme acid stress is the AR2 system. This has been shown to confer the highest survival rates when cells are subjected to extreme pH (Lin *et al*., 1995, Castanie-Cornet *et al*., 1999). The regulation of the AR2 system is also the most understood. The function is performed by two decarboxylases, GadA and GadB, and a Glutamate/GABA antiporter, GadC (Castanie-Cornet *et al*., 1999) (Figure 1.2.1). In turn *gadA* and *gadB/C* are regulated by GadE , the main regulator of the AR2 system, in a heterodimer with RcsB, targeting a sequence called the *gad* box just upstream of the translational start site (Ma *et al*., 2003, Hommais *et al*., 2004, Castanié-Cornet *et al*., 2010, Johnson *et al*., 2011, Aquino *et al*., 2017).

Continuing from *gadE,* the regulation of the AR2 system is extensive, presenting a highly complex network of regulation controlling for the appropriate response when subjected to acid stress. Considering the regulation involved in the AR2 response two two-component systems (EvgAS and PhoPQ*)* are both involved (Ma *et al*., 2004, Burton *et al*., 2010, Eguchi *et al*., 2011, Johnson *et al*., 2014). Additionally several genes are involved in regulation, including *gadX, gadW, ydeO, ydeP,* and the small regulatory RNA *gadY* (Masuda and Church, 2003, Opdyke *et al*., 2004, Sayed *et al*., 2007, Burton *et al*., 2010, Johnson *et al*., 2014, Aquino *et al*., 2017). Regulation of the AR2 varies based on whether the cells are in stationary phase, or conditioned at pH 5.5 when in exponential phase(Castanie-Cornet *et al*., 1999, De Biase *et al*., 1999, Ma *et al*., 2004, Burton *et al*., 2010, Johnson *et al*., 2014).

1.2.3.2 AR3 – Arginine dependent acid resistance

The main components of the AR3 has been identified as AdiA which is the arginine decarboxylase, and AdiC as the arginine/agmatine antiporter (Castanie-Cornet *et al*., 1999, Iyer *et al*., 2003). In the AR3 system, the main regulators are the transcription factors AdiY and CysB which have been shown to regulate the gene *adiA* in complex medium at low pH, under anaerobic conditions (Shi and Bennett, 1994, Stim-Herndon *et al*., 1996). This regulation is not yet fully understood.

1.2.3.3 AR4 – Lysine dependent acid resistance

The major components of this acid resistance mechanism are the lysine decarboxylase encoded by *cadA*, and its lysine/cadaverine antiporter, encoded by *cadB* (Soksawatmaekhin *et al*., 2004, Watson *et al*., 1992), with the main regulator being CadC (Kuper and Jung, 2005, Neely *et al*., 1994). CadC is a membrane integrated protein, believed to able to sense low pH environmental conditions required for activation (Neely *et al*., 1994, Kuper and Jung, 2005, Kanjee *et al*., 2011). In the absence of lysine, LysP inhibits *cadC* (Neely *et al*., 1994, Tetsch *et al*., 2008). In addition the alarmone, ppGpp, part of the stringent starvation response, was discovered to co-crystalise with CadA, and further analysis revealed that ppGpp was able to inhibit function of CadA (Kanjee *et al*., 2011). Although the AR4 mechanism has been shown to confer resistance towards extreme acid stress, this mechanism is also known to provide resistance under mild acid conditions (Kanjee *et al*., 2011).

1.2.4 Ammonia production: The role of deaminases

Another acid resistance mechanism is through the use of deaminases; these enzymes function by removing the amine group present upon compounds to produce ammonia (NH_3) which, in an acidic environment, will be rapidly protonated to create ammonium thus increasing pH. One mechanism which uses a deaminase is termed AR2_Q. This was identified when acid resistance was observed when *E. coli* in minimal medium supplemented with glutamine was subjected to extreme acid stress (Lu *et al.*, 2013, Pennacchietti *et al.*, 2018). This system involves the import of glutamine by the AR2 antiporter (GadC) and the action of a glutamine deaminase (YbaS). This was able to produce glutamate and ammonia from glutamine, and this confers an acid resistance phenotype. The by-product glutamate is then used in the AR2 response, further contributing to the acid resistance phenotype.

Other deaminases have been found which have been shown to confer an acid resistance phenotype in extreme acid conditions, such as with adenosine. Where the adenosine deaminase, *add,* was shown to confer an acid resistance phenotype (Sun *et al.*, 2012). Considering that these studies only focus upon the role of one metabolite to confer an acid resistance phenotype, such as the examples above, this suggests that all deaminases within *E. coli* potentially have the ability to confer an acid resistance to some extent.

1.2.5 pH Stress - Concluding remarks

In the above sections only a few of the possible mechanisms of *E. coli*'s response to acid stress have been noted. There are others, such as HdeA (Gajiwala and Burley, 2000) and HdeB (Kern *et al.*, 2007), periplasmic acid stress chaperones which protect proteins from acid stress within the periplasm. The membrane composition of *E. coli* also changes under acidic conditions, with the amount of cyclopropane fatty acids increasing upon induction at pH 5.5 (Chang and Cronan, 1999). In addition, an increase production of unsaturated fatty acid was observed when cells were subject to pH 4.2 (Xu *et al.*, 2020). An increase in enzyme activity in the TCA cycle was also observed after 2 hours of exposure to mild acid stress (Jain *et al.*, 2013).

The mechanisms mentioned above were mainly identified in the *E. coli* response to extreme acid stress. However, it is important to note that some of the mechanisms will also play a role in *E. coli* response to mild acid stress, such as the AR4 response. However, it also highlights the point, that in regard to *E. coli* response to mild acid stress, a lot is still to be discovered. In 2017, a lab based evolution experiment conducted at pH 4.6 - 4.8 , showed deletions within the main regulators of the AR systems (AR 2-4), effectively removing the activity of these systems (He *et al*., 2017), suggesting that some systems for extreme acid stress were not required in a response to constant exposure of mild acid stress, and alternative mechanisms for resistance were present.

1.3 Laboratory based evolution

Laboratory based evolution (LE) involves using evolution to adapt a population to defined conditions for a prolonged period of time within a laboratory setting. As the amount of generations increase, selection will occur, which will lead to a population becoming better adapted to these conditions (Figure 1.3.1). Typically, once an experiment is concluded, HTS is employed to identify the mutations associated with the phenotype of the adapted populations (Figure 1.3.1d). As the time taken for these experiments is dependent on generation time, microbes are ideal and are frequently utilized within these experiments due to their quick growth rates, ability to survive on simple nutrients, and their relative ease of handling. These organisms can be cryopreserved, creating a 'fossil record' which allows mutations to be tracked over time.

The time an evolution experiment is conducted for is dependent on the user. One experiment which shows the extent of LE, is the Long Term Evolution Experiment (LTEE) conducted within the laboratory of Richard Lenski. This experiment started in 1988 and is ongoing. It involves the daily passaging of 12 independent populations of the *E. coli* strain REL606 in minimal media supplemented with glucose (Lenski *et al*., 1991, Lenski and Travisano, 1994, Good *et al*., 2017, Barrick *et al*., 2009, Blount *et al*., 2008, Blount *et al*., 2012). This experiment has provided useful insights in understanding concepts within evolution, such as historical contingency, where a mutation's contribution to fitness is dependent upon previous mutations having arisen (Blount *et al*., 2008, Blount *et al*., 2012); and diminishing returns of beneficial mutations, where over time the extent of a mutation's contribution to fitness will decline in respect of the previous mutational background (Barrick *et al*., 2009)(Figure 1.3.1c). Three recent reviews that summarise the current understanding and uses of LE are Van den Begh, (2018), Dragosits and Mattanovich, (2013) and McDonald, (2019).

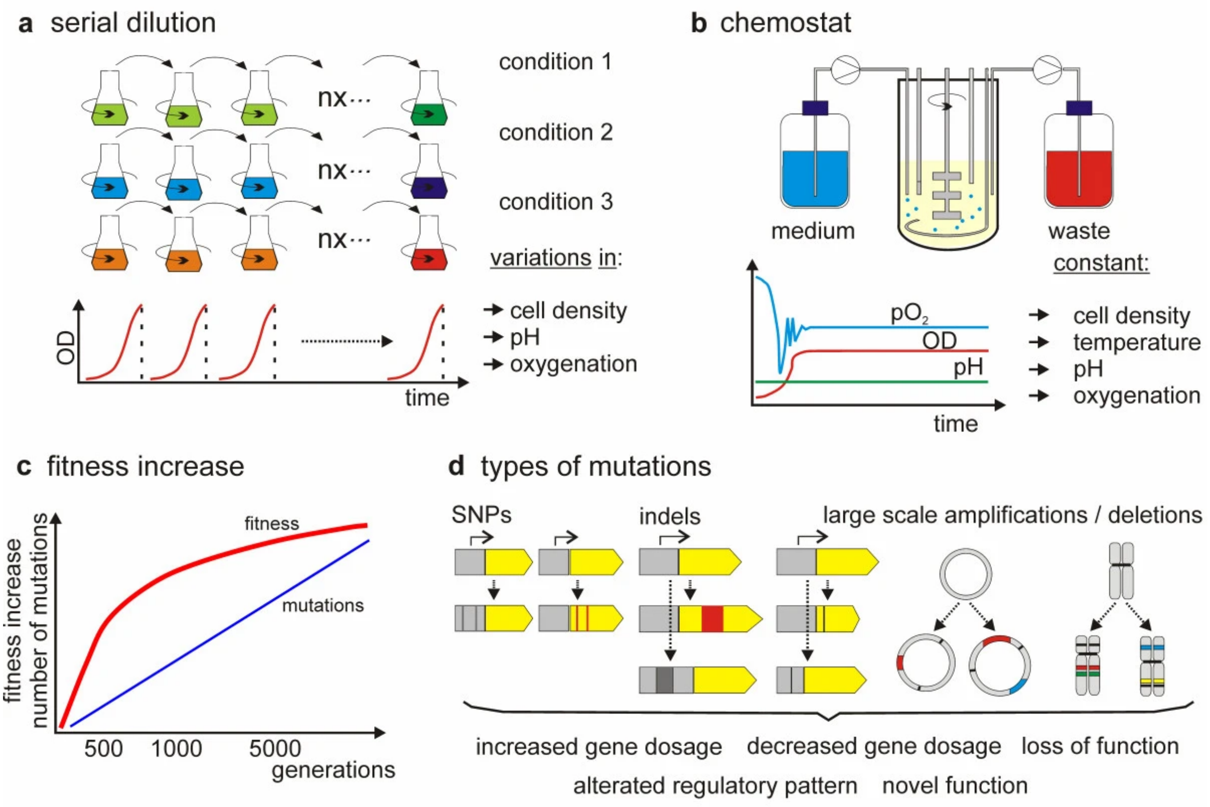


Figure 1.3.1: Overview of Laboratory based evolution**.** Two methods for conducting LE are shown in a and b. a): batch culture (serial dilution). Populations are grown in a vessel which, as the population grows the dynamics of the environment will change, such as nutrients and cell density. Populations are repeatedly passaged into fresh medium and the process repeated. b) Alternatively a chemostat can be used to maintain a constant environment during evolution. c) Over the duration of Diminishing returns of beneficial mutations: the observation that during experimental evolution mutations arise at a steady rate, however the mutational contribution to fitness, decline over time due to the changing mutational background. d) Types of mutations which are identified during experimental evolution. Taken from (Dragosits and Mattanovich, 2013).

1.3.1 Laboratory based evolution as a tool to understand stress.

Although LE has been used as a method to explore concepts within evolution, it can also be used as a tool to understand how an organism responds to stress. By performing LE using a stress this will create a selection pressure which a population will become adapted to. Once an evolution experiment is completed, an evolved population can then be used to understand how it adapted to the stress. This usually requires further experimentation and can involve using HTS to identify the mutations associated with the adaptation within the population, or clonal isolates taken from the population, and attempts to link the genotype to phenotype.

Using *E. coli,* multiple evolution experiments have been used to explore and understand adaptation to multiple conditions, some of which are highlighted within Table 1.1. When considering using LE as a tool to understand stress however, it is important to highlight the overall experimental design and how this can affect the outcome of LE. When performing LE it is important to note the whole environment will impose multiple selection pressures which will affect the outcome. Therefore, potentially adaptions which occur within LE may not be towards the target stress, but to other stresses introduced by the experimental design. Common factors such as the choice of media can play a role in driving selection and ultimately influence adaption (Knöppel *et al*., 2018). Even the volume of culture used has been shown to cause differences in the outcome of the evolution experiment (Gross *et al*., 2020).

Table 1.1: Examples of Laboratory based evolution experiments using *E. coli* under different conditions.
* Unless otherwise stated the length of the evolution experiment was estimated using Total generations/ estimated generations per day stated within the study.

Citation	Strain Used	Length of experiment	Selective Condition
Johnson *et al*. (2014)	*E. coli* MG1655	21 days	**Understanding E. *coli* response to extreme acid stress.** Grown in 5ml LB overnight at 37°C. Diluted to an 0.01 OD_{600} and grown to an 0.2 OD_{600}. A log fold dilution of culture was performed, and each tested for survival at pH 2.5 for 2 hours. Passaged at a 1: 50 dilution, Next day used the smallest log fold dilution where growth was observed.

Citation	Strain Used	Length of experiment	Selective Condition
Harden *et al*. (2015), He *et al*., (2017)	*E. coli* BW3110	2000 generations 250 days*	**Understanding *E. coli* response to mild acid stress.** Grown for ~24 hours at 37^{o}C in 200µl LBK buffered with HOMOPIPES. First 730 generations at pH4.8 using a 1:4000 dilution, the remaining 1270 generations at pH 4.6 passaged at a 1:100 dilution.
Hughes *et al*. (2007)	*E. coli* REL606 *E. coli* REL607	2000 generations 300 days*	**Understanding *E. coli* response to pH and the tradeoff observed.** Cultures grown in 10ml Davis minimal media supplemented with glucose at either pH 5.3, pH 6.3, pH 7 and pH 7.8. Grown ~ 24 hours at 37^{o}C passaged daily at a 1:100 dilution.
Goodarzi *et al*. (2010)	*E. coli* MG1655	80 generations	**Understanding *E. coli* response to Ethanol stress.** Cultures in M9 minimal media supplemented with glucose and 7% v/v ethanol. Grown at 37^{o}C for ~24 hours passaged daily at a 1:100 dilution
Minty *et al*. (2011)	*E. coli* EcNR1	430 - 500 generations 180 days	**Understanding *E. col*i isobutanol tolerance.** Cultures passaged at mid exponential phase to achieve a 0.002 OD$_{600}$ in 0.75% v/v isobutanol in 3ml of NG50 (glucose as carbon source) or NX50 medium (xylose as carbon source).. Grown at 37^{o}C
Riehle *et al*. (2001), Riehle *et al*. (2003)	*E. coli* REL606 *E. coli* REL607 after 2000 generations of the LTEE	2000 generations 300 days*	**Understanding thermal adaptation in *E. coli*.** Cultures grown in 10ml Davis minimal media supplemented with glucose at 41.5^{o}C for ~24 hours. Passaged daily at a 1:100 dilution.
Lenski *et al*. (1991), Lenski and Travisano (1994), Blount *et al*. (2008), Barrick *et al*. (2009), Blount *et al*. (2012), Good *et al*. (2017)	*E. coli* REL606 *E. coli* REL607	70,000 generations. This experiment is still ongoing	**The Long Term Evolution Experiment.** Cultures of 10ml Davis minimal media supplemented with glucose. Grown for ~24 hours at 37^{o}C passaged at a 1:100 dilution.

Citation	Strain Used	Length of experiment	Selective Condition
Bennett and Lenski (2007)	*E. coli* REL606 *E. coli* REL607 after 2,000 generations of the LTEE	2000 generations 300 days*	**The effect of temperature upon *E. coli*.** Temperature differences, evolved at 32°C, 37°C or 42°C or a daily alteration between 32 °C and 42°C. In addition, a further 2000 generations at 20°C was conducted for each temp condition. Each grown for ~24 hours in 10ml Davis minimal media supplemented with glucose. Passaged at a 1:100 dilution and grown.
Sleight and Lenski (2007), Sleight *et al*. (2008)	*E. coli* REL606 *E. coli* REL607 After 20,000 generations of the LTEE	1000 generations 300 days	**Understanding the effects of freeze thawing.** Culture frozen at -80 °C for 22.5 hours, thawed at room temperature for 1.5 hours. Passaged at a 1:100 dilution and grown for 24 hours at 37°C. The cycle was repeated 150 times. Grown in 10ml Davis minimal media supplemented with glucose.
Sen (2018). This study	*E. coli MG*1655	740 generations 150 days*	***E. coli* response to mild acid stress.** Grown in 5ml unbuffered LB at 37°C at pH 4.5 for ~24 hours. Passaged at a 1:20 dilution.
LaCroix *et al*. (2015)	*E. coli* MG1655	31 -89 days	***E. coli* MG1655 response to growth on M9 minimal media.** Growth at pH 7 in 25ml M9 minimal media supplemented with glucose at 37°C. Passaged at a 1:30 ratio when culture was in mid-exponential phase.
Du *et al*. (2020)	Adapted *E. coli MG1655* taken from study (LaCroix *et al*., 2015)	800 generations 35 days	***E. coli* acid response to mild acid stress**. Cultured at pH 5.5 in 15ml M9 minimal media supplemented with glucose at 37°C. Passaged at a 1:5 dilution when the culture was in mid-exponential phase.

Citation	Strain Used	Length of experiment	Selective Condition
Gross *et al*. (2020)	*E. coli* MG1655	64 days	**Exploring adaptations to long term stationary phase and the affect that culture volume plays**. Cultures at 4ml, 40ml and 400ml volume grown in LB at 37^{o}C. No passaging occurred. Same volume to vessel ratio used.
Kram *et al*., (2017)	*E. coli* MG1655	300 generations 120 days	**The effects of long term serial passage in a Complex medium.** Cultures of 12.5ml of LB grown at 37^{o}C were passaged at a 1: 1000 dilution every day or every 4 days.
Hamdallah *et al*. (2018)	*E. coli* BW3110	2200 generations 333 days*	***E. coli* adaptation to high pH.** Cultures of 200µl LBK media at pH 9 – 9.3 buffered with TAPS for ~24 hours at 37^{o}C. Passaged at a 1:100 dilution daily. pH increased over time.
Knöppel *et al*. (2018)	*E. coli* MG1655 *S. enterica*	1000 generations 100 days *	**Adaption to different media**. 1ml Cultures of LB MH and M9 media supplemented with either glucose or glycine. Passaged at a 1:1000 dilution every 24 hours.

1.3.2 Laboratory based evolution to understand low pH

To the knowledge of this study only four evolution experiments have been conducted on adaptation to acid stress on *E. coli*. These included one experiment focusing on extreme acid stress at pH 2.5 and three under mild acid stress, at pH 4.6 – 4.8, pH 5.3 and pH 5.5 (Johnson *et al*., 2014, Harden *et al*., 2015, Hughes *et al*., 2007, He *et al*., 2017, Du *et al*., 2020). The length of these experiments ranged from three weeks to 5 months using different nutrient media and methods to conduct the LE experiment (Table 1.1). These experiments are further considered below

The first LE experiment using acid stress was performed by Hughes *et al.,* (2007). This study evolved populations of *E. coli* REL606/REL607 over 2000 generations at 4 different pH conditions (pH 5.8, 6.3, 7 and 7.8). Using adapted clonal isolates from each pH condition Hughes *et al.,* (2015) demonstrated that a trade-off existed where acid adapted strains showed a fitness disadvantage to alkaline conditions (Hughes *et al*., 2007). This observation was also identified in another mild acid stress evolution experiment conducted by Harden *et al.*, (2015) (Harden *et al*., 2015).

In the other three evolution experiments, HTS was employed to identify mutations within the populations. The first was an evolution experiment initially aimed to be conducted at pH 4.6, however the strain used (*E. coli* BW3110) did not grow at this pH. Therefore, the experiment was conducted in LBK media at pH 4.8 initially, and then dropped to pH 4.6 once adaption had begun. In this experiment the functions of the amino acid dependent decarboxylases, (involved in AR2-AR4) were shown to be lost in adapted clonal isolates (Harden *et al*., 2015, He *et al*., 2017). RNA-seq on these strains confirmed that the decarboxylases were downregulated; however, an upregulation of genes associated with anaerobic catabolism and transport was identified (He *et al*., 2017). Mutations were also noted in the RNA polymerase subunits (*rpoB* and *rpoC*), and He *et al.*, (2017) showed that mutations in these subunits had an increase in growth rate at pH 4.6 (Harden *et al*., 2015, He *et al*., 2017). The results from this evolution experiment demonstrated that potentially other acid resistance responses existed and that some of the previously identified responses to extreme acid stress are not required and may be disadvantageous under prolonged growth at mild acid stress.

Another evolution experiment was conducted by Du *et al*. (2020) at pH 5.5 in minimal media. In this study an attempt was made to keep these cells within exponential phase for the duration of the LE experiment (Harden *et al*., 2015). Within this experiment, clonal strains were isolated from 6 independent populations. Interestingly, in 5 out of 6 strains isolated mutations within *rpoC* were observed*.* In addition, when RNA-seq was performed on these strains global differences were observed indicating an increase in expression of genes involved within amino acid transport and central metabolism, suggest that this may play a role to improved growth under these conditions.

From these studies, no specific mechanism to mild acid stress has yet been identified. Although both identify mutations in *rpoC,* the precise effects of these mutations have not been deciphered, although they improve growth rate. Interestingly, Harden *et al.* also showed downregulation of the AR 2-4 mechanisms, again suggesting that these do not play a role in prolonged exposure to mild acid stress. However, these results suggest that there is still a lot to learn about adaptation to mild acid stress.

1.3.3 The effect of mutations on gene function

During a LE experiment, when mutations occur within a population they can affect the fitness phenotype of the individual. As such mutations can be classed into three categories depending on how they affect the individual's fitness phenotype, either: advantagous, disadvantageous or neutral. If a mutation elicits an advantagous fitness phenotype then selection can act upon this causing the mutation to increase in frequency within the population, with the potential to reach fixation. Alternatively, if a mutation elicits a disadvantageous fitness phenotype, then selection can act to remove this mutation from the population. However, if a mutation is neutral, then it will have no effect on the individual's fitness meaning that selection will not act upon it. Instead, these mutations may alter in frequency in the population due to other processes, for example genetic drift. Ultimately this leads to the conclusion, that at the end of a normal LE experiment, whilst disadvantageous mutation will not be seen, typically mutations observed at high frequency in a population are most likely to be those that elicit an advantagous phenotype, though neutral mutations may also be present. Further to this a recent study, underlining the basic mathematics associated with a population undergoing LE, highlighted that the chances that a neutral mutation will occur at a detectable frequency due to genetic drift only in short-term LE experiments lasting only a few weeks or months, > 3,322 generations ,is actually quite slim (Cooper, 2018). This suggests that the majority of mutations observed at high frequency in STSE LE are advantageous mutations on which selection has acted.

What the description above does not cover is by what mechanism can a mutation cause a phenotypic change in fitness. Briefly, a genome contains two types of regions: genes, and intergenic regions. While genes are considered the functional units of the genome, encoding protein or RNA, intergenic regions also play role, in particular because contain regulatory elements required for expression of the gene. Typically, a mutation causes a change in phenotypic fitness, by affecting a gene's function or its expression and as such, the effect of a mutation upon gene function can be described in three ways: (i) No effect, in that a mutation does not change the gene's function or expression pattern; (ii) Gain of function, which describes mutations that elicit new or improved activity of a gene's function or its expression; (iii) Loss of function where mutations stop a gene from functioning or from being expressed. While “no effect” mutations are mainly associated with neutral fitness mutations, gain of function and loss of function mutations may have advantageous, disadvantageous, or neutral impacts on fitness. The frequency of loss of function mutations is likely to be higher than gain of function mutations (Gompel and Prud'homme, 2009, Hottes *et al*., 2013, Murray, 2020, Monroe *et al*., 2021). This is due to the inherent nature of a gain of function mutation, which requires the mutational target to be within a specific region of a gene or intergenic region and to be a specific mutational change that improves or alters a gene’s function. For example, in Johnson *et al*. (2014) where nonsynonymous mutation in *evgS*, the sensor component of the two component system EvgAS, caused an amino acid change of serine to isoleucine within the EvgS PAS domain, constitutively activating EvgS and subsequently the AR2 mechanism, conferring an advantagous fitness under extreme acid stress. While for loss of function mutations, the nature of the location and type of mutation is far less constrained, as it is easier to disrupt a function then to alter it.

However, a question still remains as to what contribution each type of mutation has in conferring an advantagous fitness phenotype. While gain of function mutations have mainly been associated with fitness advantage, the idea that loss of function mutations might elicit an advantageous fitness phenotype has changed over the years. In the latter half of the 20th century, it was considered that loss of function mutations would always cause loss of fitness. However at the turn of the century, this view changed with the concept of ‘less is more’ the idea that deletion of a gene and the removal of its function can be associated with an fitness advantage for the organism overall (Olson, 1999). More recently, this has been furthered with arguments suggesting that loss of function mutations are a more common adaptive strategy than gain of function mutations (Murray, 2020, Monroe *et al*., 2021). Therefore, the following section will consider the role that loss of function mutations can have within evolution experiments, by causing a fitness advantage.

1.3.3.1 Considering loss of function mutations occurrence in *E. coli*

Within the literature, multiple LE experiments have been conducted in *E. coli* under a variety of selection pressures, as illustrated by Table 1.1. Typically, after these experiments have been conducted, whole genome resequencing is employed to identify mutations which may contribute to an adaptive phenotype observed within a strain or population. However, investigating the effect that a mutation can have upon an adaptive phenotype differs within each study. To highlight this, the outcomes of laboratory evolution experiments in *E. coli* under a variety of conditions, were collated by a master's student, Melissa Lawton, and myself and summarized in Table 1.2. This highlighted the differences observed in investigating the adaptive phenotype seen after an evolution experiment, as while some LE experiments focus upon phenotype observed in the adapted strains/population (Toprak *et al*., 2011, Knöppel *et al*., 2017), other studies focus upon linking the genotype to the phenotype by investigating the individual effects that mutations have upon the phenotype. However while some studies only focus upon the individual phenotypic effect a mutation contributes to an advantageous phenotype (Conrad *et al*., 2009, Harris *et al*., 2009), others continued this further by investigating the mechanistic effects of a mutation on the molecular level and how this in turn contributes to an advantagous phenotype (Lee and Palsson, 2010, Johnson *et al*., 2014).

Therefore, when considering the role that loss of function mutations have in contributing to advantagous phenotypes in laboratory evolution experiments, a challenge is presented due to how individual studies characterise mutations. Individual examples where loss of function mutations can be identified include the LTEE, where in the first 2000 generations loss of ribose catabolism due to an IS150 insertion upstream of the *rbs* operon was shown to cause a slight fitness advantage in all 12 populations, (Cooper *et al*., 2001). Examples of other loss of function mutations can also be found within Table 1.2. However, attempting to understand how frequent loss of function and gain of function mutations are overall within LE experiments can be difficult. This is further exacerbated by the fact that most evolution experiments do not characterise all mutations identified, in part due to the high level of effort required, with many instead focusing upon only a few key mutations within genes, typically focused upon due to a large degree of parallelism (i.e., mutations in a particular gene are observed in several independent evolving populations) or *a priori* knowledge of a particular gene (Table 1.2).

Table 1.2: Overview of Laboratory evolution experiments in E. coli K-12 MG1655. Unique mutations which have been identified in the experiment have been highlighted and experimental validation of mutations have been outlined, including whether the mutations identified are associated with Gain of Function (GoF) or Loss of Function (LoF) mutations.

Reference	**Experimental Design**	**Experimental duration** What was sequenced ?	**Gene**				**Intergenic**			Large deletion	Duplication	**Experimental characterisation of mutations**
					Point mutations							
			INDEL	IS insertion	Nonsynonymous	Synonymous	Intergenic	INDEL	IS insertion			
Harris *et al*. (2009)	Stationary phase culture of *E. coli* K-12 MG1655 was exposed to ionising radiation to cause 1% survival. Irradiated cultures were then recovered by inoculation into fresh LB broth and grown overnight.	**20 cycles of selection** 9 strains (7/1/1 strains each from an independent populations)	1	0	108	59	30	1	0	1	0	Used survival assays, confirmed that evolved strains had a higher survival than WT. Used survival assays to suggest that the following individual mutations conferred resistance to the selective pressure: • Large deletion of the e14 prophage. • Two nonsynonymous mutations in *recA*
Johnson *et al*. (2014)	*E. coli* K-12 MG1655 were grown to an OD600 of 0.2 in LB broth at 37°C. Bacteria was then 10-fold serially diluted into LB adjusted to pH 2.5 for 2 hours. These cultures were diluted into fresh LB and grown overnight at 37°C. Highest dilution with growth was then used for the next passage.	**21 cycles of selection** 4 strains (Each from an independent population)	0	0	17	3	0	0	0	0	0	Used survival assays to show that evolved strains had a higher survival than WT. Used promoter reporter constructs, survival assay and protein modelling, to provide evidence that 4 nonsynonymous mutations observed in *evgS* conferred better survival when compared to the WT. Also, that mutations in *evgS* were GoF causing EvgS to be constantly active, activating the AR2 response.

Reference	Experimental Design	**Experimental duration** What was sequenced ?	**Gene**				**Intergenic**			Large deletion	Duplication	**Experimental characterisation of mutations**
			INDEL	IS insertion	**Point mutations**			INDEL	IS insertion			
					Nonsynonymous	Synonymous	Intergenic					
Harden *et al.* (2015), He *et al.* (2017)	*E. coli* K-12 BW3110 grown in LBK media buffered at pH 4.6 at 37°C. Passaging every day into fresh media at a 1in 4000 dilution.	**2000 generations** 8 strains (2 strains each from an independent population)	4	6	20	4	14	1	2	7	0	Used competition experiments to confirm evolved strains used are fitter than WT under selective conditions Used Møller broth tests of the evolved strains compared to the relevant gene deletion mutation to provide evidence that mutations observed in evolved strains were LoF mutations. Evidence provided suggests that the following are LoF mutations. • Large deletion mutation seen in the acid fitness island. • Two large deletions and IS5 insertion within *cad*. • A nonsynonymous mutation and IS5 insertion upstream of *adiY*. Recreated mutation isolated in essential gene, *rpoC* within the WT background. Showed that arginine decarboxylase activity was affected in *rpoC* mutant compared to WT. Suggesting that *rpoC* mutations observed were GoF mutations.
Hamdallah *et al.* (2018)	*E. coli* K-12 W3110 were cultured in LBK at pH 9 were serially cultured for 470 generations. After cultures were passaged at pH 9.2 in LB for 1200 generations. With the pH increasing to pH 9.3 for a further 530 generations.	**2200 generations** 8 strains (2 strains each from an independent population)	4	9	23	1	6	2	1	2	0	Used growth curves, show that evolved strains were able to grow better under selective conditions compared to WT Used growth curves and gene deletions in WT. To suggest that certain mutations were associated with LoF mutations: • IS1 insertion in *phoB* • A nonsynonymous mutations in *pcnB* and *ydcL* Using the same method also showed that gene deletion strains had no effect on growth rate (*yahO*, *mpl*, *ompT*, *slt, yfcD*) or reduced growth rate (rpoS) suggesting mutations within these genes could be potentially neutral or GoF mutations.

Reference	Experimental Design	Experimental duration What was sequenced ?	Gene					Intergenic		Large deletion	Duplication	Experimental characterisation of mutations
					Point mutations							
			INDEL	IS insertion	Nonsynonymous	Synonymous	Intergenic	INDEL	IS insertion			
Wang *et al*. (2010)	Experiment aims to adapt *E. coli* to phosphate limiting conditions. *E. coli* K-12 MC4100TF was grown overnight in T-salts and inoculated into an 80-ml chemostat containing T-salts, glucose and 30µM KH_2PO_3. The bacteria concentration in the chemostat was maintained through 37 days.	**37 days** 5 strains each showing a different colony phenotypes from a single population.	1	1	8	0	3	0	0	0	0	Using growth curves, proteomics, biolog plates and survival to osmotic and oxidative stress, differences were observed in the evolved strains compared to the ancestor. However, a definitive improved fitness phenotype could not be identified across the strains in all assays. Single nonsynonymous mutations in *hfq spot* and 3 mutations in *rpoS* were individually recreated into wildtype background and also removed from relevant evolved strains where mutation was present. Competition experiments showed that these mutations were all fitter than the WT under selective conditions and lost fitness when cured in the evolved strains.
Creamer *et al*. (2017)	*E. coli* K-12 W3110 cultures were grown in LBK media supplimented with benzoate at 37°C with 100-fold dilution into fresh media each day. The concentration of benzoate was increased throughout the duration of the experiment. Experiment was designed to adapt strains to	**2000 generations** 8 strains (2 strains each from an independent population)	12	20	21	5	11	3	8	5	0	Used endpoint OD_{600} after 16 hour to show that evolved strains grew better than WT under selective conditions. Used MIC assays of chloramphenicol with salicylate and gene deletion mutations to suggest mutations in genes associated with the mar operon, were LoF: • Large deletion in including *marRAB* operon. • Nonsynonymous point mutations in *rob* and *cpxA* regulators. Using GABA assays, showed that 2 large deletions in the gad acid fitness island removed the activity of the glutamate decarboxylase (AR2).

Reference	Experimental Design		Experimental duration What was sequenced ?	Gene				Intergenic			Large deletion	Duplication	Experimental characterisation of mutations
						Point mutations							
				INDEL	IS insertion	Nonsynonymous	Synonymous	Intergenic	INDEL	IS insertion			
Du *et al.* (2020)	Culture of *E. coli* K-12 MG1655 were independently grown at 37°C in M9 media supplemented with glucose at pH 4.6..Cultures were maintained at exponential phase continuously using an automated system.		**800 generations** 6 strains (Each from an independent population)	4	3	10	1	3	0	1	0	0	Used growth curves, to show that evolved strains had a higher growth rate than WT. RNA-seq was performed on two evolved strains. With one strain containing only a nonsynonymous mutation in essential gene, *rpoC*. RNA-seq suggests that rpoC mutation is GoF altering the transcriptome compared to the WT.
Knöppel *et al.* (2017)	Experiment aim was to profile evolved strains for antibiotic resistance adapted in the absence of antibiotics *E. col i* K-12 MG1655 were evolved within 4 different growth media (LB, MH, M9 + glucose (GLU) and M9 + glycerol(GLY). Each population was grown for 24 hours at 37°C passaging daily by 1000-fold dilution..	LB	**500 generations** 10 independent populations.	0	0	11	1	3	0	0	0	0	Used MIC assays to show that evolved populations had a greater resistance than the WT to various antibiotics. Although a gene list was generated, the affect that individual mutations had upon the genes function were not investigated neither was the mutations individual contribution to a fitness phenotype.
		MH	**900 generations** 6 strains (Each from an independent population)	8	3	10	1	4	0	0	0	0	
		M9 + GLU	**500 generations** 8 independent populations	4	0	9	0	0	0	0	0	0	
		M9 + GLY	**500 generations** 8 independent populations	4	0	22	0	0	0	0	0	0	
Anand *et al.* (2020)	Experiment aims to adapt *E. coli* to oxidative stress. *E. coli* K-12 MG1655 cultures were grown in M9 minimal media supplemented with glucose, iron sulphate and sodium citrate. Creating an oxidative stress. An automated system passed cultures to fresh media when an OD600 of 0.3 was reached to prevent entry into stationary phase.		**1000 generations** 4 strains (Each from an independent population)	1	0	11	1	1	0	0	0	0	Used growth curves to show that evolved strains had a higher growth rate than WT. Did not isolate and focus upon individual mutations. Instead used RNA-seq of the evolved strains and protein modelling. To suggest the effect of 4 nonsynonymous mutations in *oxyR* are GoF causing the constitutive activation of OxyR and its regulon.

Reference	Experimental Design		Experimental duration What was sequenced ?	Gene				Intergenic			Large deletion	Duplication	Experimental characterisation of mutations
				INDEL	IS insertion	Point mutations			INDEL	IS insertion			
						Nonsynonymous	Synonymous	Intergenic					
Lee and Palsson (2010)	Experiment was to adapt cells to use the non-native carbon source L-1,2-propanediol. *E. coli* K-12 MG1655 was evolved were derived from single colonies of the glycerol evolved strain. These were used to inoculate the M9 media supplemented with glycerol and L-1,2-propanediol. Cultures were incubated at 37°C in continuous culture to prevent cells from entering into stationary phase.		**700 generations** 3 strains (Each from an independent population)	3	0	5	1	0	0	3	0	0	Used growth curves to show that evolved populations had a faster growth rate than WT under selective conditions. Protein purification, enzyme kinetics, growth curves and qPCR suggested that individual mutations observed were GoF: • Nonsynonymous mutation within *fucO* were GoF mutations showing increased 2 fold increase in the enzymes K_m value • IS insertion in promoter region of *fucAO* increased expression of the *fucO* (Note increase in growth rate was only observed if both mutations were present, allowing for the utilisation of L-1,2-propanediol.)
Toprak *et al.* (2011)	*E. coli* K-12 MG1655 cultures were evolved under the inhibition of antibiotic either chloramphenicol, doxycycline, or trimethoprim. Cultures were maintained in morbidostat culture tubes. When an OD600 of 0.3 was reached inhibitory concentrations of antibiotic was increased.	Trimethoprim	**16-19 days** 20 strains (4 each from an independent population)	0	0	14	0	0	0	0	0	0	Using MIC assays, evolved populations were shown to have a higher resistance to antibiotic than WT. Doesn't look at the effect that individual mutations have upon fitness or of the gene's function. Instead compares trade-off between the different antibiotic stresses and identifies an evolutionary constraint to trimethoprim resistance.
		Chloramphenicol	**22 days** 26 strains (5/6 each from an independent population)	0	0	11	1	0	0	0	0	0	
		Doxycycline	**22 days** 26 strains (5/6 each from an independent population)	0	0	10	0	0	0	0	0	0	
Conrad *et al.*, (2009)	*E. coli* K-12 MG1655 cultures were grown in L-lactate minimal medium in continuous exponential growth at 30°C over 45-60 days.		**750-1100 generations** 11 strains (Each from an independent population)	10	0	29	4	0	0	0	0	5	Used growth curves to show that evolved strains had a higher growth rate than WT. Using growth curves, 10 nonsynonymous and 1 synonymous mutations in 9 different genes were re-created individually in a WT background. All showed an increase in growth rate under selective conditions.

In addition, while some types of mutation may inherently be expected to cause loss of function, such as large deletions, nonsense, and frameshift mutations, the ability to predict a mutation's effect can be difficult in other cases. For example, nonsynonymous mutations are the most frequently recorded mutations within LE experiments, however deciphering their effects on a genes function can be difficult. Examples exist that associate nonsynonymous mutations to loss of function mutations (Knöppel *et al*., 2017, Hamdallah *et al*., 2018), but gain of function caused by single non-synonymous changes can also be identified (Johnson *et al*., 2014, Anand *et al*., 2020). Although approaches now exist to be able to predict a mutation's effect *in silico*, these can still be inaccurate and depend on extensive *a prori* knowledge (Tang and Thomas, 2016). This can also be said of other mutational types such as the effects of IS5 insertions upon gene function. For example, in He *et al.,* (2017) IS5 insertions identified in two regulators of AR mechanisms, within the gene *cadC* (AR4), and an intergenic insertion upstream of *adiY* (AR3), were shown to cause loss of function, as they had an identical phenotype to the corresponding gene knockout strains. Alternatively in Lee and Palsson, (2010) an intergenic IS5 insertion upstream of the *fucAO* operon was shown cause a gain of function mutation due to increased expression of the operon.

Overall, loss of function mutations within *E. coli* (or other organisms) can be adaptive in LE experiments. Of the twelve LE experiments listed in Table 1.2, ten investigated the effect that individual mutations have upon a fitness phenotype but only eight of these studies actually considered the affect that a mutation has upon a gene's function. Of these, four studies identified at least one loss of function mutation in one gene, while five studies identified at least one mutation as gain of function. This suggests that loss of function mutations are as common as an adaptive strategy as gain of function mutations. However, when it comes to comparing frequency of loss of function compared to gain of function mutations, table 1.2 highlights the limitation that none of these studies investigated the effects of all individual mutations, and therefore attempting to consider the frequency of each of these types of mutations is difficult.

1.3.3.2 Case study: Lind *et al*., (2015) highlights that loss of function mutations are more frequent.

One particular study showing that loss of function mutations are more frequent was conducted on the wrinkly spreader morph (WS) phenotype of *Pseudomonas fluorescens* SBW25 in nutrient rich static microcosm (Lind *et al*., 2015, McDonald *et al*., 2009). Initially, a study by McDonald *et al.,* (2009) showed that in 26 independent populations, mutations which led to the occurrence of the WS phenotype were all associated with the loss of function in any one of three repressors, which in turn caused the activation of a di-guanylate cyclase domain (McDonald *et al*., 2009). This was taken further in a study by Lind *et al.*, (2015), which looked for further pathways associated with the WS phenotype. To do this, genes for the three repressors previously identified by McDonald *et al.,* (2009) were deleted, and the experiment was rerun. This experiment identified a WS phenotype in 91/200 populations after 6 days. These populations were used to identify and understand the mutations which elicited a WS phenotype. This allowed Lind *et al.,* (2015) to estimate the frequency of occurrence and type of mutation which causes a WS phenotype. In almost all cases, an activation of a di-guanylate cyclase domain was the overall response, with 95% of these mutations were estimated to be associated with a loss of function mutation of a repressor (Figure 1.3.2). This shows the role that loss of function mutations can have in evolution experiments and shows that such mutations may be a potential common adaptive strategy within evolution experiments.

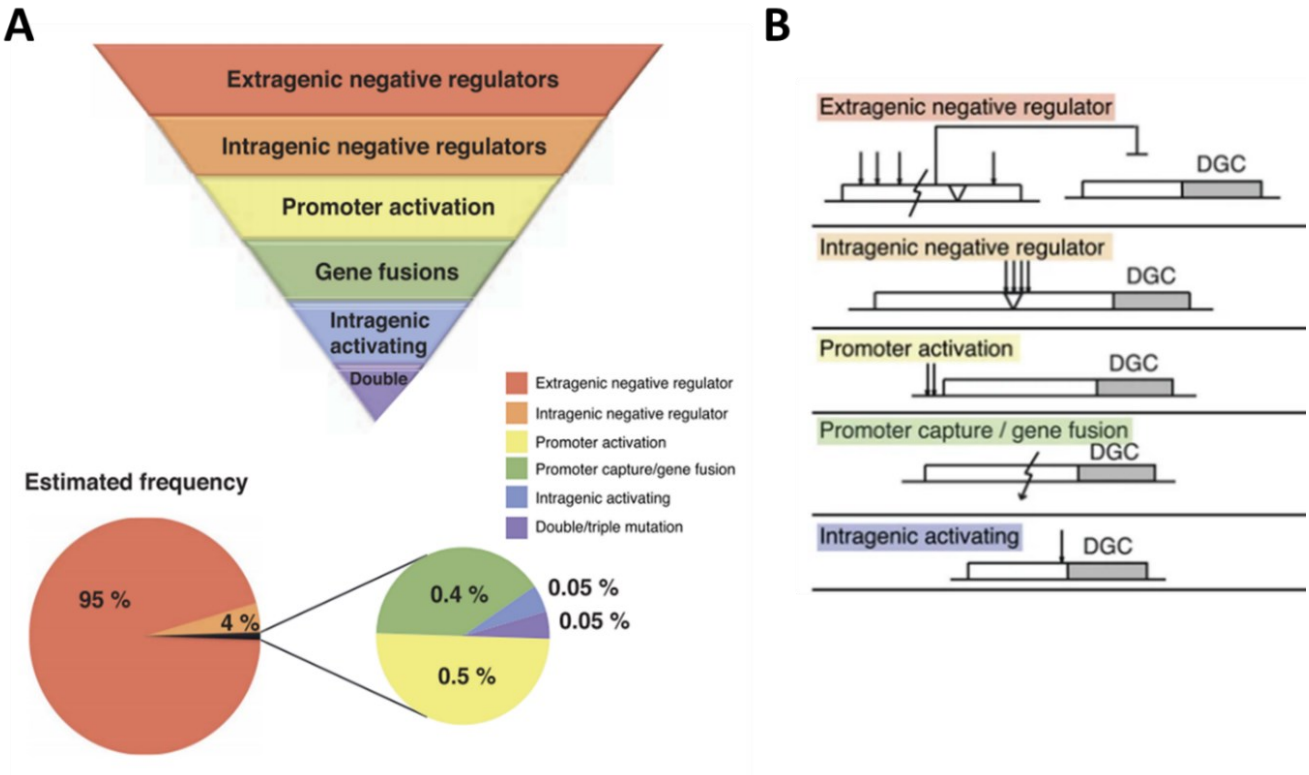

A
Extragenic negative regulators
Intragenic negative regulators
Promoter activation
Gene fusions
Intragenic activating
Double
Extragenic negative regulator
Intragenic negative regulator
Promoter activation
Promoter capture/gene fusion
Intragenic activating
Double/triple mutation
Estimated frequency
95 %
4 %
0.4 %
0.05 %
0.05 %
0.5 %
B
Extragenic negative regulator
DGC
Intragenic negative regulator
DGC
Promoter activation
DGC
Promoter capture / gene fusion
DGC
Intragenic activating
DGC

Figure 1.3.2: A) Summary of the type of mutations observed within Lind *et al., (*2015) to confer a Wrinkly spreader morphology in *Pseudomonas fluorescens* SBW25. All causes of the phenotype with the exception of double and triple mutations were found to be involved in the activation of a di-guanylate cyclase (DGC) domain function, highlighted grey in (B). Study shows that the largest frequency of mutations which cause this phenotype were loss of function mutations within repressor functions. B) Explanation of the effects of different mutations seen in this study splits the characterisation of the mutation into 6 terms, 1) Extragenic negative regulator: Mutations cause loss of function in a negative regulator, removal of this function causes activation 2) Intragenic negative regulator: A domain within the gene has a negative regulator function , mutations remove this function which causes activation the DGC domain. 3) Promoter activation: Mutations cause activation of the promoter and thus the DGC domain. 4) Promoter capture/gene fusions. Large deletions lead to the fusion of a DGC domain onto another gene/ promoter, which leads to increased expression of the DGC domain. 5) Intragenic activating: Mutations specific to a specific site on the gene, lead to the activation of the DGC domain. 6) Double and triple mutations: Multiple mutations occur in different genes, which leads to the wrinkly spreader morphology. Figures taken from (Lind *et al*., 2015).

1.3.3.3 Loss of function mutations can be a strategy for adaption under a variety of selection pressures

When considering a gene's function within a cell, the function does not sit in isolation, but typically sits in a vast regulatory network of interacting genes and proteins (A). Therefore, when considering the effect that a loss of function mutation can have upon the cell, it should not be considered in isolation, but the overall effect should be considered within this 'network'. An example of this is within a regulator, where loss of function mutation will remove the function of the regulator, and therefore can affect any gene under its control (Figure 1.3.3), which in turn may have a positive fitness effect on the organism overall in the context of the specific selection of the experiment.

The effect that loss of function mutations can have upon adaptation has been highlighted within one study conducted by Hottes *et al.* (2013). Initially, a meta-analysis was performed which identified the effects of loss of function mutations in 7 studies consisting of 144 different conditions. This showed that in 139 of these conditions at least one loss of function mutation could be identified which contributed to a fitness advantage, supporting the point that loss of function mutations which contribute to fitness are possible across multiple conditions. Furthermore, Hottes *et al.,* used a transposon library to assay the effect of fitness in minimal media supplimented with either alanine, aspartic acid, asparagine, or glutamine. Using a microarray to track the relative frequency of insertions within genes Hottes *et al.* managed to identify 397 genes which, when their function was lost, caused a fitness advantage in one of the four conditions, supporting the point that loss of function mutations can improve overall fitness under a variety of conditions.

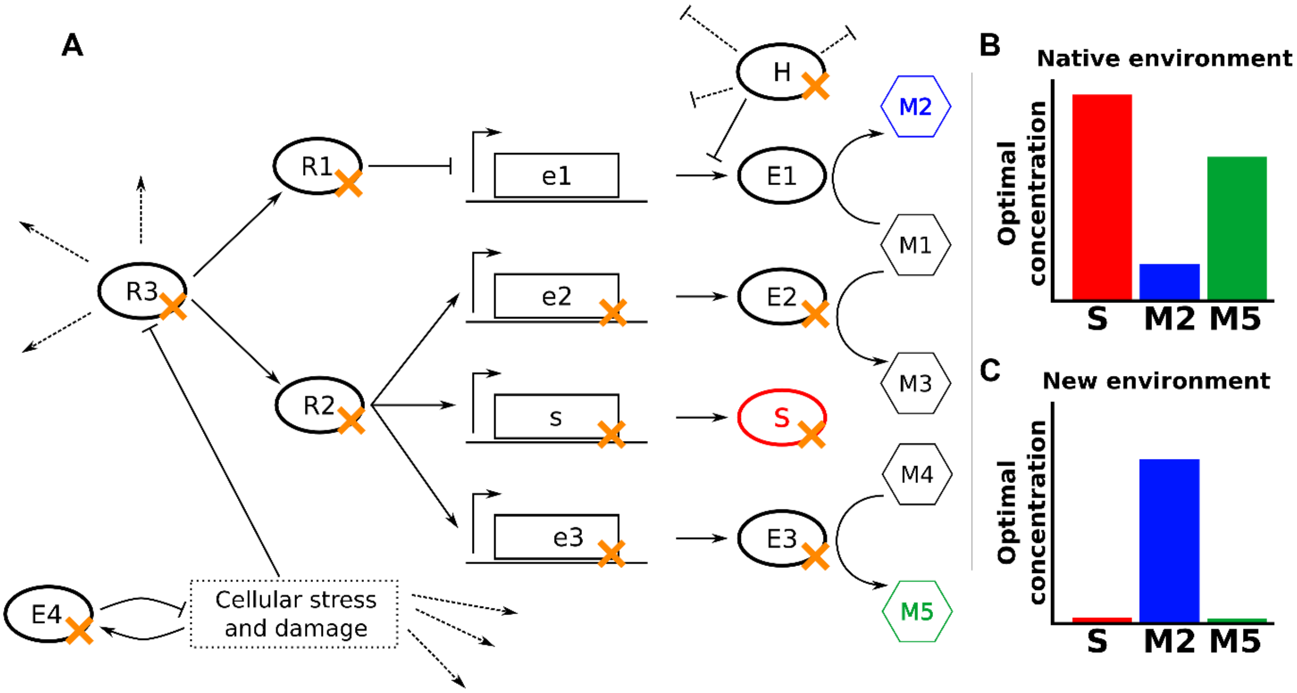


Figure 1.3.3: A Hypothetical network demonstrating the affect loss of function mutations can have A) A hypothetical network in which the nodes represent: R = Regulator, E -= Enzymes, M =Metabolites , S = Structural protein, H = Housekeeping protein which inhibits translation. The edges represent the interactions between nodes. Fitness is influenced only by the relative proportions of Metabolites M2 + M5 and Structural protein S (B,C). B) Indicates the relative concentration required for fitness under natural conditions. C) Upon being subjected to fitness under a new environment the optimal concentrations off M2, M5 and S change. Loss of function mutations can occur which causes advantageous fitness to the new condition in C, these are indicated by the yellow crosses. Figure taken from (Hottes *et al*., 2013).

1.3.3.4 Conclusions on Loss of function

Overall, this section highlights that loss of function mutations can be a common adaptive strategy within LE experiments. Examples of adaptive loss of function mutations can identified with LE experiments but quantifying the frequency of these can be difficult. Section 1.3.2 gave an example in *P. fluorescens* showing that under the selection pressures of nutrient rich static microcosms, loss of function mutations are more common than gain of function (McDonald *et al*., 2009, Lind *et al*., 2015). This is only one example, and the frequency of adaptive loss of function compared to gain of function mutations will differ for different conditions and potentially even different species (Monroe *et al*., 2021). This is in part due to differences in evolutionary landscapes and the effect that different selection pressures have in restricting paths that are available to achieve a fitter phenotype. This may explain why in some studies a large degree of parallelism is observed, as only a few routes may exist that lead to an adaptive phenotype (Johnson *et al*., 2014, He *et al*., 2017, Anand *et al*., 2020). Section 1.3.3 highlighted the wide variety of conditions where loss of function mutations can cause an advantageous phenotype, demonstrating that loss of function mutations potentially do play a significant role in adaption of an organism to selective conditions within an LE experiment.

1.3.4 Modification of laboratory based evolution to deliver outcomes in reduced time.

The use of laboratory based evolution as a tool to understand an organism's response to stress is useful. This report has discussed only a selection of the different experiments performed upon *E. coli* and its different response mechanisms to stress (Table 1.1 + 1.2). In regard to mild acid stress, only three evolution experiments have been published to date, excluding the experiment conducted by Sen, (2018), which is part of this study. One major limitation of LE experiments is that they take a long time to perform and can be quite labour intensive; this is highlighted in Table 1.1 and Table 1.2 where most evolution experiments range from a few weeks to months, while some have taken more than a year to conduct, not forgetting the LTEE, which has now been ongoing for over 30 years. As the occurrence of loss of function mutation within evolution experiments has shown to be a common adaptive strategy, this leads us the conclusion that a potential alternative methodology could be used to speed up these experiments. We propose that by evolving a population already containing all possible loss of function mutations, such as a high-density transposon library, the timescale of an evolution experiments could be reduced because all mutations which are required are already present and it is just the changing frequencies of these mutations in the evolving population that need to be tracked (for example, by using TraDIS). Therefore, a large proportion of the possible outcomes of an evolution experiment could possibly be determined over a shorter time scale.

Although previous experiments have been conducted using transposon libraries to identify loss of function mutations with an advantageous phenotype, to our knowledge no experiment has performed this comparison to predict the outcome of an evolution experiment (Hottes *et al*., 2013, Goodarzi *et al*., 2010). In addition, previous experiments have used microarrays to identify the frequency of transposons within a population, however these are limited in resolution (Goodarzi *et al*., 2010, Hottes *et al*., 2013, Girgis *et* *al*., 2009). Therefore, this study proposes the use of TIS methodologies, of which TraDIS is one, which greatly improves the resolution and the ability to identify and track the frequencies of individual transposon insertions.

1.4 Transposon Sequencing and TraDIS

The use of transposons and transposon mutagenesis to conduct genome wide phenotypic screens is not new. As early as 1995, a technique termed 'genetic footprinting', demonstrated, in *Saccharomyces cerevisiae*, its ability to identify localised regions of the genome which were important for growth under particular conditions. Using a library of transposon mutations and a PCR approach, locations of transposon insertions sites could be identified by visualisation on a gel (Smith *et al*., 1995, Murry *et al*., 2008). Since a transposon insertion disrupts a region, it can therefore affect any functionality associated with that region. Thus, when the library was challenged by a particular condition, the loss of transposon insertions from that region indicated its requirement. However, this method was limited, labour intensive, and only able to focus on individual regions.

This was taken further with the advent of microarrays with techniques such as Transposon Site Hybridisation (TraSH) allowing for identification of transposon insertions present on multiple regions of the genome simultaneously (Sassetti *et al*., 2001). However, this was plagued by issues associated with microarrays, such as the level of noise and lack of ability to detect low abundance transcripts.

2009 saw the advent of Transposon insertion sequencing (TIS) with 4 methods being released independently: Transposon Sequencing (Tn-seq) (van Opijnen *et al*., 2009,); Transposon Directed Insertion-site Sequencing (TraDIS) (Langridge *et al*., 2009); High-throughput Insertion Tracking by deep Sequencing (HITS) (Gawronski *et al*., 2009) ; and Insertion sequencing (INSeq) (Goodman *et al*., 2009). Although slightly different, these methods have the same principle. By using a high density transposon library coupled with HTS, these methods were able to simultaneously identify the insertion site of every transposon present within a library, in addition to ascertaining the relative frequencies of each insertion (Figure 1.4.1).

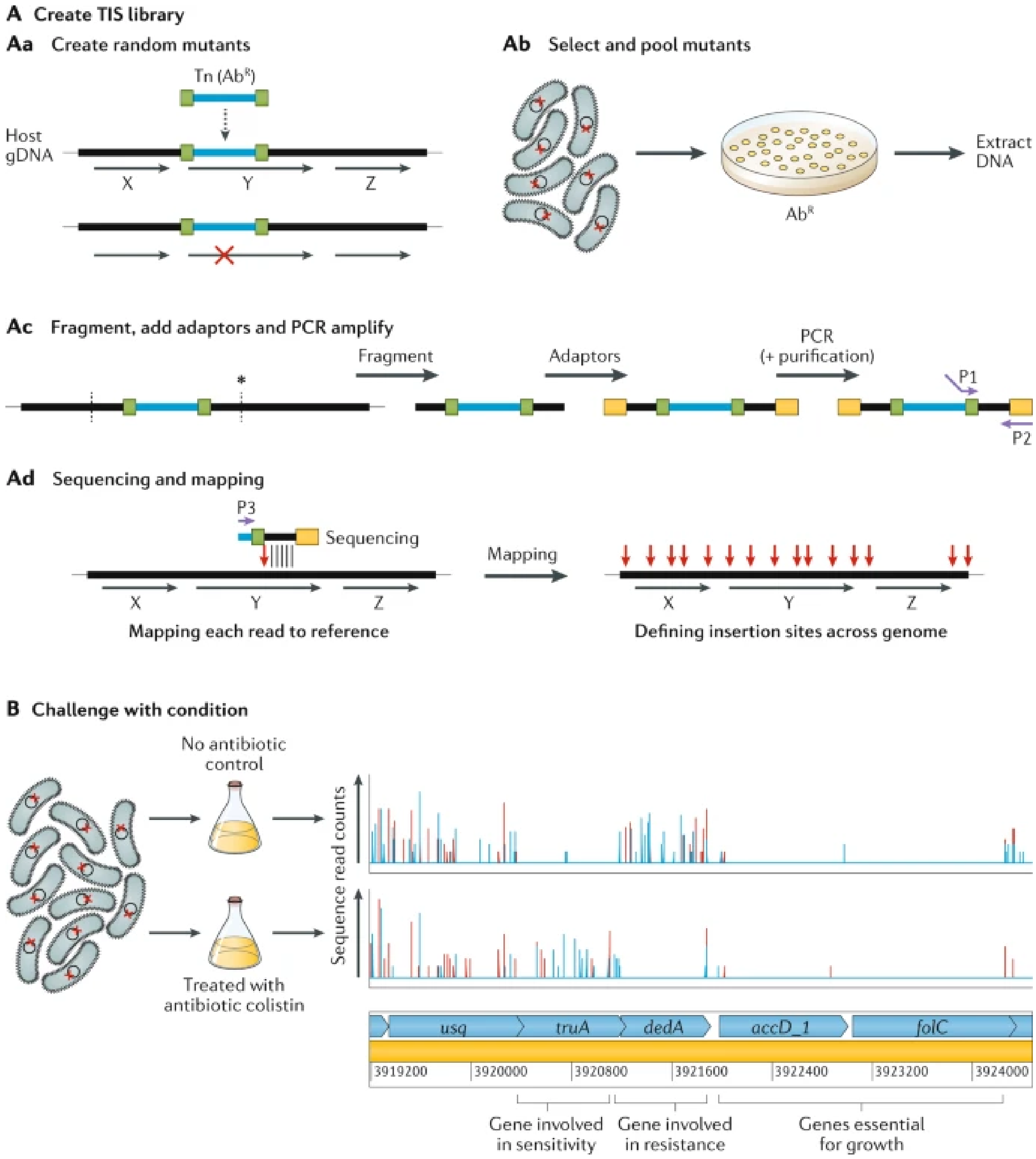

A Create TIS library
Aa Create random mutants
Tn (Ab^R)
Host gDNA
X
Y
Z
Ab Select and pool mutants
Ab^R
Extract DNA
Ac Fragment, add adaptors and PCR amplify
Fragment
Adaptors
PCR (+ purification)
P1
P2
Ad Sequencing and mapping
P3
Sequencing
Mapping
Mapping each read to reference
Defining insertion sites across genome
B Challenge with condition
No antibiotic control
Treated with antibiotic colistin
Sequence read counts
usg
truA
dedA
accD_1
folC
3919200
3920000
3920800
3921600
3922400
3923200
3924000
Gene involved in sensitivity
Gene involved in resistance
Genes essential for growth

Figure 1.4.1: An overview of transposon insertion sequencing (TIS) methodologies. A) The creation and sequencing of a transposon library. Aa) Transposon mutagenesis is performed, each transposon inserts into a random region within a genome. If a transposon insertion occurs within a gene, it can disrupt a genes function possibly causing loss of function. Ab) Transposon mutants are selected for, using an antibiotic resistance marker (Ab^R) and the library pooled together. Genomic DNA of the pooled library is then isolated. Ac) gDNA is fragmented and adaptors are ligated. A PCR reaction between the adaptor and the transposon junction can then amplify and enrich for gDNA fragments associated with a transposon junction. Ad) Transposon junction gDNA fragments are then sequenced and the gDNA associated with transposon junction aligned to a Genome. This then allows for the mapping and identification of insertion sites within the library. B) Transposon Libraries can then be subjected to different conditions and a comparison performed to identify insertions which decline or increase dependent upon the conditions. Taken from (Cain *et al*., 2020)

1.4.1 Creation of a transposon library.

Typically, most transposon libraries are created using either a Tn5 or Himar1/ mariner transposon. These two transposons have completely different origins being of prokaryotic and eukaryotic origin respectively (de Lorenzo *et al*., 1990, Lampe *et al*., 1998). Apart from their origin, these two transposons also differ in insertion site specificity, with Tn5 having a non-specific insertion site, although there may be some preference for GC rich regions (Green *et al*., 2012). This transposon is able to integrate into any region of the genome. The Himar1 mariner transposon is only able to insert into TA sites on the genome and this means that this transposon may not be suitable for GC rich genomes such as mycobacterium (Lampe *et al*., 1998).

1.4.2 Differences between TIS methods

As mentioned, each method HITS and TraDIS , Tn-seq and INSeq, have slight differences with Figure 1.4.2 highlighting these. However, each technique ultimately ends up with a fragmented gDNA library, which has been enriched for gDNA associated with Transposon junctions, before finally being prepared for HTS through the addition of sequencing adapters (Figure 1.4.1 + 1.4.2). Recognising the differences, the methods can be split up into two groups: Tn-seq and INSeq (Figure 1.4.2a); HITS and TraDIS (Figure 1.4.2b). Initially the major difference is the method used to fragment gDNA. Tn-seq and INSeq use a type II restriction enzyme, termed MmeI. Digestion using this restriction enzyme causes a staggered cut 20bp downstream of its recognition site creating a specific size of transposon containing fgDNA which only contains the transposon and 16bp of neighbouring gDNA (van Opijnen *et al*., 2009, Goodman *et al*., 2009). However, in order to be able to use MmeI digestion, these methods are constrained to a modified variant of the Himar1 mariner transposon. This transposon has a 1bp change in the inverted repeats of the transposon to introduce a MmeI restriction enzyme site (van Opijnen *et al*., 2009, Goodman *et al*., 2009). The alternative techniques, HITS and TraDIS, both randomly fragment gDNA using methods such as acoustic shearing, to generate small randomly fragmented gDNA (Gawronski *et al*., 2009, Langridge *et al*., 2009).

All of these techniques then use an adaptor ligation and a PCR based approach to be able to enrich for transposon junctions and neighbouring gDNA associated with the transposon insertion sites In addition, sequencing adaptors can then be added to ready the library for HTS.

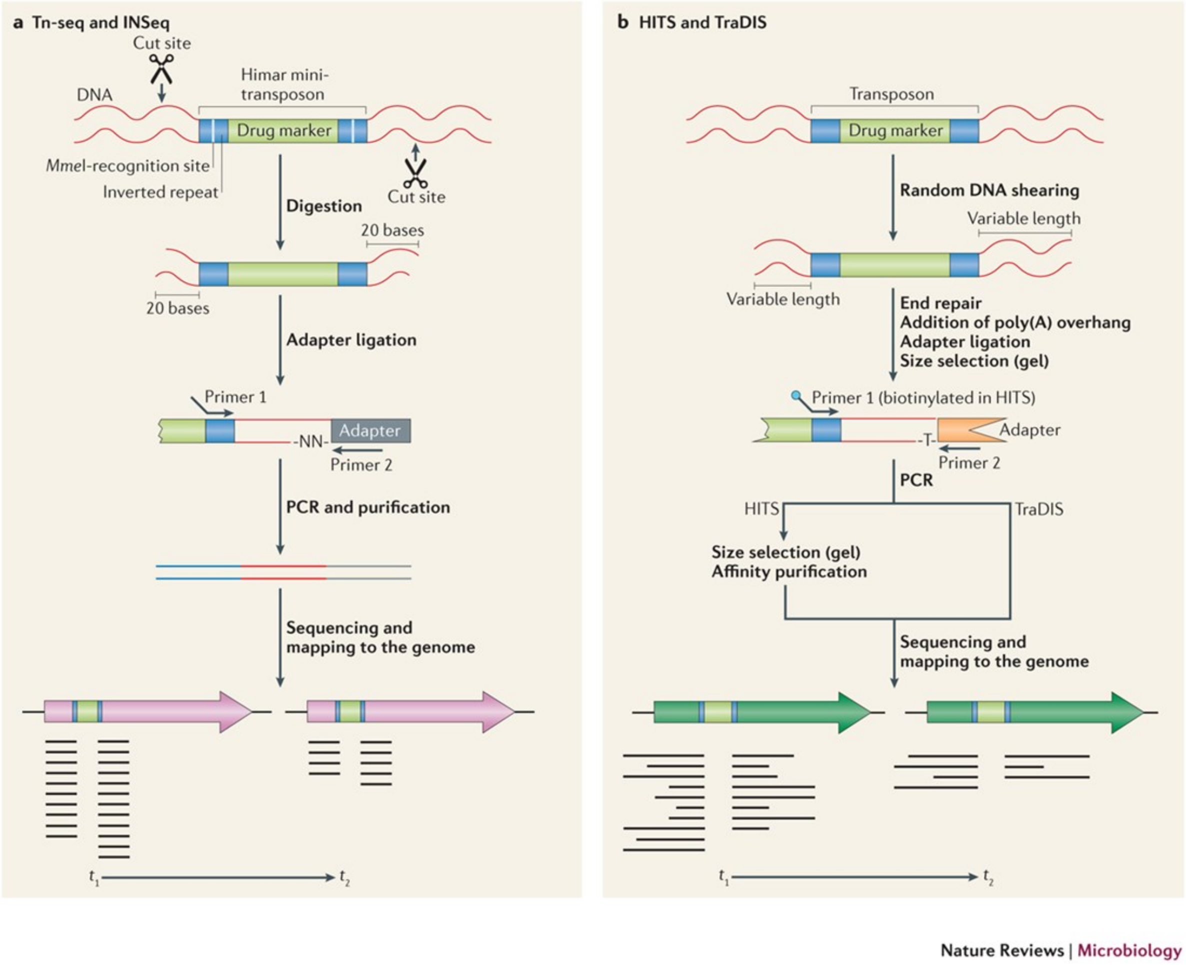


Figure 1.4.2: Diagram outlining the different methods available of TIS. a) Tn-seq and INSeq, both make use of the Himar transposon, which has been modified to contain a mmeI recognition site. Genomic DNA can then be is then fragmented using mmeI restriction digest, adaptor ligation and PCR are then performed. With INSeq and Tn-seq differing on the method used to purify the libraries either using a PAGE gel (INSeq) or agarose gel (Tn-seq) b) HITS and TraDIS fragment gDNA randomly before performing end repair and attaching the adaptor using TA ligation. Fragments are then size selected and amplifed using PCR. While TraDIS directly sequences these, HITS performs further purification using biotinylated primer and affinity capture before sequencing. Taken from (van Opijnen and Camilli, 2013)

1.4.3 Experimental use of TIS

1.4.3.1 Determining essential genes.

One major use for TIS methodologies is to identify essential genes or regions of the genome. Essential genes are usually defined as genes whose function is vital for the organism’s ability to survive. There are many reasons to identify essential genes in an organism. The most notable amongst these is that they highlight mechanisms which are fundamental for life. The use of transposon sequencing technologies has proven vital, identifying lists of essential genes in a variety of organisms such as *E. coli* (Goodall *et al*., 2018), *Pseudomonas aeruginosa* (Skurnik *et al*., 2013), *Salmonella* Typhi (Langridge *et al*., 2009), *Streptococcus pneumoniae* (van Opijnen *et al*., 2009) and *Mycobacterium tuberculosis* (DeJesus *et al*., 2017).

One study of note in regard to essentiality is by Goodall *et al.,* (2018). Using a high density transposon library in *E. coli* BW25113, they identified essential genes using TraDIS. This was then compared to two other different methods used to identify essentiality: A single gene deletion library (Keio collection) (Baba *et al.,* 2006) and the Profiling of *E. coli* Chromosome (PEC), an online database focusing on collation of information associated with understanding gene function within the *E. coli* K-12 MG1655 genome (Yamazaki *et al*., 2008). The study identified discrepancies between these three lists highlighting regions within the transposon library, where genes previously identified as essential actually had transposon inserts present. Further investigation of the TraDIS data, showed that in these genes only a domain of the gene was actually essential, and the remaining part of the gene was dispensable, and could therefore contain insertions. In addition to this, this study highlighted polar effects that can be observed with transposon insertions. This showed that in the Tn5 transposon used in this study, transcriptional and translational read-through from the transposon was observed. Using this result, by focusing on the orientation of a transposon insert, further information can be gained.

1.4.3.2 TIS applied after growth/subjected under defined conditions

In addition to identifying essential genes, one of the main use of TIS methodologies is to identify genes which are associated with a particular condition/phenotype. To do this a transposon library is challenged under a specific condition and TIS is employed before and after this. Using this approach, a gene’s contribution to a phenotype can be identified by tracking the change in frequency of insertions. This then allows the identification of two different lists of genes associated to the challenge of the condition. **1) Insertion Disadvantageous:** Genes whose function is required when the library is subjected to stress, typically determined by a reduction of insertions with the gene from the transposon library (Section 1.4.3.3). 2**) Insertion Advantageous:** Genes whose **loss of function** provides a fitness advantage under a stress, typically observed by an accumulation of insertions with a gene (Section 1.4.3.4). A third list also exists of genes in which insertions do not change in frequency, which therefore indicates that they have no effect on the phenotype under the specified condition.

1.4.3.3 Insertion disadvantageous genes and conditional essentiality

The concept of genes in which insertions are detrimental in fitness is typically the major use of transposon genes present within a library. These genes are sometimes termed conditionally essential genes based on the requirement of their function in order to be able to survive under particular conditions. However, the use of the term 'essential' when referring to a condition is a misnomer as it correctly refers to genes which are absolutely required for survival under these conditions. However, whilst this may be the case in most of these studies, what has really been shown is a mixture of both conditionally essential genes and insertions which have a disadvantageous fitness and therefore decline within the environment but are still able to show survival.

Within the literature, several different transposon libraries have been used to identify insertion disadvantageous genes under a wide variety of conditions, from simple laboratory conditions such as *E.coli* growth in LB, to complex environments such as the ability of *P. aeruginosa* to colonise a murine gastrointestinal tract (Goodall *et al*., 2018, Langridge *et al*., 2009, Skurnik *et al*., 2013, Cowley *et al*., 2018, Phan *et al*., 2013, van Opijnen *et al*., 2009, Gawronski *et al*., 2009, Goodman *et al*., 2009, van Opijnen and Camilli, 2012). A notable study involved screening of a *Streptococcus pneumoniae* transposon library under 19 different conditions (van Opijnen and Camilli, 2012). This study explored the ability of *S. pneumoniae* to colonise different niches in the host and to define the stresses that occur within these niches. By focusing on utilization of different carbon sources and other environmental stresses that could occur within a mouse model, this study identified over 700 genes in *S. pneumoniae* whose function was important in in at least 1 condition. Further to this, genotype to phenotype relationships were calculated using data across all conditions, which, when compared with the results of a transposon library which was subject to the *in vivo* conditions within a mouse, allowed the identification of stresses which were present within the *in vivo* model (van Opijnen and Camilli, 2012).

1.4.3.4 Insertion advantageous genes

Considering insertions which have an advantage under the condition, these potentially indicate insertions which cause a loss of function which contributes a fitness advantage and therefore accumulates within a population. Within the literature, however, these genes have tended not to be identified, with TIS experiments focusing more on the identification of the insertion disadvantageous genes. One experiment where these insertions were identified was within the original TraDIS paper (Langridge *et al*., 2009). Here, Langridge *et al.,* (2009) created and challenged a *S.* Typhi transposon library with 6 passages in 500ml of LB broth. During this experiment, 94 genes where insertions declined were identified, after the first day. At day 6 this increased to a total of 274 insertion detrimental genes whose function was associated with a decline in growth under these conditions. Interestingly, after 6 days of growth, Langridge *et al.,* (2009) also noted a set of genes where insertions had increased in frequency within the population, including 30 genes associated with flagella biosynthesis and assembly (Langridge *et al*., 2009). In addition to this, other studies which include the identification of enriched genes include the analysis of transposon libraries in *Vibrio cholerae* within rabbits and *Mycobacterium tuberculosis* within mice (Pritchard *et al*., 2014, Zhang *et al*., 2013, van Opijnen *et al*., 2009, van Opijnen and Camilli, 2012, Kamp *et al*., 2013).

1.4.4 Limitations and new approaches using TIS

The use of transposon libraries and TIS methodologies have been summarised in brief and this has demonstrated the use of transposon libraries to understand the genotype to phenotype relationship under a variety of conditions. As explained, when a transposon library is challenged using a given condition, genes in which insertions decline or increase in frequency can be identified and then associated with a phenotype. However, what will not be considered is the effect that essential genes will have on the phenotype, as these are identified by the absence of insertions within a library. Since essential genes will not have insertions present within them, this limits any further analysis using TIS to consider genotype to phenotype relationships, as no insertions are present and therefore cannot be tracked.

Recently, an approach termed TraDIS-Xpress combined TraDIS with a Tn5 transposon cassette constructed to include an outward facing ITPG inducible *tac* promoter (Yasir *et al*., 2020). By doing this, overexpression of genes can be considered, as transposon insertions upstream of a gene will have the potential to cause overexpression due to the outward facing promoter. Thereby, this technique provides the opportunity for essential genes to be assayed, as although an essential gene cannot be disrupted, the expression of some essential genes can be affected. A TraDIS-Xpress library, therefore, combines the effects seen within a gene disruption library and an overexpression library, with this technique being already employed to identify essential and non-essential genes of interest under a variety of conditions including triclosan resistance (Yasir *et al*., 2020), Fosfomycin resistance (Turner *et al*., 2020) and biofilm formation (Holden *et al*., 2021).

In addition to TraDIS-Xpress, new approaches combining other methods and TIS have been developed to explore biological phenomena. One such approach, termed droplet Tn-seq (dTn-Seq), combines Tn-Seq and microfluidics to allow for mutants to be assessed in isolation under various conditions by encapsulating individual cells in oil droplets containing medium (Thibault *et al*., 2019). Another approach, termed TraDISort, combines TraDIS with other methods which are able to sort a library based on physical properties, one example being the combination of TraDIS with fluorescence-activated cell sorting (FACS) to identify genes associated with the efflux of ethidium bromide in *Acinetobacter baumannii* (Hassan *et al*., 2016). Another TraDISort approach involves using a Percoll density gradient to separate a *Klebsiella pneumoniae* transposon library to identify genes associated with capsule production (Dorman *et al*., 2018).

1.5 Project aims

The aims of this thesis are to firstly continue the work conducted by Sen, (2018) in understanding the outcome of the evolution experiment conducted in unbuffered LB at pH 4.5 for 5 months. The second aim is to construct a transposon library in the same strains as Sen's evolution experiment (*E. coli* K-12 MG1655) and to identify essential genes within this strain. The third and main aim of this study is to challenge this newly created *E. coli* K-12

MG1655 transposon library to the same conditions as Sen's evolution experiment, however at a considerably shorter time frame; then to compare the outcomes to see if a short term outgrowth experiment of transposon library successfully predicts the outcomes of a longer term evolution experiment.

Chapter 2

Materials and Methods

All scripts used in this study and a copy of supplementary data can be found in this location

https://www.dropbox.com/sh/0dss5u7dcyxweeg/AACxOu_HzSRwQkp_ObAekMBaa?dl=0

2.1 Bacterial Strains and Plasmids

All bacterial strains used within this study are shown in Table 2.1. For the majority of this study, *E coli* K-12 MG1655 strain was used. The reported genotype of this strain is F$^-$, λ$^-$, *ilvG*$^-$, *rfb*-50, *rph*-1, whole genome resequencing also identified a *yeaJ::IS1* insertion present. Construction of specific knockout strains are detailed in section 2.1.1 and section 2.1.2. For storage, overnight cultures of strains were resuspended in LB supplemented with 10 % glycerol and stored at -80°C. When needed, strains were streaked onto an agar plate containing the relevant selective antibiotic if required to achieve single colonies.

Table 2.1: Strains used within this study. All kanR resistance genes are flanked by FRT sites as described in (Baba *et al*., 2006). (1) Represent strains constructed within this study.

Strain	**Description/Genotype**	**Reference**
MG1655	Major strain used within this study. Genotype *E. coli* K-12 MG1655 (F$^-$, λ$^-$, *ilvG*$^-$, *rfb-50 rph-1*, *yeaJ::IS1*)	(Blattner *et al*., 1997)
BW25113	Parent strain of Keio collection *E. coli* K-12 BW25113 (*lacI*$^+$, *rrnB*$_{T14}$, Δ*lacZ*$_{WJ16}$, *hsd*$_{R514}$, Δ*araBAD*$_{AH33}$, Δ*rhaBAD*$_{LD78}$, *rph-1*, Δ(*araB–D*)567, Δ(*rhaD–B*)568, Δ*lacZ*4787(::*rrnB*-3), *hsd*R514, *rph-1*	(Datsenko and Wanner, 2000)
TOP10	Cloning strain used within this study. *E. coli* K-12 TOP10 *(F-*, *mcrA*, Δ(*mrr-hsdRMS*-mcrBC), ф80*lacZ*ΔM15, Δ*lacX*74, *nupG*, *recA1*, *araD*139, Δ(*ara-leu*)7697, *galE*15, *galK*16, *rpsL*(StrR), *endA1*, *λ*$^-$)	Invitrogen

Strain	**Description/Genotype**	**Reference**
KH001	MG1655 *lacZ*$^{-}$ with *lac* promoter replaced with pKH3 cloning site	Courtesy of Jack Bryant.
MG1655 *ΔyjjY* [1]	MG1655 *ΔyjjY::kan*^{R}	P1 BW25113 *ΔyjjY::kan*^{R} (Baba *et al*., 2006) x MG1655.
MG1655 *ΔarcA* [1]	MG1655 *ΔarcA::kan*^{R}	P1 BW25113 *ΔarcA::kan*^{R} (Baba *et al*., 2006) x MG1655
MG1655 *ΔarcB* [1]	MG1655 *ΔarcB::kan*^{R}	P1 BW25113 *ΔarcB::kan*^{R} (Baba *et al*., 2006) x MG1655
MG1655 *ΔfimE* [1]	MG1655 *ΔfimE::kan*^{R}	P1 BW25113 *ΔfimE::kan*^{R} (Baba *et al*., 2006) x MG1655
MG1655 *ΔrpoS* [1]	MG1655 *ΔrpoS::kan*^{R}	P1 BW25113 *ΔrpoS::kan*^{R} (Baba *et al*., 2006) x MG1655
MG1655 *ΔptsP* [1]	MG1655 *ΔptsP::kan*^{R}	P1 BW25113 *ΔptsP::kan*^{R} (Baba *et al*., 2006) x MG1655. (Constructed by Wei Li)
MG1655 *ΔcadC* [1]	MG1655 *ΔcadC::kan*^{R}	P1 BW25113 *ΔcadC::kan*^{R} (Baba *et al*., 2006) x MG1655
MG1655 *ΔcytR* [1]	MG1655 *ΔcytR::kan*^{R}	P1 BW25113 *ΔcytR::kan*^{R} (Baba *et al*., 2006) x MG1655
MG1655 *picd:lacz*	*arcA* activity reporter. MG1655 *picd:lacZ kan*^{R}	(Park and Kiley, 2014)

Strain	Description/Genotype	Reference
MG1655 *ΔcadA* [1]	MG1655 *ΔcadA::kan^R*	P1 BW25113 *ΔcadA::kan^R* (Baba *et al.*, 2006) x MG1655
MG1655 *ΔcadB* [1]	MG1655 *ΔcadB::kan^R*	P1 BW25113 *ΔcadB::kan^R* (Baba *et al.*, 2006) x MG1655
MG1655 *ΔytfE* [1]	MG1655 *ΔytfE::kan^R*	P1 BW25113 *ΔytfE::kan^R* (Baba *et al.*, 2006) x MG1655
MG1655 *ΔrppH* [1]	MG1655 *ΔrppH::kan^R*	P1 BW25113 *ΔrppH::kan^R* (Baba *et al.*, 2006) x MG1655
MG1655 *ΔtreR* [1]	MG1655 *ΔtreR::kan^R*	P1 BW25113 *ΔtreR::kan^R* (Baba *et al.*, 2006) x MG1655
MG1655 *ΔptsO* [1]	MG1655 *ΔptsO::kan^R*	P1 BW25113 *ΔptsO::kan^R* (Baba *et al.*, 2006) x MG1655
MG1655 *ΔdusB* [1]	MG1655 *ΔdusB::kan^R*	P1 BW25113 *ΔdusB::kan^R* (Baba *et al.*, 2006) x MG1655
MG1655 *ΔgadX* [1]	MG1655 *ΔgadX::kan^R*	P1 BW25113 *ΔgadX::kan^R* (Baba *et al.*, 2006) x MG1655
MG1655 *ΔgadE* [1]	MG1655 *ΔgadE::kan^R*	P1 BW25113 *ΔgadE::kan^R* (Baba *et al.*, 2006) x MG1655
MG1655 *ΔrcsB* [1]	MG1655 *ΔrcsB::kan^R*	P1 BW25113 *ΔrcsB::kan^R* (Baba *et al.*, 2006) x MG1655

Strain	**Description/Genotype**	**Reference**
MG1655 *ΔfimA* [1]	MG1655 *ΔfimA::kan*^{R}	P1 BW25113 *ΔfimA::kan*^{R} (Baba *et al*., 2006) x MG1655
KH001 *ΔfimE* [1]	KH001 *ΔfimE::kan*^{R}	P1 MG1655 *ΔfimE::kan*^{R} (This study) x KH001
E1A	Strain evolved at pH 4.5 in unbuffered LB for 5 months from MG1655 ancestor.	(Sen, 2018)
E2A	Strain evolved at pH 4.5 in unbuffered LB for 5 months from MG1655 ancestor.	(Sen, 2018)
E3A	Strain evolved at pH 4.5 in unbuffered LB for 5 months from MG1655 ancestor.	(Sen, 2018)
E4A	Strain evolved at pH 4.5 in unbuffered LB for 5 months from MG1655 ancestor.	(Sen, 2018)
E5A	Strain evolved at pH 4.5 in unbuffered LB for 5 months from MG1655 ancestor.	(Sen, 2018)
E1A *ΔarcA* [1]	E1A *ΔarcA::kan*^{R}	P1 BW25113 *ΔarcA::kan*^{R} (Baba *et al*., 2006) x E1A
E2A *ΔarcA* [1]	E2A *ΔarcA::kan*^{R}	P1 BW25113 *ΔarcA::kan*^{R} (Baba *et al*., 2006) x E2A
E3A *ΔarcA* [1]	E3A *ΔarcA::kan*^{R}	P1 BW25113 *ΔarcA::kan*^{R} (Baba *et al*., 2006) x E3A
E4A *ΔarcA* [1]	E4A *ΔarcA::kan*^{R}	P1 BW25113 *ΔarcA::kan*^{R} (Baba *et al*., 2006) x E4A
E5A *ΔarcA* [1]	E5A *ΔarcA::kan*^{R}	P1 BW25113 *ΔarcA::kan*^{R} (Baba *et al*., 2006) x E5A

Strain	**Description/Genotype**	**Reference**
E1A-I1	Strain Isolated from Fossil record of E1A. Genotype: MG1655 *arcA*$_{\text{M39I}}$	(Sen, 2018) Termed F1
E1A-I2	Strain Isolated from Fossil record of E1A. Genotype: MG1655 *arcA*$_{\text{M39I}}$, *rpoA*$_{\text{N294H}}$, *cytR-priA::IS5 insertion*	(Sen,2018) Termed F2
E1A-I1 *picd:lacZ* [1]	*arcA* activity reporter. F1 picd:lacZ kan^{R}	
MG1665 *ΔyjjY::FRT* [1]	*MG1655 ΔyjjY:FRT* (102bp scar)	MG1655 *ΔyjjY::kan^{R}* x pcp20
E1A *picd:lacZ* [1]	*arcA* activity reporter. E1A picd:lacZ kan^{R}	P1 MG1655 *picd:lacz* x E1A
E2A *picd:lacZ* [1]	*arcA* activity reporter. E2A picd:lacZ kan^{R}	P1 MG1655 *picd:lacz* x E2A
E3A *picd:lacZ* [1]	*arcA* activity reporter. E3A picd:lacZ kan^{R}	P1 MG1655 *picd:lacz* x E3A
E4A *picd:lacZ* [1]	*arcA* activity reporter. E4A picd:lacZ kan^{R}	P1 MG1655 *picd:lacz* x E4A
E5A *picd:lacZ* [1]	*arcA* activity reporter. E5A picd:lacZ kan^{R}	P1 MG1655 *picd:lacz* x E5A
MG1655 *ΔarcA:FRT picd:lacZ* [1]	*arcA* activity reporter. MG1655 picd:lacZ kan^{R}	(Park and Kiley, 2014)
MG1655 *yjjY(-1)::tn5* [1]	MG1665 *yjjY:tn5*	Transposon mutant in *yjjY* isolated from D10-pH4.5-S1 of the STSE.
MG1655 *yjjY(+25)::tn5*[1]	yjjY:tn5 *yjjY::tn5*	Transposon mutant in *yjjY* isolated from D10-pH4.5 S3 of the STSE.
MG1655 f*imE(+65)::tn5*[1]	MG1655 *fimE::tn5*	Transposon mutant in *yjjY* isolated from D10-pH7-S2 of the STSE.

The plasmids used in this study are described in Table 2.2. All plasmids were stored in 10 mM Tris-HCl, pH 8.5 and stored at –20°C. When growing strains containing plasmids the culture was supplemented with the relevant selective antibiotic

Table 2.2: Plasmids used within this study

Plasmid	**Description**	**Reference**
pCP20	Contains FLP recombinase. Temperature sensitive plasmid cured at 37^{o}C. Amp^{R}	(Cherepanov and Wackernagel, 1995)
pSynP21	GFP reporter fusion contains, P21, a synthetic promoter specific for rpoD activity. Amp^{R}	(Klauck *et al*., 2018)
pSynP8	GFP reporter fusion contains, P8, a synthetic promoter specific for RpoS activity. (Note this plasmid was reconstructed in situ.). Amp^{R}	(Klauck *et al*., 2018)
pKD4	Gene Knockout plasmid. Plasmid contains kanamycin cassette flanked by FRT sites.	(Datsenko and Wanner, 2000)
pKD46	Red recombinase expression plasmid.	(Datsenko and Wanner, 2000)

2.2 Growth Conditions

2.2.1 Media

All media were made up to correct volumes using distilled water. Growth media used within this study are as follows. Lysogeny broth (LB: 10g/L - Tryptone, 5g/L - yeast extract, 10g/L - NaCl: pH7). Lysogeny broth agar (LBA) was made through the addition of 15g/L - agar to LB. LBA agar and LB (which was not used for a specific pH experiment) was made and autoclaved using an inhouse service. For P1 transductions, when determining the titer of phage, a soft agar was made by mixing melted LBA with LB at a 1:1 ratio and kept at 60^{o}C until used. For transformation (Section 2.5), super optimum media (SOC: 20g/L - tryptone, 5g/L - yeast extract, 10mM - NaCl, 10mM - $MgCl_2$, 2.5mM - KCl, 20mM - glucose) for the construction of a transposon library (Section 2.7.1) a manufactured version (Sigma-Aldrich) was used. For competitions, MacConkey agar supplemented with lactose (MacConkey:

40g/L - Difco^{TM} MacConkey Base, 10g/L - lactose, pH 7.1) was used. M9 minimal media supplemented with cas amino acid (42.3 mM - Na_2HPO_4, 22.1mM – KH_2PO_4, 8.56 mM – NaCl, 18.7mM - NH_4Cl, 0.1mM - CaCl, 2mM $MgSO_4$ 20mM – Glucose, 0.2 % w/v cas-amino acids, pH 7). 2xTY broth (2xTY: 16g/L - Tryptone, 10g/L - Yeast extract, 5g/L - NaCl, pH 6.8). Where possible all constituents of a media were mixed together before autoclaving, however where a substance was known to degrade or cause precipitation due to autoclaving, a stock solution was made, sterilized by passing the solution through a 0.22µM filter and added separately after autoclaving.

2.2.2 pH

Unless otherwise stated pH was determined with a Jenway 3510 pH meter using a Thermo Scientific Orion 815600 ROS combination probe. Media was adjusted to pre-determined pH using 1M HCl or 1M NaOH solutions before autoclaving. In certain cases where a smaller volume needed to be tested a CAMLAB pHBoy-P2 was used.

2.2.3 Buffers

Phosphate buffered saline (PBS) was made as a 10x stock solution (10x PBS: 80g/L- NaCl, 2g/L - KCl, 14.4g/L - Na_2HPO_4, 2.4g/L - KH_2PO_4), a 1x PBS solution was then made from this by dilution and adjusted to pH 7.4 before being autoclaved.

Where stated, the following buffers were added achieve to growth media : 2-(N-morpholino)ethanesulfonic acid (MES, pKa(37°C) = 5.97); 3-(N-Morpholino)propanesulfonic acid (MOPS, pKa(37°C) = 7.02); N-(1,1-Dimethyl-2-hydroxyethyl)-3-amino-2-hydroxypropanesulfonic acid (AMPSO, pKa(37°C) = 9.10); Homopiperazine-1,4-bis(2-ethanesulfonic acid) (HOMOPIPES, pKa(37°C) = 4.55). All buffers were used at a final concentration of 50mM.

2.2.4 Growth conditions

Unless otherwise stated *E.coli* was grown at 37°C in a shaking incubator when using liquid media (180 rpm) or static incubator for solid media. All cultures originated from a single colony. Selections of antibiotic markers were performed using antibiotic concentrations of 100µg/ml - Ampicillin, 30µg/ml - Chloramphenicol and 50µg/L - Kanamycin.

2.3 General Microbiology Techniques

2.3.1 Mini-preparation of plasmid DNA

A QIAprep spin Miniprep Kit (Qiagen) was used to prepare small volumes of plasmid DNA according to the protocol provided. The kit performs alkaline lysis of bacteria which causes protein denaturation and chromosomal and plasmid DNA separation into single strands. This is then neutralized and chaotropic salts are added; this causes denatured proteins and chromosomal DNA to precipitate while plasmid DNA can renature and form dsDNA which stays in solution. This solution is then passed through a silica membrane which will allow for the selective absorption of plasmid DNA in a high salt environment and elution in low salt environments. Quantification of isolated plasmid was performed using a nanodrop (Section 2.3.4.8). An average concentration 50 – 90 µg/ml was usually obtained.

2.3.2 Genomic DNA preparation and extraction

Genomic quantification of chromosomal DNA was done using the Stractec RTP Bacteria DNA Mini kit using protocol 2. This protocol performs a lysis procedure using high temperature under non-chaotropic conditions with proteinase K present. The optimal binding conditions are then obtained using a binding buffer containing isopropanol to allow efficient adsorption of genomic DNA to membrane present in a spin filter tube. An ethanol wash is then performed to remove contaminants, before elution from the membrane with 10mM Tris-Cl, pH 8.5 (Buffer EB - Qiagen)

2.3.3 Qubit Quantification of DNA

Where stated, DNA was quantified using a Qubit® 2.0 fluorometer using Qubit^{TM} dsDNA HS Assay kit (Invitrogen) with 1ul of sample.

2.3.4 PCR and Sequencing

2.3.4.1 Oligonucleotides

Oligonucleotides used in this study were synthesized by either Eurogentech (Seraing, Belgium) or Sigma-Aldrich (Dorset, UK). Primers used in this study are described in Tables Table 2.3 and 2.4, below. An exception is primers used for TraDIS method; these can be found in Table 2.14

Table 2.3: Oligonucleotides used in this study for PCR screening and/or sequencing

Name	**Sequence**	**Description**[1]
arcA_F	GATGGAAAGTGCATCAAGAACG	*arcA* -137
arcA_R	TTCACTGCCGAAAATGAAAGC	*arcA* +843
arcA_UP_F [2]	GGCTAAACTATTTCCTGACTGTACTAACGGTTGAGTTG	*arcA* -513
arcA_UP_R [2]	TCCCTGACCTGCCTGATGCATGCTG	*arcA* -48
arcB_F	AAATGATTCGCCATACGCCACC	*arcB* -286
arcB_R	GCCGCTTCAATCAGCACATTGC	*arcB* +2769
BioH_F	CAGGGAAGATGGGAGTGTGTGG	*bioH* -189
BioH_R	GTCAGAAGAAGAGTTGGCTGCG	*bioH* +833
cadA_F	TAACATCCGCAACTTTGTCAGC	*cadA* -278
cadA_R	ATCCATACCACTGAGCAGCC	*cadA* +2702
cadB_F	TTCTTGCTTCAGAATAAGTAACTCC	*cadB* -381
cadB_R	ATCCCAGTCAAAAATAACGCC	*cadB* +1594
cadC_F	TACCGAACTCAACCAGATTCTCC	*cadC* -533
cadC_R	TCCAACCCCAGATAGCAATACC	*cadC* +2033
cytR_F	ATGTAGTACGCCTGACGTGCCAG	*cytR* -38
cytR_R	CAGCGTTCTTCTGGTTTTGGTGGTAGTCCGTTTCCG	*cytR* +1365
cytR_UP_F [2]	GGCGCTCATCAATACAATACCGGATTCC	*cytR* -228
cytR_UP_R [2]	GGGATTCATTACGCTTGACGTTGCG	*cytR* +202
dusB_F	GGTCGTGGAAAAAGAAGAGTGG	*dusB* -377
dusB_R	TTGCATCACCATGTCCAACAG	*dusB* +1195
Ez-Kan_F	CGATGAATTGTGTCTCAAAATCTCTG	Internal primer for EZ-Tn5 Kan^{R} See section ….
Ez-Kan_R	CTTATACACATCTCAACCCTGAAGC	
fimA_F	TCTTCTCTCCAAAAACCACCT	*fimA* -250
fimA_R	CGGTTACTGCTGATTTGCC	*fimA* +801
FimE_F	CACTTAAAATCTCCTGTTTCGCAC	*fimE* -273
FimE_R_OFF [3]	GGTGGTTTTTGGAGAGAGAAGAATGAGG	*fimE* +1232
FimE_R_ON [3]	GACGCATCTTCCTCATTCTTCTCTCC	*fimE* +813
gadE_F	CAATATATGGAGTGCGTGATGG	*gadE* -177
gadE_R	CGAAGGGTTCTACTGGTGG	*gadE* +668
gadX_F	CACACATTATCATCCTGTTCTCC	*gadX* -119
gadX_R	CGAATGAACGACGAATGAGG	*gadX* +1235
IS186_F	GAGACGTGCCGGAAAGGGAGCTCC	Internal primers for *IS186*. See section …
IS186_R	GAGATTGCGTCTTGTCGATGGAACAGC	
IS5_F	CGATGGAAGATGCTCTGTACGAAATCGC	

Name	Sequence	Description[1]
IS5_R	AGCAGGTGGCGGAAATTCATGATGG	Internal primer for *IS5*. See section …
ptsO_F	TTCTCACCGCAACTCCTGTTCC	*ptsO* -620
ptsO_R	TTTCCGATGTCTTCCGGTGCC	*ptsO* +537
ptsP_F	GGTGCGTTGGGACACGAAGC	*ptsP* -295
ptsP_R	CCAAGACCAAACGGAATGAGTGG	*ptsP* +2816
rcsB_F	CGTGAGAAGGATGTTCCAGG	*rcsB* -88
rcsB_R	GAATCGTAGGCCGGATAAGG	*rcsB* -745
rpoA_F	GCTTCTTATCAGGTTAGTCCG	*rpoA* -250
rpoA_R	AGGCAGAGTCGTCTTGATGATTTCATGACGAACCAGTGAACCT	*rpoA* +1147
rppH_F	CATTTACAGCCTTGACGTGCG	*rppH* -447
rppH_R	TTGATGCCTGATACACCTGTTCC	*rppH* +1147
sspA_F	AACTATCATCCAATTTTCTGCCC	*sspA* -156
sspA_R	CCATCAATAACCGACATAACGG	*sspA* +1048
tnaA_F	AGCCATCACCAGAGCCAAACC	*tnaA* -207
tnaA_R	GGCACCCCAGAAAAACCAGGC	*tnaA* +1635
tnaB_F	CCGAAAGTATTGCGTCACTTCACC	*tnaB* -135
tnaB_R	AAAGCGGGACATGGGCTAAAGC	*tnaB* +1391
tnaC_F	TTCATTGTTACCACTCCTGTTATTCC	*tnaC* -198
tnaC_R	CCTGTGAATATTACATCTGCTATACC	*tnaC* +260
treR_F	ACGATGAACCAGACGGGAAGG	*treR* -438
treR_R	CCGGTTGGTTGAGGACAAAGC	*treR* -1187
yidQ_F	CGTTGATACATGACAACCTCC	*yidQ* -326
yidQ_R	CGTAATTTTCAGTGCGTTGC	*yidQ* +986
yjjY_R	CGTGTAACGATTCAGCCAATGTCG	*yjjY* -748
yjjY_F	TTTCAACGTGTTGCGTGTTACC	*yjjY* +155
yobF_F	ACGCCAGTTTAAGTATCTGCC	*yobF* -202
yobF_R	AAAGCATAGTAAAAAGCCTCGC	*yobF* +448
ytfE_F	CGTGCTTTGCTCATGGGTTGG	*ytfE* -354
ytfE_R	CGATCTACAAGATGCCGCTCGG	*ytfE* +898

1.Primers were designed for each gene considering the transcriptional direction being in the forward orientation. Therefore, in the primer names "_F" or "_R" represent the forward or reverse DNA strand where the primer anneals in the context of the gene stated translational direction.

2. Where possible, the site of the first base of where each primer anneals is recorded relative to translational start site of the gene in the MG1655 genome (Accession No: NC_000913.3)

3. UP refers to focusing on region upstream of the gene mentioned.

4. FimE OFF and ON refer to orientation of fimS for more information see section XX. Positions recorded for this gene is where *fimS* is in the OFF orientation according to the MG1655 genome

Table 2.4: Oligonucleotides used in the making of KO and RpoS reporter according to the method described in section 2.3.8. Underlined refers to the homology described in pkd4. Description highlights the final base pairing of primer annealing relative to either the translational start site or stop site of defined gene.

Name	**Sequence**	**Description**
arcB_KO_F	TCAGAATTGGGTATTATTGGGGCAGGTTGTCGTGAAGGAATTCCCTAATGTGTGTAGGCTGGAGCTGCTTC	Start (+3)
arcB_KO_R	AGTATTCGCGCACCCCGGTCTAGCCGGGGTCATTTTTTAGTGGCTTTTGCCATATGAATATCCTCCTTAGTTCC	Stop (-21)
cadA_KO_F	TTTGTCCCATGTGTTGGGAGGGGCCTTTTTTACCTGGAGATATGACTATGTGTGTAGGCTGGAGCTGCTTC	Start (+3)
cadA_KO_R	TGGCAAGCCACTTCCCTTGTACGAGCTAATTATTTTTTGCTTTCTTCTTTCATATGAATATCCTCCTTAGTTCC	Stop (-21)
cadB_KO_F	CATCATGACCCGGACTCCAAATTCAAAAATGAAATTAGGAGAAGAGCATGTGTGTAGGCTGGAGCTGCTTC	Start (+3)
cadB_KO_R	AAAGGAGGAGCCTCGGAAAATACTTTTAATTAATGTGCGTTAGACGCGGTCATATGAATATCCTCCTTAGTTCC	Stop (-21)
rpoS_synP8_F	GCTCGTATTAATCATCCGGCTGCTATACTTAATAGACGTCGACTCTCGAGTGAGATTG	See Figure 2.3.1
rpoS_synP8_R	TCTATTAAGTATAGCAGCCGGATGATTAATACGAGCGGATCCCCGGGTATTCTTGAAGACG	See Figure 2.3.1

2.3.4.2 Preparation of DNA for PCR

If the target sequence was present on the chromosome or if the purpose of the PCR was to provide confirmation based on fragment size, DNA was isolated by putting 10µl of overnight culture into a 200µl PCR tube and lysing it using a thermocycler at 98^{o}C for 10 mins before reducing the temperature to 4^{o}C. This lysate was then diluted by 1:10 by adding 90µl of nuclease free water. Otherwise if the target was present on the plasmid a miniprep was done and this was then diluted to achieve a 20ng/µl concentration. Either of these diluted mixtures were then used for subsequent PCR reactions.

2.3.4.3 Standard confirmation PCR

This method was used to confirm the presence/absence of mutations within a given sample either by differences in fragment sizes or by sequencing. For each PCR reaction, a 50µl reaction mixture was created as explained in Table 2.5. Oligonucleotides were diluted from a 100µM stock to a final concentration of 10µM. PCR was then performed using a thermocycler following the appropriate protocol described within Table 2.6. Once reaction had occurred this was then loaded onto a gel described in section 2.3.4.4.

Table 2.5: Reaction mixture for standard confirmation PCR.

Reagents	**Amount**	**Description**
Bioline 2x MyTaq ™ Red Mix	25µl	Taq polymerase 2x master mix contains red loading dye
10µM Primer 1	2.5µl	Primer designed for specific target
10µM Primer 2	2.5µl	Primer designed for specific target.
DNA template	2µl	DNA template see section 2.3.4.1
Nuclease free H_2O	18 µl	

Table 2.6: Thermocycler protocol for Standard PCR confirmation. The annealing step was determined as 3°C lower than the melting temperatures determined by benchling's Tm calculator (https://benchling.com).

CYCLE STEP	**TEMP**	**TIME**	**CYCLES**
Initial Denaturation	95°C	2 minutes	1
Denaturation	95°C	30 seconds	30
Annealing	42°C -65°C	30 seconds	
Extension	72°C	30 sec/kb	
Final Extension	72°C	5 minutes	1
Hold	4°C	∞	

2.3.4.4 PCR for cloning

Where replication of a template was required to be of high accuracy for cloning purposes, Phusion® High-Fidelity DNA Polymerase (New England Biolabs) was used. A 50µl PCR master mix was created according to Table 2.7. Template DNA was obtained as described in section 2.3.4.1. The master mix was then loaded into a thermocycler and the following protocol was performed as described in Table 2.8. These samples were then visualized upon an agarose gel as described in section 2.3.4.4.

Table 2.7: Reaction mixture for PCR with Phusion master mix. Taken according to manufacture description.

Reagents	Amount	Description
Phusion DNA polymerase	0.5µl	Phusion polymerase. 100 units/ul
10mM dNTPs	1µl	Nucleotides required for PCR reaction. (Bioline)
5X Phusion HF Buffer	10ul	
10µM Primer 1	2.5µl	Primer designed for specific target
10µM Primer 2	2.5µl	Primer designed for specific target.
DNA template	2µl	DNA template see section 2.3.4.1
DMSO	1.5ul	
Nuclease free H_2O	30 µl	

Table 2.8: Thermocycler procedure for using Phusion High fidelity polymerase (New England Biolabs). Annealing temperature was determined by 4°C lower than the melting temperatures determined by benchling's Tm calculator (https://benchling.com).

CYCLE STEP	TEMP	TIME	CYCLES
Initial Denaturation	98°C	30 seconds	1
Denaturation	98°C	10 seconds	30
Annealing	45°C - 65°C	30 seconds	
Extension	72°C	30 sec/kb	
Final Extension	72°C	5 -10 minutes	1
Hold	4°C	∞	

2.3.4.5 Agarose Gel and Electrophoresis

Agarose gels were made using TAE buffer and agarose with Midori Green Advance (Nippon) DNA stain added at a 1:25000 dilution. The amount of agarose varied depending on the difference in the size of the fragments expected. If a higher resolution was required due to small differences between the expected size of the fragments (> 500bp) a 2% w/v agarose gel was used. Otherwise if the size of expected fragments had a larger difference (greater than 500bp), or as a confirmation that amplification had worked, a 1% w/v gel was used.

For electrophoresis, agarose gels were submerged in TAE buffer in an electrophoresis tank, 10µl of PCR samples were loaded. The gels were run at a voltage of 90V – 120V for a pre-determined amount of time, depending on the expected sizes of fragments. After running these gels were then visualized in a Biorad Gel Doc XR+.

2.3.4.6 Extraction of a PCR product in a gel

Gel extraction was performed using a QIAquick gel extraction kit (Qiagen) following the instructions provided. Gel extractions were only performed when multiple fragments were observed on a gel. In order to get maximum yield of product a 1% agarose gel was loaded with 40µl of PCR sample. Samples were eluted in 30µl of Buffer EB. Samples which had been gel extracted were then processed through the PCR purification for downstream utility.

This method works in a similar way to the plasmid preparation protocol described in section 2.3.1 It involves melting agarose containing target DNA in a high salt buffer, which allows for adsorption to a silica membrane. This then allows for impurities to be washed out of the sample using ethanol, before eluting DNA off in a low salt buffer (In this study, Buffer EB).

2.3.4.7 PCR Purification by kit

PCR purification was done using a QIAquick PCR purification kit (Qiagen). If a single band was observed on an agarose gel the remainder of the PCR reaction mixture was PCR purified and was eluted in 30µl of buffer EB (Tris-HCl, pH 8.5). This method works using a similar process to that described in sections 2.3.1 and 2.3.4.6.

2.3.4.8 Quantification of DNA by Nanodrop.

Quantification of DNA was performed using a Thermos Scientific NanoDrop^{TM} ND-1000 Spectrophotometer according to the manufacturer's directions.

2.3.4.9 Sanger sequencing.

When required, PCR fragments and plasmid DNA were sequenced by Sanger sequencing using a service from Source Biosciences. Both strands of the desired product were sequenced. Samples were diluted to the company's instructions of 5µl primer at 3.2pmol/µl and 5µl of either PCR product at 10ng/µl or Plasmid at 100ng/µl.

2.3.5 Tapestation Quantification of DNA

In order to determine fragment size, DNA was quantified using an Agilent 2200 Tapestation instrument with Agilent High Sensitivity D5000 ScreenTape System. This was conducted by Faye Hughes as an in house service.

2.3.6 Transduction by P1 Phage

Transduction by P1 phage allows the transfer of genetic material of interest, which can be selected for, from one strain of *E. coli* to another. This method is only possible due to the 'sloppy' mechanism the phage uses to package its genetic material during infection, which causes, in some instances, host genetic DNA to be incorporated instead (Thomason *et al*., 2007). Host DNA packaged into a P1 phage can then be transferred, by P1 infection, into a recipient strain. Upon infection the incoming DNA can then be incorporated into the recipient's chromosome by homologous recombination. It is estimated that up to 100 kb of host DNA can be transferred.

2.3.6.1 Preparation of P1 lysates

Preparation of P1 lysates were conducted as described in Thomason *et al*. (2007). This step involves infecting *E. coli* , containing a mutation of interest, during early exponential phase and then allowing the culture grow to mid-exponential phase allowing the phage to replicate. This culture is then lysed using chloroform and SDS. It is then centrifuged to remove debris leaving a lysate containing phage some of which will contain genetic material required to be transferred.

2.3.6.2 P1 Transduction

P1 transduction was performed as described in (Thomason *et al*., 2007). This method involves infecting stationary phase cells from overnight culture with P1 lysate, in P1 buffer containing Mg^{2+} and Ca^{2+} ions to promote phage adsorption. This infection was conducted using various dilutions of phage allowing an adequate time for one infection cycle typically (30 – 45 mins at 37˚C). The addition of sodium citrate prevents secondary infections, as P1 phage requires calcium ions to be present in order to be able to adsorb onto the cell. Sodium citrate chelates these ions, preventing any additional phage infections after the initial infection. These bacteria are then spread onto agar plates with a selective marker for the desired genetic material (usually an antibiotic marker in or close to the gene of interest) and incubated. Bacteria which contain the genetic material will then form colonies. These colonies are then picked and restreaked onto plates containing sodium citrate at least 2 times to prevent residual phage contamination.

2.3.6.3 Determining of P1 titre

If P1 transduction was not successful, P1 stocks would be titred again this was performed according to (Thomason *et al*., 2007). *E.coli* cultures were grown to exponential phase, then diluted 1 in 10 with a soft agar to get 2.5 ml (see section 2.2.1) and pouring this onto LBA plates, spreading the mixture evenly across the plate by tilting. This will ensure a lawn of *E. coli* throughout the plate which will be fixed within agar. Phage stocks were then serially diluted up to a 10^9 dilution, with 10µl of each dilution spotted 3 times onto the plate. Using soft agar ensures that phage infections are limited to local cells, ensuring that plaque formation will show local infectivity. Dilutions which formed an amount of countable plaques were then used to estimate the titre. If the P1 titre was less than 10^9 phage/ml, P1 stock was discarded and a fresh P1 stock was created.

2.3.7 Gibson assembly

Sen, (2018), reported the use of an RpoS activity plasmid gifted by Gisela Klauck. This plasmid reports RpoS activity through the expression of a “superfolder” GFP (sfGFP) which is regulated by a synthetic promoter specific to RpoS. However upon further inspection glycerol stocks thought to contain RpoS activity plasmid, were actually found to contain an RpoD activity plasmid using the synP21 promoter sequence described in (Klauck *et al*., 2018). Since the RpoS specific promoter sequence, (Designated as synP8) and RpoD (synP21) sequence only differed by ~35 bp, the simplest method to obtain a RpoS plasmid was to reconstruct the plasmid using a Gibson assembly method.

As shown in Figure 2.3.1A both RpoD and RpoS reporter plasmid described in (Klauck *et al*., 2018) have the same plasmid backbone of pXG10-SF with an ampicillin cassette inserted allowing for a ~35bp specific RpoS or RpoD promoter fusion to a sfGFP. Therefore, since the specific sequence which needed changing was so small and the plasmid backbone size was reported at ~4.6 kb, oligonucleotides were designed to allow the amplification of the entire plasmid backbone and to create overhangs specific to the promoter sequence synP8, totaling 35 bp of homology at either end of the plasmid (Figure 2.3.1B). The reason for this overhang was to allow for recircularization of the plasmid with the correct promoter sequence. The plasmid backbone with overhangs was then amplified using these custom primers and Phusion polymerase and gel extracted (Sections 2.3.4.4 and 2.3.4.6). A Dpn1 digestion was then performed, to ensure the removal of all cell derived plasmid (RpoD reporter plasmid). This construct was then quantified using Qubit. A Gilson Assembly Cloning kit (New England Biolabs) was then used according to the manufacturer’s instructions for a 1-2 Fragment assembly to re-circularize the plasmid with the correct promoter sequence. Once performed this was then transformed into Top10 electrocompetent cells made as described in section 2.5.

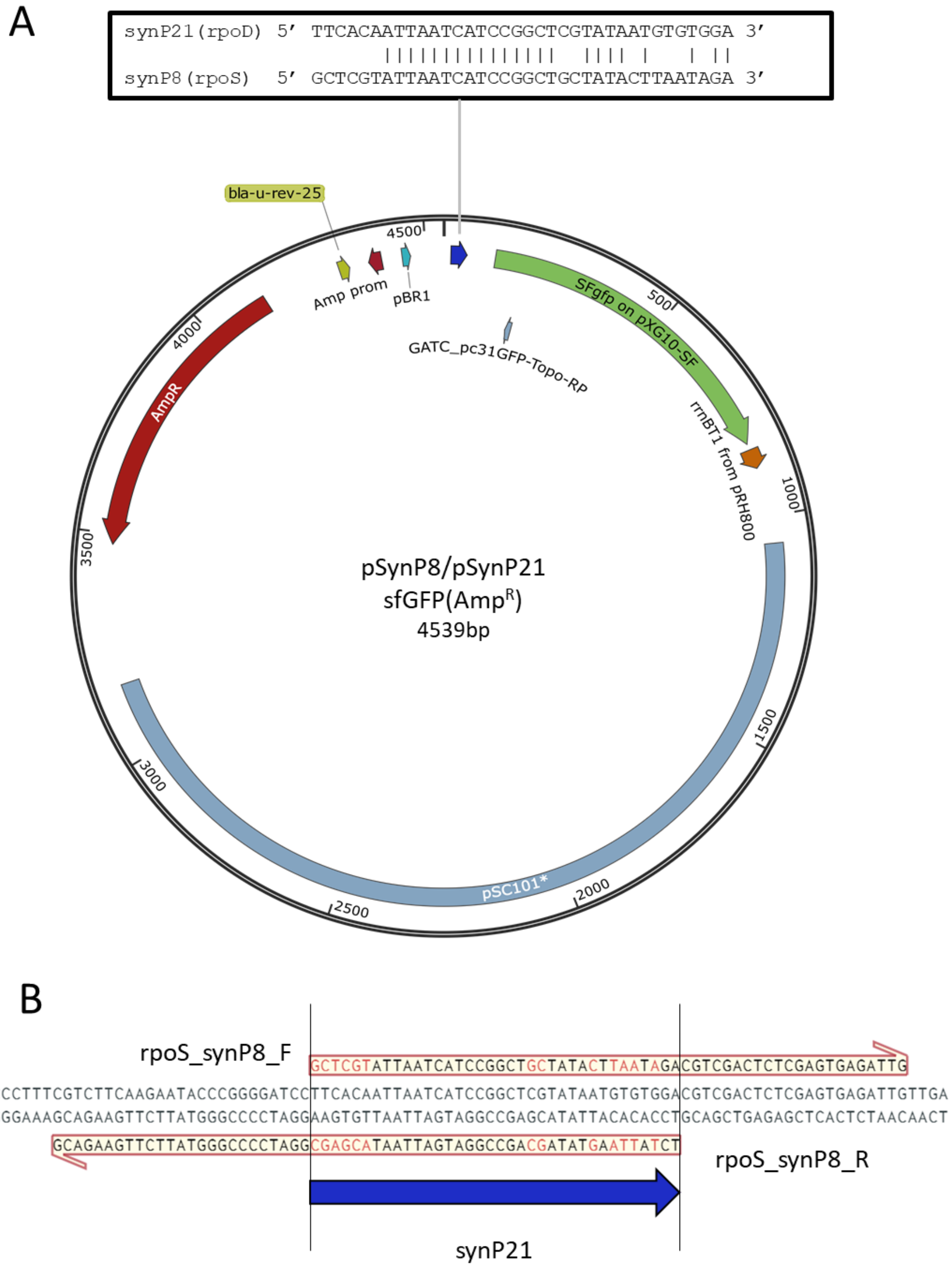

A
synP21(rpoD) 5’ TTCACAATTAATCATCCGGCTCGTATAATGTGTGGA 3’
||||||||||||||| |||| | | ||
synP8(rpoS) 5’ GCTCGTATTAATCATCCGGCTGCTATACTTAATAGA 3’
bla-u-rev-25
4500
Amp prom
pBR1
GATC_pc31GFP-Topo-RP
SFgfp on pXG10-SF
500
rrnBT1 from pRH800
1000
1500
2000
pSC101*
2500
3000
3500
4000
AmpR
pSynP8/pSynP21
sfGFP(Amp^R)
4539bp
B
rpoS_synP8_F
GCTCGTATTAATCATCCGGCTGCTATACTTAATAGACGTCGACTCTCGAGTGAGATTG
GGCCCTTTCGTCTTCAAGAATACCCGGGGATCCTTCACAATTAATCATCCGGCTCGTATAATGTGTGGACGTCGACTCTCGAGTGAGATTGTTGA
CCGGGAAAGCAGAAGTTCTTATGGGCCCCTAGGAAGTGTTAATTAGTAGGCCGAGCATATTACACACCTGCAGCTGAGAGCTCACTCTAACAACT
GCAGAAGTTCTTATGGGCCCCTAGGCGAGCATAATTAGTAGGCCGACGATATGAATTATCT
rpoS_synP8_R
synP21

Figure 2.3.1: Construction of RpoS reporter plasmid from RpoD reporter plasmid. A) Plasmid map of RpoS and RpoD reporter plasmids (psynP8 and psynP21 respectively). Plasmids differ only by small change in promoter sequences, synP8 or synP21 regulating expression of sfGFP. B) Oligonucleotides (yellow) used alter synP21 promoter sequence to synP8. Nucleotides highlighted in red indicate change.

2.3.8 Construction of gene deletion by lambda red recombination.

In certain cases, construction of a deletion of a gene was necessary, where the deletion could not be found in our version of the Keio collection (Baba *et al*., 2006). Therefore in order to construct a deletion, the protocol described in (Datsenko and Wanner, 2000) was used, using the plasmid pKD46 which encodes the lambda Red genes *exo*, *bet* and *gam* under the control of an arabinose inducible promoter. In order to make a deletion, primers were designed to create 50bp homologous flanking regions of the target gene (Table 2.4). The relevant homologous sequence was taken from (Baba *et al*., 2006) and was checked against the MG1655 genome, to ensure homology. This sequence was then used to allow amplification of a kanamycin cassette flanked by FRT sites present in pkd4.

As an overview, this method works by expressing three genes *exo*, *gam*, and *bet* from the lambda phage on a temperature sensitive plasmid within a desired *E. coli* strain. This method involves the creation of double stranded DNA (dsDNA) fragment by PCR of a deletion cassette flanked with homologous regions of the target gene to be deleted. This is then transformed into a target *E.coli* strain. Expression of *gam* prevents a linear DNA fragment from being degraded by the host nucleases while the presence of *exo*, that encodes a 5’ -> 3’ exonuclease, creates single stranded sticky ends, covering the homology region. These single stranded fragments can then be used by Bet which protects single stranded DNA and allows the annealing of the single stranded target within the cell. Homologous recombination can then occur causing the targeted incorporation of a deletion cassette into the chromosomal DNA.

2.3.9 High Throughput Sequencing (HTS) and analysis

All High Throughput Sequencing was performed by MicrobesNG (Birmingham, UK) to a depth stated literature. For variant calling, on either a population or clonal sample, the whole genome resequencing pipeline *breseq* was used on its default settings (Deatherage and Barrick, 2014, Barrick *et al*., 2014).

2.4 Competition Experiments

2.4.1 Experimental design

In this study competition experiments were done under a variety of conditions (Chapter 4). Each competition experiment requires two strains which can easily be distinguished by a single marker that does not otherwise affect their relative fitness. In this study the ability of strains to utilize lactose was used as a marker. Single colonies of each strain (Table 2.1) were separately inoculated into LB and incubated overnight and OD_{600} obtained. Unless otherwise stated, each competition was started by adding a volume of culture required to obtain a starting OD_{600} value of 0.025, for each strain. Therefore, a starting OD_{600} of 0.05 was achieved in the initial culture. The culture was then allowed to grow under defined condition. At stated timepoints, CFU/ml for each strain in culture was obtained by serial dilution in LB and plating on MacConkey lactose agar to achieve countable colonies. MacConkey lactose plates were grown at 30^{o}C for 12-16 hours to avoid overgrowth of plates. In all competition experiments CFU/ml was calculated both at 0h and at a time specified. This data was then used to calculate the relative fitness of the two strains as described in the below section.

2.4.2 Determining relative fitness

To analyse on the results of a competition experiment, typically a relative fitness metric is used. This metric provides a measure of how well one strain did compared to another in direct competition under a defined condition and set time, which is defined as the “relative fitness”. Within the literature, several different methods exist to calculate relative fitness. Therefore, in this study, several different methods to calculate relative fitness were identified and investigated in order to use the best metric to use with our data. Table 2.9 details the equations used in this investigation and a brief explanation of what the equation calculates. Each equation considers the relative fitness of strain A compared against strain B. A detailed explanation of these metrics, and determination of which metric was used overall in this study, is described in Appendix 2.

Table 2.9: Different relative fitness metrics identified in the literature.

Relative fitness metric	**Equation**	**Definition**	**Reference.**
Equation 2.1: Relative proportion of strains (RP)	$RP = \frac{A_t}{A_t + B_t}$	RP determined for each time point. Calculates the relative proportion of strain A within a population of strain A and B at time (t).	N/A
Equation 2.2: Malthusian growth model.	$N_t = N_0 e^{rt}$	Population at time (t) is dependent on exponential growth at growth rate (r) for a determined time (t). N = population size, r = population growth rate, t = time. Assumes bacterial populations are in constant exponential growth.	(Lenski *et al.*, 1991)
Equation 2.3: Malthusian parameter (r).	$r = \ln\left(\frac{N_t}{N_0}\right) day^{-1}$	Growth rate assuming Malthusian growth model. Calculated as the natural log of final population (N) at time (t) divisible by the initial population at time 0. The rate is dependent upon the time difference.	(Lenski *et al.*, 1991)
Equation 2.4: Relative fitness (W).	$W = \frac{r_a}{r_b} = \frac{\ln\left(\frac{A_t}{A_0}\right)}{\ln\left(\frac{B_t}{B_0}\right)}$	Relative fitness calculated as the ratio of the Malthusian growth parameters of strain A over Strain B. It is a dimensionless metric	(Lenski *et al.*, 1991)

Relative fitness metric	**Equation**	**Definition**	**Reference.**
Equation 2.5: Selection rate (S)	$s = r_a - r_b$ $S = \ln\left(\frac{A_t}{A_0}\right)\ day^{-1} - \ln\left(\frac{B_t}{B_0}\right)\ day^{-1}$	Selection rate is calculated as the difference between the Malthusian growth parameters of strain A against strain B.	(Travisano and Lenski, 1996)
Equation 2.6: Relative fitness estimated using Selection rate (s) between two strain A and B:	$\mathrm{W} = 1 + \frac{s}{r_{ab}}$ $\mathrm{W} = 1 + \frac{\ln\left(\frac{A_t}{A_0}\right)\ day^{-1} - \ln\left(\frac{B_t}{B_0}\right)\ day^{-1}}{\ln\left(\frac{A_t + B_t}{A_0 + B_0}\right) day^{-1}}$	A ratio of selection rate divided by the growth parameter of the entire population (Strain A + B) is calculated. This is then added to 1 (no change in fitness) to give an estimate of relative fitness	(Lenski *et al.*, 1991)
Equation 2.7: Selection Coefficient 1	$S_{(1)} = \frac{r_a - r_b}{r_b} = \frac{\ln\left(\frac{A_t}{A_0}\right) - \ln\left(\frac{B_t}{B_0}\right)}{\ln\left(\frac{B_t}{B_0}\right)}$	Proportional increase in growth rate strain A compared to Strain B	(Barrett *et al.*, 2006)
Equation 2.8: Selection Coefficient 2	$S_{(2)} = \frac{\ln\left(\frac{A_t/B_t}{A_0/B_0}\right)}{log_2\left(\frac{A_t + B_t}{A_0 + B_0}\right)}$	Creates a performance metric as the natural log of the ratio of strain A over strain B at a specific time (t) against the ratio of Strain A over B at time 0. Selection co-efficient can then be calculated using this metric per generation. By calculating the number of generations	(McDonald, 2019) Box 1

2.5 Bacterial Transformation

2.5.1 Electrocompetent cells

To make bacteria electrocompetent, overnight cultures were inoculated into fresh LB at a 1:100 dilution. This was then grown into mid exponential phase (~ 0.4 – 0.6 OD_{600}) at the required temperature (30°C / 37°C) in a suitable flask with shaking to ensure aerobic growth. Once the expected OD_{600} was reached, the culture was put onto ice for 30 mins, before being centrifuged at 4,000 x g for 10 mins to form a pellet. A wash cycle was then done by removing the supernatant and resuspending the pellet into same culture volume of 10% v/v glycerol. This wash step was done 3 times reducing the amount of ice cold 10% glycerol by half each wash cycle. At the final wash cycle, cells were pelleted again and resuspended in 10% v/v ice cold glycerol at a 100-fold concentration compared to the original culture used. An example of this would be an initial culture of 10ml would result in 100µl of competent cells. These competent cells were then aliquoted into 50µl aliquots for transformation and stored at -80^{o}C.

2.5.2 Transformation of electrocompetent cells

If frozen, electrocompetent cells were thawed on ice. DNA was the added and mixed by gently flicking the vial. This mixture was then transferred by pipetting to ice cold 1 mm electroporation cuvettes. These were subsequently electroporated at 1750V using a Eppendorf Eporator®, with 1ml of SOC media added immediately after. This was then incubated at a relevant temperature for 1 hour before plating 100µl onto LBA with the relevant selective marker.

2.6 Reporter Assays

2.6.1 β-galactosidase assay

Before a β-galactosidase assay was conducted, an ONPG solution was prepared by resuspending 0.04g of ONPG (Sigma Aldrich) in 10ml of Z buffer (60mM - Na_2HPO_4, 40mM – NaH_2PO_4, 10mM KCl, 1mM – $MgSO_4$, pH7). β-mercaptoethanol was not added for the majority of this study. When stated, 35µl β-mercaptoethanol was added to 10ml of both Z buffer and ONPG solution. Further information considering the use of β-mercaptoethanol can be found within section 3.7.4.1. Both the ONPG solution and Z buffer were incubated at 37°C for 30 mins prior to use.

To start the experiment, overnight cultures were diluted into 6 ml fresh LB in 30 ml Universals to achieve a 0.05 OD_{600}. A timepoint 0h was taken by removing 1ml of culture and measuring OD600. For any other time point, 200µl of culture was required to take measurements (100µl for OD_{600} measurement, 100µl for β-galactosidase assay). For each sample, once the OD_{600} was taken, 100µl of culture and 900µl of prewarmed Z buffer was pipetted into a 2ml microcentrifuge tube. The culture was then subsequently lysed through the addition of 100µl 0.1% w/v SDS and 100µl chloroform, before being vortexed for at least 30 seconds to ensure total lysis. 200µl of ONPG solution was then added before a timer was started and the sample incubated at 37°C. These samples were then incubated until a faint yellow colour was observed. Once the colour change was observed 500µl of Na_2CO_3 was added and the time was recorded. The sample was spun down in a microcentrifuge (7000 x g for 30s). 1ml of this sample was then transferred to a 1ml cuvette and OD_{420} was taken using 1ml of Z buffer as a blank. Miller units could then be calculated based on Equation 2.9.

Equation 2.9: Calculation of Miller units.

$$Miller\ Units = \frac{OD_{420}}{OD_{600} * Volume\ (ml) * Time\ (mins)}$$

2.6.2 GFP reporter assay

Strains containing a GFP reporter were grown overnight then diluted at a 0.05 OD600 in 5ml of LB. At each time point 1ml sample was taken, pelleted in a microcentrifuge (8000 x g, 3 mins) and resuspended in 1ml 1X PBS. At each time point, 200µl sample was taken and put into a Corstar® 96-Well Clear Bottom Black plate. A 200µl blank of LB was also plated. These samples were then read in a Molecular Devices Filtermax™ F5 multimode microplate reader recording first the OD_{600} before recording GFP fluorescence using an excitation wavelength of 475nm and emission of 535nm.

2.7 Construction of a Transposon library and TraDIS

The method described in this section was created and optimized by members of Ian Henderson's lab from the protocol described by (Langridge *et al.*, 2009, Goodall *et al.*, 2018), principally Ashley Robinson, Emily Goodall and Karl Dunne. Their descriptions can be found in their respective theses. I implemented further analysis scripts based on calculation of RPKM, and scripts that formatted GFF3 files to allow other genomes to be used with our analysis scripts. The details of these methods are described below.

2.7.1 Construction of a Transposon library

2.7.1.1 Making high efficiency electrocompetent cells.

The library in *E. coli* K-12 MG1655 was constructed as described below; this was done in duplicate. Initially, 7ml of LB in 30ml universal flasks were inoculated with a single colony of MG1655 and grown overnight at 37°C with aeration. From there, 800ml of 2xTY broth was inoculated with 7ml of overnight culture and allowed to grow to a target OD_{600} 0.3 – 0.4 at 37°C with aeration. Once the target OD_{600} was reached, 600ml of culture was divided into 12 x 50ml Falcon tubes and held on ice for 30 mins. From this point effort was made to keep all cultures and reagents at < 4°C until transformation. These cultures were pelleted by centrifugation at 5000 x *g* at 4°C for 10 mins. Once pelleted, the supernatant was removed, and the pellet was resuspended gently in 50 ml 10% glycerol. The bacteria were then centrifuged again at 5000 x *g* at 4°C for 10 mins, the supernatant was again removed, and the pellet resuspended in 50ml 10% glycerol. A further centrifugation at 5000 x *g* at 4°C for 10 mins was performed and the cells were gently resuspended in 25ml 10% glycerol. From there two lots of 25ml of resuspended cells were combined into a single 50ml Falcon tube. This wash protocol was repeated until 12 X 50 ml falcon tubes were combined into 3 X 50 ml falcon tubes. A further wash step was then done combining the 3 falcon tubes into a single tube by reducing the amount of 10% glycerol the cells were resuspended in. A further wash step was performed using 50ml 10% glycerol and the pellet was then resuspended in 1 ml 10% ice cold glycerol and split in 200µl aliquots in 1.5ml microcentrifuge tubes.

2.7.1.2 Transformation with transposome.

Transformation to construct the library was then conducted, with 0.2µl of EZ-Tn5^{TM} <KAN-2> Tnp Transposome added to each of the 10 X 200µl aliquots of electrocompetent cells, which were then incubated for 30 mins on ice. Aliquots were then transferred to 2mm electroporation cuvettes and electroporated at 22kv. Following electroporation, 1ml of SOC media was added to each transformation and cells were incubated for recovery at 37^{o}C for 2 hours. Further to this, each transformation was diluted with 5ml of LB broth. These 10 recovered cultures were then spread onto ~350 LBA plates supplemented with 30µg/ml Kanamycin. Roughly 200µl of recovered culture was plated on each plate which were then incubated at 37^{o}C overnight. After incubation, agar plates were then scraped of viable colonies by adding 500µl of LB broth. This was then collected and pooled before 100% glycerol was added to obtain a 10% glycerol final concentration of pooled transposon library which was then stored at -80^{o}C for future use.

2.7.2 TraDIS

This section describes the method used in this study to perform TraDIS. An overall concept of the method is described in Figure 2.7.1.

2.7.2.1 Isolation of DNA and Fragmentation

DNA was isolated from pellets of the initial transposon library (ITL) or from a culture after an experiment using approximately 1 x 10^{9} cells. Isolation of DNA was performed using the Stratec RTP Bacteria DNA Mini kit, following protocol 2: Isolation of DNA from bacterial pellets (1 x 10^{9} bacterial cells).

The DNA extracted was then quantified using Qubit^{TM} dsDNA HS Assay kit (Invitrogen) using 1µl of isolated DNA (Section 2.3.5). Isolated DNA was then diluted to obtain a 2µg/ml concentration in 500µl of Nuclease free H_2O. These samples were then fragmented by acoustic shearing using a Diagenode bioruptor® plus. The protocol used with the machine was 30s on, 90s off for 13 cycles at low intensity to give an average fragment size of ~350bp. Samples were then condensed to a volume of 55.5µl using a concentrator (Eppendorf).

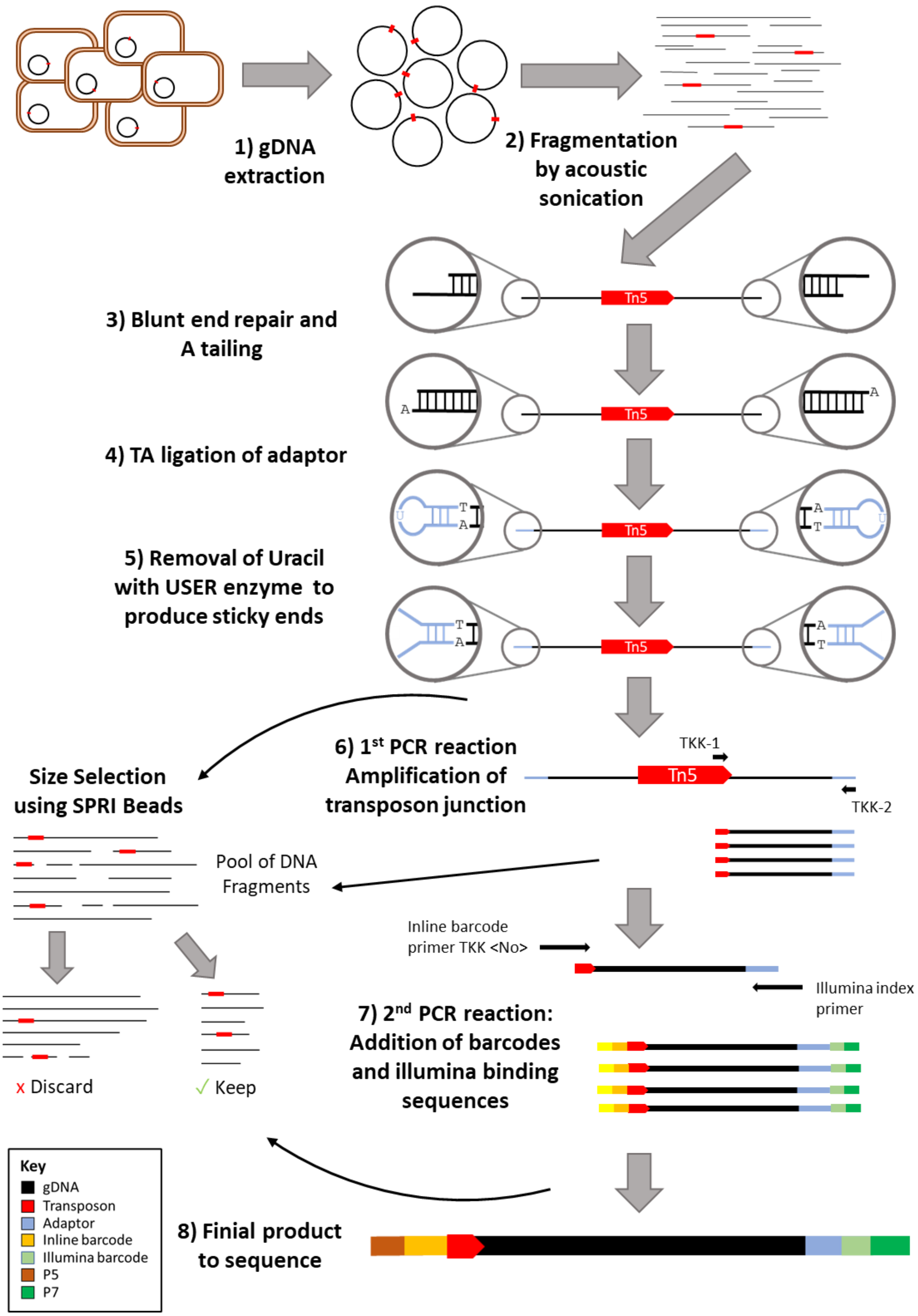

1) gDNA extraction
2) Fragmentation by acoustic sonication
3) Blunt end repair and A tailing
4) TA ligation of adaptor
5) Removal of Uracil with USER enzyme to produce sticky ends
Tn5
A
T
U
6) 1st PCR reaction Amplification of transposon junction
TKK-1
TKK-2
Size Selection using SPRI Beads
Pool of DNA Fragments
x Discard
✓ Keep
Inline barcode primer TKK <No>
Illumina index primer
7) 2nd PCR reaction: Addition of barcodes and illumina binding sequences
8) Finial product to sequence
Key
gDNA
Transposon
Adaptor
Inline barcode
Illumina barcode
P5
P7

Figure 2.7.1: Overview of gDNA library preparation for Transposon Directed Insertion Site sequencing. Note size selection steps using SPRI beads occurs three times, directly after step 5, 6 and 7.

2.7.2.2 Construction of SPRI beads

In the protocol described below, low concentrations of fragmented gDNA of transposon libraries are size selected and cleaned using Solid-phase reversible immobilization (SPRI). This method involves the use of SPRI beads which can bind to DNA in a size specific manner and are paramagnetic which allows for easy separation of beads from solution. As an overview, SPRI beads work due to the reversable binding of nucleic acids to carboxyl groups bound to paramagnetic beads in the presence of PEG and salt which act as a "crowding agent". These beads are paramagnetic, meaning they are only magnetic in the presence of a magnet, which prevents clumping when present in solution and allows controlled separation beads from solution in the presence of a magnet. The requirement for PEG and salt is necessary as it allows for nucleic acids to bind to the carboxyl molecules on the beads. Altering the relative ratios of PEG and salt concentration can alter the preference of the size of nucleic acids which bind to the carboxyl groups on the paramagnetic beads. Due to this preferential binding size selection of DNA can be differed by altering the PEG and DNA concentrations with high PEG/DNA ratios causing preference to small nucleic acid fragments.

During the first year of this study a commercial brand Agencourt AMPure XP (Beckman Coulter) was used. However, this was then swapped for SPRI beads which were made in house using a protocol described by Jolivet and Foley, (2015). The reason for this was by constructing these beads in house cost could be reduced by nearly 100 fold. This protocol essentially involves the steps of washing and resuspending commercial carboxyl coated paramagnetic beads in a buffer containing PEG and Salt. This study followed the protocol specifically for DNA, through the addition Tris base and EDTA to achieve a final concentration of 10mM and 1mM, respectively. Once constructed each batch of SPRI beads were validated using 1kb hyperladder (Bioline) as a representation of a fragmented gDNA library and compared to Agencourt AMPure XP beads (Beckman Coulter) using different ratios of DNA to beads. From there amounts of SPRI beads were altered accordingly in the protocol below to achieve the correct size selections.

2.7.2.3 Preparation of Transposon Sequencing Library

Sequencing libraries were prepared as highlighted in Figure 2.7.1 using the NEBNext Ultra^{TM} DNA Library Prep Kit for Illumina. The kit protocol was followed for blunt end repair of fragmented DNA and TA ligation of adapter to DNA fragments. Uracil base was then removed from the adaptor using USER enzyme to create overhangs. Fragments of 250bp were then selected using AMPure XP beads (SPRI beads) following the manufacturer's instructions or as described above in 2.7.2.2, before a final elution into 17µl buffer EB (Qiagen) was performed (DeAngelis *et al*., 1995). At this point an additional amplification step was done to amplify DNA fragments containing transposon and neighboring gDNA. A PCR reaction using oligonucleotides designed for Transposon and adapter sequences was done to enrich for these fragments. The following reagents used in the 1st PCR reaction are described in Table 2.10.

Table 2.10: Reagents used in 1st Amplification step to enrich for transposon containing fragments.

Reagents	**Amount**	**Description**
NEBNext Q5 Hot Start Hifi PCR Master Mix	25µl	Q5 polymerase provided in the NEBNext Ultra Kit
10µM - TKK_F1	2.5µl	Primer designed for the Kanamycin transposon sequence. See Table 2.14 Table 2.14
10µM - TKK_R1	2.5µl	Primer designed for the Adapter sequence. See Table 2.14
Sample	15ul	Sample containing fragmented DNA ~250bp fragments
Nuclease free H_2O	5µl	

PCR was then performed on this reaction mix using a thermocycler with the protocol described in Table 2.11

Table 2.11: Thermocycler protocol for 1st PCR reaction

CYCLE STEP	TEMP	TIME	CYCLES
Initial Denaturation	98°C	48 seconds	1
Denaturation	98°C	15 seconds	10
Annealing	65°C	30 seconds	
Extension	72°C	30 seconds	
Final Extension	72°C	1 minute	1
Hold	4°C	∞	

This PCR product was then purified using SPRI beads as described in the NEBNext Ultra protocol with a final elution into 17µl buffer EB (Quiagen). A second PCR step was then done to add custom inline and illumina barcodes and P5 and P7 homology sequences to allow for compatibility with the Miseq flow cell (Box 1). Barcodes were used to be able to multiplex independent samples on the same Miseq run, with custom inline barcodes present on the transposon side of the DNA fragment and Illumina barcodes on the gDNA. In addition to multiplexing samples, custom inline barcodes were of different sizes to stagger the transposon during the initial stages of sequencing to prevent the same transposon sequence from being read by all sequences at the same time (BOX1). The following reagents used in the second PCR reaction are described in Table 2.12.

Table 2.12: Reagents used in 2nd PCR amplification step, to prepare DNA fragments to be sequenced by the Miseq

Reagents	**Amount**	**Description**
NEBNext Q5 Hot Start Hifi PCR Master Mix	25µl	Q5 polymerase provided in the NEBNext Ultra Kit
10µM - TKK_Barcode	2.5µl	Adapts MiSeq flow cell sequence and a custom made in line barcode which allows staggering of the transposon and additional multiplexing of samples. Sequences of barcodes can be found in Table 2.14.
10µM – NEBNext Multiplex oligos for Illumina (Set1 or Set 2)	2.5µl	Adapts Miseq flow cell sequence and allows for multiplexing of samples.
Sample	15µl	Sample containing fragmented DNA ~250bp fragments
Nuclease free H_2O	5 µl	

From here the following cycle was performed on a thermocycler, the cycle is described in Table 2.13.

Table 2.13: Thermocycler protocol for 2nd PCR reaction.

CYCLE STEP	TEMP	TIME	CYCLES
Initial Denaturation	98°C	48 seconds	1
Denaturation	98°C	15 seconds	20
Annealing	65°C	30 seconds	
Extension	72°C	30 seconds	
Final Extension	72°C	1 minute	1
Hold	4°C	∞	

A final PCR product clean-up was done using AMPure XP beads (SPRI beads) as described in the NEBNext Ultra protocol with a final elution into 33µl buffer EB (Quiagen). Finally, 32µl of this elution was transferred to a DNase free 1.5ml microcentrifuge tube, called a library prepped sample and stored at -20°C until required.

Table 2.14: Primers used within the TraDIS method. These primers are only designed for the library prep of a kanamycin cassette Tn5 transposon. Differences in barcode sequences used for the inline barcodes are underlined

ID	Sequence
TKK_F1	ACCTGCAGGCATGCAAGCTTCAGG
TKK_R1	GACTGGAGTTCAGACGTGTGCTCTTCCGATC
TKK 6.3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTACGTAAGCTTCAGGGTTGAGATGTGTA
TKK 7.2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTAGCTAGCTTCAGGGTTGAGATGTGTA
TKK 7.4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAGCTAGAGCTTCAGGGTTGAGATGTGTA
TKK 8.2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATGCATGCAGCTTCAGGGTTGAGATGTGTA
TKK 8.3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCATGCATGAGCTTCAGGGTTGAGATGTGTA
TKK 8.4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTACGAGCTTCAGGGTTGAGATGTGTA
TKK 9.2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATCGATCGAAGCTTCAGGGTTGAGATGTGTA
TKK 9.3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGATCGATAGCTTCAGGGTTGAGATGTGTA
TKK 9.4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATCGATCAGCTTCAGGGTTGAGATGTGTA

2.7.2.4 Quantification of Sequencing Library by qPCR

To quantify the amount of library prepared DNA present within a TraDIS library prep, qPCR was employed using KAPA Library Quant Kit (Illumina) Universal qPCR mix. For each genomic library prep, three replicates were performed using dilutions of 1:50000 and 1:500000 in 10µM Tris HCl pH 8. A 10µl reaction mix for each dilution was then generated using 1µl of

DNA dilution and 9µl qPCR mix. An Agilent Mx3005P qPCR machine was used to perform the cycles described in Table 2.15. The software AriaMix (v1.1) was then used to quantify the amount of DNA in the genomic library prep.

Table 2.15: Protocol followed for qPCR to quantify library preps

CYCLE STEP	TEMP	TIME	CYCLES
Initial Denaturation	95°C	5 minutes	1
Denaturation	95°C	30 seconds	35
Annealing + Extension	60°C	30 seconds	
Dissociation melt curve	65°C to 95°C		1

2.7.2.5 Sequencing using the Miseq

Once the concentration of the genomic library preps had been determined, this was then used to dilute the genomic library prep to an 8nM stock. From here, 1.5µl of each sample was combined in a 200µl nuclease free PCR tube to create a pool of samples to be sequenced. Following the protocols for illumina this 2µl pool was then denatured using 0.1N NaOH and 98°C temperature and diluted to 16pM final concentration. with a control of 3% PhiX library and loaded onto a MiSeq 150 cycle v3 cartridge. 900 K clusters per mm^2 was considered a successful run. More information on sequencing a transposon library on an Illumina platform can be found in Appendix 1.

2.7.3 The Short Term Selection Experiment (STSE)

The Short Term Selection Experiment (STSE) was performed using the exact conditions described in Sen (2018). Twelve samples each containing 5ml of LB broth (6 at pH 7 and 6 at pH 4.5) in 30 ml universal tubes were inoculated with the MG1655 Transposon library to obtain an initial OD$_{600}$ of 0.05 (~ 500 copies of each insertion). Cultures were then incubated for 22 – 24 hours and then transferred to 15ml falcon tubes and centrifuged at 5000 x *g* at 4^{o}C for 10 mins. Once pelleted, cells were then resuspended in 2 ml of fresh LB broth at the relevant pH and 100µl of this resuspension was transferred to 30ml universal tubes containing 4.9ml of fresh LB at the relevant pH, which were then incubated at 37^{o}C for a further 22-24 hours with aeration. Overall, 1/20th of the bacterial culture was transferred to fresh media. Of the remaining culture, a fossil record was created using 500 µl of culture mixed with 500µl of 50% glycerol, to give a final concentration of 25% glycerol. This was stored at -80^{o}C. The remaining culture was then pelleted using centrifugation at 5000 x *g* at 4^{o}C for 10 mins with the supernatant removed and stored at -20^{o}C for DNA extraction. Passaging occurred using this method each day for 10 days.

2.7.4 Identification of transposon containing reads and genome alignment.

The data analysis pipeline here was initially created by Ashley Robinson and has been modified by Emily Goodall and myself. Initially illumina reads generated by sequencing on a Miseq platform are provided in a fastq format which is then run though this pipeline. All analysis was performed on a linux server provided by the MRC Cloud Infrastructure for Bioinformatics (CLIMB) project (Connor *et al.*, 2016). The whole pipeline including scripts and software can be found in the folder designated 'TraDIS_analysis_v1.8m1'. An overview of how this pipeline works is described in Figure 2.8.1.

In each fastq file, sequence checks are performed using the fastx Toolkit (version: 0.0.14). Reads are quality controlled to look for reads specifically containing inline barcode sequences and transposon sequences Figure 2.8.1A (Pearson *et al.*, 1997). Initially, reads containing the in-line barcodes are identified and pooled using fastx_barcode_splitter.pl. The barcode sequence is then removed using fastx_trimmer. The pool of reads, with barcodes identified and trimmed, then undergoes a further two rounds of filtering and trimming of sequence specific to the transposon. These two filtering steps consider the transposon sequence in two parts, depending on whether or not the sequence was homologous to oligonucleotides used in amplification of transposon fgDNA during TraDIS (Section 2.7.2.3, Table 2.14, Figure 2.8.1A). Therefore, the first filter consists of a 22bp sequence which corresponds to the primers used in amplification, due to these 3 mismatches are tolerated. The second filtering covers a 10bp sequence is amplified, therefore a more stringent cutoff value 1 mismatch is used. The output of these repetitive rounds of filtering and trimming is a fastq file containing reads of gDNA directly adjacent to the location of a transposon insertion and which due to the inline barcode are confirmed to be associated with the correct sample (Figure 2.8.1). A final quality control check is done to remove reads of poor quality or which are less than 20bp using trimmomatic (version: 0.36) (Bolger *et al.*, 2014).

These reads were then aligned to a fasta file of an annotated genome using the bwa mem aligner using default settings (version: 0.7.17-r1188)(Li and Durbin, 2009). All soft clipped reads were then binned. Aligned reads were then sorted and indexed using samtools and converted into a.bed file using the bamToBed function from Bedtools (version: 2.27.1) (Quinlan and Hall, 2010). Once converted, the annotated genome GFF3 Formatted file was intersected over the.bed to give annotation information to the aligned reads by Bedtools intersectBed function. Custom python scripts were used to collate the amount of unique transposon insertions for each coding region on the genome and to calculate the number of reads corresponding to each insertion and to each coding region. Reads that were ambiguously mapped were counted only once being randomly assigned to one location. While chimeric reads were not included within the analysis. Further processing was performed to calculate an insertion index score (Unique insertion points/ Gene length(bp)) and Reads per Kilobase per Million (RPKM) value for each coding sequence.

2.8 TraDIS analysis

At present several different methods to analyze transposon data exist, however, within the literature no consensus of which method to use has been reached. Therefore, in order to determine the analysis most suitable for this study several methods were taken from the literature based on two categories. The first refers to methods which define essentiality. These methods look specifically to define essential genes, usually based on the presence or absence of insertions. The second focus more on relative fitness, although some of the pipelines described below also consider essentiality. These different methods typically focus upon difference in normalized read counts between two conditions. Although a few different pipelines exist for this analysis, only analyses which were able identify fitness advantageous insertions as well as fitness disadvantageous insertions were considered.

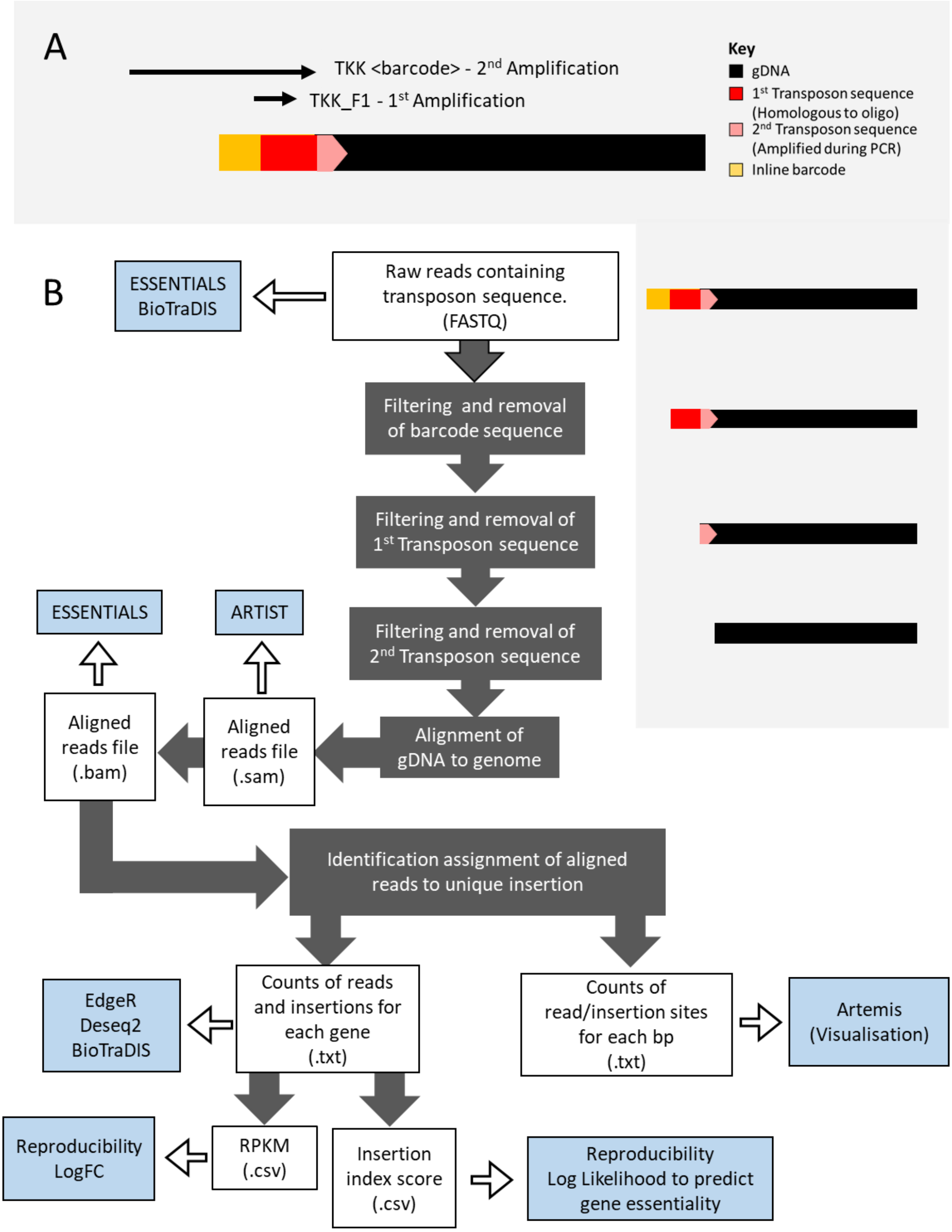

A
TKK <barcode> - 2nd Amplification
TKK_F1 - 1st Amplification
Key
gDNA
1st Transposon sequence (Homologous to oligo)
2nd Transposon sequence (Amplified during PCR)
Inline barcode
B
ESSENTIALS BioTraDIS
Raw reads containing transposon sequence. (FASTQ)
Filtering and removal of barcode sequence
Filtering and removal of 1st Transposon sequence
Filtering and removal of 2nd Transposon sequence
Alignment of gDNA to genome
ESSENTIALS
ARTIST
Aligned reads file (.bam)
Aligned reads file (.sam)
Identification assignment of aligned reads to unique insertion
EdgeR Deseq2 BioTraDIS
Counts of reads and insertions for each gene (.txt)
Counts of read/insertion sites for each bp (.txt)
Artemis (Visualisation)
Reproducibility LogFC
RPKM (.csv)
Insertion index score (.csv)
Reproducibility Log Likelihood to predict gene essentiality

Figure 2.8.1: Overview of TraDIS analysis pipeline. A) Diagram of Transposon containing gDNA fragment after sequencing and the filtering of each sequencing during subsequent analysis (grey box). Transposon sequence present in fragments is split into two sections. 1) Transposon sequence homologous to amplification oligonucleotides and 2) Transposon sequence which is amplified during amplification steps. B) Overview of TraDIS analysis in house script and input files for other TraDIS analysis pipelines. Files taken from this analysis (white boxes) to be used in other Transposon analysis (Blue boxes) are highlighted by white arrows.

2.8.1 Read Depth analysis

To be able to estimate the number of reads required to ensure sufficient coverage of each insertion present in each sample of the STSE, the following analysis was conducted. This analysis follows the procedure described in (Goodall *et al.*, 2018) and the R script used to conduct the analysis can be found in "Dropbox: Sequencing saturation script.R ". As an overview, this method assumes that the original transposon library used in other experiments has been sequenced to a depth greater than required to identify all insertions present within the library. Therefore, the maximum amount of insertions calculated represents the total amount of unique insertions present in the transposon library. Using this assumption, the number of transposon insertions identified given a defined number of reads can be estimated using Equation 2.10**Error! Reference source not found.**. The purpose of the script is to run multiple iterations of this equation using different amounts of total reads to create a curve estimating the number of transposon insertions identified compared to the number of reads used. In addition, a saturation percentage, which corresponds to the amount of library covered by a set number of reads, can be calculated by dividing the results of the above analysis by the total amount of insertions. Using this metric, a target for sequencing depth can be determined by defining a set saturation percentage.

Equation 2.10: Formula to estimate the proportion of transposon library identified, with a determined number of reads. S represent sample size I,e the total amount of insertions identified in a transposon library. n corresponds to the number of reads used

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

2.8.2 TraDIS pipelines

2.8.2.1 Log$_2$ likelihood scores to predict essentiality

The analysis performed predicts essential genes according to the method outlined in Goodall *et al.,* (2018). This was done using the script 'Langridge_R_Script_FD_v1.2.R' which was gifted by Emily Goodall and was conducted within R (version 3.5.9; http://www.r-project.org). An insertion index score is calculated for each gene based on the presence or absence of insertion sites (Equation 2.11). A full explanation on how this script was performed is described in Section 4.4.1.1.

Equation 2.11: Insertion index score calculation per gene. Total unique insertion refers to the number of unique insertions points present within a gene.

$$Insertion\ index^{gene} = \frac{Total\ Unique\ Insertions^{gene}}{gene\ length\ (bp)}$$

When considering fitness of detrimental genes in the outgrowth experiment performed in chapter 5, data from replicates were combined together. A comparison between each condition was then performed, removing essential genes and genes identified as ambiguous within the original library, to genes identified as essential within the outgrowth. This comparison is performed in 'Log-likelihood_comparision.R'

2.8.2.2 ARTIST

ARTIST is a MATLAB based package which provides two tools, EL- ARTIST and Con – ARTIST. It was designed for Tn-seq data, however it can also be used with TraDIS data (Pritchard *et al.*, 2014). El -ARTIST was run in MATLAB, splitting the genome up into 50bp windows; underrepresented windows were determined using a pval < 0.05 cutoff. With a sliding window value used of 7 consecutive windows. All further comparative analysis was conducted when the annotations of the genome were overlayed onto the Artist analysis. A full description of the methodology can be found in section 4.4.1.3

Con-artist is part of the Artist pipeline used to detect conditionally essential and transposon enrichments in an annotation independent manner within transposon sequencing experiments (Pritchard *et al.*, 2014). Scripts for this analysis can be found in the 'Artist_analysis' folder. The analysis was conducted in MATLAB. with aligned reads file in a SAM format being imputed for each replicate and condition of the STSE as well as the original ITL. The genome was split into 50 bp windows. Artist was run for each replicate separately. A gene which differed in read count was only considered if it was classed as fully or domain different in all three replicates.

2.8.2.3 EdgeR

The EdgeR package (version: 3.26.4) was performed within R using raw read counts for each gene, as well as the total read counts for each replicate generated in the TraDIS analysis pipeline described in section 2.7.2 (Robinson *et al.*, 2010, McCarthy *et al.*, 2012). The script 'EdgeR_script. R', was used to perform this analysis. In this analysis two conditions, pH 7 and pH 4.5, were each compared against the ITL. A false discovery rate using a Benjamini Hochberg procedure was applied, a p.val < 0.05 and a LogFC +/- 2 to determine significant differences.

2.8.2.4 Deseq2

As with edgeR, Deseq2 is a differential expression analysis software designed for RNA seq (Love *et al.*, 2014). Deseq2 was performed within R using raw read counts generated for each gene in the TraDIS pipeline (Figure 2.8.1). Raw read counts for all samples were placed into a data frame. In our data set no prefiltering of reads was conducted. Deseq2 log FC and statistical testing is done according to the script " Deseq2.R". When performing Deseq2, all steps were performed according to the paper specifications following the Wald test as a method for statistical testing (Love *et al.*, 2014). The P values generated from this test were then adjusted using a Benjamini-Hochberg procedure. To determine significant differences a pval < 0.05 with a LogFC +/- 2 was used.

2.8.2.5 Essentials

ESSENTIALS is an open source, TraDIS analysis pipeline which was web based (Zomer *et al.*, 2012). However, since early 2019 , the website was not actively maintained and since 2020 is unavailable. The source code can be obtained from https://trac.nbic.nl/essentials/. ESSENTIALS is a pipeline which begins by filtering reads by barcode and transposon sequence to ensure a read corresponds to transposon sequence before being removed to allow only gDNA present (Zomer *et al.*, 2012). In the case of our data which has a three step filtering process, step 1 and step 2 were combined together (barcode and transposon homologous to amplification primer) and step3 (transposon not covered by amplification primer) using fastx_trimmer. This was then aligned to the *Escherichia coli* MG1655 genome using PASS. Read counts and Insertion sites for each gene were then identified. ESSENTIALS has the ability to not consider the last 10% of each gene. In this study the last 10% was not removed. A LOESS regression was then performed on the read count data to normalise for insertion and read bias seen within Transposon data. This data was then passed through edgeR which was used to identify changes in fitness between two samples.

In addition, edgeR was made to identify essential genes through a comparison of expected read counts with actual measured read counts. This used a pvalue < 0.05 and LogFC +/- 2 as a cutoff. More information identification of essential genes using ESSENTIALS can be found in Section 4.4.1.4

2.8.2.6 BioTraDIS

BioTraDIS like ESSENTIALS is a pipeline designed for the easy analysis of TraDIS data (Barquist *et al*., 2016). This was run on the CLIMB (Connor *et al*., 2016). This can be used to identify essential genes using a log-likelihood method described in section 2.8.2.1 and 4.4.1.2. Note when using this method to compare different samples, the genes identified as both essential and ambiguous in the MG1655 library, were removed from the genes identified as essential in the comparative sample

In addition, BioTraDIS also utilizes the RNAseq software, edgeR, as a method to compare read count differences between two different samples (Robinson *et al*., 2010). A false discovery rate using the Benjamini Hochberg procedure was applied and significant genes were determined using a LogFC cutoff of +/- 2 and an adjusted pvalue < 0.05.

2.8.3 Gene ontology and KEGG pathway enrichment analysis

GO ontology and Kegg ontology enrichment analysis was performed using the R webservice David according to the package defaults. In all cases an enrichment was only classed as significant if it had a p value of < 0.05 after adjustment using a Benjamini Hochberg procedure. The scripts used to perform this analysis can be found in the GO_KEGG_ontology_analysis.R script.

2.8.4 GSEA

Gene set enrichment analysis (GSEA) was performed using the GSEA software downloaded from http://www.gsea-msigdb.org/gsea/index.jsp, (GSEA 4.1.0)(Subramanian *et al*., 2005, Mootha *et al*., 2003). With databases created based on outputs from RegulonDB and Kegg pathway.

2.8.5 Comparative analysis

For Figure drawing and comparative analysis, was performed in R and Microsoft Excel. In addition, two online browsers were used. These were Venny 2.1 which is a software for comparison of and drawing of 3 and 4 way Venn diagrams using gene lists (https://bioinfogp.cnb.csic.es/tools/venny/). To perform larger comparisons and to draw upset plots the Intervene tool, using its ShinyApp companion (Khan and Mathelier, 2017). (https://asntech.shinyapps.io/intervene/)

Chapter 3

Analysis of an Evolution Experiment

3.1 Introduction

pH stress is considered in a variety of discrete categories. Firstly, pH as stress is split in two, low and high pH stress, in which these two stresses can be considered separately and each will elicit a unique stress upon the cell, and in turn an appropriate response. Focusing on low pH this is then further is categorized into two, referring to extreme and mild low pH stress. Although the boundaries of these groups are not clearly defined in terms of pH, extreme pH stress can be defined as a 'strong' stress, which typically causes lethality within bacteria, and as such the cell response is to survive the environment until the stress is reduced and the organism is able to proliferate. Alternatively, mild pH stress can be considered as a low pH environment where the stress is 'weaker' to the point where the organism is able to proliferate within the environment.

The current understanding of how *E. coli* survives under extreme acid stress (< pH 3) has been intensively studied. Indeed under aerobic conditions the main mechanisms involved in eliciting a response have been mostly identified, as has the complex regulatory network required for an appropriate response (Lin *et al.*, 1995, Lin *et al.*, 1996, Castanie-Cornet *et al.*, 1999, Johnson *et al.*, 2014, De Biase *et al.*, 1999, Hersh *et al.*, 1996). Although there are still some unknowns in the response of extreme acid stress, mild acid stress has been less focused upon and in turn the understanding behind the mechanism of this response is less understood.

Laboratory based evolution has been very successful as a tool to understand an organism's response under multiple different stresses. With regards to acid stress, only a few laboratory based evolution experiments have been conducted, with very few focusing on acid stress (Hayes *et al.*, 2006, Johnson *et al.*, 2014, Harden *et al.*, 2015, He *et al.*, 2017, Du *et al.*, 2020). To address this a laboratory based evolution experiment was conducted to study *E. coli* response to mild acid stress. This lab based evolution experiment was performed by Hrishiraj Sen and the outcome of his results can be found in his Thesis (Sen, 2018).

3.2 An overview: Laboratory based evolution at pH 4.5

In order to provide the reader context, for the remainder of this chapter, a description of how the evolution experiment was designed and conducted by Sen is provided. It is important to stress, that none of the work undertaken in this subheading was conducted by myself, all information provided within this Section was taken from Sen (2018).

The overall goal of this evolution experiment was to look at how *E. coli* K-12 MG1655 adapts to acid stress when it is still able to proliferate (i.e., under mild acid stress). Therefore, in order to determine a pH which the evolution would be conducted at, growth curves were performed over 8 hours using MG1655 in 100ml of unbuffered LB in 250ml Erlenmeyer flasks using different pH's ranging from pH 3 to pH 7 (Figure 3.2.1A). Using these growth curves, it was determined that pH 4.5 would be the pH used in the evolution experiment as it was the lowest pH where growth was observed within the 8 hours.

Using this pH, an evolution experiment was then conducted. This experiment was designed by setting up 6 independent populations derived from a single *E. coli* MG1655 clonal ancestor. Each individual population was started by inoculating 5 ml of unbuffered LB at pH 4.5 with an overnight culture of MG1655 to achieve an OD_{600} of 0.05. This was then allowed to grow for 24h at 37°C with shaking before being passaged into fresh media. To passage, Sen pelleted the total 5ml culture and resuspended in 2ml of LB before passaging 100µl of the resuspension into 4.9ml of fresh unbuffered LB at pH 4.5 which was then allowed to grow for 24 hours at 37°C. By doing this, the amount of overnight culture passaged into fresh media constituted a 1 in 20 dilution. This experiment was then allowed to continue with repeated passaging every 24 hours for 5 months, totaling approximately 740 generations. During the evolution experiment a fossil record was created for each culture, by glycerol stocking populations every 15 days. Upon completion of the evolution experiment each individual population was isolated and labeled E1P – E6P. In addition to this a single clone was isolated from each population and labeled E1A-E6A as a representation of each population

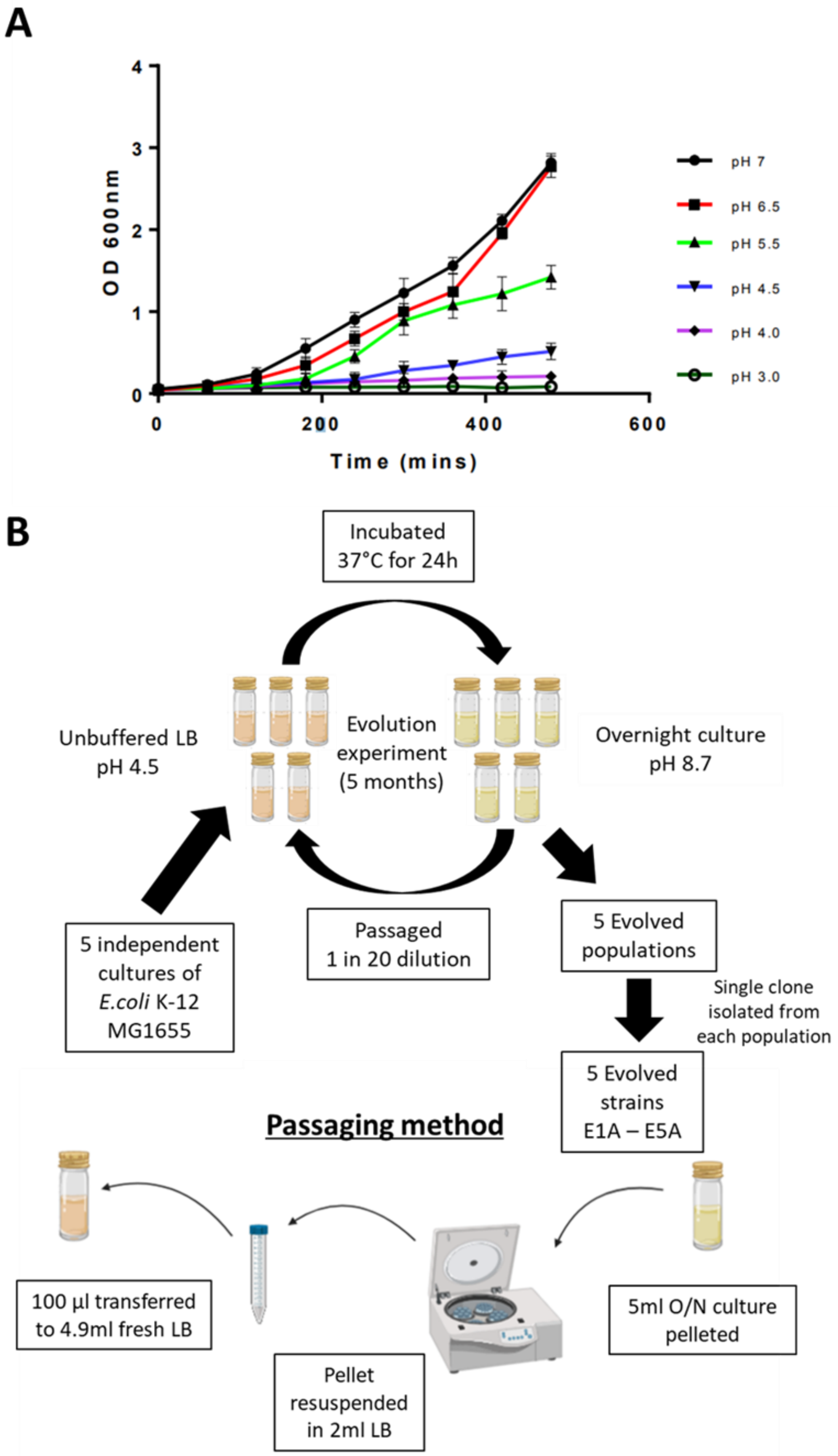


Figure 3.2.1: Evolution experiment conducted by Sen. A - Figure taken from (Sen, 2018). 8 hour growth curves of *E. coli* MG1655 comparing growth at different pH. OD_{600} was taken every hour with 3 replicates for each pH condition. Each replicate consisted of growth in 100ml of unbuffered LB in 250ml Erlenmeyer flasks at different pH. Note y axis is not Log10 to show differences in growth. B – Schematic diagram of Evolution experiment conducted by Sen (2018).

3.3 Consideration of the use of unbuffered LB

The description of this evolution experiment above sets out an experiment to study adaptation at mild acid stress. However, within this experiment several additional stresses could be identified in which the organism could adapt to, such as stress incurred due to culture being at stationary phase. In practice this stress is present in any batch culture evolution experiment grown for 24 hours, however within this experiment this is more prominent as the passaging dilution equates to 1/20 of overnight culture transferred to fresh media so the time taken to reach stationary phase was relatively short.

In addition to this the evolution experiment was conducted in unbuffered LB. LB consists of yeast extract, tryptone and NaCl, and this medium itself can elicit particular stresses in the cell. These components create an environment which main carbon source is that of oligopeptide chains and free amino acids and only a finite amount the divalent cations such as Mg^{2+} are present, which can cause stress upon the cell (Sezonov *et al*., 2007, Nikaido, 2009). It is the utilization of oligopeptides, however which highlights a potential issue with the design of the evolution experiment. In order for the cell to obtain energy in LB the cell must catabolize amino acids. This catabolism leads to ammonia being produced as a byproduct, which in turn is then released into the environment as a waste product. This release of ammonia can cause an increase in environmental pH. Indeed, an increase in pH has been demonstrated in media in which *E. coli* was grown in LBK, a media similar to LB but replacing NaCl with KCl, started at pH 7 but changed to a pH greater than 8.5 after 24 hours (Stancik *et al*., 2002). As mentioned before Sen used unbuffered LB, meaning that the pH during the experiment wasn't controlled and was likely to rise.

To investigate whether the environment pH did indeed change within the conditions of the evolution experiment, the pH was measured in the current study in 5ml of LB initially at pH 4.5 after 24 hours growth of *E. coli* MG1655, which resulted in a final environmental pH 8.74 ± 0.03. In addition, as a comparison this experiment was also conducted in unbuffed LB starting at pH 7, which yielded a final pH of 9.01 ± 0.02. This result demonstrated that an environmental pH change was occurring during each passage of the evolution experiment, indicating that the evolution experiment was not conducted in an environment constantly at pH 4.5, but a dynamic change in pH (pH 4.5 – pH 8.7).

To investigate the dynamics of the environmental pH change further, 50ml cultures of *E. coli* MG1655 in unbuffered LB at pH 7 and pH 4.5 *E. coli* were grown over 24 hours monitoring the OD_{600} and environmental pH after every hour timepoint (Figure 3.3.1). The results of these experiments demonstrated that at pH 4.5 growth in the culture was slower compared to pH 7, indicating that the cell was under more stress. However, after 24 hours the final OD600 in each culture was comparable suggesting that the low pH stress was not a limiting factor for growth. In addition, the change in environmental pH lagged behind growth, suggesting that environmental pH change is dependent upon growth. With the largest environmental pH change occurring between 3h -6h when each culture was in mid exponential phase – early stationary phase (pH change at 3h - 6h: pH4.5 = pH 4.7 - pH 7.1, pH 7 = pH 6.9 - pH 8.1) When the culture was in stationary phase the pH then generally increased over the remaining time with unbuffered LB at pH 4.5 finally reaching pH 8.7 and pH 7 finally reaching pH 9.0, highly similar to the final environmental pH observed in 5ml cultures. In addition as seen in (Stancik *et al*., 2002), in the first three hours of the pH 7 culture there is a slight drop in pH from pH7 – pH6.8, suggesting a small amount of glucose is present in the media. Therefore, a drop in environmental pH is observed due to catabolism of glucose producing acetate. An initial delay in the change in environmental pH was also observed in the pH 4.5 culture, which suggests catabolism of glucose or a similar carbon source, although a pH drop was not observed.

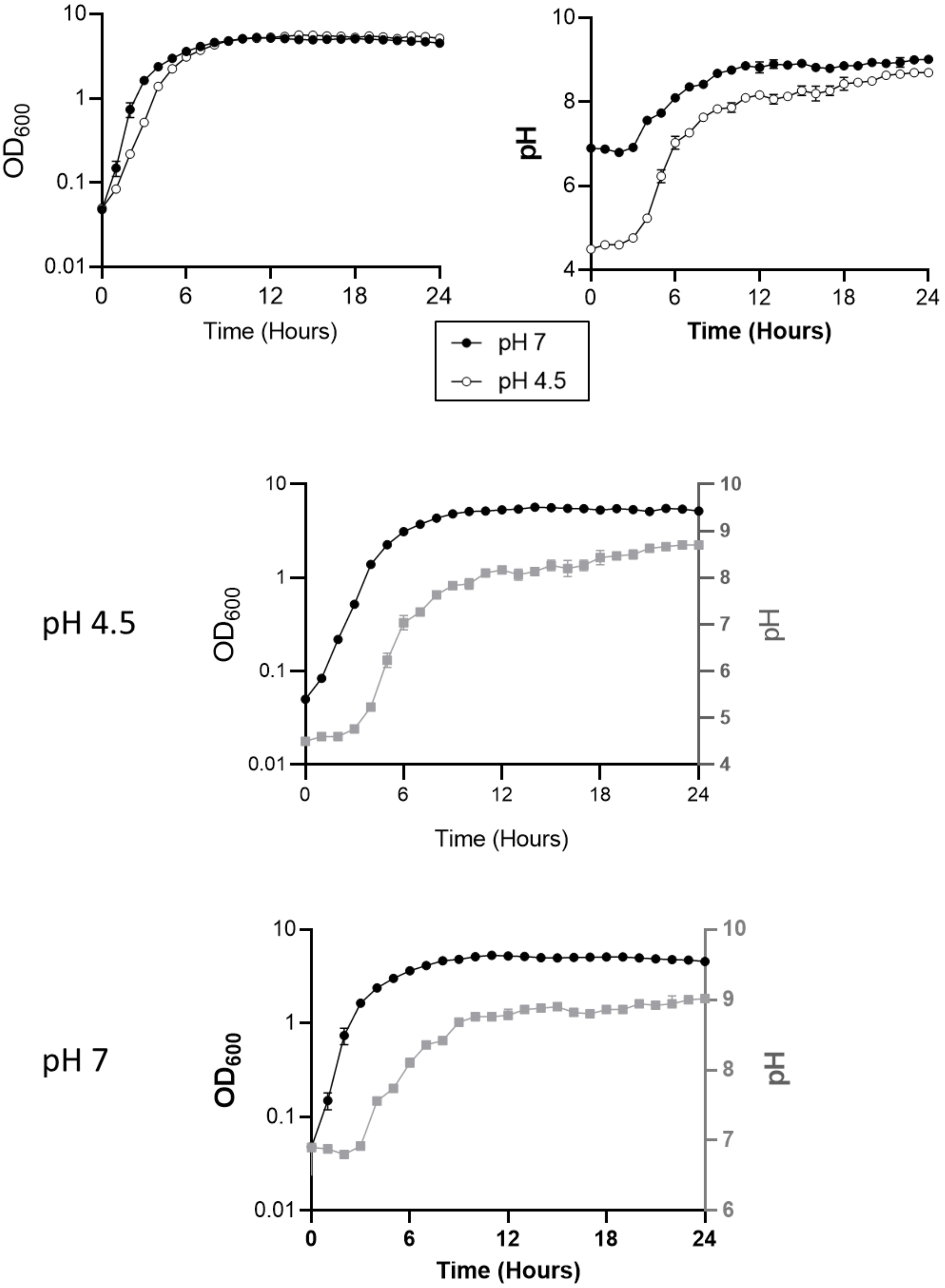


Figure 3.3.1: Growth over 24 hours of *E. coli* MG1655 , tracking both the environmental pH and OD_{600} , hourly in cultures of 50ml unbuffered LB with a starting pH at either pH 4.5 or pH7. Each culture was diluted to a starting 0.05 OD_{600} from an overnight culture and incubated at 37^{o}C with shaking. Top graphs represent a comparison of environmental pH and OD600 change at pH7 or pH4.5. Bottom graphs represent, growth (black) and environmental pH (grey) alongside each other. The data presented is the mean +/- SD, in unbuffered LB at either pH 4.5 or pH 7, using 3 and 4 replicates respectively

3.4 Genome sequencing of Evolved strains

Populations and evolved clonal isolates from the Sen evolution experiment described above were then sequenced using an illumina platform to produced 150 bp paired end reads by MicrobesNG. The initial analysis variant-calling analysis using VarScan was performed by Dr Joshua Quick and can be found in Sen (2018). However, a common problem with variant calling using short reads is the difficulty in identifying structural variants within a genome. Therefore, a limitation of the previous analysis was that only SNP/base substitutions and small indels were identified. Since this analysis was conducted, several other variant calling pipelines have been made available which are also able to effectively to identify larger structural variation, such as movement of insertion sequences and larger deletions. Once such package is the whole genome resequencing pipeline *breseq*, developed by Jeffrey Barrick to analyse the LTEE. Its pipeline is able to identify reads at the junction/s of structural variants (Deatherage and Barrick, 2014). It does this by developing a method for identifying reads which have a chimeric alignment, where a single read will align to two different regions on the genome. By identifying these reads, which may correspond to a junction site of a structural variation, the variation can then be predicted based on the read evidence identified. Using *breseq*, the evolved clones, populations and MG1655 ancestor were reanalyzed aligning reads with the MG1655 genome (Accession no: U00096.3). The results of this analysis are described below.

Although in this chapter, 6 evolved populations are mentioned, after further inspection of the variant calling from *breseq* for the E6 population and clone, issues arose with the validity of E6A clone and its origin with the E6P population. In addition, RNAseq data described in Section 3.9 was not generated for the clonal strain. Therefore, it was decided to not include further analysis of this strain and the analysis below refers only to the populations E1P – E5P and the clonal strains derived from them E1A – E5A.

3.4.1 Evolved Clonal Strains

To begin, each evolved strain was labeled E1A-E5A, and the MG1655 wild type ancestor, were each sequenced to a 30X depth. The results of this analysis can be found in Table 3.1 Overall, the MG1655 ancestor used in this study revealed 3 additional mutations when compared to the published MG1655 genome (U00096.3). These were: an additional IS1 insertion within *yeaJ*, a non-synonymous mutation in *cbtA* and GC insertion in the intergenic region between *gltP* and *yjcO*. Considering the effect that these mutations will have upon the strain, *yeaJ* is predicted to encode a diguanylate cyclase, a deletion in *yeaJ* has been shown to increase motility by 2 fold compared to MG1655 wildtype (Pesavento *et al*., 2008, Sanchez-Torres *et al*., 2011). Therefore, an *IS1* insertion in *yeaJ* could lead to increased motility. The nonsynonymous mutation in *cbtA* is thought to have very little effect. The gene *cbtA* encodes a toxin which belongs to the type IV toxin-antitoxin CbtA-CbeA. This toxin-antitoxin system works by an indirect mechanism where CbeA does not directly interact with CbtA but interacts with the targets of CbtA. Therefore, mutations seen in *cbtA*, will have very little effect upon the cell, as CbeA will counteract the toxin effect. The *gltP* and *yjcO* intergenic +GC insertion is also predicted to have null effects upon the phenotype as the insertion is downstream of both gene transcriptional terminal sites. Overall, the mutations seen within the wildtype are predicted to have very little if any effect on the outcome of the evolution experiment.

Having identified the mutations within the WT, variants that arose within the evolved strains were then identified. A total of 26 variants were identified within the evolved strains in 17 different gene or intergenic regions, with the exception of two large deletions in the evolved strain E1A and E4A. The 5 evolved strains generally showed a similar evolutionary trajectory, with mutations associated with *arcA* and *cytR* occurring in all 5 evolved strains and a mutation in *rpoA* within 4 evolved strains (described in more detail below). In order to confirm the genotypes described by *breseq* in the evolved strains, experimental validation was performed on 10 different mutations using PCR to assess structural variants and PCR and Sanger sequencing for nonstructural (example shown in Figure 3.4.2).

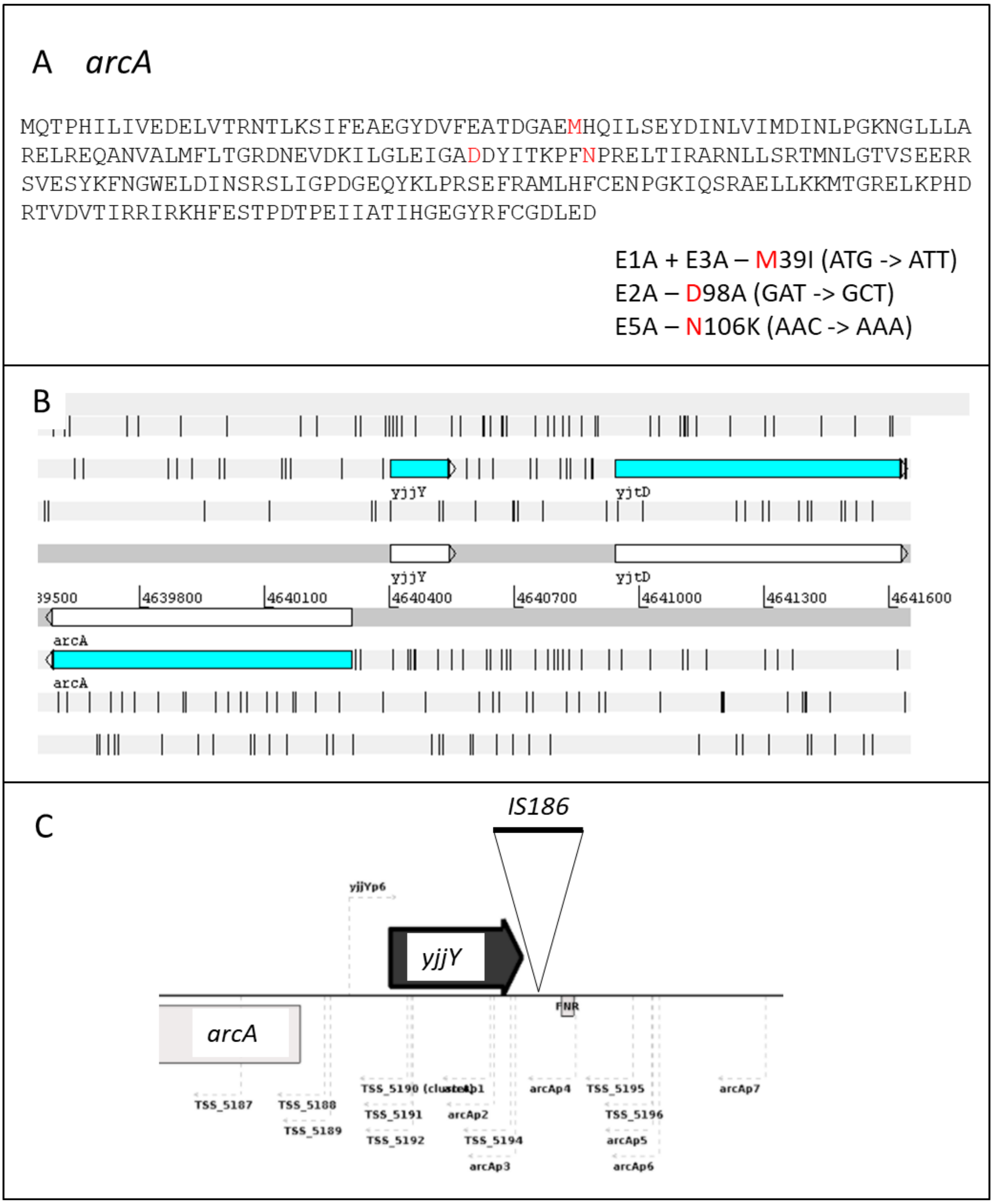
Table 3.1: Genotypes of *E. coli* MG1655 ancestor and evolved clonal strains E1A – E5A.

Position	E1A	E2A	E3A	E4A	E5A	Mutation	Annotation	Gene	Description
1196325	■					Δ 15,088 bp	IS-mediated	*[icd] - mcrA*	Large deletion : *[icd], ymfD, ymfE, lit, intE, xisE, ymfI, ymfJ, cohE, croE, ymfL, ymfM, oweE, aaaE, ymfR, beeE, jayE, ymfQ, stfP, tfaP, tfaE, stfE, pinE, mcrA*
1969584				■		Δ 8,919 bp	IS-mediated	*[tap] - flhD*	*Large deletion : [tap], tar, cheW, cheA, motB, motA, flhC, flhD*
3035546		■				T → C	T246A (ACG → GCG)	*prfB* ←	peptide chain release factor RF-2
3377183				■		C → T	D80N (GAT → AAT)	*sspA* ←	stringent starvation protein A
3440150	■	■	■		■	T → G	N294H (AAC → CAC)	*rpoA* ←	R polymerase alpha subunit
3544482	■		■			IS186 (-) +5 bp::Δ1 bp	coding (359-363 / 771nt)	*bioH* ←	pimeloyl-ACP methyl ester carboxylesterase
3888473	■					C → A	F13L (TTC → TTA)	*tnaC* →	tryptophase leader peptide
3888626			■			IS5 (-) +4 bp	intergenic (+117 / -101)	*tnaC* → / → *tnaA*	tryptophase leader peptide/tryptophase/L-cysteine desulfhydrase; PLP-dependent
3888873					■	IS1 (+) +9 bp	coding (144-152 / 1416nt)	*tnaA* →	tryptophase/L-cysteine desulfhydrase; PLP-dependent
4106379			■			G → T	intergenic (+59 /-90)	*cpxP* → / → *fieF*	inhibitor of the cpx response; periplasmic adaptor protein/ferrous iron and zinc transporter
4123585					■	G → A	P291L (CCG → CTG)	*cytR* ←	Anti-activator for CytR-CRP nucleoside utilization regulon
4124389		■				31bp x 2	duplication	*cytR* ←	Anti-activator for CytR-CRP nucleoside utilization regulon
4124521	■		■			IS5 (+) +4 bp	Intergenic (-65 / +88)	*cytR* ← / ← *priA*	Anti-activator for CytR-CRP nucleoside utilization regulon/Primosome factor n' (replication factor Y)
4124521				■		IS5 (-) +4 bp	intergenic (-65 / +88)	*cytR* ← / ← *priA*	Anti-activator for CytR-CRP nucleoside utilization regulon/Primosome factor n' (replication factor Y)
4360793				■		A → C	L381* (TTA → TGA)	*cadC* ←	cadBA operon transcriptional activator
4542161			■	■		IS5 (-) +4 bp	coding (125-128 / 597 nt)	*fimE* →	tyrosine recombinasse/inversion of on/off regulator of fimA
4545086		■				C → A	intergenic (+57 / -10)	*fimC* → / → *fimD*	periplasmic chaperone/fimbrial usher outer membrane porin protein; FimCD chaperone-usher
4639989		■				G → T	N106K (AAC → AAA)	*arcA* ←	response regulator in two-component regulatory system with ArcB or CpxA
4640014					■	T → G	D98A (GAT → GCT)	*arcA* ←	response regulator in two-component regulatory system with ArcB or CpxA
4640190	■		■			C → A	M39I (ATG → ATT)	*arcA* ←	response regulator in two-component regulatory system with ArcB or CpxA
4640557				■		Δ2bp::IS186 (-) +7bp::Δ1bp	intergenic (+15 / -379)	*yjjY* → / → *yjtD*	uncharacterized protein/putative methyltransferase

3.4.1.1 ArcA – The anaerobic switch regulator

ArcA is the regulator of the Arc two component system in conjunction with the sensor kinase ArcB. Termed the anoxic redox control system, this two component system is involved in regulating the change of the cell from aerobic to anaerobic respiration. ArcB kinase activity has been shown to respond to a change in redox states of the ubiquinone and methoquinone pools present in the inner membrane under anaerobic conditions (Bekker *et* *al*., 2010). Upon sensing this change ArcB the undergoes a 3 step phosphorelay, beginning with an autophosphorylation event of His^{294} in ArcB , before ending with phosphorylation of Asp^{54} ArcA (Iuchi and Lin, 1992, Tsuzuki *et al*., 1995). ArcB transfers this phosphate to ArcA, and the phosphorylation of ArcA then goes on to cause a global affect in the cell, repressing genes associated with aerobic metabolism , such as genes involved in the TCA cycle and the glyoxylate shunt, and activating genes associated with anaerobic metabolism (Park *et al*., 2013, Perrenoud and Sauer, 2005, Liu and De Wulf, 2004). However, although the majority of the literature focus upon phosphorylated ArcA which readies the cell for anaerobic metabolism. ArcA has also shown to have a role in aerobic conditions, indeed Perrenoud and Sauer (2005) showed that upon deletion of *arcA* the metabolic flux in the TCA cycle increased by 60%. In addition to this, Park *et al*. (2013) used CHIP-chip and CHIP-seq to identify ArcA binding sites within *E. coli* K-12 MG1655 to understand the ArcA regulon. Although the majority of genes where ArcA binding sites were identified under anaerobic conditions, a set of genes were also identified under aerobic conditions only. Unfortunately, no further focus was conducted on these genes, however these results suggest evidence of a set genes regulated by ArcA under aerobic conditions.

Within all of the evolved strains, mutations associated with *arcA* were observed, a nonsynonymous mutation observed in four strains (E1A - E3A, E5A, Figure 3.4.1A). In addition, in E4A, although there was no mutation in *arcA*, a IS186 insertion upstream of *arcA* was observed downstream of a putative gene termed *yjjY*. Further inspection of the *arcA* promoter reveals that the arcA promoter region extends into and beyond *yjjY*, a putative gene (Figure 3.4.1C). Therefore, this E4A insertion was predicted to have an effect on the expression of the *arcA*.



```
MQTPHILIVEDELVTRNTLKSIFEAEGYDVFEATDGAEMHQILSEYDINLVIMDINLPGKNGLLLA
RELREQANVALMFLTGRDNEVDKILGLEIGADDYITKPFNPRELTIRARNLLSRTMNLGTVSEERR
SVESYKFNGWELDINSRSLIGPDGEQYKLPRSEFRAMLHFCENPGKIQSRAELLKKMTGRELKPHD
RTVDVTIRRIRKHFESTPDTPEIIATIHGEGYRFCGDLED
```

E1A + E3A – M39I (ATG -> ATT)
E2A – D98A (GAT -> GCT)
E5A – N106K (AAC -> AAA)

Figure 3.4.1: Mutations observed in evolved strains associated with arcA. A – Amino acid sequence of arcA mutations observed in Evolved strains highlighted red. B Diagram of arcA region taken from Artemis browser. Top 3 and bottom 3 strands represent the forward and reverse 3 reading frames with black lines representing consecutive stop codons. Middle two strands represent forward and reverse strands (Top and Bottom, respectively). C. Schematic illustrating the promoter region of arcA taken from RegulonDB. E4A IS186 insertion is highlighted. Diagram not to scale.

3.4.1.2 RpoA – RNA polymerase α subunit

In the same 4 out of the 5 strains which possessed non-synonymous mutations in *arcA*, an identical amino acid substitution of asparagine to histidine at position 294 of *rpoA* was found. The gene *rpoA* encodes for the RNA polymerase α subunit of the RNA polymerase complex which is responsible for transcription. The function of RpoA is defined by two domains. The N-terminal domain (α-NTD) provides a scaffold in which the RNA polymerase complex can form. A flexible linker then links the α-NTD to the C terminal domain (α-CTD) which comprises a 4 helices structure with hydrophobic cores involved in DNA site recognition (Jeon *et al.*, 1995). Further understanding of the function of the C terminal domain reveals that in the RNA polymerase complex the C-terminal tail of RpoA is associated with recognition and binding to an UP element, a 9bp long sequence roughly 20bp upstream of the -10 and -35 sites which when RpoA binds to this site can affect transcription, by altering the RNA polymerase – DNA interaction (Browning and Busby, 2004, Estrem *et al.*, 1999, Lloyd *et al.*, 2002). In addition to this, several transcription factors (TF) have been shown to interact with the C terminal domain of RpoA in what is known as Class I activation. Additionally, since the RNA polymerase complex is in a dimer formation, each C-terminal tail is able to independently associate to TF or UP sites (Lloyd *et al.*, 2002, Estrem *et al.*, 1999).

In the evolution experiment described in this study, the location of the *rpoA* (N294H) mutation is within the C-terminal tail. Within the literature, a focus upon the N294 site of the α-CTD have been observed and investigated before in the form of alanine substitutions at this position. These α-CTD substitutions have been investigated in the context of cspD expression and within rrnBP1 and T7 phage D promoter regions (Uppal *et al.*, 2014, Ozoline *et al.*, 2001). In both of these studies, several alanine substitutions within the α-CTD were compared to the expression strength of the promoter and both showed that a N294A substitution did not affect expression, although other alanine substitutions within the α-CTD did.

As mentioned above the α-CTD is comprised of 4 helices labeled 1 – 4, in which helix 1 and the region between helix 3 and 4 have been shown to be involved in DNA recognition of the rrnBP1 promoter (Jeon *et al.*, 1995, Ozoline *et al.*, 2001). Interestingly further inspection of these sites on the rrnBP1 promoter by (Ozoline *et al*., 2001) showed that alanine substitutions (L295A, G296A and K297A) in the regions between helix 3 to 4, all showed repression of the rrnBP1 promoter, indicating that RNA polymerase was unable to sufficiently bind. Although N294A did not have an effect upon expression in this study, considering the differences of a histidine substitution compared to an alanine, one could suggest that this would have a larger effect upon the region.

More recently the same RpoA (N294H) substitution has been observed in a 200 generation evolution experiment to establish suppressor mutants in the "mammary pathogenic" *E. coli* M12 in a zinc uptake deletion background (*ΔznuABC*) evolved in LB in the presence of 2% bile salts (Olson *et al*., 2020). Although observed no further characterization of these mutants was performed. From the results of the literature, one could hypothesize that due to the location and nature of the mutation, differences in expression would occur globally due to associations with UP elements and TF. However, they may also be subtle specific interactions with particular UP elements due to this mutation.

3.4.1.3 cytR – cytidine regulator

In addition, in all 5 strains, mutations in or upstream of *cytR* were observed. CytR is the cytidine regulator and controls a set of genes involved in the acquisition and utilization of ribonucleosides from the environment. Its regulon includes genes whose products are involved in nucleotide transport, nucleotide biosynthesis and nucleoside catabolism, as well as the autoregulation of *cytR* itself. In order for CytR to act as a repressor, it typically forms a dimer which will then form a complex with CRP; this CytR – CRP repression complex prevents CRP activation and thus in turn represses expression. (Søgaard-Andersen *et al*., 1990, Kristensen *et al*., 1997). The inducer of CytR is environmental cytidine. In the presence of cytidine, CytR repression is removed allowing for the activation of the gene by CRP (Barbier *et al*., 1997).

Several site directed mutagenesis studies of *cytR* have allowed the characterization of and identification of functional domains within CytR (Barbier and Short, 1993). In the evolved strains a 32bp duplication (D34DDRLPSINESR*) was found in strain E2A, while an amino acid substitution P291L was found within E5A. The mutation in E2A creates an early termination site within what has been identified as the DNA interaction subunit of CytR. Given the location and nature of this mutation, E2A is likely to have a loss of function of *cytR*. Loss of function of *cytR* could also be hypothesized in E5A, as the location of the mutation is within the oligomerization domain of CytR, and previous mutations seen within this region of *cytR* result in inactive repressors, as the mutated forms of cytR are unable to form dimer complexes and are therefore unable to function. Given the nature of the mutation seen in strain E5A it is likely that this mutation would disturb this site and thus cause loss of function.

In addition to seeing mutations within *cytR* in E2A and E5A, in the remainder of the evolved strains an IS5 insertion upstream of *cytR* was identified. In three out of the 5 strains, an IS5 insertion event occurred in the exact same location at position -65 from the *cytR* start site. The location in this start site is in the CRP recognition site of the *cytR* promoter (Figure 3.4.2A). An insertion at this site would therefore be expected to disrupt CRP binding so no repressor or activator effect would occur at the *cytR* promoter as both are dependent on CRP. One could assume therefore that the effect of this mutation would be the dysregulation of cytR probably causing little expression of *cytR* and therefore a subsequent activation of the normally repressed *cytR* regulon. Interestingly, although the location of insertion did not differ, the orientation of the IS5 element did, with *breseq* reporting E1A and E3A to have a positive orientation while E4A a negative orientation. In order to confirm that this orientation specific differences were actually seen within the evolved strains a PCR based approach was conducted using flanking and IS5 internal primers (Figure 3.4.2B).

Recently a laboratory based evolution experiment looking at long term stationary phase in LB found that a series of mutations within *cytR* occurred within this experiment (Kram *et al.*, 2017). Here, Kram *et al*. (2017) showed that mutations in *cytR* had a greater fitness in competition with the wildtype in long term stationary phase conditions. Results from this study suggested that activation of the *cytR* regulon could lead to higher fitness as it allows the cell to utilize free nucleotides which would have been present when the cells would normally be undergoing death phase. Although in the conditions of our evolution experiment the daily passaging would be too short for death phase to occur, the utilization of nucleosides could still provide a benefit as although not as present in the environment as they would be during death phase, there still would be a cellular pool which could be tapped.

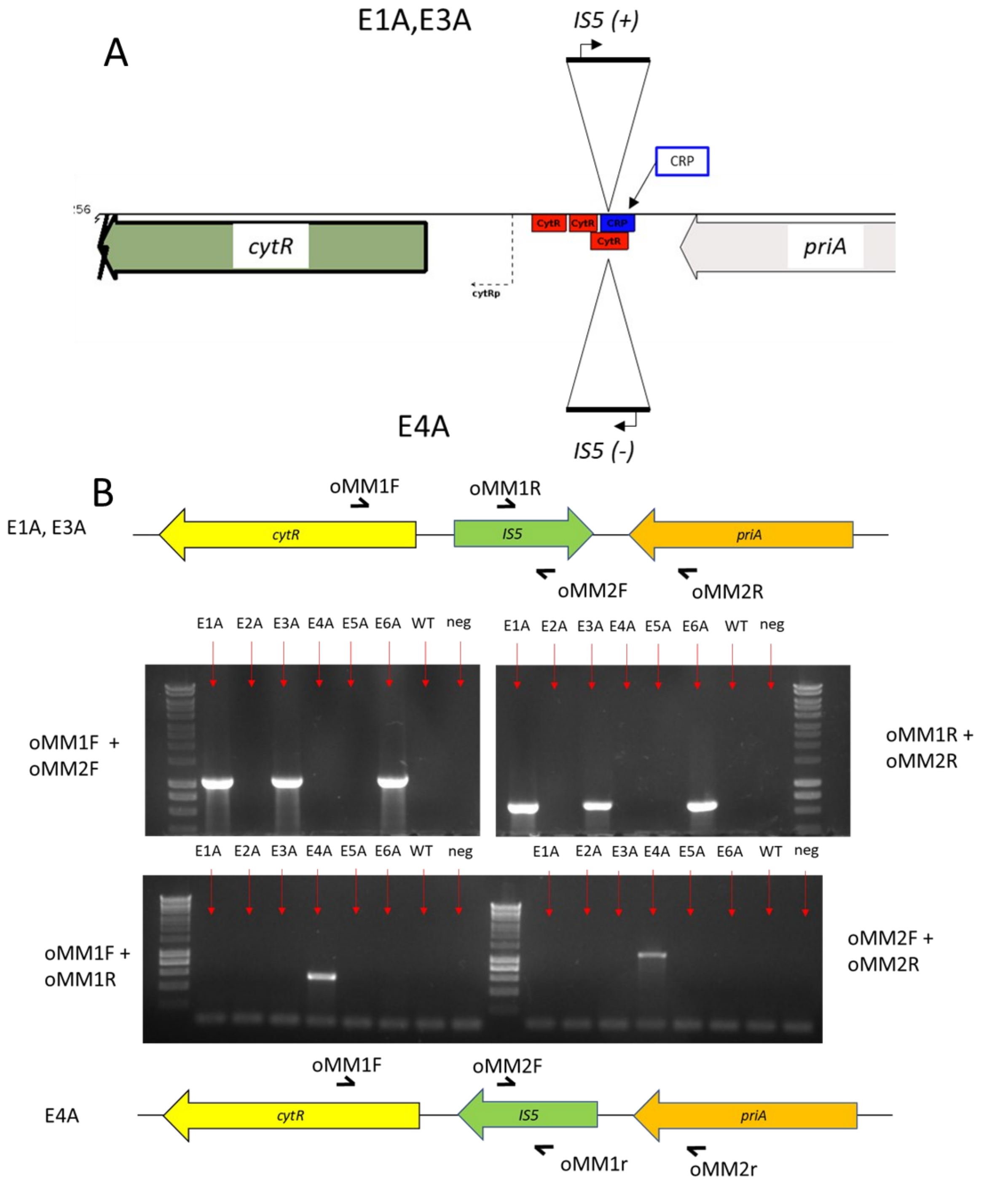


Figure 3.4.2: Evolved strain IS5 mutations upstream of cytR. A) Diagram taken from RegulonDB, illustrating the regulatory region upstream of cytR and the relative location of IS5 insertions. IS5 insertion different in orientation but not location between evolved strains are highlighted; diagram is not to scale. Transcription factor binding sites for cytR (Red) and CRP(Blue) that regulate the expression of cytR. IS5 insertions insert within the CRP and cytR binding cite as indicated. B) Confirmation of IS5 insertions orientation upstream of cytR. Ladder corresponds to 1kb hyper ladder (bioline). Schematic diagrams are shown indicating whether a band would be expected based on the primer pair in the evolved strain. E6A did indeed have a IS5 insertion at the positive orientation.

3.4.1.4 Other mutations seen within the evolved strains

As mentioned above within the evolved strains, mutations in *arcA*, *rpoA* and *cytR* showed a degree of parallel evolution. In addition to this, this section discusses the other mutations which occurred in the evolved strains. One such example is in the *tna* operon where mutations occurred in strains E1A, E3A and E5A. The *tna* operon includes *tnaA*, which encodes tryptophanase and is responsible for converting tryptophan present within the environment to indole, and *tnaB*, encoding a tryptophan antiporter. During late exponential phase early and stationary phase, indole environmental and intracellular concentration dramatically increases before declining again, which is termed the “indole spike”. Indole has been shown to be involved in the signalling of a variety of different cellular responses and has been proposed to be a signal for the cell to enter stationary phase (Gaimster *et al.*, 2014, Gaimster and Summers, 2015). Deletion of the tryptophanase *tnaA* has shown that the cell can enter stationary phase later than its wildtype equivalent and thus loss of TnaA could potentially confer a fitness benefit to these strains under some selective conditions.

Further to this, in E2A, E3A, and E4A, mutations within the *fim* operon are observed. The *fim* operon encodes the type 1 fimbriae, with *fimA*, *C + D* encoding the main subunits of the type1 fimbriae. The regulation of *fim* is by phase variation with *fimE* and *fimB* encoding recombinases, which regulate the *fim* operon by altering the orientation of *fimS,* a region of DNA with flanked inverted repeats that contains the promoter for the operon. While FimB can switch the expression of the type 1 fimbrae from OFF to ON and ON to OFF, FimE is only able to turn *fimS* from ON to OFF. Therefore, deletions or IS insertion observed within fimE may thought to alter this phase variation, which may promote the expression of fimE.

In two strains (E1A and E4A) large deletions were observed. In E4A this resulted in the loss of the flagella motor and regulator genes, *motAB* and *flhCD*. In addition, the large deletion observed in E1A results in loss of the e14 prophage, present within MG1655.

3.4.2 Evolved populations E1P-E5P

Further to sequencing of the evolved clonal strains, the evolved populations from which these strains were isolated were also sequenced to understand what mutations were observed in the evolved populations, and at what frequency. Evolved populations were sequenced by MicrobesNG (Birmingham, UK) to a minimum of 80X sequencing depth, overall, however the average sequencing depth of all evolved population was 92X. Variant calling was performed using *breseq* using a 5% cutoff for any reported variant call within the population. Overall, a total of 69 different mutation events were identified, within 40 genes or intergenic regions with the results presented in Table 3.2.

Table 3.2: Mutation frequency of Evolved populations E1P -E5P. Duplicate reads were removed using FastUniq (Xu *et al*., 2012). Grey box represent mutation present, with the number indicating the proportions of reads corresponding to variant called, this in turn can relate to the frequency of the genotype in the population. Genes present in wildtype with a total variant call of 1 were removed from this analysis.

Position	E1P	E2P	E3P	E4P	E5P	Mutation	Annotation	Gene	Description
70371		0.055				T → C	intergenic (-323 / -16)	*araB* ← / → *araC*	L-ribulokinase/ara regulon transcriptional activator; autorepressor
70581		0.054				C → T	V65V (GTC → GTT)	*araC* →	ara regulon transcriptional activator; autorepressor
71109		0.067				T → C	T241T (ACT → ACC)	*araC* →	
71175		0.096				A → G	R263R (CGA → CGG)	*araC* →	
71214		0.071				T → C	F276F (TTT → TTC)	*araC* →	
109384	0.098					A → G	N369S (AAC → AGC)	*secA* →	preprotein translocase subunit; ATPase
200228					0.071	T → C	R767R (CGT → CGC)	*bamA* →	BamABCDE complex OM biogenesis outer membrane pore-forming assembly factor
414138			0.058			A → G	V539A (GTT → GCT)	*sbcC* ←	exonuclease; dsDNA; ATP-dependent
579159			0.115			C → T	Intergenic (-266 / +25)	*borD* ← / ← *ybcV*	DLP12 prophage; putative lipoprotein/DLP12 prophage; uncharacterized protein
723528				0.500		(T)8 → 9	coding (887 / 2685 nt)	*kdpD* ←	fused sensory histidine kinase in two-component regulatory system with KdpE: signal sensing protein
932562			0.714			IS1 (+) +9 bp	intergenic (-512 / -25)	*trxB* ← / → *lrp*	thioredoxin reductase; FAD/NAD(P)-binding/leucine-responsive global transcriptional regulator
962854			0.063			T → G	L287R (CTG → CGG)	*rpsA* →	30S ribosomal subunit protein S1
966454					0.096	T → A	I712N (ATT → AAT)	*ycaI* →	ComEC family inner membrane protein
968206		0.056				A → C	N529T (AAC → ACC)	*msbA* →	lipid ABC transporter permease/ATPase
1028046			0.071			A → G	K34R (AAA → AGA)	*yccU* →	putative CoA-binding protein
1152651			0.067			T → C	M238T (ATG → ACG)	*fabF* →	3-oxoacyl-[acyl-carrier-protein] synthase II

Position	E1P	E2P	E3P	E4P	E5P	Mutation	Annotation	Gene	Description
1296620				0.200		Δ12 bp	coding (1502-1514/2676 nt)	*adhE* ←	fused acetaldehyde-CoA dehydrogenase / iron-dependent alcohol dehydrogenase
1969201	0.210					A → G	V55A (GTT → GCT)	*cheR* ←	chemotaxis regulator; protein-glutamate methyltransferase
2193791				0.222		T → C	V126V (GTA → GTG)	*mrp* ←	antiporter inner membrane protein
2483872			0.054			C → T	R40W (CGG → TGG)	*evgA* →	response regulator in two-component regulatory system with EvgS
2765456			0.123			(TATGGCAC)6 → 5	intergenic (-303 / +50)	*yfjL* ← / ← *yfjM*	putative defective prophage/polymorphic toxin family protein/ prophage; uncharacterized protein
3035546		1.000				T → C	T246A (ACG → GCG)	*prfB* ←	peptide chain release factor RF-2
3209081			0.176		0.184	G → A	G353E (GGA → GAA)	*ttdT* →	L-tartrate/succinate antiporte
3214770			0.277		0.296	A → C	E575A (GAA → GCA)	*rpoD* →	RNA polymerase; sigma 70 (sigma D) factor
3237853		0.071				T → C	G181G (GGT → GGC)	*ygjR* →	putative NAD(P)-dependent dehydrogenase
3277167	0.064					C → T	R56* (CGA → TGA)	*prlF* →	antitoxin of the SohA(PrlF)-YhaV toxin-antitoxin system
3351726			0.215			C → A	E434* (GAA → TAA)	*arcB* ←	aerobic respiration control sensor histidine protein kinase; cognate to two-component response regulators ArcA and RssB
3353003			0.154			G → A	A8V (GCG → GTG)	*arcB* ←	
3377183				1.000		C → T	D80N (GAT → AAT)	*sspA* ←	stringent starvation protein A; phage P1 late gene activator; RNAP-associated acid-resistance protein; inactive glutathione S-transferase homolog
3440150	1.000	1.000	0.750		0.867	T → G	N294H (AAC → CAC)	*rpoA* ←	RNA polymerase; alpha subunit
3544482	0.813					IS186 (–) +5 bp :: Δ1 bp	coding (359-363 / 771 nt)	*bioH* ←	pimeloyl-ACP methyl ester carboxylesterase
3888473	1.000					C → A	F13L (TTC → TTA)	*tnaC* →	tryptophanase leader peptide
3888626			0.078			IS5 (–) +4 bp	intergenic (+117 / -101)	*tnaC* → / → *tnaA*	tryptophanase leader peptide/tryptophanase/L-cysteine desulfhydrase; PLP-dependent
3888873					0.700	IS1 (+) +9 bp	coding (144-152 / 1416 nt)	*tnaA* →	tryptophanase/L-cysteine desulfhydrase; PLP-dependent
3890555			0.054			IS5 (+) +4 bp	coding (320-323/1248 nt)	*tnaB* →	tryptophan transporter of low affinity
4093990			0.055			C → T	V95I (GTC → ATC)	*rhaD* ←	rhamnulose-1-phosphate aldolase
4101089			0.062			A → G	T94A (ACC → GCC)	*sodA* →	superoxide dismutase; Mn
4123585					0.467	G → A	P291L (CCG → CTG)	*cytR* ←	Anti-activator for CytR-CRP nucleoside utilization regulon
4124179			0.165			G → T	A93E (GCG → GAG)	*cytR* ←	
4124389		0.667				31bp x 2	duplication	*cytR* ←	
4124521	0.782		0.629			IS5 (+) +4 bp	Intergenic (-65 / +88)	*cytR* ← / ← *priA*	Anti-activator for CytR-CRP nucleoside utilization regulon/Primosome factor n' (replication factor Y)
4124521				0.938		IS5 (–) +4 bp	Intergenic (-65 / +88)	*cytR* ← / ← *priA*	
4296060		0.141	0.209	0.308		C → T	intergenic (+266 / +376)	*gltP* → / ← *yjcO*	glutamate/aspartate: proton symporter/ Sel1 family TPR-like repeat protein
4296268		0.323	0.308			T → C	intergenic (+474 / +168)	*gltP* → / ← *yjcO*	

Position	E1P	E2P	E3P	E4P	E5P	Mutation	Annotation	Gene	Description
4296286		0.187	0.179			C → T	intergenic (+492 / +150)	*gltP* → / ← *yjcO*	
4296303		0.121				A → G	intergenic (+509 / +133)	*gltP* → / ← *yjcO*	
4296304		0.121				A → C	intergenic (+510 / +132)	*gltP* → / ← *yjcO*	
4360758			0.075			IS5 (–) +4 bp	coding (1174-1177 /1539 nt)	*cadC* ←	cadBA operon transcriptional activator
4360793				0.600		A → C	L381* (TTA → TGA)	*cadC* ←	
4408494					0.096	A → G	Q614R (CAG → CGG)	*rnr* →	exoribonuclease R; RNase R
4542161				1.000		IS5 (–) +4 bp	coding (125-128/597 nt)	*fimE* →	tyrosine recombinase/inversion of on/off regulator of fimA
4542689					0.214	C → A	intergenic (+56/-426)	*fimE* → / → *fimA*	tyrosine recombinase/inversion of on/off regulator of fimA/major type 1 subunit fimbrin (pilin)
4639989	0.179	1.000				G → T	N106K (AAC → AAA)	*arcA* ←	response regulator in two-component regulatory system with ArcB or CpxA
4640014					0.706	T → G	D98A (GAT → GCT)	*arcA* ←	
4640108			0.719			G → T	R67S (CGT → AGT)	*arcA* ←	
4640190	0.656		0.106			C → A	M39I (ATG → ATT)	*arcA* ←	
4640557				1.000		Δ2 bp :: IS186 (-) +7 bp :: Δ1 bp	Intergenic (+15/-379)	*yjjY* → / → *yjtD*	uncharacterized protein/putative methyltransferase

3.4.2.1 Diversity within populations

Inspection of the mutations in the evolved populations revealed that some samples had more variation within population than others, with E3P showing the most diversity of 25 mutations within 21 different genes/intergenic regions while E1P only showing 9 mutations within 8 gene/intergenic regions (Table 3.2 + 3.3). These differences in diversity indicate that independent populations could have different population dynamics, with those with more mutational diversity potentially having the opportunity to have a more diverse population substructure.

Therefore, in order to investigate population substructure further, the relative frequency of each mutation was considered. Relative frequency was determined as the proportion of sequencing reads which align to a region specified that are able to identify a variant within the region. This in turn can then be used as proxy to the proportion of the population which have that variant. Relative frequency can range from 0.05 (5%), where a 'mutation' is below the threshold value required to predict a variant, to 1 (100%) where a mutant is considered fixed, referring to it being present in all cells within the population. It was found that in 3 independent populations, 8 mutations observed went to fixation (E1P, E2P and E4P, Table 3.3). Further to this, within these 3 independent populations, fixation events were associated with the same gene in 2 populations, namely *arcA* (E2P + E4P) and *rpoA* (N294H) (E1P + E2P). Each individual population then also had fixation events in individual genes, as follows: E1P in *tnaC* (E1A-I13L), E2P in *prfB* (T246A) and E4P in *sspA* (D80N). In addition to fixations, determination of relative frequency revealed that the same subset of genes typically had higher frequency of mutations than others. For instance, RpoA $_{N294H}$, which in two populations was fixed, had reached a relative frequency of > 75% within the other populations where the mutation was identified (E3P + E5P). Also, in *arcA,* where again two fixations occurred in independent populations, in the other independent populations a least 1 variant in *arcA* had a frequency greater than 65%. In addition, within two population, 2 different variants in *arcA* were identified (E1P + E3P). The same could also be seen with mutations in or upstream of cytR, with the exception of E5P which had a mutation greater than 65% of the population. Multiple different mutations associated with *cytR* were found within E3P. Considering that within each population mutations in the same 3 genes were at larger frequencies than others, this provides further evidence of parallel evolution occurring within these populations. In addition, the relative frequencies of these mutations and the frequency of occurrence of these mutations could be a consequence of the extent of fitness advantage that these mutations can provided to the cell in these conditions.

Table 3.3: Summary of genes identified from evolved populations

	E1P	**E2P**	**E3P**	**E4P**	**E5P**
Total number of mutations present	**9**	**16**	**25**	**8**	**10**
Fixation (100%)	2	3	0	3	0
Very high frequency (75% ≤ X < 100%)	2	0	1	1	1
High frequency (50% ≤ X < 75%)	1	1	3	2	2
Low frequency (50% > X ≥ 25%)	0	1	2	1	3
Very low frequency (25% ≥ X≥ 5%)	4	11	19	1	5
Number of genes or intergenic regions mutations are present	**8**	**9**	**21**	**8**	**10**
Number of genes with mutations present	7	7	15	5	9
Number of Intergenic regions with mutations present	1	2	6	3	1

The description above focuses on mutations which have a higher relative frequency within the population. In each population however, there are also a number of mutations which are low in frequency. In particular two evolved populations E2P and E3P have a larger proportion of these with the majority of mutations identified being present at less than 25% of the population. However, the exclusion of these mutations indicates that within E2P, like other populations, very few mutations have occurred within other genes during the space of the evolution experiment (Table 3.4). While in E3P, the mutations identified typically correspond to different genes and intergenic regions, interestingly, in E2P, a total 15 mutations were identified within 9 different genes/intergenic regions (Table 3.3). Further investigation of the E2P population reveals that in two cases, multiple low frequency mutations within a particular gene/ intergenic region were observed. In E2P, 4 synonymous mutations were observed within *araC* all at low frequency (< 10%), in addition, 5 independent base pair substitutions occurred in the intergenic region between *gltP* and *yjcO* with a relative frequency ranging from 12.1% - 32.3%. Mutations in the *gltP – yjcO* intergenic region were also observed in E3P (3 different mutations) and E4P (1 different mutation), at low frequency. Although unusual, further inspection of the reads associated with these variants showed that where identification was possible the mutations were located within the same reads which had different alignment positions, indicating that these variants are inherited together. This characteristic of the E2P population is unusual and the reason why this has occurred is unknown, however the exclusion of these mutations indicates that within E2P, like other populations, very few mutations have occurred within other genes during the space of the evolution experiment.

Table 3.4: The type of mutations observed in Evolved populations

Type of mutations observed:	**E1P**	**E2P**	**E3P**	**E4P**	**E5P**
Bp substitutions	6	15	19	4	9
Nonsynonymous	6	4	15	2	7
Synonymous	0	5	0	1	1
Intergenic	0	6	4	1	1
IS insertions	2	0	5	3	1
Indel	0	1	1	1	0

Interesting, as highlighted in table 3.4 there was small number of synonymous mutations identified within the evolved population but although this number was low, it was not surprising. Although there are examples which have demonstrated synonymous mutations to alter phenotypic fitness (Lind *et al*., 2010, Lebeuf-Taylor *et al*., 2019). Typically, synonymous mutations are considered to be neutral, and therefore selection will not upon them, instead neutral mutations can increase in frequency due to other processes mainly genetic drift. Given that genetic drift is a slow process, the chances that numerous neutral mutations will occur to a frequency that is detectable (>5%) within the timeframe of our evolution experiments is slim. An opinion, presented by Cooper (2018) supports this conclusion, arguing that in the average timeframe of an evolution experiment (Less than a year), the majority of mutations which are at a frequency which can be detected selection will have acted upon them and that it is unlikely that these mutations have accumulated, within this time frame by the effects of genetic drift alone. Although other processes such as genetic hitchhiking, may increase the number of neutral mutations observed, it would still be unexpected to see a lot of neutral mutations arising within the timeframe of our laboratory evolution experiment.

Overall, by looking at the relative frequencies of mutations within each independent population it is clear each is characterized by a series of mutations which are found at high frequency. Using these relative frequencies one can start to assume what variants might coincide together within a single cell to represent a large proportion of the population. It is evident that within each population a dominant genotype would mainly consist of mutations associated with *arcA*, *cytR* and *rpoA* (E4P as an exception), in addition to other variants which are specific to each population. Evidence of this can be clearly seen in E1P, E2P and E4P where with the amount of fixation events and large relative frequency variants a representative genotype can be predicted. However, within E3P and E5P, since a lot of variants are at low frequency and as no fixation events have occurred, considering which mutations might occur together is not possible. However, since the evolved strains were isolated from each population these can be used as examples of which genes are inherited together.

3.4.2.2 Comparing Evolved populations to Evolved Strains

Since a set of fixations and high frequency mutations were identified in the population, the question was asked how well do the mutations in the evolved strains represent the mutations in the population (Table 3.1 + 3.2)? Initially, each evolved strain was compared back to its population sequence to ask if they correlated well. The result of this showed that in strains, E1A, E4A and E5A, all mutations, with the exception of large deletions, identified in the clonal samples were also identified in the evolved populations. In addition, for all the mutations identified in E1A, E4A and E5A within the respective populations the relative frequency of these mutation was > 60%. The only instance where this was not the case was the cytR (P291L) recorded in E5P which showed a relative frequency of 46.7 %. In E2A a similar result was seen, with 4 out of the 5 mutations being identified and with a relative frequency > 60%. However, in E2A a base substitution seen in the intergenic region between *fimC* and *fimD* was not observed in the population E2P. Further to this, when comparing E3A and its population, E3P, out of the 7 mutations identified in E3A only 4 were observed within the population. The 3 mutations not identified included two IS insertions *bioH::IS186* and *fimE::IS5* and a point mutation in the intergenic region of *cpxP* and *feiF*.

The differences between the evolved strains and the populations from which they were isolated could simply be due to random chance in which a mutation has arisen at low frequency and due to chance, this has been present in the clone selected for sequencing. This could be the case with E2A, where the majority of mutations identified are at high frequency within the population. However, in E3A the fact that three such mutations identified makes this explanation less likely. Furthermore, in E3A the mutations which did match were in *rpoA* and *cytR*, two mutations already at high frequency in the population. The further two matching mutations were within *tnaA* and *arcA* at a 7.8% and 10.6% relative frequency, respectively. It is the mutation in *arcA* which potentially highlights a situation that the E3A strain does not represent the dominant genotype within E3P. Within E3P two nonsynonymous mutations within *arcA* were identified, R67S and a M39I substitution, with a relative frequency 71.9% and 10.6% respectively. Considering the mutations seen in E3A, the *arcA* mutation which was observed is the one with the smaller frequency (M39I), which indicates that E3A represents a smaller subpopulation within E3P. In addition to this, within E3P, another mutation reported at high frequency (71.4%) that is not reported in E3A is

IS1::(-25)*lrp*. The fact that this mutation is not seen in the E3A genotype, supports the idea that E3A represents minority subpopulation within E3P. A possibility arises with regards to the other mutations observed within E3A which do not correlate with E3P, which is that they may be missed in E3P due to being at a lower frequency than is detected by *breseq* (5% of reads), as well as two of these mutations being IS insertions which are harder to identify at lower frequencies since junction evidence from both sides of the insertion is required. Indeed, a manual visualization of the alignment using the Integrative Genomics Viewer (IGV), confirmed that reads associated with these mutations could still be identified (DATA NOT SHOWN).

Overall, mutations identified within the evolved strains correlated well with the populations, and with the exception of E3A, each clonal strain seem to be a representative of the dominant mutations present within the populations. Although E3A did not cover all dominant mutations it still did have mutations within the major three genes associated with all populations, *arcA*, *rpoA* and *cytR*. The isolation of these evolved strains each gave a representation of which mutations in the populations occur together. Since no other strains were isolated, however, for mutations which did not occur in the strains their co-inheritance (i.e. whether these mutations occur together) cannot only be inferred and not determined.

3.4.3 The E1A-I1 Intermediate.

In addition to the evolved strains and populations sequenced, Sen also isolated and sequenced a single strain from the E1P population, from stock frozen on day 10 of conducting the evolution experiment. Reanalysis using *breseq* was also performed on this sample which concluded that this clone contained only an *arcA* (M39I) mutation, identical to that seen within E1A, as well as the mutations attributed to the MG1655 ancestor. Although a detailed analysis of the occurrence of mutations was not conducted, , the isolation of this strain demonstrated that mutations in *arcA* arose early within the E1P population.

3.5 Growth curves of Evolved strains

Section 3.3 showed that during 24 hour of growth of *E. coli* MG1655, the environmental pH increased when grown in unbuffered LB at pH 4.5, due to the catabolism of amino acids. Since the evolved strains were adapted to this dynamic pH environment, and several mutations identified had occurred in TF which were associated with metabolism, it was therefore of interest to consider whether any differences between growth could be observed between the evolved strains and MG1655 ancestor. Therefore, growth experiments were done for each evolved strain, with OD_{600} and environmental pH recorded every hour for 24 hours (Figure 3.5.1). Due to the amount of volume required to take measurements and pH hourly, 50ml of LB was used in a 250ml Erlenmeyer flask for each replicate, and therefore the volume does not directly relate to the evolution experiment.

Overall, no large differences were observed between the evolved strains and ancestor. However, the evolved strains, E1A, E2A, E3A and E4A looked to enter exponential phase slightly earlier than MG1655 and E5A. A slight decline in OD_{600} was also seen in the evolved strains when the cultures were in stationary phase. Comparison of the environmental pH revealed that the overall “dynamic of change” in pH, was highly similar to MG1655, with the largest increase in pH occurring between 3 and 8 hours of growth, and then started to slowly increase over the remaining time to reach a pH 8.7 ± 0.1. Only E2A showed a difference, with a pH 0.2 – 0.3 increase at 6 – 9 hours compared to MG1655 before matching the pH at 10 hrs.

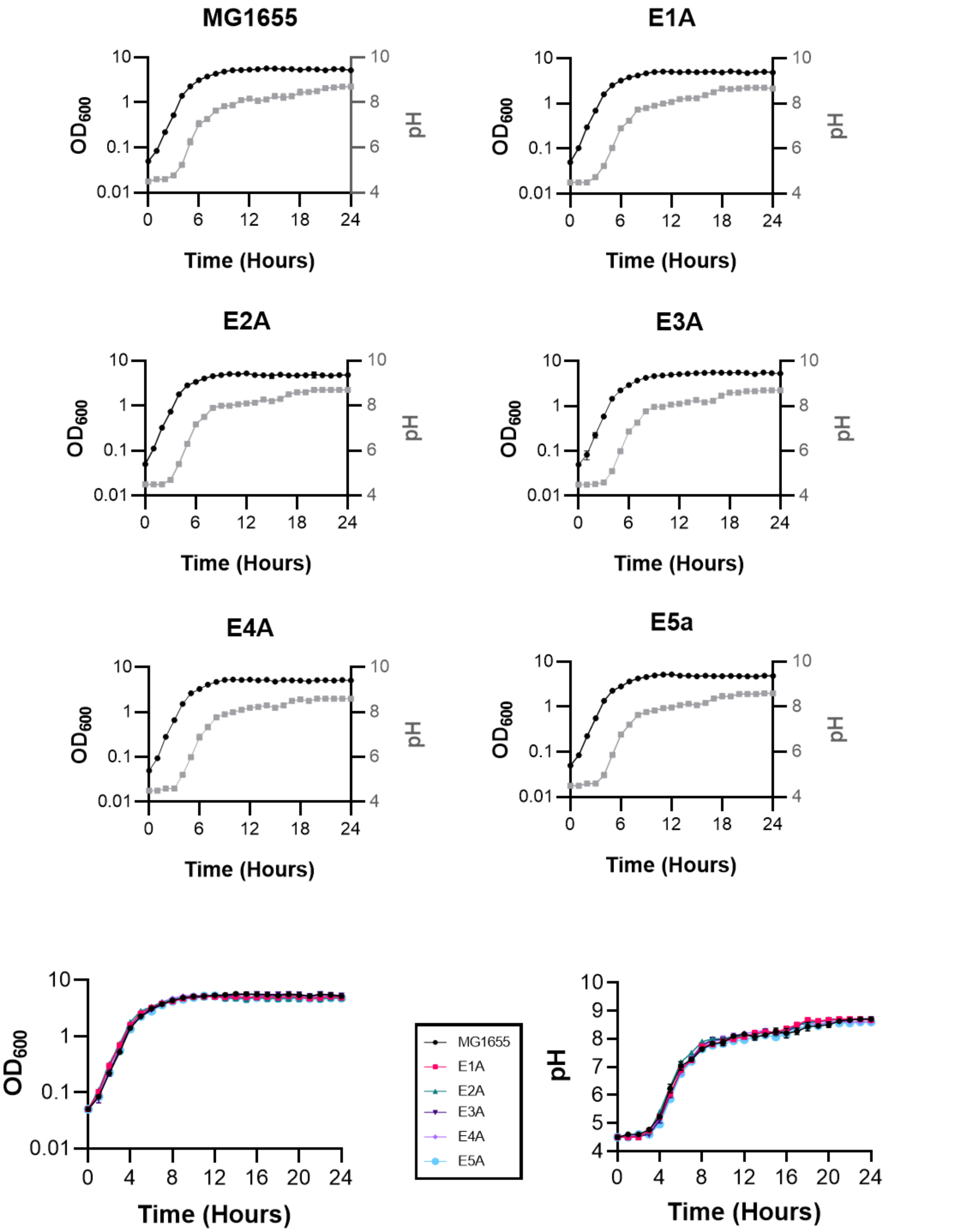


Figure 3.5.1: Growth curves of evolved strains and MG1655 ancestor over 24 hours in 50ml unbuffered pH LB starting at pH 4.5. OD$_{600}$ (Black) and pH (Grey) of culture was taken at each hour timepoint. A comparison of pH and evolved strains can be found in the bottom two graphs. The data presented is the mean +/- SD of 3 independent replicates.

3.6 Understanding fitness within the evolved strains

Growth curves of the evolved strains showed that at pH 4.5 there were no large differences in growth rate between strains (Figure 3.5.1). Therefore, in this study, competition experiments were conducted to assess the evolved strains' relative fitness under a variety of different conditions. To determine evolved strain fitness, the ideal comparison would be competitions with the MG1655 ancestor that each evolved strain was derived from. However, it is not easy to quickly distinguish between the evolved strain and MG1655. To resolve this, in each competition a MG1655 lac^{-} derivate, KH001, was used as a proxy for MG1655, allowing each strain to be distinguished by plating on MacConkey lactose. To confirm that KH001 could be used as a proxy for MG1655 in each condition, MG1655 was competed with KH001 under all conditions.

Once CFU/ml for each strain at each timepoint has been obtained, a comparison can then be made to determine whether one strain is fitter than the other. This comparison usually results in a single metric being calculated which describes fitness of strain A relative to strain B. However, within the literature there are several different 'relative fitness' metrics available with each calculating a metric which reports fitness differently. In addition to this, different metrics make different assumptions about the data, which can lead to different interpretations of results from the same data generated from a competition experiment. Therefore, in order to choose the right fitness metric for this study, a comparison of five different relative fitness metrics was done, to ensure that the most appropriate fitness measurement was used in this study. Appendix 2 highlights the details of this comparison, ultimately the metric selection rate (s) was chosen as it was able to demonstrate linear correlation between fitness reported and relative change in proportions between strains, irrespective of whether a population was growing or in decline.

In this control, if a selection rate of 0 was obtained, this would mean that there were no differences in fitness between MG1655 and KH001 indicating that KH001 can be used a proxy to MG1655. The results present in this section below, with the exception of the condition unbuffered LB at pH 4.5 with 1mm Acetic acid, showed no difference in fitness was observed between MG6155 and KH001 and therefore in these conditions the KH001 could be used as a proxy to the MG1655 ancestor (Figure 3.6.1 – 3.6.10). The difference observed in unbuffered LB at pH 4.5 with 1mm Acetic acid, is discussed in section 3.6.4.4

3.6.1 Assessment of adaptation in the evolved strains.

In laboratory based evolution experiment, the first question which arises is whether selection and adaptation to the conditions of the experiment has occurred. To address this, competition experiments were conducted of the evolved strains against KH001 under the same conditions as the evolution experiment (5 ml unbuffered LB pH 4.5 at 37^{o}C, Figure 3.6.1). Throughout these experiments, the control competition of MG1655 vs KH001 indicated no differences in relative fitness, confirming that KH001 can be used a proxy for the MG1655 ancestor. After 1 day of competition, as was expected, evolved strains E1A – E4A showed increased fitness compared to KH001.Unexpectedly however, for E5A, no difference in fitness was observed. This suggests E5A had not adapted to the conditions of the evolution experiment.

The previous result indicated that under 24 hours of the conditions of the evolution experiment, E5A did not show any fitness advantage. Although unusual, two possibilities arose which could provide an explanation, the first was considering the evolution experiment consisted of subsequent passaging of repeated batch culture there was the possibility that the evolved strains had adapted to any stage of this protocol, such adaptations providing fitness towards lag and stationary phase, as well as exponential phase. The second was that the difference in fitness seen may have been too small during one day of growth to observe within the resolution of the fitness metric. Therefore, to address these possibilities, the competition experiment was continued further, for a total 5 days with passaging using the same dilution series as was used in the evolution experiment, i.e. a 1 in 20 dilution.

At days 3 + 5 of the competition experiments, cells were plated and counted, and the selection rate was calculated for each evolved strain by comparing counts to the initial population. On day 3 all evolved strains, including E5A, showed increased fitness relative to KH001, thereby indicating that E5A adaptions may be to other conditions present in the evolution experiment and not just growth in unbuffered LB (Figure 3.6.1). Alternatively, they may be relatively small and so require amplification over several days of repeated growth and dilution to be seen above background variation.

Differences in selection rates generated at Day 1 ,3 and 5 indicated that different strains did better or worse during the competition at different times. For strain E2A, E3A, and E4A the selection rate per unit of time, highlighting that within E1A and E5A, however, no difference in selection rate was observed between Day 3 and Day 5 although they saw large increases at Day1. This pattern of selection rate suggested that these strains were fitter at different times of the evolution experiment.

The main conclusion of these experiments is that calculations of selection rate over 5 days of competition do show that the evolved strains are fitter than their parent under the conditions of the evolution experiment. This indicates that selection and adaptation had occurred during the evolution experiment.

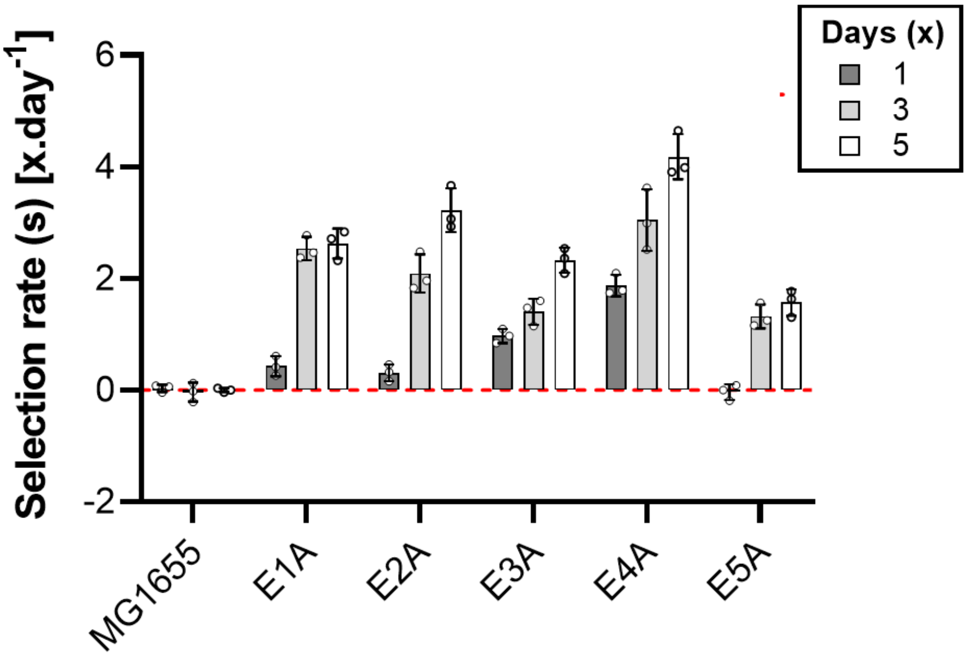


Figure 3.6.1: Selection rates from competition experiments done over 5 days in 5ml unbuffered LB at pH 4.5. Every 24 h, strains were passaged into fresh media at a 1 in 20 dilution. Timepoints were taken at Day 1, 3 and 5. Selection rates presented in the figure are calculated in units per time according to their final time point (Day1 [s 1Day^{-1}], Day 3 [s 3Days^{-1}] and Day5 [s 5Days^{-1}]. Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

3.6.2 Evolved strains show increased fitness at pH 7

Since growth of *E. coli* in unbuffered LB at pH 4.5 changes the environmental pH, the evolved strains would be subjected daily to different pH at different stages of growth. Therefore, to begin to investigate the effect of pH on the fitness of the evolved strains, competition experiments were then done in unbuffered LB at pH 7. By using unbuffered LB at pH 7, the evolved strains would not undergo lag and exponential phase at low pH. Instead they would undergo these at neutral pH, whilst spending stationary phase under alkaline conditions (Figure 3.3.1). Thereby any increase in fitness seen within these experiments would indicate that the evolved strains have not adapted solely to the conditions of low pH, but instead indicate they had also adapted to alternative stresses present within the conditions of the evolution experiment.

As shown in Figure 3.6.2, after 1 day of growth using unbuffered pH at pH7, no significant differences were seen in the fitness of the evolved strains competed against KH001. To further examine this, selection rates were compared for competitions using unbuffered LB at either pH 7 or pH 4.5 (Figure 3.6.3). Analysis of data from day 1 revealed that there was a significant difference in relative fitness of E3A and E4A at pH 4.5 compared to pH7, indicating that these strains had adapted and show increased fitness during the initial stages of growth in LB at low pH (Figure 3.6.3A).

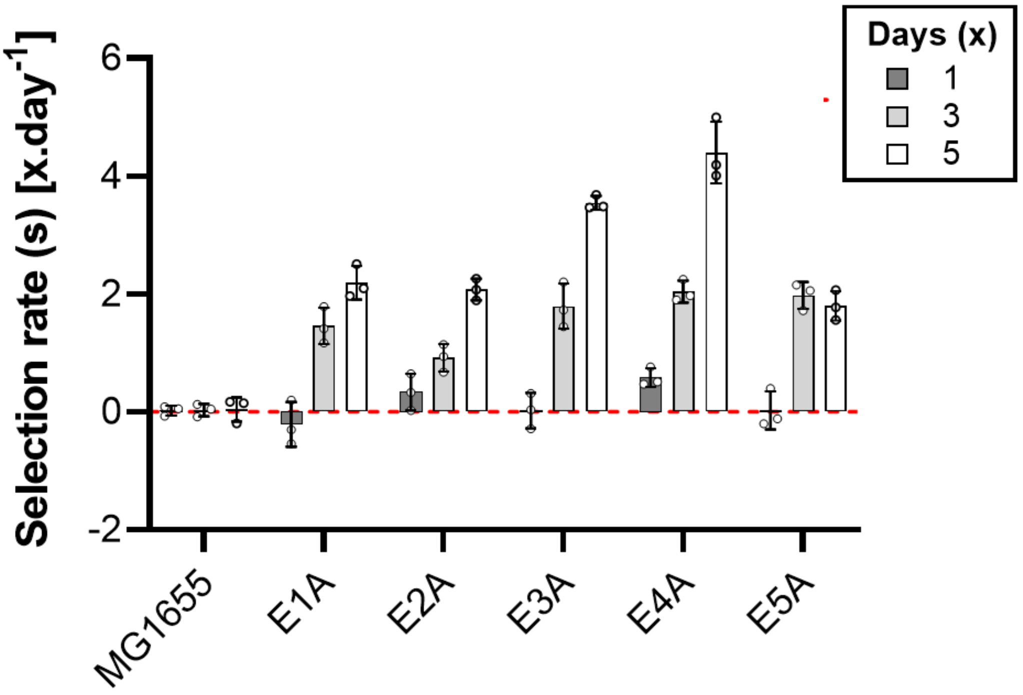


Figure 3.6.2: Selection rates from competition experiments done over 5 days in 5ml unbuffered LB at pH 7. After each day, strains were passaged into fresh media at a 1 in 20 dilution. Timepoints were taken at Day 1, 3 and 5. Selection rates presented in the figure are calculated in units per time according to their final time point (Day1 [s = 1Day^{-1}], Day 3 [s = 3Days^{1}] and Day5 [s = 5Days^{-1}]. Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

As with the competitions in unbuffered LB at pH 4.5, the competition experiments at pH 7 were continued for 5 days to mimic the conditions of the evolution experiment and to include selective conditions provided by subsequent passaging. As with pH4.5, selection rates, in unbuffered LB at pH 7 increased, indicating that evolved strains continued to be fitter over subsequent days (Figure 3.6.2). This increase in selection rates suggests that the evolved strains have not solely adapted to conditions of initial low pH and instead have adapted to other conditions within the evolution experiment which arise at pH7 as well as pH 4.5.

Comparison the selection rate for each condition at pH 4.5 and pH 7 for each day, indicates that the evolved strains have adapted differently with the evolution experiment to stresses in addition to low pH. This is demonstrated particularly in strains E3A and E5A, which at Day 3 and at Day 5 show a higher selection rate at pH 7 (Figure 3.6.3 B + C). The differences seen in these strains, suggest that these evolved strains have adapted to conditions in the evolution experiment, which are not associated with initial low pH,. but are present in both pH experiments and potentially have a stronger selective pressure at pH 7, such a length of time the cultures spend at high pH, or at stationary phase.

Interestingly, throughout each experiment, one strain (E4A) consistently had the highest selection rate, indicating that this strain had particularly adapted better to these conditions than the other evolved strains. However, since the dynamics of the evolution experiment is so diverse, using just these two experiments to decipher the particular adaptions to the evolved conditions, is near on impossible. In summary, by conducting competition experiments at pH 7 and pH 4.5, the differences in fitness observed confirm that the adaptions within these strains are to conditions in addition to those present at low pH.

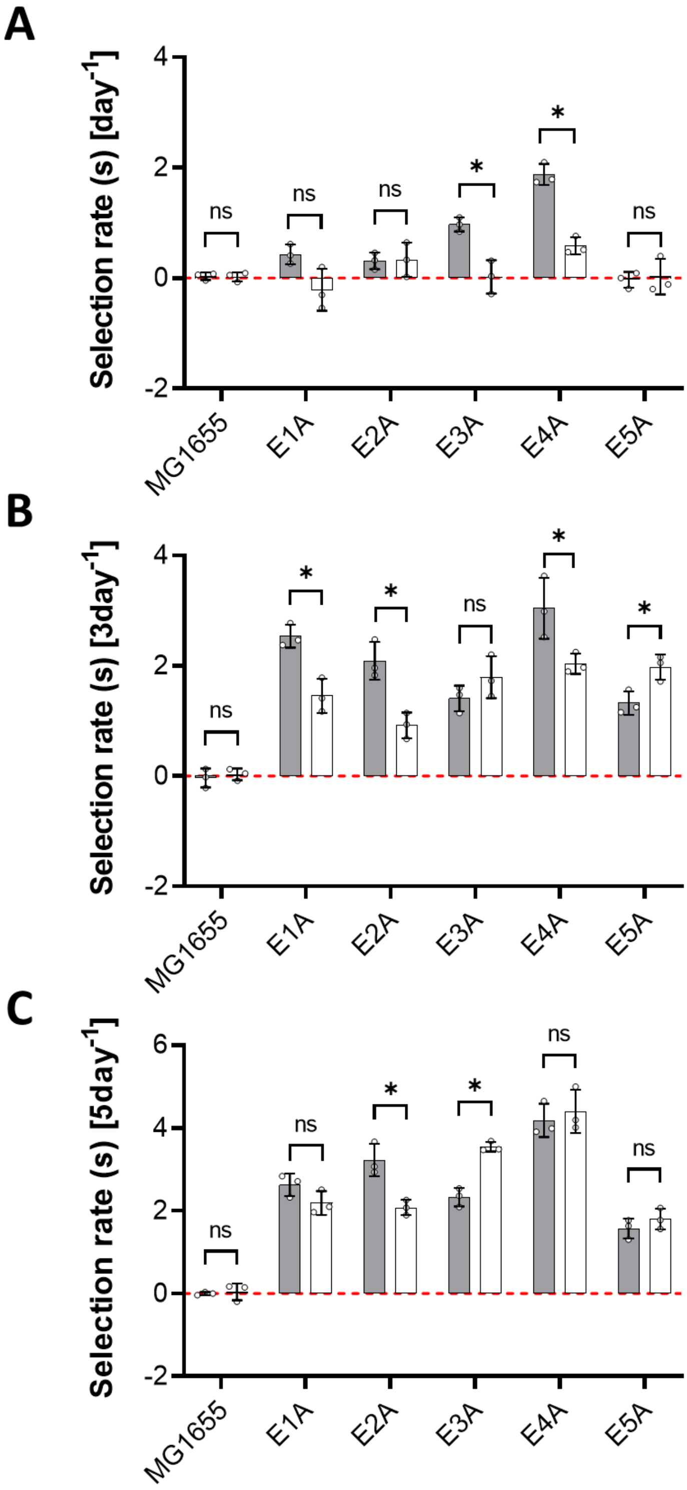


Figure 3.6.3: Comparison of selective rates of evolved strains against KH001 at either pH 4.5 (grey) or pH7 (white) at Day 1 (A), day 3 (B) or day5 (C) of a 5 day competition experiment. Statistical significance of differences was determined using multiple pairwise t test, with a false discovery rate of q < 0.05 using a Benjamini Hochberg procedure. Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

3.6.3 Impact of culture volume on measurements of fitness

The growth curves shown in Figure 3.3.1 + Figure 3.5.1 were conducted within 50ml of unbuffered LB at pH 7 or pH 4.5. As mentioned above, this was due to the amount of media required to accurately measure the OD_{600} and pH at each timepoint. Within the literature, difference in the size of the vessel and the volume of LB has been shown to have considerable effects on *E. coli* MG1655 phenotype, particularly altering the extent and time of the death phase and long term stationary phase (Kram *et al*., 2017, Gross *et al*., 2020, Kram and Finkel, 2014). Interestingly both, (Kram and Finkel, 2014) and (Gross *et al*., 2020) demonstrated that changing the volume of LB and size of the vessel showed differing extents of alkalization of the media after a single day growth. Although every attempt was made to maintain the conditions as close as possible to those in the original evolution experiment from which these strains arose, the amount of volume of media did differ. However, as described in section 3.3 the final pH and OD_{600} were highly similar between 5ml and 50ml conditions. To see whether any differences were observed in the phenotype of the evolved strains, competition experiments of the evolved strains using 50ml unbuffered LB at either pH 4.5 or pH 7 were conducted in 250ml Erlenmeyer flasks.

Initially, similar to the 5ml conditions, increased fitness of the evolved strains relative to KH001 was observed at both pH 4.5 and pH 7 in 50 ml of unbuffered LB (Figure 3.6.4). In addition, selection rates indicated that E4A was the fittest strain in both pH 7 and pH 4.5 unbuffered conditions. However, no significant differences between the fitness at pH 7 or pH 4.5 was observed within each evolved strain, indicating that the evolved strains had not adapted specifically to the initial low pH condition.

The selection rates reported in 50ml of LB after 1 day growth were similar to the those reported for 5ml conditions at Day 5 and higher than the selection rates at Day 1. This result this suggested that the evolved strains show higher fitness in 50ml of LB compared to in 5 ml LB. What is the factor that causes this difference in fitness? Although not much differs between these two conditions, what would be expected within a 50ml culture is a more efficient gaseous exchange with the media due to its larger surface area. Thus, it may be that the evolved strains could show greater fitness in more aerobic cultures.

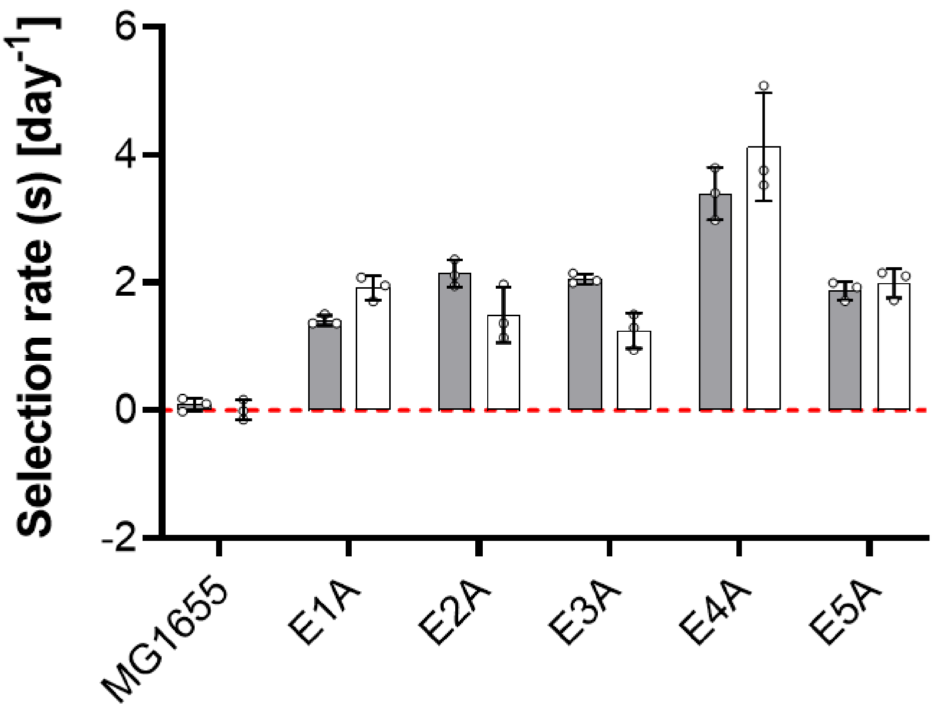


Figure 3.6.4: Competition of evolved strains against KH001 in 50 ml of unbuffered LB grown for 1 day at 37°C. An initial pH at either pH 4.5 (grey), pH 7 (white) was used for each evolved strain. The red line indicates no difference in fitness between the evolved strain and KH001 (selection rate of 0). No significance in fitness was observed between the two selection rates for any strain (2 ways t tests, corrected by a Benjamini Hochberg procedure). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

3.6.4 Cross protections and fitness trade-offs in the evolved strains.

The competition experiments described previously demonstrate that the evolved strains have adapted to conditions of the evolution experiment. One fundamental concept within evolutionary theory is that of cross-stress protections and tradeoffs that occur following adaptations. Briefly, this means organisms adapted to one stress, when exposed to a different stress, may also be fitter under that stress (cross-stress protection), or may lose fitness (trade-off). In the evolution experiment described in this study, the conditions of the experiment are variable such that the organism undergoes several different stresses throughout the experiment. Therefore, within this section, competition experiments were performed to identify the evolved strains fitness to defined conditions, both associated and not associated with the evolution experiment, to see whether evidence for either cross-protection or for fitness trade-offs could be seen.

The experiments conducted in this section were performed by myself and two final year undergraduate students, Katherine Cannings (KC) and Kryiaki Symrilli (KS), whom I supervised during their time in the laboratory. Due to the effort required to test all evolved strains in a single condition it was decided to only perform competitions for two of the five evolved strains, and for MG1655 as a control against KH001. The evolved strains chosen were E4A (as within the previous conditions E4A always showed the strongest fitness phenotype) and E1A (as a strain ancestral to E1A , E1A-I1, containing the arcA mutation had been isolated by Sen (2018)).

3.6.4.1 Using 50ml instead of 5ml

In order to assess the fitness of the chosen evolved strains it was decided to perform each competition using 1 day of growth in 50ml of culture in 250 ml Erlenmeyer flasks, instead of the multiple day's growth and passaging in 5ml volume in 30ml universals. This was to ensure each competition experiment was as consistent and as simple as possible by focusing on a single 24h cycle of growth. Within the 5ml competition experiments at Day 1, the majority of the evolved strains indicated very little change in selection rates indicating no difference in fitness compared to KH001. Therefore, to keep within the conditions of the evolution experiment each evolved strain was passaged under the same conditions.

The alternative was to use 50ml cultures over 1 day, since although this does not replicate the precise conditions used in the evolution experiment, it removes the complexity associated with stress associated with passaging (Figure 3.6.4). By comparing the results of 50ml at Day 1 with 5ml at Day 5 at both pH 4.5 and pH 7, similar trends between these two experiments were observed (Figure 3.6.3C + Figure 3.6.4). As both experiments indicated the evolved strains to be fitter relative to KH001 they also indicated E4A to be the fittest out of the evolved strains, with the largest selection rate. Although there were differences between the order of evolution strains in terms of selection rate, the overall trend was the same. Taking these points into consideration it was determined to use 50ml of LB, in order to be able to focus upon the effect that a single 24 hour growth in LB has an effect upon the cell.

3.6.4.2 Improved fitness in evolved strains is mainly at stationary phase

The previous competition experiments focused on measuring relative fitness after 24 hours of growth. This measurement therefore includes fitness differences that may act during any or all of the three stages of growth (lag, exponential and stationary phase). To begin to deconvolve the differences these stages of growth have on the evolved strains, a 6 hour timepoint was taken, which would provide a rough measure of altered fitness effects as the strains were exiting exponential phase and entering stationary phase (Figure 3.6.5). This experiment showed that the majority of improved fitness observed in the evolved strains E1A and E4A at both pH7 or pH 4.5 is due to differences in stationary phase. This was particularly the case for experiments at pH 7, indicating that the evolved strains E1A and E4A had adapted to conditions applying at high pH even though the experiment was originally devised to select for resistance to low pH (Figure 3.6.5B). Interestingly when the experiment was done in unbuffered LB at pH 4.5, some increased fitness was observed in both evolved strains after the 6 hour time point, but not when it was done in unbuffered LB at pH7. This suggests the evolved strains had some adaptations to lag or exponential growth in low pH conditions (Figure 3.6.5).

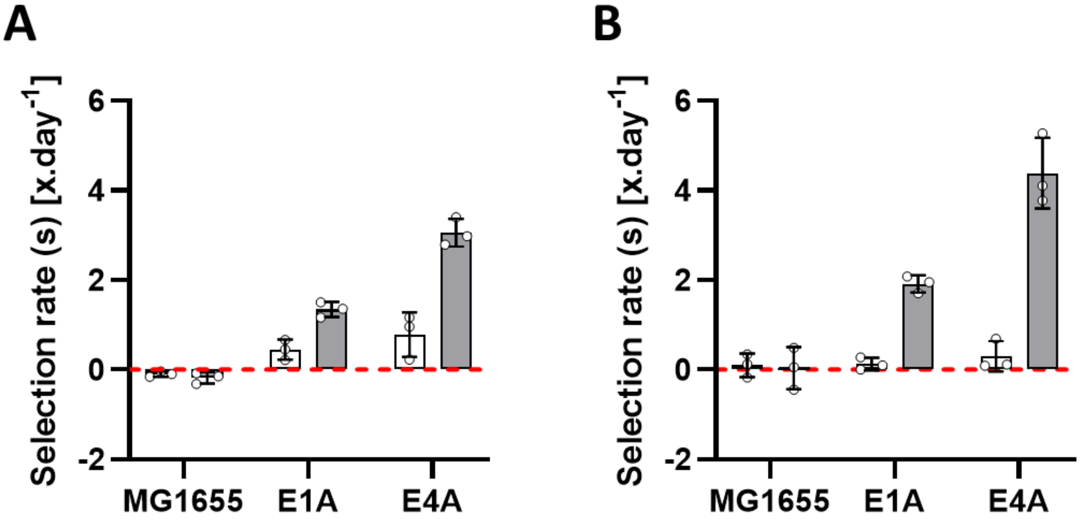


Figure 3.6.5: Competition experiments over 24 hours in 50ml of unbuffered LB at either pH 4.5 (A) or pH 7 (B) grown at 37°C with shaking. Selection rates were calculated at 6h (white bar) and 24h (grey bar). Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

3.6.4.3 Evolved strains show fitness at low and high pH

Since in unbuffered LB the main fitness observed in the evolved strains was during stationary phase and knowing that the use of unbuffered LB creates a pH change, we wished to determine whether fitness was due to stationary phase per se, or would only be seen in stationary phase at high pH. Therefore, to start to more fully decipher the effect of pH on fitness at stationary phase, competition experiments were performed with LB buffered using 50mM MES and MOPS at pH 4.5 and pH 7 (Figure 3.6.6). The use of MES and MOPS was expected to be unable to effectively buffer at pH 4.5 and indeed the buffer was shown to 'break' under these conditions (final media pH: pH 4.5 = 6.16 ± 0.04, pH 7 = 7.72 ± 0.03). However, the presence of the buffers prevented the cultures from reaching pH > 8. Overall, improved fitness was still observed in the evolved strains with a larger fitness observed during stationary phase as before. However, at pH 7, the selection rates for E1A and E4A after 24 hours were dramatically reduced in the buffered media compared to the unbuffered media, suggesting that adaptations which confer fitness in unbuffered LB at pH7 are somewhat associated with stress at high pH in stationary phase. Interestingly however, this was not seen within competition in LB buffered at pH 4.5. In this case, both strains showed similar fitness compared to unbuffered LB during exponential growth, and improved fitness at stationary phase, indicating that the evolved strains are fitter under acidic conditions.

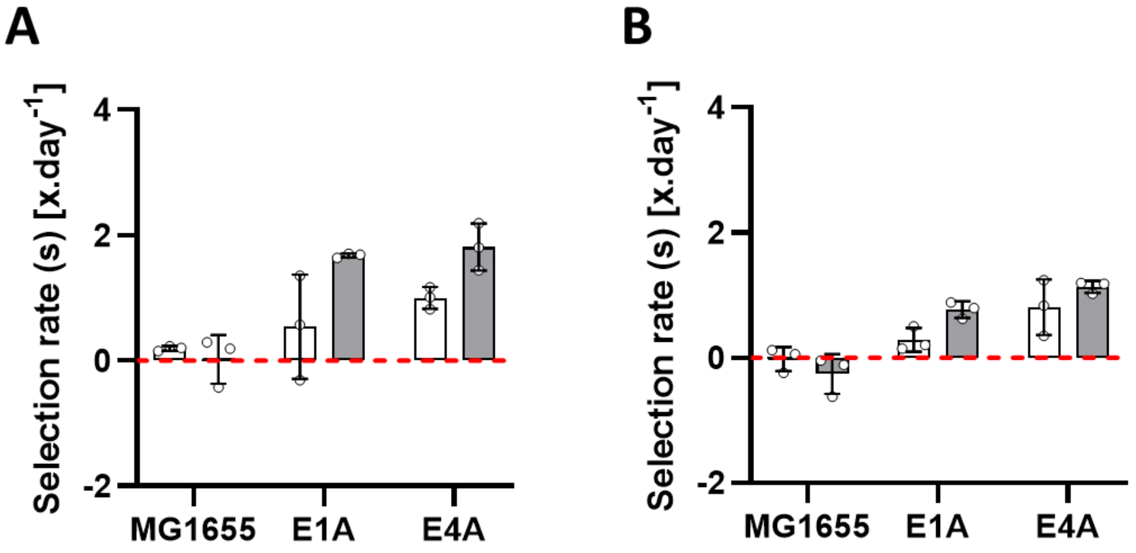


Figure 3.6.6: Competition experiments over 24 hours in 50ml of LB buffered with 50mM MES and MOPS at either pH 4.5 (A) or pH 7 (B) grown at 37°C with shaking. Selection rates were calculated at 6h (white bar) and 24h (grey bar). Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

Since, the above competition experiments suggested that adaptations of the evolved strains have occurred that improve fitness at low pH, but also at stationary phase at high pH, a further attempt was made to explore fitness specifically at low pH and high pH (pH 4.5 and pH 9, respectively) using more appropriate buffers to maintain the environmental pH (pH4.5 – HOMOPIPES, pH 9 – AMPSO) (Figure 3.6.7). Measurement of the pH after growth found that the capacity of the HOMOPIPES buffer was exceeded somewhat but acidic conditions were still maintained, while the AMPSO buffer held (Final Media pH: pH 4.5 = pH 5.27 ± 0.04 , pH 9 = 9.14 ± 0.01). Unfortunately, due to the COVID-19 pandemic the 6 hour time point of each condition was not obtained, however after the 24 hour growth the evolved strains were shown to be fitter than the KH001 at both acidic and alkaline pH (Figure 3.6.7). The results of this competition demonstrate that indeed each evolved stain has adapted to conditions both at low pH and at high pH, however unfortunately as a 6 hour timepoint was unable to be completed, we were unable to distinguish between exponential and stationary phase adaptation. Interestingly at pH 9, E4A shows a larger fitness than E1A indicating that this strain has adapted more to high pH (Figure 3.6.7B).

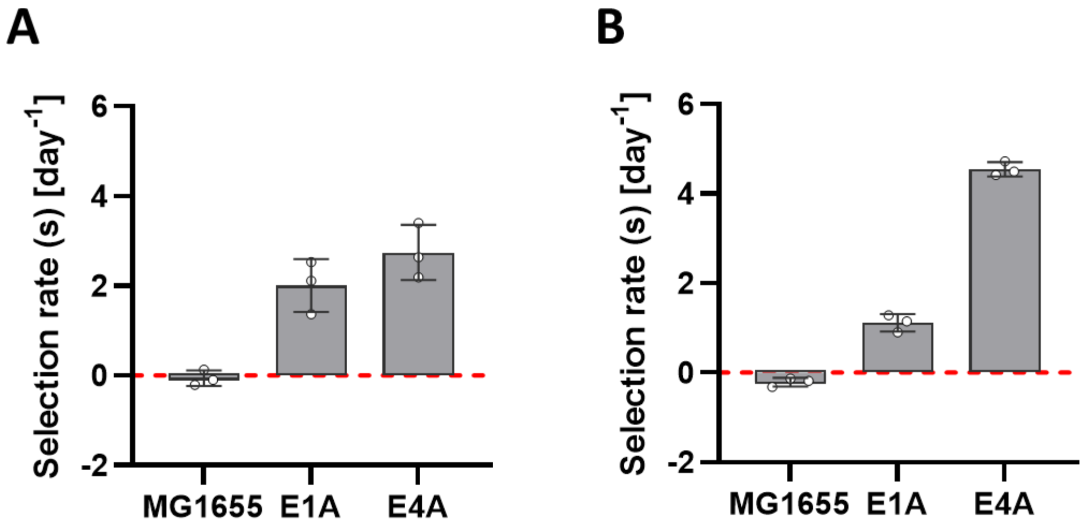


Figure 3.6.7: Competition experiments over 24 hours in 50ml of LB using buffers 50mM HOMOPIPES at pH 4.5 (A) and AMPSO at pH9 (B). Grown at 37ºC with shaking. Selection rates were calculated for 24 hours only. Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

3.6.4.4 The evolved strains fitness phenotypes are associated with growth in LB

All the above competitions consider the evolved strains within the context of LB. As explained above, LB is a medium consisting primarily of oligopeptides, the catabolism of which provides the main carbon source for the organism and also leads to alkalization of the media. To understand fitness with regards to the media the evolved strains were also competed with KH001 in M9 media supplemented with glucose and casamino acids at both pH 4.5 and pH 7 (Figure 3.6.8). Using M9 instead of LB changes several variables within the environment, the major one being glucose being used as the carbon source instead of oligopeptides. Since glucose was present, amino acids were not extensively catabolized, therefore the alkalization of environmental pH would not occur. Instead, acidification of the media would occur due to the production of acetate as a byproduct of glucose catabolism . To prevent this acidification from occurring M9 was supplemented with MES and MOPS. As observed with LB, at pH 7 MES and MOPS maintained the environmental pH, while at pH 4.5 this buffer combination broke causing the media to acidify slightly (M9 Final pH: pH 4.5 = 3.94 ±0.01 , pH7 = 6.84 ± 0.01). Overall competition experiments within this media showed there was a subtle increase in relative fitness of the evolved strains E1A and E4A. This fitness advantage was observed to occur within the first 6 hours of growth and did not significantly change after 24 hours indicating that this fitness advantage present within M9 occurs during exponential or lag phase, very different to the situation in LB. In addition, the selection rate did not increase above 1 for either strains at either pH 7 or pH 4.5 suggesting that the adaptations which give an increase in fitness in LB have a smaller effect in the supplemented M9 medium.

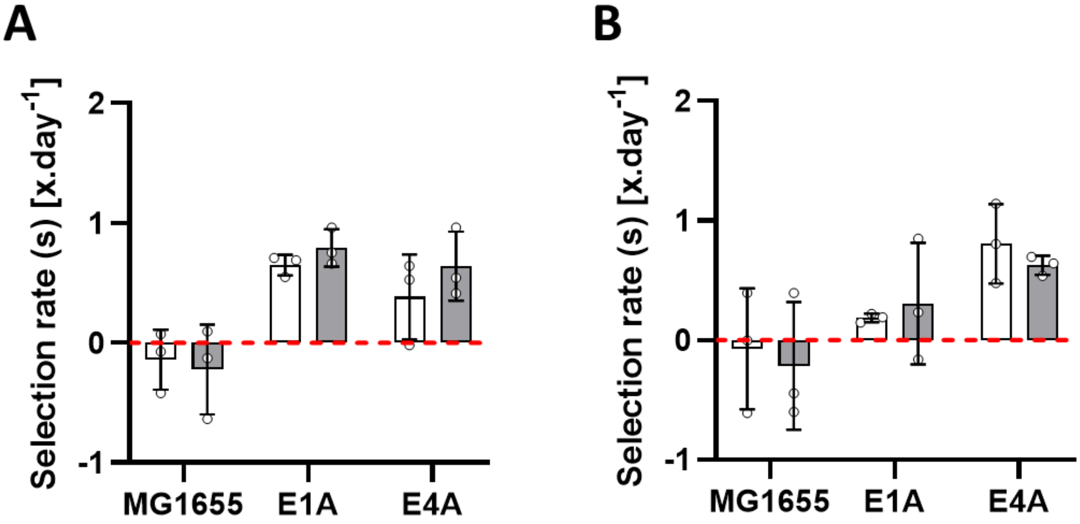


Figure 3.6.8: Competition experiments over 24 hours in 50ml of M9 media supplemented with 0.2% w/v glucose buffered with 50mM MES and MOPS at either pH 4.5 (A) or pH 7 (B) grown at 37ºC with shaking. Selection rates were calculated at 6h (white bar) and 24h (grey bar). Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

3.6.4.5 Fitness of the evolved strains in acetic acid

The evolved strains were also tested in the presence of 1mM acetic acid at both pH 7 and pH 4.5. Acetic acid is a weak organic acid and is likely to cause different stresses for the cell compared to strong acids such as HCl used in this study. Organic acids do not completely dissociate in water but exist in an equilibrium in solution between their fully ionized and unionized forms. The unionized form can cross the lipid membrane and once within the cytoplasm the acid can dissociate releasing a proton and lowering the cell cytoplasmic pH, thus bypassing some of the mechanisms *E. coli* uses to prevent this, such as alteration of the inner membrane. Organic acids can therefore have a larger effect on the cell than strong acids at the same pH, as the cell attempts to maintain a constant pH within the cytoplasm and pH gradient across the inner membrane. This effect of the organic acids can be observed at pH 4.5 as neither evolved strain, E1A or E4A, showed an increase in fitness compared to KH001, after 6 hours growth (Figure 3.6.9). Surprisingly however, E1A was found to be fitter than E4A in the presence of acetic acid at pH 4.5 and pH7 after 24 hours. This suggest subtle differences between the adaptions within E1A and E4A that allow for E1A to have a better response to the effects of acetic acid.

In the conditions of unbuffered LB at pH 4.5 with 1mM acetic acid, the selection rate of the MG1655 vs KH001 competition was significantly different (Figure 3.6.9A). Therefore, under this condition, KH001 cannot be considered as a good approximation to the MG1655 ancestor. Although fitness of the evolved strains can still be compared and described within this condition. When comparing the fitness stated under this condition to other conditions, the selection rates presented in Figure 3.6.9A cannot be directly compared.

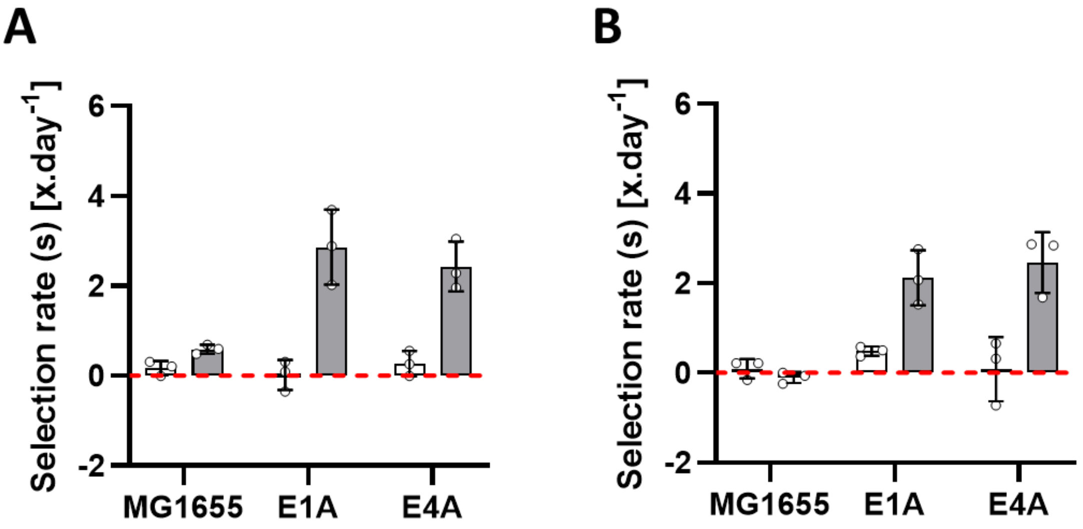


Figure 3.6.9: Competition experiments over 24 hours in 50ml of LB supplemented with 1 mM acetic acid at either pH 4.5 (A) or pH 7 (B) grown at 37°C with shaking. Selection rates were calculated at 6h (white bar) and 24h (grey bar). Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

3.6.4.6 Overall comparison of conditions.

The sections above describe an assessment of the fitness of the evolved strains over a range of conditions. The data show that fitness under the conditions of the evolution experiment at pH 4.5 can be associated with two broad categories: fitness at exponential phase associated with low pH and fitness at stationary phases associated with high pH. These fitness phenotypes also depends on the media used, as shown with the reduced fitness seen in M9 media relative to LB. Figure 3.6.10 shows a comparison of the selection rates generated at 24 hours for each condition. This shows that overall, for growth in unbuffered LB at pH 4.5 there is strong selective advantage, and E4A generally shows greater relative fitness than E1A. Considering the degree of fitness for each evolved strain individually, the conditions that showed the greatest increase in fitness was different between the two strains. In E4A, the three conditions with the highest relative fitness were at pH 4.5 and pH 7 unbuffered and at pH9 buffered with AMPSO, while on the other hand in E1A the highest relative finesses were seen in acetic acid at pH 7 and in pH 7 buffered with 50mM MES and MOPS. These differences indicate that each evolved strain has adapted differently to the environment of the evolution experiment.

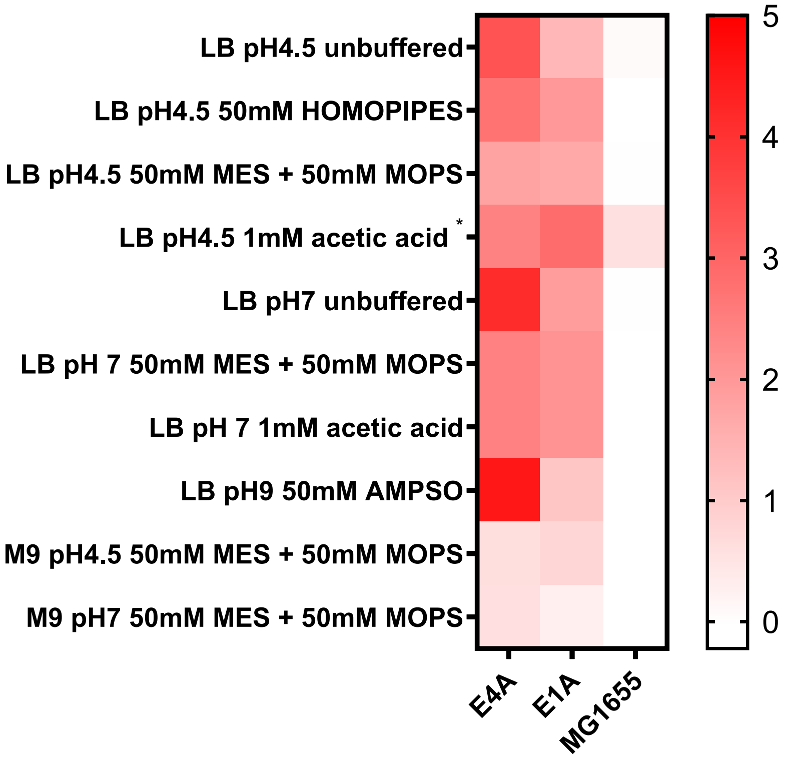


Figure 3.6.10: Heat map of selection rates observed from each competition experiment after 24 hour growth. With dark red indicating the larger differences. Note LB pH 4.5 with 1mM Acetic acid denoted with an Asterix *, the MG1655 vs KH001 control, indicated that MG1655 was fitter, therefore cannot be directly compared.

3.7 Investigating the effects of ArcA on fitness

As reported in section 3.4.1.1, mutations associated with *arcA* were observed in all evolved strains, and multiple mutations in *arcA* were seen within the evolved populations. *arcA* is part of a two component system with ArcB acting as sensor, detecting changes in the redox states of the cell, which can occur under anoxic conditions. This in turn creates a phosphorelay which results in the phosphorylation and subsequent activation of ArcA. Activation of ArcA has been mainly associated with *E.coli* switching from aerobic to anerobic respiration. Interestingly, since the evolution experiment was conducted under aerobic conditions, mutations in *arcA* therefore warranted further study described within this section.

3.7.1 How does loss of *arcA* affect the fitness phenotype of the evolved strains?

In each evolved strain, a mutation in *arcA* (E1A - E3A + E5A) or close to *arcA* (E4A) was observed. However, these mutations were not the only ones in the genotype of each evolved strain. Therefore, to examine the effect that each mutation associated with *arcA* had upon the evolved strains' fitness phenotypes, *arcA* was deleted in each evolved strain and from the MG1655 ancestor. Competition experiments were then performed using these *ΔarcA* strains against KH001 in unbuffered LB at pH 4.5 and these were compared back to their WT equivalent (Figure 3.7.1A). Unexpectedly this comparison revealed that in unbuffered LB at pH 4.5 no significant differences in fitness were observed between the evolved strains and their *ΔarcA* equivalents well as MG1655 ,which suggests that the mutations in *arcA* do not play a significant role in the growth phenotype of the evolved strains at this pH.

As the results showed previously, the evolved strains also had higher relative fitness in unbuffered LB at pH 7. A competition experiment using the *arcA* deletion strains was therefore also conducted in unbuffered LB at pH 7. Under these conditions, a difference in the fitness phenotypes was observed between the *arcA* deletion and WT strains (Figure 3.7.1). When *arcA* was deleted in the evolved strains, a decline in fitness at pH 7 was observed. Although this was only statistically significant in three strains, this suggests that the arcA mutations contribute to the fitness phenotype observed in the evolved strains at pH 7. In addition, a large decline in fitness was observed for MG1655 *ΔarcA* compared to WT, indicating that ArcA function is necessary for growth in unbuffered LB at pH 7.

The results above suggest that the mutations in *arcA* within the evolved strains, do not simply cause loss of function of ArcA. The reasoning for this is that if these mutations did in fact cause loss of function of ArcA, then in competition experiments MG1655 *ΔarcA* would be fitter than the MG1655 ancestor. In the data this is not the case, as at pH 4.5 no change in fitness of the MG1655 ancestor compared to MG1655 *ΔarcA* was observed, and at pH 7 a decrease in fitness occurred. The deletion of *arcA* within the evolved strains would also be predicted to show different results, as if the mutations that arose in evolution did cause of function, no change in fitness would be expected if these mutated *arcA* genes were then deleted. Although this was seen in the results in unbuffered LB at pH 4.5, the *arcA* deletions within the evolved strains had a fitness disadvantage at pH 7 suggesting that it was very likely that the mutations within *arcA* do not cause loss of function. Instead, the results above indicate that mutations in (or associated with) *arcA*, alter the function of ArcA in a way that enhances the fitness of the evolved strain.

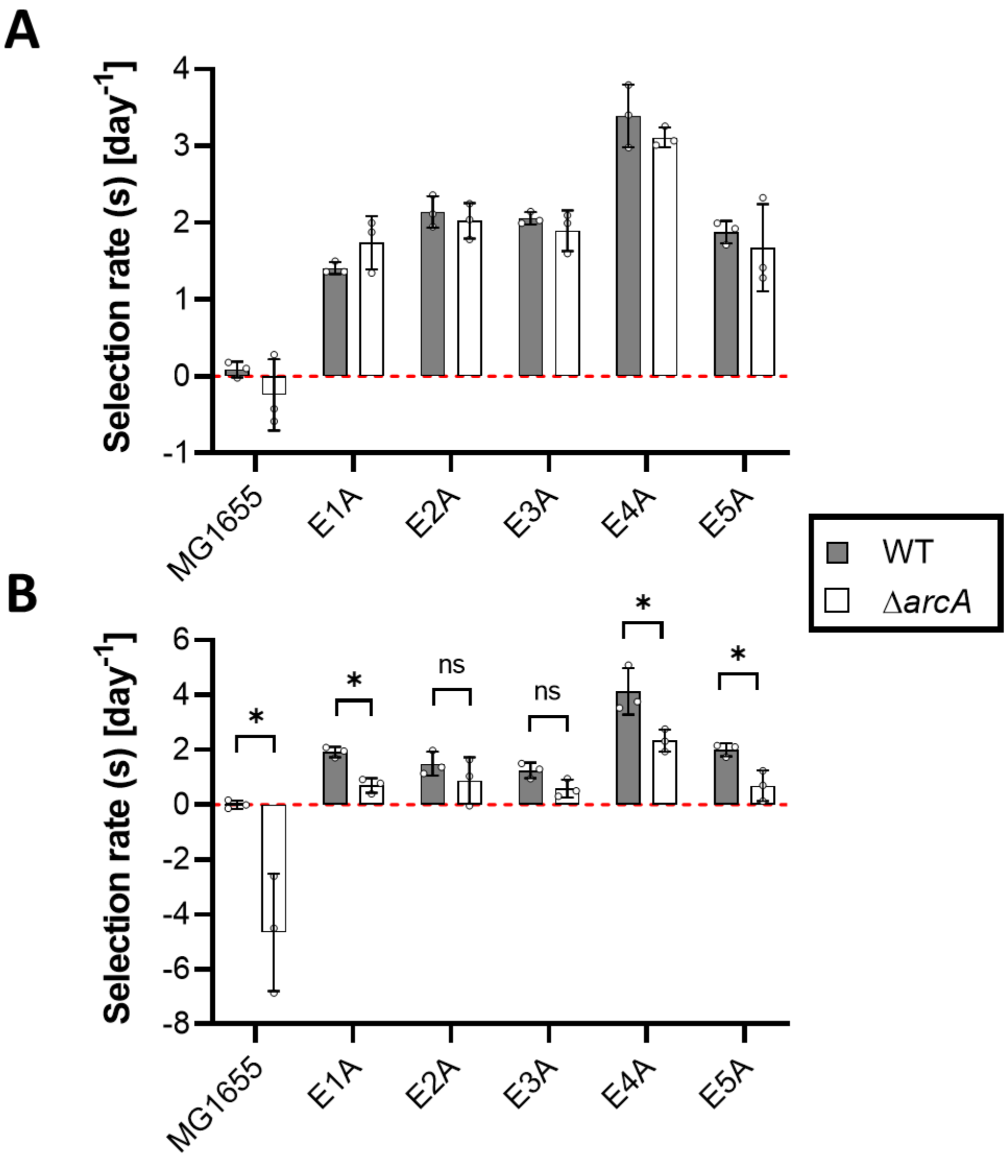


Figure 3.7.1: Competition experiment comparing the effect of an *arcA* deletion in the evolved strains and MG1655 ancestor after 24 hours of culture at 37°C, in 50ml of unbuffered LB with an initial pH at either pH 4.5 (A) or pH 7 (B). Statistical significance was determined using multiple t tests with a false discovery rate determined by the Benjamini Hochberg procedure. At pH 4.5 no significant differences were observed between the evolved strain and the *arcA* deletion. Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points. .

3.7.2 Considering the loss of function of *arcA* within the parameters of the evolution experiment

The experiment above indicated that *arcA* has a significant role in the fitness phenotype observed at pH 7 but probably not at pH 4.5. Since the conditions of these competitions did not directly reflect conditions under which the evolution experiment was done, it was therefore decided to assess the role that loss of function of *arcA* has upon fitness under the conditions of the evolution experiment. Therefore, competitions of MG1655 *ΔarcA* with wild type MG1655 was done over 5 days in 5ml of unbuffered LB at pH 4.5 or pH 7 using the same passage series as the original evolution experiment (Figure 3.7.2)

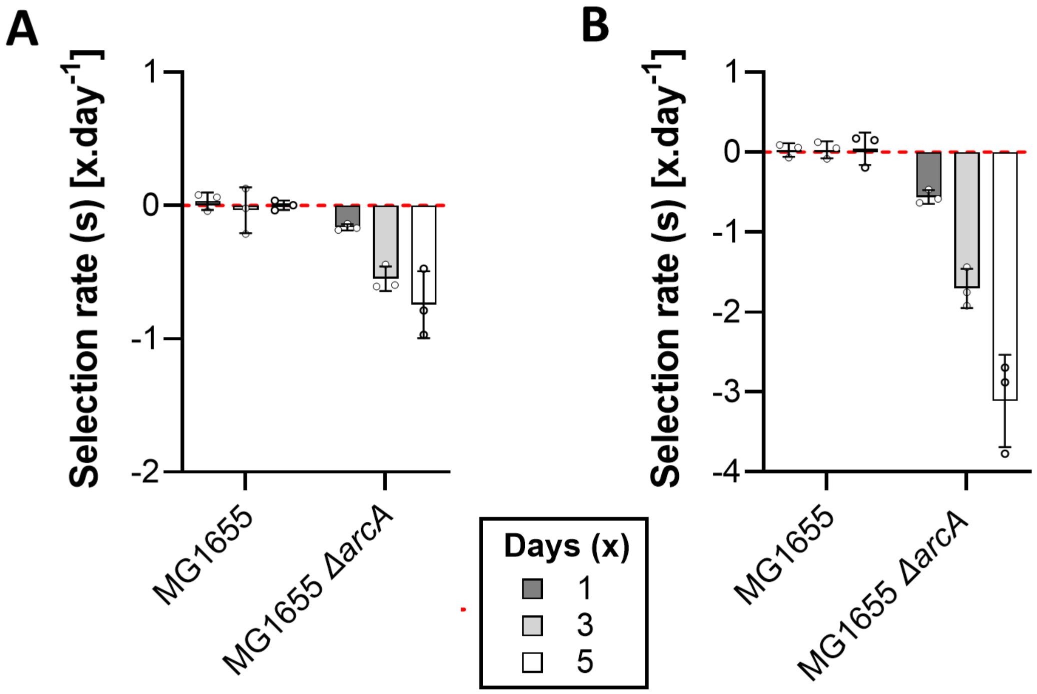


Figure 3.7.2: Competition of MG1655 ancestor and MG1655 *ΔarcA* against KH001 over 5 days in 5ml of unbuffered LB at either pH 4.5 (A) or pH 7 (B). After 24 hours cultures were passaged into fresh media at a 1 in 20 dilution. Selection rates were calculated by comparison with final counts at time 0. Therefore, the rate described change for each time point (Day 1 = Day^{-1} , Day 3 = 3 Days^{-1}, Day 5 = 5 Days^{-1}). Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

These results show that at both pH values, functioning ArcA is required for full MG1655 fitness under these growth conditions. As expected from our earlier data, the results of these competitions show that at pH 7, loss of *arcA* causes a decrease in fitness (Figure 3.7.2B). However, at pH 4.5, a fitness decrease in MG1655 *ΔarcA* was also observed over time. The difference in selection rate observed between pH 7 and pH 4.5 was significant and indicated that *arcA* function is required more at pH 7 under these conditions. Although growth in unbuffered LB at pH 4.5 and pH 7 generates different changes in environmental conditions, the results suggest that some aspect of *arcA* function is required for maximum fitness under a specific condition or conditions that occur more during growth in unbuffered LB at pH 7 compared to pH 4.5.

3.7.3 Assessing the role of the intermediates in ArcA function.

The above section demonstrates that the mutations that were selected for in the lab evolution experiment and are associated with *arcA* contribute to the fitness phenotype observed in the evolved strains at pH 7 but not pH 4.5 (Figure 3.8.1). During the evolution experiment Sen, (2018) identified two precursor strains associated with the E1A. The first, termed E1A-I1, was a strain isolated 10 days after the evolution experiment began. Genome sequencing of this strain was showed it contained the single *arcA* mutation M39I (Section 3.4.3). In addition to this, another strain termed E1A-I2 with genotype of only arcA(M39I), rpoA, (N294H), cytR::IS5 was isolated from a frozen fossil record approximately 4 months into the duration of the experiment. These two strains are precursor genotypes to the E1A evolved strain described in this study. Therefore, in order to assess the effect of these genotypes on fitness, each strain E1A-I1 and E1A-I2 was competed against KH001 and compared to the evolved strain's (E1A) fitness phenotype in unbuffered LB at pH 7 and pH 4.5.

The results, presented in Figure 3.7.3, showed that the E1A-I1 strain was fitter at pH 7 and pH 4.5, demonstrating that the M391I mutation in *arcA* contributed to fitness at both these pH conditions. In addition, the E1A-I2 strain was also shown to be fitter than KH001. However, although both strains were fitter at both pH values, the strength of selection rate differed in an unusual manner. Considering these intermediates as precursors to the evolved strain, the expectation was that the fitness observed would relate to the genotype, with E1A-I1 having only a single arcA mutation having the smallest increase in fitness, with

E1A having the largest increase. Indeed, for growth in unbuffered LB pH 4.5 (i.e., the conditions used for selection) this was the case, with the E1A-I1 strain being less fit than E1A-I2 and E1A. Interestingly the fitness did not change between E1A-I2 and E1A, suggesting the three mutations seen in *arcA*, *rpoA* and *cytR* in the E1A-I2 strain are responsible for the majority of the fitness phenotype observed in E1A.

This incremental relationship between the fitness values of these strains in unbuffered LB at pH 4.5 was however not observed at pH 7. Instead E1A-I1 had the largest increase in fitness, which declined with E1A-I2 and further with E1A. The differences in fitness between these three strains suggests in E1A, mutations in different genes have been selected for due to adaptation to different selective factors in the evolution experiment conditions. This suggests the possibility of a fitness tradeoffs, where the mutation in arcA enables adaptation to a particular selective factor (which is potentially present in both conditions, but is more selective at pH7), but where the other mutations acquired in E1A confer fitness at pH 4.5 but are deleterious at pH7.

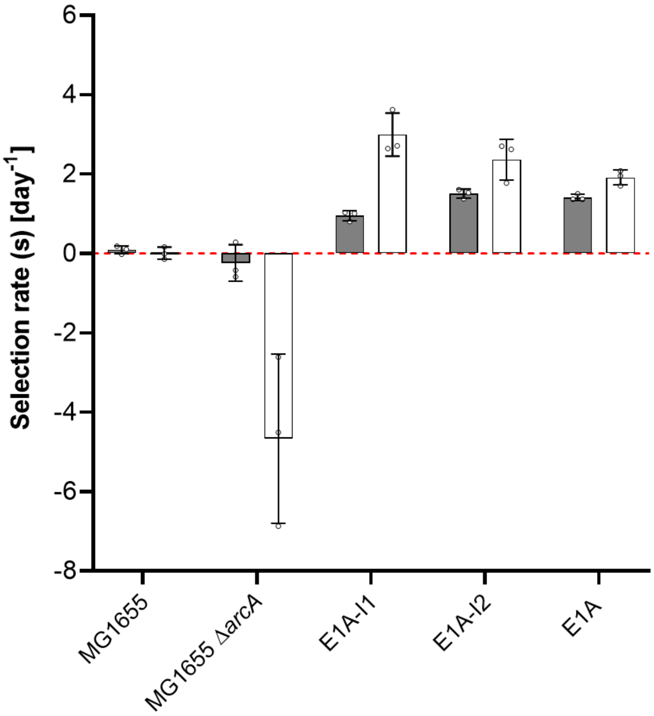


Figure 3.7.3: Competition of E1A and its precursor strains E1A-I1 and E1A-I2 with KH001 after 24 hours growth at 37ºC in 50 ml unbuffered LB at either pH 4.5 (Grey bars) or pH 7 (white bars). Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

3.7.4 Investigating ArcA activity within evolved strains and intermediates.

The above results suggest that within the parameters of the evolution experiment, the mechanism of how the mutations observed in (or associated with) *arcA* confer fitness is complex. The results so far demonstrated that ArcA function is required for fitness, and that mutations associated with *arcA* within the evolved strains must alter the function in some manner. ArcA is a regulator of gene expression. Therefore, to investigate this further, an ArcA activity reporter described in Park and Kiley (2014) was used. Using this reporter, ArcA repression activity is reported using a *lacZ* fusion with the *icdA* promoter. Isocitrate dehydrogenase, *icdA*, is involved in the TCA cycle converting D-isocitrate to 2-oxoglutarate, and its repressed by ArcA and FruR. To ensure only ArcA activity was reported, Park and Kiley (2014) removed FruR repression by site directed mutagenesis. The ArcA activity reporter was then recombined into location of the lac operon to ensure only ArcA activity was reported, with a kanamycin cassette marker to allow for selection. The results presented below describe the use of this ArcA activity reporter to begin to understand the affect mutations present in the evolution experiment had upon ArcA activity.

3.7.4.1 Use of β-mercaptoethanol within β-galactosidase assays

To report on the activity of a promoter fused with lacZ, a β-galactosidase assay is typically conducted. β-galactosidase assay reports activity based on the hydrolysation of ortho-Nitrophenyl-β-galactoside (ONPG) to galactose and ortho-nitrophenol which can be quantified using a colourmetic assay. To perform a β-galactosidase assay, a bacterial lysate and ONPG is suspended in “Z buffer” (Section 2.6.1), which provides an environment where β-galactosidase is able to function, buffering the pH and providing the relevant metal ions required for function. In addition to this, β-mercaptoethanol (BME) is added at low concentration to further replicate the reducing environment seen in the cytoplasm to increase stability of β-galactosidase. However, BME is toxic causing side effects such as headaches, vomiting and nausea, and some cases individuals (myself included) can be more susceptible to the effects BME than others.

Since BME is only present in Z buffer to increase stability, the question was asked about whether addition of BME was necessary to conduct and accurately report a β-galactosidase assay. Therefore *E.coli* K-12 MG1655 and its lac$^-$ derivate overnight cultures were assayed for β-galactosidase activity with or without BME. This comparison showed that overall, no significant difference was observed in the β-galactosidase activity reported with or without BME (Figure 3.7.4). Therefore, in all β-galactosidase assays performed no BME was used.

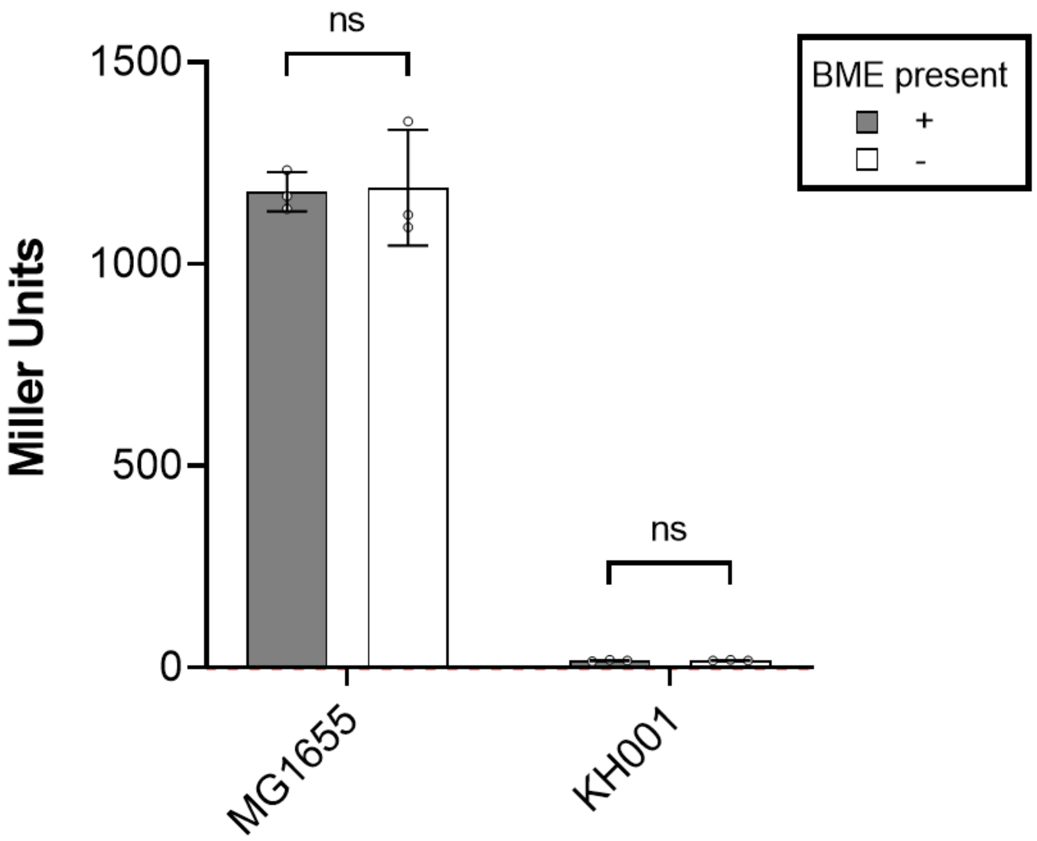


Figure 3.7.4: Investigating the effect of β-mercaptoethanol (BME) upon β-galactosidase assays using *E.coli* MG1655. Statistical significance was reported using a pairwise t test (pval 0.05). The mean +/- SD are plotted alongside the individual data points.

3.7.4.2 ArcA activity within the E1A intermediates

In Section 3.7.3, the intermediates of E1A presented different fitness phenotypes when subjected to pH 4.5 compared to pH 7. Therefore, considering that E1A-I1 was the only strain isolated which a single *arcA* mutation was present, the *picdA::lacZ* construct was introduced into E1A and the E1A-I1 intermediate by P1 transduction. These strains were then grown over 24 hours using 5 ml cultures in unbuffered LB at pH 4.5 and pH 7 taking a time point at 6 and 24 hours (Figure 3.7.5). Since the construct *picdA::lacZ* measures ArcA activity by measuring repression, the expression of LacZ reported in a *ΔarcA* background should represent the maximum possible activity (i.e., there will be no ArcA mediated repression in this strain). In MG1655, ArcA will be able to repress therefore expression of LacZ should be reduced. At pH 4.5 the activity of the *arcA* mutation observed within the E1A-I1 strain, showed similar repression activity to that of MG1655, indicating that ArcA repression functioned normally under these conditions (Figure 3.7.5A). At pH 7 at 6 hours, ArcA activity was similar to MG1655. After 24 hours, in the E1A-I1 strain ArcA activity to repress was reduced compared to the wildtype indicating that in some manner the mutation present within *arcA* alters the function, in this instance reducing repression activity after 24 hours growth at pH 7 (Figure 3.7.5B).

Unusually, within the evolved strain E1A, at both pH conditions, the activity reported was higher than the *arcA* deletion strain. This result therefore reports a difference in ArcA activity between the evolved strain and E1A-I1 strain even though they both possess the same *arcA* mutation. This result could have three explanations. First, the proportion of phosphorylated ArcA present within the E1A-I1 strain may be higher than E1A. In this first instance however, the increase in activity of the E1A compared to the *arcA* deletion, would not be accounted for. The second instance suggests that other mutations present in E1A have an effect upon the reporter, therefore ArcA activity is not truly reported here. Although every attempt was made by (Park and Kiley, 2014) to ensure ArcA specificity of the reporter, one mutation in E1A which may cause an affect upon this reporter is within the RNA polymerase subunit α, (RpoA$_{N294H}$) which since it is required for transcription could be predicted to have an effect upon expression.

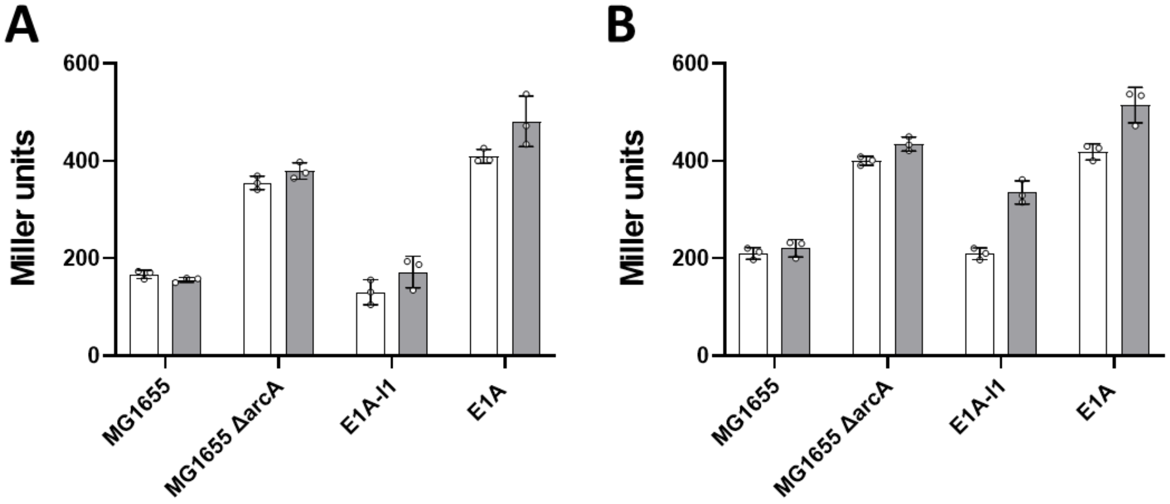


Figure 3.7.5: ArcA activity of the Evolved strain E1A and its intermediate strain E1A-I1 over 24 hours of culture in 5ml of unbuffered LB at pH 4.5 (A) and pH 7 (B). Timepoints were taken at 6 hours (white bars) and 24 hours (grey bars). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

3.7.4.3 Considering ArcA activity within the evolved strains at pH 4.5

As reported in Section 3.7.4.2, the differences observed between the E1A-I1 and Evolved populations indicated that *picd::lacZ* may not just be reporting the activity of ArcA. However, since different mutations were observed within (and associated with) *arcA,* the evolved strains were subsequently assayed using the ArcA activity reporter in 5 ml unbuffered LB at pH 4.5. The results presented in Figure 3.7.6 show, that with the exception of E4A, the activity of ArcA within the evolved strains was greater than the *arcA* deletion. This indicates that with regards to the promoter of *icdA*, expression is increased in the evolved strains compared to the MG1655 ancestor. This result again suggests that ArcA activity reporter is not only reporting the activity of ArcA.

As mentioned above one particular mutation which may cause this increase in activity, is within *rpoA*. Interestingly the one strain which did not possess a *rpoA* mutation, E4A, also did not see an increase in activity above that of the *arcA* deletion. This result suggests that mutations not present in E4A, but present within the other evolved strains, such as *rpoA,* had an effect upon *icdA* expression.

Overall, however the use of this reporter to determine ArcA activity is contextual, since it only reports activity specific to one promoter, so using it to determine the exact effect that ArcA has upon global gene expression within an evolved strain is limited. For this reason, it would be of great interest to obtain RNA-seq data for these strains grown under the different conditions used here. This is discussed below (Section 3.9). However, the result presented in Figure 3.7.6 suggests that within the evolved strains expression of *icdA* is increased. An exception would be in E1A whose genotype includes a deletion of the endogenous gene.

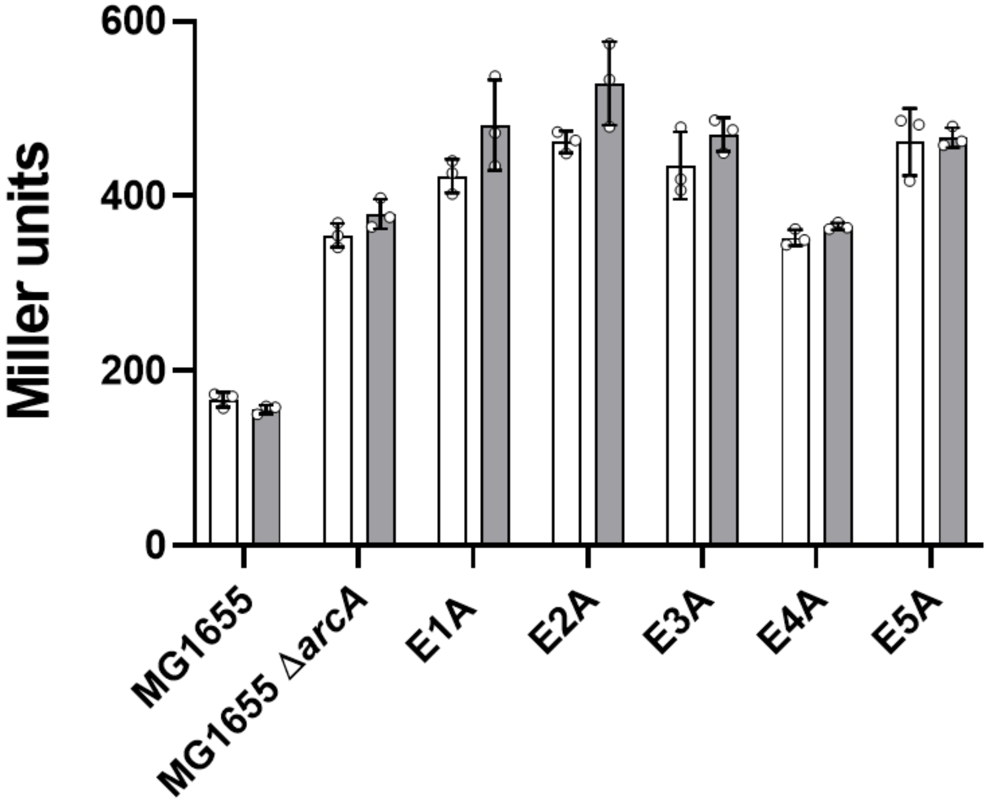


Figure 3.7.6: ArcA activity of the evolved strains at pH 4.5 in 5ml unbuffered LB starting at pH 4.5 in unbuffered pH. Timepoints were taken at 6 hours (White bars) and 24 hours (Grey bars). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

3.8 The effect of rpoS on the evolved strains.

Considering the fitness of the evolved strains discussed in Section 3.6, the majority of fitness observed within the evolved strains occurred when the strains were in stationary phase. One major factor which is involved in initiating the general stress response of the cell is the alternative sigma factor, *rpoS,* a subunit of the RNA polymerase complex. Levels of *rpoS* and RpoS activity typically increase when the culture enters stationary phase. Therefore, it was decided to find whether RpoS activity was affected within the evolved strains. To do this a RpoS reporter (psyn8::GFP) was created as described in Section 2.3.7. This RpoS reporter plasmid consisted of a GFP fusion with a synthetic promoter sequence specific to RpoS activity as described in (Klauck *et al.*, 2018). To ensure that the reporter was able to report RpoS activity, normalised fluorescence of MG1655 and MG1655 *ΔrpoS* was compared. A significant difference was observed between these two strains indicating that RpoS activity was being reported (Figure 3.8.1).

Using this reporter, RpoS activity was determined within E1A and E4A of the evolved strains after 6 hours and 24 hours of growth in 5ml of unbuffered LB at pH 4.5 (Figure 3.8.1). The results show that at both 6 hours and 24 hours of growth, no significant difference was observed in activity between MG1655 and the evolved strains E1A and E4A. The results presented suggest that differences in rpoS activity do not occur within the evolved strains. The conclusion observed in this study differ to those reported in Sen (2018). However, analysis of the reporter plasmid used in Sen (2018) showed that the *rpoS* promoter sequence was incorrect, which is why it was re-engineered within this study.

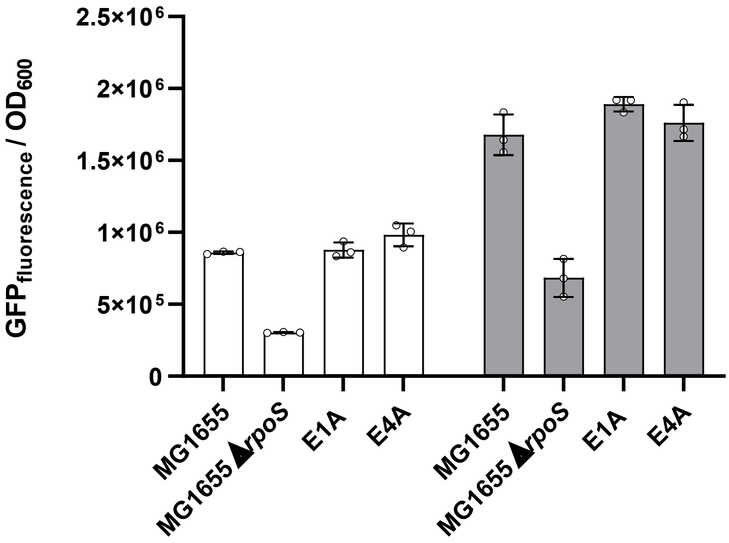


Figure 3.8.1: RpoS activity of evolved strains at pH 4.5 over 24 hours growth in 5ml unbuffered LB at pH 4.5. Activity was reported as $GFP_{fluorescence}$ normalised by OD_{600}. Activity was calculated at timepoints of 6 hours (white bars) and 24 hours (grey bars). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

3.9 RNA-seq of the evolved strains.

In the above sections the promoter fusions only describe the activity one transcription factor in the context of one promoter. What they do not capture is the overall global effects in expression which would occur due to these mutations within the evolved strains. To begin to address this, Sen (2018) sampled the transcriptome by performing RNA-seq on the evolved strains and MG1655 ancestor during mid/late exponential phase in unbuffered LB at pH 4.5 (0.8 OD_{600}). At this $OD_{600,}$ the culture of the pH would be predicted to be pH 5.2 – pH 5.4.

Once RNA-seq was performed, as described in Sen (2018), further analysis was performed by John Herbert of the University of Liverpool. This analysis included performing differential expression analysis using *edgeR* and following this with Gene Set Enrichment Analysis (GSEA) (Sen, 2018, Robinson *et al*., 2010, Subramanian *et al*., 2005). GSEA, in simple terms, uses predefined gene sets created with a biological rationale, and looks for differences in levels of expression of gene sets between two sets of data. GSEA is then able to determine statistically whether a gene set differs in expression level between the two groups, by assigning an enrichment score. With regards to expression data, the enrichment score can also determine whether a statistically different gene set is upregulated or downregulated.

The results below describe the GSEA analysis performed using differential expression data from Sen (2018). The results described in Section 3.10.1 were generated by John Herbert using a gene set describing transcription factor targets from RegulonDB, while in Section 3.9.2 GSEA analysis was performed by myself using metabolic pathway gene sets generated from the KEGG database.

3.9.1 GSEA using Transcription Factor Targets.

Using differential expression data of the evolved strains compared to the MG1655 ancestor, GSEA was performed using gene sets taken from RegulonDB, with each set referring to the target of transcription factors (Figure 3.9.1). Initially the two gene sets which on average showed the highest enrichment were ArcA and CytR. This enrichment indicated that an upregulation of expression was observed within these gene sets. Considering how this result relates to the function of mutations observed, since CytR only acts as repressor, upregulation of expression within CytR regulon observed suggests that the mutations in or upstream of *cytR* potentially cause loss of function, and this removes CytR repression. This same observation however cannot be applied to the mutations within (and associated with) *arcA*, as ArcA has shown to be an activator and repressor. However, the upregulation of the ArcA regulon corresponds to the increase in ArcA activity within the evolved strains using the *picd::lacZ* reporter described in section 3.7, using the ArcA activity reporter within the evolved strains. In addition to CytR and ArcA, other gene sets were shown to be upregulated, in particular in transcription factors associated with central metabolism. In particular several gene-sets had high enrichment scores associated with metabolism in particular carbon and nitrogen metabolism including SprR, SrlM, PhdR, CRA, GlcC, GalS and CRP.

As well as upregulation, other gene-sets were shown to be downregulated such as the regulons FlhC and FlhD. These two regulators function in a complex to regulate the expression of the flagellum. Interestingly in the genotypes of the evolved strains, only one strain, E4A, had a deletion of these two transcription factors. Therefore, the results suggest that the mechanism whereby these regulons are downregulated may be different in different evolved strains. The downregulation of the expression of the flagellum predicts that the evolved strains would be non-motile.

In addition to FlhC and FlhD, several transcription factors involved in the AR2 acid resistance mechanism, in particular the gene sets of GadW, GadX, YdeO, GadE and EvgA, were downregulated in at least 1 evolved strain. This downregulation suggests that the AR2 is not required within the evolved strains, at least at the time sampled. However how this downregulation relates to the fitness of the evolved strain is unknown, since the AR2 is mainly associated with survival at extreme pH (e.g. pH 2.5), which would not occur within the conditions of the evolution experiment. However, when the cultures for RNAseq were sampled, the environmental pH would be approximately pH 5.2 -5.4 based on the OD_{600} described by Sen (2018). Since AR2 can be induced at mild pH (~pH 5.5), these cultures would have already exposed to mild pH when being sampled. It could be that the evolved strains, therefore, have been selected to reduce or prevent the induction of the AR2. The physiological reason for this is not known.

Considering the extent of the differences involved in the evolved strains, based on this analysis out of the 83 gene sets analyzed in this GSEA analysis, 63 of these gene sets were shown to significant in at least one evolved strain, indicating that within the evolved strains a global change in expression has occurred. In addition, within each evolved strain, the majority of the same gene sets were significantly different in each evolved strain indicating an overall similar response between the strains. However, there were also gene sets which are significantly enriched unique to each strain. In particular E4A and E5A were shown to have more transcription factors regulons significantly enriched in the GSEA analysis potentially, suggesting subtle differences in responses between strains.

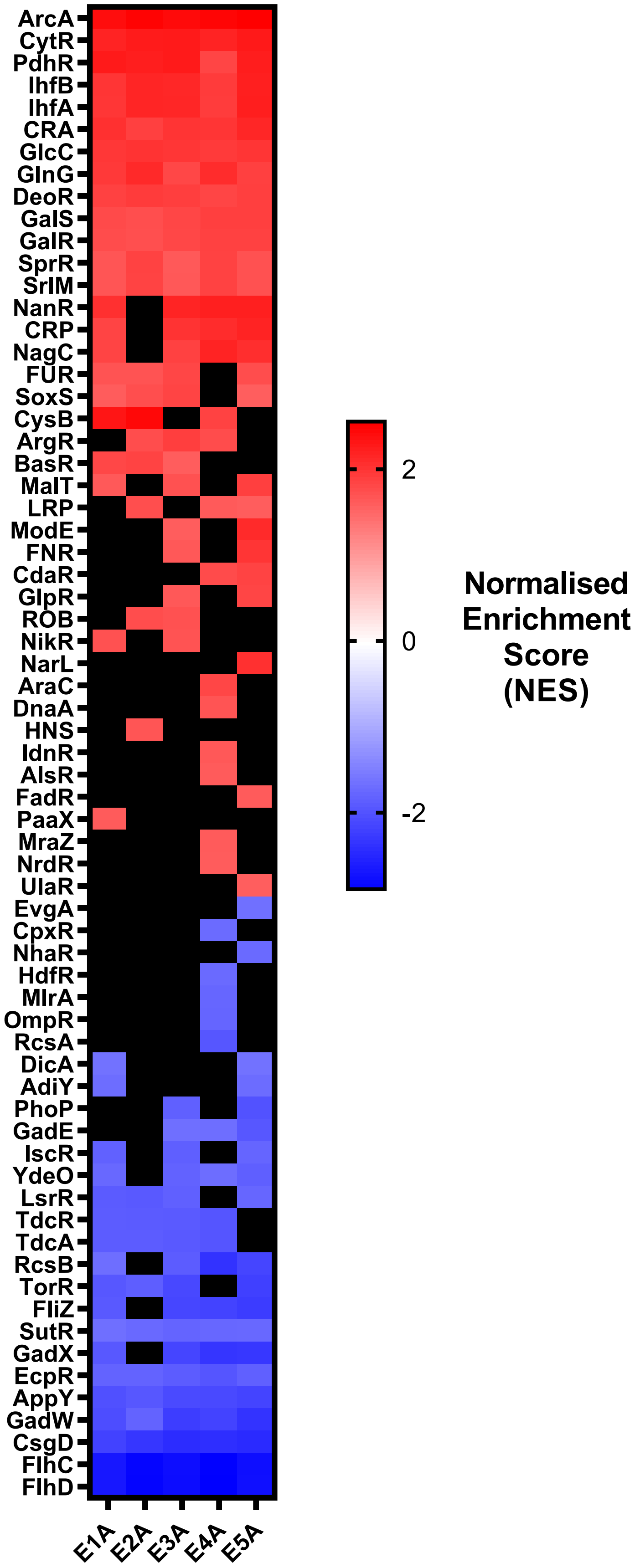


Figure 3.9.1: Heatmap illustrating the outcome of GSEA of differential expression data of the evolved strains against MG1655 ancestor using a gene set of Transcription factor regulons from Regulon DB. Gene-sets were ranked based on the average NES score across all 5 evolved strains. Gene sets predicted as significant using an FDR < 0.05, the normalised enrichment scores are presented to indicate upregulation (red) or downregulation (blue) of the gene-set. Gene-set which were not significant in any evolved strain are not shown, and where not significant in at least one strain this is indicated by a black entry. The analysis presented was performed by John Herbert.

3.9.2 GSEA using Kegg pathway analysis

The previous GSEA described above indicated that within the evolved strains, several regulons of transcription factors associated with metabolism were upregulated. Therefore, to consider the effects of this upregulation observed within the expression data, GSEA was performed using gene sets referring to different metabolic pathways from the KEGG database (Figure 3.9.2).

As expected from the previous analysis, several gene sets shown to be associated with cell motility, particularly the flagella, were also observed to be strongly down-regulated using this alternative gene set. The results of this analysis also show that within the evolved strain carbon metabolism is upregulated, in particular the TCA cycle, glyoxylate and dicarboxylate metabolism (referring to the glyoxylate shunt), and oxidative phosphorylation, indicating that the evolved strains may have adapted by altering patterns of nutrients utilization in the media. In addition, several additional carbohydrate and fatty acid metabolism pathways were enriched as well as amino acid metabolism and degradation pathways, indicating that the evolved strains are promoting pathways involved in creating molecules which cannot be obtained directly from the media. When considered with the result described in Section 3.6.4.4 where little difference in fitness was observed in the evolved strains competed in M9 minimal media, this suggests that the evolved strains have adapted towards utilization of nutrients within LB media. In addition, in 3 out of the 5 evolved strains the phosphotransferase system gene set was also shown to be enriched suggesting active transport of nutrients from the environment into the cell was being promoted. In conclusion the results of this analysis potentially suggested that during mid-late exponential phase transcription was altered in the evolved strains to allow for more efficient uptake and utilization of nutrients available in the environment, than the MG1655 ancestor.

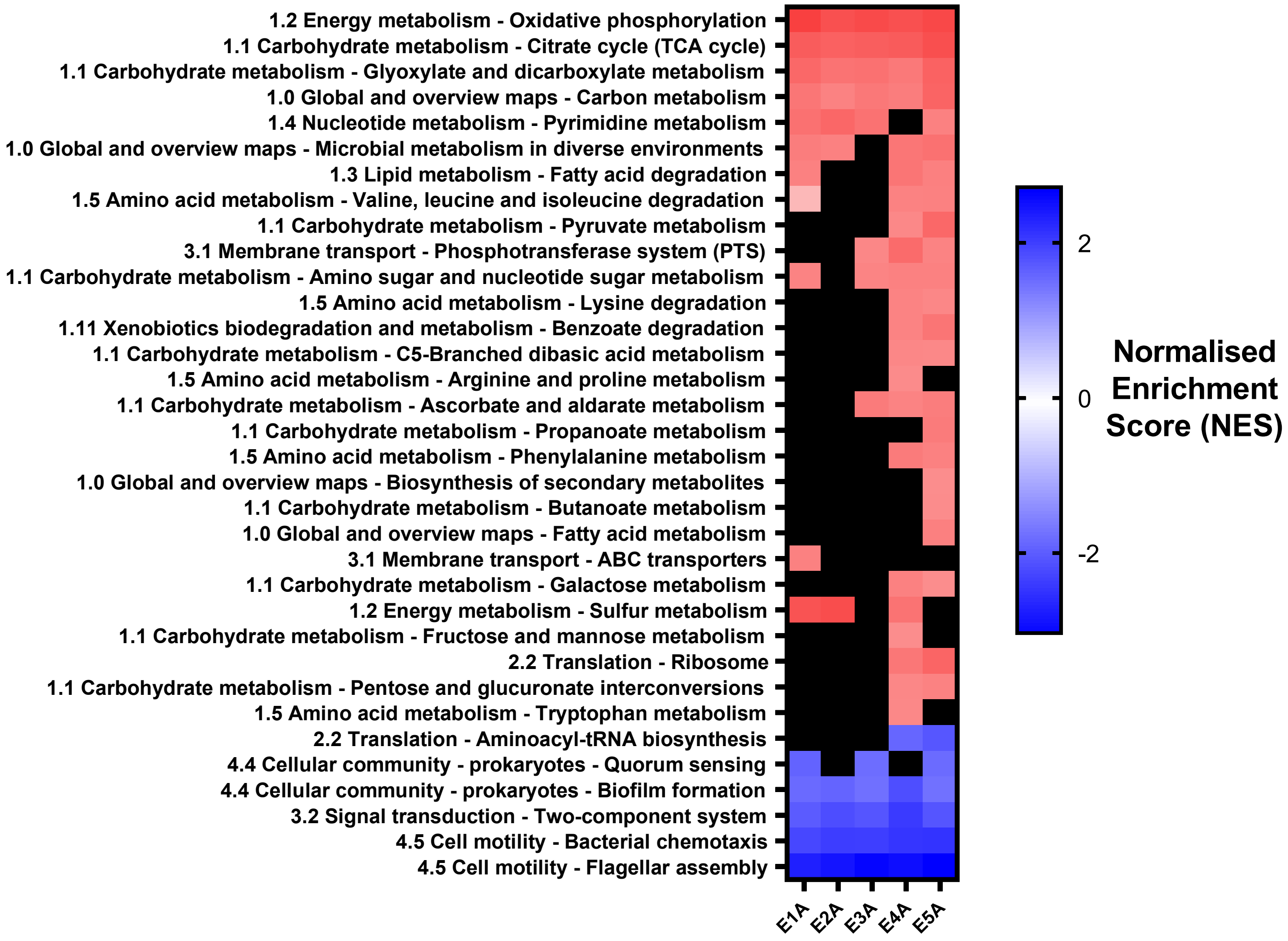


Figure 3.9.2: Heatmap illustrating the outcome of GSEA of differential expression data of the evolved strains against MG1655 ancestor using a gene set using the Kegg pathway database. Gene sets were ranked based on the average NES score across all 5 evolved strains. Gene sets predicted as significant using an FDR < 0.05, the normalised enrichment scores are presented to indicate upregulation (red) or downregulation (blue) of the gene set. Genest which were not significant in all evolved strains are not presented, and where not significant in one evolved strain is indicated by a black entry.

3.10 Summary

In summary this chapter continued the analysis of the evolution experiment described by Sen (2018), by further characterizing the evolved strains isolated. Below the results described in this chapter are summarized:

1) It was demonstrated that the conditions of the evolution experiment described by Sen (2018) was not constantly under low pH. Instead, due to the catabolism of

amino acids within unbuffered LB, the environmental pH varied from approximately pH 4.5 to pH 8.7, within each passage of the evolution experiment.

2) It presented analysis of the genotypes of the 5 evolved strains and evolved populations previously isolated and sequenced by Sen (2018) including the identification of structural genome variants using breseq.
3) It demonstrated that within the conditions of the evolution experiment mutations were repeatedly observed in the same genes within independent clones and populations. In particular, in all evolved strains mutations were observed in genes *arcA* and *cytR* and in 4/5 evolved strains the same *rpoA* mutation N294H was observed.
4) It tested different relative fitness metrics using experimental and hypothetical data for competition experiments. Selection rate was determined to be the optimal choice for reporting fitness within this study.
5) It was concluded that in direct competition, the evolved strains were fitter than the MG1655 ancestor under the conditions of the evolution experiment.
6) However, it was demonstrated that the evolved strains had did not adapt specifically to low pH. Instead, competition experiments performed under different experimental conditions revealed the evolved strains had adapted differently and to other conditions which occurred using unbuffered LB at pH 4.5. Examples of these include fitness at high pH, fitness against acetic acid stress, fitness specifically to the LB media, etc.
7) It was concluded that mutations within *arcA* did not cause loss of function. Competitions and ArcA activity measurements showed that the effect of these mutations is complex particular in different environments, indicating a gain of function of the *arcA* mutations overall.
8) It was demonstrated that RpoS activity did not alter within the evolved strains compared to the MG1655 ancestor
9) It was demonstrated that within the transcriptome of the evolved strains global effects were observed, indicating change associated with the regulons of the majority of transcription factors and upregulating of genes involved in central metabolism, in the evolved strains compared to the MG1655 ancestor.

Chapter 4

Construction and Analysis of a High Density Transposon Library in *E. coli* MG1655

4.1 Introduction

As stated above, the main purpose of this study was to consider if the outcome of a short term selection experiment using a transposon library as the starting point would be similar to that of the evolution experiment described in Chapter 3. If this indeed turns out to be the case, it means that the outcome of evolution experiments could be reported in dramatically shorter time compared to a conventional evolution experiment. In order to perform this experiment, a transposon library needed to be constructed using the MG1655 strain that was the "ancestor strain" (i.e., the starting strain) in the evolution experiment. Therefore, the main aim of this chapter is to describe the construction and subsequent analysis of such a high density *E. coli* MG1655 transposon library.

Upon construction of a transposon library, TraDIS can then be applied to map every unique insertion and its relative frequency within the population. Once this has been conducted, an analysis which is then typically performed is to identify essential genes within the genome (Langridge *et al*., 2009, DeJesus *et al*., 2017, Goodall *et al*., 2018, van Opijnen *et al*., 2009). As mentioned above, an essential gene can be described as a gene whose function is required for the organism to survive and grow. Identification of essential genes can be useful in understanding an organism's requirements for viability. Therefore, within this chapter the identification of essential genes within MG1655 will be described.

In addition to this, within these chapter two comparisons of essential gene lists will be described. One will be a comparison with the *E*. *coli* K-12 BW25113 transposon library, to ask if there are any differences in essential genes between two E. coli K-12 strains, both of which are widely used in the lab. The other will provide a brief overview of a current ongoing project conducted by Fatima Alattar, a project which I have assisted on. This was to consider the differences in the "essentialome" observed between *E coli* strains of different origin, both commensal and pathogenic. The results of this comparison are described below.

4.2 Construction of a Transposon Library in MG1655

To construct a transposon library in the *E. coli* MG1655 strain described in chapter three, a mini-Tn5 transposon was used, as Tn5-based transposons have no specific insertion site preference, therefore in theory would be able to insert anywhere into the genome where they do not disrupt an essential function. Specifically, EZ-Tn5^{TM} <KAN-2> Tnp is a transposome, which is a complex containing purified EZ -Tn5 transposon with Tn5 transposase bound to the transposon inverted terminal repeats (ITR), which can be directly transformed by electroporation into competent cells (Goryshin *et al*., 2000). The transposon itself consists of a modified version of a Tn5 transposon, containing only a kanamycin resistance gene, flanked by 19 bp ITR which are recognized by the Tn5 transposase. Once transformed into the cell, transposition of the transposon can then occur into a random location of the genome. Since the transposon has no Tn5 transposase gene, no additional Tn5 transposase can be produced after insertion. Therefore, once the pre-existing Tn5 transposase is degraded within the cell, this prevents additional transposition events from occurring, which means that upon creation of a transposon library the vast majority of cells within the library will only contain one transposition event.

The overall protocol that was used to create the transposon library was given in Section 2.7.1. In outline, construction of a transposon library involved making *E. coli* MG1655 ultra-competent and then several transformations were performed using the EZ-Tn5^{TM} <KAN-2> Tnp. Once transformed the culture was then recovered and plated on 30 µg/ml kanamycin LBA plates, before being pooled. Following creation of the library, it was then sequenced, and the transposon locations mapped to a genome using a TraDIS procedure described in Section 2.7. To ensure that the results were reproducible, the MG1655 library was sequenced in duplicate with each sample processed using the same protocol as described in Section 2.7. Of the two samples sequenced, reads which were associated with a transposon junction were then separated and aligned to *E.coli* MG1655 genome. For Sample 1 and Sample 2, respectively 3503736 and 2507491 reads were aligned to the genome which corresponded to 465013 and 411444 unique insertions sites, respectively. In order to compare the samples, two metrics were generated for every gene on the genome using custom scripts. These were insertion index score, (amount of unique insertions per gene normalized for gene length) and RPKMs (amount of reads per gene, normalized by gene length, and overall read depth). The results, presented in Figure 4.2.1, show that there was high correlation between the two samples indicating high technical reproducibility, for both RPKMs and Insertion index, with both having a Pearson’s correlation coefficient greater than 0.98. Since high correlation was observed, the reads of the two samples were then combined and the custom scripts rerun to give a data set of 6011223 reads which were confirmed to have 574250 unique insertion sites throughout the genome. Overall, this amount of unique insertions approximated to an insertion every 8.85 bps. This combined data set and individual samples were concurrently used in the analysis below.

.

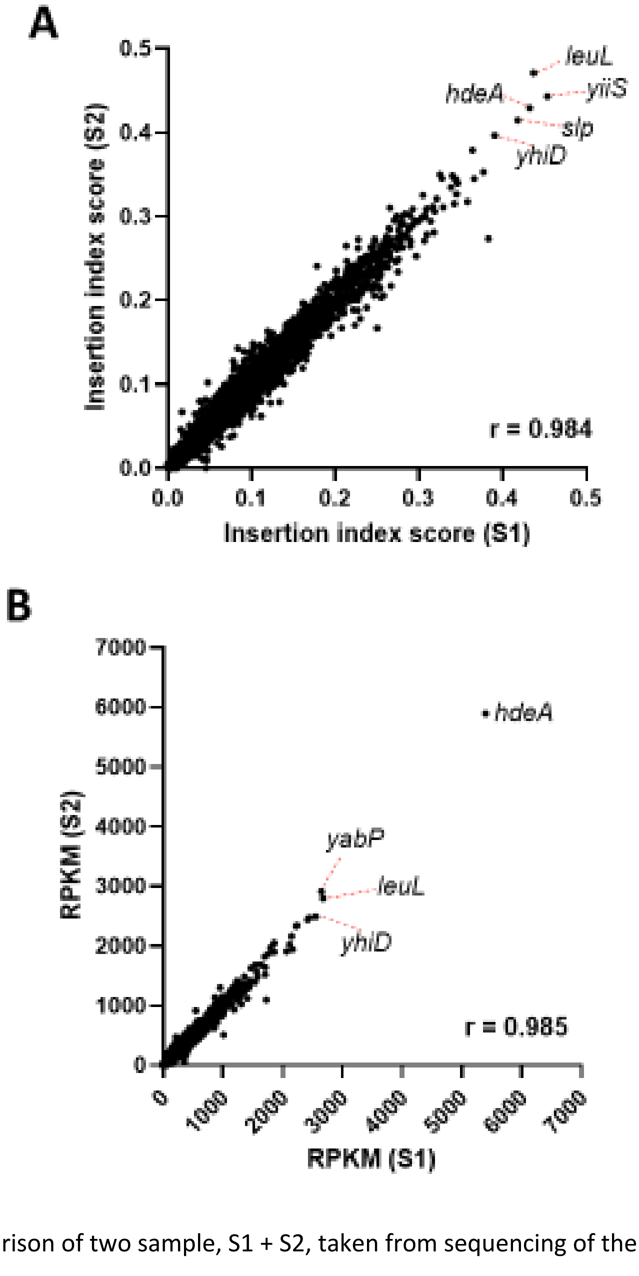


Figure 4.2.1: Comparison of two sample, S1 + S2, taken from sequencing of the MG1655 transposon library, for either insertion index scores (A) or RPKMs (B). To show similarity the Pearson's correlation coefficient was generated for each metric and is shown in the bottom right of each graph.

4.3 Distribution of insertions within the MG1655 Transposon library

Interestingly upon comparison of the two samples, an enrichment of reads and insertions were observed in individual genes that are highlighted within Figure 4.2.1. These enrichments suggested that there may be preferential sites of insertions within the genome in these genes, or alternatively this could indicate enhanced biological fitness of strains containing these insertions under the selection which occurs during the construction of the library. Considering the function of these genes observed, three genes were predicted as genes of unknown function (*yiiS, yhiD* and *yabP*), while *hdeA* is an acid stress chaperone, *leuL* is the leader peptide for the leucine biosynthesis operon, as well as the outer membrane protein, *slp*. However, considering the function of these genes, no straightforward biological reason as to why these genes would be enriched was found.

As mentioned in section 1.4.1, the Tn5 transposon has been reported to have an insertion bias toward GC rich regions (Green *et al*., 2012). Considering this, we sought to look at whether there was a higher GC content in these genes identified in this study. Since *E. coli* K-12 MG1655 has been reported to have an average GC content of 50.8% (Milo et al., 2010) we subsequently calculated the GC content of each of these genes (GC content: *yiiS* = 46.3%, *yhiD* = 45.5%, *yabP* = 45.2%, *hdeA* = 42.9%, *leuL* = 47.1%, *slp* = 45.9%) These results show that the GC content of all genes identified was lower than the average GC content of *E. coli* K-12 MG1655, suggesting GC content was not a factor as to why more insertions in these genes was observed.

As shown previously, an uneven distribution in read and insertion count was also observed in other transposon libraries (Langridge *et al*., 2009, Goodall *et al*., 2018, Ruiz *et al*., 2017). This uneven distribution is associated with genome replication, since replication begins at the origin of replication (*ori*C) in a bidirectional manner, so there is a likelihood that more copies of region close to *oriC* will be present than those further away. Therefore, when a transposon library was sequenced distribution of insertions and reads were shown to be unevenly distributed towards the origin of replication, and this was proposed to be due to more copies of that region of the chromosome being present when the library was constructed. This uneven distribution was also observed within our MG1655 transposon library (Figure 4.3.1A).

Inspection of insertions on the genome showed that there was no particular region of the MG1655 genome lacking in insertions, although a decline in insertions and read count was observed (Figure 4.3.1). In order to investigate this further, the read count metrics for each gene were plotted based on their position in the genome (Figure 4.3.1 B+C). This result shows the classic trough and peak, with the peak centred around the origin of replication, as seen in other studies. In addition, of the 5 genes which showed high in insertion count and RKPMs, 4 of these were not further than 250kbp from the origin of replication. The exception to this was *leuL* which was further away from the oriC (~750kbp). Further to this three of these genes *leuL*, *hdeA* and slp, high insertion index scores were also seen within other libraries constructed within the laboratory, including a *E. coli* BW25113 kanamycin transposon library and a *P. aeruginosa* BX25 (Discussion with members of Ian Henderson's Lab). Therefore, this result suggests that the bias seen within these genes is in part potentially due to their location within the genome being close to origin of replication. However, it should be mentioned that not all genes present around *oriC* were overrepresented, therefore potentially a bias in insertions does exist within these genes.

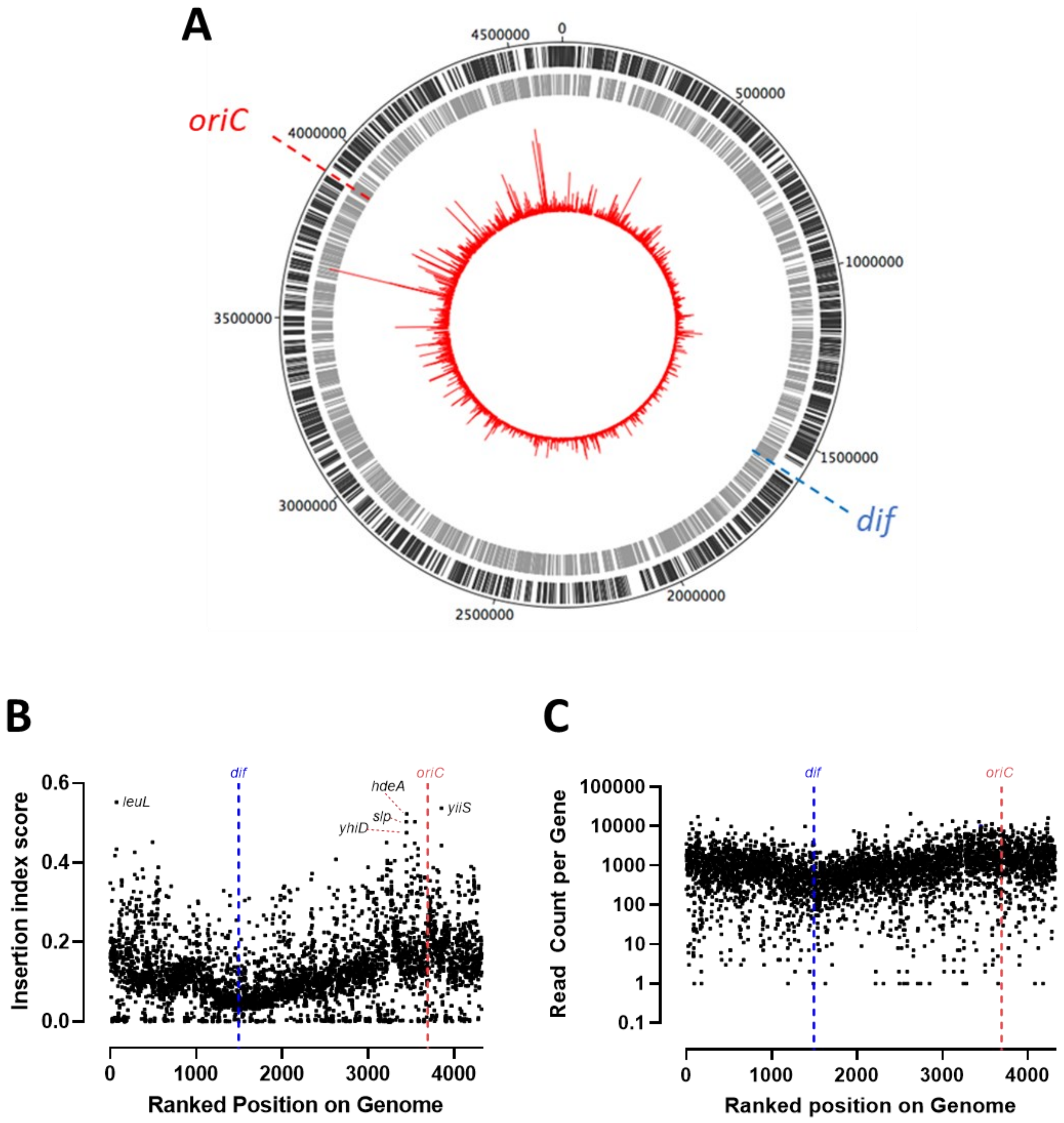


Figure 4.3.1: The distribution of Tn5 insertion positions observed within the MG1655 transposon library. The origin of replication and terminator region are highlighted by the *oriC* (red line) and *dif* (blue line) sites, respectively. A) Genome map of MG1655 transposon library. The black and grey outer circles illustrate CDS annotations for forward and reverse strands respectively. The red inner circle represents the relative number of reads corresponding to transposon insertions, for that region. B) Distribution of insertion index score per gene, ordered by the position on the genome. Genes observed in Figure 4.2.1 were highlighted. C) Distribution of read counts per gene, ranked by position on genome. Note Due to the logarithmic scale 272 genes were not plotted due to having a 0 read count.

4.4 Defining the essential genome of *E. coli* MG1655

As mentioned above, one of the most common purposes for sequencing a transposon library, is to identify essential genes present within the genome. Considering this, an essential gene can be defined as a gene whose function is required by the organism to survive and proliferate. To identify essentiality using TraDIS data involves the process of identifying regions which are significantly absent of insertions (Figure 4.4.1 A). Absence of insertions suggests that the gene (or location) is essential. The reasoning behind this is that if a transposition event occurred within an essential gene, the transposon would disrupt the coding sequence of the genes which would completely or partially prevent function of the gene. Disruption of normal gene expression could have a similar effect. Therefore, if the function is disrupted, since the gene is essential, the organism will not survive/proliferate and therefore will drop out of the population and not be sequenced.

4.4.1 Different methods used to identify essential genes

To identify essential genes using TraDIS data several methods exist. One such method would be to manually inspect the TraDIS data using a genome browser (Figure 4.4.1). However manual inspection, although useful, will take time, and the potential to miss something due to human error is always a problem. Alternatively, within the literature several computational methods have been developed to identify essential genes. However, these also can have some difficulties in identifying essential genes, as these methods attempt to quantify the absence of insertions within a gene, but within a transposon library insertion patterns in essential genes may be complex. For example, strains carrying insertions at the 3’ end of certain ”essential” genes are able to survive, as seen in Figure 4.4.1A. In addition to this, other insertion patterns can be observed within gene where only part of the gene is considered essential. The analysis performed by Goodall *et al*. (2018) demonstrates this in more detail. An example of this is shown in Figure 4.4.1B where the gene *rne* would be classed as essential, due to the clear absence of insertions, however this refers to only part of the gene. Therefore, in many cases, automatic analysis of these genes would consider this non-essential. Since several different computational analysis methods exist to determine gene essentiality, this study compared four different methods with the results as described below.

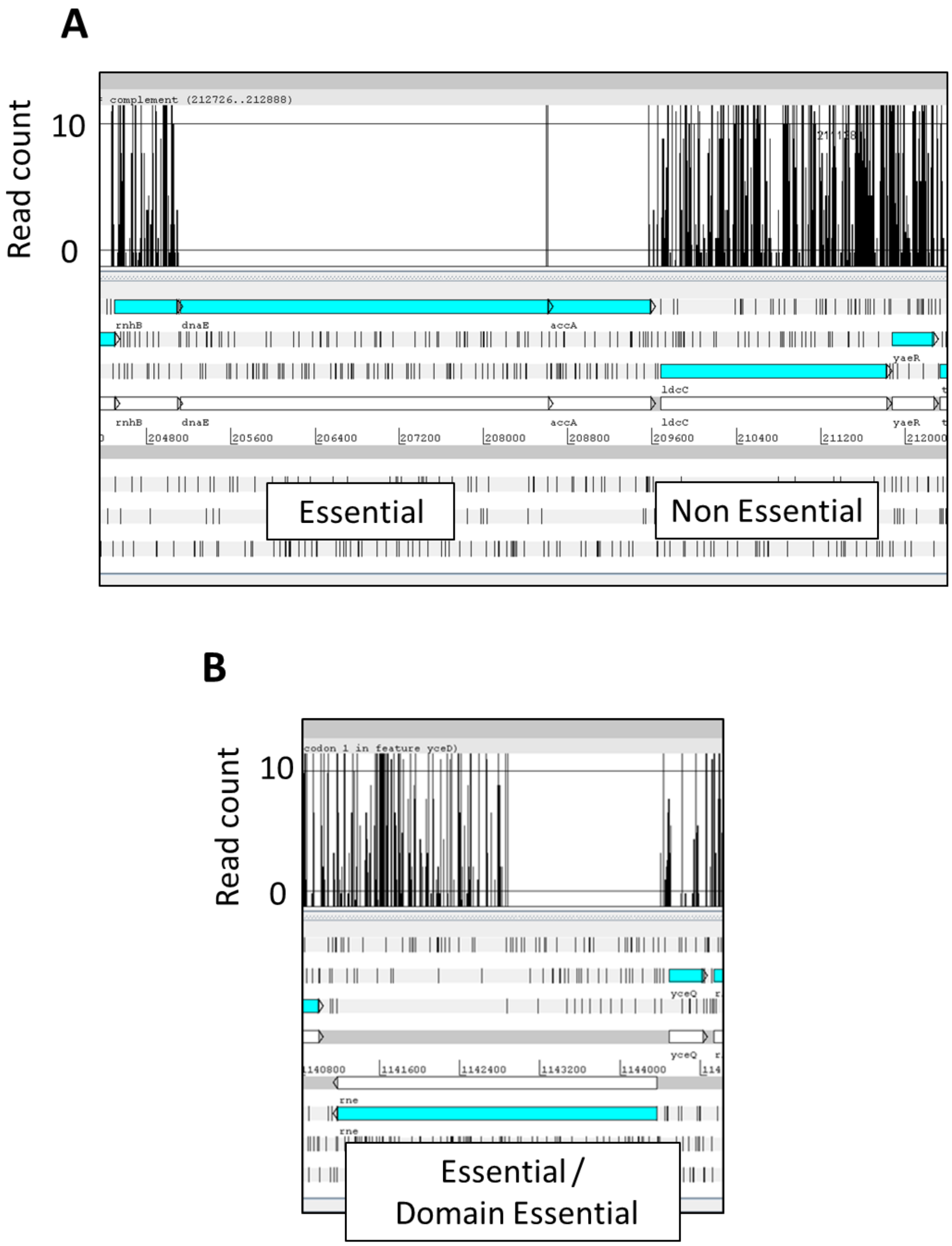


Figure 4.4.1:Visualization of MG1655 TraDIS data using the Artemis genome browser, to show examples of (A) Essential and non-essential genes , and (B) where only a domain is observed as essential within a gene. Each line refers to an individual transposon insertion, with the length of line inferring the read count. The scale was set to a max of 10, to look at only presence and absence of insertions.

4.4.1.1 Log$_2$ likelihood scores to predict essentiality

This method follows a protocol described in Goodall *et al*. (2018) with the scripts used to perform this analysis gifted by Dr Emily Goodall. This method uses a log likelihood estimation to identify essential or nonessential genes. To do this the insertion index score is used. Typically in a normal transposon library, when a histogram is created of the insertion index score a bimodal distribution is observed (Figure 4.4.2). It is in these two modes where essential genes (lacking insertions, low insertion index score) and non-essential genes (insertions present, high insertion index score) can be distinguished. Using the R package, R MASS library (version: 7.3-51.4) a gamma probability distribution can be fitted to each mode, using a user dependent cutoff value to define the modes. Essential genes can be defined by fitting an exponential distribution, a special case of a gamma distribution, to the smaller mode and non-essential genes can be defined using a gamma distribution for the larger mode. Essential and non-essential genes can then be determined by calculating a log$_2$ likelihood ratio for each gene between the two distributions. Genes could then be split into three groups, defined as (1) essential if the log$_2$ likelihood ratio was less than -3.6, indicating that they were 12 times more likely to be in the essential mode than the non-essential mode; (2) Non-essential if the log$_2$ likelihood ratio greater than 3.6; and (3) ambiguous if the log$_2$ likelihood ratio did not hit the threshold of either other groups.

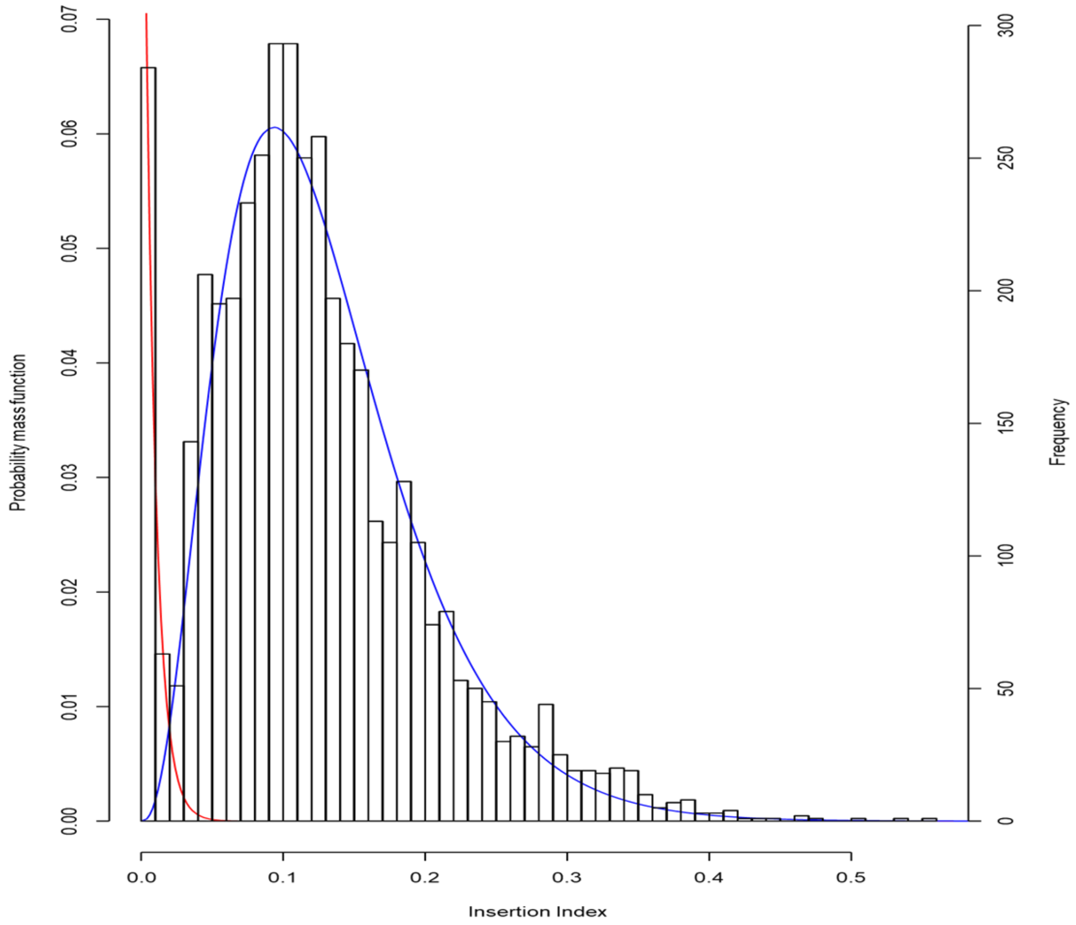


Figure 4.4.2: Bimodal distribution seen within a Transposon library. Coloured lines represent two gamma distribution models fitted to each modal data. The red line is an exponential distribution model fitted to the essential gene data. The blue line is a gamma distribution model fitted to the non-essential gene data.

4.4.1.2 Biotradis

Biotradis is a pipeline designed for the easy analysis of TraDIS data, in which pipeline is provided with raw read data and an annotated genome. More information about this protocol can be found within the Method section 2.8.2.7 (Barquist *et al*., 2016). Overall following this analysis, a total of 563490 insertions were identified. Once the reads were aligned to the genome and the insertion index score obtained for each gene, gene essentiality was then predicted using a method similar to described above in Section 4.4.1.1. In addition to this Biotradis uses a lower threshold for determining essential genes using a log likelihood cutoff of -2 and 2 to define essential, nonessential, and ambiguous genes.

4.4.1.3 El-Artist

ARTIST is a MATLAB based package which provides two tools, EL- ARTIST and Con – ARTIST. It was designed for Tn-seq data, however it can also be used with TraDIS data (Pritchard *et al.*, 2014). EL-Artist focuses upon identifying essential genes defined by a transposon library and its methodology will be described in this section. Con-artist focuses upon identifying differences in insertions in a transposon library, before and after growth in specific conditions; the subsequent analyses will be described in Chapter 5, below.

EL- ARTIST looks at transposon data in an annotation independent manner and is able to identify essential genes but also goes further than this by identifying genes which contain an essential domain as seen in Figure 4.4.1B. Of the four analyses described here, identifying, and classing essential domain genes is something which is unique to the ARTIST method. It does this by firstly identifying all potential insertion sites upon the genome for the transposon library. In order for ARTIST to properly work, EL-ARTIST requires a high saturation of transposons at these insertion sites. Since ARTIST focuses on Tn-seq which utilizes the mariner transposon and therefore inserts at TA sites, this is simple. However, in the case of using TraDIS data, which in this study uses the Tn5 transposon, insertion sites can be anywhere along the genome, therefore a saturation of insertion sites will be low if you consider every position in the entire genome as a potential insertion site. Therefore, in the analysis within this study the genome was split into 50 bp "transposon windows" with all transposon count data present in the window averaged together in the middle base of the window. The read count associated with each 50bp window was then normalized to correct for the read count bias seen towards the origin of replication. This normalization was done in local 100kb windows across the genome, generating a scaling factor by taking the average of read counts seen in each window and comparing it to the overall average across the genome. This scaling factor was then multiplied with the value for each 100kb local window to normalize the data.

Essential regions were then defined by using a sliding window analysis. A null distribution was generated corresponding to the mean of read counts for 1000 independent randomly chosen sliding windows. This sliding window analysis was then performed taking a sliding window for every 10bp in the genome. This then allowed for the generation of a p-value by ranking the mean of the window against the null distribution and dividing by the amount of independent simulations used to generate the null distribution. An FDR was then applied using a Benjamini-Hochberg procedure. Under-represented windows could then be identified using a cut off value of 0.05 for the FDR.

The results of the analysis above was then used to train a Hidden Markov model (HMM). This process allowed for refinement of Transposon windows to be classed as essential or non-essential. Once essential or non-essential transposon windows were classified the genome annotation was then overlaid on top of this data. This created a simpler view in which each annotation and intergenic region was either classed as essential, non-essential or had a domain present within the annotation which is essential. It was this simpler view was then used for further comparative analysis.

4.4.1.4 ESSENTIALS

ESSENTIALS is an open source TraDIS analysis pipeline which was web based (Zomer *et al.*, 2012). However, since early 2019 , the website was not actively maintained and since 2020 is unavailable. This pipeline takes raw reads data and aligns them to the genome before performing comparative analysis of different conditions which will be described below in chapter 5. In order to identify essential genes this pipeline generates an expected number of reads per gene using the amount of insertion sites per gene, the total amount of unique insertions presents in a genome, and the sequencing depth. It then performs an analysis similar to a differential expression analysis using the software, edgeR, which was developed to analyse RNAseq data, comparing the expected read counts with actual measured read counts. Essential genes are then be identified as any gene which showed a decline in read count using a predefined value as a cutoff.

4.4.2 Comparison of different pipelines.

4.4.2.1 User input into the different pipelines

Upon performing the different analyses described above, the first thing which was noticed was the amount of user input required in generating lists of essential genes. The analysis pipeline of Biotradis generated a list with no user input at all and only little input was required using the log likelihood method to define the cutoff value. ARTIST and ESSENTIALS had a more hand on approach requiring the user to alter parameters such as p values and sliding window scales to suit the user's requirement and data. Therefore, to assess these two analyses properly for ESSENTIALS and ARTIST, the analysis was performed multiple times altering the size of the sliding window, ranging from 5 – 9, 50bp windows (ARTIST only) and adjusting the pvalue required to determine essential genes or underrepresented windows (ESSENTIALs and ARTIST, respectively). Ultimately choosing variables which would generate a list of genes that would be classed as essential within a range of 300 – 400 genes. This range was chosen as the amount of essential genes identified within *E. coli* K-12 has been shown to vary between different studies ranging from 303 – 620 genes (Martínez-Carranza *et al*., 2018). However, one particular a study conducted by Goodall *et al.,* (2018) used TraDIS to identify 358 essential genes in *E. coli* K-12 BW25113. Therefore, using the Goodall *et al.,* (2018) as a rough estimate of the number of essential genes which may be present in *E. coli* MG1655, an overall range of 300-400 genes was defined, to consider reasonable variability within the different methods and pipelines.

4.4.2.2 Reporting on the results generated from each pipeline

In the analysis pipelines described above, with the exception of ESSENTIALS, each data analysis did not split genes clearly between essential and non-essential categories, therefore another pool of genes was also created by these analyses. As mentioned above, in the case of Biotradis and log likelihood, genes in this set were classed as "ambiguous", and this covered genes which these two methods were unable to class as essential or non-essential based on the log likelihood cut off values. In the case of Artist, a second pool of essential genes was determined which were termed 'domain essential', i.e. these were genes for which the sliding window analysis had confirmed the presence of an essential domain in the gene but the gene also had at least one non-essential domain. In addition to this, Artist was also able to identify intergenic regions which could be classed as essential; the other pipelines could not do this. This pipeline therefore adds another feature which may be useful in determine genome essentiality. Further inspection of this revealed that, with the exception of 13 intergenic regions, the majority intergenic regions classed as essential these were related to a defined essential gene, suggesting that the promoter region of an essential gene is also likely to be essential, as would be predicted. Of the 13 intergenic regions were classed as essential which were not related to an essential gene. A manual inspection revealed that 8 of these regions had no insertions present, of these, two were associated with domain essential genes, one corresponded to a tRNA being present, *asnU* and the remaining 5 were not associated with any other neighbouring feature. In addition, of the 5 intergenic regions also classed as essential still had insertions present when manually visualized, although at a lower frequency, therefore these intergenic regions would not be actually considered essential.

4.4.2.3 To cutoff or not to cutoff that is the question

Within two analysis pipelines, Biotradis and ESSENTIALS, an option existed to not include 10% of the 3' end of each gene. As mentioned above, it is common to observe insertions within the 3' end or 5' of essential genes and within a computational pipeline these insertions could actually affect the classification of a gene's essentiality. Therefore, by ignoring insertions in the 3' end it removes this problem, and in theory allows better determination of essentiality. The comparative analysis described below does not use this feature, as we wanted to see the effects that this option may have upon defining gene essentiality. Since ESSENTIALs is no longer being maintained, below I have described the use of the Biotradis pipeline to perform this comparison.

Results shown in Figure 4.4.3 show a high similarity in the essential genes identified irrespective of whether or not the 3' end was removed. However, a set of genes were also unique to each analysis. A manual inspection of these genes showed that in the case of not having the 3' end removed, the genes still had insertions present. This indicated that these genes probably were not essential in the first place, they just had a lower insertion density than the average causing it to be classed as essential (e.g. in Figure 4.4.3: *cmk)* and while some of these genes were close to the terminus, not all were. In some cases, this was also true in essential genes when the analysis did not include the 3' end, as some low density genes were classed as essential (e.g. in Figure 4.4.3: *tktA*). However in other cases, it was clear that the gene was essential and removing the 3' end allowed for the gene to be classed as essential (Figure 4.4.3: *ribB*). Overall, the results of this analysis show that in the generation of the essential gene list, removal of the 3' end of a gene does not affect the majority of genes classed as essential. However, for a small number of genes removal of the 3' end can allow for the inclusion of certain genes which would be considered as essential, and removal of other genes that would not be considered. Even with this analysis some false positives would still be present.

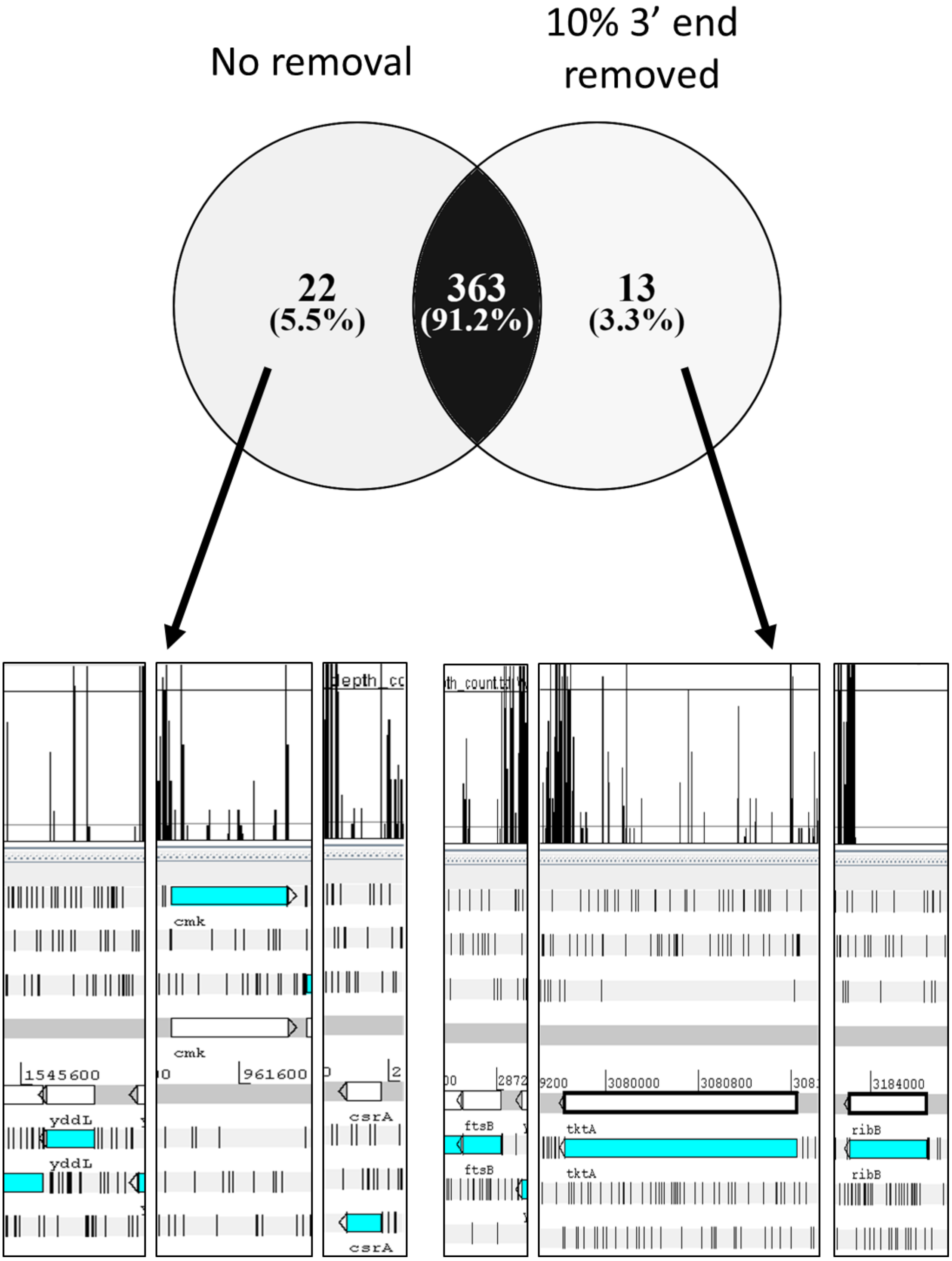
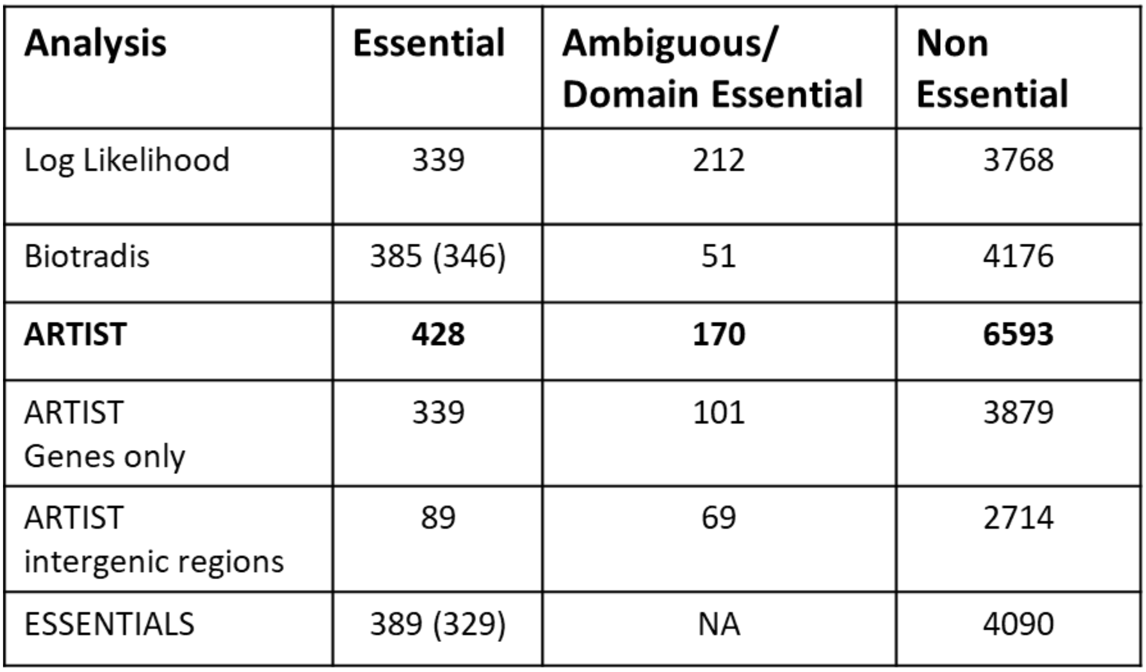


Figure 4.4.3: Comparison of the essential genes reported using Biotradis, reporting insertions throughout all the gene or when 10% of the 3' end removed. Genes which differed were inspected using the Artemis visualization browser, with examples as provided.

4.4.2.4 Differences in the annotations reported as essential within each pipeline.

Interestingly comparing the number of genes included in the analysis upon initial inspection, when using the Biotradis pipeline 4612 genes were reported, compared to 4319 genes reported within the other analysis methods (Table 4.1). Further inspection of this discrepancy revealed that within the Biotradis analysis, any feature which was annotated was included within this analysis, while the other analysis focused only on features labelled as 'CDS' within the annotation folder. Therefore, a "perk" of the Biotradis analysis was that essentiality was reported on genome features in addition to genes encoding proteins. Of the 389 genes reported as essential, 39 genes which were reported as essential corresponded to features not included within the other analysis, these included features such as transfer RNAs and small regulatory RNAs. The same was true of the ESSENTIALS analysis, although in addition to this, in the case of genes which had the exact same annotation, which on the MG1655 genome was only true of genes present within IS elements (such as *insA* and *insB*), all duplicates were classed as essential. This suggests the way ESSENTIALS deals with duplicate annotations is to pool all reads into one entry. Therefore, on this basis, duplicated genes were removed from the essential gene list.

Table 4.1: Summary of outputs generated from each analysis. In the case of Log likelihood and Biotradis, the middle column refers to genes which were ambiguous, i.e. cannot be classed as essential or non-essential. In the case of Artist, the number refers to domain essential, i.e. genes which were determined to have a domain which is essential. In addition, Biotradis and ESSENTIALs reported essentiality on other features, the essential genes reported in brackets include annotations which are only present in the other analysis.



Analysis	**Essential**	**Ambiguous/ Domain Essential**	**Non Essential**
Log Likelihood	339	212	3768
Biotradis	385 (346)	51	4176
ARTIST	**428**	**170**	**6593**
ARTIST Genes only	339	101	3879
ARTIST intergenic regions	89	69	2714
ESSENTIALS	389 (329)	NA	4090

4.4.2.5 Comparison of essential genes from different pipelines.

After generating a list of essential genes from each of the pipelines, these gene lists were then compared. The results of this comparison are summarized in Figure 4.4.4A. As would be expected, there was a large proportion of genes identified as essential within all pipelines. Interestingly in only one pipeline, the log likelihood method, all genes identified as essential in this method were also identified within at least one other pipeline. However, considering that the Biotradis and log likelihood pipelines are highly similar methods to identify essential genes, it would also be expected they would correlate well and indeed they were shown to have the largest overlap with 338 genes being identified as essential by both methods. Since these methods also generate a list of ambiguous genes for visual inspection, these were analysed, and it was found that the genes which didn’t match in the essentials lists were present within the respective ambiguous gene lists. This indicates that the log likelihood threshold difference probably causes the inclusion or exclusion of these genes.

Interestingly, considering the number of ambiguous genes which were described within these two analyses shows that the lower thresholding within the Biotradis causes more genes to be binned into either essential or non-essential groups, than the ambiguous. Therefore, suggests that due to Biotradis lower thresholding, the inclusion of more false positives occurs within this analysis. Indeed, the comparison in Figure 4.4.4A reveals of the 4 genes that were only specific to the Biotradis pipeline, namely *tusE, flhE, fdx* and *ilvX*. Manual inspection of these genes using the Artemis browser indicated that these genes are probably non-essential.

Considering genes which were unique to each pipeline, ESSENTALS showed the largest difference with 37 genes not identified in any other pipeline. As mentioned above this, could be due to user bias being introduced when the cut off is selected. Further inspection of the data revealed known non-essential genes being classed as significantly essential such as *arcA* and *aceE*, as loss of function of these genes have been identified within this study and others (Section 3.7) (Nizam *et al*., 2009, Baba *et al*., 2006). This suggests that use of ESSENTIALs to identify essential genes may not be as robust as the other methods described within this comparison.

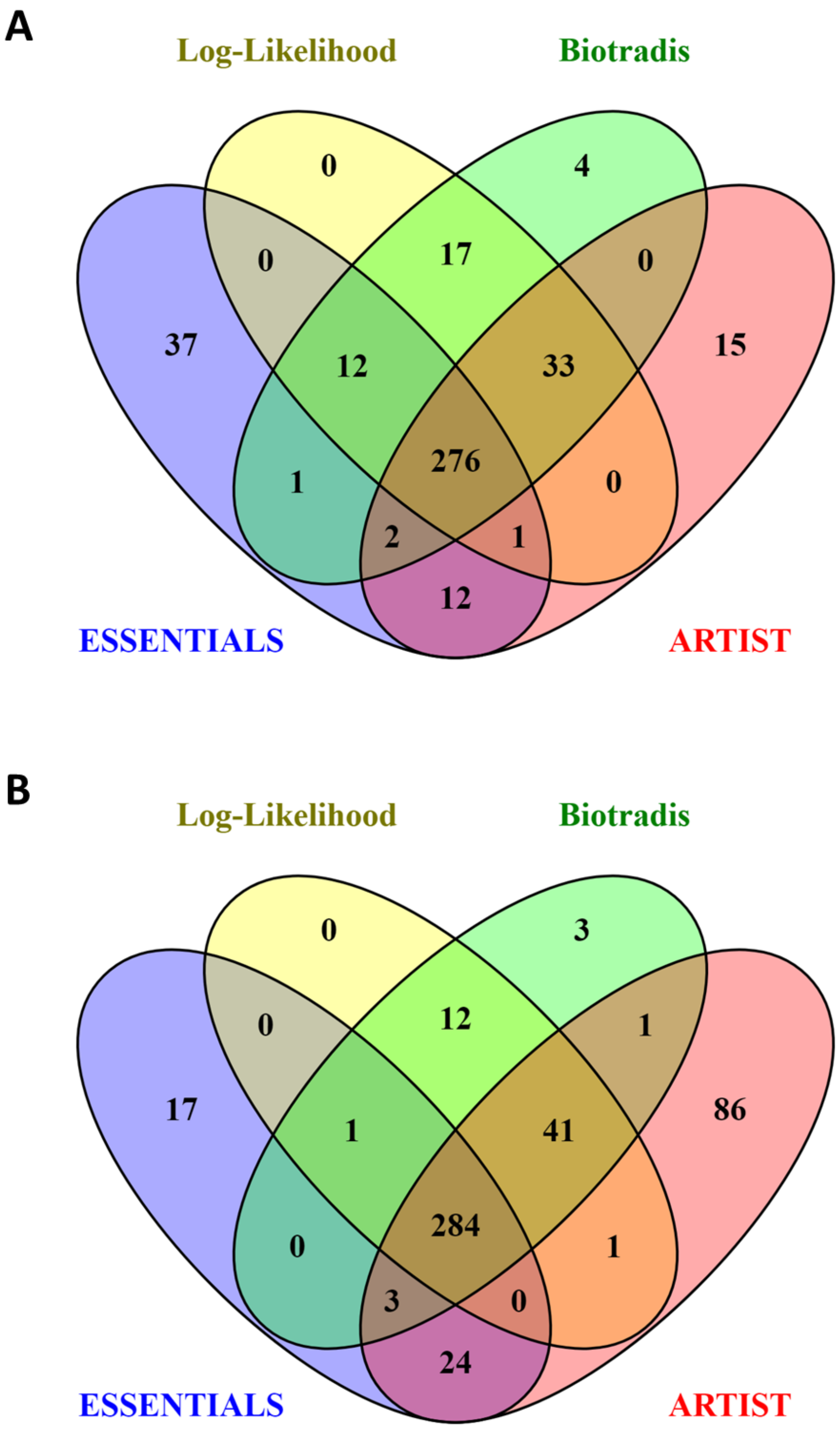


Figure 4.4.4: Four-way Venn diagram detailing the comparison of essential genes predicted by 4 pipelines, ESSENTIALS, Log likelihood, Biotradis and ARTIST. A) Details the comparison considering only genes to be predicted as essential B) includes Artemis prediction for domain essential genes also.

4.4.2.6 Considering domain essential genes identified with Artist.

ARTIST also identified a list of what it termed ‘domain essential’ genes. This describes a gene where only part of the gene, a domain, is classed as essential by the absence of insertions, however the remaining part of the gene, insertions can be present. Within the other pipelines described, identification of these genes would be difficult as insertions are still present in them. However, in theory these genes would be classed as essential, as if the overall function was removed the cell would be unable to survive/ grow. Therefore, a separate comparison was performed including these genes as essential within ARTIST, the results of which is presented in Figure 4.4.4B. From this comparison it can be seen that the overall core set of essential genes increased when the ARTIST ‘domain essential’ genes were included, showing that within the Artist analysis, some genes which are classed as domain essential are identified using the other methods. However, a substantial proportion of the “domain essential” genes were also not found in any other analysis pipeline. Therefore, to further investigate this, a manual inspection of some of genes classed as domain essential was performed. In Figure 4.4.5 A+B, genes which were classed as essential within other pipelines and classified by ARTIST as ‘domain essential’, no insertions were present in the majority of the gene, only at the 3’ and 5’ end. Interestingly, of the genes classed as “domain essential” only by Artist, in certain cases on visual inspection genes were identified which actually did have an essential domain (Figure 4.4.5 C). However, in other cases after a manual inspection it was still difficult to clarify whether or not an essential domain was present (Figure 4.4.5 D+E).

Overall inspection of these results suggested that ARTIST is capable of identifying essential genes within TraDIS data and is also capable of identifying some domain essential regions. However, “domain essential” regions identified by Artist do also require a manual inspection of the transposon data to confirm or not confirm the presence the potential presence of an essential domain. For actual confirmation additional experimentation is needed. It must be stated that the analysis performed in this study may not be the most efficient possible when using the Artist pipeline, as due to computational power constraints the resolution of essential domains was limited. It may be that a better analysis could be performed with more computational power.

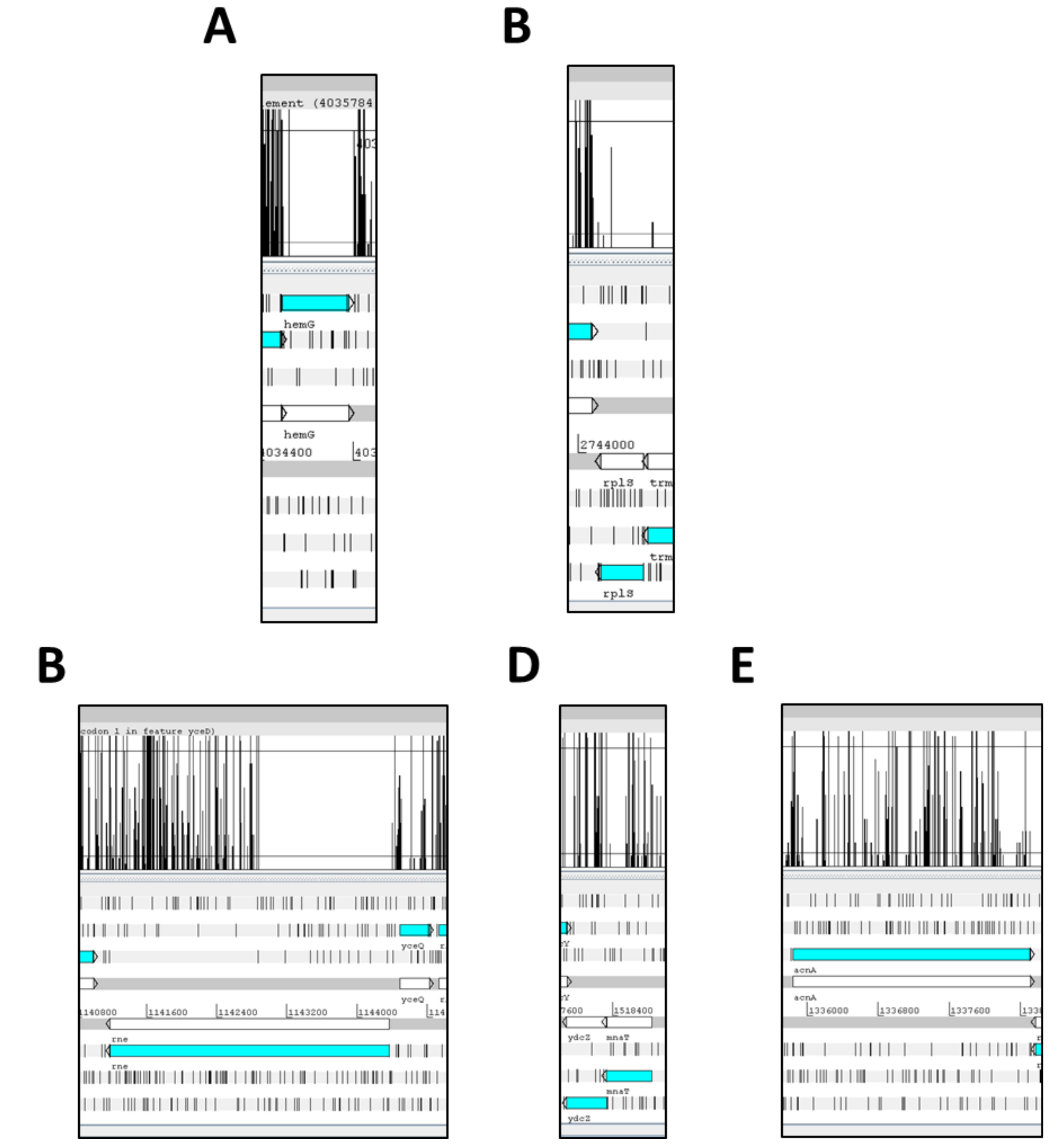


Figure 4.4.5: Visualization of transposon data in the genome browser Artemis. The genes present in this figure are genes which are classed as “domain essential” by ARTIST. A+B are genes which are also classed as essential in the other 3 pipelines (Biotradis, ESSENTIALS and Log likelihood). C-E are examples of genes which are predicted to be domain essential by ARTIST and not by any other software.

4.4.3 Conclusion: use of different pipelines

In the comparison above, it was probably inevitable that within the gene list produced by each analysis, false positives and negatives would occur. However, in this analysis, a core set of essential genes consisting of 276 genes, was identified by four different pipelines. Considering that different methods and metrics were used to identify these same genes, this increases the likelihood of these being essential.

Upon focusing on the effectiveness of each individual pipeline, it was determined that the ESSENTIALs pipeline ability to predict essential genes is not as effective as the other methods. This is probably due to its approach in comparing a gene's total read counts to an expected read count, by performing a type of differential expression analysis. Therefore, this pipeline is only predicting a reduction in read count from an expected value, which although will include essential genes, can also include genes which still have insertions, just at a lower amount than the average, and these genes may therefore be wrongly classified as essential or non-essential. In addition, it should be stated that this observation is not just restricted to this study, with other studies reporting similar issues when using ESSENTIALS to identify essential genes (DeJesus *et al*., 2013, Solaimanpour *et al*., 2015)

The ARTIST pipeline performed an annotation independent analysis, which allowed the identification of essentiality within the intergenic regions within the genome. This is something that is not investigated by the other pipelines. In addition to this it provided a "domain essential" category which is able characterize genes with essential domains, that wouldn't be picked up by the other four pipelines. Although it will be highly useful to identify these regions as these genes will also be essential, particularly if the user wanted to identify the essentialome of an organism, it does require a large amount of computational power. In addition, further focus upon its "domain essential" bin would require a manual inspection of all genes to actually confirm domain essential genes. Therefore, although highly useful, for a quick comparison of multiple genomes, this method would not be the best

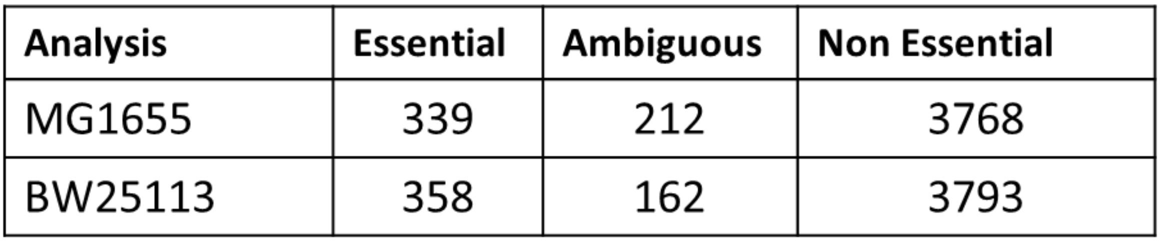
Biotradis and log likelihood were the most user friendly requiring the least user input. Both use the same method to determine gene essentiality generating a score based on the presence of insertions and not read counts, which is then able to focus more specifically on essential genes. The main difference between these two analyses was the stringency used to determine essential genes with the log likelihood method being more stringent, therefore it is more likely within Biotradis that false positives will be identified. However, it should be mentioned that in order to use these methods a certain density of transposon library is required to be able to accurately predict essentiality.

Overall, based on this analysis, no one method is the ‘best’ choice, instead each analysis is able to identify essential genes, although a number of false positives will be introduced. In addition to this each pipeline have individual features which may be useful to the user needs such as ARTIST in looking at intergenic regions. It should be mentioned that there are other methods to identify essential genes within the literature (Hubbard *et al*., 2019, McCoy *et al*., 2017, Page *et al*., 2020, Zhao *et al*., 2017). Therefore, the results of this comparison should only be considered within the context of this study.

4.5 Comparison of transposon libraries in *E. coli* MG1655 and *E. coli* BW25113

Goodall *et al*., (2018) previously described an in detail analysis of the essential genome as reported by an *E.coli* BW25113 transposon library. The features of this library was that it was of high density with 901,383 unique insertions points and was constructed using a Mini Tn5 transposon containing a chloramphenicol resistance cassette (Goodall *et al*., 2018). Since the *E. coli* strains BW25113 and MG1655 are both of K-12, it was decided to compare these two libraries to see if any differences existed between these two strains with regard to the essential genome. In addition, within this study, to identify essential genes, Goodall used the log likelihood method described in Section 4.4.1.1. Therefore, since the same method was used to predict essential genes within MG1655, a summary of this analysis performed in both strains is presented in Table 4.2. Overall, there were similar numbers of essential genes identified with both transposon libraries, however the number of ambiguous genes differed, with more present in the MG1655 library. This was probably due to the overall densities of the libraries with MG1655 being of lower density than BW25113.

Table 4.2: Summary of the genes reported as Essential, Ambiguous and Non-essential using the log likelihood method described in 4.4.1.1. The data presented of BW25113 transposon library was taken from (Goodall *et al*., 2018)



Analysis	Essential	Ambiguous	Non Essential
MG1655	339	212	3768
BW25113	358	162	3793

In Figure 4.5.1 the essential genes of the BW25113 and MG1655 were then compared. Although there was a large crossover in the genes predicted as essential, there were also some differences observed, with 10% and 5% of genes in BW25113 and MG1655 respectively not being classed as essential in the other strain.

However, within the analysis used to predict essential genes, genes which are unable to be classed as essential or nonessential are binned into a pool termed “ambiguous”. Several factors can affect why a gene would be classed as ambiguous, but overall, it is likely that insertions in these genes are at a lower density than insertions in the non-essential genes but not low enough for the genes to be classed as essential. It suggests the possibility that genes classed as essential in one strain are actually classed as ambiguous in the other strain. Therefore, the essential genes which were unique to the strain were then compared with the ambiguous genes of the other strain. The result of this comparison is presented in Table 4.3. Of the 41 and 28 genes essential only in BW25113 and MG1655, 10 and 5 genes remained which were not classified as ambiguous in the other strain. These therefore were genes which were classed as essential in one strain and nonessential in the other, suggesting for these genes the function was required in one library and was not in the other.

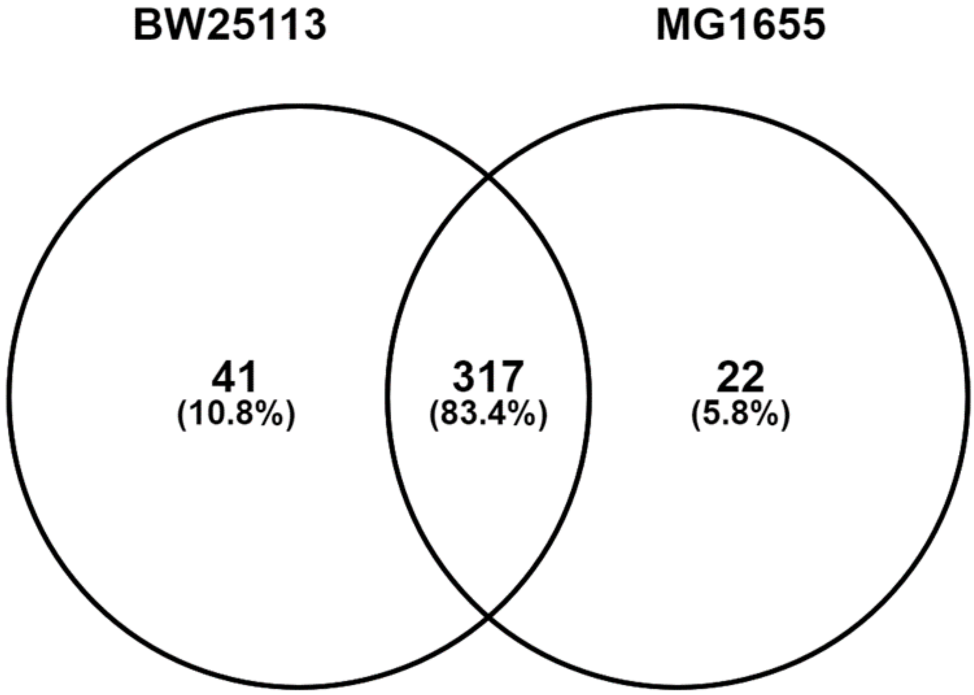


Figure 4.5.1: Comparison of the essential genes of two transposon libraries in *E. coli* K-12 BW25113 and MG1655.

Table 4.3: Comparison of genes identified essential in one strain compared to genes which are ambiguous in the other. Genes which were not identified as ambiguous were then compared to the non-essential grouping. All genes described in this table were identified in both strains *E. coli* K-12 MG1655 and BW25113, for genotypes see Table 2.1 .

Condition	**Gene Name**
Essential in MG1655 Ambiguous in BW25113	*rseP, secD, secF, ppiB, ybeD, ybfB, cmk, gnsA, narJ, tpr, yciM, yddK, ynfN, cspI, zwf, azuC, rpmJ*
Essential in BW25113 Ambiguous in MG1655	*secA, coaE, aceF, ybeY, sucA, tusE, mnmA, tonB, ydaS, yedN, ptsI, fdx. hscA, ftsB, ygeN, ubiH, tktA, higA, rpsO, nusA, obgE, def, rpsG, rnpA, ubiE, rplA, yjbS, psd, rsgA, hold*
Essential in MG1655 Non essential in BW25113	*cysB, ydaC, ydiE, ptsH, polA*
Essential in BW25113 Non essential in MG1655	*cydX, rpmF, ihfA, yqeL ,ygeF, ygeG, rbfA, rpe, gpsA, glmS, dapF*

Therefore, a manual inspection of insertions within these genes was conducted using the Artemis browser. The results presented in Figure 4.5.2 provides examples of the genes which were manually inspected. Upon inspection of these genes, the first thing that was noticed was that in some genes, reported as non-essential in BW25113, an essential domain was observed. This was the case in Figure 4.5.2A when inspecting *polA*. *polA* encodes DNA polymerase I and in BW25113, an essential domain can be observed indicating that although the gene is essential, only part of the gene is required for the essential function. In MG1655, although insertions were observed within *polA* gene they are a considerably lower frequency, this may be due to the lower density of library, but also could suggest a more detrimental effect on fitness of insertions within *polA* within MG1655

Focusing on, *gspA,* a Glycerol-3-phosphate dehydrogenase, essentiality was observed in BW25113 and non-essentiality in MG1655 (Figure 4.5.2B). Interestingly the downstream gene is a serine acetyltransferase, *cysE,* involved in the first step of cysteine biosynthesis. It has a reduced amount of insertions, therefore is classed ambiguous in MG1655 but is non-essential within BW25113. This reduction in insertions suggests that within the MG1655 strains, insertions within *cysE* may have a reduced fitness. Why these differences would be observed between the two libraries is unknown. However, one gene which did occur as essential in MG1655 and non-essential in BW25113 was *cysB* which regulates the sulfonate-Sulphur and Sulphur utilization via cysteine biosynthesis (Figure 4.5.2C). The fact that these two genes, *cysE and cysB* were either reduced in insertions or essential, suggests a potential difference between these two libraries.

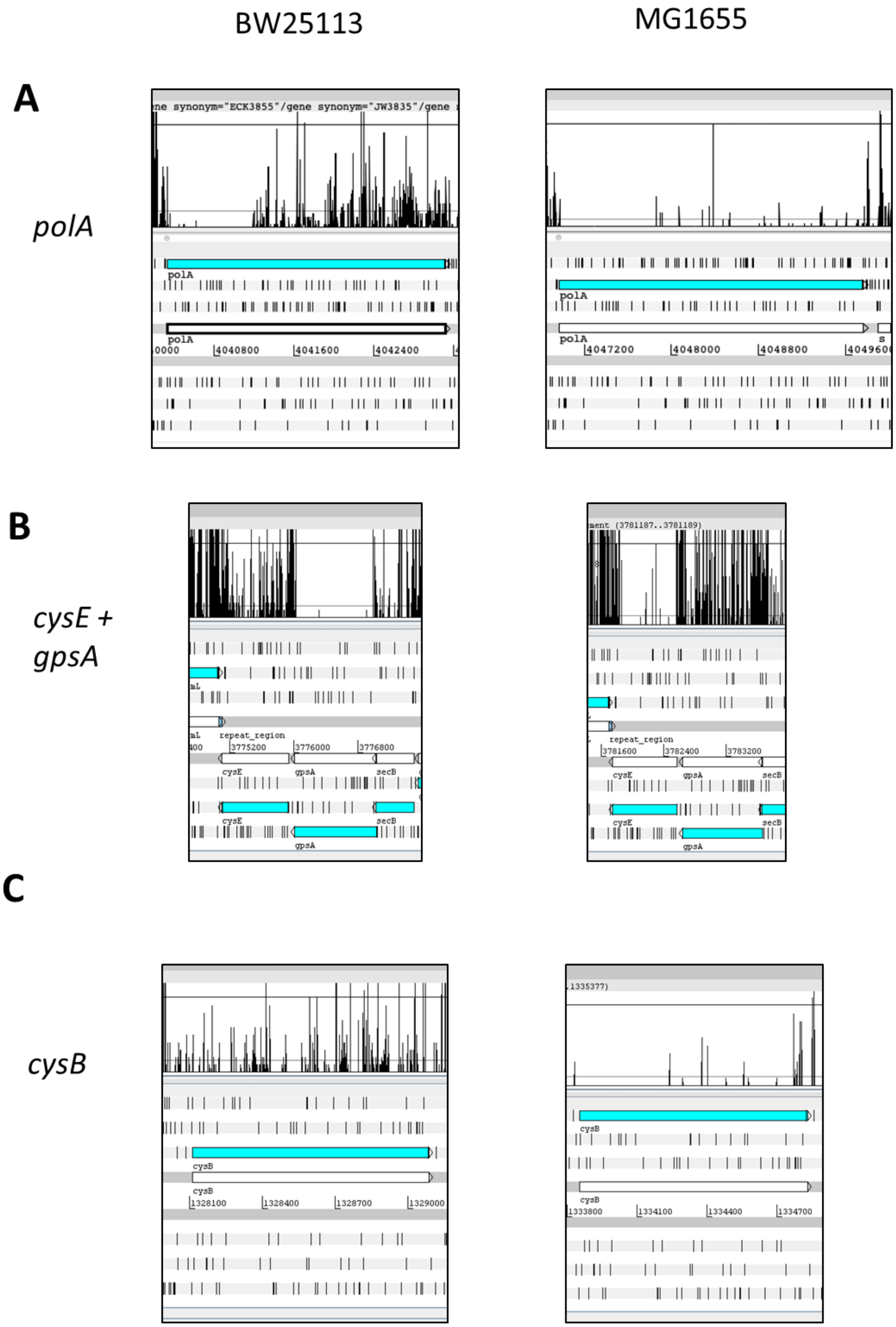


Figure 4.5.2: Manual inspection of transposon inserts using the Artemis Browser. Diagrams on the left are from the BW25113 transposon library from (Goodall *et al.*, 2018). Diagrams on the right are from the MG1655 transposon data.

Although the examples provided suggest that there are potentially some differences observed between the “essentialomes” of BW25113 and MG1655. Before this can concluded, several factors must be considered. The first is the conditions in which the libraries were created, as selection pressures would have been applied during the creation of the library, such as recovery after transformation, and selection of strains containing insertions. Therefore, inserts which were susceptible to these conditions may have been selected for, before the library was sequenced and therefore could be misinterpreted as essential. In addition, since these two libraries are constructed using different antibiotic resistance cassettes, the potential also exists for genes to be classed as essential due to the different stresses occurred due to the antibiotic. As such, considering the way these two library were constructed, although the conditions for construction of the MG1655 library are reported in this study (Section 2.7.1), the precise conditions used for the construction of the BW25113 library are not clearly reported in Goodall *et al*. (2018), including the concentration of chloramphenicol used to select for the library. Therefore, this study cannot determine whether or not the differences in putative essential genes observed above, are due to a strain specific difference or differences in the parameters associated with construction. In order properly determine strain specific effects in essentialome, this study believes that either a library would need to be constructed, using the same transposon and experimental conditions, or perform further experimental validation of these results, such as by attempting to knockout these genes within both strains.

4.6 Summary

Overall, the results described within this chapter cover the construction of a MG1655 library and its subsequent comparison to other libraries. The key points in this chapter are summarized below.

1) It presented the data associated with construction of a transposon library within *E. coli* MG1655, the ancestor strain used in the evolution experiment described in Chapter 3.
2) It demonstrated that there was a high correlation between technical replicates and that potential insertion bias was observed within the library. It also demonstrated that insertion bias occurred around the origin of replication.

3) It compared different computational pipelines used to identify essential genes within a transposon library. The results demonstrated that each pipeline is able to identify a similar core set of essential genes, however they do differ.
4) It concluded that the log likelihood method was the best choice for identifying essential genes within this study.
5) It demonstrated that a comparison of two *E. coli* K-12 transposon libraries in BW25113 and MG1655 detected some differences in the essential genes identified between the two strains.

Chapter 5

Evaluating the Ability of TraDIS to Predict Evolutionary Outcomes using a Short Term Selection Experiment

5.1 Introduction

As mentioned before, the main purpose of this study was to determine whether a short term selection of a transposon library produces results that overlap to some extent with those from a long term evolution experiment. The previous chapters began to address this, as chapter 3 described the details and further analysis of an evolution experiment of *E. coli* MG1655, under the conditions of unbuffered LB beginning at pH 4.5, while Chapter 4 details the construction of a high density transposon library in the *E. coli* MG1655 ancestor used in Chapter 3. Chapter 5 describes the attempt to test this hypothesis, by performing a mini selection experiment (termed the STSE) using the evolution conditions of the experiment described in chapter 3 and the *E coli* MG1655 transposon library described in chapter 4. Chapter 5 therefore details the results of TraDIS employed on the outcome of the STSE to understand the frequency of different insertions which occur under different parameters.

To our knowledge, only one study has so far used TIS to understand an organism's response to acid stress. This study performed TraDIS using a *Salmonella* Derby 14T transposon library, which was challenged to growth in unbuffed LB at pH 4.0 for 12 hours, including a relevant control at pH 7 (Gu *et al*., 2021).The outcome of this study, being the identification of genes, which are implicated in the acid stress response of *S.* Derby, such as *cpxAR*, an envelope stress reponse two component system, and *casC* and *casE*, the type I-E CRISPR-Cas system. Overall, this demonstrates that TIS can be used to identify genes whose function is required for fitness under acid stress. In addition to this, outgrowth of a transposon library has also been performed in LB at pH 7, for a single passage (Langridge *et al*., 2009, Goodall *et al*., 2018) and for a longer 6 day experiment (Langridge *et al*., 2009). Therefore, in the STSE, TraDIS data, was analysed in a standard manner to identify insertion detrimental genes, and thus to reveal genes which function was required for fitness to the organism when grown under varying pH, enabling further understanding of the *E. coli* response.

Further to this, another way to look at transposon sequencing data is to look for genes where insertion of the transposon conveys a fitness advantage, in other words where transposons within a gene are enriched within a population indicating that a loss of function of that gene causes a fitter phenotype under the defined stress. Although this data could be obtained in most Transposon sequencing experiments, very few experiments have actually used enrichment data (Pritchard *et al*., 2014). We propose that by looking at transposon enriched genes in more detail, additional useful information can be obtained from transposon sequencing experiments.

5.2 The Short Term Selection Experiment (STSE)

Therefore, to begin with this chapter, the high density *E. coli* MG1655 transposon library was subjected to the same conditions as the evolution experiment conducted by Sen (2018). This was termed the Short Term Selection Experiment (STSE). The STSE describes the experiment conducted using the same conditions described in Sen, (2018), however instead of just one initial condition of pH 4.5, a control experiment was done where the initial pH was pH 7, to make it possible to determine or identify the pH specific effects present within the evolution strains (Figure 5.2.1). Therefore, the STSE comprised two conditions (starting pH 4.5 or pH 7) in 5ml of unbuffered LB broth. Six independent populations were established for each condition, with each population inoculated with the MG1655 Transposon Library to achieve a starting OD_{600} of 0.05. Each population was then 'evolved' over 10 days at 37^{o}C with aeration, passaging at a 1 in 20 dilution every 22-24 hours. Due to the limitations of cost, for each pH condition only the lineages of three populations were sequenced at Day 1, Day 5, and Day 10. After the completion of STSE, TraDIS analysis was conducted on a total of 20 individual samples, including the two replicates of the MG1655 initial transposon library discussed in Chapter 4. In regard to the MG1655 transposon library this was renamed the Initial Transposon Library (ITL) for the remainder of this study as this referred to the original composition of the library before any condition had been applied to it. For clarification, the samples used in this study are summarized in Table 5.1.

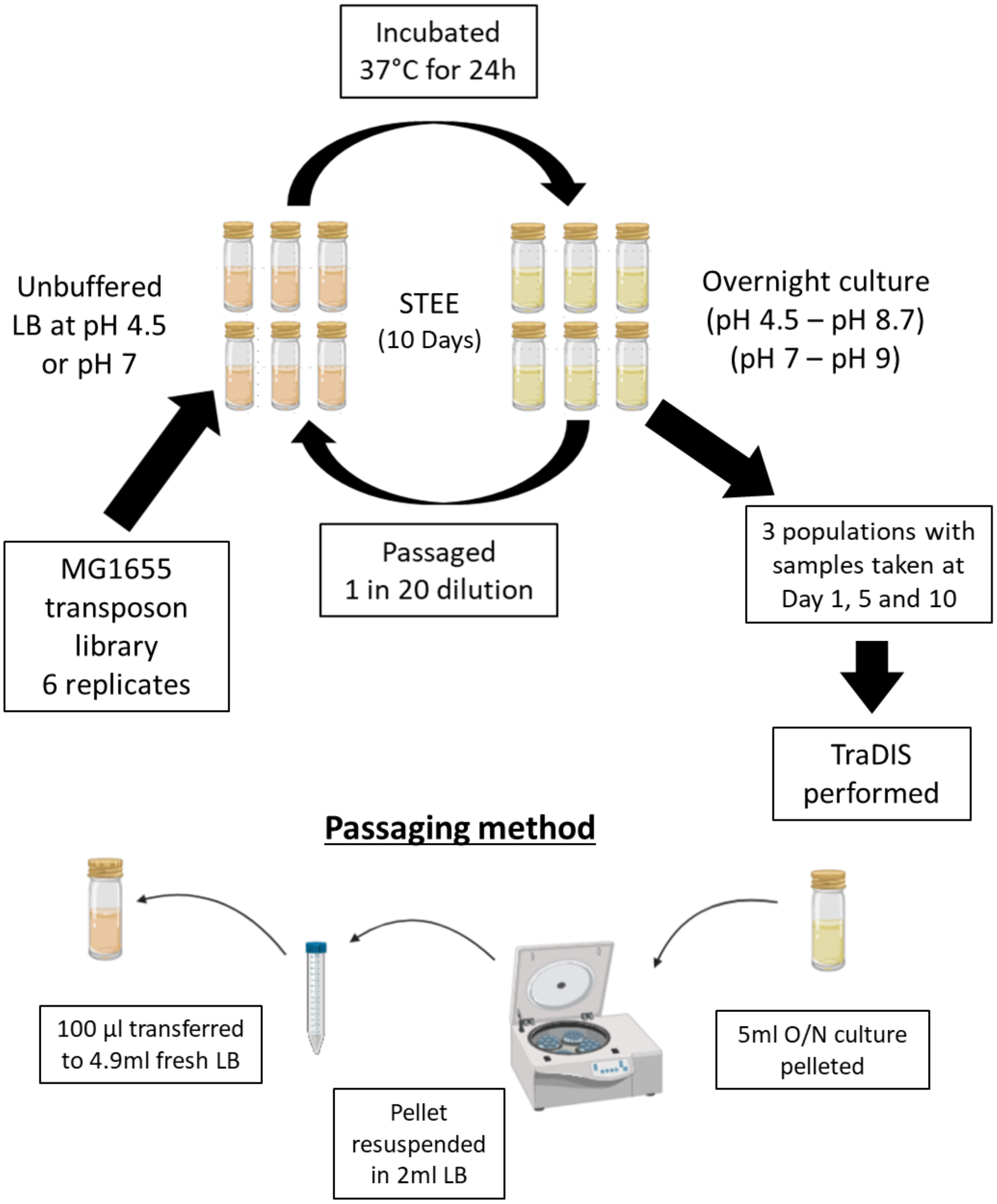


Figure 5.2.1: Overview of the Short Term Selection Experiment (STSE). Overall, a total of 6 replicates were conducted using evolution conditions of Sen, (2018) including an additional control of pH 7 unbuffered LB. The passaging method used in the STSE is described below.

Table 5.1: Samples sequenced in the Short Term Selection Experiment and used for further analysis

Experiment	**Term**	**Samples**	**Description**
Initial Transposon Library	ITL	2	Technical repeats. The MG1655 transposon library before any further growth or stress is performed. All subsequent samples originate from this library
Day 1 pH 4.5	D1-pH4.5	3	Biological repeats. Samples grown in 5 ml unbuffered LB at pH 4.5 for 24 hours.
Day 5 pH 4.5	D5-pH4.5	3	Biological repeats. Samples grown in 5ml LB at pH 4.5 for 5 days, passaging at a 1 in 20 dilution every 24 hours. Samples are direct decedents of D1-pH4.5
Day 10 pH4.5	D10-pH4.5	3	Biological repeats. Samples grown in 5ml LB at pH 4.5 for 10days, passaging at a 1 in 20 dilution every 24 hours. Samples are direct decedents of D1-pH4.5 and D5-pH4.5
Day 1 pH 7	D1-pH7	3	Biological repeats. Samples grown in 5 ml LB at pH 7 for 24 hours.
Day 5 pH 7	D5-pH7	3	Biological repeats. Samples grown in 5ml LB at pH 7 for 5 days, passaging at a 1 in 20 dilution every 24 hours. Samples are direct decedents of D1-pH7
Day 10 pH 7	D10-pH7	3	Biological repeats. Samples grown in 5ml LB at pH 7 for 10 days, passaging at a 1 in 20 dilution every 24 hours. Samples are direct decedents of D1-pH7 and D5-pH7

5.3 Considering the sequencing depth required for each sample

Before performing TraDIS, a consideration was made on the amount of sequencing depth required for each sample. This was to ensure adequate representation and sampling of transposons present within each sample of the STSE. Typically, within normal genome sequencing an average coverage is chosen such as 30X or 100X coverage. However, recently in Goodall *et al.*, (2018) an equation was presented that can be used to estimate the number of insertions, covered by a number of reads. The precise detail of this method is described in Section 2.8.1. Briefly, this method uses iterative input to an equation which estimates the number of insertions site identified, given a set number of reads. In order to do this, it assumes that the original MG1655 transposon library had been over sequenced i.e sequenced to a depth where all insertions present within the library have been identified. Therefore, considering that within the MG1655 transposon library a total of 574250 unique transposon sites were identified using 6011223 reads which were aligned to the genome, the estimation of sequencing depth to unique insertion count was conducted (Figure 5.3.1). Considering this simulation, it was decided to aim for a minimum of 1,730,000 reads per sample, resulting in an estimate 95% overall coverage of insertions within each sample (Table 5.2). Overall, the total number of reads used for each sample of the STSE is summarized in Table 5.3.

Table 5.2: Estimated percentage saturation given a specific amount of read depth, based on the estimate described in Section 2.8.1.

Corresponding estimated insertions	Percentage saturation	Reads
460550	80%	930000
493989	90%	1330000
546018	95%	1730000
568562	99%	2650000

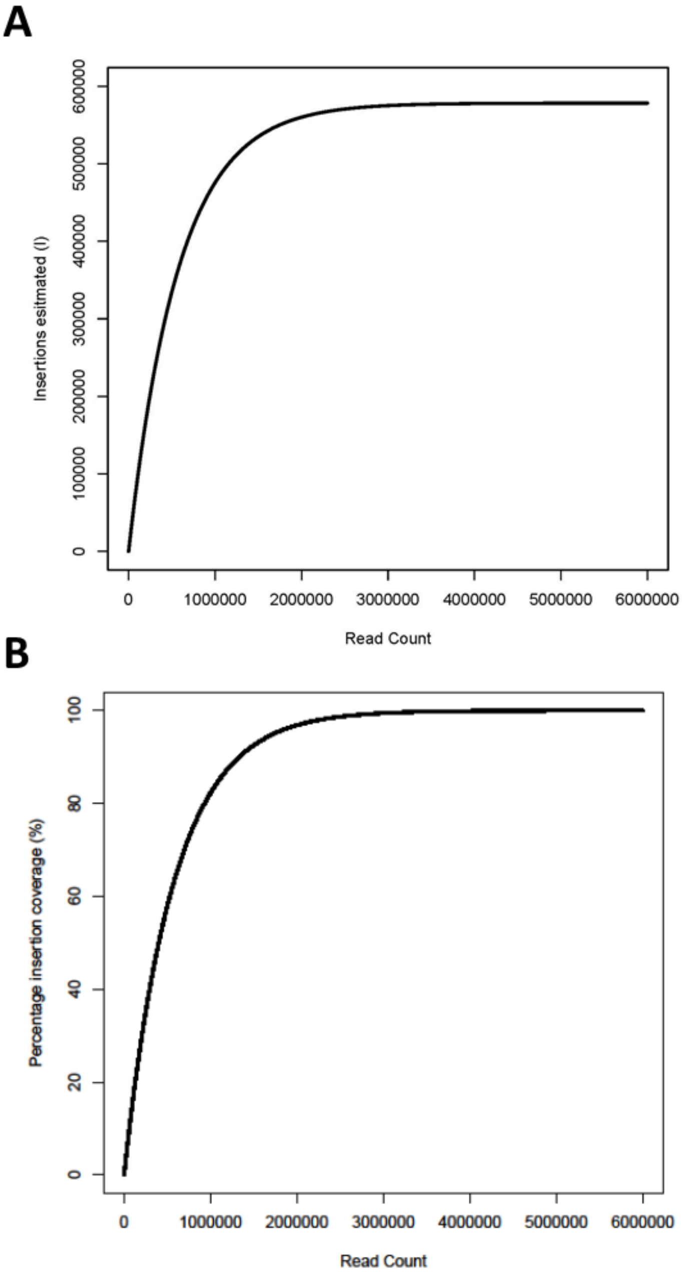


Figure 5.3.1: Results of an iterative process described in Section 2.8.1, to estimated the number of reads required to obtain a set amount of unique insertions (A). Once this was calculated a percentage insertion coverage could be created, respresenting the amount of insertion sites covered by the level of sequencing. These results were generated based on the assumption that the read depth produced for the MG1655 library covers all possible insertions present within the transposon library.

Table 5.3: The amount of reads aligned to the genome corresponding to an insertion for each sample of the STSE. A total describes the number of reads aligned to the genome for each condition of the STSE.

	Sample 1	Sample 2	Sample 3	Total
D1–pH 7	2154574	1801965	2625292	6581831
D5–pH 7	3267871	3362694	1933887	8564452
D10–pH 7	1848517	1855796	1814309	5518622
D1–pH 4.5	1828429	2049927	2840494	6718850
D5–pH 4.5	1734502	1887332	1948216	5570050
D10–pH 4.5	1763807	2361200	1793427	5918434

5.4 Initial look at the transposon data produced by the STSE

Following analysis of the TraDIS data generated in the STSE, an initial overview of the data was then performed. This was done to begin to understand how the composition of transposon library looked after growth under the conditions of the STSE. To do this, reads generated using TraDIS were passed through custom scripts described in Section 2.7.4, which allowed for the identification of reads that corresponded to transposon junctions present on the genome. From this, as with all TraDIS experiments, two features of transposons within a transposon library were obtained. The first is the position of insertions sites on the genome and the second is the number of reads which corresponded to each insertion position. Using these two features in some manner should provide the basis for all further analysis described within this Chapter. To begin, these features were generated for each replicate for each condition and timepoint in the STSE, resulting in 18 individual data sets. This initial overview of the STSE uses this data to analyse the results from replicated experiments.

5.4.1 Loss of complexity within the STSE

In Transposon sequencing experiments, loss of complexity within a library is typically seen upon outgrowth of this library under a defined condition. This is due to loss of fitness caused by insertions in specific genes ("Conditionally essential genes") under condition of the outgrowth. Strains carrying these mutations will decline and eventually may disappear from the population. However, in certain cases the extreme loss of complexity can provide difficulties in downstream analysis, in particular, in determining whether insertion loss within a sample is due to fitness differences to the stress, due to independent factors within the experiment or due to stochastic loss.

Therefore, upon completion of TraDIS, the first thing that was determined was the extent of the loss of complexity within the STSE for each condition and timepoint. To do these the unique insertion sites were identified as if at least one read corresponded to a unique position on the genome. The total amount of unique insertions sites were then generated for each condition and time point of the STSE (Figure 5.4.1). The major conclusion of this analysis was that overall, complexity of the library dropped over the 10 days of evolution in both pH conditions. Particularly, the largest loss of complexity was observed at Day 10, with an average of 100898 (+/- 53119, SD) insertions remaining at pH 4.5. and a larger decline of insertions at pH 7 with an average of 18735.33 (+/- 19835, SD) insertions remaining (Figure 5.4.1). Overall, this decline in insertions suggests that selection is acting upon the library and causing strains with specific insertions to be lost from the population. In addition to this as indicated by the large standard deviation the magnitude of insertion losses between samples were not uniform with insertions lost varying between samples at day 1, 5 and 10.

Considering the implications of the decline, several possible reasons could explain this. One could be the size of the bottleneck in this experiment: a small bottleneck could lead to loss of diversity by random chance. However, considering the amount of culture passaged was large (1 in 20 dilution) it would be highly unlikely that a bottleneck effect is causing this reduction. Alternatively, there could be a biological reason, where insertions in some specific genes cause a fitness advantage which allows strains with these inserts to grow to dominate the population. Therefore, further investigations were done as described in the section below.

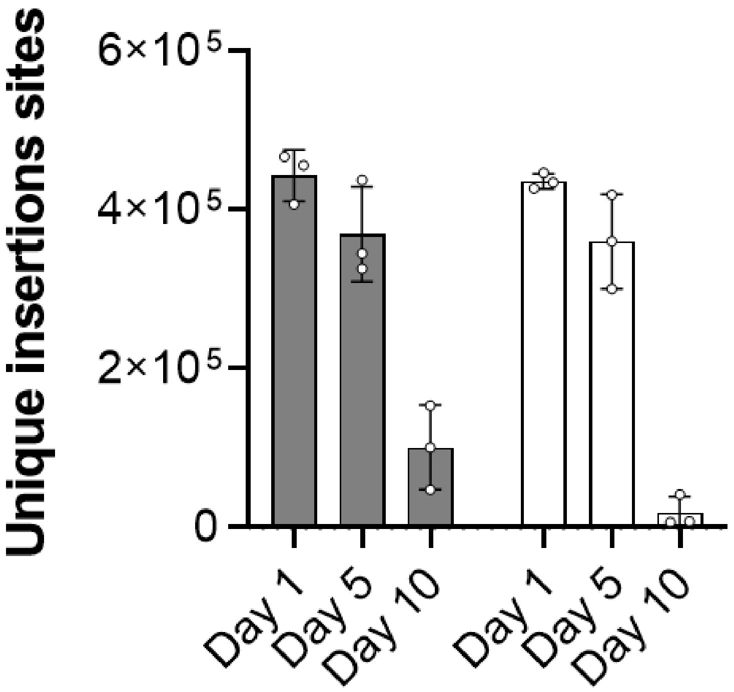


Figure 5.4.1: Average unique insertion points calculated for each condition of the STSE. Error bars show standard deviation. Grey bars refer to the decline in insertions under the pH 4.5 condition and white bars refer to pH 7. The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

5.4.2 Considering reproducibility of the STSE

A decline in insertions was observed in the replicates of the STSE, and a large variation in this decline was observed between replicates. In many transposon sequencing studies, the reproducibility of the study is addressed, with high reproducibility generally observed between replicates (van Opijnen *et al.*, 2009, Phan *et al.*, 2013, Goodall *et al.*, 2018). Therefore, the reproducibility of the STSE was considered.

To consider reproducibility, the two features of each transposon library were used to generate two normalized metrics referring to different aspects of the data for every gene for each replicate within the STSE. These were insertion index and RPKM, referring to normalized metrics of unique insertion sites and read count, respectively. Considering these two metrics should also allow for better identification of different population dynamics within each transposon library.

Using these metrics, two questions were asked. The first was to address the decline observed in Figure 5.4.1, by asking whether the loss of insertions observed over the time course of the STSE was uniform (insertions lost across all genes) or showed a specific pattern (insertions lost in some genes, but not others)? The second was to ask how reproducible the results observed within each timepoint and condition of the STSE were. To consider both these questions simultaneously a pairwise comparison of the different replicates of the STSE was conducted, using both these metrics (Figure 5.4.2 + Figure 5.4.3). This allows us to understand how well each replicate is correlated, and in addition to this allows us to begin to understand the transposon insertion dynamics occurring within TraDIS experiment at the gene level.

5.4.2.1 Variability within insertion indices

The first consideration was the pairwise comparisons of insertion index scores of different replicates. In previous studies, unique insertion sites were found to be highly correlated and this was then used as a reason to combine data sets together (Langridge *et al.*, 2009, Phan *et al.*, 2013, Goodall *et al.*, 2018) As mentioned in Chapter 4, when comparing differences in insertion index score this refers to the number of unique insertions, determined by their location within each gene. It does not refer to the relative frequency of insertion within each gene.

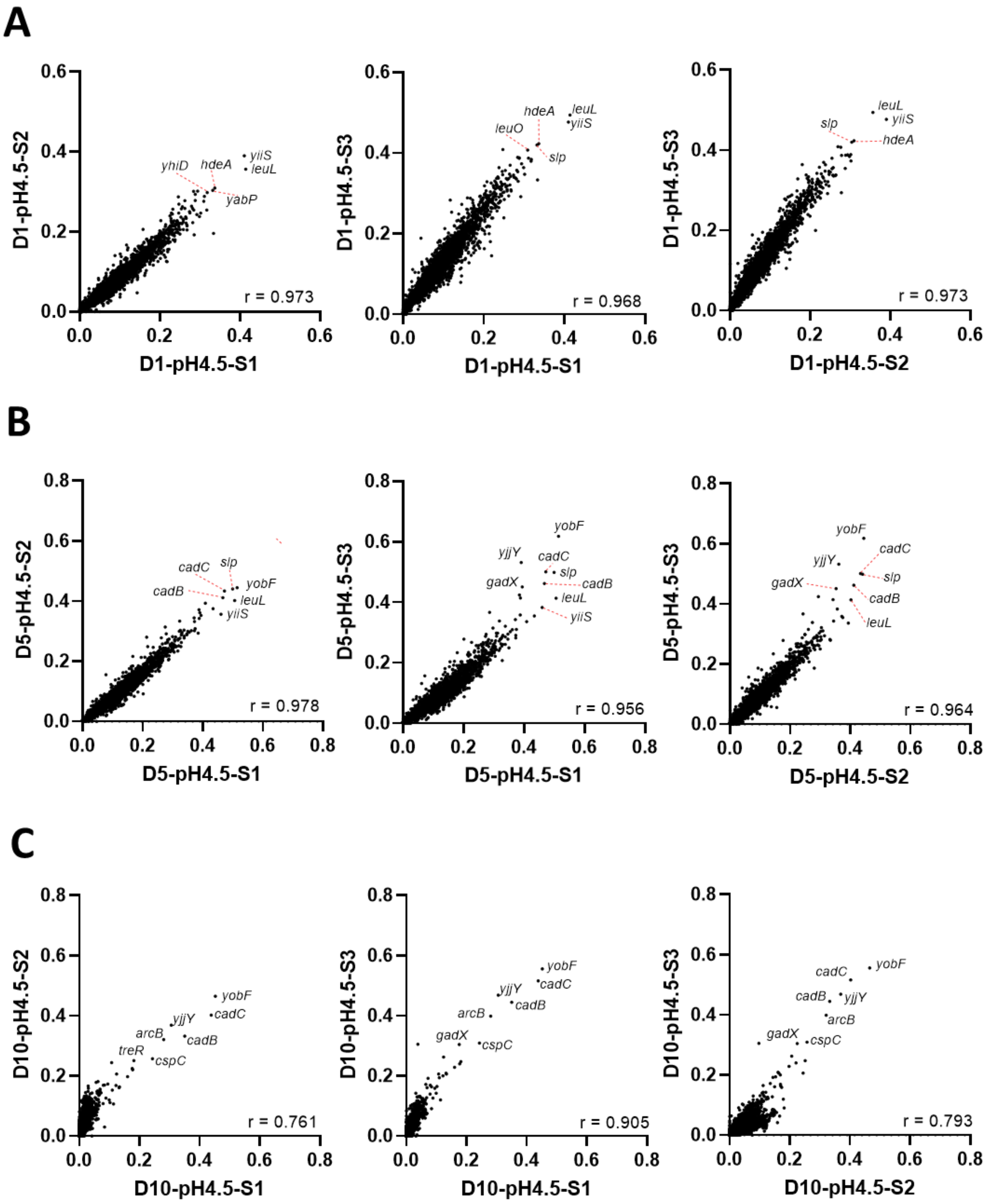
Insertion index scores are calculated as the number of unique insertion sites divided by gene length. Considering that the Tn5 transposon has no specific insertion site bias, the value produced actually represents the proportion of possible insertion sites within gene that actually has insertions present. For example, if a gene had an insertion index score of 0.5, this would actually indicate that 50% of all bases present within a gene will have at least one insertion present. Therefore as shown in Figure 5.4.2 the insertion index score of each sample was calculated and a pairwise comparison conducted.

Upon performing the pairwise comparison of insertion index scores, all pairwise comparisons conducted had a high correlation (defined by a Pearsons correlation coefficient greater than 0.75) (Figure 5.4.2). Further to this, at Day 1 and Day 5, under both pH conditions the samples correlated highly with a correlation coefficient greater than 0.9. The exception was at D1 pH7, where samples S1 and S3 had a correlation coefficient of 0.866.

Although these samples still indicate strong correlation, observation of this pairwise comparison showed more differences in insertion index score between these replicates than in the other examples (Figure 5.4.2D). Interestingly, the genes which were easily identifiable to be different in insertion score were recognized as belonging to *ilv* regulon, involved in the synthesis of isoleucine from L-Threonine. Why these genes might have a higher insertion count in one replicate than the other is unknown; however inspection of RPKM data also revealed these genes to be higher in read count in this sample (Discussed below section 5.4.2.2). Overall, at Day 1 and Day 5 the coefficient correlation indicated that at pH 4.5 the samples correlated better than at pH 7.

As shown in chapter 4, within the library certain genes were shown to have a greater relative frequency of insertions compared to other genes. After Day 1 these same genes were found to still be saturated (*yiiS, slp, leuL* and *hdeA*). However, at Day 5 and Day10, other genes, with some specific to the pH condition, started to show a high insertion density (Figure 5.4.2 BCEF). The genes found with high insertion index score at day 5 were also found at day 10 in their respective condition and population. Of the day 10 samples, pairwise comparisons of the insertion index score reveal that in the majority of genes, the number of insertions had declined, while high insertion density only remained for a few specific genes, which were similar between replicates (Figure 5.4.2C+F). This result showed that the loss of insertions within the samples at day 10 of the STSE was not uniform. Instead within each replicate of the STSE at day 10, a few genes were found with a high insertion index score, potentially indicating that these strains carrying these insertions potentially had a high fitness advantage. Indeed, some of these genes were familiar from the STSE experiments such as *fimE*, *yjjY* and *tnaA*. In addition to this, the loss of insertions was observed at a greater extent at pH 7 compared to pH 4.5, corresponding to the observation shown in Figure 5.4.1, where a greater loss of complexity was seen in strains grown at pH 7. This result suggested to us that over the 10 days of evolution strains with insertions in a few specific genes start to dominate the population, hence leading to an overall decrease in population diversity. However, it should be stated that considering insertion index scores only considers the presence and absence of insertions and not their relative amounts within the population. This point is discussed further in the section below.

Interestingly, as stated above, genes specific to different conditions were sometimes identifiable by index score by day 5 as well as at Day 10. For these genes the unusual result was seen that insertion index score was seen to increase from day 5 to day 10 within replicates. In addition to this, these insertion index scores reported at Day 10 were also higher than the insertion index score reported in the ITL. This indicated that more unique insertions were being identified in the library after day 5 and 10, something which in theory should not happen unless addition transposition events were occurring within the condition of the STSE (this should not be possible as there is no transposase encoded in these strains). An example of this can be seen for the *fimE gene*. In the ITL, this has an insertion index score of 0.293, while in the pH7-S1 population of the STSE, insertion index scores of 0.291 , 0.444 and 0.539 were reported at days 1, 5 and 10, respectively. This result implies that more transposon insertion sites are present within *fimE* at Day 10 than when the library was created, either suggesting that somehow within the transposon library new insertion events were occuring within this library, or that the original library was not sequenced to a sufficient depth to detect all insertions. This unexpected result warranted further study, and the results of this are discussed in Section 5.5 of this chapter.


A
D1-pH4.5-S2
D1-pH4.5-S1
yiiS
leuL
yhiD
hdeA
yabP
r = 0.973
D1-pH4.5-S3
D1-pH4.5-S1
hdeA
leuL
yiiS
leuO
slp
r = 0.968
D1-pH4.5-S3
D1-pH4.5-S2
leuL
yiiS
slp
hdeA
r = 0.973
B
D5-pH4.5-S2
D5-pH4.5-S1
cadC
slp
yobF
cadB
leuL
yiiS
r = 0.978
D5-pH4.5-S3
D5-pH4.5-S1
yobF
yjjY
cadC
slp
gadX
cadB
leuL
yiiS
r = 0.956
D5-pH4.5-S3
D5-pH4.5-S2
yobF
cadC
yjjY
gadX
slp
cadB
leuL
r = 0.964
C
D10-pH4.5-S2
D10-pH4.5-S1
yobF
cadC
yjjY
arcB
cadB
treR
cspC
r = 0.761
D10-pH4.5-S3
D10-pH4.5-S1
yobF
cadC
yjjY
cadB
arcB
gadX
cspC
r = 0.905
D10-pH4.5-S3
D10-pH4.5-S2
cadC
yobF
cadB
yjjY
arcB
gadX
cspC
r = 0.793
0.0
0.2
0.4
0.6
0.8

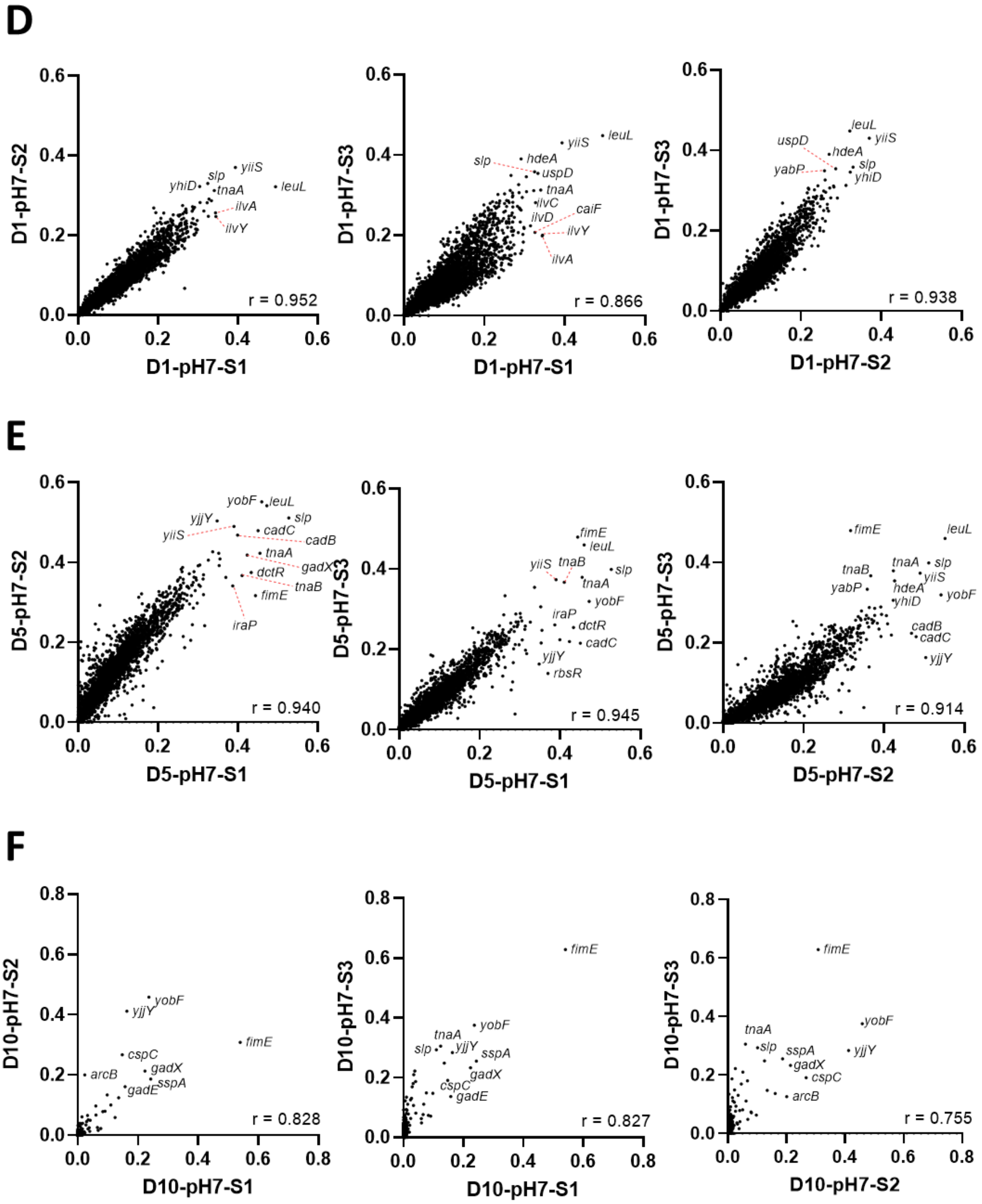


Figure 5.4.2: Pairwise comparison of insertion index scores between replicates of each sample of the STSE. Samples A-C represent the pH 4.5 condition while sample D-F the pH 7 condition, in addition to each comparison of the different timepoints D1 (A+D), D5 (B+E) and D10 (C+F). Each point represents a gene and the genes with highest insertion index score have been labelled (where necessary a red dash line has been placed to ensure clarity of points). The Pearson's correlation coefficient is reported for each pairwise comparison in the bottom right hand corner of each graph, as indicated by the R value.

5.4.2.2 Variability with RPKMs

As stated above, RPKMs refers to the total read count assigned to a gene which has been normalized for gene length and overall sequencing depth. Since the TraDIS method allows for the amplification and specific selection of reads associated with the junction of a transposon, an RPKM value calculated using TraDIS data can therefore represent the relative frequency of insertions within a gene in a given population. Using this metric, it is therefore possible to begin to identify genes where the insertions either decline or increase in the population. Therefore, using pairwise comparisons, differences in population dynamics can start to be inferred between each replicate of the STSE. The result of this pairwise comparison is presented in Figure 5.4.3 and the conclusions of this comparison are presented below.

5.4.2.2.1 Large RPKM values suggest enrichments of insertions within the STSE

As seen above when comparing the insertion indices, at day 10 the population was defined by strains with insertions within a relatively small set of individual genes, while the strains with insertions in other genes declined. This potentially suggests that insertions within these genes, have accumulated within the population, however, insertion indices alone are unable to determine relative frequencies of insertions within the population.

Therefore, the first question which was asked is, can an accumulation of strains in a population with insertions in a particular gene be inferred from the RPKM values? Interestingly the first thing that was noticed was the scale of RPKM values being identified, with the highest values being at least 100 fold higher in all conditions of the STSE after Day 5 , and in D10-pH7, up to 1000 fold larger, than at Day 1. Focusing upon the Day 10 samples of the STSE, this increase in scale was shown to be due to small set of genes which could be clearly defined as having large RPKM values within each replicate (Figure 5.4.3, C+F). This was true at both pH 4.5 and pH 7 with some genes being identified in both conditions but others differing between pH condition. In addition to this, the magnitude of the RPKM value in these genes differed between replicates, indicating that strains carrying insertions within these gene accumulate within the population to different degrees in replicate experiments. This can clearly be seen by comparing the D10-pH7 replicates, where the gene which showed the largest RPKM value differed between *fimE* and *yjjY* in different replicates (Figure

5.4.3 F). This suggested that variations in population structure was being observed between the replicates, with strains carrying insertions in different genes varying in relative frequency.

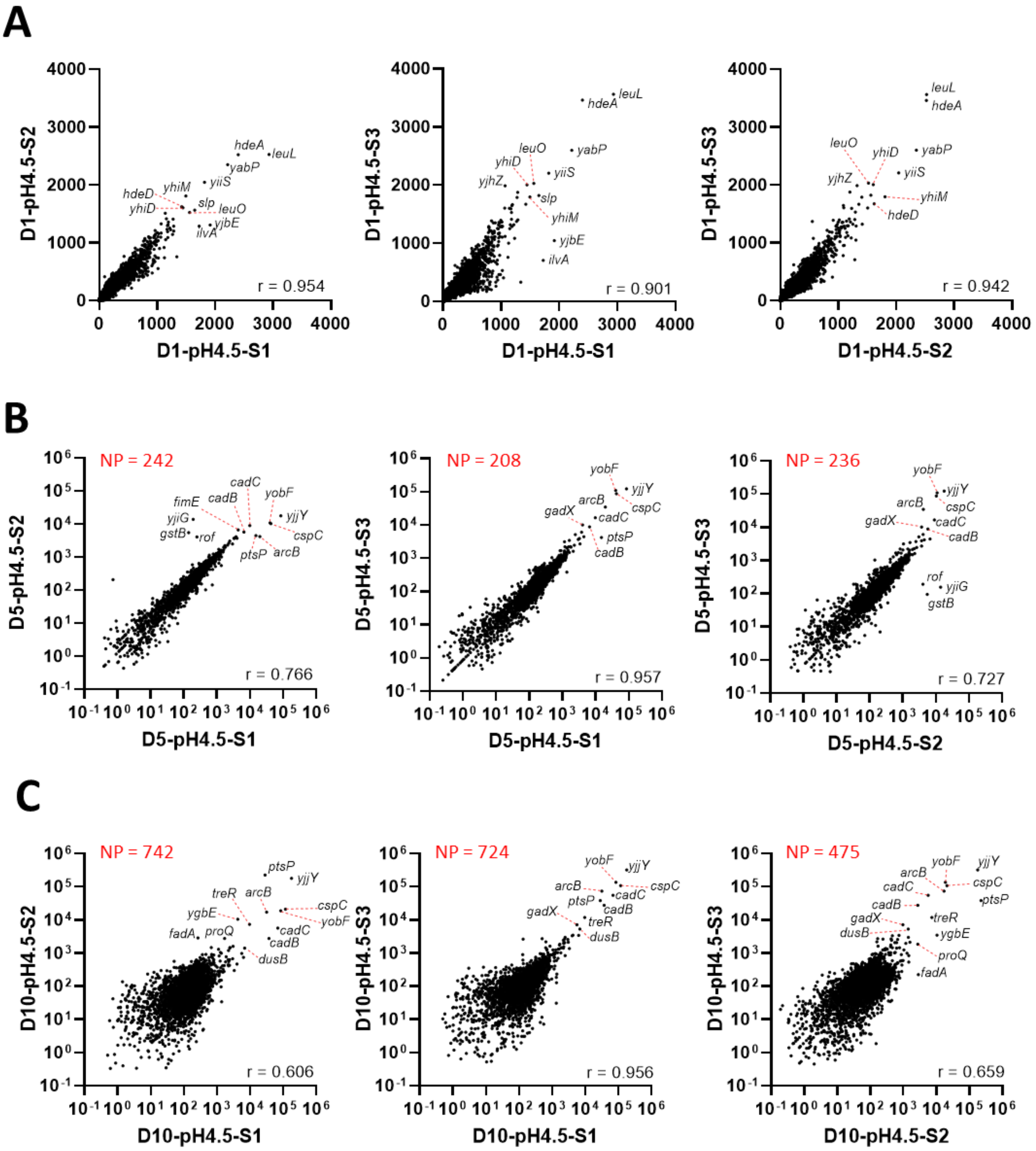
Strikingly, some of the genes identified as enriched by high RPKM values within replicates were genes which had been identified previously in the evolution experiment described in chapter 3. This was the first indication that our overall hypothesis was correct: i.e. that short term selection experiments with TraDIS libraries could to some extent predict the result of longer term experiments which start with a single genotype. Overall, these results suggested that within the STSE, in particular at Day 10 and to some extent day 5, strains with insertions within particular genes come to make up a large proportion the population.

5.4.2.2.2 Considering correlation within STSE of read count

Once enrichments were identified, the next step was to consider how well the RPKM values of the STSE replicates under different conditions correlate with each other? To address this, a Pearson's correlation Coefficient was generated for each pairwise comparison (Figure 5.4.3). This showed, unexpectedly, that replicates generated at pH 4.5 correlated better than replicates generated at pH 7, and this was true for all time points. However, at Day 1 a strong correlation was observed for both conditions, with the Pearson's correlation coefficients ranging from 0.613-0.886 at pH7 to 0.901-0.954 at pH 4.5. For the D1 sample at pH4.5, the RPKM values were highly correlated; in particular the same genes, *leuL, hdeA, slp* and *yjiS*, were identified as having the highest RPKM values (Figure 5.4.3 A). These genes were also identified to have high RPKM values at pH 7, but with higher variability between replicates. Interestingly, these genes were also observed to have the highest RPKM values within the Initial Transposon library (ITL) (Figure 4.2.1). This result implies that the length and extent of the stress provided by the condition of the STSE was not strong enough to be able to observe accumulation of insertions specific to the condition of the STSE after only one day of selection.

However, in the D1-pH7 samples, in addition to similar genes identified in the ITL between replicates, another distribution was observed indicating a change in population levels of strains carrying insertions in these genes (Figure 5.4.3 D). An initial inspection of this showed that within two samples S1 and S2 at D1-pH7, a different distribution of high RPKM values was observed in the genes *ilvACDY*. A similar observation was also observed with the insertion index score (Figure 5.4.2 + Figure 5.4.3 , D). As stated before, these genes are involved in the synthesis of isoleucine from L-Threonine. High RPKM values in *ilvA* were also observed in all pH conditions however to a lesser extent. Therefore, this result potentially indicates that after 1 day of growth under the conditions in the STSE, the make-up of the population has already begun to change. The level of variation within this second distribution differs between samples, with D1-pH7-S1 having a larger difference than D1-pH7-S3S3. In addition, a slight effect is also observed in the populations grown at pH4.5 at Day1, although to a lesser extent than at pH 7.

Now considering the correlations between replicates of the STSE after day 1, we observed that correlation between replicates decreased over time, with replicates differing more at Day 10 than at other time points. At pH 4.5, each replicate still showed moderately strong correlation, with the Pearson's correlation coefficient greater than 0.6 for all pairwise comparisons; however, variation between replicates increased at pH 4.5 with samples at day 10 showing the least correlation. (Figure 5.4.3: Min pairwise comparison r value: Day 1 = 0.901, Day 5 = 0.727 , Day 10 = 0.606). This increase in variation over time was also true at pH 7, and variation was larger for samples grown at pH 7 than for those grown at pH 4.5 (Figure 5.4.3: : Min pairwise comparison r value: Day 1 = 0.613, Day 5 = 0.055 , Day 10 = 0.034). Despite the low correlation overall at Day 5 and Day10, some similarity between samples could be seen within the pairwise comparison of these results and this is described in the next section (Figure 5.4.3 B, C, E + F).

A
D1-pH4.5-S2
D1-pH4.5-S1
hdeA
leuL
yabP
yiiS
yhiM
hdeD
yhiD
slp
leuO
yjbE
ilvA
r = 0.954
D1-pH4.5-S3
D1-pH4.5-S1
leuL
hdeA
yabP
leuO
yhiD
yjhZ
yiiS
slp
yhiM
yjbE
ilvA
r = 0.901
D1-pH4.5-S3
D1-pH4.5-S2
leuL
hdeA
leuO
yhiD
yabP
yjhZ
yiiS
yhiM
hdeD
r = 0.942
0
1000
2000
3000
4000
B
D5-pH4.5-S2
D5-pH4.5-S1
NP = 242
cadC
fimE
cadB
yobF
yjiG
yjjY
gstB
rof
cspC
ptsP
arcB
r = 0.766
D5-pH4.5-S3
D5-pH4.5-S1
NP = 208
yobF
yjjY
arcB
cspC
gadX
cadC
ptsP
cadB
r = 0.957
D5-pH4.5-S3
D5-pH4.5-S2
NP = 236
yobF
yjjY
arcB
cspC
gadX
cadC
cadB
rof
yjiG
gstB
r = 0.727
C
D10-pH4.5-S2
D10-pH4.5-S1
NP = 742
ptsP
yjjY
treR
arcB
cspC
ygbE
yobF
cadC
fadA
proQ
cadB
dusB
r = 0.606
D10-pH4.5-S3
D10-pH4.5-S1
NP = 724
yobF
yjjY
arcB
cadC
cspC
ptsP
cadB
gadX
treR
dusB
r = 0.956
D10-pH4.5-S3
D10-pH4.5-S2
NP = 475
yobF
yjjY
arcB
cadC
cspC
ptsP
cadB
treR
gadX
ygbE
dusB
proQ
fadA
r = 0.659
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
10^{6}

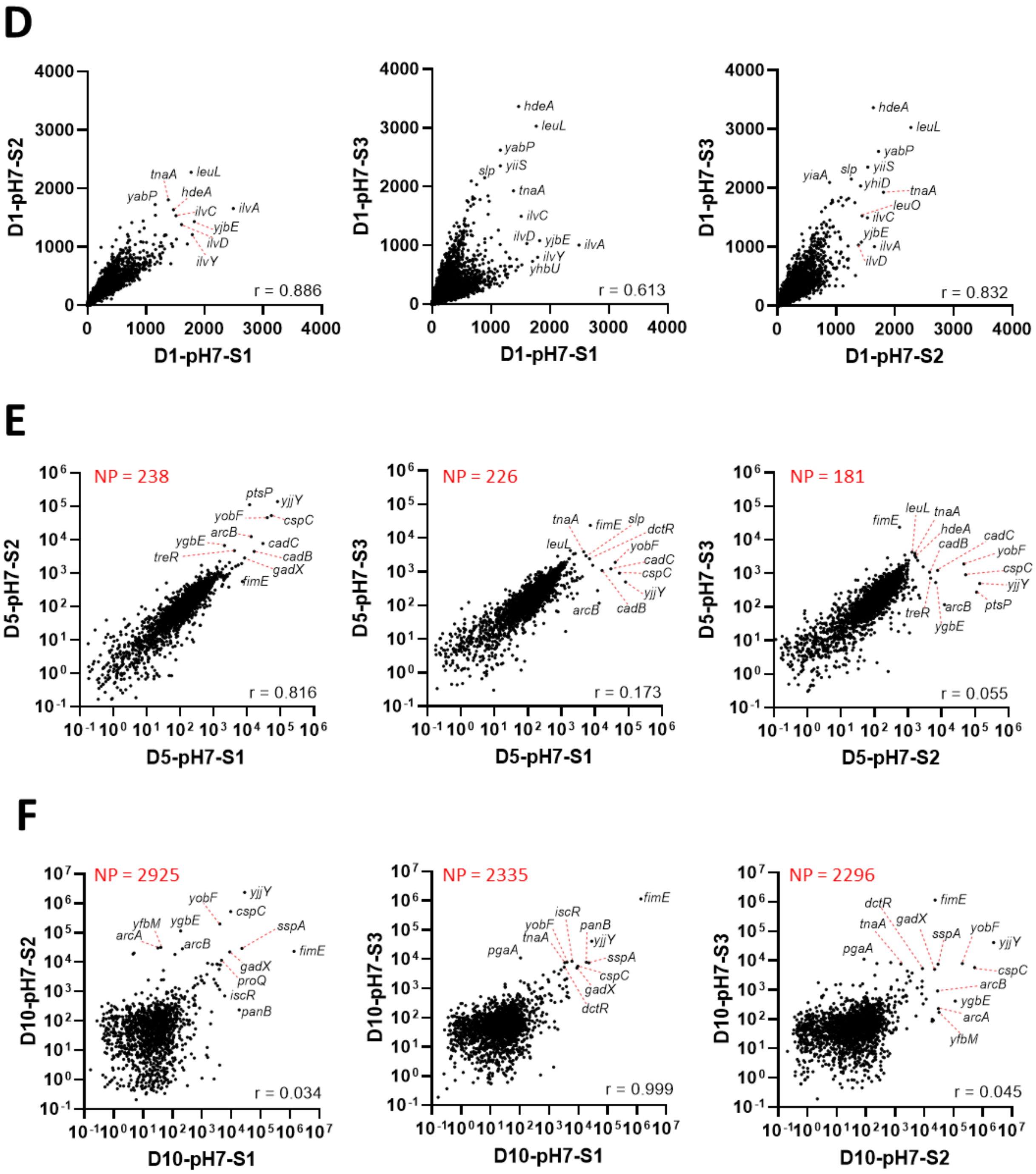


Figure 5.4.3: Pairwise comparisons of RPKM values for each gene in replicates within the STSE. Comparisons of replicated are shown for each pH condition, pH 4.5 (A-C) or pH7 (D-F), as well as timepoints of Day 1 (A+D), Day 5 (B+E) and Day 10 (C + F). Pearson correlation coefficients were calculated for each pairwise comparison. Data was plotted on a logarithmic scale for Day 5 and Day 10 as high RPKM values observed for a few individual genes, which compressed the majority of data on a linear scale. Note that in the graphs with logarithmic scales, genes with RPKM of 0 value cannot be plotted, and the number of genes which were not plotted (NP) because of this is indicated on the graphs in red. Note that at Day 10 pH 7 (F), an unusual result in Pearson's correlation coefficient is identified this is discussed further in Section 5.4.2.2.4

5.4.2.2.3 Accumulation of insertions differs between replicates: a focus upon results at pH7 in the STSE

As stated above, correlation of RKPM values between replicates decreased over the timepoints with the correlation decreasing more under the pH 7 condition of the STSE than the pH 4.5 condition. Further inspection of replicates at pH 7 Day5 and Day 10 revealed an interesting observation. As described in Section 5.4.2.2.1, after day 5 and day 10 enrichments of inserts within particular genes were observed, within each replicate, however interestingly at pH 7 the order and magnitude of RPKM scores of different genes differed considerably between the replicates.

To illustrate this, the ten genes with the highest RPKM were ranked for each replicate in each condition at Day 10 of the STSE (Table 5.4). Considering this in the case of pH 4.5, the enrichments were similar for the same genes, and to the same magnitude, and thus correlated well. However at pH 7, large differences between the replicates could be observed for the genes with the highest RPKM values, with one replicate (D10-pH7-S2) being particularly different from the other two (D10-pH7-S1 + S3; Figure 5.4.3 F). However, despite this difference, the majority of genes found to be enriched in D10-pH7-S2 were also found to be enriched to some extent within the other two replicates. In other words, rather than the identity of the genes, it was the magnitude of the enrichment observed which differed more. However as stated above the gene with the highest enrichment of inserts differed between the replicates, with the gene being *fimE* or *yjjY* at pH 7. This overall showed the population at day 10 was dominated by strains containing insertions within either *fimE* or *yjjY*. Interestingly within the samples where *fimE* had the highest frequency of insertions, the second most enriched gene was *yjjY*, and in the sample where *yjjY* was the most enriched, *fimE* had the 9th highest RPKM value. This indicated that still strains with inserts in the alternative gene were being selected for but to a lesser extent.

Further investigation into these replicates indicated that of the two samples where *fimE* dominated, the RPKM values reported for the other enrichments declined dramatically, showing that within these replicates, strains carrying insertions within *fimE* were present in the population at such as high frequency that the frequency of other strains was reduced.

This was also seen for D10-pH7-S2 where *yjjY* dominated, however the extent of the dominance, may not be as great as within the other replicates, as the 2nd and 3rd highest RPKM value genes (*cspC* and *yobF*), the extent of RPKMs for these genes were 10 fold greater than other replicates (Table 5.4, Figure 5.4.4). Therefore, these results suggest that within different replicates although overall similar genes were being identified the extent of the enrichments differed.

Table 5.4: Top ten genes with the highest RPKM ranked for each replicate of the STSE at Day 10 at pH 4.5 or pH7. Genes which were present in multiple samples are colour coded accordingly. The RPKM value for each gene is stated within the brackets

Rank	D10-pH 4.5			D10-pH 7		
	S1	S2	S3	S1	S2	S3
1	*yjjY* (177889)	*ptsP* (222982)	*yjjY* (319452)	*fimE* (1367628)	*yjjY* (2366795)	*fimE* (1159431)
2	*cspC* (117001)	*yjjY* (176949)	*yobF* (135155)	*yjjY* (28575)	*cspC* (527797)	*yjjY* (40803)
3	*yobF* (84315)	*cspC* (20771)	*cspC* (106968)	*sspA* (23052)	*yobF* (201088)	*pgaA* (10995)
4	*cadC* (68664)	*yobF* (18596)	arcB (73213)	*panB* (18759)	*ygbE* (115182)	*iscR* (8501)
5	*cadB* (37013)	*arcB* (16961)	*cadC* (54973)	*cspC* (9591)	*yfbM* (31665)	*panB* (7779)
6	*arcB* (31225)	*ygbE* (10391)	*ptsP* (38059)	*gadX* (8876)	*arcA* (30465)	*yobF* (7777)
7	*ptsP* (28085)	*treR* (7327)	*cadB* (27368)	*iscR* (5972)	*sspA* (29585)	*tnaA* (7620)
8	*treR* (9533)	*cadC* (5642)	*treR* (11782)	*proQ* (4703)	*arcB* (28802)	*sspA* (7478)
9	*dusB* (6828)	*fadA* (2847)	*gadX* (7086)	*rcsB* (4246)	*fimE* (23286)	*cspC* (5770)
10	*yjdQ* (6258)	*cadB* (2750)	*dusB* (5260)	*yobF* (4093)	*gadX* (22479)	*dctR* (5289)

To explore these different population dynamics further, we wanted to consider how RPKM values changed with time went by, as this shows the relative numbers of strains with insertions in a specific gene present within the population. Therefore, for each replicate, the total RPKMs for the whole population were generated and a percentage RPKM value for each gene was then calculated, to represent the relative percentage of strains with inserts within that specific gene in the population at a given time. Using this the percentage value, the results for *fimE* and *yjjY* were then plotted to visualize their dynamics throughout the duration of the STSE at pH 7 (Figure 5.4.4). This result clearly demonstrated the difference within replicates, with regards to *fimE* and *yjjY*, with the pH7-S2 population not showing an increase in *fimE* at Day5, while the other replicates did. Interestingly, as shown in Figure 5.4.3 E, at day 5 an increase in *yjjY* was observed in the pH7-S2 population, and also in pH7-S1 which ultimately becomes enriched in *fimE* insertions at day 10. Therefore, this result suggests that early on in the pH7-S2 replicate, an event occurred which prevented the selection of strains containing insertions in *fimE*. Interestingly, in pH 7- S2 a slightly larger proportion of RPKM was attributed to *yjjY* compared to the other two samples at Day1, indicating a slight increase in insertions which may explain why *yjjY* then accumulated and *fimE* did not. How this would influence the overall result observed within this replicate is however remains unclear.

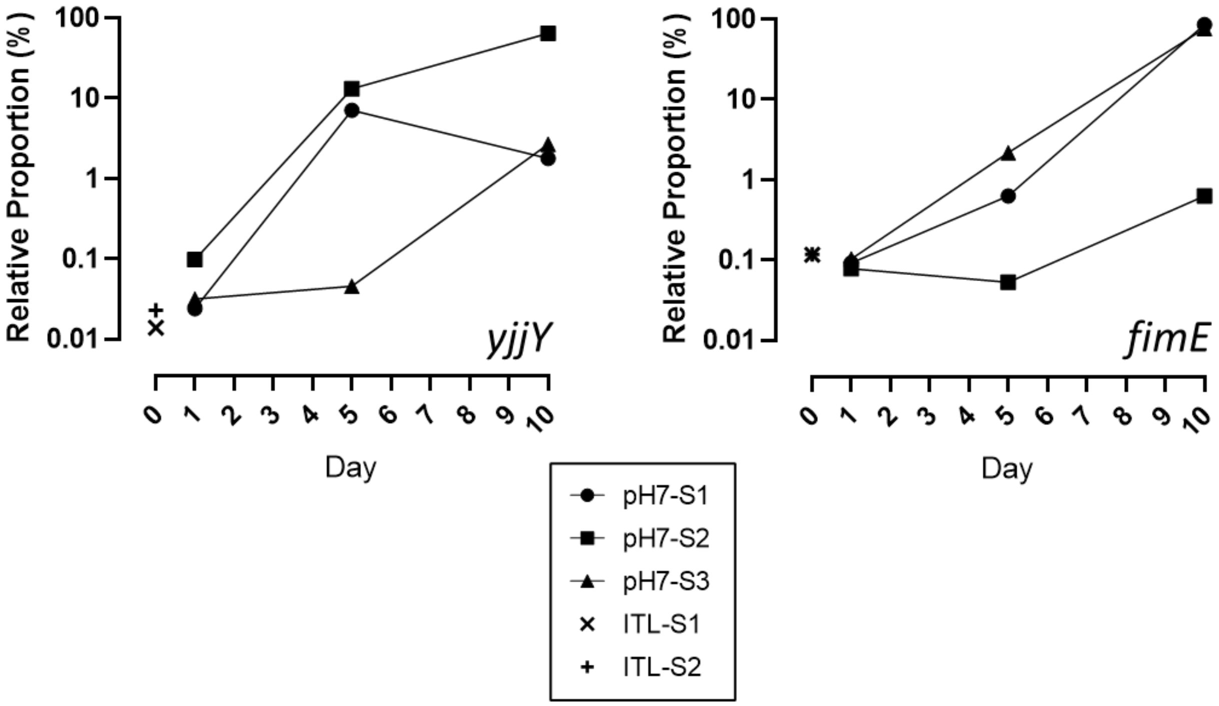


Figure 5.4.4: Relative percentages of RPKM values for *yjjY* and *fimE* over the time course of the STSE. Each replicate of the STSE is plotted independently. The proportions of each gene in the original MG1655 transposon library are indicated by the ITL point.

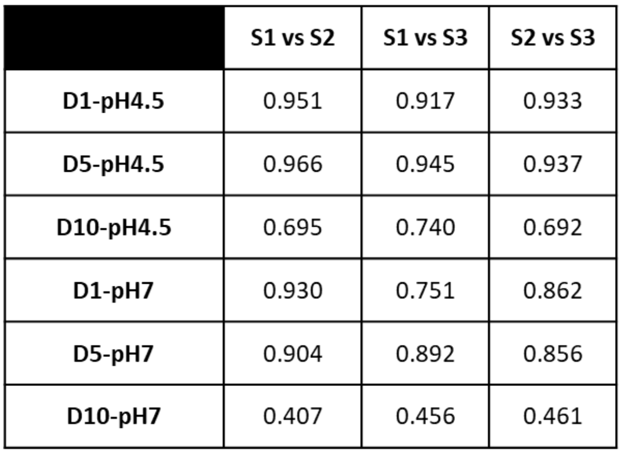
5.4.2.2.4 Correlation considered by rank and not by value

The section above showed that for the STSE at pH 7 Day5 and Day 10, interesting differences in population dynamics were seen between the samples. Further focus upon these replicates showed that the Pearson's correlation coefficient did not represent what was actually occurring between replicates. The Pearson's correlation coefficient is calculated by using the sum of RPKM values between replicates and this is how linear correlation is reported. As described in the above sections, at day 5 enrichments of strains with inserts within particular genes were starting to be observed. At pH 4.5, these enrichments were similar between genes, and occurred to approximately the same magnitude. Thus, they correlated well as shown by a high Pearson's correlation coefficient (see Figure 5.4.3). In the pH 7 condition, comparison of replicates showed that in the majority of genes, to an extent a linear correlation could be observed. However, in the day 5 and 10 a set of genes differed in the magnitude of enrichment, between replicates. A clear example of this can be found within D5-pH7-S2 vs D5-pH7-S3 comparison where the majority of RPKMs can be observed to be highly correlated, although for a handful of genes each replicate differed in the magnitude of enrichments observed, with D5-pH7-S2 showing an enrichment in genes such as *ptsP, cspC, yobF, yjjY, arcB* while in D5-pH7-S3 an enrichment is only observed in *fimE*. In both these cases the RPKM values of these genes were also high within the other sample, however the RPKM value of these genes was 10 – 100 fold greater in one replicate than the other.

Since the Pearson's correlation coefficient is calculated using the sum of all values, significant outliers can have a profound effect upon the coefficient. An example of this can be seen with the comparison of D5-pH7-S2 vs D5-pH7-S3 for which a Pearson's correlation coefficient of 0.055 is reported. However, removal of the five enriched genes reported above (*fimE, ptsP cspC, yobF and yjjY*) gives a revised Pearson's correlation coefficient of 0.550, indicating stronger linear correlation between replicates in most of the other genes in the population. This result thus suggests that the samples do largely correlate, however outliers distort the Pearson's correlation coefficient. Therefore, within replicates of the STSE at day 5 and 10 timepoints, the effects that these enrichments have upon the calculated correlation is significant. At pH 4.5 enrichments are fairly consistent between the replicates, in terms of the genes enriched and the extent of the enrichment as indicated by the RPKM values. But at pH 7 differences were observed not only in the genes enriched, but also the extent of the enrichment differed to a much larger degree; hence why the Pearson's correlation coefficient in some replicates suggested no correlation (Figure 5.4.3F).

We therefore looked at the effect of using ranking rather than actual numbers to look at correlation between replicates. To do this, a Spearman's correlation coefficient was also calculated for each pairwise comparison, with the results presented in Table 5.5. Spearman's correlation coefficient calculates correlation differently, by ranking the RPKM values from highest to lowest before essentially calculating a Pearson's correlation co-efficient. Therefore, Spearman's coefficient does not report linear correlation but reports a monotonic relationship between two data sets. Therefore, in the RPKM data this removes the variability introduced by the extent of the enrichment and therefore only reports correlation based on the order of RPKM values. The results of this analysis indicate that in all conditions of the STSE, the ranked order of the RPKM showed good correlation between the replicates. This is particularly prominent at day 5 condition at pH 7 where the order of ranked RPKMs is largely the same, although the magnitude of the enrichments is very different.

Table 5.5: Spearman's rank correlation coefficient, calculated for each replicate for each condition and timepoint of the STSE.



	S1 vs S2	S1 vs S3	S2 vs S3
D1-pH4.5	0.951	0.917	0.933
D5-pH4.5	0.966	0.945	0.937
D10-pH4.5	0.695	0.740	0.692
D1-pH7	0.930	0.751	0.862
D5-pH7	0.904	0.892	0.856
D10-pH7	0.407	0.456	0.461

5.4.3 A discussion on insertion drop at D10- pH 7

Considering the results shown in Table 5.5, the condition where the smallest correlation was reported using both Pearsons and Spearman's coefficient was D10-pH7. These samples showed the largest variation in enrichments, and the largest reduction in insertions within the STSE, with the number of insertions remaining being 6922, 7649 and 41635 insertions for the three samples S1 ,S2 ,S3 respectively. This roughly equates to < 10% of total insertions in the ITL being present in all the replicates of Day10-pH7 condition. Given the extent of the insertion number drop, it would be expected that within these conditions, reproducibility would also drop.

Whether strains carrying these insertions have completely disappeared from the population cannot be determined. As shown above, within these samples a large proportion of the population at Day 10 pH7 was dominated by strains with insertions within a very small set of genes. The sequencing depth used within this study is limited in its ability to detect insertions present at low frequency, so a possibility exists that insertions which were not identified may still be present within the STSE at low frequency, which is unable to be detected due to the accumulation of large numbers of strains with insertions in a small number of genes. How would this affect further interpretation of the results of the STSE ? Based on this observation, identification of genes where insertions have a detrimental effect on strain fitness will be impossible at Day10-pH7 and to a degree D10-pH4.5, as it would be impossible to distinguish between genuine loss of fitness, and stochastic loss caused by proliferation of a small number of strains with insertions in a few genes. Therefore, only genes which showed an accumulation of insertions will be discussed below.

5.4.4 Easy isolation of transposons from condition of the STSE

Once enrichments of insertions had been identified within the STSE, an attempt was made to try and isolate strains carrying some of these insertions for further analysis later. Isolating particular insertions from a high density transposon library can be difficult as each insertion will be at low frequency. However, under the condition of the STSE, since enrichments of strains with insertions in a few genes were observed, it should be feasible to take a simple PCR approach to identify inserts. Therefore, glycerol stocks of each replicate population from Day 10 of the STSE were streaked out onto LB agar plates and 10 single colonies from each replicate were used to set up overnight culture. From there a PCR approach with flanking primers was used to see if strains with insertions in *fimE* and *yjjY* could be identified. Of the 60 colonies which were screened, 2 contained a *yjjY::Tn5* insertion (isolated from D10-pH4.5-S1 and D10-pH4.5-S3) and one *fimE::Tn5* insertion was isolated from D10-pH7-S3. This result, although not properly quantified, suggests that the accumulation of insertions observed within the TraDIS experiment does actually equate to enrichments observed within the population. These were subsequently sequenced to reveal Tn5 insertions in *yjjY* at positions +25 and -1, both in the opposite orientation to the *yjjY* gene. The *fimE* insertion was at +65 in *fimE*, in the same orientation as the *fimE* gene.

5.5 Enrichments highlight a flaw in TraDIS method – The 60 insertion library.

As highlighted in section 5.4.2.1, over the time course of the evolution experiment, under both pH conditions, the insertion index score within some specific genes started to increase to the point where it was higher than the initial transposon library (Figure 5.5.1, Figure 5.4.1). This was unexpected, since insertion index describes, the amount of unique insertions presents in the library, and on the face of it this increase suggests that new transposon insertion sites were being identified during the STSE which were not identified within the initial library. An example of this is shown in Figure 5.5.1, where a clear apparent increase in insertions is observed within *sspA* at Day10-pH7 of the STSE compared to the ITL. Within some of the genes where this phenomenon was observed, approximately a doubling in the number of unique insertions present was seen at Day10 when compared to the ITL ().

Table 5.6: Examples of genes with an increased insertion index score at Day 10 in one replicate of the pH 4.5 or pH 7 condition of the STSE compared to their insertion index score in the ITL.

Gene	ITL	Day 10	Fold Change
fimE	0.350	0.628	1.79
yjjY	0.170	0.468	2.75
yobF	0.257	0.556	2.17
cspC	0.152	0.310	2.04
sspA	0.110	0.255	2.32
cadC	0.325	0.516	1.59

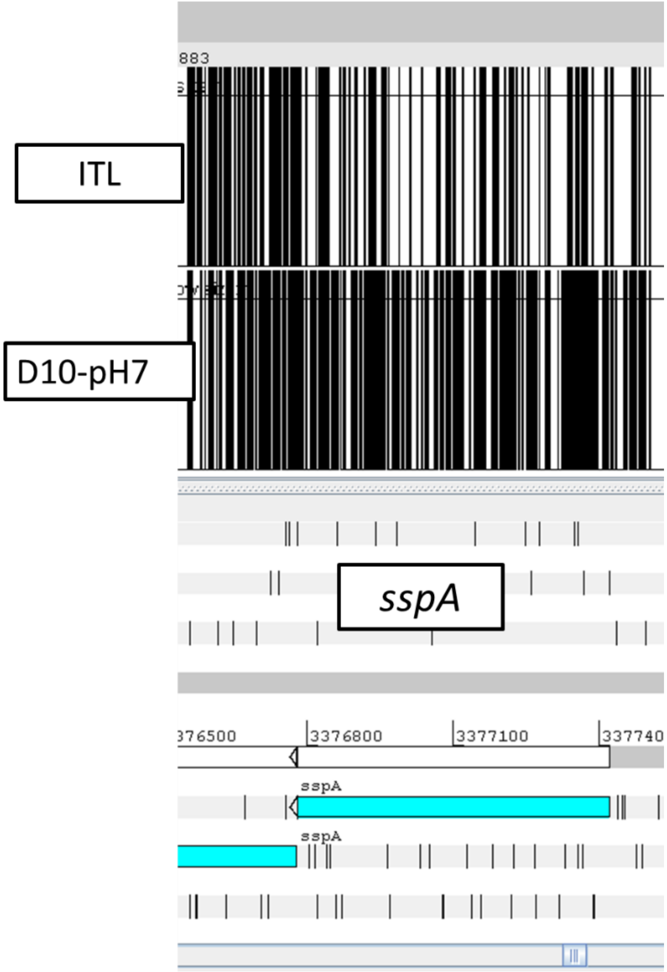


Figure 5.5.1: Transposon insertion sites apparently increase during the STSE. Insertions visualized using Artemis. The example here is *sspA* with TraDIS data taken from the Initial transposon library (ITL) and at D10-pH7 of the STSE. Black lines represent an insertion site. Scale is set to 0-1.

Interestingly, genes where an increased insertion index score was seen also have a high RPKM count, indicating that this phenomenon was occurring in genes where insertions have higher fitness under the conditions of the STSE. The simplest explanation for this was that the original MG1655 library was not sequenced to a sufficient depth, and therefore due to the accumulation of insertions which was occurring during the STSE, more insertions were identified. An alternative explanation is that this observation was due to artefacts introduced in the method used to generate TraDIS data.

Although background noise within TIS data and in general HTS data has been reported before (Chao *et al*., 2013, Chao *et al*., 2016, Barquist *et al*., 2016, Park *et al*., 2017). To distinguish these possibilities within our study, the question was asked: does increased insertion count occur within a library with a known number of insertions? To address this the concept of the "60 insertion library" was created. Here, the original MG1655 transposon library was spread onto plates to achieve single colonies and 60 colonies were picked and grown individually in a 96 well microtiter plate overnight at 37^{o}C. 50µl of overnight culture from each well was then pooled and the library and TraDIS was performed. Therefore, within this library only 60 insertions were present. If sequencing causes artefacts, more than 60 insert positions would be counted. A small volume of library was therefore sequenced, and the data analysed. Overall, this analysis yielded 17640 reads which when aligned to the genome corresponding to 169 insertion sites, an insertion frequency practically tripled compared to what the library consisted of. Therefore, further investigation of the insertion sites was performed. The first thing that was to be considered was the level of noise within the library, with 23 insertions only reporting to have 1 read count and 87 insertions with less than 10 read counts. In addition to this, 57 insertions sites had a read count over 100 reads suggesting that these insertions were the ones that were the genuine insertion positions in the library (Figure 5.5.2: Red line). However, determining the other 3 insertions was difficult as there was no clear boundary of read count to detect the other 3 other insertions.

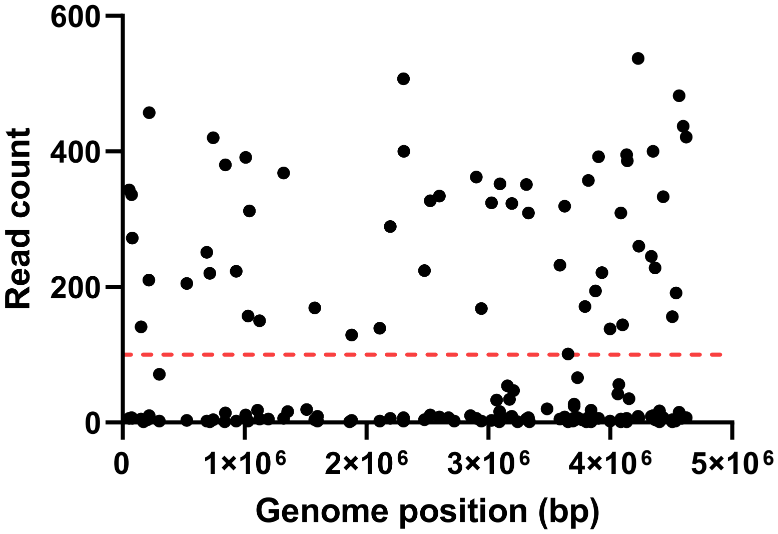


Figure 5.5.2: analysis of the “60 insertion” library, with the number of reads corresponding to each unique insertion site identified. Each point represents the location of the first base of a unique insertion site. The red line indicates 100 reads cut off, which identifies 57 insertions.

Since more insertion sites were reported than should be present in the library, a manual inspection of each one was then performed using the Artemis genome browser. Upon this inspection, a surprising pattern started to emerge, with two insertions present in the same location precisely 9bps from each other. This was seen for 60 different locations. In addition to this, all these insertions had the same pattern in read counts, with one insertion with a high read count, and the other lower with a low read count (Figure 5.5.3). A 9bp duplication event occurs upon transposition of the Tn5 transposon (Reznikoff *et al.*, 1999). Therefore, these two apparent sites could potentially represent reads referring to the flanking junction of a transposon.

Further to this, the orientation of the read alignment was also considered. This was done as previously reported in Goodall *et al.*, (2018). Since within the TraDIS analytical method only one flanking region of the transposon is sequenced, the orientation of the insertion can be deciphered based on the orientation of DNA strand that the read aligns to (Figure 5.5.4A). With regards to the unusual two peak insertion, it was always seen that these two peaks were represented apparent insertions in opposite orientations. In addition to this, the smaller peak was always downstream of the larger peak’s orientation of the transposon (Figure 5.5.3). This result therefore strongly suggests that the larger peak observed within this two insertion pattern corresponds to the actual insertion present within the 60 insertion library, while the second peak, which is 9bp downstream, relates to the other non-target transposon junction which is an artefact being introduced at some point in the generation of TraDIS data.

In addition to this, in 2/60 locations where this 9bp difference was identified, a further insertion site was identified between the two peaks, 1 bp or 2bp away from the insertion with the larger read count. In both cases these insertion sites only had 1 read assigned to them, in the same orientation as the insertion with the higher read count. Further analysis of this was not conducted as this was a rare event. Indeed, this observation has been reported before, with the literature suggesting a minimum read threshold to determine transposon insertions, in order to remove this noise (Zhao *et al*., 2017, Page *et al*., 2020).

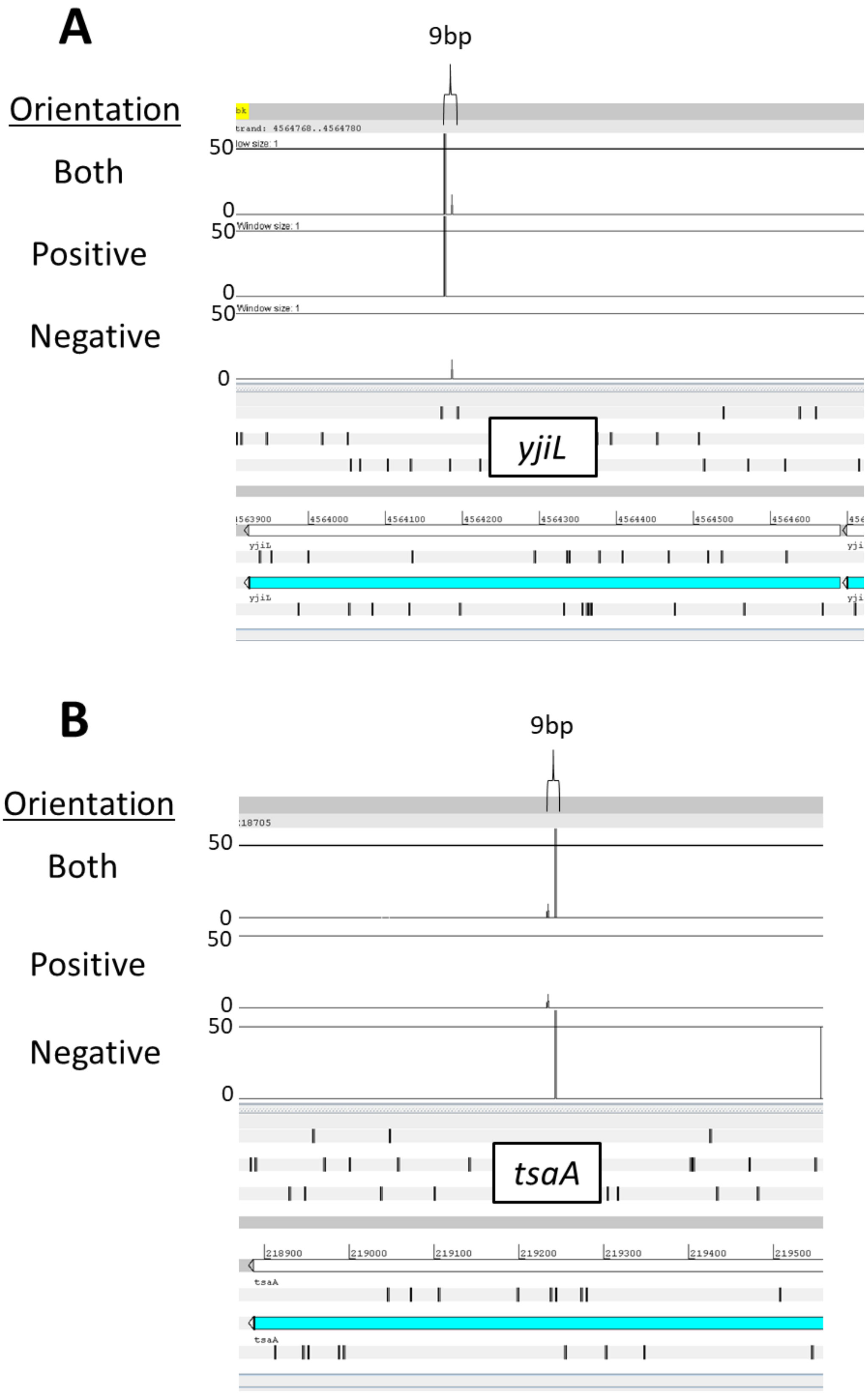


Figure 5.5.3: Examples of inserts in the 60 insert library, visualised in the Artemis genome browser. Histograms represent insertion site locations with two genes *yjiL* and *tsaA*. Positive and negative, refer to insertions present only in one orientation while both refers to both orientations combined together.

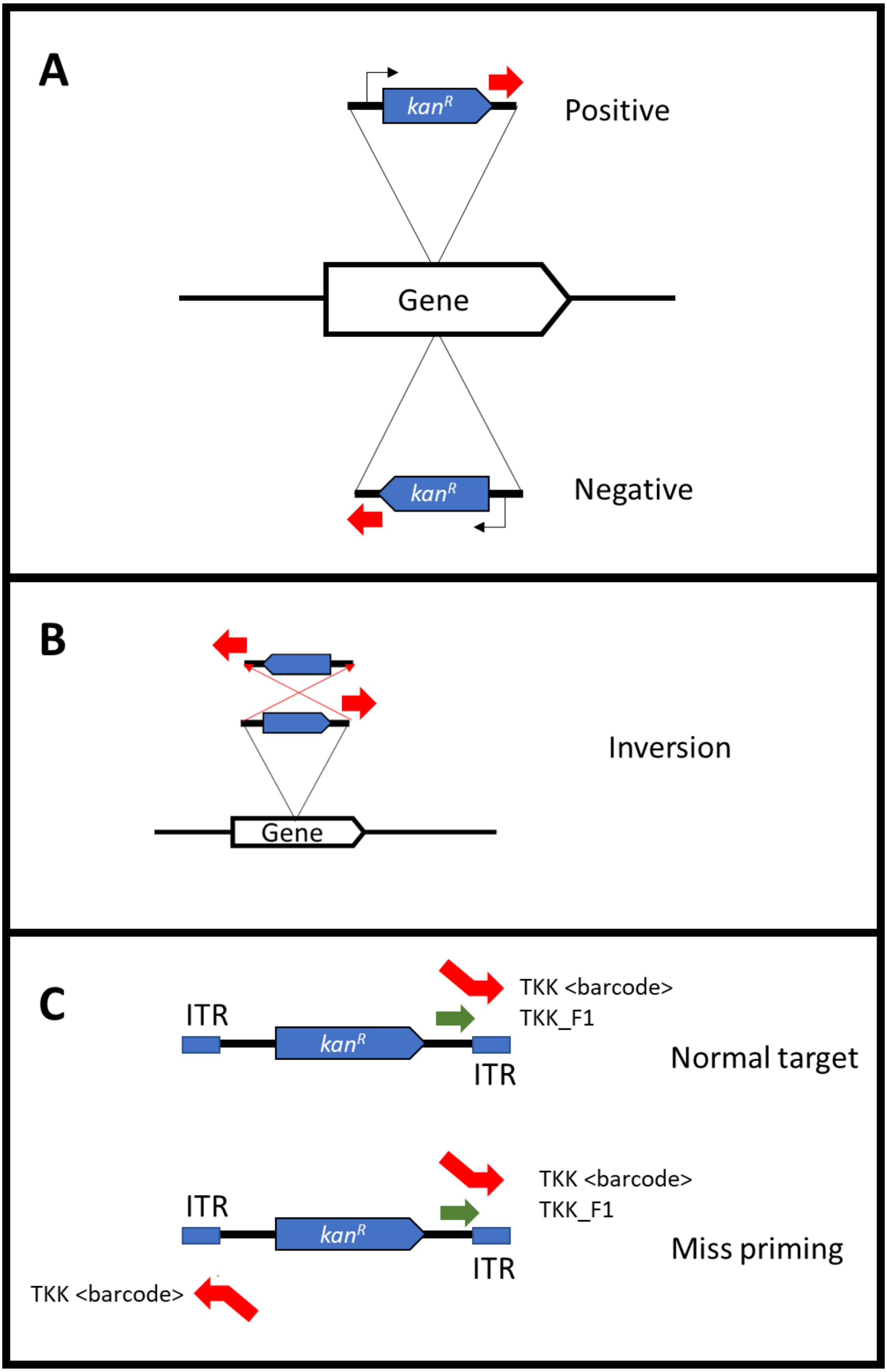

A
kan^R
Positive
Gene
kan^R
Negative
B
Gene
Inversion
C
ITR
kan^R
ITR
TKK <barcode>
TKK_F1
Normal target
ITR
kan^R
ITR
TKK <barcode>
TKK_F1
Miss priming
TKK <barcode>

Figure 5.5.4: Understanding the 60 insertion library. A) Schematic diagram demonstrating that different orientation of transposon insertions identified from the TraDIS analysis. The red arrow indicates the transposon junction sequenced. The orientation of insertion is termed positive and negative referring to the sense of the insertion based on the MG1655 genome. Two possible explanations to describe the two insertion pattern observed in the 60 insertion library are shown in (B) and (C). In (B), an inversion event occurs causing the transposon orientation to invert. Red arrows indicate transposon junction sequenced. In (C) The two-step PCR process used to amplify the transposon may be sometimes mis-priming from the inverted terminal repeat. Step one uses the primer TKK_F1 (Green arrow). Step two uses the TKK <barcode> primer (red arrow), of which 14bp binds to the inverted terminal repeat (ITL), therefore potentially mis-priming at the opposite transposon junction occurs.

Two explanations were considered for what might cause the two insertion observation described above, and these are shown in Figure 5.5.4. The first was that inversion events in the Tn5 transposon have previously been reported at low frequency with the orientation of the transposon inverting due to recombination (Weber *et al*., 1988, Weber, 1995). Therefore, the possibility exists that this second smaller peak could be reporting rare cases of the transposon having inverted. The second proposes an error during the two step PCR amplification of the TraDIS method (Section 2.7.2). These PCR steps enrich DNA fragments containing the transposon junction and genomic DNA in addition to providing additional sequences to prepare the library for sequencing. In the primer used in the second PCR step (designated TKK <barcode>) there is 14 bp homology at the 3' end which anneals within the 19bp inverted repeat of the transposon. Therefore the possibility exists that these primers can rarely mis-prime off the other inverted terminal repeat of the transposon (Figure 5.5.4), which would lead to the spurious identification of an insert in the reverse orientation to that which is actually the case.

Unfortunately, due to time limitations caused by the COVID-19 lockdown, further investigations could not be conducted on this in detail. However, in an attempt to see if inversion could be identified, a single transposon insertion mutation in *yjjY* isolated from the STSE was used (see Section 5.4.4). Using a PCR with a combination of flanking and transposon internal primers, it should be possible to see if inversion occurs in this strain. Using 6 overnight cultures of this *yjjY* insertion and a MG1655 control, PCR reactions using different primer combinations were performed. The results presented in Figure 5.5.5 show that transposon inversions could not be identified, however this does not completely rule out the possibility that inversions were still occurring as they could occur at lower frequency than the PCR method was able to detect in this limited study.

Overall, the result with the 60 insertion library experiment show that in a low complexity library an artefact is introduced which causes more insertions to be identified than are actually present. Whether this is biological (due to inversions occurring at small frequency), or technical (due to the way in which TraDIS sequencing libraries are generated in this study) is unclear. However, considering how this related to the observation within the STSE, this result suggests that the increase in insertion sites within particular gene is probably due to the same artefact observed within the 60 insertion library. The fact that this increase in insertions is observed within genes where high read counts are present supports this, since as these insertions are at a high frequency in the population, this increases the likelihood that either mis-priming or inversions would occur in these cases. It should be noted here that this observation does not rule out the possibility that not all insertions site within this library had been identified in the original library, and that within the STSE in these enrichments some additional insertion sites still might be identified which could also contribute to this phenomena. This results also highlights a potential flaw in that if this phenomenon is due to PCR mis-priming the relative frequency generated using read count may be overestimated due to this artefact.

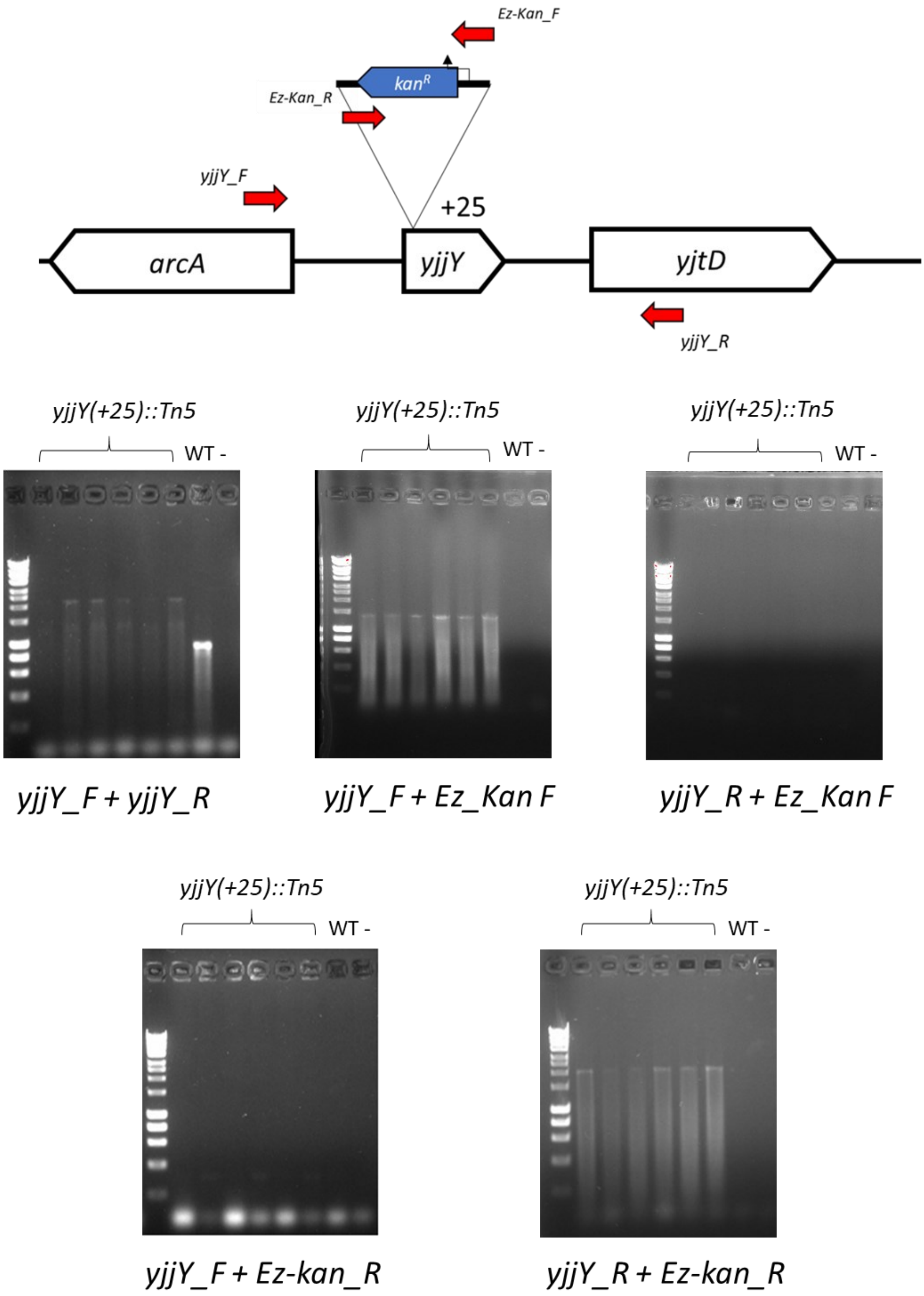


Figure 5.5.5: Attempting to identify inversion events occurring with one transposon insertion. A schematic of the primers used in the experiment is presented above. Six overnight cultures of MG1655 *yjjY(+25)::Tn5* and MG1655 control (WT), were set up and PCR reactions were performed using different primer pairs., Visualization of PCR products by gel electrophoresis was then done with the results seen. The DNA ladder used was Bioline 1kb hyper ladder, each primer pair used is describe at the bottom of the gel.

5.6 Analysis of the STSE

The sections above show that under different conditions of the STSE, there is a high degree of correlation between replicates although some differences are observed. Although the analysis above starts to identify genes of interest within each condition it is so far only qualitative, as it does not allow for proper identification of genes which are significantly different between each condition of the STSE. Therefore, to be able to compare different conditions within the STSE a more rigorous method of analysing the data was required. To this date, several pipelines exist which have been designed for transposon insertion sequencing. The comparison of these pipelines is discussed in Appendix 3.

When comparing transposon sequencing samples, a typical analysis is to compare relative frequencies of insertions in one sample against the other, to identify whether or not a given insertion has dropped or increased within a sample, which will give information about the impact on fitness of the insertion. Typically, this is done by considering all insertions within a gene using a normalized read count, as this is the most relevant for these studies. Comparison of read counts between samples then allows for the creation of a fold-change metric for each gene, which can then be used to describe whether insertions have significantly declined or increased in one sample relative the other (Figure 5.6.1). This then allows for two different types of genes to be identified in the comparison: genes containing insertions which become more common in the population due to the insertions causing an increase in fitness, or genes where insertions are detrimental to fitness and where the strains containing them hence decline in the population. From the comparison in Appendix 3, it was decided to use two different analysis methods. These were edgeR which was able to identify both advantageous fitness and detrimental fitness genes. However, when considering detrimental genes in the Day 1 of the STSE, edgeR ability to identify detrimental genes was limited, which was potentially due to this study not removing genes which had a zero read count. Therefore, a log-likelihood method, which considers the presence and absence of insertions only, was also used as it was able to better identify detrimental fitness genes.

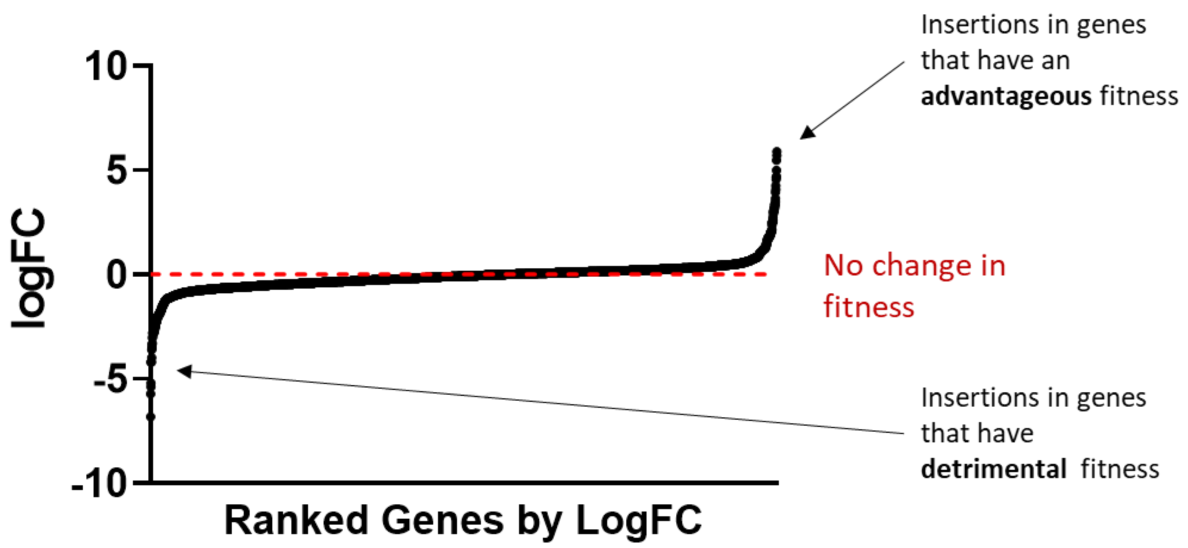


Figure 5.6.1: Overview of a typical transposon analysis, where insertion frequencies within genes are compared between samples to produce a Fold change (FC) metric which describes an increase or decrease of insertions. This then can be attributed to their effects on fitness, where insertions which decrease have a detrimental impact on fitness, and insertions which increase having a beneficial effect.

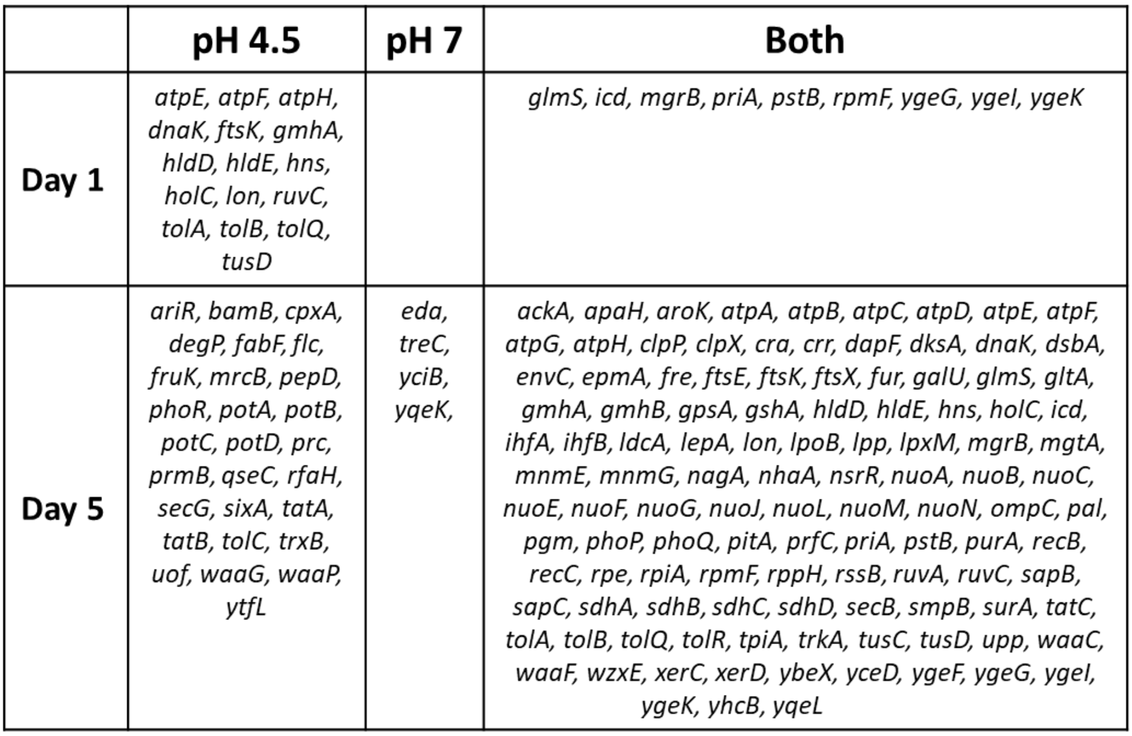
5.6.1 Genes where insertions show a fitness disadvantage in the STSE

As described in section 5.4.3, genes with insertions that were determinantal in fitness were identified for Day 1 and Day 5 only, since the amount of decline observed by Day 10 made it difficult to determine whether gene loss was for biological reasons, or due to stochastic loss. Once these gene lists for day 1 and day 5 were generated, since the conditions of the STSE were highly similar it was decided to compare these lists to see which genes were specific to either pH 7 or pH 4.5, and which were detected in both conditions. In Table 5.7, the results of the log likelihood analysis are shown. This analysis will typically indicate that strains with insertions in specific genes were becoming rarer in the population, to the point that the gene would be considered conditionally essential. In Table 5.8 the edgeR analysis indicates which genes are on the decline.

Overall, of the genes identified at day 1 in this analysis, no gene was specific to the pH7 condition only. Instead, the genes identified at pH 7 which showed a decline in insertions were also identified at pH 4.5. This result correlates with the expectation that under both pH conditions of the STSE, similar stresses were occurring. Of the genes identified within both these pH conditions, genes associated with metabolism (*glmS* and *icd*), regulation of the *phoPQ* two component system (*mgrB*), and DNA replication were identified (*priA)*. In addition to these three uncharacterized genes (*ygeG, ygeI* and *ygeK*) were also identified as having their function required for survival under the conditions of the STSE.

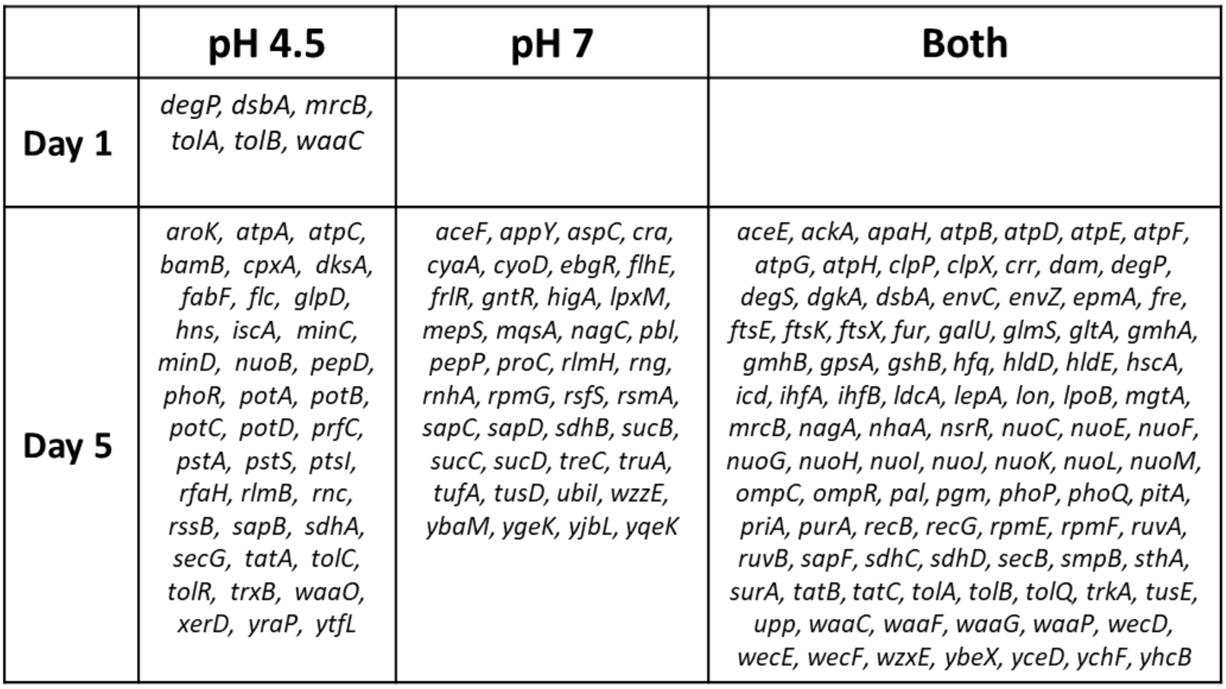
At pH 4.5 at day 1, however, a set of genes unique to this pH condition were identified. This indicated that the function of these genes was important specifically for being grown in unbuffered LB at pH 4.5. The function of these genes suggests that at pH 4.5 in unbuffered LB, cell division and outer membrane stability is of particular importance, with the genes *ftsK* and *tolABQ* genes being identified as having a reduction in overall insertion count. In addition to this, genes involved in generation of the LPS were also required (*gmhA*, waaC *hldDE*), as well as genes involved in stress response such as *dnaK* (a molecular chaperone) and genes involved in ATP synthase, *atpEFH*. The global regulator *hns* function was also shown to be required at pH 4.5 of the STSE; this regulator is associated with a variety stress responses. Overall, these genes appear to be important for growth at pH 4.5. However, at day 5 all these genes were also identified in the pH7 condition as well as the pH 4.5 condition, suggesting that these genes may play a role in fitness under the overall stress presented by growth in unbuffered LB.

Table 5.7: Genes identified as insertions with a detrimental fitness using the log likelihood method. Further information on these genes can be found in Supplementary Table S1



	pH 4.5	pH 7	Both
Day 1	*atpE, atpF, atpH, dnaK, ftsK, gmhA, hldD, hldE, hns, holC, lon, ruvC, tolA, tolB, tolQ, tusD*		*glmS, icd, mgrB, priA, pstB, rpmF, ygeG, ygeI, ygeK*
Day 5	*ariR, bamB, cpxA, degP, fabF, flc, fruK, mrcB, pepD, phoR, potA, potB, potC, potD, prc, prmB, qseC, rfaH, secG, sixA, tatA, tatB, tolC, trxB, uof, waaG, waaP, ytfL*	*eda, treC, yciB, yqeK,*	*ackA, apaH, aroK, atpA, atpB, atpC, atpD, atpE, atpF, atpG, atpH, clpP, clpX, cra, crr, dapF, dksA, dnaK, dsbA, envC, epmA, fre, ftsE, ftsK, ftsX, fur, galU, glmS, gltA, gmhA, gmhB, gpsA, gshA, hldD, hldE, hns, holC, icd, ihfA, ihfB, ldcA, lepA, lon, lpoB, lpp, lpxM, mgrB, mgtA, mnmE, mnmG, nagA, nhaA, nsrR, nuoA, nuoB, nuoC, nuoE, nuoF, nuoG, nuoJ, nuoL, nuoM, nuoN, ompC, pal, pgm, phoP, phoQ, pitA, prfC, priA, pstB, purA, recB, recC, rpe, rpiA, rpmF, rppH, rssB, ruvA, ruvC, sapB, sapC, sdhA, sdhB, sdhC, sdhD, secB, smpB, surA, tatC, tolA, tolB, tolQ, tolR, tpiA, trkA, tusC, tusD, upp, waaC, waaF, wzxE, xerC, xerD, ybeX, yceD, ygeF, ygeG, ygeI, ygeK, yhcB, yqeL*

Table 5.8: Genes identified as insertions with a detrimental fitness using edgeR. Further information on these genes can be found in Supplementary Table S2

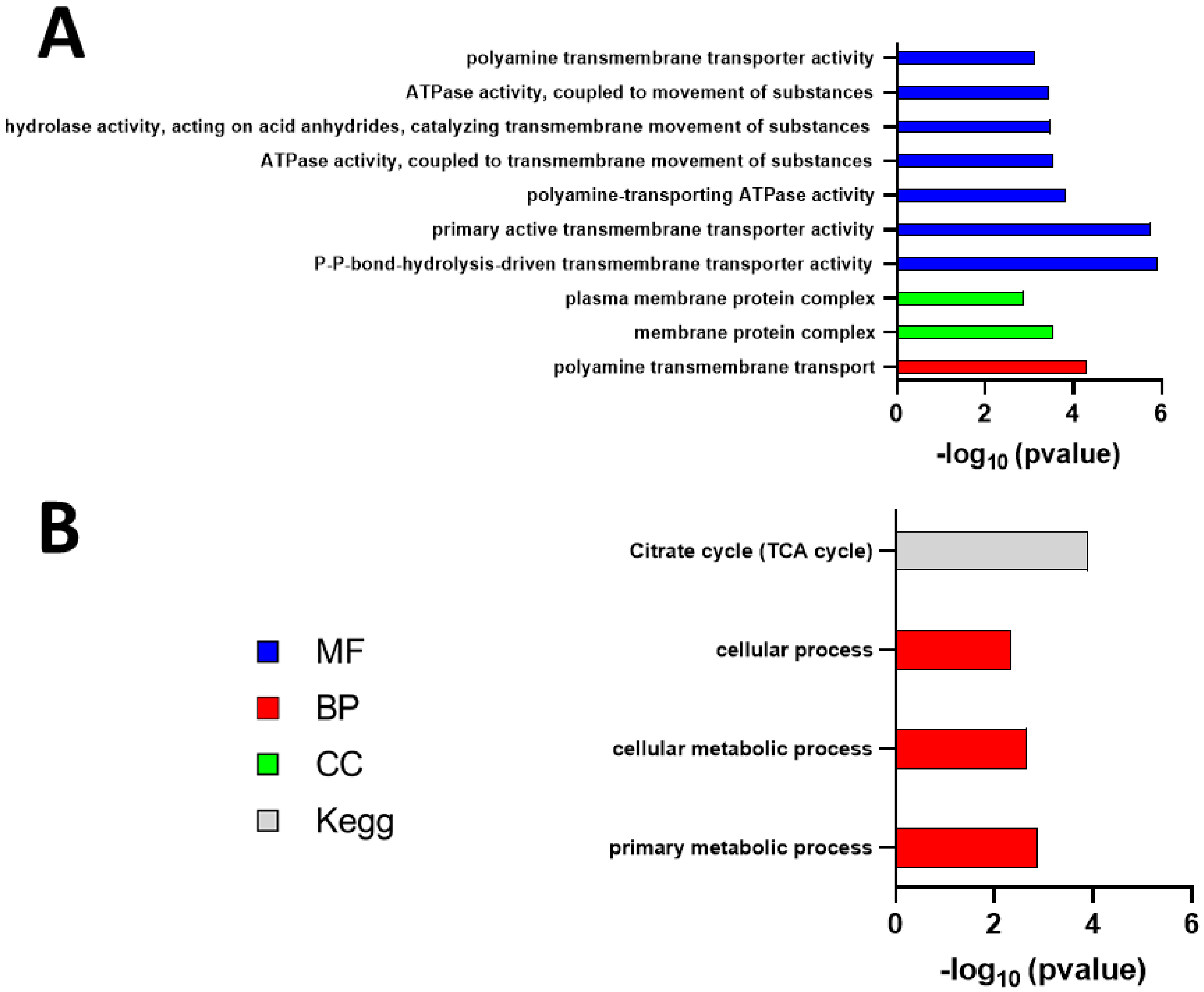


	pH 4.5	pH 7	Both
Day 1	*degP, dsbA, mrcB, tolA, tolB, waaC*		
Day 5	*aroK, atpA, atpC, bamB, cpxA, dksA, fabF, flc, glpD, hns, iscA, minC, minD, nuoB, pepD, phoR, potA, potB, potC, potD, prfC, pstA, pstS, ptsI, rfaH, rlmB, rnc, rssB, sapB, sdhA, secG, tatA, tolC, tolR, trxB, waaO, xerD, yraP, ytfL*	*aceF, appY, aspC, cra, cyaA, cyoD, ebgR, flhE, frlR, gntR, higA, lpxM, mepS, mqsA, nagC, pbl, pepP, proC, rlmH, rng, rnhA, rpmG, rsfS, rsmA, sapC, sapD, sdhB, sucB, sucC, sucD, treC, truA, tufA, tusD, ubiI, wzzE, ybaM, ygeK, yjbL, yqeK*	*aceE, ackA, apaH, atpB, atpD, atpE, atpF, atpG, atpH, clpP, clpX, crr, dam, degP, degS, dgkA, dsbA, envC, envZ, epmA, fre, ftsE, ftsK, ftsX, fur, galU, glmS, gltA, gmhA, gmhB, gpsA, gshB, hfq, hldD, hldE, hscA, icd, ihfA, ihfB, ldcA, lepA, lon, lpoB, mgtA, mrcB, nagA, nhaA, nsrR, nuoC, nuoE, nuoF, nuoG, nuoH, nuoI, nuoJ, nuoK, nuoL, nuoM, ompC, ompR, pal, pgm, phoP, phoQ, pitA, priA, purA, recB, recG, rpmE, rpmF, ruvA, ruvB, sapF, sdhC, sdhD, secB, smpB, sthA, surA, tatB, tatC, tolA, tolB, tolQ, trkA, tusE, upp, waaC, waaF, waaG, waaP, wecD, wecE, wecF, wzxE, ybeX, yceD, ychF, yhcB*

At day 5, as was expected, the number of genes identified as being detrimental to fitness when inactivated by insertion increased dramatically. More genes were identified as being specific to either pH 7 or pH 4.5 as well as within both conditions. Interestingly at pH 7, more specific genes were identified with the edgeR analysis than with the log-likelihood analysis, indicating that in the majority of these genes, insertions were declining in frequency however, had not been completely lost from the population. To begin to consider how the role of these genes for the cell at Day 5 of the STSE, the gene lists present in Table 5.7 and Table 5.8 were combined and gene ontology and Kegg pathway enrichment analysis was performed on each (Figure 5.6.2).

The results of this analysis show that at pH 4.5 genes identified were associated with active transport of substances across the membrane, in particular polyamines (Figure 5.6.2A). Further focus upon pH 4.5 revealed that all insertions declined in the genes *potABCD* which encode a high affinity transporter of polyamines, specifically spermidine. Spermidine has been shown to be implicated in acid resistance by making the membrane less permeable (Yohannes *et al*., 2005). Additionally polyamines have also shown to be implicated in the proliferation and viability of the cell and have been shown to be involved in the expression of RpoS (Igarashi and Kashiwagi, 2018).

Of the genes at pH7, not many terms were classed as enriched (Figure 5.6.2B). However, the terms that were identified were often of genes with a function in central metabolism. Of the genes identified in both conditions, the analysis suggested that these genes were associated also with central metabolism and overall energy generation (Figure 5.6.2). This result suggests that under both conditions there is still a need to maintain overall metabolism and energy generation, which would be expected. It was also obvious that strains with insertions in the NADH dehydrogenase I operon *nuo* as well as the ATP synthase genes *atp* all decline at day 5.



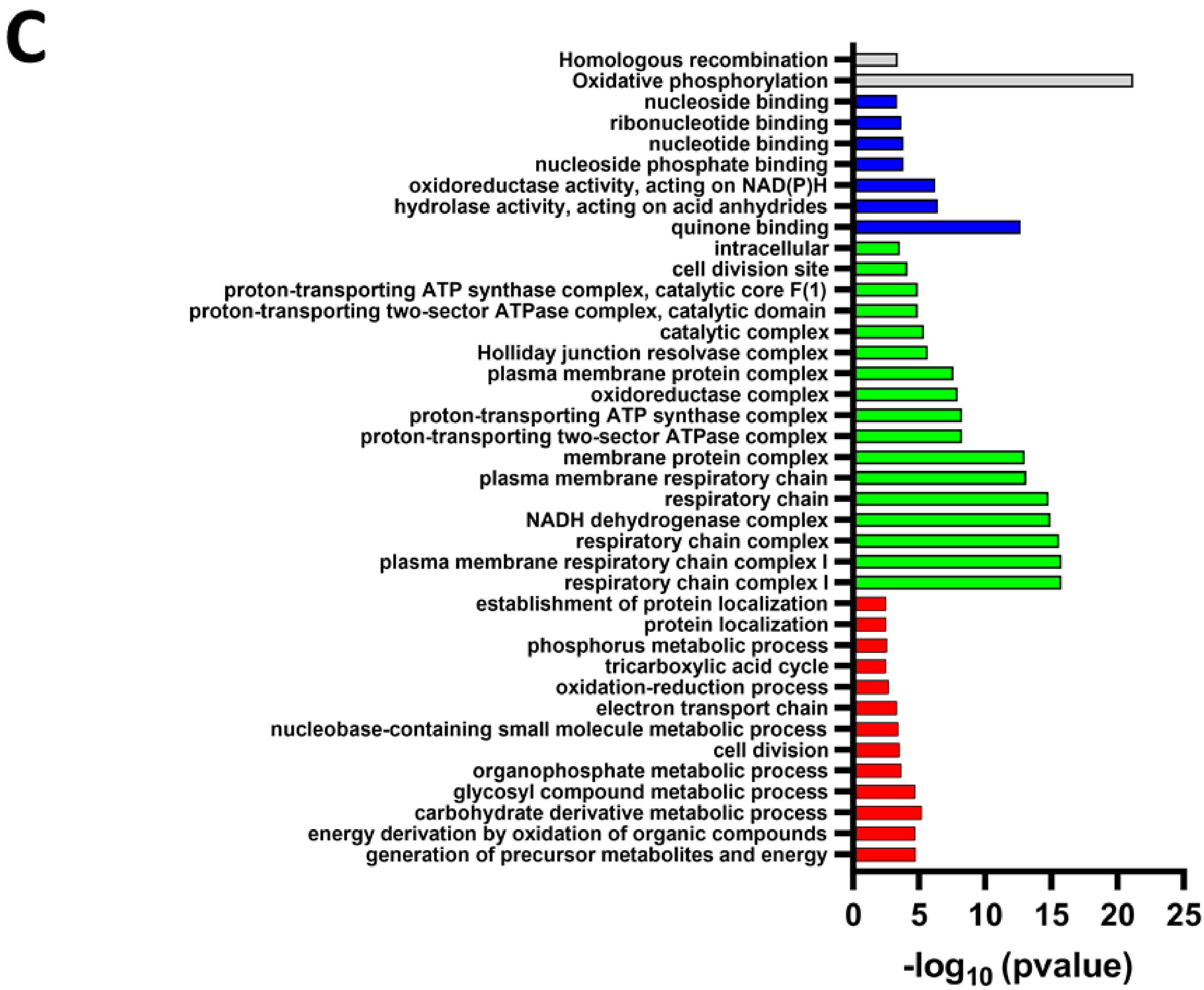


Figure 5.6.2: Results of Gene ontology and Kegg pathway enrichment analysis using DAVID, for genes identified as insertion detrimental at Day 5 in either the log likelihood analysis or edgeR analysis. Bars are colour coded to represent GO terms annotations from Molecular Function (MF), Biological Process (BP) and Cellular Component (CC) ontologies in addition to Kegg pathway (Kegg). Gene lists were split based on whether they were identified in pH 4.5 (A), pH 7 (B), or both conditions of the STSE (C). Functional annotation was considered significant using a FDR < 0.05 after a Benjamini Hochberg procedure was performed. C only the lower levels of the GO ontology are presented.

5.6.2 Conditional enrichments within the Transposon library

The second and for this study more important focus of the STSE analysis was to identify genes where insertions appeared to cause a gain in fitness. As previously reported within Section 5.4, at Day 10 and to some extent day 5, insertions within genes were starting to be identified as accumulating within the population in the STSE replicates. The genes identified as showing a significant accumulation of insertions (based on using edgeR analysis) are shown in Table 5.9.

Table 5.9: Genes identified as having a significant accumulation of insertions under the conditions of the STSE. No enrichments were observed at Day 1. Note that essential genes were not removed in this analysis. Genes associated with transcription regulation are underlined. Further information on these genes can be found in Supplementary Table S3

	pH 4.5	**pH 7**	**Both**
Day 5	*yjiG, der*	*sspA*	*yjjY, cspC, arcB, yobF, ptsP, proQ, slyA, treR, apaG*
Day 10	*ptsP, cadC, treR, cadB, dusB, ptsO, rpoC, yjdQ, ytfE*	*fimE, sspA, iscR, panB, yfbM, yfbK, creD, rnlA, rpoS, gadE, wcaJ, gadX, rpoZ, sspB, rbsR, uidC, cytR, yaiA, tnaA*	*yjjY, arcA, cspC, ygbE, proQ, yobF, slyA, arcB, rcsB, dnaK, apaG, der*

These results identified a core set of genes at day 5 that showed insertion accumulation at both pH 4.5 and pH 7, while only a few enrichments were observed that were specific to the pH condition. This changed at day 10, where more enrichments were identified that were specific to the pH condition, as well as under both conditions. A sizable proportion of these genes were associated with regulation, either being subunits of RNA polymerase, or a regulator of transcription (Table 5.9. underlined). These insertions would be expected to have effects on gene expression in each strain. These genes, and others identified as enriched, are discussed in the section below.

5.6.2.1 Differences between day 5 and day 10

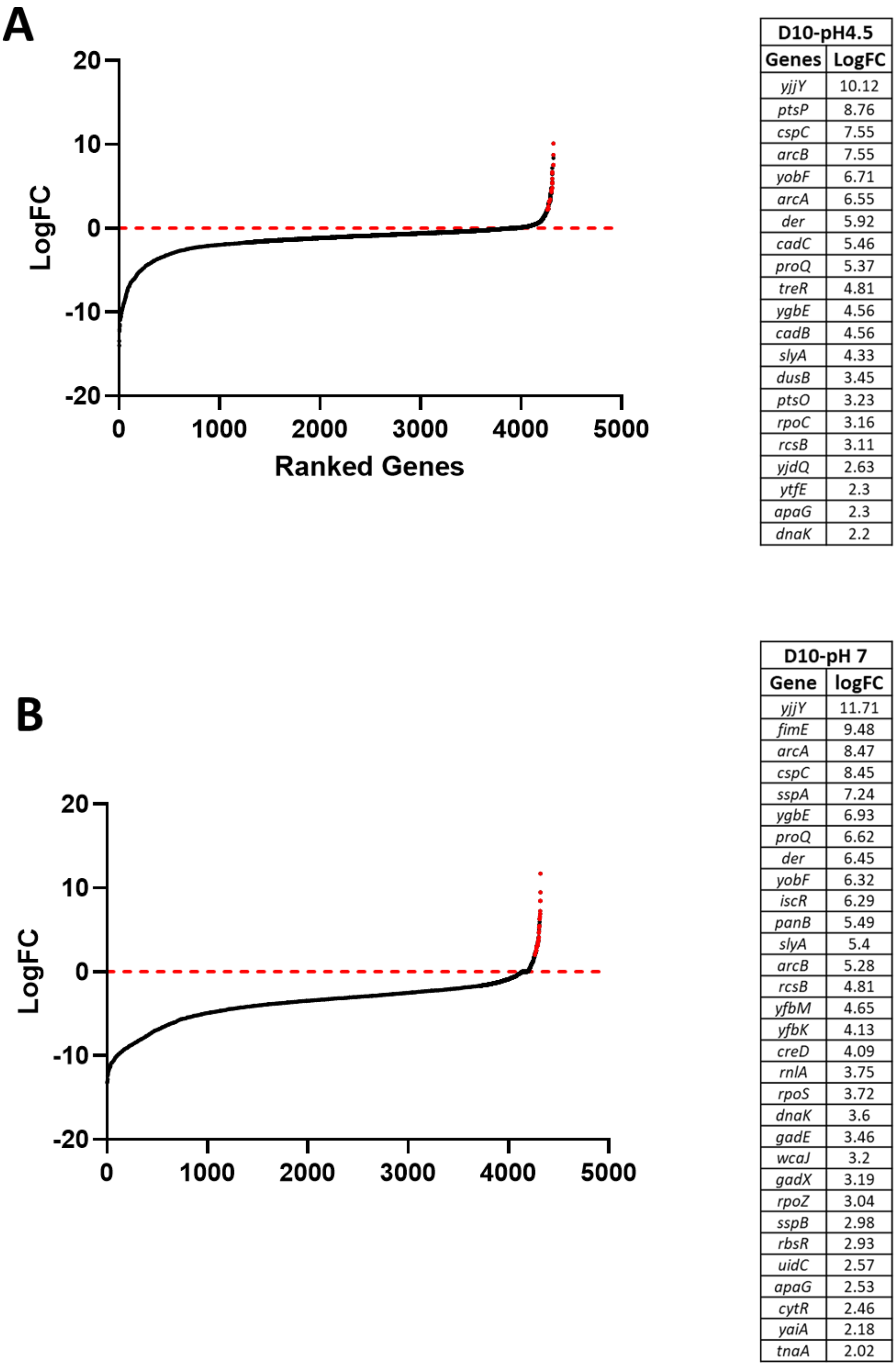
The majority of genes containing insertions that were identified as enriched at Day 5 also were enriched at Day 10. However, one gene, *yjiG* (which encodes a predicted inner membrane protein) was enriched at Day 5 pH 4.5 but then at Day 10 was not identified. Further investigation of this gene showed that at Day 10 this gene had actually declined in overall abundance when compared back to the original ITL. Why this gene did not accumulate further during the STSE is unknown, but this suggests a complexity in population dynamics, which could not be captured with the current timepoint resolution of the STSE.

In addition to these two other genes, *ptsP* (part of the nitrogen phosphotransferase system), and *treR* (the repressor of trehalose transport and degradation), were found to be enriched at both pHs at Day 5, but at day 10 were only enriched at pH 4.5. Further focus on these genes showed that insertions within these genes were still present within the library at Day 10 -pH7 although at a low frequency. There by potentially indicating that they still had fitness, however to a lesser extent than some of the enrichments seen. Therefore, potentially insertions within these genes are still fit at pH 7, although the extent of the fitness observed is not enough to be identified as significantly different, within the analysis.

5.6.2.2 Enrichments observed differ in magnitude

As previously reported in Section 5.4.2.2, the extent of normalized read count differed between genes, showing that insertions within different genes accumulated to different amounts. Although the table above indicates the genes which have been classed as enriched within this study, it does not quantify the extent of fitness differences caused by insertions in the different genes. One metric which can be used to describe a relative quantification of fitness is the LogFC, describing the Fold change difference of normalized reads under the conditions of the STSE compared to the ITL. Therefore to show this, the LogFC was calculated for the two pH conditions at Day 10 of the STSE using edgeR, and this data is presented in Figure 5.6.3.

Considering the results presented in Figure 5 6.3, the overall effect of insertion loss at Day 10. It was clear to see that for the majority genes the insertions showed a fitness disadvantage, reporting a LogFC < 0. The extensiveness of this insertion drop could be clearly seen at Day 10 pH 7 with majority of genes reporting a LogFC change less than 0 (B). This was also true at Day 10 pH 4.5, which also saw a loss of overall insertion frequency in most genes but to a lesser extent (A). This further highlights the observation described in Section 5.4, where the majority of the population is defined by enrichments within only a few genes



D10-pH4.5	
Genes	**LogFC**
yjjY	10.12
ptsP	8.76
cspC	7.55
arcB	7.55
yobF	6.71
arcA	6.55
der	5.92
cadC	5.46
proQ	5.37
treR	4.81
ygbE	4.56
cadB	4.56
slyA	4.33
dusB	3.45
ptsO	3.23
rpoC	3.16
rcsB	3.11
yjdQ	2.63
ytfE	2.3
apaG	2.3
dnaK	2.2

D10-pH 7	
Gene	**logFC**
yjjY	11.71
fimE	9.48
arcA	8.47
cspC	8.45
sspA	7.24
ygbE	6.93
proQ	6.62
der	6.45
yobF	6.32
iscR	6.29
panB	5.49
slyA	5.4
arcB	5.28
rcsB	4.81
yfbM	4.65
yfbK	4.13
creD	4.09
rnlA	3.75
rpoS	3.72
dnaK	3.6
gadE	3.46
wcaJ	3.2
gadX	3.19
rpoZ	3.04
sspB	2.98
rbsR	2.93
uidC	2.57
apaG	2.53
cytR	2.46
yaiA	2.18
tnaA	2.02

Figure 5.6.3: Results of edgeR plotted for Day 10 conditions of the STSE at pH 4.5 (A) or pH7 compared against the ITL (B) identifying genes which insertions had accumulated in the population over the 10 days in this study . Only genes which had a positive LogFC values and were classed as statistically significant are coloured red with the table on the right reporting the LogFC values of these genes. Genes which were identified as statistically significant with a negative LogFC value are not highlighted.

These results show that under both conditions, insertions within *yjjY* show the largest fitness increase compared to any other gene, and this is true at both pH 7 and pH 4.5, although this represents the averaged data across the replicates; analysis of the individual replicates showed a slightly different story in regard to the observation within different replicates (Section 5.4.2.2).

5.6.2.3 The Arc Two component system- *arcA*/*arcB* and *yjjY*.

arcA and *arcB* encode a two-component system involved in switching gene expression in *E. coli* as it moves from aerobic to anaerobic growth (Lee *et al*., 2001, Perrenoud and Sauer, 2005). In addition to this, the relationship between *yjjY* and the ArcAB two component system is that *yjjY* is within the large promoter region of *arcA* (Figure 5.6.4). Within all conditions of the STSE, an increase in insertions in *yjjY*, *arcB* and *arcA* was observed at Day 10, with insertions in *yjjY* showing overall the largest increase in fitness in both condition at Day 10 (). In addition, *yjjY* also showed an accumulation of insertions at Day 5. Therefore, under the conditions of the STSE, insertions in *yjjY* appear to have a large fitness advantage.

A more detailed manual inspection of the insertion position genome showed that although insertions in *arcA* are significantly enriched, this was due to insertions enriched at the 3' and 5' ends of gene, with no insertions present in the middle of *arcA* (Figure 5.6.4A). It was also noticed that there was a large enrichment of insertions upstream of *arcA* within the predicted gene, *yjjY* (Figure 5.6.4A), and these were found to be not confined to the *yjjY* gene, but in the location of the promoter. Therefore, this observation suggested that insertions seen in *arcA* did not cause loss of function and that perhaps in strains carrying these insertions, alternative regulation of *arcA* was occurring to elicit the phenotype which provides a fitness advantage in the STSE.

Additionally as explained before in section 5.5, , transcription and translation has been shown to occur from one end of the mini-Tn5 transposon used to construct the library (Goodall *et al*., 2018). Therefore, it was important to also consider the orientation of the transposon insertions, to see whether polar effects can be observed. By visualizing orientation of insertions in Artemis, an enrichment of transposon insertions in one orientation could be clearly seen in the upstream region of *arc*A (Figure 5.6.4 C+D). The orientation of this insertion enrichment was in the direction which would promote readthrough from the transposon into the *arcA* gene. This orientation bias was seen both in populations grown at pH 4.5 and pH 7 in all three replicates at Day 10 (Figure 5.6.4 C+D). This suggested that the fitness advantage seen due to the accumulation of the insertions in *arcA* and *yjjY* is not eliciting loss of function, but instead it indicates that the function of ArcA needs to be maintained, although the alternative regulation caused by the insertions leads to a fitness advantage in the STSE. Corresponding to our previous findings that in competition experiments in Figure 3.7.2 + 3.7.3 which show loss of function of *arcA* causes a disadvantageous fitness under the conditions of pH 4.5 and pH 7, while mutations identified in *arcA* in the evolution experiment confer advantagous fitness (Figure 3.7.3).

Since an accumulation of insertions was also seen with *arcB*, this was also manually inspected. The result presented in Figure 5.6.4 B revealed no unusual insertion or orientation pattern, just an accumulation of insertions throughout the gene at both pH conditions at day 10. The fact that insertions within *arcB* (which encodes a histidine kinase specific for ArcA) also increased suggests that phosphorylation of ArcA may not be required to lead to a fitness advantage.

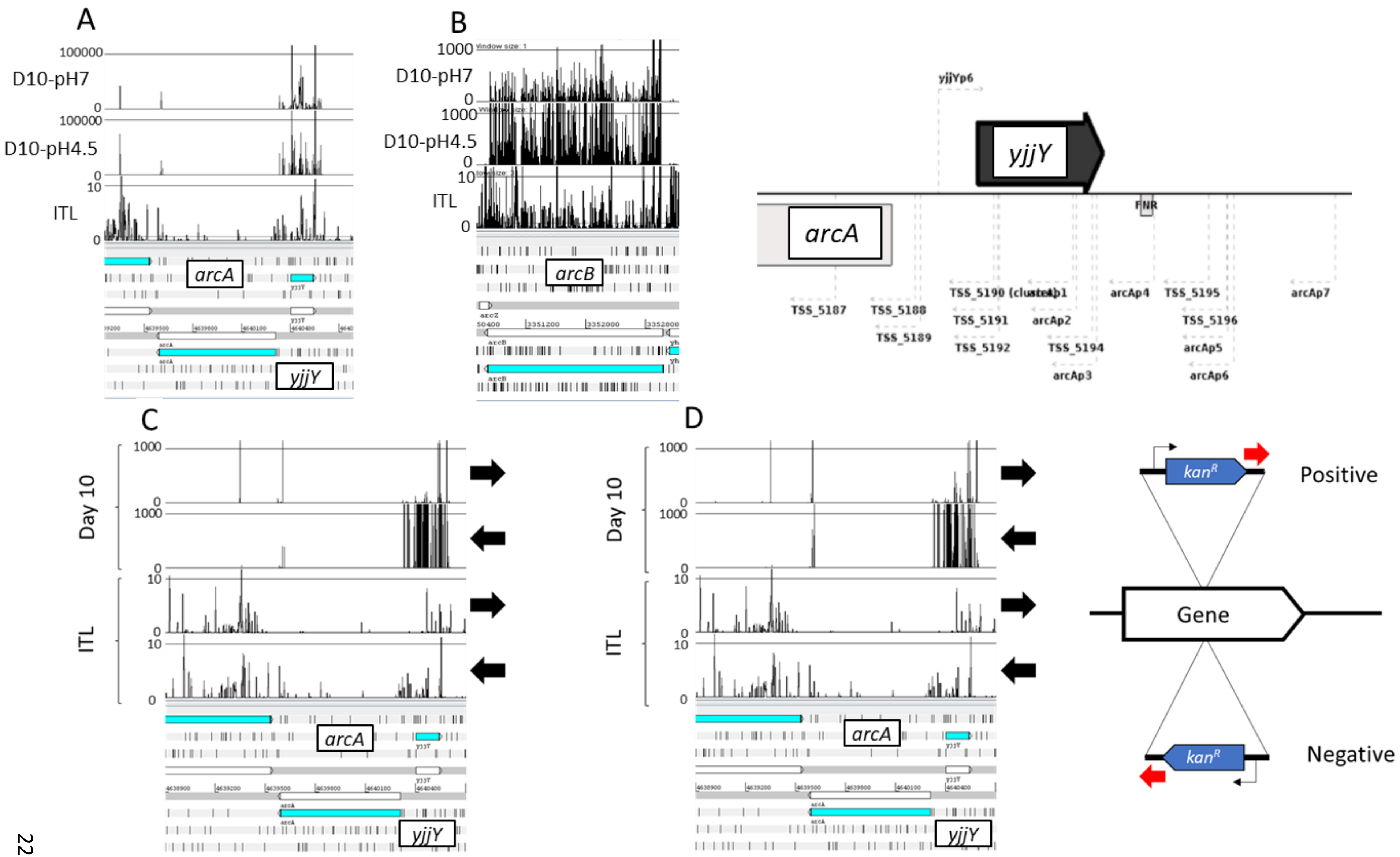

A
100000
D10-pH7
0
100000
D10-pH4.5
0
10
ITL
0
arcA
yjjY
B
1000
D10-pH7
0
1000
D10-pH4.5
0
10
ITL
0
arcB
yjjYp6
yjjY
arcA
FNR
TSS_5187
TSS_5188
TSS_5189
TSS_5190 (cluster)
arcAp1
TSS_5191
TSS_5192
arcAp2
TSS_5194
arcAp3
arcAp4
TSS_5195
TSS_5196
arcAp5
arcAp6
arcAp7
C
Day 10
1000
0
1000
0
ITL
10
0
10
0
arcA
yjjY
D
Day 10
1000
0
1000
0
ITL
10
0
10
0
arcA
yjjY
kanR
Positive
Gene
kanR
Negative

Figure 5.6.4: Unique insertion pattern in the two component system *arcAB* visualized in Artemis. Examples are provided which occur in all three replicates. Initially, insertions in sample Day 10 pH7 and pH4.5 and ITL were visualized (Top track to bottom track respectively) for *arcA* (A) and *arcB* (B). Insertion counts were then split between orientation, for Day 10 pH 4. 5 (C) and Day 10 pH 7 (D). Orientation is defined by the arrows; with arrow indicating the direction of transcriptional and translational readthrough of the transposon. This is highlighted in the diagram bottom right which highlights the differences in orientation where the red arrows show the direction of this readthrough. The transcription sites are taken from Regulon DB for *yjjY* and *arcA* is shown top right.

5.6.2.4 The AR response regulators

A lot of genes associated with *E. coli* acid response mechanisms were identified as showing enrichment, particularly genes from the amino acid dependent acid resistance mechanisms. In particular, the AR4 main regulator *cadC* and antiporter *cadB* were identified under pH 4.5 conditions at Day 10. Interestingly, loss of function of *cadC* has also been seen a previous evolution experiment conducted in buffered LB at pH 4.6 (Harden *et al*., 2015, He *et al*., 2017).In addition to this, regulators of the AR2 response (*gadE, gadX* and *rcsB*) were shown to have accumulated insertions at pH 7, and *rcsB* was also shown to be enriched at pH 4.5. All these genes have been shown to directly activate the AR2 mechanism encoded by *gadA, gadB and gadC*. The accumulation of insertions within these genes may lead to similar conclusion described in He *et al.,* (2018) that the AR responses, which are required for survival under extreme acid, may be detrimental for prolonged growth under mild stress.

5.6.2.5 Insertions associated with RNA polymerase

Three genes (*rpoC, rpoZ* and *rpoS*) were noted to show an accumulation of insertions in the STSE at Day 10. All these genes are associated with RNA polymerase. However, what was unusual about this result was that *rpoC* is a known essential gene, which encodes the RNA polymerase subunit β’, yet within the STSE, it was reported show increased insertions within the population. Since it is essential, no insertions should occur within this gene, so a manual inspection of this gene was conducted. This revealed that in all three replicates insertions were found within the 3’ end of *rpoC* which were enriched (Figure 5.6.5C). Within evolution experiments, mutations in genes for different sub-units of the RNA polymerase complex are quite common, indeed within this study the evolution experiment by Sen, (2018) revealed mutation in *rpoA* and *rpoD*. In addition to this some mutations in *rpoC* have been shown to confer a fitness advantage towards particular stresses including at mild acid stress (Harden *et al*., 2015, Knöppel *et al*., 2018, Du *et al*., 2020). Therefore, these insertions within the 3' end of *rpoC* presumably do not prevent RNA polymerase from functioning and could also elicit a fitness advantage.

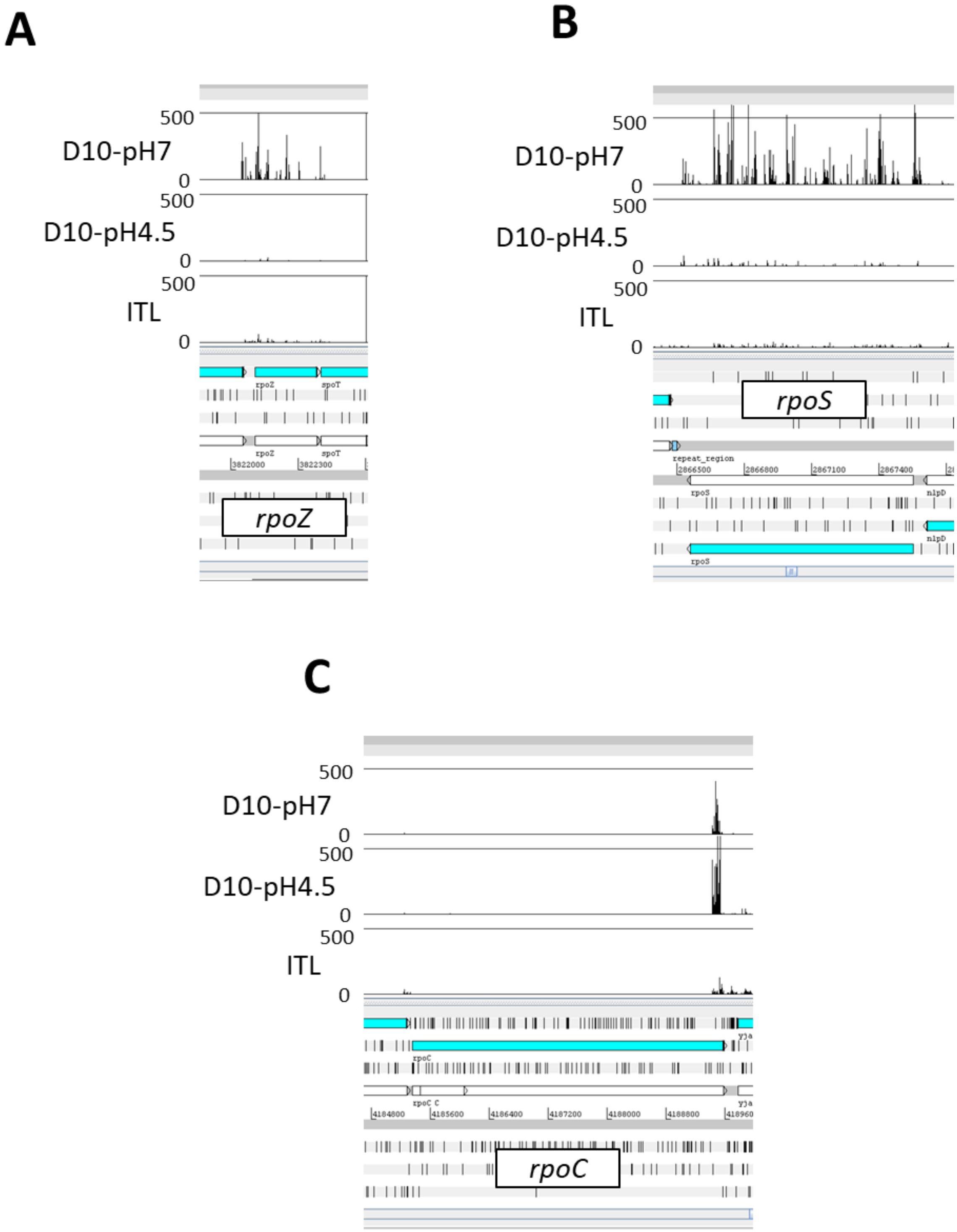


Figure 5.6.5: Genes associated with the RNA polymerase complex that accumulate insertions during the STSE. A) *rpoZ,* B) *rpoS,* c) *rpoC*. The scale of each histogram is shown on the left of each graph.

Additionally, insertions were significantly enriched within *rpoZ* which encodes the RNA polymerase subunit ω (Figure 14 B). This non-essential gene is part of the RNA polymerase complex and deletion of this gene can lead to global changes in transcription (Geertz *et al*., 2011). In addition, insertions in the *rpoS* sigma factor gene were also found to be enriched at the D10-pH7 time point of the STSE. Interestingly, *rpoS* mutations have found to provide a fitness advantage in nutrient-limited slow growing environments, although they are more sensitive to other stressors, and these loss of function transposon insertions may provide a fitness phenotype within the STSE only at pH 7 (Notley-McRobb *et al*., 2002).

5.6.2.6 der – single insertion is enriched

Another essential gene with enriched insertions under both conditions of the STSE at Day 10 and Day 5 was *der.* Der is a GTPase, which is required for stability of the large ribosomal subunit and therefore is essential for growth. However, within our data, an enrichment of insertions was observed within *der*. Since this should not be happening a manual inspection of the *der* for insertions was conducted using the Artemis browser (Figure 5.6.6). This revealed that within the *der* gene a single insertion at +954bp from the translational start site was seen. This insertion had an approximately 100 fold increased read count at Day 10 in both conditions of the STSE, implying that strains with this insertion were fitter in the STSE. This is unlikely to be due to PCR bias, as this same increase was observed within all replicates of the STSE at day 10. In addition to this, the library prepared by Goodall *et al.* (2018) in *E. coli* K-12 BW25113 strains also saw an insertion at this precise location. Therefore, these results suggest that this is a real insertion.

Der has two GTP binding domains, and disruption of either one (but not both) can still allow Der to function and growth to be observed, although it creates a temperature sensitive strain, able to grow at 42°C and not at 30°C (Hwang and Inouye, 2006). The insertion identified in the transposon library is 6 amino acids away from the 2nd GTP binding site (321-324aa), and does not cause loss of function, so this suggests it only disrupts only 1 GTP binding site. Why this insertion provides a fitness advantage in the STSE is unknown.

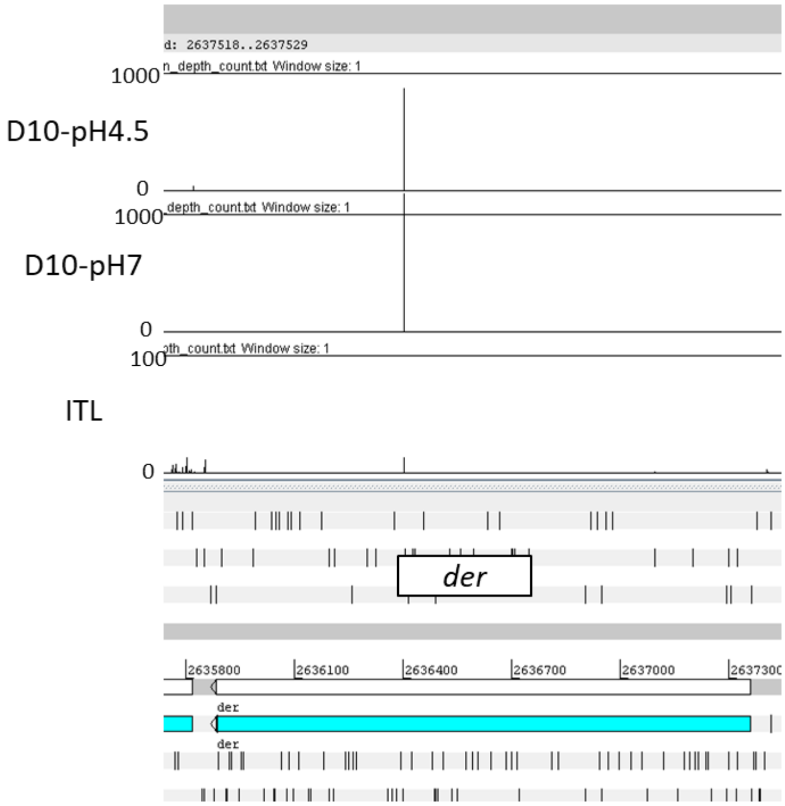


Figure 5.6.6: Position of insertion with the essential gene *der* shown using Artemis. Examples of pH 4.5 and pH7 at Day 10 of the STSE show that the single insertion present within *der* is enriched relative to the ITL.

5.6.2.7 Genes associated with stationary phase in LB

As shown above, strains with insertions within *rpoS* accumulated in the population at D10-pH7. RpoS is a sigma factor which is a key regulator of transcription in stationary phase, so this suggested that a fitness advantage may be caused by some insertions in genes which are involved in regulating gene expression in stationary phase. In addition to *rpoS,* insertions within the *tnaA* gene were also found to be enriched at pH 7. These genes encode tryptophanase which is involved in the synthesis of indole and pyruvate from tryptophan. Indole is an important signalling molecule. In LB broth when cells reach a high concentration, a pulse of indole occurs just before the bacteria enter stationary phase. It is thought that this pulse inhibits cell growth and division (Gaimster *et al*., 2014, Gaimster and Summers, 2015). Due to this, loss of function mutations in *tnaA*, which do not produce an indole pulse, have been shown to move into stationary phase later than their WT counterparts. In addition to this other genes such as *cytR* and *sspA* were identified in the enrichment analysis for cells grown at pH 7, and *yobF* and *cspC* for cells grown under both pH conditions; all these have been shown to play a role in long term stationary phase adaptation in LB (Kram *et al*., 2017, Gross *et al*., 2020). During the STSE each culture will be in stationary phase for the majority of the time, so the fact that these insertions are enriched is consistent with these earlier findings.

5.7 Comparison of the results of the STSE to the results of the evolution experiment

One of the key aims for doing the STSE was to determine the extent to which the results seen in a short-term evolution of a transposon library resembled the results seen in a normal evolution experiment of a longer length. As discussed previously, the growth conditions in the STSE were chosen so that a comparison could be made to the previous evolution experiment conducted in LB at pH 4.5 for 5 months described in Sen (2018) (Chapter 3). This evolution experiment was only performed at pH 4.5, with no additional control performed at pH 7. In the following section a comparison of the two data sets will be conducted.

5.7.1 Comparison to the evolved strains.

In chapter 3, the majority of the work described towards understanding fitness in the longer term evolution experiment was with the evolved strains E1A-E5A, each of which represented a single clone isolated from each of the five independent populations of Sen's evolution experiment. Here, this data is compared to the results from the STSE. Only two genes found in the evolved clones were identified in the "insertion detrimental" list from the STSE; these were isocitrate dehydrogenase, *icd*, which was identified on Day 1 of the STSE under both conditions and was part of the large deletion observed in E1A, and *priA, for* which insertions declined in both conditions of the STSE by day 5. The mutations associated with *priA* in the Sen experiment were all IS5 insertions downstream of the stop codon of *priA* and are hence more likely to have an effect upon on the next gene downstream, *cytR,* due to the insertion being within its promoter (Section 3.4.1.3). Of the large deletions seen in strains E1A and E4A, apart from *icd* mentioned above, none of the genes involved showed either a decline or an increase in insertions in the STSE.

The remaining genes which were identified in the evolved strains were then investigated to see how well they correlated to genes where insertions led to a fitness advantage in the STSE. To begin this comparison, the question of how to consider mutations present within the intergenic regions had to be considered. It is most likely that a mutation in an intergenic region would have an effect if it was within the promoter region of a gene. Using this assumption, genes with mutations in intergenic regions were considered only if these mutations were upstream of the translational start site. Using this assumption an initial comparison was conducted comparing the evolved strains to the Day 10 enrichment with the STSE (Figure 5.7.1).

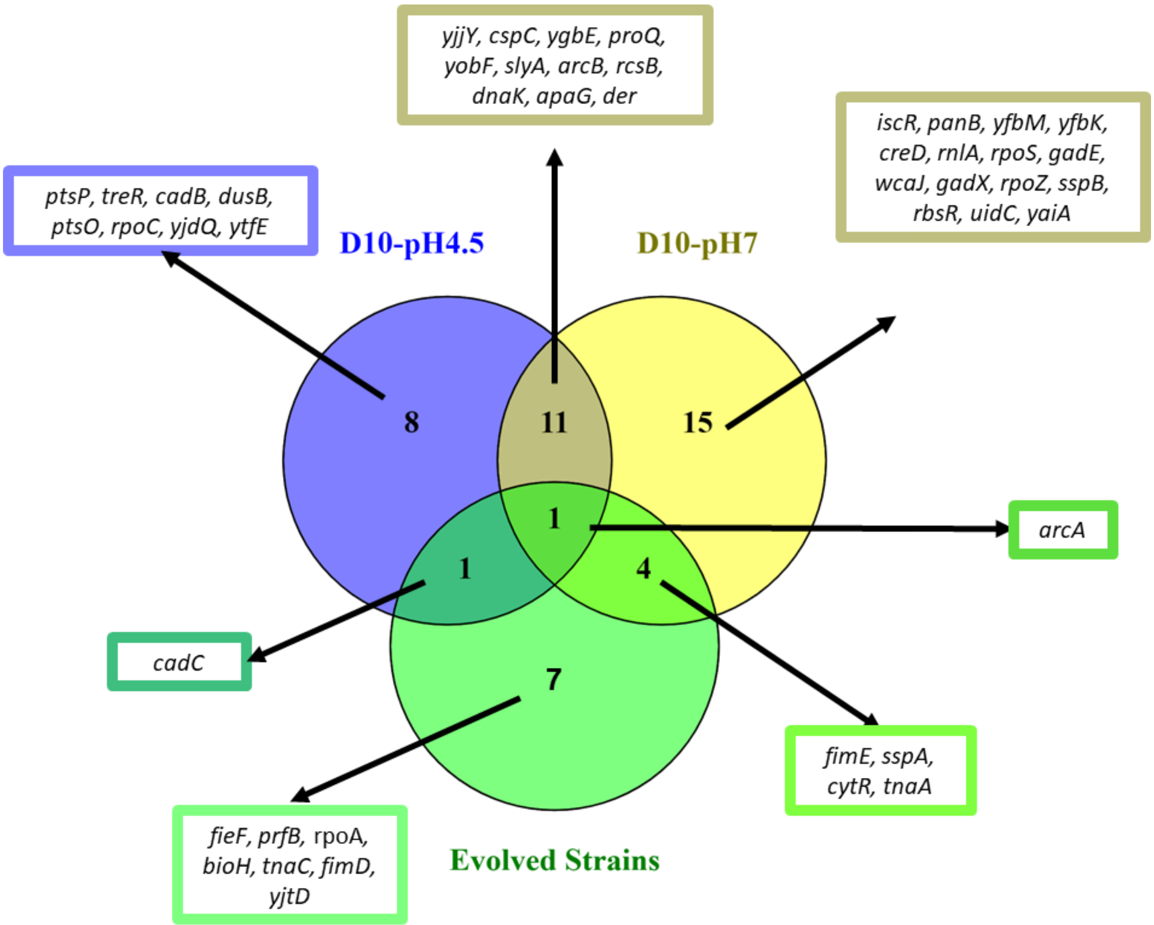

yjjY, cspC, ygbE, proQ, yobF, slyA, arcB, rcsB, dnaK, apaG, der
iscR, panB, yfbM, yfbK, creD, rnlA, rpoS, gadE, wcaJ, gadX, rpoZ, sspB, rbsR, uidC, yaiA
ptsP, treR, cadB, dusB, ptsO, rpoC, yjdQ, ytfE
D10-pH4.5
D10-pH7
8
11
15
1
arcA
1
4
cadC
7
fimE, sspA, cytR, tnaA
fieF, prfB, rpoA, bioH, tnaC, fimD, yjtD
Evolved Strains

Figure 5.7.1: Comparison of genes where mutations were observed in the evolved strains to genes shown as enriched in the conditions of the STSE.

As can be seen, of the genes found both in the STSE and in the evolved strains, more matches occurred with genes identified at D10-pH7 of the STSE, than with genes identified in D10 -pH 4.5. Only one gene (*cadC*) matched from the evolved strains to the STSE at D10-pH 4.5, but not D10-pH 7,. This result supports the point made in section 3.7.2, namely that adaptions of the evolved strains are likely to be to conditions other than the low pH.

Overall, some similarities were observed, although there were more genes different than matching in this comparison. Of the genes identified in the evolved strains, three were associated with intergenic mutations (*fimD, fieF* and *yjtD*). Of these, one *yjtD* was associated with an IS insertion, while the other two were single point mutations. Considering the IS186 insertion associated with *yjtD* is in the intergenic region between *yjjY* and *yjtD*, and this mutation has been shown to affect *arcA* expression (Section 3.4.1.1 and 5.6.2.3). Therefore, the impact of the mutation associated with *yjtD* in the evolved strain may be related to the increase in insertions observed in *yjjY* in the STSE.

Following this preliminary analysis, the next question was: of the genes which had mutations in the evolved strains, which genes were the most important with regards to the evolved strain fitness? This really can only be determined experimentally, but the degree of parallel evolution occurring in particular genes suggests that mutations of these genes confers a fitness advantage under the experimental selection conditions. Therefore, a comparison was done only including genes which were mutated in more than one strain in the Sen experiment. The result showed that of the 6 genes which occurred in more than one strain, 4 were also identified in the STSE (Figure 5.7.2). Again, these matched better to the STSE results at pH 7, with only one (*arcA*) identified at both pH values in the STSE. Only two genes (*rpoA* and *bioH*) did not show any associated enrichments in the STSE (Figure 5.7.2). *rpoA* is an essential gene, and this highlights a limitation of the use of TraDIS in this way, in that only genes where insertions are viable are able to be detected. *bioH* was not recognized within the STSE, which is interesting as the mutation in the Sen experiment was an IS186 insertion, which should be comparable to Tn insertion.

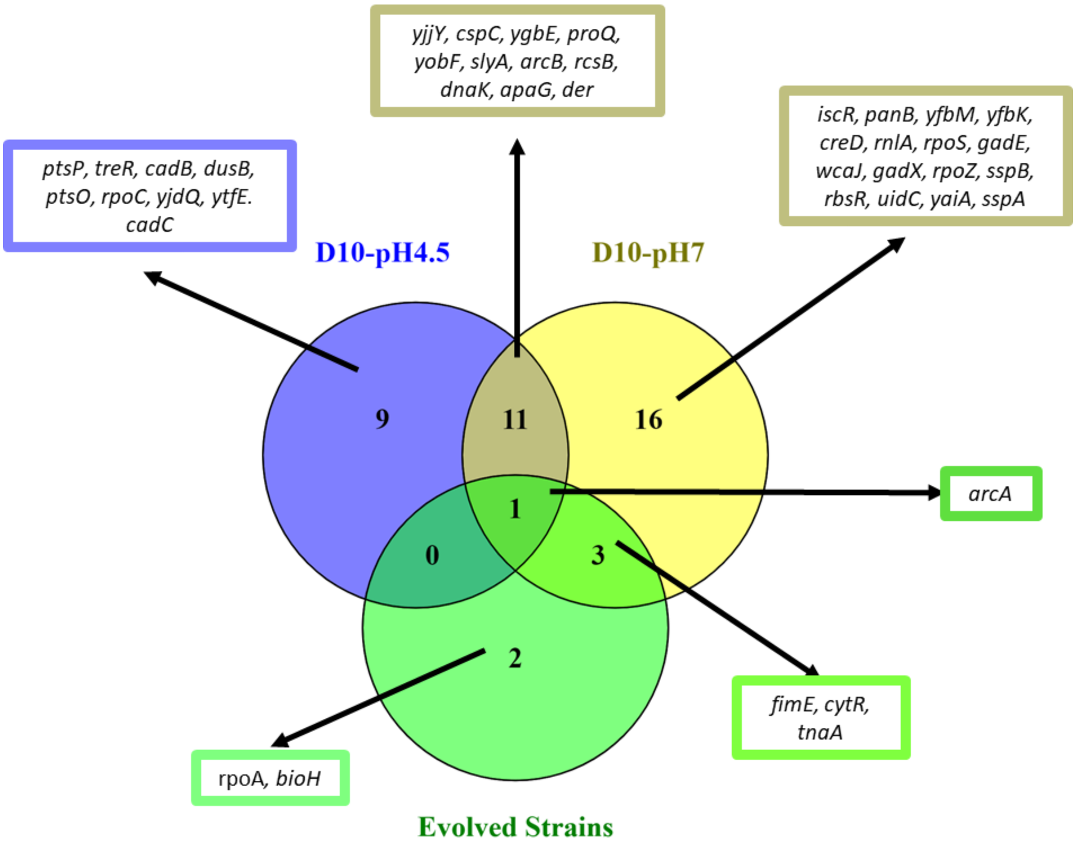

yjjY, cspC, ygbE, proQ, yobF, slyA, arcB, rcsB, dnaK, apaG, der
ptsP, treR, cadB, dusB, ptsO, rpoC, yjdQ, ytfE. cadC
iscR, panB, yfbM, yfbK, creD, rnlA, rpoS, gadE, wcaJ, gadX, rpoZ, sspB, rbsR, uidC, yaiA, sspA
D10-pH4.5
D10-pH7
9
11
16
1
arcA
0
3
2
fimE, cytR, tnaA
rpoA, bioH
Evolved Strains

Figure 5.7.2: Comparison of results of evolved strains to enrichments observed at day 10 of the STSE where mutations occur in more than one strain in the evolved strains. Intergenic mutations which were upstream of the gene were considered in this analysis.

Overall considering the differences observed between the evolved strains, it was difficult to ascertain which mutations contributed the most to fitness. However, it was noted that some genes had mutations identified in nearly all strains, namely *arcA* (5/5 strains), *rpoA* (4/5) strains and *cytR* (5/5) strains. With the exception of *rpoA* (due to it being essential), the other two genes were identified by the STSE. This strongly suggests that the STSE is at least partly able to predict the outcomes of longer term evolution experiments.

5.7.2 Comparison to the evolved populations

As stated in chapter 5, in addition to the clones E1A – E5A independent populations from the STSE were sequenced and breseq was performed to identify the mutations present in these (Table 3.2). A problem with population data is determining which mutations in which genes contributed to the increased fitness observed in the evolution experiment. An initial comparison was made between the STSE data and the population data, including all genes where a mutation was present in the population with the result seen in Figure 5.7.3. Not much has changed in the genes identified within the evolved strains as compared to the evolved clones E1A – E5A, with only one extra gene *arcB* being identified in the STSE.

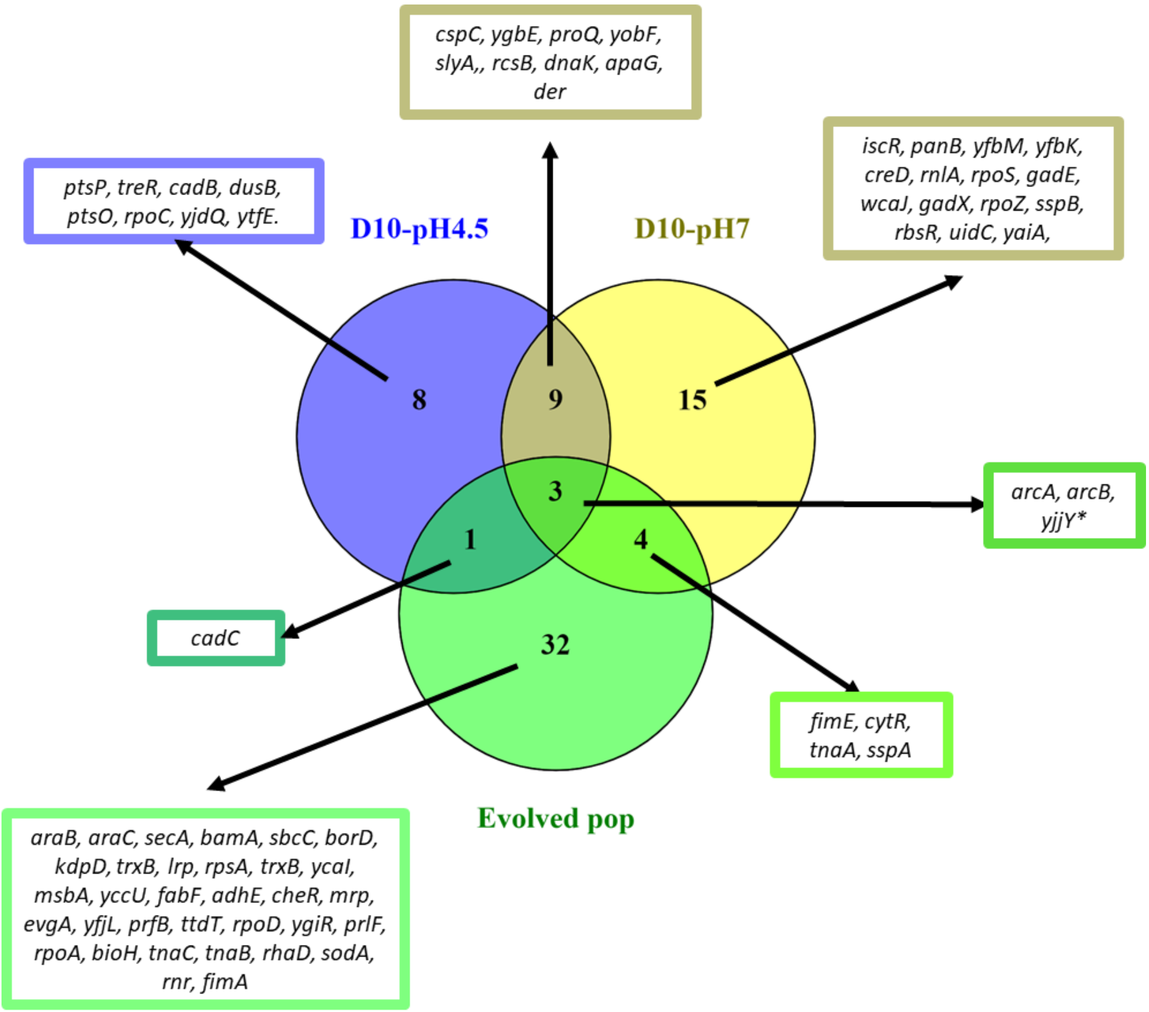

cspC, ygbE, proQ, yobF, slyA,, rcsB, dnaK, apaG, der
ptsP, treR, cadB, dusB, ptsO, rpoC, yjdQ, ytfE.
D10-pH4.5
D10-pH7
iscR, panB, yfbM, yfbK, creD, rnlA, rpoS, gadE, wcaJ, gadX, rpoZ, sspB, rbsR, uidC, yaiA,
8
9
15
3
1
4
32
arcA, arcB, yjjY*
cadC
fimE, cytR, tnaA, sspA
Evolved pop
araB, araC, secA, bamA, sbcC, borD, kdpD, trxB, lrp, rpsA, trxB, ycaI, msbA, yccU, fabF, adhE, cheR, mrp, evgA, yfjL, prfB, ttdT, rpoD, ygiR, prlF, rpoA, bioH, tnaC, tnaB, rhaD, sodA, rnr, fimA

Figure 5.7.3: Genes identified in the evolved populations, compared to the enrichments at Day 10 pH 4.5 and pH 7 of the STSE. * Note the yjjY > / > yjtD IS186 insertion event associates with the insertion accumulation seen in *yjjY* D10 of the STSE.

As with the evolved strains, the next question was could we identify genes where mutation led to the largest fitness increases within the evolved populations, to see if these genes were also identified by the STSE. As stated in Section 3.4.2 a lot of low frequency mutations were observed which were specific to one gene and only seen in one independent population indicating that mutations within these genes were less likely to have a large contribution to increased fitness. However, using population data to try and identify genes which mutations may confer a larger fitness that others is difficult within a population sequenced at a single timepoint. Therefore, bearing this in mind, two attributes were considered which may indicate increased fitness. The first was parallel evolution: whether a mutation in the same gene arose in multiple independent populations thereby possibly indicating similar strategies of adaptation within independent populations. The second was considering the frequency of the mutations, i.e. if a larger proportion of the population has a mutation, this suggests that this mutation confers a fitness advantage that has caused it to accumulate. Although these two methods can be used to infer genes which when mutated can confer a fitness advantage, they do not confirm this and really only experimentation can determine this. However, to start to try and identify genes where mutations may increase fitness a cutoff value was used, in that a gene was only considered if a mutation occurred in more than one strain or the enrichment observed was greater than 50% of the population (Figure 5.7.4).

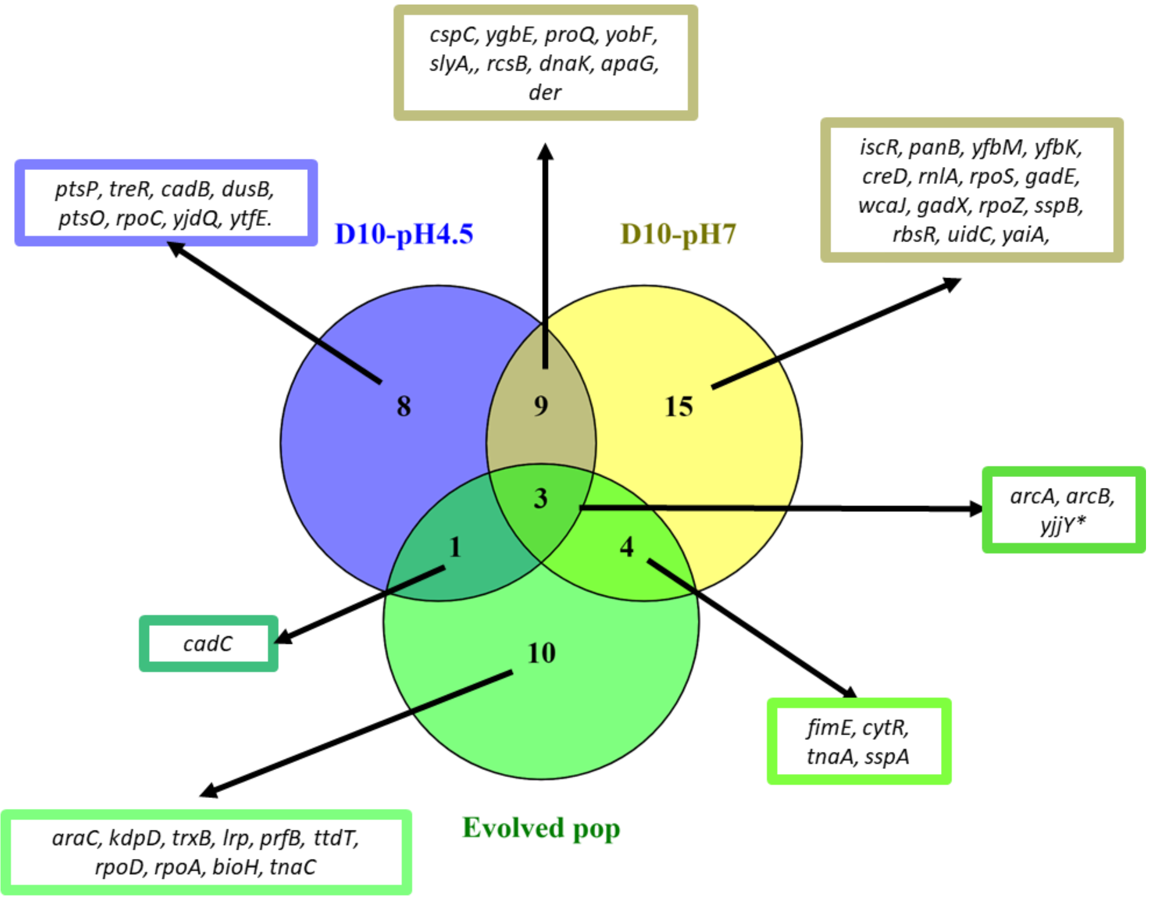

cspC, ygbE, proQ, yobF, slyA,, rcsB, dnaK, apaG, der
ptsP, treR, cadB, dusB, ptsO, rpoC, yjdQ, ytfE.
D10-pH4.5
D10-pH7
iscR, panB, yfbM, yfbK, creD, rnlA, rpoS, gadE, wcaJ, gadX, rpoZ, sspB, rbsR, uidC, yaiA,
8
9
15
3
1
4
10
arcA, arcB, yjjY*
cadC
fimE, cytR, tnaA, sspA
araC, kdpD, trxB, lrp, prfB, ttdT, rpoD, rpoA, bioH, tnaC
Evolved pop

Figure 5.7.4: Comparison of evolved population data to the genes identified with accumulation of insertions at Day 10 of the STSE at pH 4.5 or pH 7. A cutoff was used to attempt to determine genes which when mutated contributed more to increased fitness, as described in the text. * Note the *yjjY* > / > *yjtD* IS186 insertion event associates with the insertion accumulation seen in *yjjY* D10 of the STSE.

Although this cutoff criteria are only a rudimentary method to determine genes where mutations are fitter, this approach starts to highlight that the genes which are identified by the STSE do not include those observed at relatively low frequency in evolving populations. Instead they are in genes, which in an evolution experiment either occurred multiple times in independent populations or have accumulated to a level consistent with adaptation. Indeed, if one was to consider only fixation events within the evolution experiment of Sen, (2018) one would be able to identify or associate 5/7 fixation events using the output of the STSE. These are (i) in the genes *arcA, fimE,* and *sspA,* corresponding to their fixation, (ii) the *yjjY > \ < yjtD:: IS186* insertion relating to the accumulation of insertions within *yjjY* and (iii) the mutation in *tnaC* which is the leader peptide of the *tna* operon, associated with accumulation of insertions with *tnaA.* Therefore, this result indicates that the STSE is able to identify genes which when mutated are involved in increased fitness, as with the evolved strains in the longer term experiment. As with the evolved strains, the majority of these genes identified are only identified in the pH 7 condition of the STSE.

5.7.3 Consideration on the comparison

Overall, considering the similarities, the results above suggest that the genes identified within the STSE, that are associated with genes identified in the evolution experiment, are often in genes which when mutated confer a large fitness advantage, detected either by the degree of parallelism or by the increased frequency within independent populations. Particularly, it is notable that two out of the three genes previously highlighted in chapter 3 with a high degree of parallelism (*arcA, rpoA*, *cytR*) which are mutated in nearly all populations from the longer term experiment, are also identified using the STSE (Table 3.1 + 3.2).

A limitation with this method is that genes can only be included when insertions are present. In addition to this, analysis of evolution experiments cannot directly quantify fitness, while in the STSE this can be done based on the LogFC provided.

It is important to note that genes identified with accumulation of insertions within the STSE, and also identified in the evolution experiment, are not causing increased fitness simply because of loss of function. As shown in section 5.6.2.3, polar effects due to the insertion orientation can occur, which actually can cause a gain of function phenotype, as seen with *arcA* (Figure 5.6.4). Therefore, it is important to visualize insertions as well as calculating LogFC to improve interpretation of the STSE TraDIS results.

Finally, since the STSE was conducted at pH 7 and pH 4.5 more information can be obtained on how the nature of mutations observed in genes within the evolution experiment might contribute to the fitter phenotype. What was initially unexpected was that, of the genes seen within the evolution experiment that were enriched in both the evolved population and clonal isolates, only one gene, *cadC*, was enriched in just the pH 4.5 condition (Figure 5.7.1, 5.7.3 + 5.7.4). Insertions in *yjjY* and *arcA* and *arcB* were enriched in both conditions while all other genes seen (*tnaA, sspA*, *fimE,* and *cytR*) were enriched only at pH7 (Figure 5.7.1- 5.7.4. This indicates that it is likely that the mutations which arose in the evolution experiment were not due to low pH stress (even though growth was initially at pH 4.5) and instead may be due to other stresses which occurred in the evolution experiment, such as prolonged duration in stationary phase, or advantage for growth in LB broth or even possibly adaptation to higher pH. Indeed, in the case of *arcA* differences in fitness have already been observed (Section 3.8). Due to this, the STSE approach could also be used to further understand the reasons as to why mutations are selected for within an evolution experiment.

It also should be noted that the genes which were identified in the evolution experiment were not in the list of conditionally detrimental genes identified, which typically TraDIS is used for, but within the conditionally enriched gene list (Section 5.7). Moreover, the conditionally enriched gene list has more members than the list of genes identified in the longer term evolution experiment, therefore providing more potential candidates which may possess greater fitness in the conditions of the evolution experiment (Table 5.9). Experiments described in chapter 6 tested whether mutations in the additional genes identified in these lists were indeed fitter under the selection conditions used.

5.8 Summary.

The major achievements and conclusions in this chapter are summarized as follows:

1) It presented a short-term selection experiment (STSE) using a MG1655 transposon library which was done at a starting pH of 4.5 or 7 for 10 days with sampling at days

1, 5, and 10. The same growth conditions were used as Sen (2018) to allow for the comparison of results to this experiment.

2) A large decrease in transposon library complexity after growth at pH 4.5 and pH 7 was observed. The largest loss of complexity was seen after growth at Day 10 pH 7
3) Reproducibility within replicates varied greatly at pH 4.5 and pH 7 at Day 5 and Day 10 of the STSE. The differences were found to be caused by transposons inserts in a single gene being highly enriched. This enriched single gene differed between replicates particularly at D10 pH7 with an enrichment seen in either *fimE* or *yjjY*. Overall, however the top genes with transposons enriched within the population were the same.
4) Within genes that were enriched, an unusual insertion bias was seen in that a second minor peak of insertions could be observed which was shown to represent the other end of the transposon. Precisely what caused this has yet to be identified.
5) A comparison of transposon sequencing analysis pipelines was performed. Overall, a combination of log likelihood (an annotation-independent analysis) and edgeR (an RNAseq differential expression analysis pipeline) was used for further analysis.
6) Gene lists of conditionally essential genes, and transposon enriched genes, were generated. Gene ontology and Kegg pathway enrichment analysis was performed.
7) Genes which showed an enrichment in transposon sequencing were also in some cases manually inspected, revealing some patterns probably associated with polar effects.
8) It was demonstrated that transposon enrichments can be identified within essential genes, only if transposons are present within the initial library. Transposon enrichments at pH 7 and pH 4.5 of the STSE were seen within the RNA polymerase subunits and sigma factors
9) It was demonstrated that genes associated with transposons enriched in the population within the STSE overlapped significantly with the genes identified in the evolution experiment conducted by Sen (2018).

Chapter 6
Experimental Confirmation of the STSE

6.1 Introduction

In chapter 5, the results of the Short Term Selection Experiment (STSE) on an MG1655 Transposon library were presented. The overall results of this experiment concluded that after 10 days of 'evolution' in 3 independent populations in unbuffered LB at either pH 7 or pH 4.5, a specific set of genes were identified which indicated that transposon insertions within these genes had accumulated within the population (Figure 6.1.1). Insertions in these genes were considered to lead to increased fitness relative to the wild type in the conditions of the STSE that they were identified in. No experimental confirmation had been performed to confirm this.

As discussed in Section 5.8, these genes identified were also correlated to the evolution experiment described in chapter 3. This experiment was conducted under the same conditions as the pH 4.5 part of the STSE. The results of the comparison suggested that the genes where mutations were selected in the evolution experiment which were considered to have higher fitness, did correspond to genes with an accumulation of insertions in the STSE, although interestingly this was found more in the pH 7 condition than the pH 4.5 condition. However, the results of this comparison also identified a large number of genes with an accumulation of insertions in the STSE that were not seen in the evolution experiment. Therefore, this left the question: did insertions within the genes identified in the STSE actually confer increased fitness? This chapter aims to experimentally test this question.

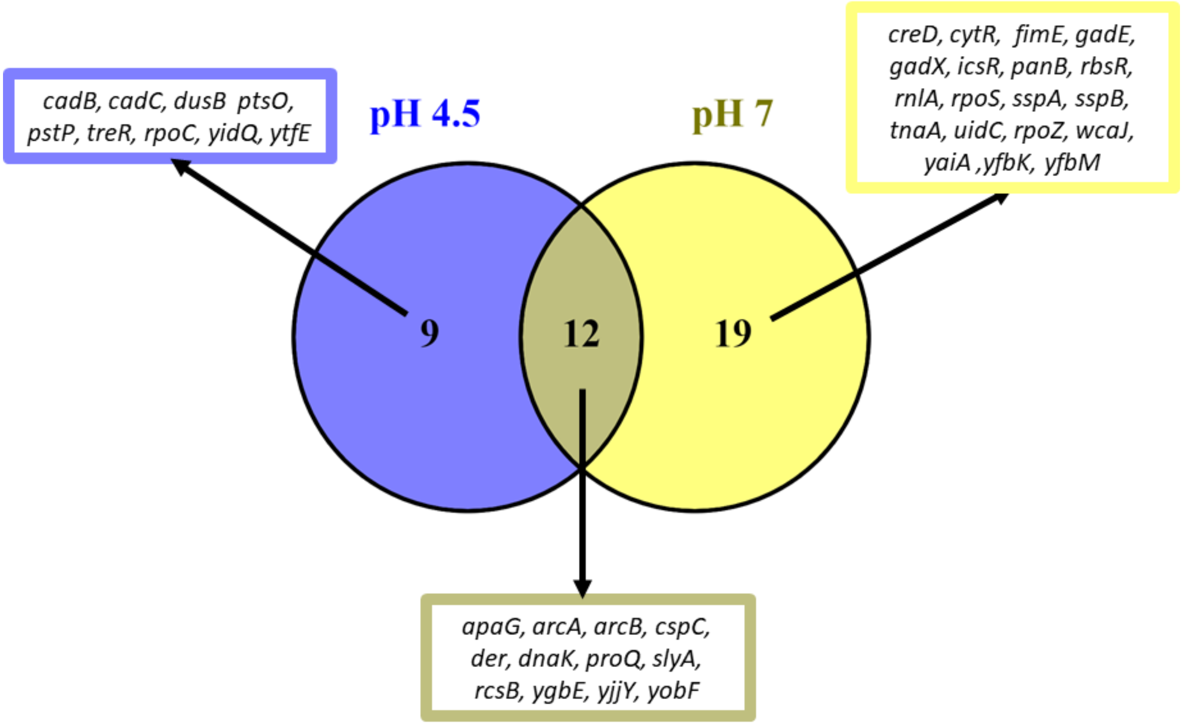

cadB, cadC, dusB ptsO, pstP, treR, rpoC, yidQ, ytfE
pH 4.5
pH 7
creD, cytR, fimE, gadE, gadX, icsR, panB, rbsR, rnlA, rpoS, sspA, sspB, tnaA, uidC, rpoZ, wcaJ, yaiA ,yfbK, yfbM
9
12
19
apaG, arcA, arcB, cspC, der, dnaK, proQ, slyA, rcsB, ygbE, yjjY, yobF

Figure 6.1.1: Overview of genes where insertions accumulated at Day10 of the STSE. Genes are split according to whether they were identified at pH 4.5, pH7 or both.

6.2 Experimental confirmation of genes identified within the STSE.

6.2.1 Confirming fitness at pH 4.5

The precise effect that insertions that accumulate in specific genes as seen in the STSE have upon the function of the gene is unknown. However, the likelihood is that a transposon insertion will cause disruption of the gene function, therefore effectively creating a loss of function. Therefore, in order to experimentally test the results seen within the STSE, a deletion of the gene identified was used as proxy for the overall effect that transposon insertions are expected to have upon the gene.

Unfortunately, due to time, not all genes could be assayed and therefore only a subset of genes was chosen. It was initially decided to only consider genes which were identified in the pH 4.5 condition of the STSE. Therefore, for each gene identified at Day 10 pH 4.5 of the STSE only, a deletion mutant in MG1655 was constructed, using P1 transduction from the Keio collection of deletion mutants (Baba *et al*., 2006). Out of the 9 genes which were identified at Day 10 pH 4.5 of the STSE as accumulated in insertions, only 7 were able to be deleted. The exceptions were *rpoC* (due to it being an essential gene) and *yidQ*, as after several attempts to construct the deletion, a deletion could not be made (Figure 6.1.1(pH4.5) + 6.2.1).

A standard way to assess fitness of a mutation is through competition experiments. This involves competing two strains together, which can then be distinguished based on a selectable marker. As previously described in Chapter 3, in order to be able to distinguish between two strains, this study used a lac^{-} derivative of *E. coli* MG1655, KH001, as a proxy for the *E. coli* MG1655 strain. Therefore, to begin to understand the effects on fitness of insertions, the deletions in genes identified under the conditions of the STSE, were competed with KH001 over 5 days using the same conditions as the STSE: 5ml unbuffered LB at pH 4.5. The results are shown below (Figure 6.2.1).

The results presented in Figure 6.2.1 show that in unbuffered LB at pH 4.5, all genes identified (with the exception of *cadB*,) which showed an increase in insertions in the STSE showed increased fitness after 5 days of competition of the MG1655 ancestor. This result confirms the prediction made from the STSE, in that insertions within these genes accumulated due to an increase in fitness. Interestingly however in Day 1 of the competition experiment, fitness in the majority of the deletion mutations was mostly neutral or even slightly negative. This result suggests that the fitness contribution of these genes is not towards a single day's growth of competition, but towards a subsequent passaging series with the condition of the STSE.

In this competition experiment, deletions of two genes (*ptsP* and *ptsO*) showed the largest increase in fitness. These two genes are the first and second enzymes in the nitrogen phosphotransferase system, which results in the phosphorylation of PtsN, a regulator thought to be involved in nitrogen metabolism and potassium transport (Pflüger-Grau and Görke, 2010). The fact that deletion of either of these genes gave a fitness advantage highlights the fact that the nitrogen phosphotransferase system plays a role in determining fitness under the conditions of the STSE at pH 4.5. PtsN has been shown to bind and phosphorylate KdpD, the sensor part of a two component system *kdpDE* which regulates a high affinity potassium transporter (Lüttmann *et al.*, 2009). Interestingly, a mutation within this gene was also shown seen at high frequency in the E3P evolved population.

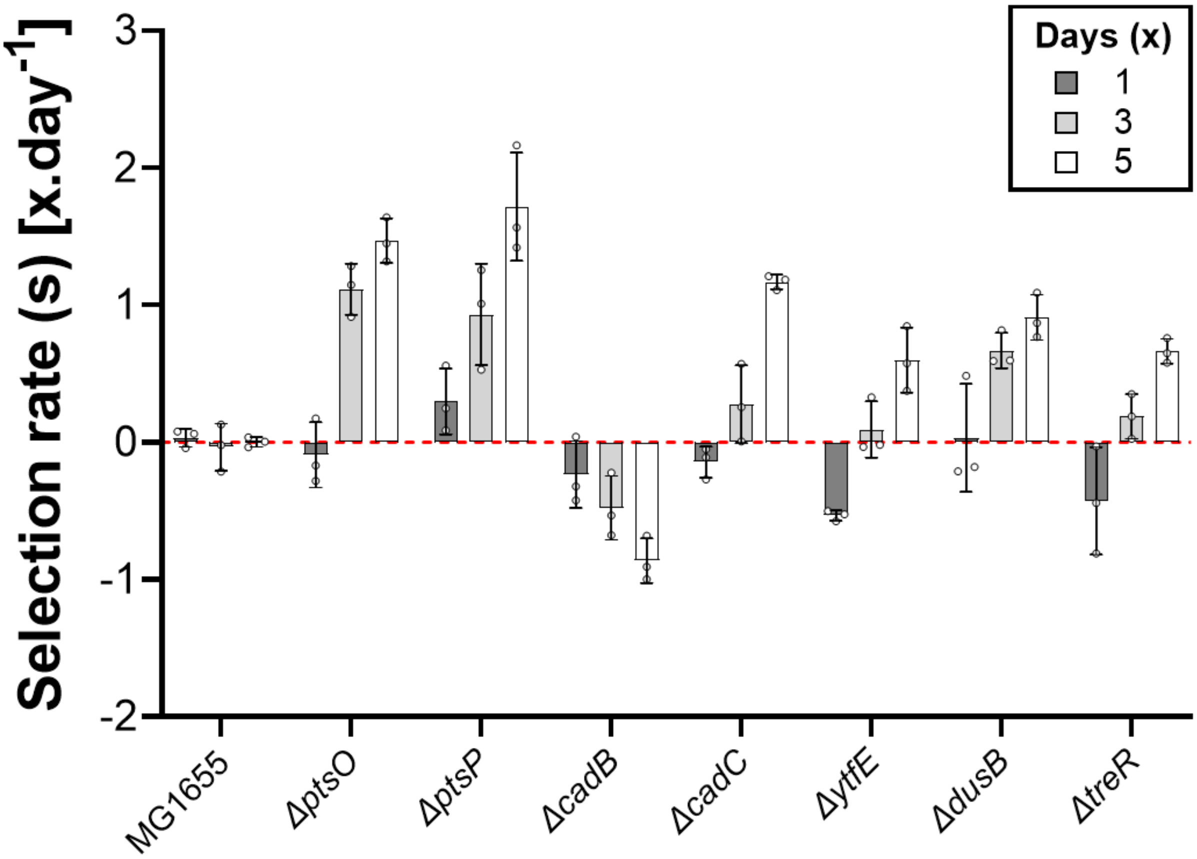


Figure 6.2.1: Competition experiments of deletions of genes identified specifically in Day 10 pH 4.5 of the STSE. Competition was performed in 5ml of unbuffered LB at pH 4.5 passaging at a 1in 20 dilution every 24 hours. Timepoints were samples on Day 1, 3, and 5. Selection rates were calculated by comparison with final counts at time 0. Therefore, the rate described the change for each time point (Day 1 = Day^{-1} , Day 3 = 3 Days^{-1}, Day 5 = 5 Days^{-1}). Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

6.2.2 Deletion of *cadB* causes a decline in fitness.

The two genes *cadB* and *cadC* were insertion enriched in the STSE at pH 4.5. Both these genes belong to the AR4 acid resistance mechanism, where *cadC* is the main regulator of AR4 activating the *cadBA* operon which encodes the mechanism for the AR4 response. *cadB* encodes the L-lysine/ cadaverine antiporter. Interestingly, *cadA* (which encodes the decarboxylase of the AR4 response, which is able to maintain cytoplasmic pH, through the reaction of L-lysine with H$^+$ to produce CO_2 and cadaverine) was not enriched in the STSE, However in the competition experiment shown in Figure 6.2.1, only one gene deletion, , that of *cadC*, shows an increase in fitness in unbuffered LB at pH 4.5, while a *cadB* deletion actually shows a decrease fitness. This does not correspond to the results seen within the STSE. A manual inspection of the insertions identified at D10-pH 4.5 within *cadBA* and *cadC* in the STSE was therefore done, including noting the orientation of the transposon insertion The results of this analysis showed that the orientation of the transposons presented a clear bias, where the transposon was higher in frequency in the opposite orientation to transcription of the *cadBA* operon. It is possible that this could create an antisense RNA, which can inhibit/reduce transcription or translation in a variety of ways (Thomason and Storz, 2010). Therefore, potentially the result indicated that the transposon accumulation observed within *cadB* at D10-pH4.5 of the STSE, could cause an inhibition of the entire *cad* operon including *cadA*.

To look at this experimentally, *cadA* was also deleted using P1 transduction from the Keio collection, and competition performed in unbuffered LB at pH 4.5 for 5 days (Baba *et al.*, 2006). The results of this experiment for all *cad* gene deletions are shown in Figure 6.2.2B. Interestingly a *cadA* deletion showed a fitness advantage under these conditions which correlated with the similar fitness advantage seen for Δ*cadC*. This is consistent with the transposon insertions present within *cadB* also affecting expression of *cadA*. However, this doesn't explain why insertions were found accumulated only within *cadB*. It could be that a larger fitness advantage is seen when the entire operon function is removed. This would be consistent with the results also seen with *cadC* which is the activator of this operon, as when this is deleted the *cadBA* genes would not be activated. This hypothesis could be tested using a double deletion of *cadA* and *cadB*. Unfortunately, due to time constraints this could not be done.

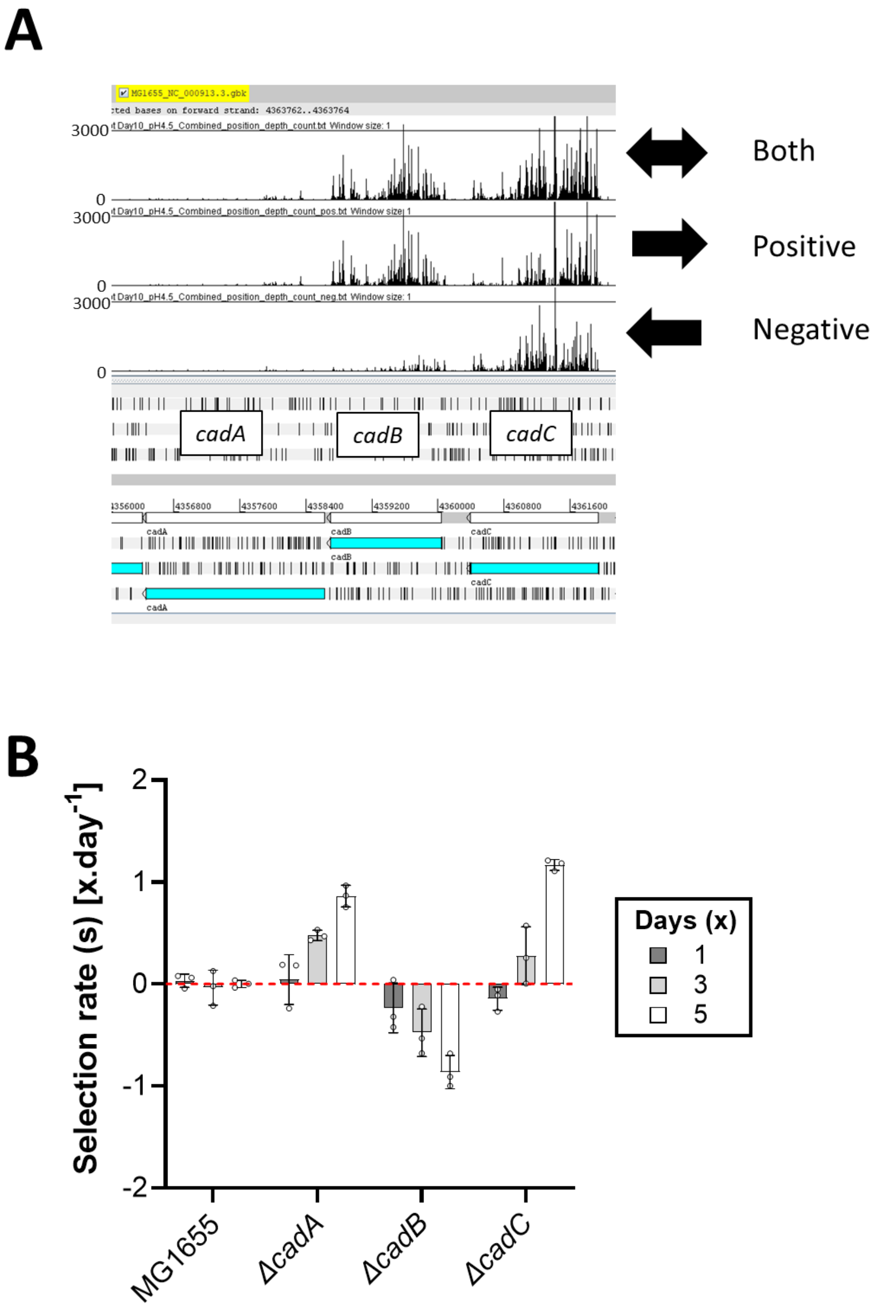


Figure 6.2.2: Focus upon the genes involved in the AR4 mechanism *cadABC*. A) Manual visualisation of insertions found at D10-pH4.5 of the STSE (An example of one replicate is provide). Insertions are split according to whether they were in the positive or negative orientation with the black arrows indicating the direction of the transcription readthrough. B) Competition experiment over 5 days in unbuffered LB at pH 4.5, passaging ever 24 hour at a 1 in 20 dilution. Selection rates were calculated by comparison with final counts at time 0. Therefore, the rate describes the change for each time point (Day 1 = Day^{-1} , Day 3 = 3 Days^{-1}, Day 5 = 5 Days^{-1}). Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

6.2.3 Deletion of genes associated with AR2 gives an increase in fitness at pH 7

In section 5.6.2, genes involved in the regulation of the AR2 were seen to have an increase in insertions at D10-pH 7 of the STSE. These specifically were *gadX* and *gadE*; *rcsB* was found have an accumulation of insertions at D10-pH4.5 as well. To determine whether these insertions do lead to increased fitness, deletions of these genes were constructed in MG1655 using P1 transduction using the Keio collection (Baba *et al*., 2006). These deletions were then competed with KH001 under conditions of the STSE at pH 7. The results of this analysis showed that in all three genes, an increase in fitness was observed, consistent with the result observed in the STSE (Figure 6.2.3). The main regulator of the AR2 mechanism (*gadE*) had the largest fitness advantage at all days of the competition compared to the other two genes *gadX* and *rcsB.* It is difficult to explain precisely the effect that loss of function of these genes would have on the cell, as although all three genes are associated with the AR2 response, they have also been shown to be involved in the regulation of several other genes, involved in a variety of functions throughout the cell (Hommais *et al*., 2004, Seo *et al*., 2015, Wall *et al*., 2018).

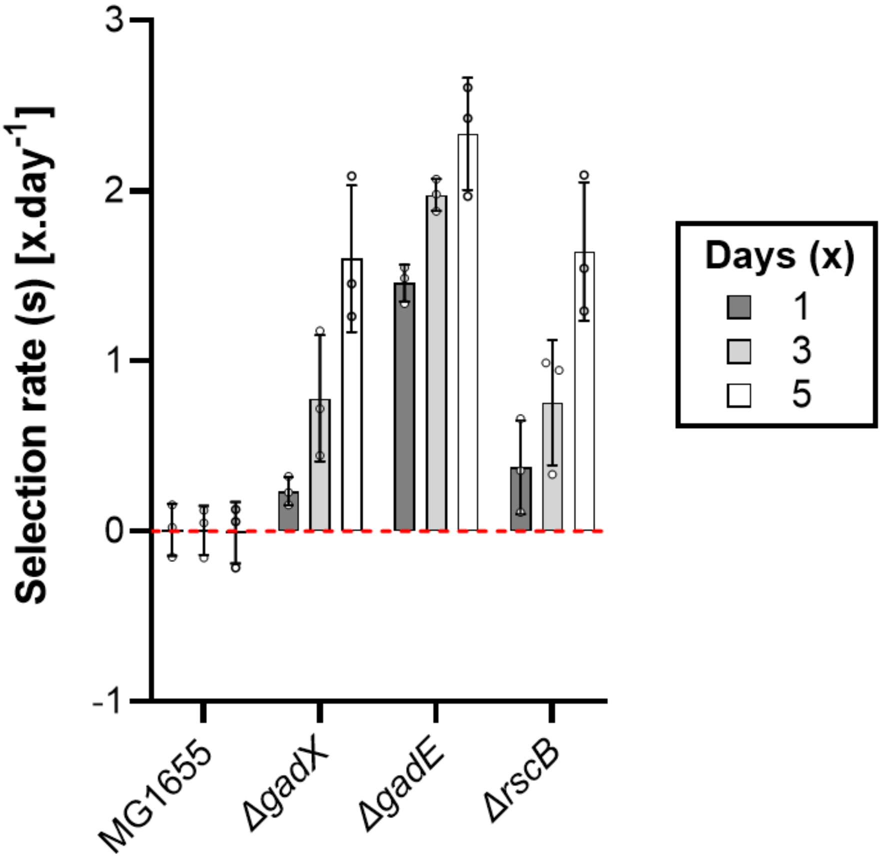


Figure 6.2.3: Competition of deletions in genes associated with the AR2 mechanism Competition was performed in 5ml of unbuffered LB at pH 7 passaging every 24 hours at a 1 in 20 dilution. Red line indicates selection rate which would be observed if no difference in fitness existed between strains. Selection rates were calculated by comparison with final counts at time 0. Therefore, the rate describes change for each time point (Day 1 = Day^{-1} , Day 3 = 3 Days^{-1}, Day 5 = 5 Days^{-1}) The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

6.3 The *arcAB* two component system.

Chapter 3 described how mutations were identified in the two component system *arcB* and *arcA* in the evolved populations generated in unbuffered LB at pH 4.5 (Section 3.4). In addition to this an *IS186* insertion downstream of *yjjY* was also observed and shown to be fixed within one evolved population. This insertions was likely to affect *arcA* expression as the *arcA* promoter extends 500bp upstream from the translational start site (Figure 6.3.1A). These three genes all also showed increased insertions in the STSE, at Day 10 at both pH7 and pH 4.5,. In particular, insertions associated with *yjjY* were not limited to the *yjjY* gene itself but extended to the boundaries of the *arcA* promoter gene (Figure 6.3.1B). In addition, as described in Section 5.7.2.3 an unusual pattern of insertions in *arcA* was observed, in which insertions were not seen throughout the gene but only in the 5' and 3' ends of the gene (Figure 6.3.1B). Moreover, inspection of *yjjY* insertions showed an orientation bias in *yjjY* and the *arcA* promoter, indicating that read though from the transposon into *arcA* was possibly occurring (Figure 6.3.1B). As previously reported in Section 3.8, this result corresponds to what was observed for *arcA* deletions in the MG1655 ancestor which showed a fitness disadvantage in unbuffered LB at both pH 4.5 and pH7.

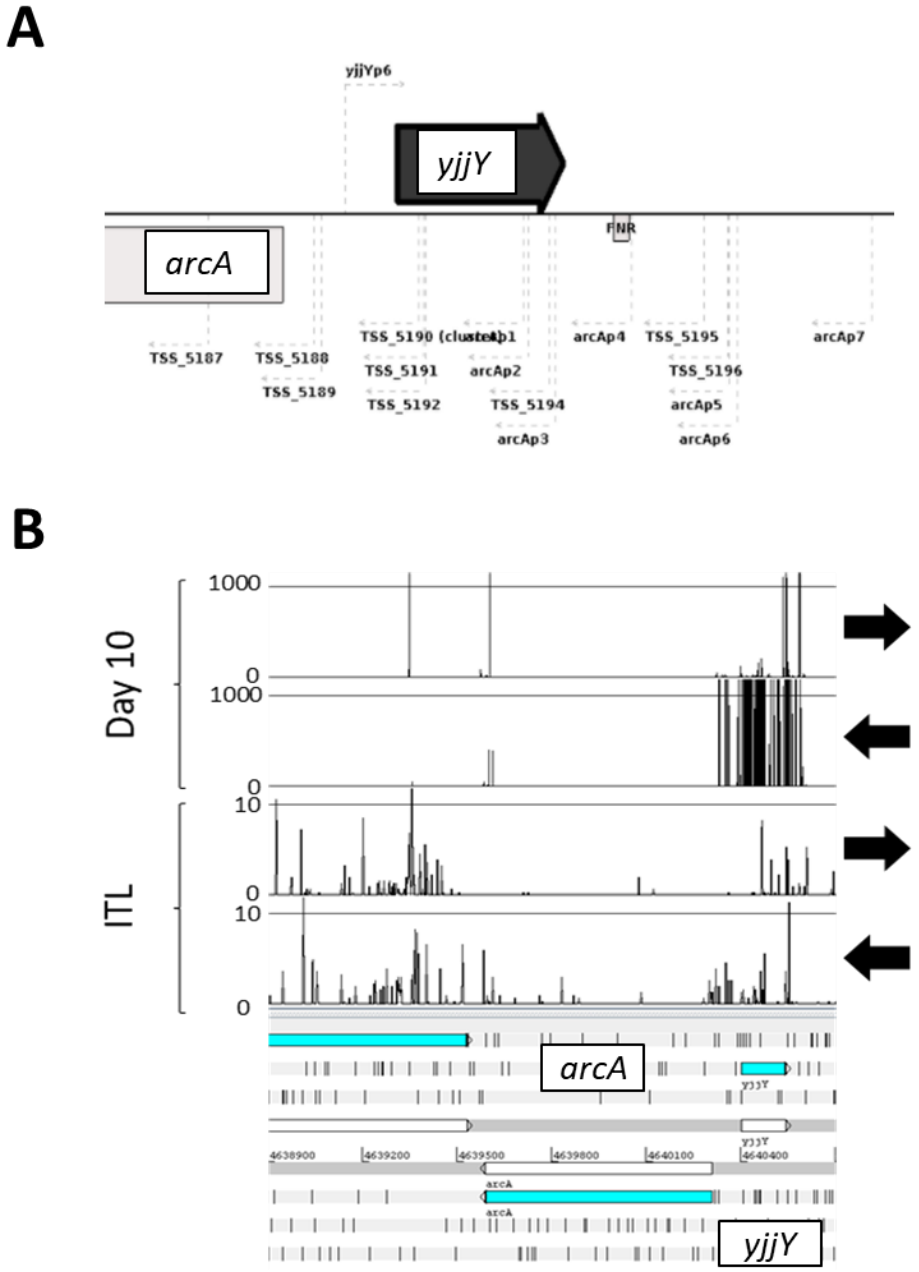


Figure 6.3.1: Figure showing the relationship between *arcA* and *yjjY.* A) Diagram taken from RegulonDB, showing the *arcA* promoter, is overlaps with *yjjY*. B) An example of transposon insertion pattern (shown in Artemis) at Day10 of the STSE at pH4.5. Transposons have been split based on their orientation. Black arrows indicate the orientation of the transposon.

6.3.1 Deletions of *arcAB* and *yjjY* at pH 4.5

These results suggest that *arcA, arcB* and *yjjY* have a role in fitness in unbuffered LB at both pH 4.5 and pH 7. It was initially decided to ascertain the effect that gene deletions have on fitness at pH 4.5. Therefore, deletions of *arcA, arcB* and *yjjY* were made by P1 transduction from the Kieo collection into MG1655 (Baba *et al*., 2006). These were then competed with KH001 in unbuffered LB at pH 4.5 over 5 days (Figure 6.3.2). This confirmed that *ΔarcA* shows a fitness disadvantage under these conditions, and this was also found to be true of *ΔyjjY*. This result therefore showed that loss of function of *yjjY* could not explain result observed in the STSE, so the pattern of transposon insertions present in the STSE must be due to a different effect.

However, *ΔarcB* showed an increase in fitness, although this was only observed at Day 5, with *ΔarcB* actually showing a fitness decrease at Day1 and (to a lesser extent) Day3. What causes this initial decrease in fitness is unknown. It has been previously reported that an *arcB* deletion strain had a slower growth rate in LB and glucose over 24 hours (Mika and Hengge, 2005, Perrenoud and Sauer, 2005). Deletions in *arcB* have also been shown to have a profound effect on central carbon metabolism under aerobic conditions, causing up-regulation of central TCA cycle genes under these conditions (Perrenoud and Sauer, 2005, Nizam *et al*., 2009). It is possible that a fitness advantage exists which cannot be detected after a single day's growth, such as a shorter lag phase.

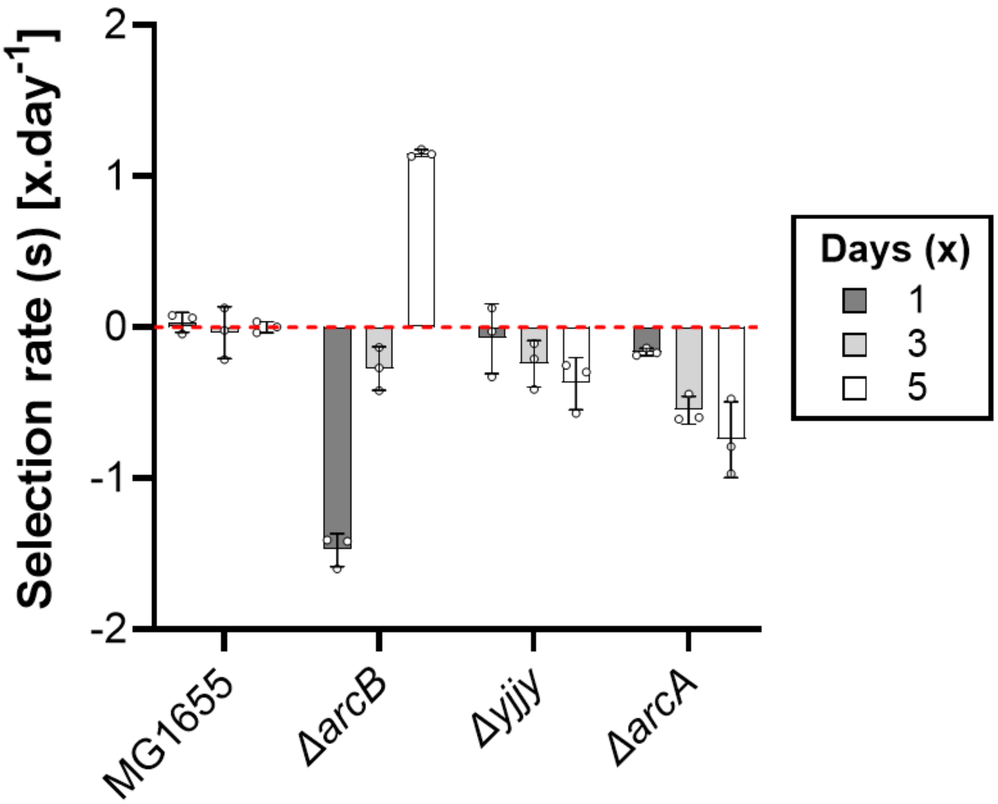


Figure 6.3.2: Competitions of gene deletions associated with the *arcAB* two component systems. Competition was performed in 5ml of unbuffered LB at pH 4.5 passaging every 24 hours at a 1 in 20 dilution. The red line indicates the selection rate which would be observed if not difference in fitness was occurred between strains. Selection rates were calculated by comparison with final counts at time 0. Therefore, the rate described change for each time point (Day 1 = Day^{-1} , Day 3 = 3 Days^{-1}, Day 5 = 5 Days^{-1}). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

These results therefore suggest something unusual, since *yjjY* and *arcA* show a fitness disadvantage under these conditions. This shows that *arcA* function is required for full fitness in unbuffered LB at pH 4.5, as well as in unbuffered LB at pH 7 (Section 3.8). As *yjjY* overlaps with the promoter of *arcA*, so the region controlling expression of arcA is within or upstream of *yjjY.* Considering that when *arcB* is deleted this can confer a fitness advantage, this result corresponds to the result seen in the STSE, where insertions were found throughout the gene with no insertion bias. As mentioned previously loss of ArcB function would mean that ArcA would be unable to be activated by phosphorylation through the ArcB kinase activity. Therefore, this suggests that the effect that *arcA* has upon fitness is associated with its unphosphorylated form, unless ArcA can be activated by some alternative means.

6.3.2 The impact of mutations in *yjjY* on fitness

In the STSE, at Day 10 insertions within *yjjY* had the largest increase in LogFC, suggesting that these had the greatest fitness advantage. However, an *yjjY* deletion showed a decrease in fitness. As discussed above, this may be explained if (as suggested by the orientation of insertions within *yjjY*) the insertions in *yjjY* elicit expression of the downstream *arcA* gene. An *yjjY* deletion from the Keio collection would remove any *arcA* promoter region within the *yjjY* gene itself. In addition to this, it would also disrupt any promoter which was at the 3' end of the *yjjY* gene. Therefore in an attempt to see if the kanamycin cassette has any effect, it was removed using a FRT recombinase leaving only a 102bp scar (Baba *et al.*, 2006). In addition to this, as mentioned in section 5.5, *yjjY*-associated transposon insertions were isolated from the STSE population. Five day experiments competing *ΔyjjY::kan, ΔyjjY::FRT* and *yjjY(+25)::TN5* against KH001 were then done. As both pH conditions saw an accumulations of insertions within *yjjY* , competitions were performed at both pH 4.5 and pH 7 with the results shown in Figure 6.3.3.

The results presented in Figure 6.3.3 confirm that the insertion isolated from the STSE in *yjjY* did indeed confer an large fitness advantage at both pH 4.5 and pH 7 over the 5 days of competition. Interestingly, a larger fitness increase was observed at pH 4.5 than pH 7 suggesting that *yjjY* insertions are fitter at pH 4.5 than pH 7. In addition to this the Δ*yjjY::FRT* mutation, which replaced *yjjY* with an 102bp scar, also showed higher fitness, while *ΔyjjY::Kan* (the kanamycin-marked deletion from the Keio collection) had a loss of fitness. As to whether *yjjY* is an actual gene or not, this study does not precisely know, analysis using NCBI blastn revealed that although *yjjY* was highly conserved across the main phylogroups of *E. coli*, so was *arcA* and it's promoter (which includes *yjjY*) (DATA NOT SHOWN). However, the result in Figure 6.3.3 suggests that fitness observed due to insertions in *yjjY* is probably not due to any protein generated by *yjjY* as if this were the case the same level of fitness would be expected between these two deletions. It also suggests that the kanamycin cassette disrupts the *arcA* promoter, and also suggests that a region downstream from the 3' end of *yjjY* has an effect on *arcA* expression. In addition, this result shows that the insertions upstream of *arcA* including those in the *yjjY* gene are causing increased fitness, as if normal regulation was maintained the fitness reported would be the same as that of MG1655.

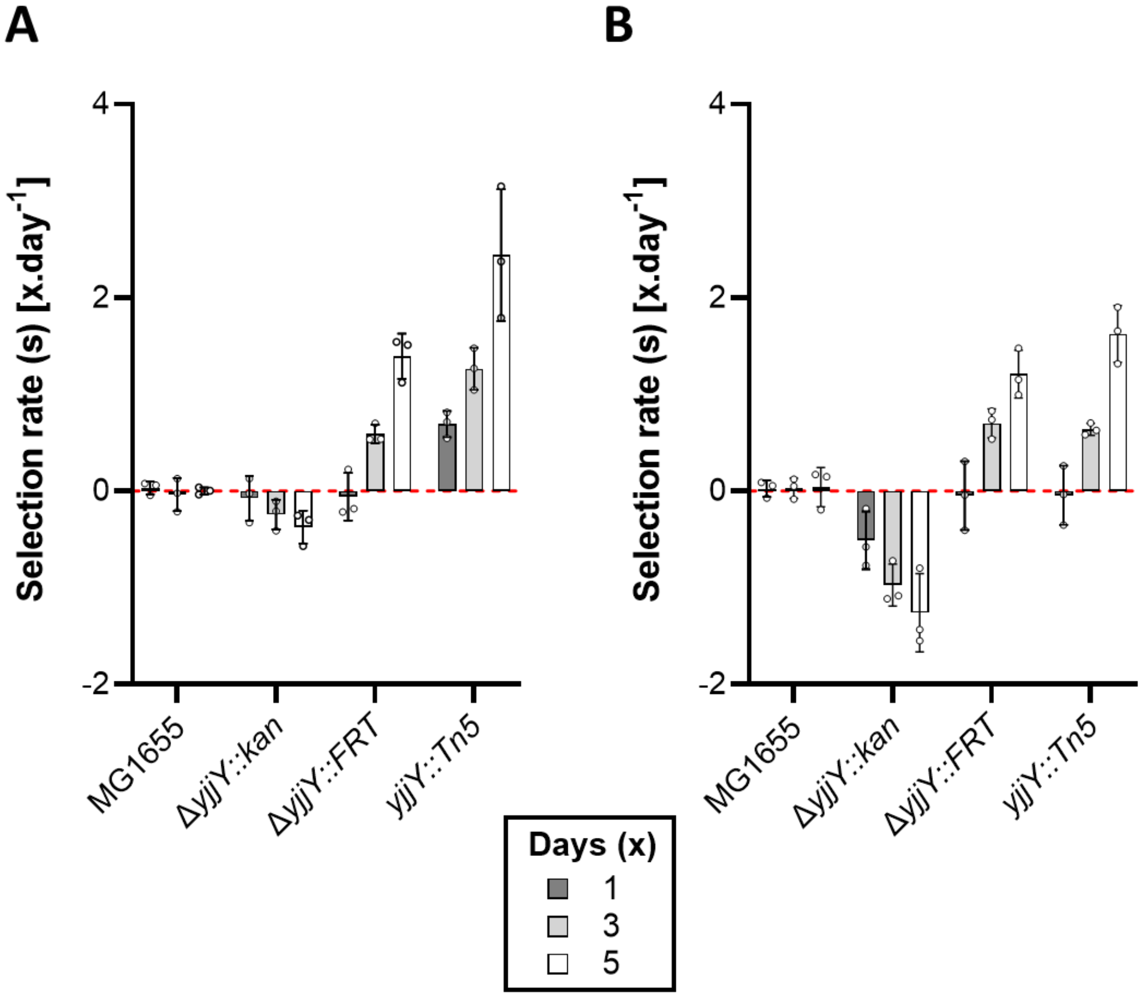


Figure 6.3.3: 5 day competition experiments of different types of *yjjY* mutation.. Competitions were performed in 5ml of unbuffered LB at either pH 4.5 (A) or pH 7 (B). After 24 hours each culture was passaged at a 1 in 20 dilution. Selection rates were calculated by comparison with final counts at time 0. The rate describes the change for each time point (Day 1 = Day^{-1} , Day 3 = 3 Days^{-1}, Day 5 = 5 Days^{-1}). Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

Overall, this result shows clearly that different mutations can have different effects even when they are at the same location. It also suggests that mutations in *yjjY* which cause a fitness advantage to affect the promoter region and alter the expression/regulation of *arcA* so that it is no longer WT regulation. However, what this result does not tell us is whether in aerobic conditions *arcA* is expressed higher than wildtype levels, or is reduced to below wildtype levels, whether a dysregulation occurs and the overall affect that differential expression level might have upon the cell.

6.4 YjjY vs FimE: a tale of different replicates

At Day 10 in the STSE at both pH 4.5 and pH7, insertions in *yjjY* appeared to have the largest increase in fitness. At Day 10 pH7 however, a slightly different story was seen in different replicates, with two replicates showing insertions in *fimE* to be the most fit, whilst in the other it was *yjjY*. In this section, an attempt was made to start to understand this phenomenon.

FimE is involved in the regulation of the type 1 fimbrae, whose function is closely associated with biofilm formation and pathogenicity of *E. coli* (Rodrigues and Elimelech, 2009, Hultgren *et al*., 1985, Avalos Vizcarra *et al*., 2016). Type 1 fimbrae undergo phase variation meaning that in a population a proportion of *E. coli* will express type 1 fimbrae and a proportion will not (Abraham *et al*., 1985). This phase variation occurs as the genes which encode for the type 1 fimbrae are in an operon *fimAICDFGH*, controlled by a promoter which is flanked by two 9bp inverted terminal repeats (Figure 6.4.1). This region of DNA has been labelled the fim switch, *fimS,* as it can invert through the actions of two recombinases *fimB* and *fimE* (Klemm, 1986, Gally *et al*., 1996, McClain *et al*., 1991). Expression of the fimbrae genes is dependent on whether the *fimS* is in the ON or OFF orientation. While the recombinase FimE can only invert *fimS* from ON to OFF, FimB can invert *fimS* in both direction (Figure 6.4.1) (McClain *et al*., 1991).

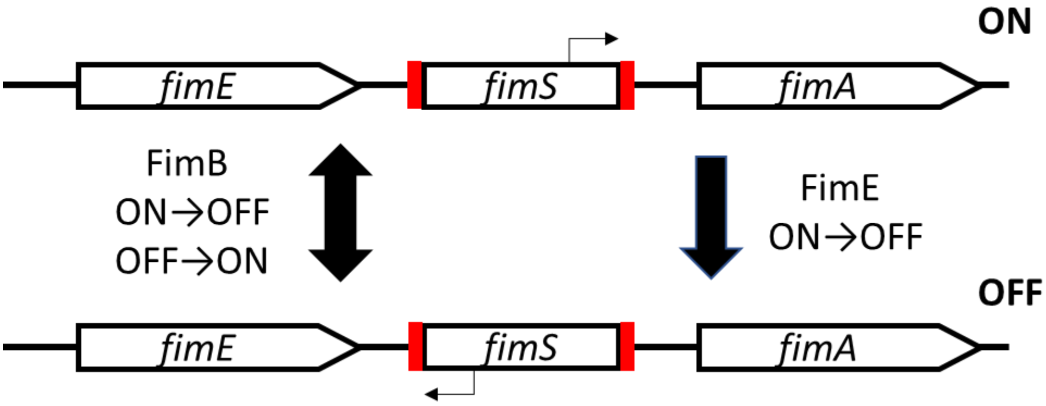


Figure 6.4.1: The phase variation of the type 1 fimbrae. Regulation of the major component of the type 1 fimbrae *fimA* is by a promoter which is in a region of DNA, *fimS*, flanked by 9 bp inverted terminal repeats (red lines). Two recombinases, *fimE* and *fimB*, act upon *fimS* inverting the promoter so expression of *fimA* is either ON or OFF. The black arrows indicate which direction each recombinase is able to go. Downstream of *fimA* an operon exists consisting of genes *fimAICDFGH. fimA* is the major component for the type 1 fimbrae, *fimFGH* are the minor components and *fimICD* is for transport and assembly. Upstream of *fimE* is *fimB;* both are recombinases

6.4.1 Relative effects on fitness of transposon inserts in *fimE* and *yjjY*.

In the STSE, insertions within *fimE* and *yjjY* were shown to be fitter in all conditions at D10. Accumulation of insertion within these two genes was also seen at D5-pH4.5 in some replicates, but not at D10-pH4.5 where inserts in *yjjY* were found to dominate (Figure 5.4.3). To understand how insertions within these genes contribute to fitness, a deletion of *fimE* was created using P1 transduction from the Keio collection (Baba *et al*., 2006). It should be stated that within the STSE conditions no insertion orientation bias was observed, suggesting that the insertions within *fimE* elicited a loss of function. A five day competition experiment of *ΔfimE* against KH001 was performed under the same conditions as the STSE. In order to compare between *fimE* and *yjjY* , the result from competition of the *yjjY(+25)::Tn5* are presented alongside the *ΔfimE* competition (Figure 6.4.2).

The results of this analysis show that *ΔfimE* was fitter at both pH 4.5 and pH7 (Figure 6.4.2). At pH 4.5 at Day 5, *yjjY::Tn5* had a larger fitness advantage than *ΔfimE*, corresponding to what was observed at pH 4.5, where insertions within *yjjY* eventually showing the highest fitness. The fact that *ΔfimE* showed a fitness advantage at all in unbuffered LB at pH 4.5 confirms that accumulation of strains with insertions in *fimE* accumulation at day 5 is understandable; however, it does not answer why *fimE* insertions were not seen enriched at Day 10 pH 4.5 of the STSE.

The results at pH 7 suggest that *ΔfimE* is fitter than *yjjY* which corresponds to the results observed within two of the replicates at D10-pH7 of the STSE where the largest accumulation of insertions was within *fimE*. However, after 1 day growth in unbuffered LB at pH 4.5 or pH7, Δ*fimE* shows a greater fitness advantage than *yjjY::Tn5.* This result indicates that these two different mutations may confer fitness, to different stresses in the STSE. The precise nature of these are unknown. However, mutations within *fimE* have been observed in several evolution experiments involving unbuffered LB, or LB buffered at pH9, possibly indicating an advantage specific to LB (Hamdallah *et al.*, 2018, Knöppel *et al.*, 2018, Behringer *et al.*, 2018).

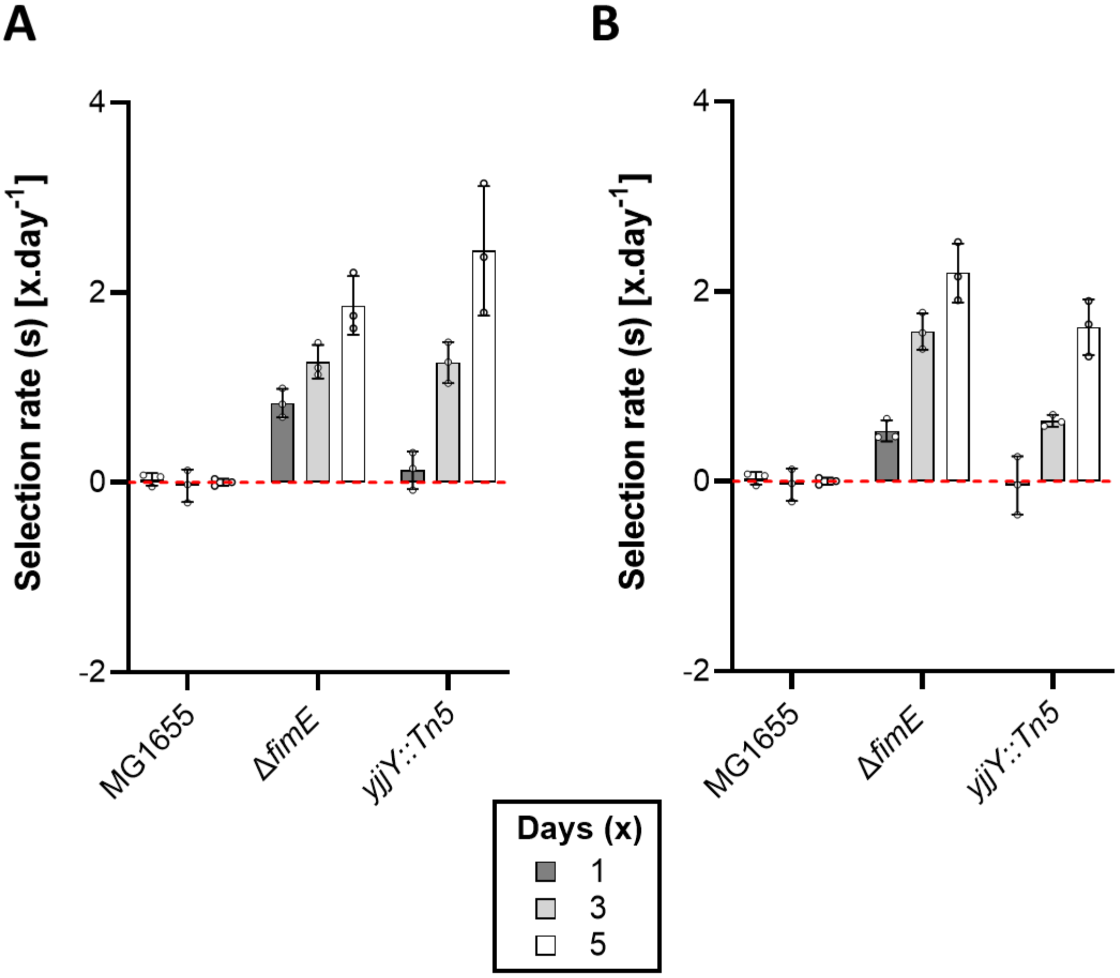


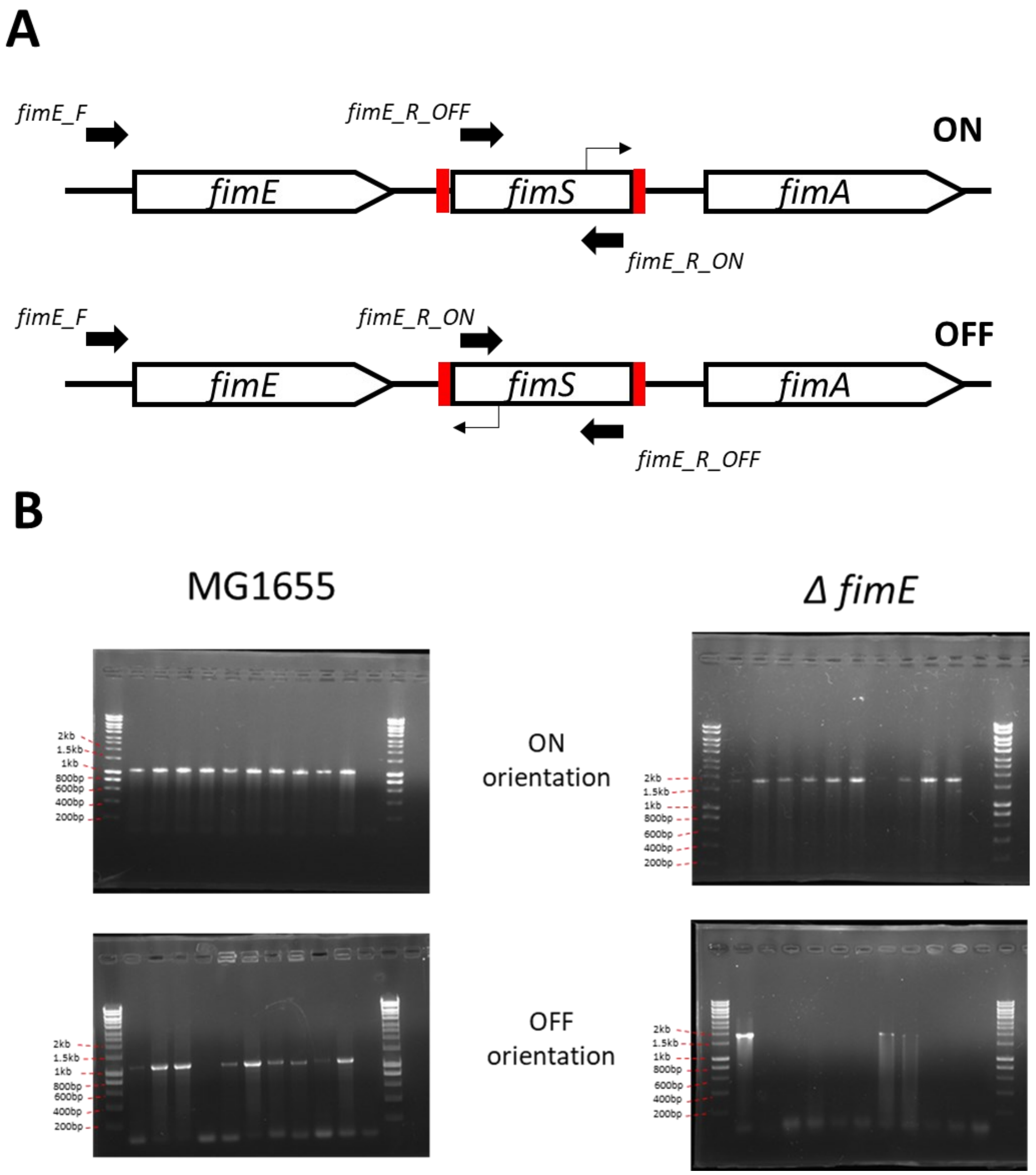
Figure 6.4.2: Comparison of the fitness of *ΔfimE* and *yjjY::Tn5* under the conditions of the STSE. Competitions were performed over five days in unbuffered LB at either pH 4.5 (A) or pH7 (B). Every 24 hours the culture was passaged into fresh media at a 1in 20 dilution. Selection rates were calculated by comparison with final counts at time 0. Therefore, the rate described change for each time point (Day 1 = Day^{-1} , Day 3 = 3 Days^{-1}, Day 5 = 5 Days^{-1}). Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

6.4.2 Deletion of *fimE* enhances expression of type 1 fimbrae

As mentioned above *fimE* is only able to invert *fimS* from the ON to OFF configuration, while the other recombinase *fimB* can invert *fimS* in both orientations (McClain *et al*., 1991). Considering that *fimB* has the capacity to regulate *fimS* independently to *fimE*, the question was asked how would type 1 fimbrae expression be affected in a *ΔfimE* strain?

To establish this, PCR was used to determine the orientation of the *fimS* within a population. This involved using a flanking primer within *fimE* and orientation specific primers for *fimS* (Figure 6.4.3A). Using these primers would indicate the orientation of *fimS* depending on whether or not a fragment was observed, (Figure 6.4.3A). Ten independent cultures of MG1655 and *ΔfimE* were allowed to grow in 5ml unbuffered LB at pH 7 for 24 hours at 37^{o}C. PCR was then performed on these cultures with the results shown in Figure 6.4.3B. As expected, the results for MG1655 show that in a population, some of the population had *fimS* in the ON orientation, while the remainder had it in in the OFF orientation. However, in the *ΔfimE* strain, in the majority of cultures, a large proportion of the population had the *fimS* in the ON orientation, although a few cultures were more predominantly in the OFF orientation (Figure 6.4.1). This shows that in *ΔfimE* cultures, in the majority of the population *fimS* was in the ON orientation.

The orientation of *fimS* was also considered in the previously isolated *fimE (+65)::Tn5* mutation (isolated from the Day 10-pH7 STSE). 16 independent cultures of *fimE(+65)::Tn5* were grown for 24 hours in unbuffered LB at pH 7. The results of this experiment show that in the majority of cultures *fimS* is on the ON orientation, where type 1 fimbrae are expressed (Figure 6.4.3C). Type 1 fimbrae have been associated with biofilm formation, but no biofilm was seen in any of the cultures. This result is limited by the resolution of the PCR test and cannot show if the entire population are in to one orientation. It also does not confirm that the type 1 fimbrae are expressed, only that *fimS* is in the orientation that makes it possible. However, this result does confirm that for a large proportion of the population with a *fimE* mutation, the *fimS* is in the ON orientation (Figure 6.4.3.BC).


A
fimE_F
fimE_R_OFF
ON
fimE
fimS
fimA
fimE_R_ON
fimE_F
fimE_R_ON
OFF
fimE
fimS
fimA
fimE_R_OFF
B
MG1655
Δ fimE
ON orientation
OFF orientation
2kb
1.5kb
1kb
800bp
600bp
400bp
200bp

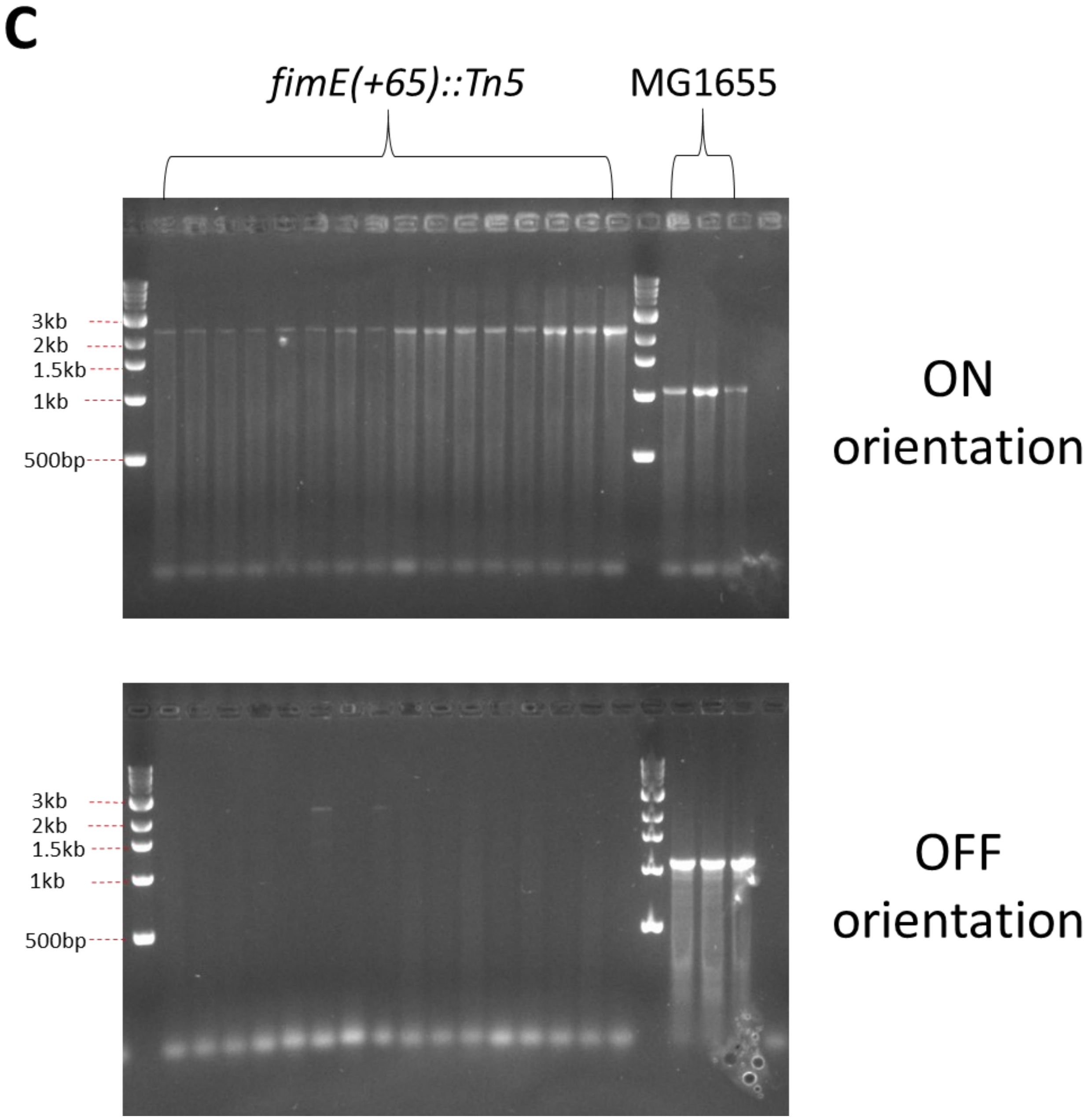


Figure 6.4.3: PCR determination of the orientation of *fimS* in a *ΔfimE* background. A) Primer pairs used to determine whether *fimS* is ON (Expression of the type1 fimbrae) or OFF (Not Expressing type 1 fimbrae). The ON orientation is detected by primers (*fimE_F + fimE_R_ON)* and OFF orientation by primers (*fimE_F + fimE_R_OFF)*. B) 10 independent cultures of MG1655 and *ΔfimE* grown for 24 hours in 5ml unbuffered LB at pH 7. For each culture PCR was done to determine the orientation of the *fimS.* The DNA ladder used is the Bioline 1kb hyper ladder. C) 16 independent cultures of *fimE(+65)::Tn5* were grown overnight and the orientation of *fimE* was determined

6.4.3 Type 1 fimbrae are required for fitness

As the above result cannot confirm whether type 1 fimbrae are expressed, the effect of removing the major component of fimbrae, *fimA,* was measured. The results of this analysis saw a loss of fitness in *ΔfimA* over the course of the competition (Figure 6.4.4). This strongly supports the hypothesis that expression of type 1 fimbrae is required to observe fitness for a *ΔfimE* mutation, and for insertions in *fimE.*

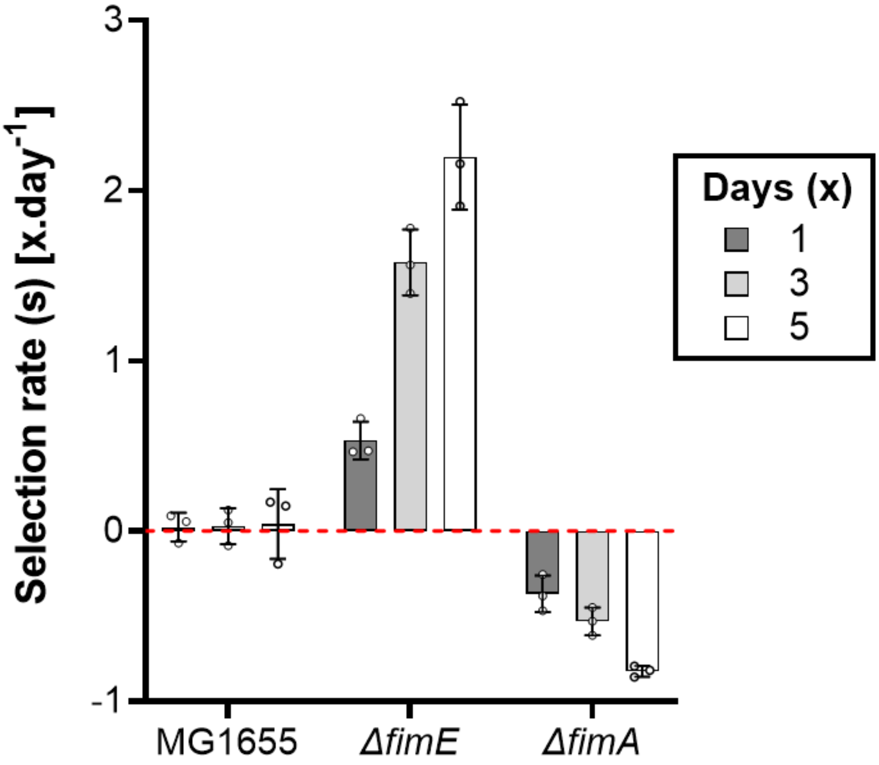


Figure 6.4.4: The effect of type 1 fimbrae on fitness. Strains were competed against KH001 in 5ml of unbuffered LB with an initial pH of pH 7. Every 24 hours the culture was passage into fresh media at a 1 in 20 dilution. Selection rates were calculated by comparison with final counts at time 0. Therefore, the rate describes changes for each time point (Day 1 = Day^{-1} , Day 3 = 3 Days^{-1}, Day 5 = 5 Days^{-1}). Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points.

6.4.4 Direct competition between *yjjY* and *fimE* mutants.

The results above show that insertions in *fimE* contribute to fitness, but not why differences between replicates were observed at pH 7 (Section 5.4), where insertions in either *fimE* or *yjjY* were enriched in different replicates. As shown in Figure 6.4.2B, at pH 7 when *yjjY::Tn5 and ΔfimE* are competed against KH001, both show a fitness advantage, although *ΔfimE* has a larger fitness advantage than *yjjY::Tn5*. We would therefore predict that in a direct competition, *ΔfimE* would outcompete *yjjY::Tn5.* To address this, the *ΔfimE* mutant was created in KH001, so that we could directly compare fitness between *fimE* and *yjjY* mutants. (Figure 6.4.5).

As an initial control, to ensure that KH001 *ΔfimE* could be used a proxy for MG1655 Δ*fimE* these two strains were competed. In both conditions (pH 4.5 and pH 7) no difference in fitness was observed between these two strains. Overall, the results of the competition showed that at pH 4.5, the *yjjY::Tn5* insertion at Day 5 had higher fitness than *ΔfimE*. However, at day 1 of the competition at pH 4.5, Δ*fimE* had a fitness advantage over *yjjY::Tn5*. This is similar to what was observed in competition against KH001, (see Figure 6.4.2), and suggests that these two mutations potentially confer fitness to different stresses/ selection pressures within the STSE and these may operate at different timepoints. At pH 7 however, *ΔfimE* showed increased fitness relative to both *yjjY* mutants over 5 days, indicating that it is fitter than *yjjY* mutations at pH 7. The fact that both *yjjY* mutants declined in fitness indicates a similar situation to what was observed within two replicates of the STSE where insertions in *fimE* come to dominate the population.

In addition to comparing the *yjjY::Tn5* mutation, the *yjjY::FRT* which was observed to have a fitness advantage against KH001 was also tested. This result was validated in the direct comparison as *yjjY:FRT* mutation did not show the same fitness as the Tn5 mutation, with Δ*fimE* actually showing a fitness advantage at pH 4.5. This was also true at pH 7 with *yjjY::FRT* showing a larger fitness decline than the *yjjY::Tn5* mutant, providing more evidence that the *yjjY::Tn5* confers a larger fitness advantage than *yjjY::FRT*. Overall this result further supports the hypothesis that the two *yjjY* mutations have differences in relative fitness.

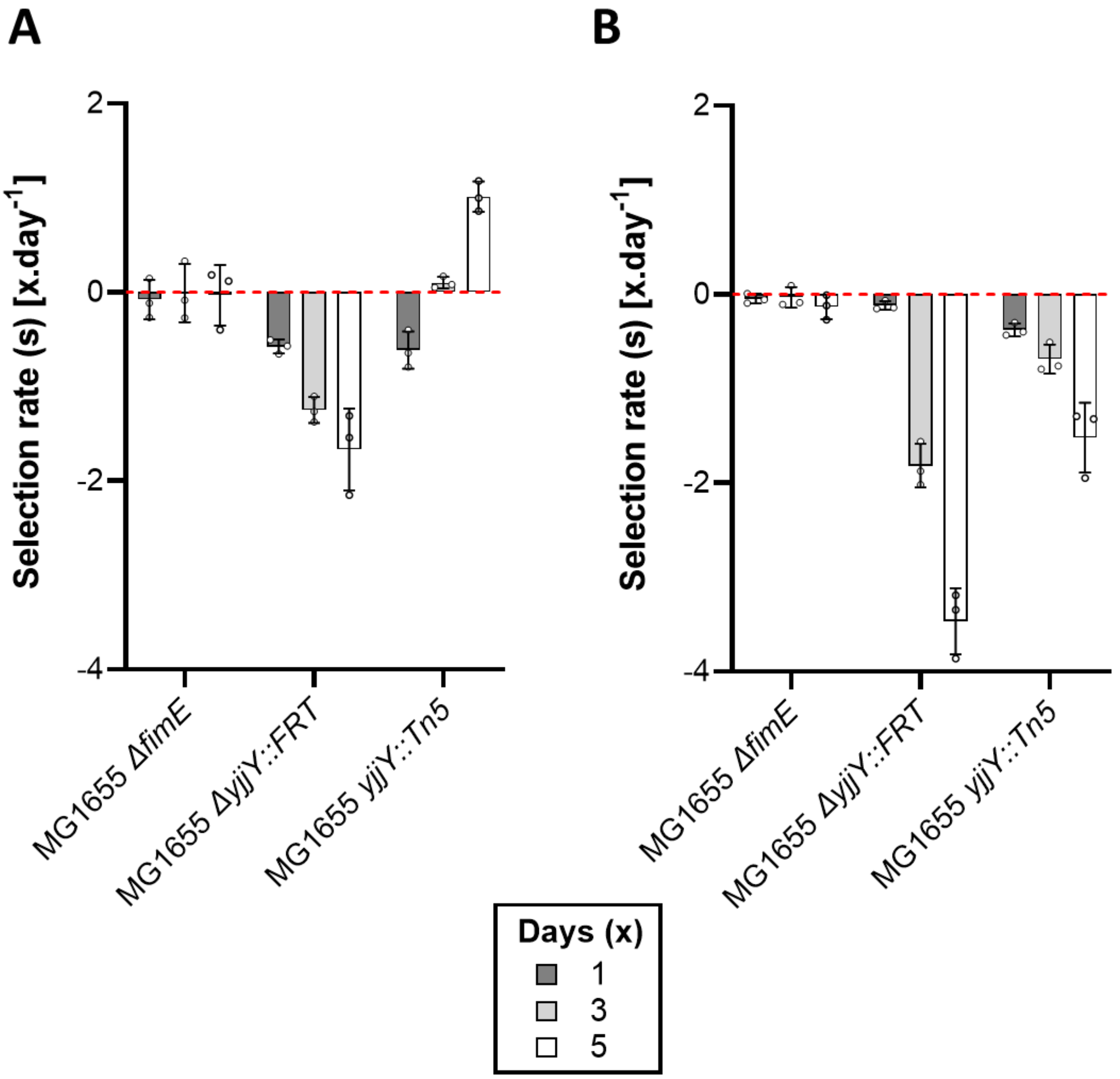


Figure 6.4.5: Competition of *yjjY::FRT* and *yjjY::TN5* against KH001 *ΔfimE*. Each competition involved equal mix of strains competed in 5ml of unbuffered LB with an initial pH at either pH 4.5 (A) or pH (7) over five days of competition, each culture was passaged every 24 hour at a dilution of 1:20. Selection rates were calculated by comparison with final counts at time 0. Therefore, the rate described change for each time point (Day 1 = Day^{-1} , Day 3 = 3 Days^{-1}, Day 5 = 5 Days^{-1}). Red dashed line indicates the value if there was no difference in fitness observed between strains (selection rate of 0). The mean +/- SD of three independent replicates are plotted, alongside the individual data points. Note that a negative value indicates a fitness advantage for *ΔfimE* while a positive value shows a fitness advantage for *yjjY* mutations.

A simple explanation for what was observed in the other D10-pH7 replicate of the STSE, where strains carrying inserts in *yjjY* dominated could not be found from these results. Most of the data predicts that insertions within *fimE* should be the most enriched as they show the higher fitness at pH 7. However, the experiments consider only two competing strains, which does not fully represent what would occur in the STSE. Here, competition would be occurring between strains all with insertions in multiple different genes with multiple different fitness's and at different initial frequencies. The results of the STSE at D10-pH7 in the one replicate where insertions in *yjjY* came to predominate, may perhaps be explained by a selective event occurring in one population early on (hence the lower accumulation of insertions in *fimE* at day 5), which favoured the strains with *yjjY* insertions. What this event might have been is unknown and unfortunately, due to the pandemic, further investigation to this could not be performed.

6.5 Summary.

The key points of the chapter are highlighted below

1. .Deletion of most of the genes where strains carrying insertions dominated by Day 10 pH 4.5 of the STSE gave a fitness advantage under the same conditions, confirming that the insertions accumulated in the STSE do relate to their effects on relative fitness.
2. In addition to this, genes associated with the AR2 response at pH 7 also showed a fitness advantage when deleted.
3. Insertions in *cadB* were potentially also causing loss of function in the entire *cadBA* operon. A *ΔcadA* mutation showed increase fitness at pH 4.5 in support of this, but the fact that *ΔcadB* showed a fitness disadvantage requires further study.
4. Insertions within *arcA* and *yjjY* selected in the STSE did not cause a loss of function. Loss of function of *arcB* led to higher fitness, over prolonged exposure to the conditions of the STSE
5. The insertions within *yjjY* observed within the STSE cause a gain of function mutation. These insertions within *yjjY* potentially alter the expression of the *arcA* gene.
6. *ΔfimE* has a fitness advantage at pH 4.5 and pH 7 and that this is associated with expression of the type 1 fimbrae, as deletion of *fimA* leads to loss of fitness under the condition of the STSE at pH 7.

7. Direct competition of *yjjY* against *fimE* supports the observation on pH 7 Day 10 of the STSE in two replicates which saw a large accumulation of *fimE*. The increase in *yjjY* inserts in the third replicate is hypothesised to have been caused by a different selective event in this replicate.

Chapter 7

Discussion

7.1 Overview of project.

As discussed in section 1.3.3., it is argued that within evolution experiments, loss of function mutations contribute to an adaptive phenotype more commonly than gain of function mutations (Olson, 1999, Murray, 2020), although this argument is still disputed, with examples of laboratory based evolution being identified which both support and oppose this statement (McDonald *et al*., 2009, Lind *et al*., 2015, He *et al*., 2017, Johnson *et al*., 2014, Anand *et al*., 2020, Blount *et al*., 2008). However, this study sought to capitalise on this observation, since if it is correct that loss of function mutations are the more common adaptive strategy, a possible alternative to laboratory based evolution could be proposed. In this, by starting an experiment with all possible loss of function mutations already present and tracking the frequency of these mutations at individual timepoints, similar results could be obtained in a considerably shorter time frame to that of a conventional lab-based evolution experiment, as time normally spent waiting for the mutations to arise would not be needed.

Therefore, this thesis sought to answer the question, "Can transposon directed insertion site sequencing (TraDIS) be used to determine the possible outcomes of evolution?". To do this, the fundamental aims of this study were: 1) To continue to link the genotype to the phenotype of a previous 5 month laboratory evolution experiment conducted in unbuffered LB at pH 4.5 (Chapter 3). 2) Create a transposon library within *E. coli* K-12 MG1655, the strain previously used within the evolution experiment (Chapter 4). 3) Perform an outgrowth experiment for 10 days under the same conditions as the previous evolution experiment (STSE) using this transposon library (Chapter 5). 4) Compare the results between these two techniques (Chapter 5.7 + 6). This work aimed to provide an example in which this approach could be used to as an alternative to conventional laboratory based evolution.

7.2 Adaptation has occurred to conditions other than low pH within an evolution experiment

Laboratory based evolution has been used as a tool to understand stress for many years, with table 1.1 and 1.2 highlighting a few of the many stresses that *E. coli* has been adapted to in such experiments. In this thesis, a previous laboratory evolution experiment using unbuffered LB at pH 4.5 provided the foundation with which to test our hypothesis (Chapter 3). Although the original experiment considered low pH to be the main stress (Sen, 2018), further analysis in the present work showed that the use of unbuffered LB actually created an environment with a dynamic pH (Section 3.3), with competition experiments demonstrating that adaptation within the evolution experiment was not just towards stress at low pH (Figure 3.6.3). In particular, results presented in section 3.6.4 suggested conditions such as growth in LB media (Figure 3.6.5 + 3.6.8), and stresses associated with stationary phase (Figure 3.6.5), to be some of the more prominent selective conditions which our evolved populations within this study had adapted to.

Further to this, genes where adaptations had occurred within this study, did not correlate well to a previous evolution experiment conducted under highly similar conditions in LBK medium buffered at pH 4.6, with only mutations within *cadC* and in *rpoD*, in our study correlating (He *et al*., 2017, Harden *et al*., 2015), further suggesting that within our experiment low pH was not the dominant selection pressure. However, when comparing our results to other evolution experiments under selective conditions, other than low pH, many similar genes could be identified. For example, mutations identified in *cytR* and *sspA* can be linked to a fitness advantage under long term stationary phase in LB media (Kram *et al*., 2017, Ratib *et al*., 2021), as well as mutations within *sspA* being associated with adaptation at pH 9 (Hamdallah *et al*., 2018). Mutations in *arcA* have also been identified within evolution experiments conducted in unbuffered LB at pH 7 (Knöppel *et al*., 2018), correlating to our results that *arcA* mutations were selected more in unbuffered LB at pH 7 rather than pH 4.5 (Figure 3.7.1 - 3.7.3). In addition, mutations within *fimE* also indicated that adaption had occurred towards off-target stresses present within the evolution experiment such as growth in LB, as discussed below.

Overall, these observations demonstrated that in the evolution experiment using unbuffered LB at pH 4.5, low pH was probably not the defining stress, showing that within every evolution experiment designed to understand a specific stress, there will also exist multiple additional stresses present due to the experimental conditions, which will also have selective pressure. Therefore, when using laboratory evolution to understand an organism's response to stress, it is important to consider whether an adaptation corresponds to the target stress, or an alternative stress. Although attempts have been made to identify only the adaptive genotype to a specific target stress either by performing parallel control experiments (Du *et al*., 2020, Knöppel *et al*., 2018), or by using an organism already pre-adapted to experimental conditions without the target stress (Riehle *et al*., 2001, Bennett and Lenski, 2007), a large proportion of laboratory based evolution experiments are typically conducted under one condition (as is the case with this study), where both target and off-target stresses are present together. Therefore, this study shows that when attempting to experimentally confirm that a particular mutation provides an adaptation to a defined stress, it is important to confirm this under other conditions involving that stress and not just within the same experimental conditions.

7.2.1 In laboratory evolution, adaptations to off-target conditions are common.

Considering this further, this study also shows that in laboratory evolution experiments, adaptations in a subset of genes can be identified which confer advantagous fitness to common off-target conditions. For example, our study identified adaptations within *fimE* in both our evolution experiment and the STSE pH 7 condition, which when compared to the literature suggested adaptation to a common off-target stress. FimE is a recombinase, which is involved in the regulation of the type I fimbrae, primarily turning expression OFF (Figure 6.4.1) and we have shown that a *fimE* deletion is associated with increased fitness when *E. coli* is grown in unbuffered LB at pH 4.5 or pH 7 (Figure 6.4.2) potentially by causing an increase in expression of the type I fimbrae (Figure 6.4.3). However, within the literature, mutations in *fimE* have been identified under a variety of conditions, such as growth at pH 9 (Hamdallah *et al*., 2018), tolerance to benzoate (Creamer *et al*., 2017), microgravity (Tirumalai *et al*., 2017), rifampicin resistance (Matange *et al*., 2019) and triclosan resistance (Leyn *et al*., 2021). In addition, a recent preprint describing a TraDIS-Xpress experiment showed that a deletion of *fimE* is associated with advantagous fitness in biofilm formation

(Holden *et al*., 2021). However, a single common condition can be identified in all these experiments in that they were all done using LB media. Experiments which have focused solely on growth in LB media have also identified mutations within *fimE* (Behringer *et al*., 2018, Knöppel *et al*., 2018). Furthermore, in Knöppel *et al.* (2018), evolution experiments were performed in different laboratory media demonstrating that *fimE* mutations only occurred within complex media (MH and LB) experiments, while evolution under defined media (M9 supplemented with glucose or glycerol) did not select for mutations in *fimE*.

In addition, mutations within the core RNA complex (*rpoA, rpoB, rpoC, rpoZ*) are routinely identified within laboratory evolution. Some of these have been characterised in detail, with different mutations in these genes demonstrating different effects on global transcription, as well as different advantagous fitness phenotypes (Conrad *et al*., 2009, González-González *et al*., 2017, Du *et al*., 2020, Harden *et al*., 2015, He *et al*., 2017). Considering the observations in our study and in others, again this supports the point made above that, there is a potential requirement to identify genes and adaptations that can confer a fitness advantage to off-target conditions. Identifying such adaptations could then be used to analyse the output of laboratory evolution when it is used to understand stress using only one condition, enabling the identification of adaptations associated to the target stress and not to off-target conditions. Evolution experiments already exist which can assist in this identification, such as studies which focus on common off-target conditions such as growth in specified media for which the LTEE can also be useful (Conrad *et al*., 2009, Knöppel *et al*., 2018, Barrick *et al*., 2009). The current study suggests that the collation of evolution experiment genotypes alongside metadata, which should include a precise description of experimental conditions, would allow the identification of genes whose adaptations are associated to off-target conditions. Attempts to do this have already started, such as the ALEdb database which has begun to collate the final genotypes of laboratory evolution experiments (Phaneuf *et al*., 2019). However, at present, this database only provides the genotypes identified within evolution experiments and, although the potential exists, has yet to correlate experimental conditions to particular adaptations.

7.3 Outgrowth of a transposon library identifies similar genes to those found in an evolution experiment.

To address the main hypothesis of this study, the results presented in section 5.7 demonstrated a significant overlap between genes identified as mutated in a 5 month evolution experiment (Table 3.1 and 3.2) and genes identified in a short term 10 day selection experiment of a transposon library (STSE: Table 5.7 – 5.9). This included genes which, in the evolution experiment, were identified as the most important for fitness advantage, based on mutations reaching higher frequency or occuring within multiple independent populations (Figure 3.7.2 + 3.7.4). This included two of genes (*arcA* and *cytR*) which were mutated in all five independent populations in the five month evolution experiment. (The third gene which was mutated in all five populations, *rpoA*, was not seen in the TraDIS experiment as it is essential). Therefore, although the use of TraDIS cannot identify every gene where mutations confer fitness advantage in an evolution experiment, it is able to identify what we would consider the majority of key genes of interest in a laboratory evolution experiment over a significantly shorter time period.

7.3.1 Overlap of evolution experiment to conditions of the STSE, were towards insertion accumulation genes

As stated previously (section 1.4.3), TraDIS can be used to identify two sets of genes associated with fitness under a given condition: “insertion detrimental” or “insertion advantagous”. Within the STSE, this study identified both sets of genes, and showed that the overlap observed with the evolution experiment was, as expected, with genes identified as insertion advantagous at Day 10 of the STSE (Table 5.9, Section 5.7). In addition, genes which were identified as “insertion detrimental”, suggesting that their function was required for fitness under our selective conditions, as expected, did not correlate well to the results of our evolution experiment (Section 5.7.1). Further to this, manual inspection of the TraDIS data and experimental validation of these insertion advantagous genes demonstrated that loss of function of the majority of these genes was associated with advantagous fitness (Chapter 6).This observation is consistent with previous findings which identified “insertion advantageous” genes to be correlated with advantagous loss of function mutations (Hottes *et al*., 2013, Goodarzi *et al*., 2010) and supports our hypothesis that loss of function adaptations will be identified within the scope of the STSE.

In addition, this study also identified an example of gain of function, with this study, observing an unusual insertion pattern, including insertion orientation bias, in the 3’ end and in the promoter region of *arcA* (which included the gene *yjjY*). This observation suggested polar effects of these insertions were causing alteration of expression of *arcA*, but not loss of function as no insertions were present within the gene itself (Section 5.6.2.3). Further experimentation also confirmed this by showing that loss of function of these genes indicated a disadvantageous fitness (Figure 6.3.2), while insertions isolated from the library demonstrated an advantagous fitness (Figure 6.3.3). This study therefore shows that STSE-like experiments can not only identify loss of function mutations, but also can identify mutations causing a gain of function which has a selective advantage.

However, there was also a set of genes identified only in the evolution experiment which were nonessential but were not identified in the STSE. We propose three reasons why these genes were not seen in the STSE:

1) In the evolution experiment, adaptions in these genes in isolation, were either neutral or only caused a slight fitness advantage. Therefore, within the conditions of the STSE, these genes may be identified if the STSE was continued for a longer duration.
2) Alternatively, these genes may never be identified within the STSE. Since growth of a transposon library can be considered as a competition experiment, it is possible that insertions within these genes could be outcompeted by fitter mutants that are already present (Langridge *et al*., 2009, van Opijnen *et al*., 2009).
3) The mutations selected in the evolution experiment may be gain of function mutations. Although this study demonstrated that insertions that gain of function can be detected in some cases in the STSE (Sections 5.6.2.3 + 6.3), it is unlikely that all gain of function mutations will be identified in a STSE-like experiment. For example, in the evolution experiments described by Johnson *et al*. (2014) and Anand *et al*. (2020) nonsynonymous mutations with a specific effect were seen which caused the constitutive activation of a regulator and consequently its regulon. Replicating the precise effect of such a mutation using an insertion, is unlikely,

although as noted above in the case of *arcA*, insertions may be able to produce similar phenotypic effects to gain of function mutations by affecting the expression downstream genes.

As the precise effects of adaptions that were only identified in the evolution experiment were not considered in this study, we do not know whether mutations in these genes were loss of function or gain of function. However, as discussed in section 1.3.3, we expect loss of function mutations to be more commonly selected for in an evolution experiment. Considering that the overlap observed between the STSE and the evolution experiment was mostly in genes which are "insertion advantagous", we can infer the type of mutations identified in our evolution experiments. These results thus suggested that loss of function mutations played a key role in contributing to advantagous fitness. However, as noted above, of the three key genes which were mutated in at least four out of the five independent populations in the evolution experiment, this study found that mutations in two of these, *arcA* and *rpoA*, were likely to be gain of function, with only mutations associated with *cytR* being loss of function. Therefore, although loss of function mutations are a more common adaptive strategy, gain of function mutations do play an important role.

7.3.2 Most of the overlap with the evolution experiment occurred in the pH 7 condition of the STSE

The STSE experiment had two conditions, one of which replicated the conditions of the evolution experiment using unbuffered LB at pH 4.5, and the other acted as a control condition using unbuffered LB at pH 7. This was in an attempt to identify genes specifically associated with low pH. Genes identified under the conditions of the STSE were grouped according to whether they were found at pH 7 only, pH 4.5 only, or both conditions. Unexpectedly the largest overlap with the findings of the 5 month evolution experiment was observed within the pH 7 only group, followed by both conditions, with the smallest overlap being with the pH 4.5 only group (Figure 3.7.1 - 3.7.4). This suggested that our evolved populations had adapted to a condition than the condition (low pH) that it was initially selected under. As discussed in Section 7.2, results presented within chapter 3 of this study suggested that under the conditions of the evolution experiment using unbuffered LB at pH 4.5, adaptation of the evolved strains and populations had not occurred specifically to low pH, but rather to other stresses that are present within the experimental conditions.

It is not clear why some genes identified in the evolution experiment were identified only in the "pH7 only" group of the STSE, considering that the stresses present during growth in unbuffered LB at pH 7 and at pH 4.5 are similar. However, there are some differences in these conditions. For example, this study showed that when starting growth in unbuffered LB at pH 7, a population will be exposed to alkaline conditions for longer than a culture where growth is started in unbuffered LB at pH 4.5 (Figure 3.3.1). In addition, a stronger selection pressure is indicated in the STSE conditions at day 10, where a larger drop of insertions in genes occurs in the pH 7 cultures than the pH 4.5 cultures (Figure 5.4.1 + 5.6.3). Overall, these observations suggest that mutations in the genes identified in the STSE at pH 7 only which overlap with those identified in the evolution experiment may represent adaptations to a stress, which is stronger in the cultures started at pH 7 than those started at pH 4.5, although it is still present for these cultures. Therefore, we would predict that increasing the duration of the STSE may allow these genes to be identified also in the cultures grown in unbuffered LB initially at pH 4.5. It is possible however that although insertions within these genes confer a fitness advantage under a stress present at pH 4.5, these insertions may be outcompeted by other mutations in the TraDIS population which are fitter, and therefore they may never be observed. This is further discussed in the limitations section (Section 7.4)

7.3.3 Additional sets of genes are identified by the STSE.

This study also identified a set of insertion advantagous genes which were not found in the five month evolution experiment. When these genes were deleted from the *E. coli* K-12 ancestor, we found that the majority conferred a fitness advantage under the relevant experimental conditions (Figure 6.2.1-6.2.3, 6.3.2 + 6.4.4). While a simple explanation for why mutations in these genes were not seen in the evolution experiment is that mutations within these genes may have occurred but due to clonal interference they may have been outcompeted. Alternatively, they may have been identified if the experiment had been continued for longer or more independent replicates done. This explanation is perhaps unlikely, as lower frequency mutations already occuring within our evolved populations identified by breseq did not correspond to these genes (Figure 5.7.3 – 5.7.4). An alternative explanation could be that in an environment where different stresses occur transiently, a mutant will encounter both negative and positive selection at different timepoints. Results from competition experiments demonstrated that, while loss of function of these genes, ultimately did show a fitness advantage after five days of passaging, in the first day of competition a loss of fitness was observed (Figure 6.2.1 + 6.3.2). Therefore, in an environment where positive and negative selection exist , while in an evolution experiment a mutation may arise within a single cell but be removed by negative selection. In a transposon library, since a proportion of mutants already exists in the population, the possibility exists that some of these mutants will survive the negative selection, and they can then be acted on by positive selection which will cause them to accumilate until the next round. What selection pressure might be is unknown, but within the dynamic pH condition of our evolution experiments two stresses which could potentially act as negative and positive selection are alkaline and acid stress. Previous evolution experiments have demonstrated that these two stresses are antagonistic in terms of fitness, showing that in strains adapted to one stress (acid or alkaline) a trade-off is observed to the other stress (Hughes *et al*., 2007, Harden *et al*., 2015, Hamdallah *et al*., 2018).

7.4 Limitations of the use of TraDIS in evolution experiments

As discussed in the introduction, TraDIS and TIS in general are powerful tools which can report the location and relative frequency of every insertion present within a Transposon library simultaneously (van Opijnen *et al*., 2009, Langridge *et al*., 2009). A limitation of TraDIS is (as previously discussed) that it cannot be used to characterise essential genes, other than by identifying them. This limitation also transfers to our technique, with the STSE being unable to identify essential genes that were mutated in our evolution experiment, such as *rpoA, rpoD* and *prfB* (Figure 5.7.2). However, essential genes can have insertions present in regions which will not completely disrupt gene function, typically at the 3' end of a gene (Goodall *et al*., 2018). Therefore, the limitation described above may not apply to all essential genes, as if an insertion is present within an essential gene, its effects on fitness can still be characterised Therefore, these genes may be identified within the STSE, as was shown this study showed such as in the genes r*poC* (Section 5.6.2.5) and *der* (Section 5.6.2.6).

An STSE-like experiment is essentially a competition experiment on a large scale. Of the mutations initially put into the experiment in the form of a transposon library, the relative frequencies of insertions will change and in some cases may decline to zero, but no novel mutations will be detected. Random mutation and selection will still be on going in these experiments, however due to the analytical method being employed, only changes in insertion frequency will be detected. This highlights a limitation mentioned previously in that only the fitness effects which are caused by insertions will be considered, and therefore STSE-like experiments may miss advantageous phenotypes which are identified within laboratory evolution caused by other types of mutations such as nonsynonymous mutations and duplication events.

In addition, other evolutionary processes that occur in a laboratory evolution experiment may also not be considered in a single round of an STSE-like experiment. In particular potential effects that arise when multiple mutations are present within a single genotype, such as historical contingency, where a mutation's contribution to fitness is dependent upon other mutations and the previous genetic background the mutation occurs in. For example, within the LTEE a novel phenotype of *E. coli* being able to utilise citrate aerobically was observed. It has been demonstrated that using evolutionary re-runs from different timepoints of the LTEE that a citrate utilisation phenotype could only be recreated using clones isolated from the 20,000 generations or later. Re-runs using earlier generations unable to recreate this phenotype. This shows that a particular genetic background was required in order for the phenotype to arise (Blount *et al*., 2008, Blount *et al*., 2012). An STSE-like experiment, where the transposon library consists of a population where each cell contains a single insertion, is equivalent to an evolution experiment where one round of mutation has occurred. As such, determining the epistatic effects of two or more mutations is not possible in a single round of an STSE-like experiment.

7.5 Future work

In this thesis, the STSE was conducted for 10 days, showing an overlap with a 5 month evolution experiment. Further questions have arisen from this experiment, such as how would variation of duration of each of these experiments, STSE and evolution, affect the overlap observed between these two conditions? For example, in the evolution experiment if the different samples were taken during the experiment, would there be a particular timepoint with a better overlap with the STSE? Could investigation of the temporal population dynamics of mutations within the evolution experiment, such as which mutations arose first and how quickly they accumulated, provide a better measure to identify genes of particular interest, and how well would this relate to the genes identified within the STSE? The same questions can be applied to the STSE, such as if it were allowed to continue, although more insertion advantagous genes would probably be identified, would these genes overlap with the evolution experiment? In addition, this study predicts that there will come a point, where competition observed in the STSE will be between the fitter insertion advantagous genes and therefore this list of genes will become smaller. This study predicts that a "king of the hill" scenario may occur, where insertions in genes that cause the highest fitness advantage will compete to the point where the effects of background evolution will also be detected within the TraDIS data, with individual insertions in the insertion advantagous genes increasing within the STSE population. This will eventually result in a population represented by a "king" (or "kings" if stable subpopulations occur), indicated by a single insertion peak or peaks within the TraDIS data, acting as a tag, corresponding to a lineage of clones with an underlying genotype.

Although an STSE-like experiment could be conducted in isolation, this study also envisages the potential for a STSE-like experiment to be used as a companion tool alongside the use of laboratory evolution to understand a organism's reponse to stress. As discussed previously in Section 1.3.3.1 when using laboratory evolution as a tool to understand a reponse to stress, determining the effect that every mutation within a genotype has upon a genes function and in turn the adaptive phenotype, is not routinely considered, with many studies just focusing upon a handful of mutations, as well as identifying adaptations which are specific to a target stress and not an off-target stress. By performing parallel STSE-like experiments, these could then to be used to address these issues, using an STSE-like experiment to infer whether an adaptation was a loss of function or gain of function, based on an insertion pattern identified within insertion advantagous genes, such as within our experiment where insertions identified throughout a gene will typically indicate loss of function (Section 5.6.2), while other unusual insertions patterns, indicating polar effects, suggest gain of function (Section 5.6.2.3 - 5.6.2.6). By performing STSE-like experiments under off-target stress conditions this could then be used to identify off-target adaptations within the laboratory based evolution experiment, in addition to potentially identifying further insertion advantagous genes where loss of function may confer fitness towards a stress of interest as well as insertion detrimental genes identifying genes whose function is required for the stress response. Overall, by using STSE-like experiments as a companion tool to understand stress, not only does it identify addition gene targets, but has the opportunity to provide further characterisation of a laboratory evolution experiment, without the need for much extra effort.

As stated previously, one of the limitations of TraDIS and in turn the STSE is that in the majority of cases, fitness effects involving essential genes cannot be determined. Recently a method has been developed, termed TraDIS-Xpress, which as described in section 1.4.4 uses a modified Tn5 transposon containing an outward facing *tac* promoter, which when induced is able to cause overexpression of the neighbouring gDNA (Yasir *et al*., 2020). If this transposon inserts upstream of a gene and is in the correct orientation and ITPG is present, overexpression of the gene can be achieved, and by tracking the relative frequency of insertions in libraries grown with and without IPTG, its effect on fitness can be determined. Therefore TraDIS-Xpress is able to provide a potential method to assay the fitness effects of essential genes, as well as combining an overexpression and gene disruption library into one experiment. As such, the use of TraDIS-Xpress in an STSE-like experiment in addition to identifying genes identified by conventional TIS technologies, will also identify genes, including essential genes, whose expression when affected may cause advantagous or disadvantageous fitness.

In this study the STSE used a transposon library and TraDIS, to identify key genes present in an evolution experiment. However, it should be stated that an STSE-like experiment is not just limited to TIS methodologies. For example, we believe similar effects could be achieved through the use of CRISPR interference libraries. This method uses dCas9, a version of Cas9 which had had its endonuclease activity removed, as such when dCas9 is targeted to a gene of interest using a small guided RNA (sgRNA) it is able to act as a repressor, as it is thought to prevent RNA polymerase from binding (Bikard *et al*., 2013). Genome-wide screens are therefore possible using whole arrays of sgRNAs which can be designed and synthesised to target every gene within a given genome and which are placed into a plasmid vector and transformed into a strain of interest (Gilbert *et al*., 2014). High throughput sequencing can then be employed to simultaneously identify and map the location and relative frequency of every single sgRNA present in the library. CRISPR interference does have its limitations, such as off target effects and a larger upfront cost than TIS methodologies since the sgRNAs need to be designed and synthesised (Zhang *et al*., 2021). Like TraDIS-Xpress it is also able to identify the fitness effect of changes in expression of essential genes: by using an inducible promoter to express the dCas9, repression effects on gene expression can be tuned, therefore not completely silencing an essential gene's function (Li *et al*., 2016, Zhang *et al*., 2021).

As highlighted in section 1.3, the majority of laboratory evolution experiment have been conducted manually using serial transfer of batch cultures. This method does create issues, firstly by generating a dynamic environment, in part due to the changes in population density creating stresses, such as nutrient limitation, which therefore has the potential to cause adaptations to off-target stresses, as well as being labour intensive and with the potential to introduce human error. However, these limitations can potentially be overcome through the use of automation (Wang *et al*., 2010, Lee and Palsson, 2010, Du *et al*., 2020, Anand *et al*., 2020). By automating a laboratory evolution experiment, it firstly can allow cultures to be kept in exponential growth, maintaining a constant environment, and removing all limiting factors from the environment. In particular this could also remove the physical and molecular changes that occur when a culture enters stationary phase (Gaimster and Summers, 2015, Jaishankar and Srivastava, 2017, Gross *et al*., 2020). Therefore, really the environment is made simpler removing a lot of the off target stresses associated with batch culture as well as increasing the number of generations observed over time. There is also the potential to conduct more experiments ,with some automated systems using multiwell microtiter plates to perform evolution. Considering this, the use of automation coupled with the STSE or laboratory evolution, could potentially remove the majority of off-target adaptations seen within our experiment, focusing on the target stress only. In addition, since more generations can be conducted within a shorter timeframe, this can allow for results of the STSE or laboratory evolution to potentially be identified in quicker time.

7.6 Concluding remarks

Overall, this thesis demonstrated that the use of the STSE over 10 days, is able to identify most of the key genes found in a 5 month evolution experiment. In addition, this study has identified further key genes associated with fitness under a given condition, including insertion advantagous genes which in the majority of cases indicate the loss of function of these genes confers a fitness advantage, as well as insertion detrimental genes, which normal TraDIS approaches will usually identify. Therefore, when considering understanding an organism's response to stress, this study suggests that an STSE-like experiment may be a better tool to provide insight than conventional laboratory evolution experiment, potentially providing more information and in a significantly shorter timescale

Appendix

Appendix 1: A consideration of TraDIS using illumina sequencing

8.1 Overview of sequencing by synthesis.

The TraDIS described here utilizes an illumina Miseq to sequence transposon libraries. This sequencing method is described as Sequencing by Synthesis. In brief, sequencing by synthesis involves taking prepared single stranded DNA libraries which have illumina specific adaptor at both 5’ and 3’ ends and anneals them to a flow cell which has the complementary adaptor sequence already immobilized on the surface. Once annealed these sequences are then amplified and the complementary strand removed to create localized “cluster “of clonal sequence for each DNA fragment. Sequencing by synthesis uses fluorescently tagged terminator nucleotides, which can block replication due to the location of the fluorophore tagged to the nucleotide. In order to distinguish between nucleotides, each nucleotide has a different fluorophore which, when excited, fluoresces at a different wavelength. Sequencing by synthesis can then begin by washing a combination of universal sequencing primer, DNA polymerase and fluorescently tagged nucleotides over the flow cell containing clusters of clonal DNA fragments. The first base of each cluster can be then sequenced, as the fluorescent nucleotide is incorporated into the synthesized DNA fragment and unbound nucleotides removed by washing. Each nucleotide incorporated into a clonal cluster can then be determined by exciting the fluorophore with laser and imaged accordingly. The fluorescent molecule is then enzymatically cleaved, allowing the DNA polymerase to continue with replication and the next cycle of fluorescently labeled nucleotides being washed over the flow cell. Since the clusters are fixed to the flow cell, each cluster is individually tracked to generate a sequence. The amount of cycles performed in one run is usually determined by the reagent kit provided, for the Miseq 50, 300, 500 and 600 cycle kits are available. These cycle numbers also correspond to the maximum length of sequence which can be returned for each cluster.

8.2 How TraDIS can affect illumina sequencing

As a sequencing run progresses images acquired are analyzed in real time in order to call the sequence accurately and confidently for each cluster. This method requires that the sample present will have diversity, meaning there will be a proportion of clusters fluorescing for each nucleotide during each cycle. The ability for the software to distinguish and accurately call each nucleotide in a sequence is important. Therefore, during each sequencing run several quality control steps are performed which in turn are used to provide confidence on the base called and these are typically reflected in quality scores which are provided for each sequence read.

It is during these quality control steps where transposon sequencing can be an issue. As mentioned above illumina software is designed to have a degree of diversity in sequence being called by each cluster. This diversity is necessary for performing quality control in particular within the first 25 cycles, where the majority of quality control steps occur. However, as described, TraDIS involves the amplification of the transposon in fgDNA to enrich for transposon containing DNA fragments. In our method, as part of the transposon is also sequenced, this means that for the first part of each fragment sequenced, the sequence will be homologous and therefore during each cycle the pattern of nucleotides fluorescing for each cluster will be uniform. Due to this low diversity, difficulties with quality control will arise, which in turn, can result in reduced accuracy of each sequence called, reduced yield as the number of clusters determined to produce viable signal will be reduced and can even result in failure of the run.

8.3 Solutions in our method for TraDIS

Usually, the recommendation from illumina for low diversity libraries is to mix sample DNA with a high diversity control phiX at larger ratio than usual. However, this could also reduce yield of reads corresponding to samples as more clusters would correspond to the control. As an alternative, 2 steps have been introduced in our protocol which introduce diversity but reduce the amount of yield loss. Firstly, a custom inline barcode is introduced when preparing our library on the transposon side of the sequence; these barcodes are of different length and therefore introduce a stagger in which cycle the transposon sequence will be called. This is to prevent all clusters calling the same base at the same time, and this introduces diversity. These barcodes range from 6-9bp and have variants which differ in sequence to prevent the sample low diversity from occurring within the barcode (Table 2.1.4). This variation can also allow for additional multiplexing of each sample. A typical run will aim to have 25% of each sample corresponding to each length barcode to introduce some level of diversity in this sequence. Secondly the run is loaded with less DNA than recommended, 16pM instead of 20pM typically used in a normal run for a Miseq 150 V3 cartridge. Therefore, the cluster density during a run is reduced from 1200 -1400 K/mm^2 (K – 1,000) to aim for a lower cluster density of 1000 K/mm$^{2.}$.

Appendix 2: Considering relative fitness

Growth curves as shown in section 3.5 indicate that there are subtle differences that can be observed between the evolved strains and MG1655 ancestor in terms of growth. Growth curves, however, will only describe the growth dynamics of each strain when grown alone. Therefore, when trying to determine whether a strain is fitter than its ancestor, since these strains are not in direct competition only an indirect comparison can be made. An alternative way to assess each evolved strain's relative fitness is through competition experiments. These competition experiments provide a method of directly competing two strains together to ascertain the relative fitness of one strain compared to other. In such experiments, the two strains are mixed together in equal amounts and competed, under conditions defined by the user. Counts of each strain can then be obtained at different timepoints within a mixed culture and the CFU/ml calculated, using a selectable or screenable marker to distinguish between strains.

Once CFU/ml for each strain at each timepoint has been obtained, a comparison can then be made to determine whether one strain is fitter than the other. This comparison usually results in a single metric being calculated which describes fitness of strain A relative to strain B. However, within the literature there are several different 'relative fitness' metrics available with each calculating a metric which reports fitness differently. In addition to this, different metrics make different assumptions about the data, which can lead to different interpretations of results from the same data generated from a competition experiment. Therefore, in order to choose the right fitness metric for this study, a comparison of five different relative fitness metrics was done, to ensure that the most appropriate fitness measurement was used in this study. An overview of these metrics and the equations used in this analysis in be found in Table 2.9 within the Methods section. An abbreviated version of this has been provided in Table 9.1 for the readers convenience.

9.1 Data used in the assessment

In order to effectively assess each relative fitness metric, data from a competition experiment provided from this study was used. This data relates to a competition of evolved strains competed with KH001 (an MG1655 lac$^-$ derivate) in 50ml of unbuffered LB at pH7; the conclusions of this competition can be found in section 3.6.2. As an overview of this experiment Figure 9.2.1A shows the relative proportions of evolved strains at time 0 and 24 hours. A control of MG1655 in competition with KH001 showed no change in relative proportions between time 0 and 24 hours, indicating that there was no difference in fitness. The evolved strains and KH001 at time 0 were of equal proportions within mixed culture, but the proportions of the evolved strains then increased after 24 hours, indicating that the evolved strains were fitter than KH001.

In addition to this, in order to explore the parameter space of each relative fitness metric, a synthetic dataset was generated which modelled competition of two strains A and B. As with competitions, at time 0 each strain was set to be mixed in equal amounts (10^6 CFU/ml); the final populations varied between strains. Since all relative fitness metrics calculate a value that they define as “fitness” which represents logarithmic changes occurring within a competition experiment, final populations were generated on a 2 fold scale considering a competition experiment where one strain’s growth would be constant, defined by 8 doubling events of the population (2^8), while the other was varied to simulate growth or decline of that strain by varying the power (2^x). This created a synthetic dataset which could be used to investigate each metric’s ability to report changes in final populations on a logarithmic scale. Since each relative fitness metric performs a comparison of one strain in terms of the other, data was generated for where strain A would be compared in terms of strain B, in two scenarios considering the final population of one strain growing at a constant rate , while the other varied (Figure 9.2.2). Equation 9.1 illustrates the methods used to calculate final populations for each scenario. By considering the variation of two strains separately, the effect of how each strain within the comparison would affect the outcome of each metric could be explored.

Equation 9.1: Generation of final populations for hypothetical data, 2 scenarios where one strain was kept constant and the other varied based on a scale factor of 2. X indicates the variable that was changed.

$$\text{Scenario 1:}\ \ \mathrm{A_t} = \mathrm{A_0} \times 2^{\mathrm{x}}\,, \qquad \mathrm{B_t} = \mathrm{B_0} \times 2^{8}$$

$$\text{Scenario 2:}\ \ \mathrm{A_t} = \mathrm{A_0} \times 2^{8}\,, \qquad \mathrm{B_t} = \mathrm{B_0} \times 2^{\mathrm{x}}$$

9.2 Evaluation of relative fitness metrics.

Table 9.1: Equations used to determine relative fitness, shorted version of Table 2.9

Relative fitness metric	**Equation**	**Reference.**
Equation 9.2: Relative proportion of strains (RP)	$RP = \frac{A_t}{A_t + B_t}$	N/A
Equation 9.3: Malthusian growth model.	$N_t = N_0 e^{rt}$	(Lenski *et al.*, 1991)
Equation 9.4: Malthusian parameter (r).	$r = \ln\left(\frac{N_t}{N_0}\right) day^{-1}$	(Lenski *et al.*, 1991)
Equation 9.5: Relative fitness (W).	$W = \frac{r_a}{r_b} = \frac{\ln\left(\frac{A_t}{A_0}\right)}{\ln\left(\frac{B_t}{B_0}\right)}$	(Lenski *et al.*, 1991)
Equation 9.6: Selection rate (S)	$s = r_a - r_b$ $S = \ln\left(\frac{A_t}{A_0}\right) day^{-1} - \ln\left(\frac{B_t}{B_0}\right) day^{-1}$	(Travisano and Lenski, 1996)
Equation 9.7: Relative fitness estimated using Selection rate (s) between two strain A and B:	$\mathrm{W} = 1 + \frac{s}{r_{ab}}$ $\mathrm{W} = 1 + \frac{\ln\left(\frac{A_t}{A_0}\right) day^{-1} - \ln\left(\frac{B_t}{B_0}\right) day^{-1}}{\ln\left(\frac{A_t + B_t}{A_0 + B_0}\right) day^{-1}}$	(Lenski *et al.*, 1991)
Equation 9.8: Selection Coefficient 1	$S_{(1)} = \frac{r_a - r_b}{r_b} = \frac{\ln\left(\frac{A_t}{A_0}\right) - \ln\left(\frac{B_t}{B_0}\right)}{\ln\left(\frac{B_t}{B_0}\right)}$	(Barrett *et al.*, 2006)
Equation 9.9: Selection Coefficient 2	$S_{(2)} = \frac{\ln\left(\frac{A_t/B_t}{A_0/B_0}\right)}{log_2\left(\frac{A_t + B_t}{A_0 + B_0}\right)}$	(McDonald, 2019) Box 1

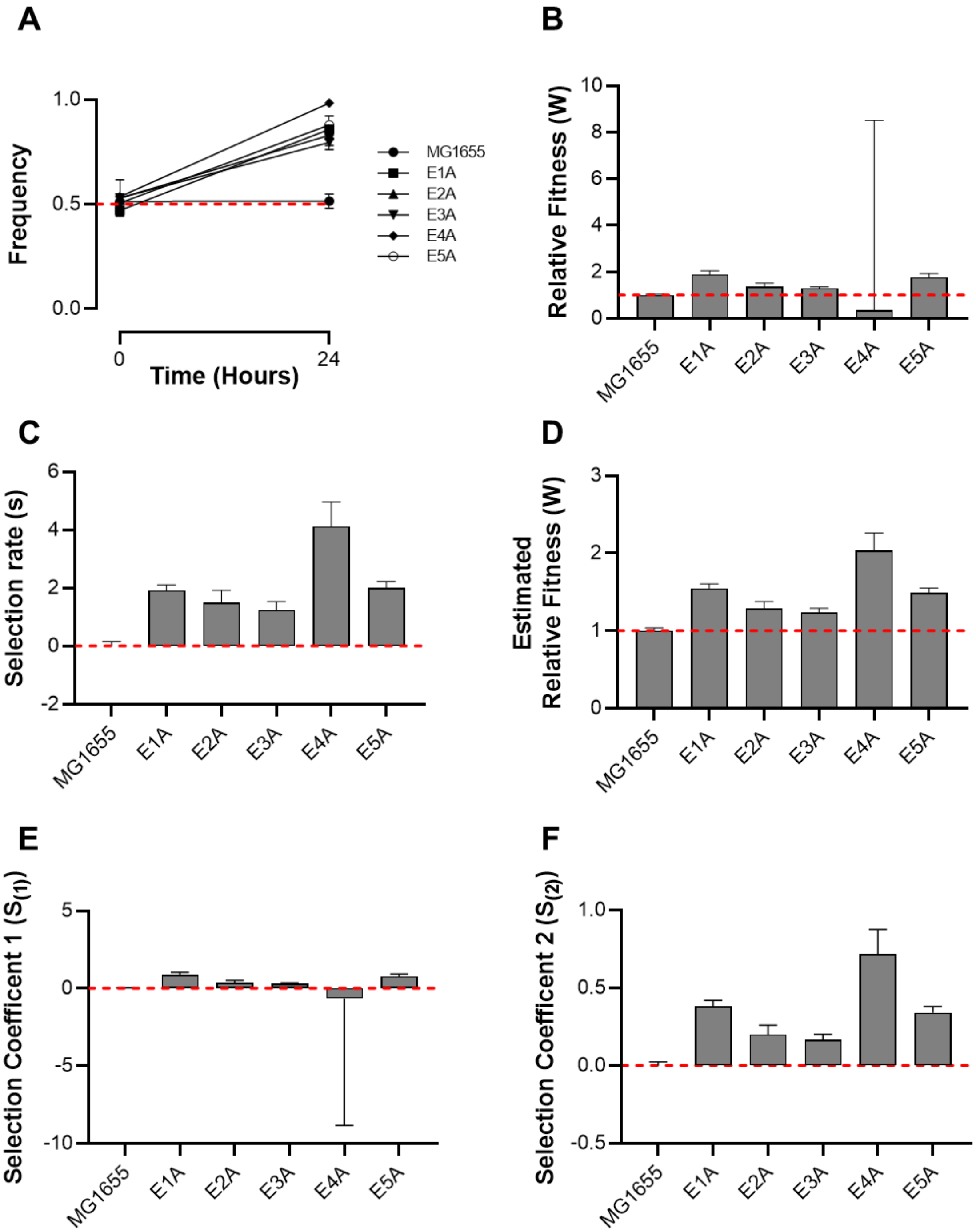


Figure 9.2.1: **Different measures of fitness.** Data from competitions of Evolved strains and MG1655 ancestor competed against KH001 in 50 ml LB at pH7 for 24 hours. Red dashed line indicates the value expected if no change in fitness was observed between two strains. Equations used to determine fitness can be found in the Table 2.9 and Table 9.1 above).**A)** Relative frequency of evolved strain in a population at each time point Table 9.1: Equation 1). **B)** Relative fitness value (Table 9.1: Equation 4).. **C)** Selection rates (Table 9.1: Equation 5). **D)** Relative fitness values which have been estimated using selection rates (Table 9.1: Equation 6). **E)** Selection co-efficient 1 based on relative fitness (Table 9.1: Equation 7). **F)** Selection co-efficient 2 based on ratio of relative frequencies over generation time (Table 9.1: Equation 8).

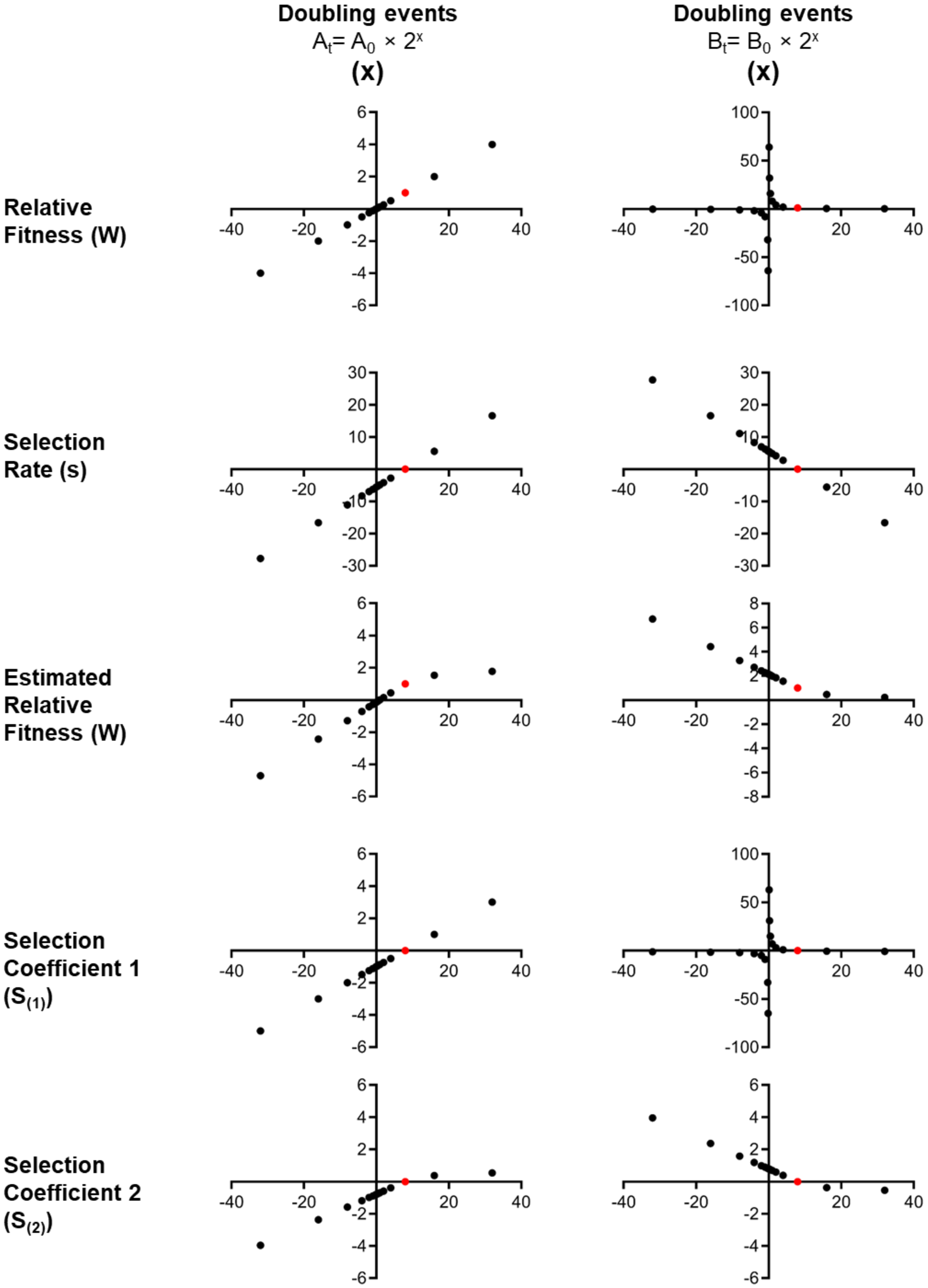

Doubling events
$A_t = A_0 \times 2^x$
(x)
Doubling events
$B_t = B_0 \times 2^x$
(x)
Relative Fitness (W)
Selection Rate (s)
Estimated Relative Fitness (W)
Selection Coefficient 1 ($S_{(1)}$)
Selection Coefficient 2 ($S_{(2)}$)
-40
-20
20
40
6
4
2
-2
-4
-6
100
50
-50
-100
30
20
10
-10
-20
-30
8
6
4
2
-2
-4
-6
-8

Figure 9.2.2: Testing different methods of determining fitness. Simulated data were generated by creating a set of hypothetical data comparing strain A in relation to strain B. The value reported by each relative fitness metric is described in the Y axis of all graphs. Two sets of hypothetical data were created by defining each strain as having an initial population of 10^6 CFU/ml cells. In the two data sets the final populations for each strain was then determined in the amount of 2 fold increments/decrements. In each dataset one strain had a fixed final population defined by 8 doubling events while the other strain the final population was altered by altering the amount of 2 fold increments/decrements. The increments/decrements used in the varying strains are defined by the X axis. The equations at the top of the figure indicate the strain which was varied in each data set, strain A (left graphs) and strain B (right graphs). Strain A and B refer to the fitness metric equations described in Table 9.1

9.2.1 The Malthusian growth model and the Malthusian parameter (r)

Of the 5 relative fitness metrics identified, 4 of these measurements determine fitness by estimation of the relative growth rates. Typically, the simplest way to determine growth rate is to assume a population follows the Malthusian growth model, otherwise known as the simple exponential growth model, which considers an organism to be growing constantly in an exponential way (Table 9.1: Equation 2). From this equation the growth rate can be estimated. Practically this could be done using CFU/ml counts at the beginning and end of a competition experiment, by determining what is known as the Malthusian parameter (r,Table 9.1: Equation 3). Since the Malthusian parameter is generated assuming exponential growth, any changes in population that are a consequence of differences in lag or stationary phase, cannot be determined separately and therefore are also reported within the parameter. Therefore, relative fitness metrics which use the Malthusian parameter to generate metrics have an underlying assumption, which is that any change in population between two strains will occur when the populations are in exponential phase (Section 9.2.2– 9.2.6).

Considering a competition experiment, these experiments can be conducted using batch cultures (indeed this is true of all the competition experiments performed in this study). Using batch cultures, the strains will not be in constant exponential phase, instead will undergo the different phases of growth (lag, log, and stationary phase). Therefore, when using a Malthusian parameter within a fitness metric the assumption above is broken.

However, this does not mean that the metric is useless, it indicates that the fitness reported is not a true estimate of growth rates.

9.2.2 Relative fitness (w)

Relative fitness (w) is a metric which describes fitness as the ratio of two Malthusian parameters within a competition experiment. Since a ratio is used, the metric is dimensionless. (Table 9.1: Equation 4). This metric was championed by Richard Lenski to ascertain fitness within the LTEE, and is used predominately as the gold standard to report fitness within competition experiments (Lenski *et al*., 1991). Using this metric, if fitness is reported with a value of 1 it indicates no difference in fitness between the two strains, with anything larger than 1 representing an increase in fitness and less than 1 a decrease in fitness (see Figure 9.2.2)

However, by calculating fitness as a ratio, this introduces a limitation of this metric which is that within a competition both strains must undergo growth, i.e the final populations of each strain must be larger than their initial populations. In our hypothetical data, this limitation was also identified, as although relationship between a population decline in strain A, showed a linear relationship with fitness, when a final population decline in strain B occurred an inverse relationship between 'fitness' and logarithmic growth was observed. This 'inverse of fitness' occurs when a Malthusian parameter of strain B is less than 0, introducing a negative value in the relative fitness which implies that strain A is less fit than strain B, when in reality this should be the opposite. A negative Malthusian parameter can only be introduced when a population is in decline i.e. the ratio of final to initial population size is less than 1. The effect of this limitation can also be clearly seen within our competition data, as although all the evolved strains had a positive fitness (w > 1), in experiments with E4A there was one replicate where KH001 was in decline introducing a negative relative fitness value. The effect of this can then be seen in the E4A data with large size in the error bars ().

It is important to note here, that a decline in population size could be due to a biological reason, but also a decline could be introduced artefactually through limitation in sampling sizes. For example, in our study, counts were determined by plating a 10 fold dilution series, counting the dilution where 30 – 200 colonies were present. By considering a situation like

E4A where 98% of colony counts corresponds to one strain, if 100 colonies were counted at a 10^6 fold dilution only 2 would correspond to the competing strain. Due to the count being so small, an addition of a single colony to that count (which could happen easily with random sampling) would cause a 1 x 10^6 change in final CFU/ml calculated and this would be sufficient to change the measured behavior of the population from growing to being in decline.

9.2.3 Selection rate (s)

Selection rate (s) reports the difference between the two Malthusian parameters of the two strains in competition, creating a metric which is defined by units of inverse time (Table 9.1: Equation 5). Also used by Lenski, by this method a value of 0 indicates no difference in fitness between two competing strains, with anything greater than or less than 0 representing advantageous or detrimental fitness (Travisano and Lenski, 1996). Unlike the measure of relative fitness referred to above, this metric reports the differences of the growth rate, therefore unlike relative fitness no assumption had to be made, that both populations are growing exponentially. Indeed, this can be seen in our hypothetical data, where a linear relationship between growth and fitness is seen (Figure 8.2.2). In addition, selection rate, gives a result more in line with what would be expected within our evolved strain data, with all evolved strain being shown as fitter and E4A showing the largest fitness ()

9.2.4 Relative fitness estimated using Selection rate

Since both selection rate and relative fitness use the Malthusian parameter to report fitness, these two metrics are related, and therefore an approximation of relative fitness using selection rate can be calculated. This is because relative fitness is the ratio of growth rates, and selection rate is the difference of the growth rates. By creating a ratio of selection rate and the Malthusian parameter using the total population, this would create an approximation of the difference between strains in terms of relative fitness. Relative fitness can then be estimated by adding the ratio calculated using selection rate to a value of 1, where in terms of relative fitness no difference in growth rate would be expected. This approximation allows for relative fitness to be calculated which removes the assumption that both strains should be growing within a competition experiment. Upon exploring how fitness is reported using this metric within our hypothetical data, where either strain A and B is declining, a linear relationship is observed between fitness and growth (Figure 8.2.2) However, a skew is observed when the CFU/ml of the strain which is varying increases above the strain with a constant CFU/ml, indicating the relationship between relative proportions of strains and fitness is not linear.

This is due to the calculation estimating the average Malthusian parameter of the entire population in competition. Since Malthusian parameter value represents the logarithmic ratio of the final population compared to the initial population. If one population is in decline, the effect that this would have on the overall population would be limited as there would be no logarithmic difference and therefore would not be reflected when calculating the average Malthusian parameter. However, when the population of a strain doubles, the Malthusian parameter doubles indicating that fitness determined doubles indicating that when both populations are growing a linear relationship is restored. When this metric is applied to the competition data, the fitness value does vary as expected, with all evolved strains showing fitness and E4A having the largest fitness increase.

9.2.5 Selection coefficient 1

The selection coefficient described by (Barrett *et al*., 2006), describes the increase in growth rate of one strain relative to another. As with the other methods described above, this metric calculates the Malthusian parameters of two strains. It then creates a ratio of the difference in growth rate between strain A and B, relative to the strain B. By this method, no fitness difference between strains has a value of a 0 value and anything above or below this indicates a higher lor lower fitness. However, like relative fitness, the assumption is made that both strains within the competition must be under constant growth. Again, if this assumption is broken, in particular when strain B enters a decline in population, the linear relationship between fitness and relative proportions breaks and the fitness relationship is inverted. This is also seen when this method is applied to the competition data, in particular within E4A ().

9.2.6 Selection coefficient 2

The second selection coefficient, described by (McDonald, 2019), is the only relative fitness metric identified which does not calculate the Malthusian parameter for each strain and therefore does not focus directly on the difference in growth rates. Instead it is a metric which uses the natural logarithm of the change in relative proportions strains at the initial and final time points, which is then divided by the number of generations which occurred, to create a number which has a per generation unit. By creating a parameter which is approximated using the total population, in this case the number of generations, this metric skew fitness described when a population goes into decline, due to the reduction of the total final population masking the overall decline of one strain. Therefore as with Estimation of Relative fitness using selection rate (Section 9.2.4), the skew indicates that the fitness based on relative proportions of cells is not a linear relationship. However, with our data the interpretation of fitness within the evolved strains is not affected.

9.3 Selection rate – The relative fitness metric for this study

A relative fitness metric is a measure of the fitness of one strain in relation to another. Each of the metrics tested above describe fitness using slightly different parameters measured in a competition experiment. In this study two fitness metrics, relative fitness, and selection coefficient 1, both use an assumption that in a competition experiment both strains are growing. If this assumption is broken, then the ability for each metric to define fitness can invert the relationship within a competition, where one strain which is fitter actually being measured as being less fit. As seen within our data, some populations may decline when in competition, therefore for this reason it was decided not to use these two metrics to estimate fitness.

Two other metrics, Estimated Relative fitness and Selection coefficient 2, both did not show a truly linear relationship between each strain's final proportions and relative fitness when tested using our synthetic data, Instead each metric involves an additional parameter, that included information about the total final population which in the hypothetical data provided did not change a lot when one population was in decline and therefore the effect this parameter had could only be seen when the population was increasing. In both these metrics however, this parameter of the final populations is only approximated for each strain by considering the total population and not each strain individually. Therefore, these two metrics were also not used.

The remaining parameter, selection rate, refers to only the difference in Malthusian parameters between strains, per unit of time. As such, this does not make any assumptions about whether or not within a competition experiment a strain is growing or declining. In addition, since it does not include approximation of a parameter assessing total population growth, no estimates are introduced. However, when comparing selection rates caution must be taken to consider the unit of time that the selection rate is described in. Within our hypothetical data, a linear relationship between fitness and relative proportions were observed. Since we sometimes saw potential declines within strains in our study, the selection rate metric was the most appropriate choice. Therefore, throughout the remainder of this study all competition experiments have been reported using selection rate.

Appendix 3: Considering different analysis pipelines to explore the STSE

To this date, several pipelines exist which have been designed for transposon insertion sequencing. Therefore, within this appendix the aim was to identify and test some of these pipelines using data from the STSE.

When comparing transposon sequencing samples, a typical analysis is to compare relative frequencies of insertions in one sample against the other, to identify whether or not a given insertion has dropped or increased within a sample, which will give information about the impact on fitness of the insertion. Typically, this is done by considering all insertions within a gene using a normalized read count, as this is the most relevant for these studies. Comparison of read counts between samples then allows for the creation of a fold-change metric for each gene, which can then be used to describe whether insertions have significantly declined or increased in one sample relative the other (Figure 5.6.1 10.0.1). This then allows for two different types of genes to be identified in the comparison: genes containing insertions which become more common in the population due to the insertions causing an increase in fitness, or genes where insertions are detrimental to fitness and where the strains containing them hence decline in the population.

Therefore, in analysis of the STSE, two genes' lists should be identified, the first where insertions decline in the population, and therefore have a detrimental fitness on the strain. As stated above, within the literature these are also defined as conditionally essential genes and are typically the genes which are the main focus of studies using transposon libraries (Phan *et al*., 2013, Wong *et al*., 2016, Zhao *et al*., 2017, Goodall *et al*., 2018). The second is genes where insertions have a beneficial effect on the fitness of strains that carry them. Only a few studies consider these (Langridge *et al*., 2009, Cowley *et al*., 2018, Yasir *et al*., 2020). Below, we compare different pipelines in terms of their ability to separately identify genes where insertions decline, and genes where insertion accumulate in the population. In order to keep things consistent, for all comparisons only the comparison of the pH 7 data of the STSE compared to the ITL was studied.

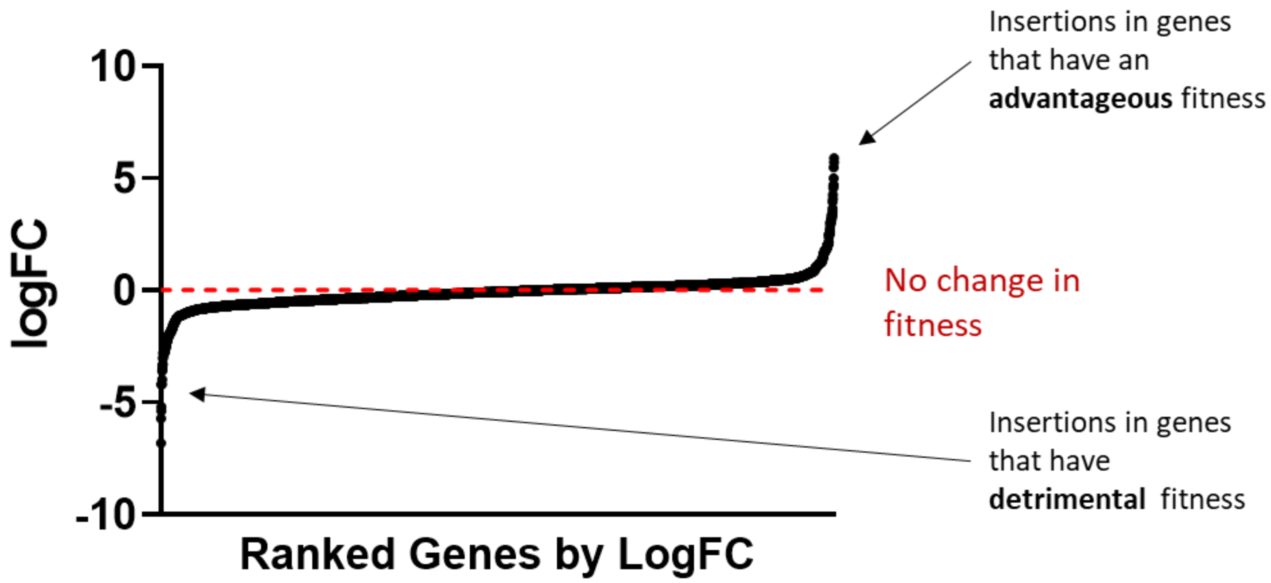


Figure 10.0.1: Overview of a typical transposon analysis, where insertion frequencies within genes are compared between samples to produce a Fold change (FC) metric which describes an increase or decrease of insertions. This then can be attributed to their effects on fitness, where insertions which decrease have a detrimental impact on fitness, and insertions which increase having a beneficial effect. This Figure is it the same as Figure 5.6.1, repeated to assist the reader.

10.1 Pipelines used within this analysis.

Of the different analyses described below, only Biotradis used unprocessed sequencing reads, which were then subsequently processed through the pipeline, although it is also able to accept pre-processed and mapped reads as input also (Barquist *et al*., 2016). Another analysis pipeline previous mentioned in this study (ESSENTIALS) was discontinued from this analysis, as comparison of the Day 10 -pH7 against the ITL was unable to generate a comparison of this sample with the STSE. As the essential gene list generated by ESSENTIALs also showed the largest variation, it was decided not to include this within the comparison (Section 4.4).

Alternatively, the reads were processed using custom scripts described in Section 2.7.4 which were designed for the method of TraDIS conducted on the STSE and then run through each of the following pipelines: edgeR, deseq2, log-likelihood and ARTIST. Descriptions of each of these pipelines are found below, with the exception of the log-likelihood method which has been described previously in section 4.4.1.1.

10.1.1 RNA-seq analysis.

With the exception of ARTIST, all the different analysis pipelines used software which was initially written to conduct differential expression analysis of RNA-seq data. Biotradis and ESSENTIALS both use the differential expression analysis pipeline EdgeR to perform the comparative analysis of two conditions (Robinson *et al*., 2010). Deseq2 has been used as an alternative within several transposon studies (Love *et al*., 2014). How these methods were run is described below.

10.1.1.1 EdgeR

Read counts were normalized by overall sequencing depth using a trimmed mean of M-values model. For each condition, normalized read counts were then fitted to a negative binomial model. Biological variation between replicates was then calculated on a gene by gene basis ("tagwise") using empirical Bayes Moderation. This moderation considers reproducibility between genes, allowing genes which are more reproducible to be ranked higher than genes which are not. The exact test within edgeR was then used to test for differentially expressed genes. Once performed, differentially expressed genes were determined using a cut off p-value of 0.05 (following adjustment using the Benjamini-Hochberg procedure). These significant differentially expressed genes were then separated into enriched genes with a positive LogFC value and conditional essential genes with a negative LogFC value.

10.1.1.2 Deseq2

Deseq2 begins by firstly normalizing the data to correct for differing sequencing sample sizes. It does this by generating and applying a size factor to each sample. This size factors is estimated as the "median of the ratio, for each gene, of its read count over its geometric mean across all lanes" (Love *et al*., 2014). A generalized linear model is then fitted on a gene by gene basis, across all samples. In order to understand and include the variability that occurs between biological replicates of the same condition, a dispersion parameter is then estimated. This dispersion parameter is estimated on a gene-wise basis using an empirical Bayes approach which provides a good estimate of dispersion for genes which have a given size in read count. The generalized linear model can then be used to ascertain LogFC and significance for any gene between two samples. Firstly, Deseq2 addresses the issue associated with the potential artefact that can arise for genes where the read count is low, where simple sampling error (the variability of which increases as the sample size decreases) could lead to artificially high LogFC values. It does this by shrinking LogFC based on the amount of information present using an empirical Bayes procedure. Therefore, low read count genes which have a high LogFC are extensively reduced to prevent false positives from arising. Statistical testing of LogFC is performed using a Wald test. The P values generated from this test were then adjusted using a Benjamini-Hochberg procedure.

10.1.2 TraDIS analysis pipelines,

10.1.2.1 Con Artist

Con-artist is part of the Artist pipeline used to detect conditionally essential and transposon enrichments in an annotation independent manner within transposon sequencing experiments (Pritchard *et al*., 2014)

The analysis was conducted in MATLAB. To begin with an aligned reads file in a SAM format for each replicate and condition of the STSE as well as the original ITL was uploaded into MATLAB. In addition, MG1655 annotations were also uploaded in a GFF3 Format. As with EL-ARTIST, the MG1655 genome was then spilt into 50bp windows. Con-artist works by comparing a control library against an experimental library and requires the maximum amount of insertions present within the control library in order to perform a normalization based on simulations of the bottleneck seen within an experimental sample. Due to this, fastq files from different replicates of the ITL were combined to create an analysis file, which in theory should contain the maximum amount of possible insertions within our transposon library. This limits the amount of comparisons as conditions can only be compared to the ITL. All reads from the ITL.sam file were then assigned to each 50 bp window. From there comparisons could be made against condition files, with each replicate loaded in and tested separately. The bottleneck observed in each replicate was then simulated 100 times within the ITL. Mann-Whitney U analysis was then performed looking for significant differences based on the bottleneck simulations of the ITL control library. This Mann-Whitney analysis was then used to train a Hidden Markov Model in order to predict whether a region of the genome is conditionally essential or has an enrichment of transposons. Predictions were then allocated to annotations on the genome, including intragenic regions, and annotations were grouped depending on whether the complete annotation, or a particular region within the annotation, was classed as conditionally essential, or alternatively transposons were enriched relative to the ITL.

Due to variation between replicates for each condition, each replicate was run separately. The predictions from each replicate were then combined into three categories: conditionally essential; no change; and enrichment. This study only looked at the annotation level and did not consider essential regions identified within an annotation separately. The replicates for each condition were then compared and genes or intergenic regions which had either full, or domain essential/enriched in all 3 replicates were used in further analysis.

10.1.2.2 Biotradis

BioTraDIS like ESSENTIALS is a pipeline designed for the easy analysis of TraDIS data. This approach allows for the filtering and removal of transposon containing fragments, which are then aligned to the genome using SMALT using default parameters. From here Bam files are then processed to produce several files relating to transposon insertions and read counts on an individual level but also on a gene by gene level. Further to this, the difference between two samples can also be considered using edgeR. This follows the same procedure described in section 10.1.1.1. A p value of 0.05 and a LogFC +/- 2 was used after FDR correction using a Benjamini-Hochberg procedure. In addition to this Biotradis is also able to identify essential genes looking at unique insertion sites; this is described in more detail in Section 4.4.1.2.

10.2 Insertion detrimental fitness/ conditionally essential genes

As stated above, all pipelines were assessed on their ability to identify genes where insertions showed a fitness advantage or a disadvantage in the conditions of the STSE. This section only describes the genes where insertions begin to decline in the population. Since as previously reported in Section 5.4.3 a large loss of insertions at Day 10 occurred for both pH conditions, determining fitness genes within these samples was pointless as the majority of genes on the genome had no insertions present within this sample (as strains containing them had fallen to too low a level in the population to be detected). So, in the comparison described below only the D1-pH7 and D5-pH7 conditions of the STSE are described. Since essential gene lists were also generated in this analysis, these essential genes were removed before the comparison was conducted.

10.2.1 Considering the differences between pipelines which use RNAseq software

As within the literature RNAseq Differential expression software was used to determine significant genes, it was decided initially to compare the different outputs, as slightly different methods are used to normalize the data and determine significance, and it was important to understand the effect of these. To keep thing consistent the same false discovery procedure was used for each analysis (Benjamini-Hochberg procedure) and the same cut-off was used, which was an FDR < 0.05 with a LogFC cutoff of -2. In addition to this, as described previously in Section 4.4, Biotradis considered differences within all annotations, while the other methods focus only on CDS annotations within the genome. Therefore, to ensure a valid comparison, 9 and 11 annotations were not considered in the comparison for Day 1 and 5 of the STSE below as they were associated with regulatory RNAs.

The overview of this analysis is provided in Figure 10.2.1. At Day1, the Biotradis pipeline identified the most significantly different genes, while with edgeR no genes were identified as significant. This was interesting as Biotradis uses edgeR running nearly exactly the same protocol. However further inspection of the detailed steps in these methods revealed that within the Biotradis pipeline, genes which had a 0 value in at least one replicate were removed from the analysis before edgeR was conducted, while with edgeR and deseq2, this removal did not happen. This would affect the calculation of statistical significance between these two pipelines and probably explains the difference in significant genes identified between these two pipelines. Indeed, the LogFC values reported by these three tools were highly similar, it was just the p-values reported which were different.

This could be better observed in the Day 5 pH7 sample, as although a large proportion of genes were identified as fitness detrimental in all three methods, there were more genes identified as significantly different by Biotradis compared to EdgeR and Deseq2. Manual inspection of genes which didn’t matched between the methods showed that for the genes specific to edgeR and deseq2, in at least one replicate no insertion was observed within that gene, while in Biotradis of the 36 genes specific to this pipeline, an general decline was observed, but not to the extent of genes identified within all three analysis methods. This suggested that the main overall difference between these methods was the stringency of the p-values generated. Indeed, when comparing the LogFC values at both day 1 and 5 these were highly similar in all analysis pipelines suggesting that it was only the statistical analysis which differed them.

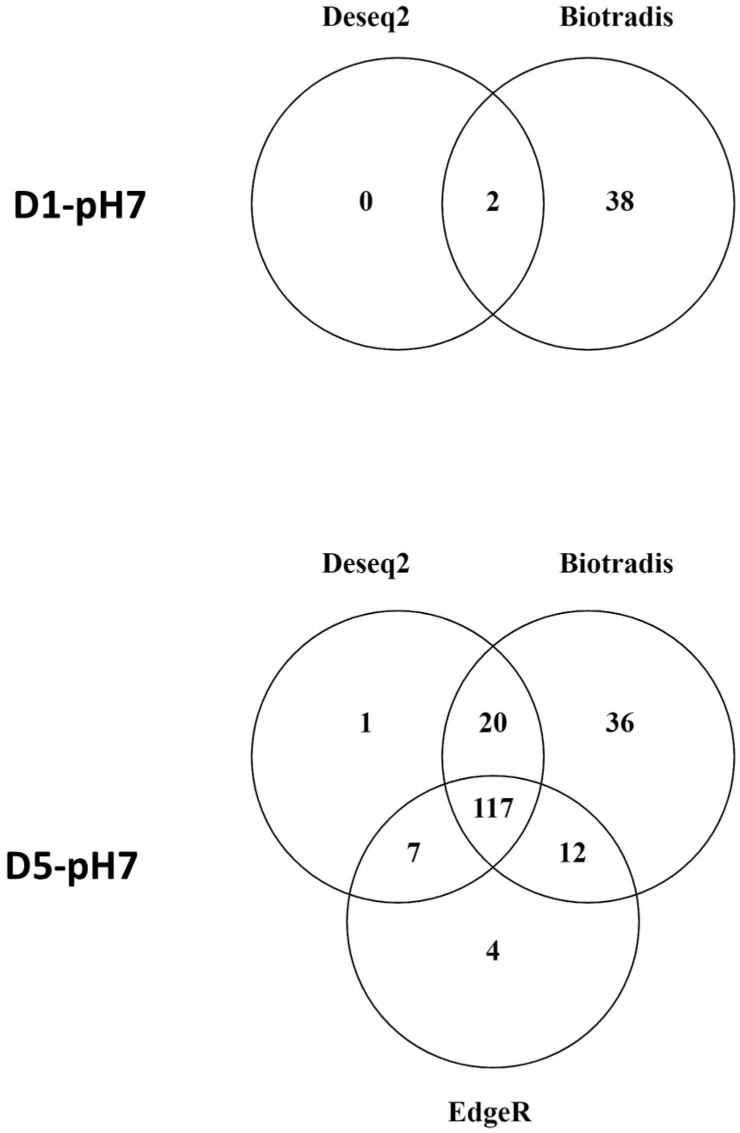


Figure 10.2.1: Comparison of pipelines which used RNAseq methods to determine inserts with a detrimental effect on fitness in the pH 7 condition of the STSE. Note that Biotradis uses edgeR to determine differential expression.

10.2.2 Looking at conditional essentiality.

The RNAseq analysis pipelines above can be used on TraDIS data to identify genes where reads have declined within the population, indicating that insertion with these genes is detrimental for fitness. However other methods exist, which only focus upon the presence or absence of insertions within genes within a sample, these same methods are used to define essential genes as discussed in Chapter 4. These methods are the Biotradis and log likelihood methods, of which Biotradis is a less stringent than the log likelihood which is highlighted in section 4.4. These methods can also be applied to find conditional essential genes, by identifying conditionally essential genes in the outgrowth experiment and then removing all essential genes identified in the ITL. However, one issues arises which is how to consider the ambiguous set of genes identified using these analysis. Since these genes can neither be defined as essential or non-essential, it is difficult to determine what they are. Therefore, to remove this problem one can define genes which decline in insertions during the condition of the STSE if they move from the non-essential in the ITL to essential within the conditions of the STSE. Again, since Biotradis uses all annotations on the genome, 11 and 18 annotations were removed from this comparison for the day1 and 5 conditions respectively, as these annotation are to regulatory RNAs which were not included in the other analysis described in the previous section.

In addition to this, the pipeline ARTIST was also considered as it has a comparative pipeline termed Con-Artist which is able to compare outgrowth experiments to the ITL to identify conditional essential genes in an annotation independent manner (see section 10.1.2.1). Due to this ARTIST is able to consider the intergenic regions between genes, therefore a further 12 and 48 annotation from the day 1 and 5 analysis were not included in this comparison, due to these annotations corresponding to intergenic regions. In addition to this it considers these regions on the genome as fully essential or domain essential, as described in section 10.1.2.1. These were both considered.

Genes identified as having a decline in insertions using these different analyses were compared to each other at Day 1 and Day 5 (Figure 10.2.2). Overall, in both these analyses, it was clear to see that the Biotradis method was a less stringent version of the log likelihood method, with every gene identified as essential in the log likelihood method being identified in the Biotradis. However as seen in Chapter 4, ARTIST identified a further set of genes unique to this pipeline as having a decline in insertions. Further inspection of these genes using a manual analysis, showed that in some cases these genes could be considered to have a conditionally essential domain (i.e., part of the gene was conditionally essential but not all of it),. while in others it was difficult to see any difference that would indicate essentiality. In addition to this the Biotradis log likelihood method generated a large number of genes unique to this analysis. Further inspection of these genes revealed that indeed some were essential, however in other cases insertions were present at low frequency, although there was still a high read count.

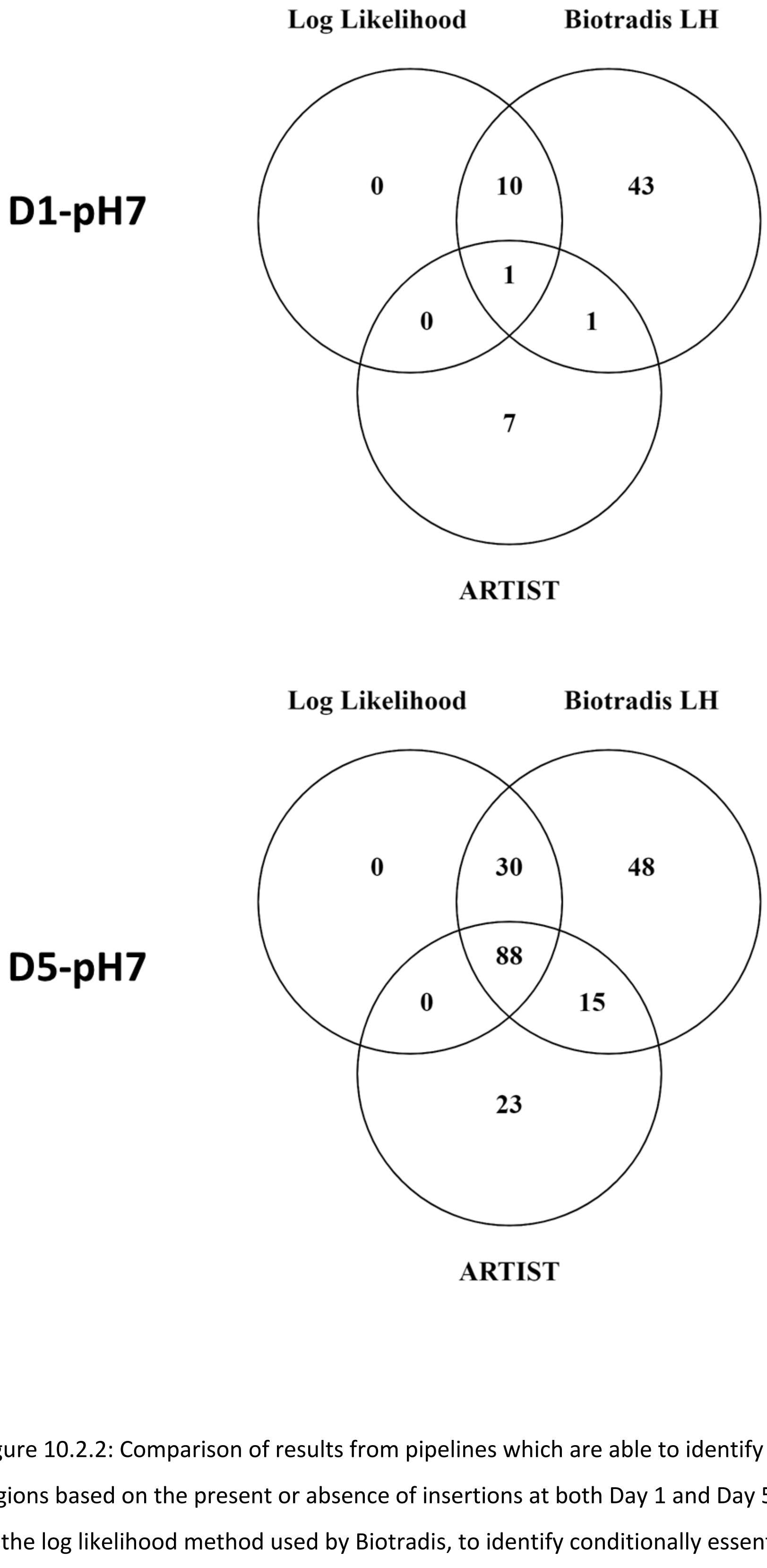


Figure 10.2.2: Comparison of results from pipelines which are able to identify conditionally essential gene regions based on the present or absence of insertions at both Day 1 and Day 5 of the STSE. Biotradis -LH refers to the log likelihood method used by Biotradis, to identify conditionally essential genes.

10.2.3 Overall comparison of the “Fitness detrimental” genes.

The analysis pipelines described in Section 10.2.1 + 10.2.2 identify genes which decline in insertion frequency following grown in a specific condition. However, the approaches that they use are towards different attributes of a transposon library. While some use read count to detect decline of insertions within a population, the others look at the presence and absence of insertion sites within a given gene. Therefore, to start to identify the differences between these two types of analysis an overall comparison was conducted (Figure 10.2.3).

This comparison highlights two points. The first is that Biotradis pipeline on its default settings is less stringent than any of the other pipelines, with both the RNA-seq comparison and the log likelihood comparison generating the most genes which were unique to that analysis at day 1 and day 5. The second is that at day 1 the RNA-seq pipelines do not detect as many differences between the ITL and outgrowth as other pipelines, with the exception of Biotradis, focusing upon present and absence of insertions. This suggests that when considering genes where insertions have a detrimental impact on fitness, different analyses are able to identify different types of these detrimental genes. Overall, however not one pipeline stood out as the most ideal pipeline.

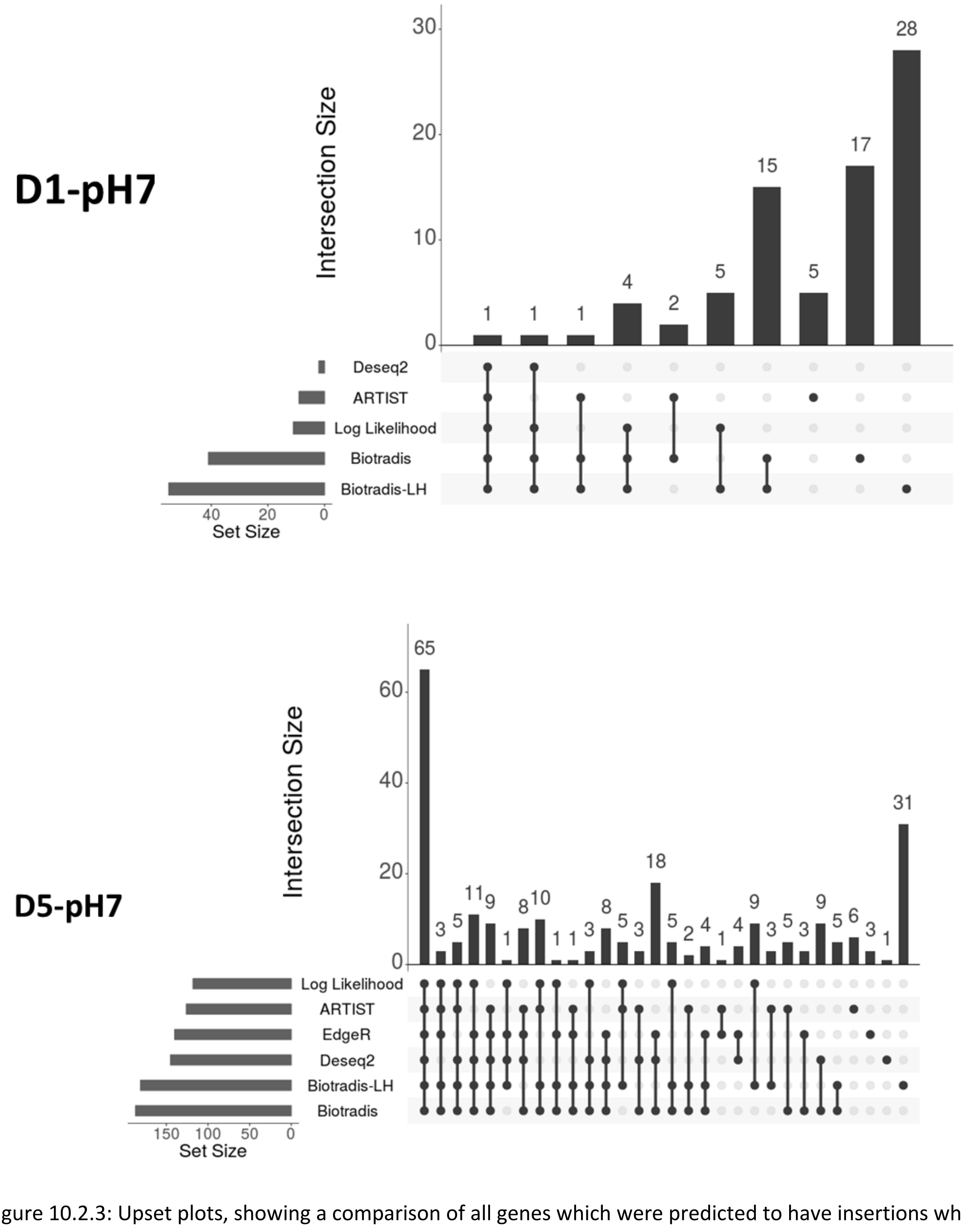


Figure 10.2.3: Upset plots, showing a comparison of all genes which were predicted to have insertions which were fitness detrimental. Each comparison was done on the datasets from cells grown at pH 7 at Day 1 and Day5 of the STSE, compared against the ITL. Biotradis pipeline analysis was split into a comparative analysis by edgeR (Biotradis) and a log likelihood analysis (Biotradis LH).

10.2.4 Considering “Insertion advantageous” genes.

The other set of genes which needed to be identified from the STSE was genes containing insertions which increased within the population, presumably due to strains containing these having a fitness advantage relative to other strains in the library. Only a few pipelines were able to identify this set, by analysis the difference between overall read count between samples. As expected, the previous analysis methods which utilize RNAseq differential expression analysis were able to do this. In addition to this the ARTIST pipeline was also able to identify regions which showed an increased in reads. Therefore, a comparison of these pipelines was conducted. As in the previous section, in the RNA-seq pipelines genes were only considered which showed an FDR < 0.05 and a LogFC +/- 2 to ensure significant difference between samples had been observed, while the ARTIST threshold is described in Section 10.1.2.1. As with the other comparisons, the Biotradis and Artist pipeline included annotations which could not be compared to the other analysis, therefore they were excluded from the results. With regards to Biotradis, these were three sRNAs at Day 5 and two srRNAs at day 10, while for Artist a total of 14 intergenic regions were excluded. Notably, all of these intergenic regions were related to the genes which were already identified.

Since enrichment could be expected at any timepoint of the STSE at pH 7, each timepoint was assessed individually. At Day 1, only one analysis pipeline, Biotradis, identified 5 genes whose insertions start to show a fitness advantage. Therefore, comparison at only Day 5 and Day 10 are presented in Figure 10.2.4. At day 5, Artist was not able to identify any enrichments within these samples, indicating that Artist may be more stringent than the other methods. At Day5 the same set of genes were identified in Deseq2, edgeR, and Biotradis, with again Biotradis showing less stringency. At day 10 however, all pipelines used were able to identify genes as enriched. Unusually, although the Biotradis pipelines identified the most genes in all other time points, at day 10 it did not. At this time point, there were a lot more genes identified as enriched in the deseq2 and edgeR pipelines, with deseq2 identifying the largest number of genes classified as enriched. However, as with day 5, in each RNA-seq analysis the same gene set was identified, with only the cut-off of these genes differing between these pipelines.

ARTIST identified a set of genes, which were different from the remaining pipelines at Day 10. Manual inspection of these genes showed that insertions within these genes were enriched but only within a particular region of the gene. Interestingly these were mainly within two operons: *rbs* (an ABC transporter for ribose import of which the repressor *rbsR* was identified as insertion advantageous in the other pipelines) and *hdeA*, an acid chaperone protein previously identified to have a high read count in both ITL, D1 and D5 timepoints, but not at D10. However, upon a manual inspection of these genes, the extent of the enrichment observed was not as high as some genes identified in the RNA-seq and not identified within Artist. What this showed was that in these genes at lower frequency there was still a considerable amount of insertions present. Since there was a large drop in insertions at the Day 10 pH7 and since there was a consistent amount of insertions within these genes. This could potentially indicate that insertions within this region have a fitness advantage. Although since it was at a lower frequency, not a large fitness advantage just enough to be able to remain within the population.

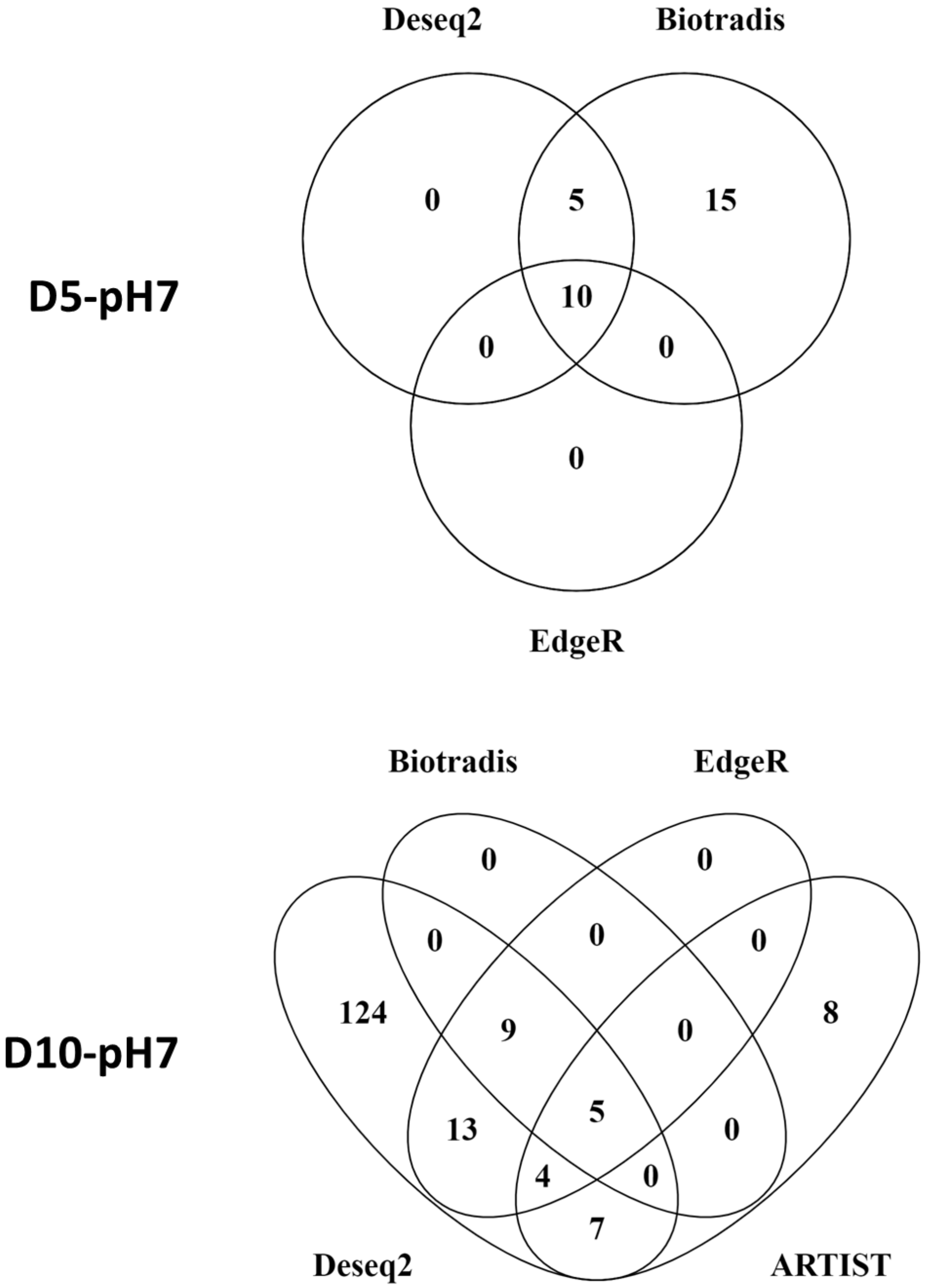


Figure 10.2.4: Comparison of different pipelines for identification of genes where insertions give an increase in fitness and therefore accumulate within the population.

10.3 Choosing the analysis for this study.

This section gives a comparison of different analysis pipelines which have previously been used to compare transposon sequencing analysis outgrowth experiments. Although some methods could only consider insertion loss from a library, others could consider also increase in insertions. Overall, however, the conclusion is that there is not one method which stood out amongst the rest. Therefore, with regards to continuing of the analysis of the STSE, it was decided that two different pipelines would be used: log likelihood and *edgeR*.

The reasoning for this was that when considering genes with insertions that were detrimental in fitness, the RNAseq pipelines at day 1 only identified a few genes, however with regards to genes which saw an increase in insertions, they could identify these genes clearly. One pipeline which could do both of these was the ARTIST pipeline. However, ARTIST was designed for Tn-seq, and using the method for TraDIS can reduce the resolution. Results which were specific to ARTIST did indeed reveal some genes which were not picked up by the other pipelines, but it also introduces false positives, as observed with insertion detrimental genes. In addition to this, this pipeline required significant computational power and a lot of user input therefore it was decided not to continue with this analysis.

Therefore, it was decided to use one RNAseq pipeline, together with another method which was better at identifying declines in insertions. In regard to insertion declines, this would be a form of the log likelihood analysis, either using the different thresholds described by Goodall *et al*., (2018) or in Biotradis. One thing that was noticed with this comparison was that with regards to Biotradis both methods used within this pipeline were less stringent, and therefore it was concluded that using these would risk introducing more false positives into the analysis. Therefore, with regard to genes where insertions were fitness detrimental, the log likelihood method was chosen. In regard to RNAseq the variability observed in the statistical significance in Biotradis between days, particularly within the enrichments, dissuaded the use of it. Therefore, this left deseq2 and edgeR. As both of these pipelines produced similar results, it was decided to use edgeR, which was slightly more stringent compared to deseq2 using the same data.

References

ABRAHAM, J. M., FREITAG, C. S., CLEMENTS, J. R. & EISENSTEIN, B. I. 1985. An invertible element of DNA controls phase variation of type 1 fimbriae of *Escherichia coli*. *Proc Natl Acad Sci U S A,* 82**,** 5724-7.

ANAND, A., CHEN, K., CATOIU, E., SASTRY, A. V., OLSON, C. A., SANDBERG, T. E., SEIF, Y., XU, S., SZUBIN, R., YANG, L., FEIST, A. M. & PALSSON, B. O. 2020. OxyR Is a Convergent Target for Mutations Acquired during Adaptation to Oxidative Stress-Prone Metabolic States. *Mol Biol Evol,* 37**,** 660-667.

AQUINO, P., HONDA, B., JAINI, S., LYUBETSKAYA, A., HOSUR, K., CHIU, J. G., EKLADIOUS, I., HU, D., JIN, L., SAYEG, M. K., STETTNER, A. I., WANG, J., WONG, B. G., WONG, W. S., ALEXANDER, S. L., BA, C., BENSUSSEN, S. I., BERNSTEIN, D. B., BRAFF, D., CHA, S., CHENG, D. I., CHO, J. H., CHOU, K., CHUANG, J., GASTLER, D. E., GRASSO, D. J., GREIFENBERGER, J. S., GUO, C., HAWES, A. K., ISRANI, D. V., JAIN, S. R., KIM, J., LEI, J., LI, H., LI, D., LI, Q., MANCUSO, C. P., MAO, N., MASUD, S. F., MEISEL, C. L., MI, J., NYKYFORCHYN, C. S., PARK, M., PETERSON, H. M., RAMIREZ, A. K., REYNOLDS, D. S., RIM, N. G., SAFFIE, J. C., SU, H., SU, W. R., SU, Y., SUN, M., THOMMES, M. M., TU, T., VARONGCHAYAKUL, N., WAGNER, T. E., WEINBERG, B. H., YANG, R., YAROSLAVSKY, A., YOON, C., ZHAO, Y., ZOLLINGER, A. J., STRINGER, A. M., FOSTER, J. W., WADE, J., RAMAN, S., BROUDE, N., WONG, W. W. & GALAGAN, J. E. 2017. Coordinated regulation of acid resistance in *Escherichia coli*. *BMC Syst Biol,* 11**,** 1.

AVALOS VIZCARRA, I., HOSSEINI, V., KOLLMANNSBERGER, P., MEIER, S., WEBER, S. S., ARNOLDINI, M., ACKERMANN, M. & VOGEL, V. 2016. How type 1 fimbriae help *Escherichia coli* to evade extracellular antibiotics. *Sci Rep,* 6**,** 18109.

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. *Mol Syst Biol,* 2**,** 2006.0008.

BARBIER, C. S. & SHORT, S. A. 1993. Characterization of *cytR* mutations that influence oligomerization of mutant repressor subunits. *J Bacteriol,* 175**,** 4625-30.

BARBIER, C. S., SHORT, S. A. & SENEAR, D. F. 1997. Allosteric mechanism of induction of CytR-regulated gene expression. CytR repressor-cytidine interaction. *J Biol Chem,* 272**,** 16962-71.

BARQUIST, L., MAYHO, M., CUMMINS, C., CAIN, A. K., BOINETT, C. J., PAGE, A. J., LANGRIDGE, G. C., QUAIL, M. A., KEANE, J. A. & PARKHILL, J. 2016. The TraDIS toolkit: sequencing and analysis for dense transposon mutant libraries. *Bioinformatics,* 32**,** 1109-11.

BARRETT, R. D., MACLEAN, R. C. & BELL, G. 2006. Mutations of intermediate effect are responsible for adaptation in evolving *Pseudomonas fluorescens* populations. *Biol Lett,* 2**,** 236-8.

BARRICK, J. E., COLBURN, G., DEATHERAGE, D. E., TRAVERSE, C. C., STRAND, M. D., BORGES, J. J., KNOESTER, D. B., REBA, A. & MEYER, A. G. 2014. Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. *BMC Genomics,* 15**,** 1039.

BARRICK, J. E., YU, D. S., YOON, S. H., JEONG, H., OH, T. K., SCHNEIDER, D., LENSKI, R. E. & KIM, J. F. 2009. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature,* 461**,** 1243-7.

BEHRINGER, M. G., CHOI, B. I., MILLER, S. F., DOAK, T. G., KARTY, J. A., GUO, W. & LYNCH, M. 2018. cultures maintain stable subpopulation structure during long-term evolution. *Proc Natl Acad Sci U S A,* 115**,** E4642-E4650.

BEKKER, M., ALEXEEVA, S., LAAN, W., SAWERS, G., TEIXEIRA DE MATTOS, J. & HELLINGWERF, K. 2010. The ArcBA two-component system of *Escherichia coli* is regulated by the redox state of both the ubiquinone and the menaquinone pool. *J Bacteriol,* 192**,** 746-54.

BENNETT, A. F. & LENSKI, R. E. 2007. An experimental test of evolutionary trade-offs during temperature adaptation. *Proc Natl Acad Sci U S A,* 104 Suppl 1**,** 8649-54.

BIKARD, D., JIANG, W., SAMAI, P., HOCHSCHILD, A., ZHANG, F. & MARRAFFINI, L. A. 2013. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res,* 41**,** 7429-37.

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., GREGOR, J., DAVIS, N. W., KIRKPATRICK, H. A., GOEDEN, M. A., ROSE, D. J., MAU, B. & SHAO, Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science,* 277**,** 1453-62.

BLOUNT, Z. D., BARRICK, J. E., DAVIDSON, C. J. & LENSKI, R. E. 2012. Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature,* 489**,** 513-8.

BLOUNT, Z. D., BORLAND, C. Z. & LENSKI, R. E. 2008. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc Natl Acad Sci U S A,* 105**,** 7899-906.

BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-20.

BOOT, I. R., CASH, P. & O'BYRNE, C. 2002. Sensing and adapting to acid stress. *Antonie Van Leeuwenhoek,* 81**,** 33-42.

BROWNING, D. F. & BUSBY, S. J. 2004. The regulation of bacterial transcription initiation. *Nat Rev Microbiol,* 2**,** 57-65.

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. *J Mol Biol,* 401**,** 726-42.

CAIN, A. K., BARQUIST, L., GOODMAN, A. L., PAULSEN, I. T., PARKHILL, J. & VAN OPIJNEN, T. 2020. A decade of advances in transposon-insertion sequencing. *Nat Rev Genet,* 21**,** 526-540.

CASTANIE-CORNET, M. P., PENFOUND, T. A., SMITH, D., ELLIOTT, J. F. & FOSTER, J. W. 1999. Control of acid resistance in *Escherichia coli*. *J Bacteriol,* 181**,** 3525-35.

CASTANIÉ-CORNET, M. P., CAM, K., BASTIAT, B., CROS, A., BORDES, P. & GUTIERREZ, C. 2010. Acid stress response in *Escherichia coli*: mechanism of regulation of *gadA* transcription by RcsB and GadE. *Nucleic Acids Res,* 38**,** 3546-54.

CHANG, Y. Y. & CRONAN, J. E. 1999. Membrane cyclopropane fatty acid content is a major factor in acid resistance of *Escherichia coli*. *Mol Microbiol,* 33**,** 249-59.

CHAO, M. C., ABEL, S., DAVIS, B. M. & WALDOR, M. K. 2016. The design and analysis of transposon insertion sequencing experiments. *Nat Rev Microbiol,* 14**,** 119-28.

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 Res,* 41**,** 9033-48.

CHEREPANOV, P. P. & WACKERNAGEL, W. 1995. Gene disruption in *Escherichia coli*: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. *Gene,* 158**,** 9-14.

CHEVILLE, A. M., ARNOLD, K. W., BUCHRIESER, C., CHENG, C. M. & KASPAR, C. W. 1996. *rpoS* regulation of acid, heat, and salt tolerance in *Escherichia coli* O157:H7. *Appl Environ Microbiol,* 62**,** 1822-4.

CONNOR, T. R., LOMAN, N. J., THOMPSON, S., SMITH, A., SOUTHGATE, J., POPLAWSKI, R., BULL, M. J., RICHARDSON, E., ISMAIL, M., THOMPSON, S. E., KITCHEN, C., GUEST, M., BAKKE, M., SHEPPARD, S. K. & PALLEN, M. J. 2016. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. *Microb Genom,* 2**,** e000086.

CONRAD, T. M., JOYCE, A. R., APPLEBEE, M. K., BARRETT, C. L., XIE, B., GAO, Y. & PALSSON, B. 2009. Whole-genome resequencing of *Escherichia coli* K-12 MG1655 undergoing short-term laboratory evolution in lactate minimal media reveals flexible selection of adaptive mutations. *Genome Biol,* 10**,** R118.

COOPER, V. S. 2018. Experimental Evolution as a High-Throughput Screen for Genetic Adaptations. *mSphere,* 3.

COOPER, V. S., SCHNEIDER, D., BLOT, M. & LENSKI, R. E. 2001. Mechanisms causing rapid and parallel losses of ribose catabolism in evolving populations of *Escherichia coli* B. *J Bacteriol,* 183**,** 2834-41.

COWLEY, L. A., LOW, A. S., PICKARD, D., BOINETT, C. J., DALLMAN, T. J., DAY, M., PERRY, N., GALLY, D. L., PARKHILL, J., JENKINS, C. & CAIN, A. K. 2018. Transposon Insertion Sequencing Elucidates Novel Gene Involvement in Susceptibility and Resistance to Phages T4 and T7 in. *mBio,* 9.

CREAMER, K. E., DITMARS, F. S., BASTING, P. J., KUNKA, K. S., HAMDALLAH, I. N., BUSH, S. P., SCOTT, Z., HE, A., PENIX, S. R., GONZALES, A. S., EDER, E. K., CAMPERCHIOLI, D. W., BERNDT, A., CLARK, M. W., ROUHIER, K. A. & SLONCZEWSKI, J. L. 2017. Benzoate- and Salicylate-Tolerant Strains of *Escherichia coli* K-12 Lose Antibiotic Resistance during Laboratory Evolution. *Appl Environ Microbiol,* 83.

DATSENKO, K. A. & WANNER, B. L. 2000. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc Natl Acad Sci U S A,* 97**,** 6640-5.

DE BIASE, D., TRAMONTI, A., BOSSA, F. & VISCA, P. 1999. The response to stationary-phase stress conditions in *Escherichia coli*: role and regulation of the glutamic acid decarboxylase system. *Mol Microbiol,* 32**,** 1198-211.

DE LORENZO, V., HERRERO, M., JAKUBZIK, U. & TIMMIS, K. N. 1990. Mini-Tn5 transposon derivatives for insertion mutagenesis, promoter probing, and chromosomal insertion of cloned DNA in gram-negative eubacteria. *J Bacteriol,* 172**,** 6568-72.

DEANGELIS, M. M., WANG, D. G. & HAWKINS, T. L. 1995. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res,* 23**,** 4742-3.

DEATHERAGE, D. E. & BARRICK, J. E. 2014. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol Biol,* 1151**,** 165-88.

DEJESUS, M. A., GERRICK, E. R., XU, W., PARK, S. W., LONG, J. E., BOUTTE, C. C., RUBIN, E. J., SCHNAPPINGER, D., EHRT, S., FORTUNE, S. M., SASSETTI, C. M. & IOERGER, T. R.

2017. Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. *MBio,* 8.

DEJESUS, M. A., ZHANG, Y. J., SASSETTI, C. M., RUBIN, E. J., SACCHETTINI, J. C. & IOERGER, T. R. 2013. Bayesian analysis of gene essentiality based on sequencing of transposon insertion libraries. *Bioinformatics,* 29**,** 695-703.

DORMAN, M. J., FELTWELL, T., GOULDING, D. A., PARKHILL, J. & SHORT, F. L. 2018. The Capsule Regulatory Network of Klebsiella pneumoniae Defined by density-TraDISort. mBio, 9.

DRAGOSITS, M. & MATTANOVICH, D. 2013. Adaptive laboratory evolution -- principles and applications for biotechnology. *Microb Cell Fact,* 12**,** 64.

DU, B., OLSON, C. A., SASTRY, A. V., FANG, X., PHANEUF, P. V., CHEN, K., WU, M., SZUBIN, R., XU, S., GAO, Y., HEFNER, Y., FEIST, A. M. & PALSSON, B. O. 2020. Adaptive laboratory evolution of *Escherichia coli* under acid stress. *Microbiology (Reading),* 166**,** 141-148.

EGUCHI, Y., ISHII, E., HATA, K. & UTSUMI, R. 2011. Regulation of acid resistance by connectors of two-component signal transduction systems in *Escherichia coli*. *J Bacteriol,* 193**,** 1222-8.

ESTREM, S. T., ROSS, W., GAAL, T., CHEN, Z. W., NIU, W., EBRIGHT, R. H. & GOURSE, R. L. 1999. Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit. *Genes Dev,* 13**,** 2134-47.

EVANS, D. F., PYE, G., BRAMLEY, R., CLARK, A. G., DYSON, T. J. & HARDCASTLE, J. D. 1988. Measurement of gastrointestinal pH profiles in normal ambulant human subjects. *Gut,* 29**,** 1035-41.

FOSTER, J. W. 2004. *Escherichia coli* acid resistance: tales of an amateur acidophile. *Nat Rev Microbiol,* 2**,** 898-907.

GAIMSTER, H., CAMA, J., HERNÁNDEZ-AINSA, S., KEYSER, U. F. & SUMMERS, D. K. 2014. The indole pulse: a new perspective on indole signalling in *Escherichia coli*. *PLoS One,* 9**,** e93168.

GAIMSTER, H. & SUMMERS, D. 2015. Regulation of Indole Signalling during the Transition of *E. coli* from Exponential to Stationary Phase. *PLoS One,* 10**,** e0136691.

GAJIWALA, K. S. & BURLEY, S. K. 2000. HDEA, a periplasmic protein that supports acid resistance in pathogenic enteric bacteria. *J Mol Biol,* 295**,** 605-12.

GALE, E. F. & EPPS, H. M. 1942. The effect of the pH of the medium during growth on the enzymic activities of bacteria (*Escherichia coli* and *Micrococcus lysodeikticus*) and the biological significance of the changes produced. *Biochem J,* 36**,** 600-18.

GALLY, D. L., LEATHART, J. & BLOMFIELD, I. C. 1996. Interaction of FimB and FimE with the fim switch that controls the phase variation of type 1 fimbriae in *Escherichia coli* K-12. *Mol Microbiol,* 21**,** 725-38.

GAWRONSKI, J. D., WONG, S. M. S., GIANNOUKOS, G., WARD, D. V. & AKERLEY, B. J. 2009. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc Natl Acad Sci U S A,* 106**,** 16422-16427.

GEERTZ, M., TRAVERS, A., MEHANDZISKA, S., SOBETZKO, P., CHANDRA-JANGA, S., SHIMAMOTO, N. & MUSKHELISHVILI, G. 2011. Structural coupling between RNA polymerase composition and DNA supercoiling in coordinating transcription: a global role for the omega subunit? *mBio,* 2.

GIANNELLA, R. A., BROITMAN, S. A. & ZAMCHECK, N. 1972. Gastric acid barrier to ingested microorganisms in man: studies in vivo and in vitro. *Gut,* 13**,** 251-6.

GILBERT, L. A., HORLBECK, M. A., ADAMSON, B., VILLALTA, J. E., CHEN, Y., WHITEHEAD, E. H., GUIMARAES, C., PANNING, B., PLOEGH, H. L., BASSIK, M. C., QI, L. S., KAMPMANN, M. & WEISSMAN, J. S. 2014. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell,* 159**,** 647-61.

GIRGIS, H. S., HOTTES, A. K. & TAVAZOIE, S. 2009. Genetic architecture of intrinsic antibiotic susceptibility. *PLoS One,* 4**,** e5629.

GOMPEL, N. & PRUD'HOMME, B. 2009. The causes of repeated genetic evolution. *Dev Biol,* 332**,** 36-47.

GONZÁLEZ-GONZÁLEZ, A., HUG, S. M., RODRÍGUEZ-VERDUGO, A., PATEL, J. S. & GAUT, B. S. 2017. Adaptive Mutations in RNA Polymerase and the Transcriptional Terminator Rho Have Similar Effects on *Escherichia coli* Gene Expression. *Mol Biol Evol,* 34**,** 2839-2855.

GOOD, B. H., MCDONALD, M. J., BARRICK, J. E., LENSKI, R. E. & DESAI, M. M. 2017. The dynamics of molecular evolution over 60,000 generations. *Nature,* 551**,** 45-50.

GOODALL, E. C. A., 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.

GOODARZI, H., BENNETT, B. D., AMINI, S., REAVES, M. L., HOTTES, A. K., RABINOWITZ, J. D. & TAVAZOIE, S. 2010. Regulatory and metabolic rewiring during laboratory evolution of ethanol tolerance in *E. coli*. *Mol Syst Biol,* 6**,** 378.

GOODMAN, A. L., MCNULTY, N. P., ZHAO, Y., LEIP, D., MITRA, R. D., LOZUPONE, C. A., KNIGHT, R. & GORDON, J. I. 2009. Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe,* 6**,** 279-89.

GORDEN, J. & SMALL, P. L. 1993. Acid resistance in enteric bacteria. *Infect Immun,* 61**,** 364-7.

GORYSHIN, I. Y., JENDRISAK, J., HOFFMAN, L. M., MEIS, R. & REZNIKOFF, W. S. 2000. Insertional transposon mutagenesis by electroporation of released Tn5 transposition complexes. *Nat Biotechnol,* 18**,** 97-100.

GREEN, B., BOUCHIER, C., FAIRHEAD, C., CRAIG, N. L. & CORMACK, B. P. 2012. Insertion site preference of Mu, Tn5, and Tn7 transposons. *Mob DNA,* 3**,** 3.

GROSS, J., AVRANI, S., KATZ, S., HILAU, S. & HERSHBERG, R. 2020. Culture Volume Influences the Dynamics of Adaptation under Long-Term Stationary Phase. *Genome Biol Evol,* 12**,** 2292-2301.

GU, D., XUE, H., YUAN, X., YU, J., XU, X., HUANG, Y., LI, M., ZHAI, X., PAN, Z., ZHANG, Y. & JIAO, X. 2021. Genome-Wide Identification of Genes Involved in Acid Stress Resistance of *Salmonella* Derby. Genes (Basel), 12.

HAMDALLAH, I., TOROK, N., BISCHOF, K. M., MAJDALANI, N., CHADALAVADA, S., MDLULI, N., CREAMER, K. E., CLARK, M., HOLDENER, C., BASTING, P. J., GOTTESMAN, S. & SLONCZEWSKI, J. L. 2018. Experimental Evolution of *Escherichia coli* K-12 at High pH and with RpoS Induction. *Appl Environ Microbiol,* 84.

HARDEN, M. M., HE, A., CREAMER, K., CLARK, M. W., HAMDALLAH, I., MARTINEZ, K. A., KRESSLEIN, R. L., BUSH, S. P. & SLONCZEWSKI, J. L. 2015. Acid-adapted strains of *Escherichia coli* K-12 obtained by experimental evolution. *Appl Environ Microbiol,* 81**,** 1932-41.

HARRIS, D. R., POLLOCK, S. V., WOOD, E. A., GOIFFON, R. J., KLINGELE, A. J., CABOT, E. L., SCHACKWITZ, W., MARTIN, J., EGGINGTON, J., DURFEE, T. J., MIDDLE, C. M., NORTON, J. E., POPELARS, M. C., LI, H., KLUGMAN, S. A., HAMILTON, L. L., BANE, L. B., PENNACCHIO, L. A., ALBERT, T. J., PERNA, N. T., COX, M. M. & BATTISTA, J. R. 2009. Directed evolution of ionizing radiation resistance in *Escherichia coli*. *J Bacteriol,* 191**,** 5240-52.

HASSAN, K. A., CAIN, A. K., HUANG, T., LIU, Q., ELBOURNE, L. D., BOINETT, C. J., BRZOSKA, A. J., LI, L., OSTROWSKI, M., NHU, N. T., NHU, T. O. H., BAKER, S., PARKHILL, J. & PAULSEN, I. T. 2016. Fluorescence-Based Flow Sorting in Parallel with Transposon Insertion Site Sequencing Identifies Multidrug Efflux Systems in *Acinetobacter baumannii*. *mBio,* 7.

HAYES, E. T., WILKS, J. C., SANFILIPPO, P., YOHANNES, E., TATE, D. P., JONES, B. D., RADMACHER, M. D., BONDURANT, S. S. & SLONCZEWSKI, J. L. 2006. Oxygen limitation modulates pH regulation of catabolism and hydrogenases, multidrug transporters, and envelope composition in *Escherichia coli* K-12. *BMC Microbiol,* 6**,** 89.

HE, A., PENIX, S. R., BASTING, P. J., GRIFFITH, J. M., CREAMER, K. E., CAMPERCHIOLI, D., CLARK, M. W., GONZALES, A. S., CHÁVEZ ERAZO, J. S., GEORGE, N. S., BHAGWAT, A. A. & SLONCZEWSKI, J. L. 2017. Acid Evolution of *Escherichia coli* K-12 Eliminates Amino Acid Decarboxylases and Reregulates Catabolism. *Appl Environ Microbiol,* 83.

HERSH, B. M., FAROOQ, F. T., BARSTAD, D. N., BLANKENHORN, D. L. & SLONCZEWSKI, J. L. 1996. A glutamate-dependent acid resistance gene in *Escherichia coli*. *J Bacteriol,* 178**,** 3978-81.

HOBMAN, J. L., PENN, C. W. & PALLEN, M. J. 2007. Laboratory strains of *Escherichia coli*: model citizens or deceitful delinquents growing old disgracefully? *Mol Microbiol,* 64**,** 881-5.

HOLDEN, E. R., YASIR, M., TURNER, A. K., WAIN, J., CHARLES, I. G. & WEBBER, M. A. 2021. Massively parallel transposon mutagenesis identifies temporally essential genes for biofilm formation in *Escherichia coli. ibioRxiv***,** 2020.12.14.409862.

HOMMAIS, F., KRIN, E., COPPÉE, J. Y., LACROIX, C., YERAMIAN, E., DANCHIN, A. & BERTIN, P. 2004. GadE (YhiE): a novel activator involved in the response to acid environment in *Escherichia coli*. *Microbiology,* 150**,** 61-72.

HORESH, G., BLACKWELL, G. A., TONKIN-HILL, G., CORANDER, J., HEINZ, E. & THOMSON, N. R. 2021. A comprehensive and high-quality collection of *Escherichia coli* and their genes. *Microb Genom,* 7.

HOTTES, A. K., FREDDOLINO, P. L., KHARE, A., DONNELL, Z. N., LIU, J. C. & TAVAZOIE, S. 2013. Bacterial adaptation through loss of function. *PLoS Genet,* 9**,** e1003617.

HUBBARD, T. P., D'GAMA, J. D., BILLINGS, G., DAVIS, B. M. & WALDOR, M. K. 2019. Unsupervised Learning Approach for Comparing Multiple Transposon Insertion Sequencing Studies. *mSphere,* 4.

HUGHES, B. S., CULLUM, A. J. & BENNETT, A. F. 2007. Evolutionary adaptation to environmental pH in experimental lineages of *Escherichia coli*. *Evolution,* 61**,** 1725-34.

HULTGREN, S. J., PORTER, T. N., SCHAEFFER, A. J. & DUNCAN, J. L. 1985. Role of type 1 pili and effects of phase variation on lower urinary tract infections produced by *Escherichia coli*. *Infect Immun,* 50**,** 370-7.

HWANG, J. & INOUYE, M. 2006. The tandem GTPase, Der, is essential for the biogenesis of 50S ribosomal subunits in *Escherichia coli*. *Mol Microbiol,* 61**,** 1660-72.

IGARASHI, K. & KASHIWAGI, K. 2018. Effects of polyamines on protein synthesis and growth of. *J Biol Chem,* 293**,** 18702-18709.

IUCHI, S. & LIN, E. C. 1992. Mutational analysis of signal transduction by ArcB, a membrane sensor protein responsible for anaerobic repression of operons involved in the central aerobic pathways in *Escherichia coli*. *J Bacteriol,* 174**,** 3972-80.

IYER, R., WILLIAMS, C. & MILLER, C. 2003. Arginine-agmatine antiporter in extreme acid resistance in *Escherichia coli*. *J Bacteriol,* 185**,** 6556-61.

JAIN, P. K., JAIN, V., SINGH, A. K., CHAUHAN, A. & SINHA, S. 2013. Evaluation on the responses of succinate dehydrogenase, isocitrate dehydrogenase, malate dehydrogenase and glucose-6-phosphate dehydrogenase to acid shock generated acid tolerance in *Escherichia coli*. *Adv Biomed Res,* 2**,** 75.

JAISHANKAR, J. & SRIVASTAVA, P. 2017. Molecular Basis of Stationary Phase Survival and Applications. *Front Microbiol,* 8**,** 2000.

JEON, Y. H., NEGISHI, T., SHIRAKAWA, M., YAMAZAKI, T., FUJITA, N., ISHIHAMA, A. & KYOGOKU, Y. 1995. Solution structure of the activator contact domain of the RNA polymerase alpha subunit. *Science,* 270**,** 1495-7.

JOHNSON, M. D., BELL, J., CLARKE, K., CHANDLER, R., PATHAK, P., XIA, Y., MARSHALL, R. L., WEINSTOCK, G. M., LOMAN, N. J., WINN, P. J. & LUND, P. A. 2014. Characterization of mutations in the PAS domain of the EvgS sensor kinase selected by laboratory evolution for acid resistance in *Escherichia coli*. *Mol Microbiol,* 93**,** 911-27.

JOHNSON, M. D., BURTON, N. A., GUTIÉRREZ, B., PAINTER, K. & LUND, P. A. 2011. RcsB is required for inducible acid resistance in *Escherichia coli* and acts at *gadE*-dependent and -independent promoters. *J Bacteriol,* 193**,** 3653-6.

KAMP, H. D., PATIMALLA-DIPALI, B., LAZINSKI, D. W., WALLACE-GADSDEN, F. & CAMILLI, A. 2013. Gene fitness landscapes of *Vibrio cholerae* at important stages of its life cycle. *PLoS Pathog,* 9**,** e1003800.

KANJEE, U., GUTSCHE, I., ALEXOPOULOS, E., ZHAO, B., EL BAKKOURI, M., THIBAULT, G., LIU, K., RAMACHANDRAN, S., SNIDER, J., PAI, E. F. & HOURY, W. A. 2011. Linkage between the bacterial acid stress and stringent responses: the structure of the inducible lysine decarboxylase. *EMBO J,* 30**,** 931-44.

KANJEE, U. & HOURY, W. A. 2013. Mechanisms of acid resistance in *Escherichia coli*. *Annu Rev Microbiol,* 67**,** 65-81.

KASHIWAGI, K., SUZUKI, T., SUZUKI, F., FURUCHI, T., KOBAYASHI, H. & IGARASHI, K. 1991. Coexistence of the genes for putrescine transport protein and ornithine decarboxylase at 16 min on *Escherichia coli* chromosome. *J Biol Chem,* 266**,** 20922-7.

KERN, R., MALKI, A., ABDALLAH, J., TAGOURTI, J. & RICHARME, G. 2007. *Escherichia coli* HdeB is an acid stress chaperone. *J Bacteriol,* 189**,** 603-10.

KHAN, A. & MATHELIER, A. 2017. Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics,* 18**,** 287.

KLAUCK, G., SERRA, D. O., POSSLING, A. & HENGGE, R. 2018. Spatial organization of different sigma factor activities and c-di-GMP signalling within the three-dimensional landscape of a bacterial biofilm. *Open Biol,* 8.

KLEMM, P. 1986. Two regulatory fim genes, *fimB* and *fimE*, control the phase variation of type 1 fimbriae in *Escherichia coli*. *EMBO J,* 5**,** 1389-93.

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. *Front Microbiol,* 9**,** 756.

KNÖPPEL, A., NÄSVALL, J. & ANDERSSON, D. I. 2017. Evolution of Antibiotic Resistance without Antibiotic Exposure. *Antimicrob Agents Chemother,* 61.

KRAM, K. E. & FINKEL, S. E. 2014. Culture volume and vessel affect long-term survival, mutation frequency, and oxidative stress of *Escherichia coli*. *Appl Environ Microbiol,* 80**,** 1732-8.

KRAM, K. E., GEIGER, C., ISMAIL, W. M., LEE, H., TANG, H., FOSTER, P. L. & FINKEL, S. E. 2017. Adaptation of Serial Passage in Complex Medium: Evidence of Parallel Evolution. *mSystems,* 2.

KRISTENSEN, H. H., VALENTIN-HANSEN, P. & SØGAARD-ANDERSEN, L. 1997. Design of CytR regulated, cAMP-CRP dependent class II promoters in *Escherichia coli*: RNA polymerase-promoter interactions modulate the efficiency of CytR repression. *J Mol Biol,* 266**,** 866-76.

KUPER, C. & JUNG, K. 2005. CadC-mediated activation of the cadBA promoter in *Escherichia coli*. *J Mol Microbiol Biotechnol,* 10**,** 26-39.

LACROIX, R. A., SANDBERG, T. E., O'BRIEN, E. J., UTRILLA, J., EBRAHIM, A., GUZMAN, G. I., SZUBIN, R., PALSSON, B. O. & FEIST, A. M. 2015. Use of adaptive laboratory evolution to discover key mutations enabling rapid growth of *Escherichia coli* K-12 MG1655 on glucose minimal medium. *Appl Environ Microbiol,* 81**,** 17-30.

LAMPE, D. J., GRANT, T. E. & ROBERTSON, H. M. 1998. Factors affecting transposition of the Himar1 mariner transposon in vitro. *Genetics,* 149**,** 179-87.

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., WAIN, J., PARKHILL, J. & TURNER, A. K. 2009. Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. *Genome Res,* 19**,** 2308-16.

LEBEUF-TAYLOR, E., MCCLOSKEY, N., BAILEY, S. F., HINZ, A. & KASSEN, R. 2019. The distribution of fitness effects among synonymous mutations in a gene under directional selection. *Elife,* 8.

LEE, D. H. & PALSSON, B. 2010. Adaptive evolution of *Escherichia coli* K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol. *Appl Environ Microbiol,* 76**,** 4158-68.

LEE, Y. S., HAN, J. S., JEON, Y. & HWANG, D. S. 2001. The arc two-component signal transduction system inhibits in vitro *Escherichia coli* chromosomal initiation. *J Biol Chem,* 276**,** 9917-23.

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.

LENSKI, R. E. & TRAVISANO, M. 1994. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proc Natl Acad Sci U S A,* 91**,** 6808-14.

LEYN, S. A., ZLAMAL, J. E., KURNASOV, O. V., LI, X., ELANE, M., MYJAK, L., GODZIK, M., DE CRECY, A., GARCIA-ALCALDE, F., EBELING, M. & OSTERMAN, A. L. 2021. Experimental evolution in morbidostat reveals converging genomic trajectories on the path to triclosan resistance. *Microb Genom,* 7.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LI, X. T., JUN, Y., ERICKSTAD, M. J., BROWN, S. D., PARKS, A., COURT, D. L. & JUN, S. 2016. tCRISPRi: tunable and reversible, one-step control of gene expression. *Sci Rep,* 6**,** 39076.

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*. *J Bacteriol,* 177**,** 4097-104.

LIN, J., SMITH, M. P., CHAPIN, K. C., BAIK, H. S., BENNETT, G. N. & FOSTER, J. W. 1996. Mechanisms of acid resistance in enterohemorrhagic *Escherichia coli*. *Appl Environ Microbiol,* 62**,** 3094-100.

LIND, P. A., BERG, O. G. & ANDERSSON, D. I. 2010. Mutational robustness of ribosomal protein genes. *Science,* 330**,** 825-7.

LIND, P. A., FARR, A. D. & RAINEY, P. B. 2015. Experimental evolution reveals hidden diversity in evolutionary pathways. *Elife,* 4.

LIU, X. & DE WULF, P. 2004. Probing the ArcA-P modulon of *Escherichia coli* by whole genome transcriptional analysis and sequence recognition profiling. *J Biol Chem,* 279**,** 12588-97.

LLOYD, G. S., NIU, W., TEBBUTT, J., EBRIGHT, R. H. & BUSBY, S. J. 2002. Requirement for two copies of RNA polymerase alpha subunit C-terminal domain for synergistic transcription activation at complex bacterial promoters. *Genes Dev,* 16**,** 2557-65.

LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol,* 15**,** 550.

LU, P., MA, D., CHEN, Y., GUO, Y., CHEN, G. Q., DENG, H. & SHI, Y. 2013. L-glutamine provides acid resistance for *Escherichia coli* through enzymatic release of ammonia. *Cell Res,* 23**,** 635-44.

LÜTTMANN, D., HEERMANN, R., ZIMMER, B., HILLMANN, A., RAMPP, I. S., JUNG, K. & GÖRKE, B. 2009. Stimulation of the potassium sensor KdpD kinase activity by interaction with the phosphotransferase protein IIA(Ntr) in *Escherichia coli*. *Mol Microbiol,* 72**,** 978-94.

MA, Z., GONG, S., RICHARD, H., TUCKER, D. L., CONWAY, T. & FOSTER, J. W. 2003. GadE (YhiE) activates glutamate decarboxylase-dependent acid resistance in *Escherichia coli* K-12. *Mol Microbiol,* 49**,** 1309-20.

MA, Z., MASUDA, N. & FOSTER, J. W. 2004. Characterization of EvgAS-YdeO-GadE branched regulatory circuit governing glutamate-dependent acid resistance in *Escherichia coli*. *J Bacteriol,* 186**,** 7378-89.

MARTINEZ, K. A., KITKO, R. D., MERSHON, J. P., ADCOX, H. E., MALEK, K. A., BERKMEN, M. B. & SLONCZEWSKI, J. L. 2012. Cytoplasmic pH response to acid stress in individual cells of *Escherichia coli* and *Bacillus subtilis* observed by fluorescence ratio imaging microscopy. *Appl Environ Microbiol,* 78**,** 3706-14.

MARTÍNEZ-CARRANZA, E., BARAJAS, H., ALCARAZ, L. D., SERVÍN-GONZÁLEZ, L., PONCE-SOTO, G. Y. & SOBERÓN-CHÁVEZ, G. 2018. Variability of Bacterial Essential Genes Among Closely Related Bacteria: The Case of *Escherichia coli*. *Front Microbiol,* 9**,** 1059.

MASUDA, N. & CHURCH, G. M. 2003. Regulatory network of acid resistance genes in *Escherichia coli*. *Mol Microbiol,* 48**,** 699-712.

MATANGE, N., HEGDE, S. & BODKHE, S. 2019. Adaptation Through Lifestyle Switching Sculpts the Fitness Landscape of Evolving Populations: Implications for the Selection of Drug-Resistant Bacteria at Low Drug Pressures. *Genetics,* 211**,** 1029-1044.

MCCARTHY, D. J., CHEN, Y. & SMYTH, G. K. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res,* 40**,** 4288-97.

MCCLAIN, M. S., BLOMFIELD, I. C. & EISENSTEIN, B. I. 1991. Roles of *fimB* and *fimE* in site-specific DNA inversion associated with phase variation of type 1 fimbriae in *Escherichia coli*. *J Bacteriol,* 173**,** 5308-14.

MCCOY, K. M., ANTONIO, M. L. & VAN OPIJNEN, T. 2017. MAGenTA: a Galaxy implemented tool for complete Tn-Seq analysis and data visualization. *Bioinformatics,* 33**,** 2781-2783.

MCDONALD, M. J. 2019. Microbial Experimental Evolution - a proving ground for evolutionary theory and a tool for discovery. *EMBO Rep,* 20**,** e46992.

MCDONALD, M. J., GEHRIG, S. M., MEINTJES, P. L., ZHANG, X. X. & RAINEY, P. B. 2009. Adaptive divergence in experimental populations of *Pseudomonas fluorescens*. IV. Genetic constraints guide evolutionary trajectories in a parallel adaptive radiation. *Genetics,* 183**,** 1041-53.

MIKA, F. & HENGGE, R. 2005. A two-component phosphotransfer network involving ArcB, ArcA, and RssB coordinates synthesis and proteolysis of sigmaS (RpoS) in *E. coli*. *Genes Dev,* 19**,** 2770-81.

MILO, R., JORGENSEN, P., MORAN, U., WEBER, G. & SPRINGER, M. 2010. BioNumbers--the database of key numbers in molecular and cell biology. *Nucleic Acids Res,* 38**,** D750-3.

MINTY, J. J., LESNEFSKY, A. A., LIN, F., CHEN, Y., ZAROFF, T. A., VELOSO, A. B., XIE, B., MCCONNELL, C. A., WARD, R. J., SCHWARTZ, D. R., ROUILLARD, J. M., GAO, Y., GULARI, E. & LIN, X. N. 2011. Evolution combined with genomic study elucidates genetic bases of isobutanol tolerance in *Escherichia coli*. *Microb Cell Fact,* 10**,** 18.

MONROE, J. G., MCKAY, J. K., WEIGEL, D. & FLOOD, P. J. 2021. The population genomics of adaptive loss of function. *Heredity (Edinb),* 126**,** 383-395.

MOOTHA, V. K., LINDGREN, C. M., ERIKSSON, K. F., SUBRAMANIAN, A., SIHAG, S., LEHAR, J., PUIGSERVER, P., CARLSSON, E., RIDDERSTRÅLE, M., LAURILA, E., HOUSTIS, N., DALY, M. J., PATTERSON, N., MESIROV, J. P., GOLUB, T. R., TAMAYO, P., SPIEGELMAN, B., LANDER, E. S., HIRSCHHORN, J. N., ALTSHULER, D. & GROOP, L. C. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are co-ordinately downregulated in human diabetes. *Nat Genet,* 34**,** 267-73.

MURRAY, A. W. 2020. Can gene-inactivating mutations lead to evolutionary novelty? *Curr Biol,* 30**,** R465-R471.

MURRY, J. P., SASSETTI, C. M., LANE, J. M., XIE, Z. & RUBIN, E. J. 2008. Transposon site hybridization in *Mycobacterium tuberculosis*. *Methods Mol Biol,* 416**,** 45-59.

NEELY, M. N., DELL, C. L. & OLSON, E. R. 1994. Roles of LysP and CadC in mediating the lysine requirement for acid induction of the *Escherichia coli* cad operon. *J Bacteriol,* 176**,** 3278-85.

NIKAIDO, H. 2009. *The Limitations of LB Medium* [Online]. Small Things Considered. Available: https://schaechter.asmblog.org/schaechter/2009/11/the-limitations-of-lb-medium.html [Accessed 21st December 2020].

NIZAM, S. A., ZHU, J., HO, P. Y. & SHIMIZU, K. 2009. Effects of *arcA* and *arcB* genes knockout on the metabolism in *Escherichia coli* under aerobic condition. *Biochemical Engineering Journal,* 44**,** 240-250.

NOTLEY-MCROBB, L., KING, T. & FERENCI, T. 2002. *rpoS* mutations and loss of general stress resistance in *Escherichia coli* populations as a consequence of conflict between competing stress responses. *J Bacteriol,* 184**,** 806-11.

OLSON, M. A., GRIMSRUD, A., RICHARDS, A. C., MULVEY, M. A., WILSON, E. & ERICKSON, D. L. 2020. A link between zinc uptake, bile salts, and a capsule required for virulence of a mastitis-associated extraintestinal pathogenic *Escherichia coli* strain. *bioRxiv***,** 2020.06.25.172775.

OLSON, M. V. 1999. When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet,* 64**,** 18-23.

OPDYKE, J. A., KANG, J. G. & STORZ, G. 2004. GadY, a small-RNA regulator of acid response genes in *Escherichia coli*. *J Bacteriol,* 186**,** 6698-705.

OZOLINE, O. N., FUJITA, N. & ISHIHAMA, A. 2001. Mode of DNA-protein interaction between the C-terminal domain of *Escherichia coli* RNA polymerase alpha subunit and T7D promoter UP element. *Nucleic Acids Res,* 29**,** 4909-19.

PAGE, A. J., BASTKOWSKI, S., YASIR, M., TURNER, A. K., LE VIET, T., SAVVA, G. M., WEBBER, M. A. & CHARLES, I. G. 2020. AlbaTraDIS: Comparative analysis of large datasets from parallel transposon mutagenesis experiments. *PLoS Comput Biol,* 16**,** e1007980.

PARK, D. M., AKHTAR, M. S., ANSARI, A. Z., LANDICK, R. & KILEY, P. J. 2013. The bacterial response regulator ArcA uses a diverse binding site architecture to regulate carbon oxidation globally. *PLoS Genet,* 9**,** e1003839.

PARK, D. M. & KILEY, P. J. 2014. The influence of repressor DNA binding site architecture on transcriptional control. *mBio,* 5**,** e01684-14.

PARK, G., PARK, J. K., SHIN, S. H., JEON, H. J., KIM, N. K. D., KIM, Y. J., SHIN, H. T., LEE, E., LEE, K. H., SON, D. S., PARK, W. Y. & PARK, D. 2017. Characterization of background noise in capture-based targeted sequencing data. *Genome Biol,* 18**,** 136.

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. *Front Microbiol,* 9**,** 2869.

PERRENOUD, A. & SAUER, U. 2005. Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in *Escherichia coli*. *J Bacteriol,* 187**,** 3171-9.

PESAVENTO, C., BECKER, G., SOMMERFELDT, N., POSSLING, A., TSCHOWRI, N., MEHLIS, A. & HENGGE, R. 2008. Inverse regulatory coordination of motility and curli-mediated adhesion in *Escherichia coli*. *Genes Dev,* 22**,** 2434-46.

PETERSON, W. L., MACKOWIAK, P. A., BARNETT, C. C., MARLING-CASON, M. & HALEY, M. L. 1989. The human gastric bactericidal barrier: mechanisms of action, relative antibacterial activity, and dietary influences. *J Infect Dis,* 159**,** 979-83.

PFLÜGER-GRAU, K. & GÖRKE, B. 2010. Regulatory roles of the bacterial nitrogen-related phosphotransferase system. *Trends Microbiol,* 18**,** 205-14.

PHAN, M. D., PETERS, K. M., SARKAR, S., LUKOWSKI, S. W., ALLSOPP, L. P., GOMES MORIEL, D., ACHARD, M. E., TOTSIKA, M., MARSHALL, V. M., UPTON, M., BEATSON, S. A. & SCHEMBRI, M. A. 2013. The serum resistome of a globally disseminated multidrug resistant uropathogenic *Escherichia coli* clone. *PLoS Genet,* 9**,** e1003834.

PHANEUF, P. V., GOSTING, D., PALSSON, B. O. & FEIST, A. M. 2019. ALEdb 1.0: a database of mutations from adaptive laboratory evolution experimentation. *Nucleic Acids Res,* 47**,** D1164-D1171.

PRITCHARD, J. R., CHAO, M. C., ABEL, S., DAVIS, B. M., BARANOWSKI, C., ZHANG, Y. J., RUBIN, E. J. & WALDOR, M. K. 2014. ARTIST: high-resolution genome-wide assessment of fitness using transposon-insertion sequencing. *PLoS Genet,* 10**,** e1004782.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

RASKO, D. A., ROSOVITZ, M. J., MYERS, G. S., MONGODIN, E. F., FRICKE, W. F., GAJER, P., CRABTREE, J., SEBAIHIA, M., THOMSON, N. R., CHAUDHURI, R., HENDERSON, I. R., SPERANDIO, V. & RAVEL, J. 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol,* 190**,** 6881-93.

RATIB, N. R., SEIDL, F., EHRENREICH, I. M. & FINKEL, S. E. 2021. Evolution in Long-Term Stationary-Phase Batch Culture: Emergence of Divergent *Escherichia coli* Lineages over 1,200 Days. *mBio,* 12.

REZNIKOFF, W. S., BHASIN, A., DAVIES, D. R., GORYSHIN, I. Y., MAHNKE, L. A., NAUMANN, T., RAYMENT, I., STEINIGER-WHITE, M. & TWINING, S. S. 1999. Tn5: A molecular window on transposition. *Biochem Biophys Res Commun,* 266**,** 729-34.

RIEHLE, M. M., BENNETT, A. F., LENSKI, R. E. & LONG, A. D. 2003. Evolutionary changes in heat-inducible gene expression in lines of *Escherichia coli* adapted to high temperature. *Physiol Genomics,* 14**,** 47-58.

RIEHLE, M. M., BENNETT, A. F. & LONG, A. D. 2001. Genetic architecture of thermal adaptation in *Escherichia coli*. *Proc Natl Acad Sci U S A,* 98**,** 525-30.

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**,** 139-40.

RODRIGUES, D. F. & ELIMELECH, M. 2009. Role of type 1 fimbriae and mannose in the development of *Escherichia coli* K12 biofilm: from initial cell adhesion to biofilm formation. *Biofouling,* 25**,** 401-11.

RUIZ, L., BOTTACINI, F., BOINETT, C. J., CAIN, A. K., O'CONNELL-MOTHERWAY, M., LAWLEY, T. D. & VAN SINDEREN, D. 2017. The essential genomic landscape of the commensal *Bifidobacterium breve* UCC2003. *Sci Rep,* 7**,** 5648.

SANCHEZ-TORRES, V., HU, H. & WOOD, T. K. 2011. GGDEF proteins YeaI, YedQ, and YfiN reduce early biofilm formation and swimming motility in *Escherichia coli*. *Appl Microbiol Biotechnol,* 90**,** 651-8.

SASSETTI, C. M., BOYD, D. H. & RUBIN, E. J. 2001. Comprehensive identification of conditionally essential genes in mycobacteria. *Proc Natl Acad Sci U S A,* 98**,** 12712-7.

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**,** 2584-92.

SEN, H. 2018. *Escherichia coli responses to acid-stress: signal transduction and gene regulation.* PhD, University of Birmingham.

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*. *Nat Commun,* 6**,** 7970.

SEPUTIENE, V., DAUGELAVICIUS, A., SUZIEDELIS, K. & SUZIEDELIENE, E. 2006. Acid response of exponentially growing *Escherichia coli* K-12. *Microbiol Res,* 161**,** 65-74.

SEZONOV, G., JOSELEAU-PETIT, D. & D'ARI, R. 2007. *Escherichia coli* physiology in Luria-Bertani broth. *J Bacteriol,* 189**,** 8746-9.

SHI, X. & BENNETT, G. N. 1994. Effects of *rpoA* and *cysB* mutations on acid induction of biodegradative arginine decarboxylase in *Escherichia coli*. *J Bacteriol,* 176**,** 7017-23.

SKURNIK, D., ROUX, D., ASCHARD, H., CATTOIR, V., YODER-HIMES, D., LORY, S. & PIER, G. B. 2013. A comprehensive analysis of in vitro and in vivo genetic fitness of *Pseudomonas aeruginosa* using high-throughput sequencing of transposon libraries. *PLoS Pathog,* 9**,** e1003582.

SLEIGHT, S. C. & LENSKI, R. E. 2007. Evolutionary adaptation to freeze-thaw-growth cycles in *Escherichia coli*. *Physiol Biochem Zool,* 80**,** 370-85.

SLEIGHT, S. C., ORLIC, C., SCHNEIDER, D. & LENSKI, R. E. 2008. Genetic basis of evolutionary adaptation by *Escherichia coli* to stressful cycles of freezing, thawing and growth. *Genetics,* 180**,** 431-43.

SLONCZEWSKI, J. L., ROSEN, B. P., ALGER, J. R. & MACNAB, R. M. 1981. pH homeostasis in *Escherichia coli*: measurement by 31P nuclear magnetic resonance of methylphosphonate and phosphate. *Proc Natl Acad Sci U S A,* 78**,** 6271-5.

SMITH, V., BOTSTEIN, D. & BROWN, P. O. 1995. Genetic footprinting: a genomic strategy for determining a gene's function given its sequence. *Proc Natl Acad Sci U S A,* 92**,** 6479-83.

SOKSAWATMAEKHIN, W., KURAISHI, A., SAKATA, K., KASHIWAGI, K. & IGARASHI, K. 2004. Excretion and uptake of cadaverine by CadB and its physiological functions in *Escherichia coli*. *Mol Microbiol,* 51**,** 1401-12.

SOLAIMANPOUR, S., SARMIENTO, F. & MRÁZEK, J. 2015. Tn-seq explorer: a tool for analysis of high-throughput sequencing data of transposon mutant libraries. *PLoS One,* 10**,** e0126070.

STANCIK, L. M., STANCIK, D. M., SCHMIDT, B., BARNHART, D. M., YONCHEVA, Y. N. & SLONCZEWSKI, J. L. 2002. pH-dependent expression of periplasmic proteins and amino acid catabolism in *Escherichia coli*. *J Bacteriol,* 184**,** 4246-58.

STIM-HERNDON, K. P., FLORES, T. M. & BENNETT, G. N. 1996. Molecular characterization of *adiY*, a regulatory gene which affects expression of the biodegradative acid-induced arginine decarboxylase gene (*adiA*) of *Escherichia coli*. *Microbiology,* 142 (Pt 5)**,** 1311-20.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. & MESIROV, J. P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A,* 102**,** 15545-50.

SUN, Y., FUKAMACHI, T., SAITO, H. & KOBAYASHI, H. 2012. Adenosine deamination increases the survival under acidic conditions in *Escherichia coli*. *J Appl Microbiol,* 112**,** 775-81.

SØGAARD-ANDERSEN, L., MARTINUSSEN, J., MØLLEGAARD, N. E., DOUTHWAITE, S. R. & VALENTIN-HANSEN, P. 1990. The CytR repressor antagonizes cyclic AMP-cyclic AMP receptor protein activation of the deoCp2 promoter of *Escherichia coli* K-12. *J Bacteriol,* 172**,** 5706-13.

TANG, H. & THOMAS, P. D. 2016. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. *Genetics,* 203**,** 635-47.

TETSCH, L., KOLLER, C., HANEBURGER, I. & JUNG, K. 2008. The membrane-integrated transcriptional activator CadC of *Escherichia coli* senses lysine indirectly via the interaction with the lysine permease LysP. *Mol Microbiol,* 67**,** 570-83.

THIBAULT, D., JENSEN, P. A., WOOD, S., QABAR, C., CLARK, S., SHAINHEIT, M. G., ISBERG, R. R. & VAN OPIJNEN, T. 2019. Droplet Tn-Seq combines microfluidics with Tn-Seq for identifying complex single-cell phenotypes. *Nat Commun,* 10**,** 5729.

THOMASON, L. C., COSTANTINO, N. & COURT, D. L. 2007. *E. coli* genome manipulation by P1 transduction. *Curr Protoc Mol Biol,* Chapter 1**,** Unit 1.17.

THOMASON, M. K. & STORZ, G. 2010. Bacterial antisense RNAs: how many are there, and what are they doing? *Annu Rev Genet,* 44**,** 167-88.

TIRUMALAI, M. R., KAROUIA, F., TRAN, Q., STEPANOV, V. G., BRUCE, R. J., OTT, C. M., PIERSON, D. L. & FOX, G. E. 2017. The adaptation of *Escherichia coli* cells grown in simulated microgravity for an extended period is both phenotypic and genomic. *NPJ Microgravity,* 3**,** 15.

TOPRAK, E., VERES, A., MICHEL, J. B., CHAIT, R., HARTL, D. L. & KISHONY, R. 2011. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat Genet,* 44**,** 101-5.

TRAVISANO, M. & LENSKI, R. E. 1996. Long-term experimental evolution in *Escherichia coli*. IV. Targets of selection and the specificity of adaptation. *Genetics,* 143**,** 15-26.

TSUZUKI, M., ISHIGE, K. & MIZUNO, T. 1995. Phosphotransfer circuitry of the putative multi-signal transducer, ArcB, of *Escherichia coli*: in vitro studies with mutants. *Mol Microbiol,* 18**,** 953-62.

TURNER, A. K., YASIR, M., BASTKOWSKI, S., TELATIN, A., PAGE, A. J., CHARLES, I. G. & WEBBER, M. A. 2020. A genome-wide analysis of *Escherichia coli* responses to fosfomycin using TraDIS-Xpress reveals novel roles for phosphonate degradation and phosphate transport systems. *J Antimicrob Chemother,* 75**,** 3144-3151.

UPPAL, S., SHETTY, D. M. & JAWALI, N. 2014. Cyclic AMP receptor protein regulates *cspD*, a bacterial toxin gene, in *Escherichia coli*. *J Bacteriol,* 196**,** 1569-77.

VAN DEN BERGH, B., SWINGS, T., FAUVART, M. & MICHIELS, J. 2018. Experimental Design, Population Dynamics, and Diversity in Microbial Experimental Evolution. *Microbiol Mol Biol Rev,* 82.

VAN OPIJNEN, T., BODI, K. L. & CAMILLI, A. 2009. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Methods,* 6**,** 767-72.

VAN OPIJNEN, T. & CAMILLI, A. 2012. A fine scale phenotype-genotype virulence map of a bacterial pathogen. *Genome Res,* 22**,** 2541-51.

VAN OPIJNEN, T. & CAMILLI, A. 2013. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. *Nat Rev Microbiol,* 11**,** 435-42.

WALL, E., MAJDALANI, N. & GOTTESMAN, S. 2018. The Complex Rcs Regulatory Cascade. *Annu Rev Microbiol,* 72**,** 111-139.

WANG, L., SPIRA, B., ZHOU, Z., FENG, L., MAHARJAN, R. P., LI, X., LI, F., MCKENZIE, C., REEVES, P. R. & FERENCI, T. 2010. Divergence involving global regulatory gene mutations in an *Escherichia coli* population evolving under phosphate limitation. *Genome Biol Evol,* 2**,** 478-87.

WATSON, N., DUNYAK, D. S., ROSEY, E. L., SLONCZEWSKI, J. L. & OLSON, E. R. 1992. Identification of elements involved in transcriptional regulation of the *Escherichia coli* cad operon by external pH. *J Bacteriol,* 174**,** 530-40.

WEBER, P. C. 1995. Analysis of Tn5 inversion events in *Escherichia coli* plasmids. *Mol Gen Genet,* 248**,** 459-70.

WEBER, P. C., LEVINE, M. & GLORIOSO, J. C. 1988. Simple assay for quantitation of Tn5 inversion events in *Escherichia coli* and use of the assay in determination of plasmid copy number. *J Bacteriol,* 170**,** 4972-5.

WILKS, J. C. & SLONCZEWSKI, J. L. 2007. pH of the cytoplasm and periplasm of *Escherichia coli*: rapid measurement by green fluorescent protein fluorimetry. *J Bacteriol,* 189**,** 5601-7.

WONG, Y. C., ABD EL GHANY, M., NAEEM, R., LEE, K. W., TAN, Y. C., PAIN, A. & NATHAN, S. 2016. Candidate Essential Genes in *Burkholderia cenocepacia* J2315 Identified by Genome-Wide TraDIS. *Front Microbiol,* 7**,** 1288.

XU, H., LUO, X., QIAN, J., PANG, X., SONG, J., QIAN, G., CHEN, J. & CHEN, S. 2012. FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One,* 7**,** e52249.

XU, Y., ZHAO, Z., TONG, W., DING, Y., LIU, B., SHI, Y., WANG, J., SUN, S., LIU, M., WANG, Y., QI, Q., XIAN, M. & ZHAO, G. 2020. An acid-tolerance response system protecting exponentially growing *Escherichia coli*. *Nat Commun,* 11**,** 1496.

YAMAZAKI, Y., NIKI, H. & KATO, J. 2008. Profiling of *Escherichia coli* Chromosome database. *Methods Mol Biol,* 416**,** 385-9.

YASIR, M., TURNER, A. K., BASTKOWSKI, S., BAKER, D., PAGE, A. J., TELATIN, A., PHAN, M. D., MONAHAN, L., SAVVA, G. M., DARLING, A., WEBBER, M. A. & CHARLES, I. G. 2020. TraDIS-Xpress: a high-resolution whole-genome assay identifies novel mechanisms of triclosan action and resistance. *Genome Res,* 30**,** 239-249.

YOHANNES, E., THURBER, A. E., WILKS, J. C., TATE, D. P. & SLONCZEWSKI, J. L. 2005. Polyamine stress at high pH in *Escherichia coli* K-12. *BMC Microbiol,* 5**,** 59.

ZHANG, R., XU, W., SHAO, S. & WANG, Q. 2021. Gene Silencing Through CRISPR Interference in Bacteria: Current Advances and Future Prospects. *Front Microbiol,* 12**,** 635227.

ZHANG, Y. J., REDDY, M. C., IOERGER, T. R., ROTHCHILD, A. C., DARTOIS, V., SCHUSTER, B. M., TRAUNER, A., WALLIS, D., GALAVIZ, S., HUTTENHOWER, C., SACCHETTINI, J. C., BEHAR, S. M. & RUBIN, E. J. 2013. Tryptophan biosynthesis protects mycobacteria from CD4 T-cell-mediated killing. *Cell,* 155**,** 1296-308.

ZHAO, L., ANDERSON, M. T., WU, W., T MOBLEY, H. L. & BACHMAN, M. A. 2017. TnseqDiff: identification of conditionally essential genes in transposon sequencing studies. *BMC Bioinformatics,* 18**,** 326.

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.