UNIVERSITY OF LIVERPOOL

Epistasis Within the Arabinose Operon and its Regulatory Sequences

Thesis submitted in accordance with the requirements of the University of Liverpool for the

degree of Doctor in Philosophy by Jack James Fitzpatrick.

March 2023

Table of Contents

List of Figures

List of Tables

Glossary

Term	**Definition**
Fitness	The ability of an organism to survive and reproduce in a given environment.
Pairwise interactions	When two mutations interact causing their effects on fitness to change.
Epistasis	When the effect of multiple mutations differ from the additive expectation of those mutations.
Additive effects/expectation	When the fitness (ω) of a double mutation AB is equal to ωA*ωB.
Non-additive effects	When the fitness of a double mutation AB differs from the expectation of ωA*ωB.
Magnitude epistasis	When the scale of the fitness effect of a double mutant is greater than the additive expectation.
Positive epistasis	When the fitness of a double mutant is higher than the additive expectation.
Negative epistasis	When the fitness of a double mutant is lower than the additive expectation.
Sign (of a mutation)	Whether a mutation has a positive (+) or negative (-) effect on fitness.
Sign epistasis	When a mutation changes its sign in the presence of another mutation.
Reciprocal sign epistasis	When 2 or more mutations change their sign in the presence of each other.
Fitness landscape	A 3D landscape where the X and Y axes represent genotype space (sequential mutation from genotype A to genotype B) and the Z dimension is fitness.
Global optima	The peak(s) with the highest fitness value achievable by a genotype.

Local optima	Peaks of fitness within fitness landscapes which are lower than the global optimum but are surrounded by lower fitness genotypes, therefore causing some populations to become 'stranded' and limiting their access to more fit genotypes.
Smooth fitness landscape	A fitness landscape with one fitness peak that can be accessed from anywhere in the genotype space.
Rugged fitness landscape	A fitness landscape with multiple fitness peaks, some being local optima, separated by valleys of lower fitness.
Higher order epistasis	Epistasis among more than two mutations than cannot be decomposed into pair-wise epistasis.

Table of Abbreviations

Term	**Meaning**
CRE	*Cis*-regulatory element
HGT	Horizontal Gene Transfer
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
mRNA	Messenger Ribonucleic Acid
ATP	Adenosine Triphosphate
TF	Transcription Factor
LB	Lysogeny Broth
E. coli	*Escherichia coli*
SOC	Super Optimal broth (with Catabolite repression)
PCR	Polymerase Chain Rection
Venus-YFP	Venus-Yellow Fluorescent Protein
OD_{600}	Optimal Density (at 600nm)
KO	KEGG Orthology

Acknowledgements

I would firstly like to thank my supervisors; Jay Hinton, for being an invaluable mentor throughout my PhD and for helping me to complete my thesis in a supportive and understanding way. Jon Bollback, for seeing potential in me and providing guidance throughout my time at Liverpool. Lastly, Kate Baker, for helping me understand the world of bioinformatics and for being the best karaoke partner to sing Alanis Morissette with. I am also extremely grateful to the other academic staff on the Lab H corridor, Heather for your moral support and energetic personality, Jamie for your scientific guidance and Mal for good chats in the kitchen area!

I am extremely fortunate to have met so many incredible people over my PhD journey and would like to thank every member of Lab H with who I have interacted with during my time here. Paul Loughnane is the best lab manager anyone could ask for and there is no better person to attend a beer festival with. For everyone who I've enjoyed lunch with over the years; you are the reason I was able to keep going. I'd like to give a huge thank you to Charlotte Chong who has been one of the best friends I could ask for and has endlessly supported and helped me when bioinformatics software was not behaving! I equally could not have pushed through the struggles without Jordan's jokes at lunchtime- I have made two lifelong friends so don't go anywhere!

Speaking of lifelong friends, I want to thank Ross Mulhall; we started our PhD journey together and now we are taking on a new adventure, one that I couldn't imagine doing with anyone else. You and Josh have provided endless laughs and friendship and I can't wait to see where we go next. Another friend who has been there from start to finish is Rama Bhatia; we started our PhDs at the same time and have been there for each other through thick and thin. I wish you all the best in your new job and look forward to a long friendship. I'd like to extend another thank you to another very close friend, Tanya Horne. Although we only met halfway through my PhD journey, our friendship has quickly become one I'll never forget. From our 'quick' catch ups to the more direct support you've shown me, I am extremely grateful to have met such an inspiring and steadfast friend. I am also extremely grateful for Blanca Perez Sepulveda, someone who has provided me support and been my biggest cheerleader. I consider it a privilege to have been supported by you over the years and look up to you with the utmost admiration; you are a natural at inspiring and lifting people up when they need it.

I would like to take this chance to thank so many others I was lucky to have met during my studies; Matt for our endless Pokemon discussions; Malaka for our mutual love of McCoys; Becky Bennett for our mutual love of Marvel; Hannah for enlightening me about poo; Arthur for giving me body inspiration; Lewis for our mutual addiction to league of legends; Kier for his fabulous taste in coats; Léo for his fabulous taste in cheese; Aisling for her fabulous taste in beer; Yan for his infinite bioinformatic wisdom; and Simon and Nico for their limitless lab wisdom. I would also like to thank Pavel Payne and Hande Acar for their guidance in the early stages of my journey.

Aside from the friends I made during my time in Liverpool, I also have a support network at home that I am grateful for. Thanks to my best friend Steve for much needed down time and the rest of the 'gaming' guys, Josh, Jms and Mike who filled my evenings with laughter, mostly at Mike's expense. Thanks also to all my friends in the 'Wirral squad' as well as the day tripping group. I'd also like to acknowledge my friends from undergraduate, Shaz and Rhiannon, for many fun trips to Leeds and Lauren, who has been on an almost identical journey with her PhD and submitted her thesis just days before me. We made it.

Lastly, I would not be anything without my immensely supportive family. Mum, you have supported me my whole life and no less during this PhD, whether through trips to TGI Fridays or a much-needed hug. You are my rock. Dad, you've always been a voice of reason and stability through the chaos and provided much needed comedic respite when times were tough. Olivia, I am lucky to have a sister as close as you, someone to complain to if Mum annoys us, and someone to share a KFC with when its

needed. Thanks to Jo, Si, Dan and Alex who add some crazy into Christmas day! Additional thanks to Carl, Debs and Byron who have also supported me throughout my journey.

This paragraph is reserved for a special person. My partner Fiona. I can not put into words how much you have helped me. Not only did we meet as we both started our studies, but you have been there for me every step of the way. Your limitless kindness and selflessness has made this entire process infinitely more bearable and I am eternally grateful for your support and love. I would not be here without you.

I was told to leave the most important thank you until last so thank you to Oreo, Felix, Fletcher and Milo whom without I could not have completed this PhD. One cat cuddle per day is a necessity to staying sane.

PhD Timeline

Year	*Month*	*Events*
2017	September	Started lab work
	October	Official start date
2018		
2019		
2020	January	
	February	
	March	COVID UK Lockdown 1
	April	
	May	
	June	
	July	Lab reopened part time
	August	Cessation of consumables funding
	September	
	October	
	November	COVID UK Lockdown 2
	December	
2021	January	COVID UK Lockdown 3
	February	Lab reopened full time
2022	January	
	February	
	March	Original submission deadline
	April	
	May	Change of primary supervisor
	June	
	July	
	August	
	September	Started full time job as Lecturer
	October	
	November	
	December	
2023	January	
	February	
	March	Thesis submission

Thesis Abstract

Title of thesis: Epistasis within the arabinose operon and its regulatory sequences

Bacteria have evolved to fill almost every niche on the planet, ranging from sea ice to thermal vents. These organisms can be used to our advantage to break down pollutants and produce industrially relevant enzymes but can also be harmful to us, by rapidly developing antibiotic resistance and increased virulence in short timeframes. Bacteria evolve through many different mechanisms, one such mechanism is epistasis. Epistasis, or gene interactions, is the phenomenon by which the effect on fitness of a gene is dependent on the genetic background in which it is present. In this thesis I aimed to study epistasis within the arabinose operon and its regulatory regions, specifically, the cis-regulatory element (CRE) of the *pBAD* promoter.

In the second chapter, I investigated whether patterns of epistasis amongst double mutants in the *pBAD* CRE would be consistent when measuring fitness when compared with expression measurements from a previous study. I adapted the mutant library from the previous study and inserted the *araBAD* genes in place of the fluorescent protein on the plasmid construct. I then measured the growth rate of mutants on arabinose as a sole carbon source and calculated epistasis values from the relative growth rates of mutants. I found that, when using growth rate data, no significant epistasis could be detected. I concluded that this was likely due to limitations of the approach and equipment used.

In the third chapter, I wanted to see if patterns of epistasis for the previously created mutant library differed between environments. I grew the mutant library in four different environments, varying in either sugar or temperature and used growth rates to calculate epistasis. I found that some epistatic effects changed between environments but others remained constant. I concluded that some epistatic interactions were strong enough to resist changes in the environment, whilst others were affected non-additively by environmental factors.

In Chapter 4 I took a different approach and searched for 'footprints' of epistasis within the arabinose operon genes. Horizontal gene transfer (HGT) usually indicates an absence of epistasis and so I wanted to find evidence of the occurrence of HGT or the absence of HGT within the arabinose operon. I firstly analysed the gene neighbourhood of the arabinose operon in multiple bacterial species and found large amounts of conservation in the Enterobacteriaceae family, with less conservation when analysing the Gammaproteobacteria class. I then undertook gene cluster analysis to study the chromosomal organisation of the arabinose genes. I found that *araA* and *araB* often occur together and are co-linear, however, the synteny *araC* and *araD* was less conserved. I suggest this could be due to the presence of epistatic interactions between *araA* and *araB*, reflecting the roles of the encoded enzymes in arabinose catabolism. I also analysed the phylogenetic trees of the arabinose genes compared to a core species tree to identify incongruence which can be an indicator of HGT. I generated branch score values for each gene tree but these provided limited insight and I concluded that there may not be enough diversity in the arabinose genes to provide an accurate phylogeny for comparison.

Taken together, my findings provide insight into the epistatic forces that shape the evolution of the arabinose operon and its regulatory sequences. I find that epistasis, both environment-dependent and environment-independent, shape the evolutionary potential of the *pBAD* CRE and that epistatic interactions between select genes of the arabinose operon are likely to limit potential rearrangements of gene co-linearity. Overall, epistasis is a significant evolutionary force that acts on the arabinose operon and this may extend to other bacterial operons, an interesting area to explore going forward.

CHAPTER 1

General Introduction

1.1. Bacterial Evolution

Bacteria are one of the most diverse groups of organisms, having evolved to adapt to various environments ranging from hot springs and sea ice to plant and animal microbiomes (Arrigo, 2014; Davenport *et al.*, 2017; Marsh & Larsen, 1953; Trivedi *et al.*, 2020). Indeed, bacteria provide some of the most important functions to sustain life on the planet such as oxygenating the atmosphere and fixing nitrogen for plants to use (Madigan *et al.*, 2012). The diversity of bacteria on the planet can also be useful for human existence and we can often study and use them to our advantage. Examples include the use of thermostable enzymes from thermophilic bacteria. Thermophilic bacteria live in high temperature habitats and can grow at temperatures exceeding 100°C (Takai *et al.*, 2008). Due to evolving in harsh environments, thermophilic bacteria have evolved enzymes that are extremely heat stable which can be isolated and harvested for industrial use (Chien *et al.*, 1976; Kambourova, 2018). Not only can such heat stable enzymes be isolated directly but also studied to use directed evolution in the lab to create enzymes with very specific traits (Reeve & Fuller, 1995). However, not all bacteria are beneficial with some evolving to become pathogenic and infecting a wide diversity of hosts (Casadevall & Pirofski, 1999). This pathogenic evolution can prove detrimental to humans and has presented a unique challenge for societies throughout history.

The study of bacterial evolution has historically presented greater challenges than that of plants and animals. Without the distinct morphological characteristics observed in modern animal and plant species and that of the fossil record it was difficult for scientists to phylogenetically classify bacteria, let alone investigate bacterial evolution. It was only the advent of DNA sequencing technology that allowed scientists to begin to decipher the relationships between bacteria and study the evolutionary steps involved (Woese, 1987).

Studying bacterial evolution is key in helping tackle challenges such as antibiotic resistance and bacterial virulence. Antibiotic resistance is driven by rapid acquisition and spread of resistance genes amongst as bacterial population and understanding the mechanisms by which this spread occurs is crucial to preventing the long-term redundancy of antibiotics (Waclaw, 2016). Bacterial virulence can also be studied through an evolutionary lens, complementing the study of pathogenesis by viewing virulence as a factor in pathogen spread and survivability in the host (Diard & Hardt, 2017). Virulence

is the result of fitness trade-off between the host and pathogen so is intrinsically linked to bacterial evolution (Alizon & Michalakis, 2015).

These examples highlight how the study of bacterial evolution can help better our understanding of bacterial ecology and pathogenicity, advancing the fields of biotechnology (Mavrommati *et al.*, 2022) and medicine (Christaki *et al.*, 2020). One aspect to consider when studying the evolution of bacteria is what mechanisms shape the evolutionary potential of organisms and how they restrict or allow organisms to evolve certain traits. One such mechanism is epistasis.

1.2. Epistasis

Epistasis (genetic interactions) is the phenomenon by which the fitness effect of a given mutation depends on the genetic background on which the mutation appears (Domingo *et al.*, 2019; Phillips, 2008). A prime example of epistasis is in bacterial toxin-antitoxin systems. The toxin gene is lethal when present on a background lacking the anti-toxin gene and, conversely, the cell is wasting energy producing anti-toxin if no toxin is present, therefore decreasing its fitness. However, when both genes are present there is a marked fitness increase for the cell by eliminating its competitors (Unterholzner *et al.*, 2013).

Epistasis as a term has been around for many years and so there can be confusion when it is used in different contexts. Whilst the definition above is the most widely encompassing one, epistasis can also be considered to be an interaction between two mutations. In this way, epistasis is described as any interaction in which the fitness of the double mutant differs from the expected additive effects of the constituent single mutants (Fisher, 1919; Wong, 2017). This will be the definition of epistasis referred to in **Chapter 3** and **Chapter 4**.

Epistasis occurs in different forms, the most common being magnitude epistasis. This is where the fitness of a double mutant is greater or lesser than the additive expectation. Positive epistasis refers to an increase in fitness whereas negative epistasis refers to a decrease in fitness. Another form of epistasis is sign epistasis (Weinreich *et al.*, 2005) where the fitness 'sign' of a mutation is reversed, going from positive to negative or vice versa. Reciprocal sign epistasis occurs when both mutants respectively change signs in the presence of one another (Phillips, 2008). **Figure 1.1** shows the fitness landscapes of these forms of epistasis.

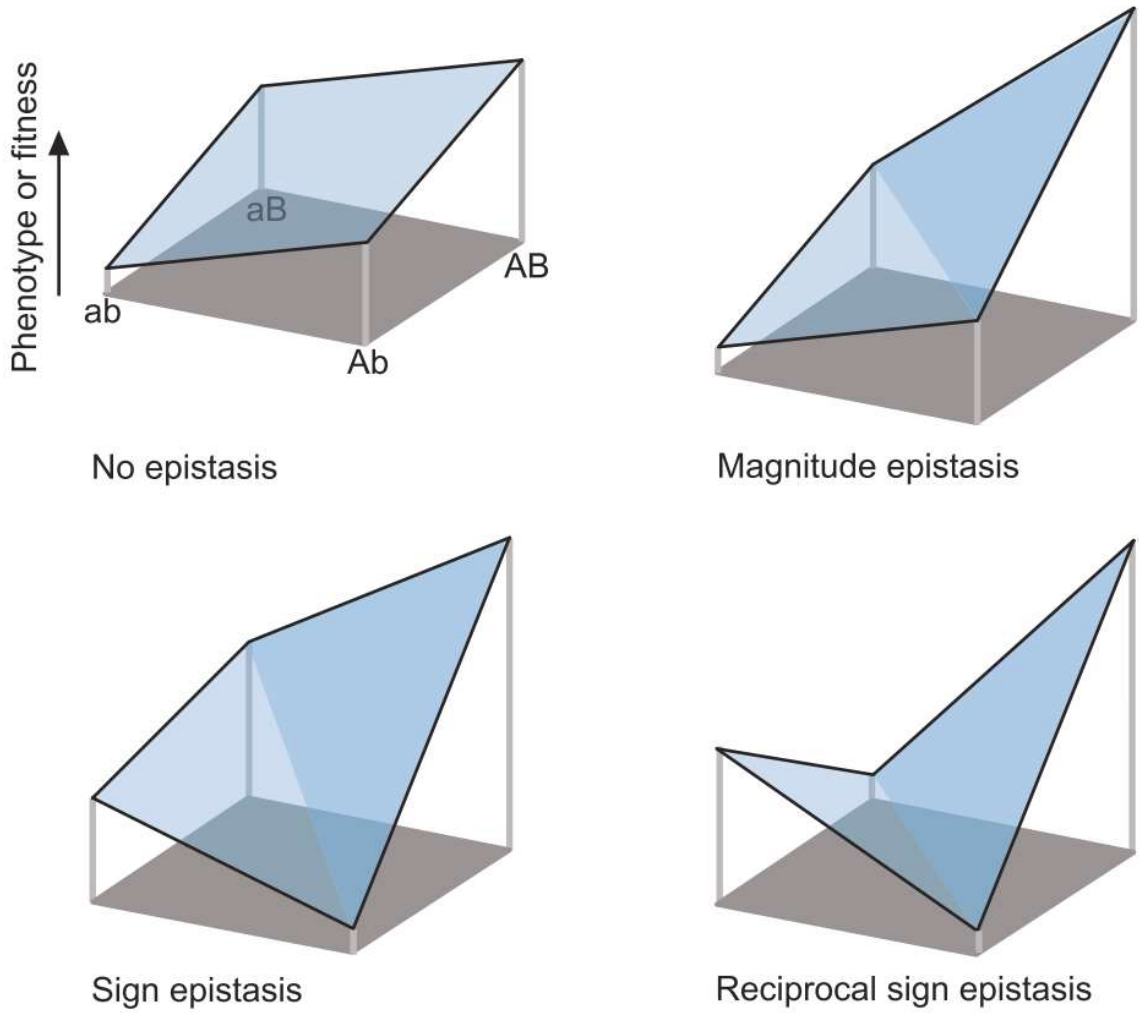


Figure 1.1 Graphical representation of different forms of epistasis and associated fitness landscapes. *Alleles are denoted by a, b, A and B. Capitalised alleles indicate the mutant allele. No epistasis shows two alleles (A and B) conveying an increase in fitness which is additive when the alleles are combined. Magnitude epistasis shows two alleles that provide a greater increase in fitness when combined than the additive effects of each individual allele. Sign epistasis shows two alleles of opposing fitness effects coming together to provide an overall increase in fitness therefore changing the sign of one of the alleles. Reciprocal sign epistasis demonstrates two alleles which have negative fitness effects combining to give a fitness benefit. Adapted from: (Dawid et al., 2010).*

Epistasis is inherently a complex process and is generally poorly understood due to the combinatorial nature of interactions. Epistasis can exist between mutations both within genes (intragenic) and between genes (intergenic), which generally occur at similar frequencies when studying compensatory mutations (Poon & Chao, 2005). Intergenic epistasis is extremely prevalent and affects quantitative traits among organisms through pleiotropy (Flint & Mackay, 2009). Intergenic epistasis is more easily conceptualised as it can be the result of two proteins physically interacting, this has been shown by observation of compensatory mutations appearing in the L19 protein of the large ribosomal subunit of *Salmonella enterica* serovar Typhimurium to compensate for fitness costs in the S12 protein of the small subunit arising due to antibiotic resistance (Maisnier-Patin *et al.*, 2007). Intergenic epistasis is also evidenced in metabolic pathways where the activity of upstream enzymes is increased but the downstream enzymes create a bottleneck in the system and so the increased metabolic flux of the system is not realised. The increase can only be realised if both enzymes increase activity proportionally (Kacser & Burns, 1981).

Intragenic epistasis is somewhat more complex than intergenic epistasis as intragenic epistasis involves interactions between individual mutations within genes. One form of intragenic epistasis is threshold epistasis, which relates to the fact that some proteins have a threshold of stability and that most proteins have excess stability beyond this threshold. Consequently, individual mutations would have little effect but in combination can affect the stability of the protein (Bershtein *et al.*, 2006). It is proposed that threshold epistasis could be a defence against stochastic change to ensure protein stability (Lehner, 2011). Another major source of intramolecular epistasis is conformational epistasis, which occurs when a conformational change is required alongside a residue change to result in novel functionality. Conformational epistasis has been shown to occur in glucocorticoid receptors (Ortlund *et al.*, 2007). Epistatic effects can also exist in the form of 'global suppressors' where a mutation acts to suppress the effects of all destabilising mutations by increasing protein stability universally (Shortle & Lin, 1985). Similarly, universal 'enabling' mutations have been found in bacterial toxins that enable several subsequent mutations to arise, allowing the toxin to resist disruption by the anti-toxin (Ding *et al.*, 2022).

Epistasis can also occur within non-coding regulatory sequences. The regulation and expression of a gene can often play an important role in the resultant phenotype and this explains why epistatic effects are observed in non-coding, regulatory sequences. An example of epistasis within non-coding regulatory sequences is where one mutation reducing the activity of an enzyme may be compensated for by a mutation in the promoter sequence causing increased expression levels. Regulatory epistasis has been shown empirically in relation to *Salmonella enterica* antibiotic resistance (Paulander *et al.*, 2010) and within *E. coli* enzyme evolution (McLoughlin & Copley, 2008). Epistasis also has effects on

global gene regulation through the action of global regulator proteins (Srinivasan *et al.*, 2013) and so can influence gene regulation in different ways.

1.3. Fitness Landscapes

Due to the effects of epistasis on fitness, epistatic effects could be expected to influence evolutionary outcomes. By nature, epistasis shapes the evolutionary landscape rendering some routes to higher fitness inaccessible due to unfavourable mutation combinations. Limiting evolutionary pathways can restrict the ways in which organisms can adapt and react to certain evolutionary pressures (Weinreich *et al.*, 2005).

It is generally accepted that natural selection works by selecting individuals that are the most well adapted to their environment and therefore have an increased chance to survive and reproduce (B. K. Hall *et al.*, 2014). The measure by which suitability of an individual to an environment is quantified is termed 'fitness' (Thoday, 1953). The individuals with the highest fitness in a given environment are more successful and so natural selection favours mutations which increase the ability of an organism to survive and reproduce in its environment. It could therefore be expected that every beneficial mutation would be selected for unconditionally until an individual reaches the peak possible genotype for a given environment, thereby maximising its fitness.

However, when epistasis is considered, the route to the highest fitness may not be a straightforward one. When the fitness of a mutation is dependent on the background the mutation appears on (meaning the other mutations that preceded them) then each mutation may situationally be adaptive or not, having a different fitness value depending on background (Phillips, 2008). These genetic interactions would then create a fitness 'landscape' where acquiring mutations would move species towards or away from fitness peaks and natural selection would favour movement towards peaks of higher fitness. The concept of fitness landscapes appears straightforward, however, multiple peaks can be present within a landscape, some being higher than others. Species could consequently become stuck on a local, sub-optimal peak as natural selection will not permit the species to move to a higher peak through a 'valley' of low fitness and so reaching the global optimum becomes a difficult task involving acquiring beneficial mutation in a specific order to navigate the fitness landscape and not become stuck on a sub-optimal peak (de Visser & Krug, 2014; Fragata *et al.*, 2019; Wright, 1932).

Following this logic, fitness landscapes can either be 'smooth', containing a single global peak where beneficial mutations can be gained in any temporal order to advance a species towards the peak. Alternatively, the landscapes can be 'rough', where multiple local peaks are present as well as fitness

valleys and the route to the global optimum requires precise ‘navigation’, gaining mutations in a specific order. **Figure 1.2** demonstrates the types of fitness landscapes.

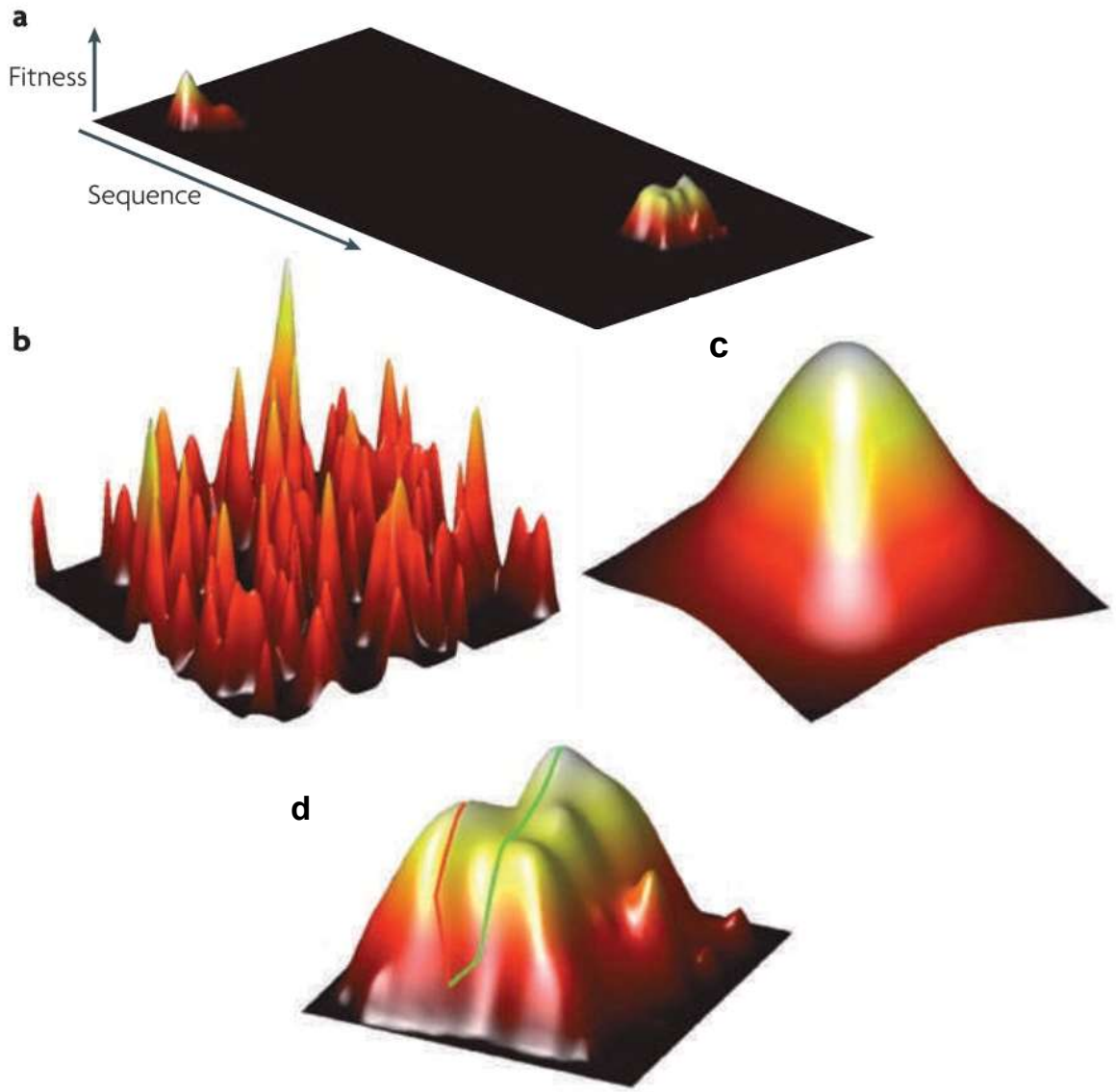


Figure 1.2 Various landscapes in the sequence fitness space. *Black indicates low fitness proceeding through red, then yellow and finally white to represent the highest fitness achievable. a) A realistic fitness landscape where a large portion of possible sequences are non-functional. b) An extremely rough landscape containing many local optima surrounded by deep fitness valleys limiting movement through sequence space. c) A smooth landscape containing one global peak where every acquired mutation increases fitness no matter the order. d) A depiction of 2 possible evolutionary pathways through sequence space; one becoming stuck on a local optimum (red) and the other reaching the global peak (green). Taken from: (Romero & Arnold, 2009).*

1.4. Operons

Given that epistasis can occur both within genes, between genes and in regulatory sequences, an ideal system to study epistasis lies in bacterial operons which are clusters of colocalised genes that are transcribed from a single promoter. Bacterial operons were initially defined as 'coordinated units of expression' (Jacob *et al.*, 1960) and the definition has since been expanded to include 'clusters of co-regulated genes with related functions' (Osbourn & Field, 2009) and 'any group of adjacent genes that are transcribed from a promoter into a polycistronic mRNA' (Fondi *et al.*, 2009). Operons are widespread amongst bacteria and archaea and are the most common form of gene organisation in prokaryotes (Koonin, 2009). Most operons are weakly conserved with a few significant exceptions including the ribosomal superoperon and proton ATPases which often encode proteins that physically interact (Brandis, 2021; Itoh *et al.*, 1999; Wolf *et al.*, 2001).

Operons account for a significant portion of protein coding genes within an average bacterial genome (Ermolaeva *et al.*, 2001; Price, Huang, Alm, *et al.*, 2005; Wolf *et al.*, 2001). *Escherichia coli* has been predicted to have 700 operons in its genome which accounts for 55% of its gene content (Salgado *et al.*, 2000). It has been speculated that operons exist to help co-regulate genes that are functionally related and would benefit from tight stoichiometric control (Rocha, 2008). The selection against rearrangement of operons supports this (Brandis, 2021; Rocha, 2006) although operons are not immune to rearrangement (Brandis, 2021). Operons often encode proteins within the same functional pathway, but not always (Rogozin *et al.*, 2002), suggesting stoichiometric balance is not necessarily the lone driver behind operon formation and co-regulation of genes. One theory for the existence of operons which contain functionally unrelated genes is that functionally unrelated genes may be required as part of a holistic function such as growth and so still benefit from being co-regulated (Price *et al.*, 2006; Rogozin *et al.*, 2002).

There is evidence that some orthologous operons are noticeably diverged from one another yet the operons still retain function (Buvinger & Riley, 1985; Leonard *et al.*, 2015). Retention of function suggests that the genes within the operons may be coevolving together to maintain function (Dover & Flavell, 1984; Lovell & Robertson, 2010), thereby giving rise to epistasis.

1.5. Operon Promoters

Promoters are the transcriptional start sites for all genes and operons and are responsible for binding RNA polymerase and unwinding the DNA duplex surrounding the transcription start site (Browning & Busby, 2004). There are two key sites involved in binding the RNA polymerase, the -10 site located 10bp upstream of the transcription start site and the -35 site, located 35bp upstream. Both the -10

and -35 sites bind domains 2 and 4 of the σ subunit of RNA polymerase, respectively. Whilst there are thousands of promoter sequences within each genome (Salgado *et al.*, 2001), there is an uneven distribution of RNA polymerase amongst said promoters. The unequal distribution of RNA polymerase indicates that mechanisms are present to control the balance of RNA polymerase between promoters at any given time. One of the ways promoters can regulate the availability of RNA polymerase is through *trans*-acting factors. There are several *trans* factors, including sigma factors, small ligands and transcription factors (Babu & Teichmann, 2003). Transcription factors are often DNA binding proteins and so have interactions with *cis* binding sites, meaning regulation via transcription factors is dependent on the DNA sequence of the cis-acting site and as a result, can be affected by epistasis. Mutations in regulatory regions have a marked effect on evolutionary outcomes (Stern & Orgogozo, 2008) and so they can often be an important target for selection to act on.

Promoters initiate transcription by interacting with transcription factor proteins (TFs), which can occur in one of two ways; the promoters either bind the transcription factor, which then helps to recruit RNA polymerase to the transcription start site by physically interacting with its subunits, or the transcription factor causes a change in conformation of the promoter region, allowing RNA polymerase to more easily recognise the transcriptional start site (Browning & Busby, 2004).

Operon promoter sequences are often described as more complex than the sequences of individually transcribed genes (Hazkani-Covo & Graur, 2005; Price *et al.*, 2006; Price, Huang, Arkin, *et al.*, 2005). The increased complexity of operon promoter sequences could help to explain why genes within operons may cluster under the control of one promoter, as this saves each gene individually needing as complex promoter (Rocha, 2008). One example of a well-studied operon that is both repressed and activated by a transcription factor is the arabinose operon (Schleif, 2010).

1.6. Arabinose Operon

One of the most well studied and widely used model operons in biology is the arabinose operon of *Escherichia coli* and this was the system chosen for study in this thesis. The arabinose operon was selected because it is inducible via a single, environmental factor; the presence of arabinose. The operon consists of one regulatory gene, *araC* and 3 metabolic genes, *araB*, *araA* and *araD*, collectively known as *araBAD*. The operon also contains several regulatory sites; *araO1L*, *araO1R*, *araO2*, as well as two component of the *cis* regulatory element (CRE) *araI1* and *araI2* (Schleif, 2000). AraC acts as a negative regulator of *araBAD* in the absence of arabinose and a positive regulator in the presence of arabinose (Englesberg *et al.*, 1965; Schleif, 2010). The mechanism of regulation involves AraC forming a homodimer which in the absence of arabinose binds to *araO2* and *araI1* causing the DNA to loop and prevent transcription of the *araBAD* genes by blocking the polymerase binding. In the presence

of arabinose, the sugar binds to the AraC dimer, changing the conformation, and causing the dimer to preferentially bind to *araI1* and *araI2* allowing the DNA loop to relax, allowing access and also recruiting RNA polymerase to the *pBAD* promoter to transcribe the *araBAD* genes (Schleif, 2000). This is shown in **Figure 1.3**. The DNA binding regions araI1 and araI2 contain nucleotide sequences that specifically bind AraC depending on the presence of arabinose (Zhang *et al.*, 2018). These regions were named 'A-box' and 'B-box'. It was found that the araI1 region contained both an A-box and B-box, whereas araI2 only contained a B-box. Conversely, araO2, which binds AraC in the absence of arabinose, only contained an A-box. The current understanding is that the A-box is responsible for binding AraC when arabinose is present, whilst the B-box is responsible for binding AraC in the absence of arabinose (Niland *et al.*, 1996).

Arabinose catabolism has been shown to be important for pathogenesis as L-arabinose is a key driver for the proliferation of *Salmonella enterica* serovar Typhimurium in the gastrointestinal tracts of superspreader hosts (Ruddle *et al.*, 2023). Arabinose catabolism results from 3 catabolic enzymes working in conjunction; AraA, AraB and AraD. The *araA* gene encodes L-arabinose isomerase which converts L-arabinose into L-ribulose. The *araB* gene encodes ribulokinase which phosphorylates L-ribulose to give L-ribulose-5-phosphate. The *araD* gene encodes L-ribulose-phosphate-4-epimerase which converts L-ribulose-5-phosphate to D-xylulose-5-phosphate which can then enter the pentose phosphate pathway (Schleif, 2022).

1.7. Arabinose Transport Operon

In contrast to the *araBAD* operon which is responsible for degrading arabinose, the *araFGH* operon facilitates the transport of arabinose into the cell (Horazdovsky & Hogg, 1987). The *araFGH* operon encodes the 'high-affinity' arabinose transport system, whilst the monocistronic *araE* gene encodes the 'low-affinity' system. AraE is a permease which is energized by proton motive force (Brown & Hogg, 1972). AraF is a periplasmic L-arabinose binding protein which is an important component of the high affinity system (Hogg, 1977). AraG is an ATP binding protein (Horazdovsky & Hogg, 1987). AraH is the permease component of the ABC transporter (Horazdovsky & Hogg, 1987). AraE, F, G and H are all membrane-bound proteins that channel arabinose into the cell (Luo *et al.*, 2014) when glucose is absent. If glucose is present, AraC represses the transcription of the *araFGH* and *araE* genes by the same mechanism as the *araBAD* operon, leading to a significant reduction in expression (Luo *et al.*, 2014).

Arabinose is a major component of hemicellulose, the major constituent of plant material (Holtzapple, 2003). Consequently, arabinose is found in many environments including soil and becomes available

to soil microorganisms following hemicellulose degradation. The human gut also contains arabinose following consumption of dietary plant material (Alam *et al.*, 2022). Although arabinose itself is poorly absorbed in the gut, the sugar is described as a microbiota-accessible-carbohydrate (MAC) (Tomioka *et al.*, 2022) that can be utilised as an energy source by the gut microbiota (Tomioka *et al.*, 2022).

1.8. Thesis Aims

In this thesis I aimed to investigate several questions regarding epistasis within the arabinose operon and its regulatory regions. These were:

1: Are patterns of epistasis consistent when considering cell fitness versus gene expression?

2: Are epistatic effects within the pBAD promoter consistent between different environments?

3: Is there evidence of epistasis between genes of the arabinose operon and what can this tell us about operon evolution?

Overall, these questions aim to investigate the effect epistasis has on the evolution of the arabinose operon and the associated regulatory sequences and could contribute to the wider understanding of operon evolution in bacteria.

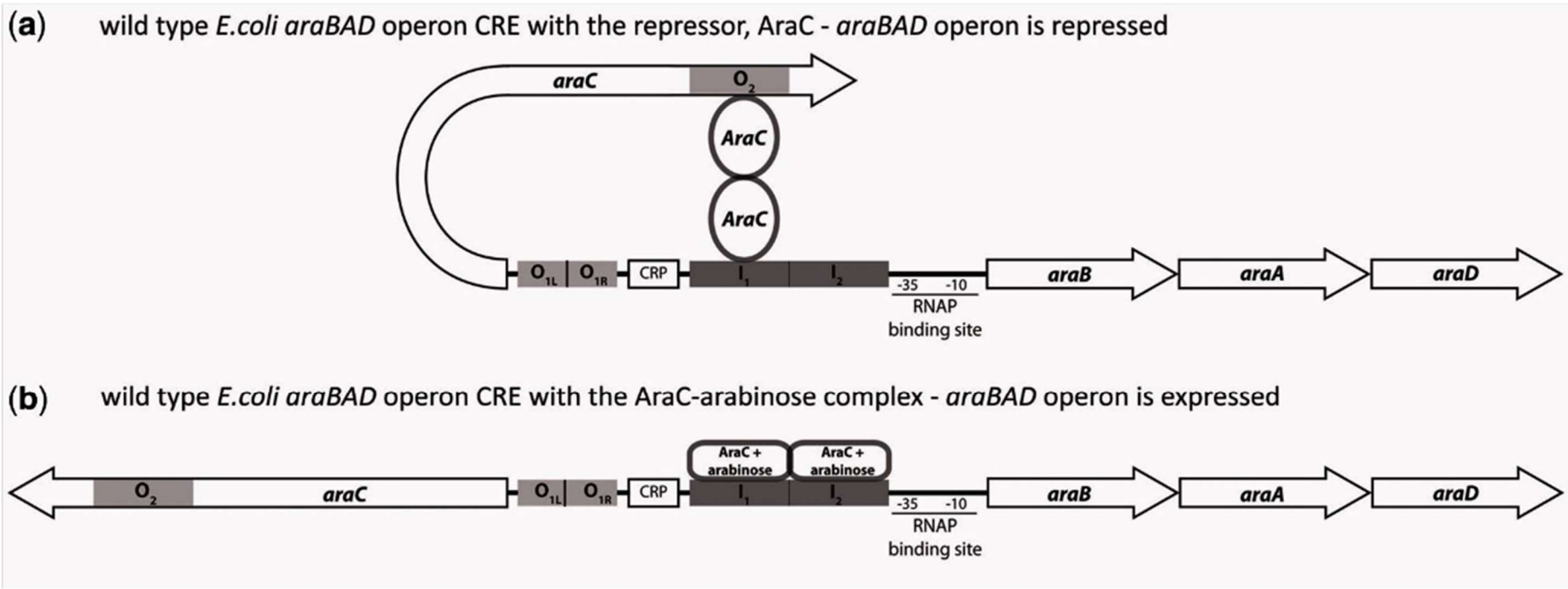


Figure 1.3 The arabinose operon. *a) The araBAD operon being repressed in the absence of arabinose via a dimerised AraC protein binding to araO2 and araI1. The CRP site and the polymerase binding site are blocked and so no transcription can take place. The araBAD genes are not expressed. b) In the presence of arabinose, AraC changes conformation and binds to araI1 and araI2. The DNA loop relaxes exposing the CRP site allowing cAMP to bind promoting the binding of DNA polymerase which initiates transcription and subsequent expression of the araBAD genes. Adapted from: (Lagator et al., 2016).*

CHAPTER 2

2.1. Chapter 3 Methods

2.1.1. Strains

Escherichia coli JW0063.1 from the Keio collection (Baba *et al.*, 2006) was used as the host strain in this study, as all the arabinose operon genes (*araA, araB, araC and araD*) were deleted from the chromosomal background of this strain and would therefore not interfere with the operon genes introduced on the plasmids. The strain carried a kanamycin resistance gene in place of the *araC* gene which was excised using lambda red recombination.

2.1.2. Culturing

Strains were streaked on LB agar plates (with or without kanamycin 50µg/ml depending on requirement) and grown overnight at 37°C. Replicate colonies were then grown up in shaking LB broth (with or without kanamycin 50µg/ml) at 37°C to obtain saturated cultures and these were then placed at -80°c in 15% glycerol for long term storage.

2.1.3. Cloning of Mutant Library

37 plasmids carrying mutant promoters were isolated from *E. coli* strain BW25113 obtained from Lagator *et al.* (2016) using [ZR plasmid miniprep – classic (Zymo Research)]. Plasmid constructs for use in this study were generated using the NEBuilder® HiFi DNA Assembly kit. Primers were created using the NEBuilder® assembly tool (**Table 2.1**).

Table 2.1 Primers used for NeBuilder assembly tool

Primer	Sequence
pZS*2-venus_fwd	AATGAGTAAAGGAGAAGAACTTTTC
pZS*2-venus_rev	CTAGATTGAGCTCTTCCTCC
araBAD_fwd	GGAGGAAGAGCTCAATCTAGATGGCGATTGCAATTGGC
araBAD_rev	GTTCTTCTCCTTTACTCATTCATCAGCTGTTTCTCCTCTTTAATTTACTGCCCGTAATATGCCTTC

Briefly, the plasmid backbone was amplified using primers **pZS*2-venus_fwd** and **pZS*2-venus_rev** in a PCR reaction and the *araBAD* insert was amplified from *E. coli* MG1655 genomic DNA using primers **araBAD_fwd** and **araBAD_rev**. The recommended concentrations of insert and vector were added in a 1:2 ratio into the NEBuilder® HiFi DNA Assembly Master Mix and incubated at 50°C for 15 minutes. This was designed to create a plasmid containing the *araBAD* genes, complete with their ribosome binding sites, alongside Venus-YFP complete with its original RBS from the original plasmid construct. The resulting plasmid can be seen in **Figure 3.3**. The plasmids (**Figure 3.3**) were then electroporated into *E. coli* JW0063.1 cells. These cells were then added to prewarmed SOC broth and incubated for 45 minutes. They were then plated on MacConkey agar base (without sugar) plates supplemented with 0.5% arabinose and kanamycin (50 µg/ml). Successful transformants were identified on agarose gel by PCR amplification to check for the expected plasmid size. The nucleotide sequence of about 95% of the *araCBAD* region of plasmids from mutants 1-8 were verified using a primer walking approach which involves using Sanger sequencing to sequence short, overlapping region of the plasmid to determine the full sequence of a given plasmid (Benes *et al.*, 1997). The DNA sequence confirmed the cloning strategy had been successful and no unexpected mutations were identified. Stocks were prepared for colonies carrying the correctly assembled plasmid in liquid LB supplemented with Kn^{50} and stored at -80℃ with 15% glycerol.

2.1.4. Growth Assays

Frozen strain stocks for the 37 mutants were streaked on MacConkey agar base plates supplemented with 0.5% arabinose and 50 µg/ml kanamycin. The plates were incubated at 37°C overnight and then a single colony was picked from each plate and inoculated into M9 media (KH_2PO_4, 15 g/L, NaCl, 2.5 g/L, Na_2HPO_4, 33.9 g/L, NH_4Cl, 5 g/L) supplemented with 0.1% arabinose, 50 ug/ml kanamycin, $MgSO_4$, $CaCl_2$ and casamino acids. Strains were inoculated in 96 deep well plates using a total volume of 1ml and incubated at 220 rpm and 37℃ for ~18 hrs. A 1:100 dilution of the 18 hr old cultures were grown in 96 deep well plates at 200 rpm and 37℃ for 4 hours to ensure exponential growth. Two different machines were used to grow cultures in 96 well plates, as defined in chapters 3.2 and 4.2.

After 4 hours of incubation, a 1:100 dilution of each mutant culture, in a total volume of 150µl, was added to a Corning® 96 well flat bottom black-walled plate (product code: 3603), with the position of each mutant being randomised, using a Python script to randomly allocate each strain to a well to avoid bias. Growth was measured in a Tecan Infinite® 200 PRO plate reader for 24 hours with OD_{600} measurements taken every 20 minutes. Each experiment was repeated to produce five biological replicates.

2.1.5. Growth Rate Calculation

To calculate growth rates the GrowthRates package in R was used (Hall *et al.*, 2013). GrowthRates was run using the parameters h=8 (sliding window of 8 data points) and quota=0.95 (extrapolate slope to windows within 95% confidence of highest growth rate window) to estimate the maximum growth rate for each mutant. Growth curves were visualised in R, using a panel plot, to check for abnormalities and to ensure growth rates were calculated from appropriate data points in the exponential section of the growth curves. Relative fitness was estimated by normalising the growth rate of the mutants with that of the 'wild type'. This produced five replicate sets of normalised growth values from which an average and standard deviation was calculated for each mutant.

2.1.6. Epistasis Calculation

Epistasis values of double mutants were calculated for each replicate, then a mean average was calculated. Formula for epistasis calculation is given below:

$$\varepsilon = \omega_{m12} - \omega_{m1} \times \omega_{m2}$$

where ω_{m12} is the fitness value of the double mutant and ω_{m1}/ω_{m2} are the fitness values of the respective single mutants.

2.1.7. Statistical Analysis

To determine whether mutants had a different growth rate compared to the wild type, one-tailed one sample t-tests were performed using $\mu_0 < 1$ and $\mu_0 > 1$. The resulting p-values were corrected for multiple tests (Benjamini & Hochberg, 1995) and $\alpha = 0.05$ was used as the level of significance.

A Pearson rank correlation was done to give an indication of the relationship expression level and fitness. This indicated whether there was a positive correlation between the two measurements. All statistical analysis were carried out using R software (4.1.1) in R Studio (Version 1.4.1717).

2.2. Chapter 4 Methods

2.2.1. Stains, Culturing and Cloning

All methods for this chapter were taken from **Section 2.1.** as the same mutant library was used for assays. Please refer to **Section 2.1.** for these experimental details (**Sections 2.1.1. to 2.1.3.**).

2.2.2. Growth Assays

Four conditions were tested; 0.1% arabinose + 37°C, 0.25% arabinose + 37°C, 0.5% arabinose + 37°C and 0.1% arabinose + 30°C. The first condition was treated as the baseline as the conditions were the same as used in the previous chapter. The sugar concentrations were selected as we theorised that below 0.1% the growth rate would be reduced and limit the ability to obtain accurate measurements due to assays needing to be run for much longer periods. Also, approaching 1.6% glucose has been shown to reduce bacterial growth rate (Kazan *et al.*, 1995). As no literature could be found describing arabinose toxicity levels, glucose toxicity levels were used as an estimation. A recent study found that mice provided with water containing 1% arabinose were able to demonstrate a phenotypic change in enteric *Salmonella enterica* serovar Typhimurium (Ruddle *et al.*, 2023), as it is unlikely that the concentration in the gut would be this high after imbibing, it was decided to use concentrations below 1%. As for temperature, the Growth profiler 960 platform (Enzyscreen) could only provide growth temperatures between room temperature and 42°C. Therefore, 37°C and 30°C were selected as they were within the equipment's capability and the temperature interval was sufficiently large so as to increase the likelihood of detecting a difference in epistasis values. 0.1% arabinose was used for the 30°C condition as this was the arabinose concentration used in the previous chapter and by Lagator *et al.* (2016).

Strains were streaked from freezer stocks onto MacConkey agar base (Difco) plates supplemented with 0.5% arabinose and 50 ug/ml kanamycin. The plates were incubated at 37°C overnight and then a single colony was picked from each plate and inoculated into M9 media supplemented with the appropriate concentration of arabinose (0.1%, 0.25% or 0.5%), 50 ug/ml kanamycin, $MgSO_4$, $CaCl_2$ and casamino acids. Each strain was inoculated into 1ml of the media within an individual well of a 96 deep well plate. The cultures were incubated overnight for 18 hours. After incubation the cultures were passaged into a fresh deep well plate using a 1 in 100 dilution and grown for four hours to ensure exponential growth. After incubation a 1 in 100 dilution of each well from the four-hour culture was added to 250µl of media in a randomized well (using the same python script as in **Section 2.1.4.**) in five independent Polystyrene greyish-white square 96-half-deepwell microplates (CR1496dg, Enzyscreen) and the position of each strain within each plate was recorded. The five plates were then placed in the Growth Profiler 960 platform (Enzyscreen) and incubated at either 37°C or 30°C,

depending on the condition being tested, for 48 hours. The Growth Profiler captured images from the underside of the plates every 20 minutes and these images were stored for analysis as described in **Section 4.2.3.**

2.2.3. Image Analysis

Images were converted into OD_{600} equivalent values by converting the green values from well images and fitting them to a calibration curve. The calibration curve was created by measuring the green values of a series of serially diluted cultures of which the OD_{600} values were known. The equation Gvalue = b*$0D_{600}$/a was used where a and b were selected to give the best fitting curve for the calibration. Green values over the 48 hours were then converted to OD_{600} values and these data were saved for growth rate analysis.

2.2.4. Growth Rate Calculation

To calculate growth rates, the R GrowthRates package (Hall *et al.*, 2013) was used. The data was imported into R and GrowthRates was run using the parameters h = 8 (sliding window of eight data points) and quota = 0.95 (extrapolate slope to windows within 95% confidence of highest growth rate window). Growth curves were visualised in R, using a panel plot, to check for abnormalities and to ensure growth rates were calculated from appropriate data points in the exponential section of the growth curves. Slope values were produced for each growth curve and these values were recorded in a new spreadsheet including the five plates from each assay. The growth rates were then normalised against the wild type growth rate by dividing all values from an individual plate by the wild type value from the same plate. This produced five replicates sets of normalised growth values per condition from which an average and standard deviation could then be calculated for each mutant.

2.2.5. Statistical Analysis

All average relative fitness values were compared against the wild type using a Student's one sample T-test comparing against a value of 1 and then FDR corrected. All epistasis values were tested for being significantly different from zero using a Student's one sample T-test and FDR corrected.

2.2.6. Epistasis Calculations

Epistasis values of double mutants were calculated for each replicate plate within each condition and a mean average was calculated for each condition. Formula for epistasis calculation is given below:

$$\varepsilon = \omega_{m12} - \omega_{m1} \times \omega_{m2}$$

where ω_{m12} is the fitness value of the double mutant and ω_{m1}/ω_{m2} are the fitness values of the respective single mutants.

2.3. Chapter 5 Methods

2.3.1. Gene Neighbourhood Analysis

Gene neighbourhood figures (**Figure 5.1, Figure 5.2**) were created using GeCoViz online software available at (https://gecoviz.cgmlab.org/) (Botas *et al.*, 2022). The KEGG Orthologous group (KO) for *araB* (K01804) was selected as the anchor gene as this is the first gene in the *araBAD* operon and is a required enzyme for the arabinose pathway (Cribbs & Englesberg, 1964) so must be present in all functional version of the arabinose operon. The list of genomes selected were the 58 genomes available within GeCoViz for the Enterobacteriaceae family (**Figure 5.1**) and the 29 genomes from 19 families available within the Gammaproteobacteria class (**Figure 5.2**).

2.3.2. Genome Selection

A list of strains to be used in analysis were selected from the NCBI genome webpage [July 2022] using the following parameters: Bacteria, complete, reference, representative. This returned 3838 candidate genomes which represented the broadest available set of bacterial genomes.

2.3.3. Gene Clustering Analysis

Using the genomes selected in **Section 2.3.2.** , a database was created using Diamond v2.0.15 (Buchfink *et al.*, 2015) for use in Cblaster. To search for the arabinose operon CBAD gene cluster across species, Cblaster v1.3.15 (Gilchrist *et al.*, 2021) was used against the database. The query sequence was the *E. coli* K12-MG1655 *araCBAD* operon sequence [which can be accessed here: RegulonDB (unam.mx)] and the following parameters were specified: minimum identity = 70%, minimum hits = 2, minimum unique hits = 2, minimum coverage = 80%. 103 genomes were returned as having Cblaster hits. The output from Cblaster was input into clinker v0.0.25 (Gilchrist & Chooi, 2021) to produce visualisations of the arabinose operon gene clusters (**Appendix 1**).

A separate analysis was done using the core genome output from Panaroo (Tonkin-Hill *et al.*, 2020). For each species, arabinose genes were identified and cross referenced with Cblaster results to define them as syntenic or not. These were then plotted against the core genome phylogeny to identify clade specific patterns of synteny (**Figure 5.3**)

2.3.4. Phylogenetic Tree Analysis

A core phylogenetic tree was required for comparison to the phylogenetic trees of the arabinose genes. The 103 genomes from **Section 2.3.3.** were input into Panaroo (Tonkin-Hill *et al.*, 2020) by Dr Charlotte Chong and a core gene alignment was produced. Dr Chong then passed the core alignment to IQ-TREE (Chernomor *et al.*, 2016) and a phylogenetic tree was produced.

For the arabinose gene trees, Panaroo was run by Dr Chong with a lower threshold of 70% so the arabinose genes would be included in its analysis. The list of genes found were then manually checked against data from Cblaster to determine which genes were part of the canonical operon in each species. The resulting genes were then extracted from their respective genomes and aligned using MAFFT (Katoh *et al.*, 2002) before being passed to IQ-TREE by Dr Chong for phylogenetic tree creation.

To compare the phylogenetic trees of the arabinose genes with the core genome tree, branch score (Kuhner & Felsenstein, 1994) values were calculated for each gene phylogeny when compared to the core phylogeny.

CHAPTER 3

3.1. Introduction

Changes to regulatory sequences can have a dramatic impact on organismal evolution, comparable or even exceeding the innovations achievable through mutations in coding sequences (King & Wilson, 1975; Wray, 2007). In fact, homologous coding sequences are often highly conserved or even identical, while their respective regulatory sequences often show substantial levels of variation (Carroll, 2008; Joshi *et al.*, 2021). *Cis*-regulatory elements (CREs) are among the most significant contributors to this regulatory divergence (Osada *et al.*, 2017; Wittkopp & Kalay, 2012). Promoters and enhancers are two of the most well studied CREs, with promoters being found in close proximity to the transcriptional start site while enhancers are often located further upstream or downstream. The primary source of regulatory divergence are enhancers (Brown & Feder, 2005; Lewis *et al.*, 2019; Wittkopp & Kalay, 2012; Wray, 2007), as they are more likely to acquire mutations compared to promoters (Naidoo *et al.*, 2018). Promoters bind to a group of highly conserved, global regulatory molecules including transcription factors and RNA polymerase, thereby having less freedom for evolutionary innovation (R. P. Brown & Feder, 2005). However, promoters can still play a significant role in evolution as mutations within these regions directly influence binding specificity and hence the expression levels of the associated gene(s) (Hammarlöf *et al.*, 2018; Islam *et al.*, 2011; Jacob & Monod, 1961). Gene expression has been shown to be a major driving force in evolution and so mutations in promoters can have marked effects on the evolutionary trajectory of an organism (Friedensohn & Sawarkar, 2014).

The non-additive effects of double mutants are defined as epistatic interactions. Epistasis is the phenomenon by which the fitness effect of a mutation is dependent on the genetic background in which it is present (Phillips, 2008). This definition can be applied to interactions between point mutations or interactions between individual mutations on varying genetic backgrounds; this study focuses on the former. Studying pairwise interactions between mutations allows for a more precise estimate of epistatic interactions that can be quantitatively measured (de Visser & Krug, 2014). Due to the effect of epistasis on phenotypic outcome and therefore fitness, epistasis can shape fitness landscapes. Negative or positive epistasis leads to single fitness peaks achievable by mutations independent of one another, whilst sign epistasis can limit evolutionary pathways. In contrast,

reciprocal sign epistasis creates multiple fitness peaks allowing for sub-optimal genotypes to become fixed in a population (Poelwijk *et al.*, 2011; Weinreich *et al.*, 2005). Epistasis can determine the evolutionary pathways that are available to genetic elements such as CREs by creating either a 'rugged' epistatic landscape, where only specific mutational pathways lead to the global fitness peak and local optima can cause populations to become 'stranded' at a sub-optimal genotype. Alternatively, epistasis can create a 'smooth' landscape where all beneficial mutations lead to the global fitness peak regardless of the order in which they are gained (de Visser & Krug, 2014; Wright, 1932).

Significant empirical evidence suggests that changes in regulatory regions can affect fitness thus impacting traits including morphology, physiology, and behaviour (Jiang *et al.*, 2019; Wen *et al.*, 2016; Wray, 2007; Young *et al.*, 2022; Zheng *et al.*, 2019). The impact of regulatory changes upon the binding specificity of various *trans*-acting factors has been studied experimentally *in vitro* (Geertz *et al.*, 2012; Maerkl & Quake, 2007) and the distribution of mutational effects have been examined *in vivo* (Brewster *et al.*, 2012; Kinney *et al.*, 2010; Patwardhan *et al.*, 2009; Sharon *et al.*, 2012). The aforementioned studies mainly focused on characterising the binding dynamics of CREs and their associated transcription factors (TFs) but did not study the pairwise nucleotide effects within CREs and how these nucleotide interactions affect expression. One investigation of the pairwise interactions between nucleotides within a mammalian CRE found that 86% of single nucleotide substitutions within the CRE had a significant impact on regulation, with double mutants showing nucleotide-specific interactions even when constituent mutations were located in separate TF binding sites (Kwasnieski *et al.*, 2012). Lagator *et al.* (2017) studied pairwise interactions in the lambda promoter and found that most pairwise interactions exhibited negative epistasis (Lagator, Paixão *et al.*, 2017). Subsequently, Lagator *et al.* (2017) found that double mutants caused more phenotypic variation than single mutants could achieve alone (Lagator, Sarikas *et al.*, 2017).

To date, studies involving epistasis in CREs have focused on characterising the effects of epistasis on gene expression levels (Lagator *et al.*, 2016; Lagator, Paixão *et al.*, 2017; Lagator, Sarikas *et al.*, 2017) but have not investigated whether such epistatic effects on expression have concomitant effects on organismal fitness. The evolutionary goal of a catabolic operon would be to have low levels of uninduced expression, limiting wasted resources in the absence of a given carbon source, and high levels of induced expression increasing metabolic output in the presence of a carbon source conferring increased growth and, by extension, fitness (Schleif, 2000). But post transcriptional effects can restrict the effect that expression has on fitness (Romeo *et al.*, 2013; Yang *et al.*, 2010).

Lagator *et al.* investigated epistasis within the *pBAD* CRE of the arabinose operon (Lagator *et al.*, 2016) (**Figure 3.1**) which is one of the most studied operons and the first example of positive regulation

found in bacteria (Hahn, 2014). The *pBAD* promoter was also the first example of DNA looping ever studied (Dunn *et al.*, 1984) and helped define the 'light switch' mechanism of regulation which is a simple yet effective method of ligand regulation (Saviola *et al.*, 1998).

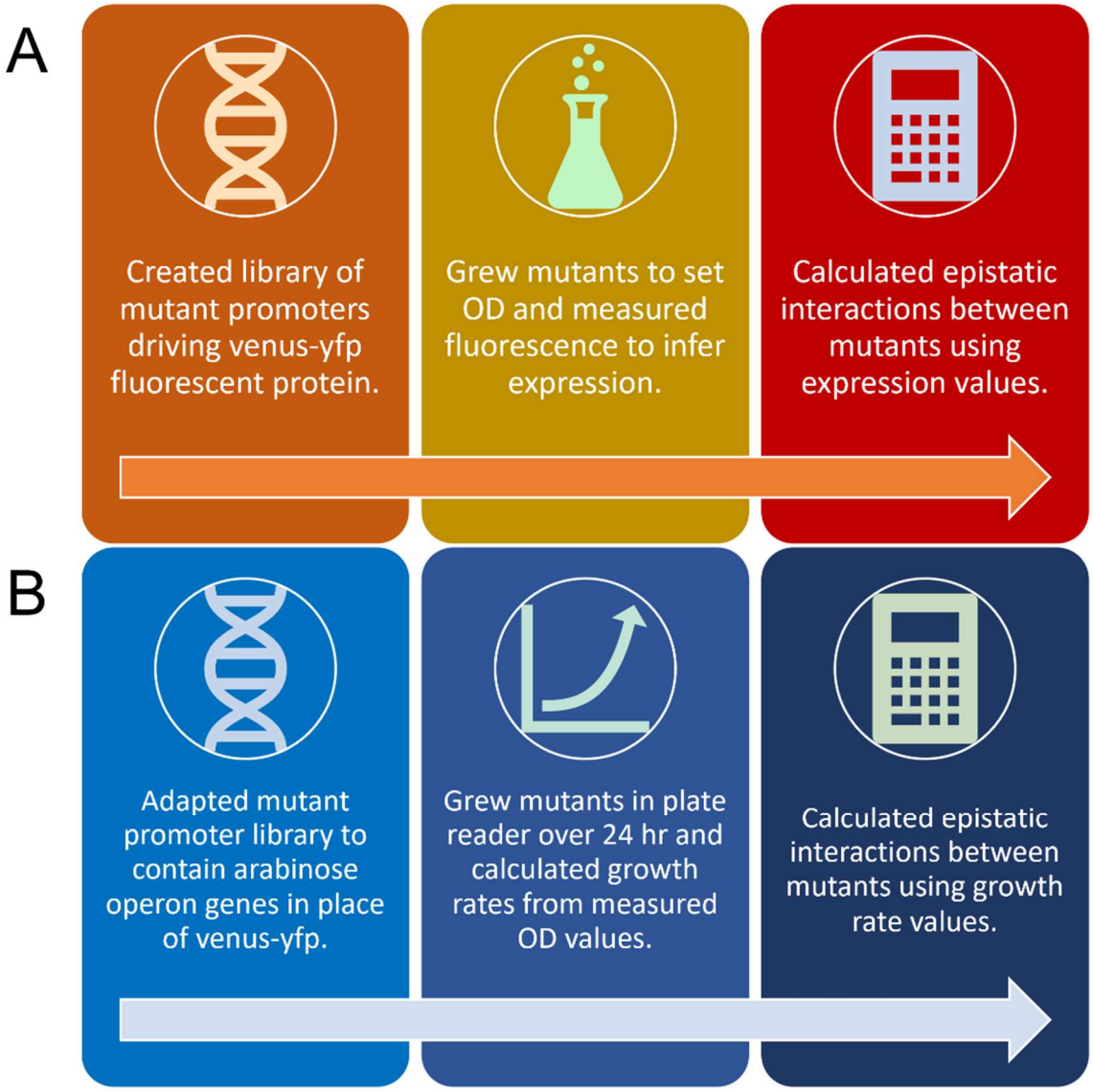


Figure 3.1 Workflow from A) Lagator et al. (2016) and B) This study. *Lagator et al. (2016) created mutant promoters driving venus-yfp and measured fluorescence levels to infer gene expression values. In this study the promoter was driving the arabinose operon genes (araBAD) and growth rates of the mutant strains were calculated as a representation of fitness. Lagator et al. calculated epistasis from expression values. This study used fitness values to calculate epistasis.*

They examined epistatic interactions by placing the arabinose operon promoter on a low copy plasmid that included the regulatory gene *araC* with the metabolic operon genes (*araB*, *araA* and *araD*) replaced with a *venus yfp* fluorescent reporter gene. A single mutant library that spanned both the $araI_1$ and $araI_2$ promoters was generated and double mutants were created through random combinations (Lagator *et al.*, 2016) of the single mutant library. Single mutants were selected based on work from Niland *et al.* which reported that mutations at the given sites resulted in at least a 10-fold reduction in AraC binding, as both $araI_1$ and $araI_2$ bind AraC in either the presence or absence of arabinose (Niland *et al.*, 1996). Expression values were estimated by normalising the fluorescence measurements of mutants against the wild type to generate values that were then used to calculate epistatic interactions based on their divergence from the expected expression of the double mutant. In this sense, expression is equivalent to the amount of fluorescent protein within the cell, this is therefore not necessarily reflective of direct levels of transcript but is affected by post-transcriptional effects. Significant negative epistatic interactions were found for half (10/20) of the double mutants (**Figure 3.2 (b)**), and significant positive epistasis found in three double mutants; one in the presence of arabinose (mutant positions '12,14' in $araI_1$ and $araI_2$, respectively) and two in the absence of arabinose (mutant positions '1,6' and '4,10' in $araI_1$ and $araI_2$, respectively). 60% of double mutants exhibited sign epistasis where a constituent mutation changed sign from positive to negative. For example, mutant 4 in the presence of arabinose has increased expression but double mutant '4,10' has a decreased expression value, changing the sign from positive to negative (**Figure 3.2)**.

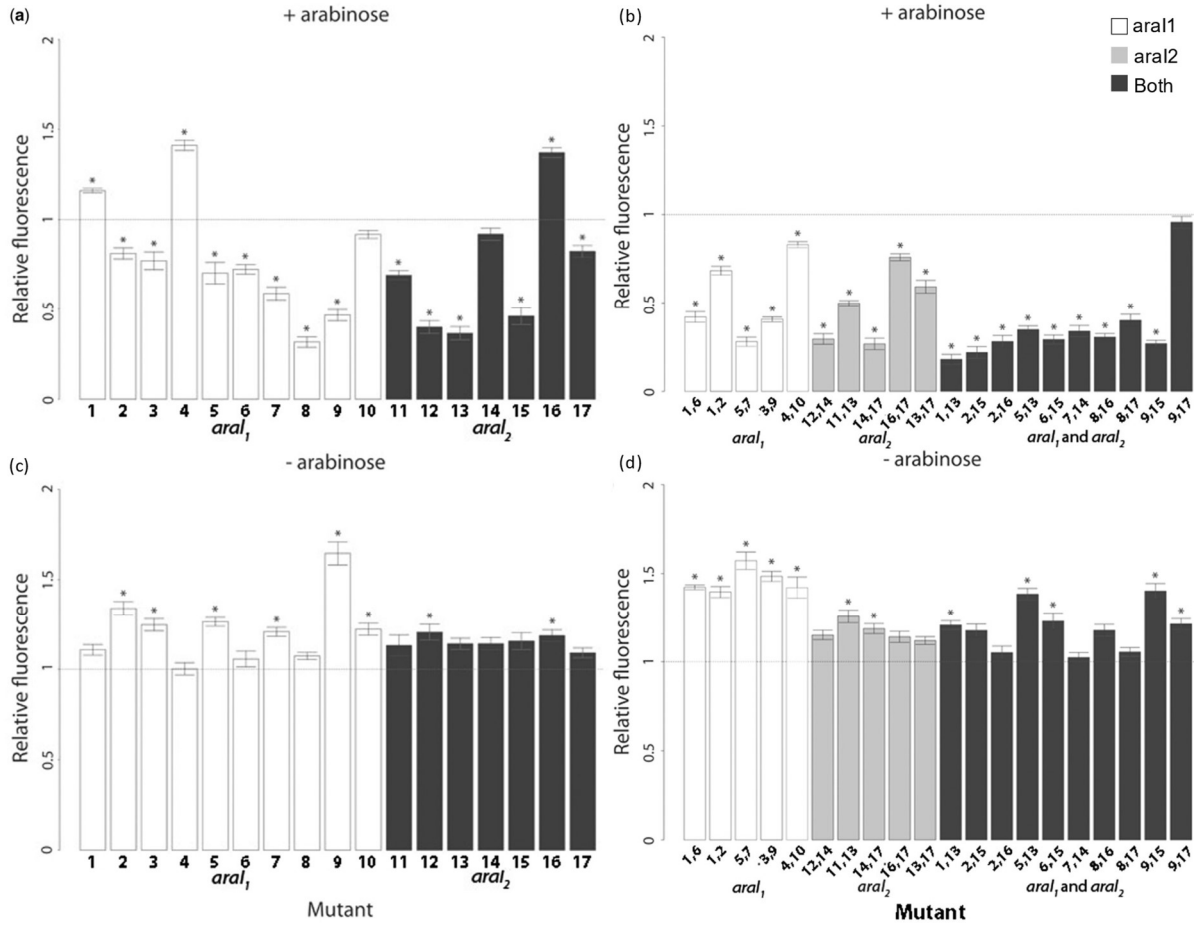


Figure 3.2 Relative Expression of mutants in different environments. *a) Relative fluorescence of single mutants in the presence of arabinose. b) Relative fluorescence of double mutants in the presence of arabinose. c) Relative fluorescence of single mutants in the absence of arabinose. d) Relative fluorescence of double mutants in the absence of arabinose. Fluorescence was measured to represent expression levels. All values are relative to the wild type value which is indicated by the horizontal line at a value of 1. Asterisks indicate significant difference from 1. Figures modified from previously published data (Lagator et al., 2016).*

These results indicate that the epistatic landscape for the *pBAD* CRE is fairly smooth showing mostly negative epistasis in both the presence and absence of arabinose. Whilst some sign epistasis was present, which can restrict evolutionary pathways, no reciprocal sign epistasis was found resulting in only one global fitness peak (Poelwijk *et al.*, 2011). Lagator *et al.* (2016) concluded that these fitness landscapes could be the result of the competing goals of optimal expression in the presence of arabinose versus tight repression in its absence with mutations either increasing or decreasing binding affinity therefore both goals are in contradiction.

The results from Lagator *et al.* (2016) are useful in elucidating the evolutionary landscape of bacterial promoters and are key to understanding the role of epistasis in microbial evolution. However, epistasis is often multifactorial and can be influenced by several factors beyond the raw expression values shown by the mutant promoters. The experimental system set up by Lagator *et al.* (2016) had significant limitations as it measured the genotype – phenotype interaction of the *pBAD* promoter using a fast-maturing yellow fluorescent protein - Venus (Nagai *et al.*, 2002), which only reflected the immediate phenotypic effect on gene expression. This system was tightly controlled and so could not detect any post-transcriptional effects. Furthermore, changes in gene expression do not necessarily result in changes in fitness (Signor & Nuzhdin, 2018). It has been demonstrated that, at high growth rates, ribosome synthesis is the rate limiting step for growth (Gourse *et al.*, 1996; Nomura, 1999) and so greater expression of metabolic genes may not cause a proportional increase in growth rate and therefore not produce greater fitness. This may result in lower penetrance of epistasis.

In natural environments, promoters experience selection pressure from several different sources and the fitness landscape may be affected by higher order epistasis (Weinreich *et al.*, 2018) when the promoters are driving the canonical metabolic genes (for example *araBAD*). Therefore, whilst absolute expression data is useful in showing epistatic effects on base promoter function, it does not necessarily reflect the effect on fitness that would be realised in nature. For example, sugar catabolism can have a significant effect on cellular response and the presence of metabolic genes can influence the fluorescence response of a reporter gene (Afroz *et al.*, 2014). Consequently, the epistatic landscape may differ when promoter mutations are investigated in the context of the complete operon background and the wider fitness of the cell is taken into consideration.

2.1.1. Aims

I wanted to discover if epistatic effects on the expression levels of the *pBAD* promoter (measured as total levels of fluorescent protein and therefore affected by post-transcriptional effects) were reflected when measuring a proxy for cell fitness directly; the ability of a cell to grow on arabinose as the sole carbon source. To investigate this, I introduced the catabolic *araBAD* genes into the constructs

from Lagator *et al.* (2016) and measured growth rates on arabinose as a proxy for fitness to determine whether epistatic effects on the expression level would be reflected at the fitness level. Due to there being no direct competition interaction between strains or any obvious way to differentiate strains, competition assays were deemed unnecessary (Hibbing *et al.*, 2010; Ram *et al.*, 2019; Wiser & Lenski, 2015).

3.2. Results

To investigate the effects of epistasis on expression versus fitness it was important to first understand the effects of the individual mutations on fitness alone. Comparing the mutational effects between this study and published expression data (Lagator *et al.*, 2016) made it possible to determine whether the mutations were having similar effects on fitness as they were on expression. Epistasis was then calculated from fitness data to compare with epistasis values based on expression. Fitness here is the ability of the cell to grow on arabinose as a sole carbon source and expression is the relative fluorescence level of Venus-yfp.

Information about the mutants used in this study is outlined below (**Table 3.1**). Mutants were labelled in numerical order to simplify the comparison between fitness and expression datasets. The genetic construct within the pZS*2 plasmid created for this study is shown in **Figure 3.3**.

Table 3.1 Mutant information. *Underlined regions in the sequence column indicate araI$_1$ and araI$_2$, respectively (see* ***Figure 1.3****). Coloured nucleotides show the individual mutations present in each strain. A= red, C= blue, T= green and G= yellow.*

Mutant	Genotype	Mutant Type	Operator	Sequence
M1	1	Single	I1	CCAGAGCATTTTTATCCATAAGATTAGCGGATCCTACCTGAC
M2	2	Single	I1	CCATACCATTTTTATCCATAAGATTAGCGGATCCTACCTGAC
M3	3	Single	I1	CCATAGGATTTTTATCCATAAGATTAGCGGATCCTACCTGAC
M4	4	Single	I1	CCATAGCCTTTTTATCCATAAGATTAGCGGATCCTACCTGAC
M5	5	Single	I1	CCATAGCATTTTTAGCCATAAGATTAGCGGATCCTACCTGAC
M6	6	Single	I1	CCATAGCATTTTTATTCATAAGATTAGCGGATCCTACCTGAC
M7	7	Single	I1	CCATAGCATTTTTATCAATAAGATTAGCGGATCCTACCTGAC
M8	8	Single	I1	CCATAGCATTTTTATCCGTAAGATTAGCGGATCCTACCTGAC
M9	9	Single	I1	CCATAGCATTTTTATCCAAAAGATTAGCGGATCCTACCTGAC
M10	10	Single	I1	CCATAGCATTTTTATCCATTAGATTAGCGGATCCTACCTGAC
M11	11	Single	I2	CCATAGCATTTTTATCCATAAGATGAGCGGATCCTACCTGAC
M12	12	Single	I2	CCATAGCATTTTTATCCATAAGATTCGCGGATCCTACCTGAC
M13	13	Single	I2	CCATAGCATTTTTATCCATAAGATTACCGGATCCTACCTGAC
M14	14	Single	I2	CCATAGCATTTTTATCCATAAGATTAGTGGATCCTACCTGAC
M15	15	Single	I2	CCATAGCATTTTTATCCATAAGATTAGCGGATCCTAGCTGAC
M16	16	Single	I2	CCATAGCATTTTTATCCATAAGATTAGCGGATCCTACATGAC
M17	17	Single	I2	CCATAGCATTTTTATCCATAAGATTAGCGGATCCTACCTGTC
M18	1,6	Double	I1	CCAGAGCATTTTTATTCATAAGATTAGCGGATCCTACCTGAC
M19	1,2	Double	I1	CCAGACCATTTTTATCCATAAGATTAGCGGATCCTACCTGAC
M20	5,7	Double	I1	CCATAGCATTTTTAGCAATAAGATTAGCGGATCCTACCTGAC
M21	3,9	Double	I1	CCATAGGATTTTTATCCAAAAGATTAGCGGATCCTACCTGAC
M22	4,10	Double	I1	CCATAGCCTTTTTATCCATTAGATTAGCGGATCCTACCTGAC
M23	12,14	Double	I2	CCATAGCATTTTTATCCATAAGATTCGTGGATCCTACCTGAC
M24	11,13	Double	I2	CCATAGCATTTTTATCCATAAGATGACCGGATCCTACCTGAC
M25	14,17	Double	I2	CCATAGCATTTTTATCCATAAGATTAGTGGATCCTACCTGTC
M27	13,17	Double	I2	CCATAGCATTTTTATCCATAAGATTACCGGATCCTACCTGTC
M28	1,13	Double	I1 and I2	CCAGAGCATTTTTATCCATAAGATTACCGGATCCTACCTGAC
M29	2,15	Double	I1 and I2	CCATACCATTTTTATCCATAAGATTAGCGGATCCTAGCTGAC
M30	2,16	Double	I1 and I2	CCATACCATTTTTATCCATAAGATTAGCGGATCCTACATGAC
M31	5,13	Double	I1 and I2	CCATAGCATTTTTAGCCATAAGATTACCGGATCCTACCTGAC
M32	6,15	Double	I1 and I2	CCATAGCATTTTTATTCATAAGATTAGCGGATCCTAGCTGAC

M33	7,14	Double	I1 and I2	CCATAGCATTTTTATCAATAAGATTAGTGGATCCTACCTGAC
M34	8,16	Double	I1 and I2	CCATAGCATTTTTATCCGTAAGATTAGCGGATCCTACATGAC
M35	8,17	Double	I1 and I2	CCATAGCATTTTTATCCGTAAGATTAGCGGATCCTACCTGTC
M36	9,15	Double	I1 and I2	CCATAGCATTTTTATCCAAAAGATTAGCGGATCCTAGCTGAC
M37	9,17	Double	I1 and I2	CCATAGCATTTTTATCCAAAAGATTAGCGGATCCTACCTGTC

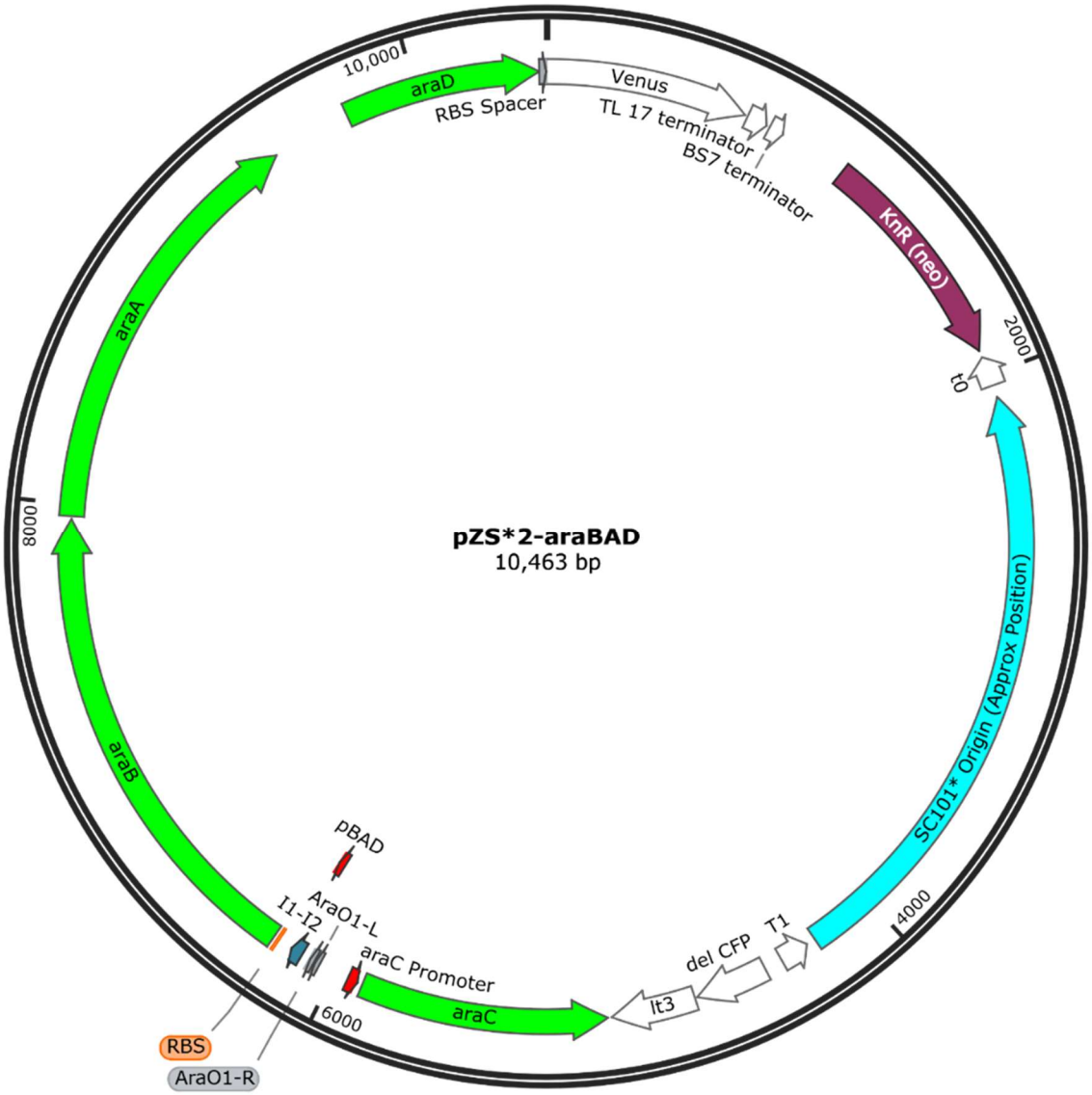


***Figure 3.3* Plasmid construct pZS*2-araBAD created using NEBuilder® HiFi DNA Assembly Kit.** *Each mutant strain contains mutations in the pBAD promoter (**Table 3.1**) show in red. araC promoter is also shown in red. Arabinose operon genes are shown in green. Kanamycin resistance gene is shown in maroon. The origin of replication is shown in blue.*

3.2.1. Relative Fitness of Single Mutants

'Fitness' of double mutants is the product of epistatic interactions between constituent single mutants. To measure the magnitude of epistasis between mutants a relative fitness value was calculated using growth rate data. For expression data, the relative fluorescence value from (Lagator *et al.*, 2016) was used. For this study, relative growth rate was used. Growth rate was selected as the measurement for fitness as exponential growth rate is a reliable measure of the ability of a cell to grow on a particular carbon source (Schaechter *et al.*, 1958; Wang *et al.*, 2019). Examples of the growth curves observed for wild type and mutant strains are shown below (**Figure 3.4**). The strains were grown in M9 minimal media with arabinose as the sole carbon source to ensure that growth was driven by arabinose catabolism. The resulting fitness distributions of single mutants are shown in **Figure 3.5**.

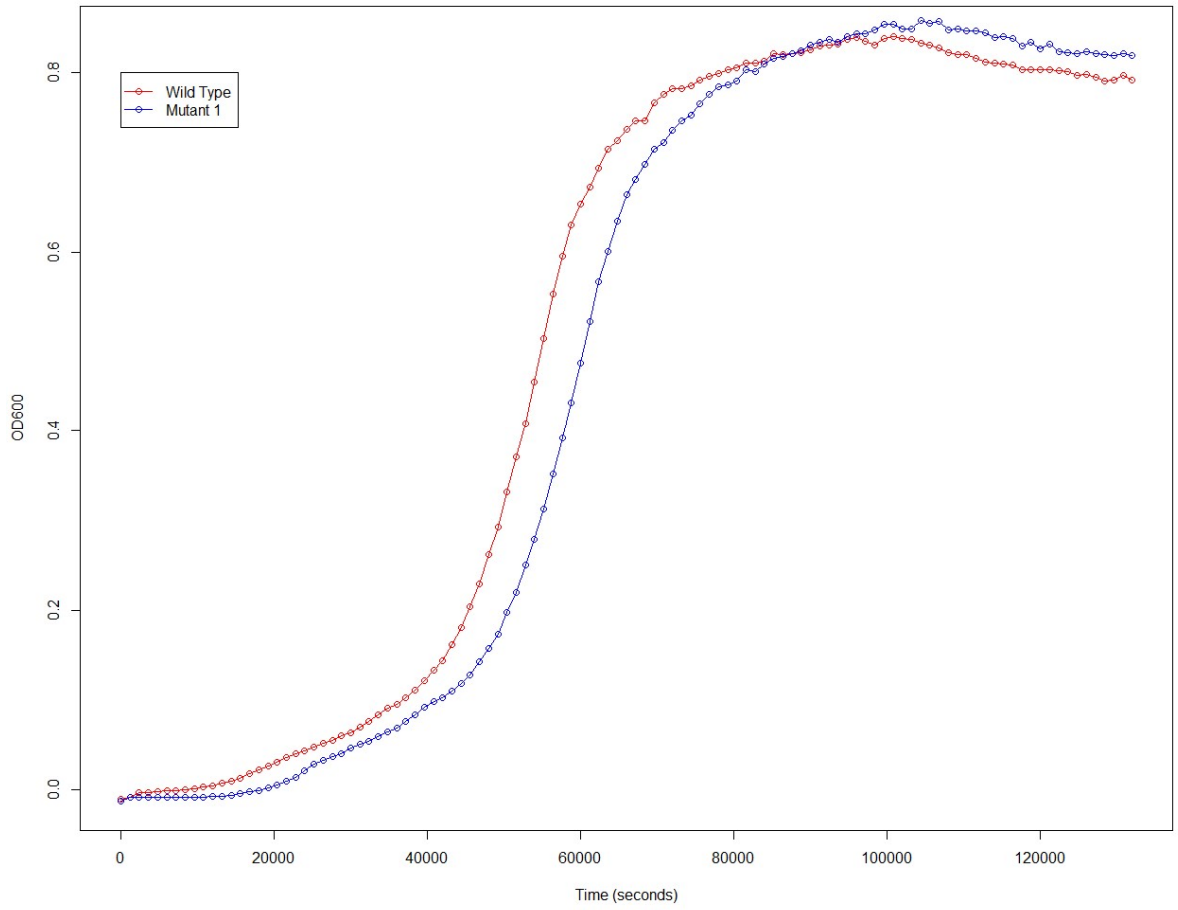


Figure 3.4 Example growth curves from wild-type and Mutant 1 strains. *Wild-type OD600 values are plotted in red, while the Mutant 1 (Table 3.1) data are plotted in blue.*

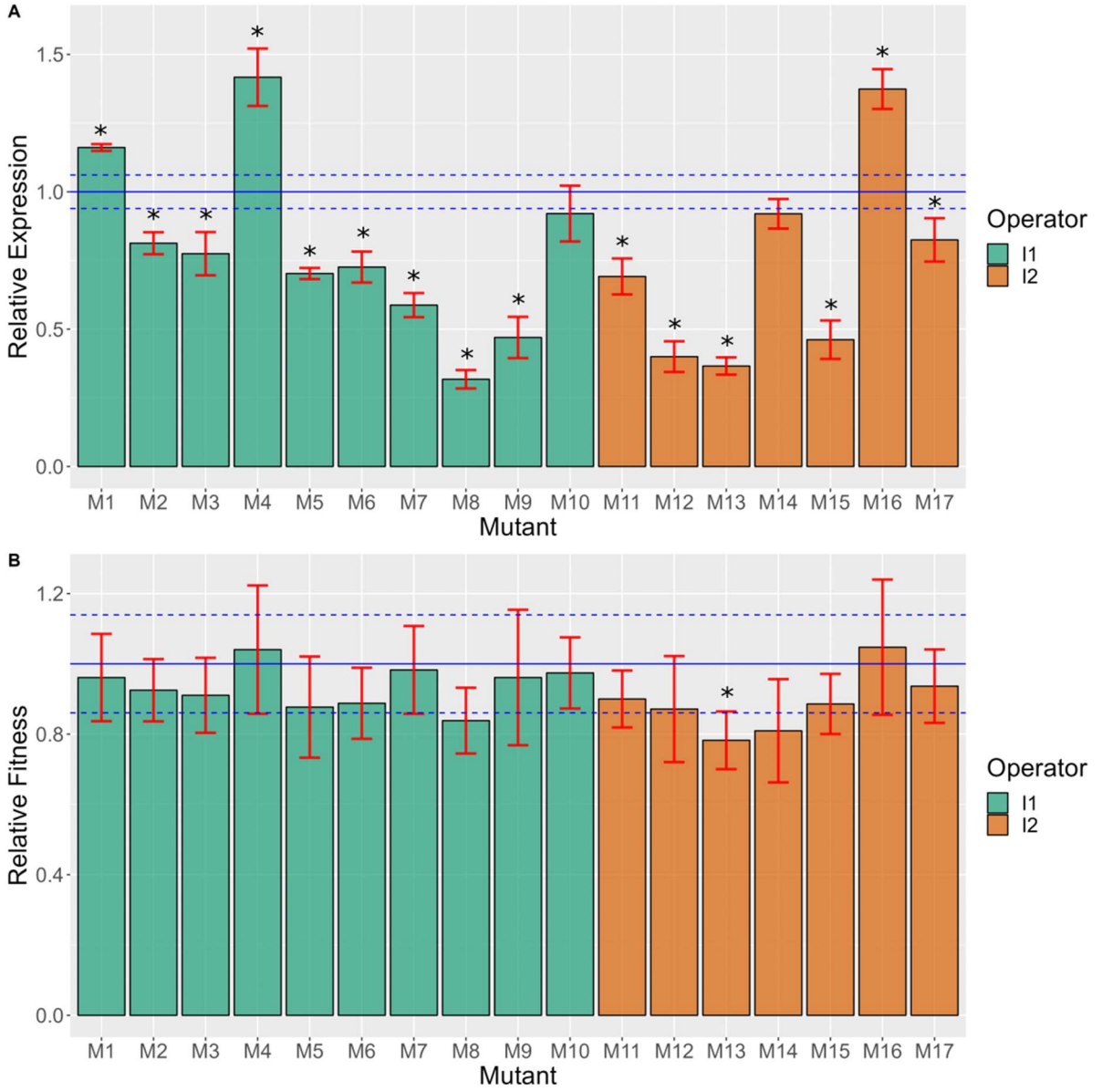


Figure 3.5 Relative fitness of single mutants varies less from wild type value than relative gene expression. *A) The relative fluorescence of pBAD CRE single mutants compared to the wild type strain indicated with the blue line. Data from (Lagator et al., 2016). Error bars represent standard deviation. B) Relative growth rates of pBAD CRE single mutants compared to the wild type (methods and media outlined in Section 2.1.4.) indicated with the blue line. Error bars represent standard deviation. Asterisks signify significant difference from wild type value ($p < 0.05$).*

When measuring relative fluorescence, 15 out of 17 mutants were found to be significantly different from the 'wild type'. Of those 15 mutants, 12 had significantly lower fluorescence than the wild type. Mutants 1, 4 and 16 showed higher fluorescence than the wild type suggesting the mutations within the promoter led to higher expression. When observing growth rates, most mutants did not significantly differ from the wild type, with the exception of mutant 13. The comparison between the two datasets (**Figure 3.5**) indicates that single mutation effects are more pronounced when measuring absolute expression levels but are heavily masked when measuring a more general trait of cell fitness, such as growth rate. An ANOVA test was performed that identified statistically significant differences between both expression and fitness values of individual mutants (p-value <0.001), and between the expression and the fitness of mutants overall (p-value <0.001). The ANOVA also showed that there was a significant interaction between mutants and the variable being measured (expression vs growth rate) (p-value <0.001).

3.2.2. Relative Fitness of Double Mutants

To determine epistasis, the expression and fitness of double mutants needed to be measured to determine if the values differed from the additive expectation. When measuring expression of double mutants, 18 out of the 19 mutants (M30) had a statistically significant difference from the wild type (**Figure 3.6 (a)**), this is a similar proportion to that of the single mutants where 15 out of 17 mutants were significantly different (**Figure 3.5 (a)**). However, when measuring fitness of double mutants there were 10 of the 19 mutants showing a significant difference from the wild type (**Figure 3.6 (b)**). This finding was in marked contrast to the single mutant fitness data where only one mutant showed a significant difference (**Figure 3.5 (b)**).

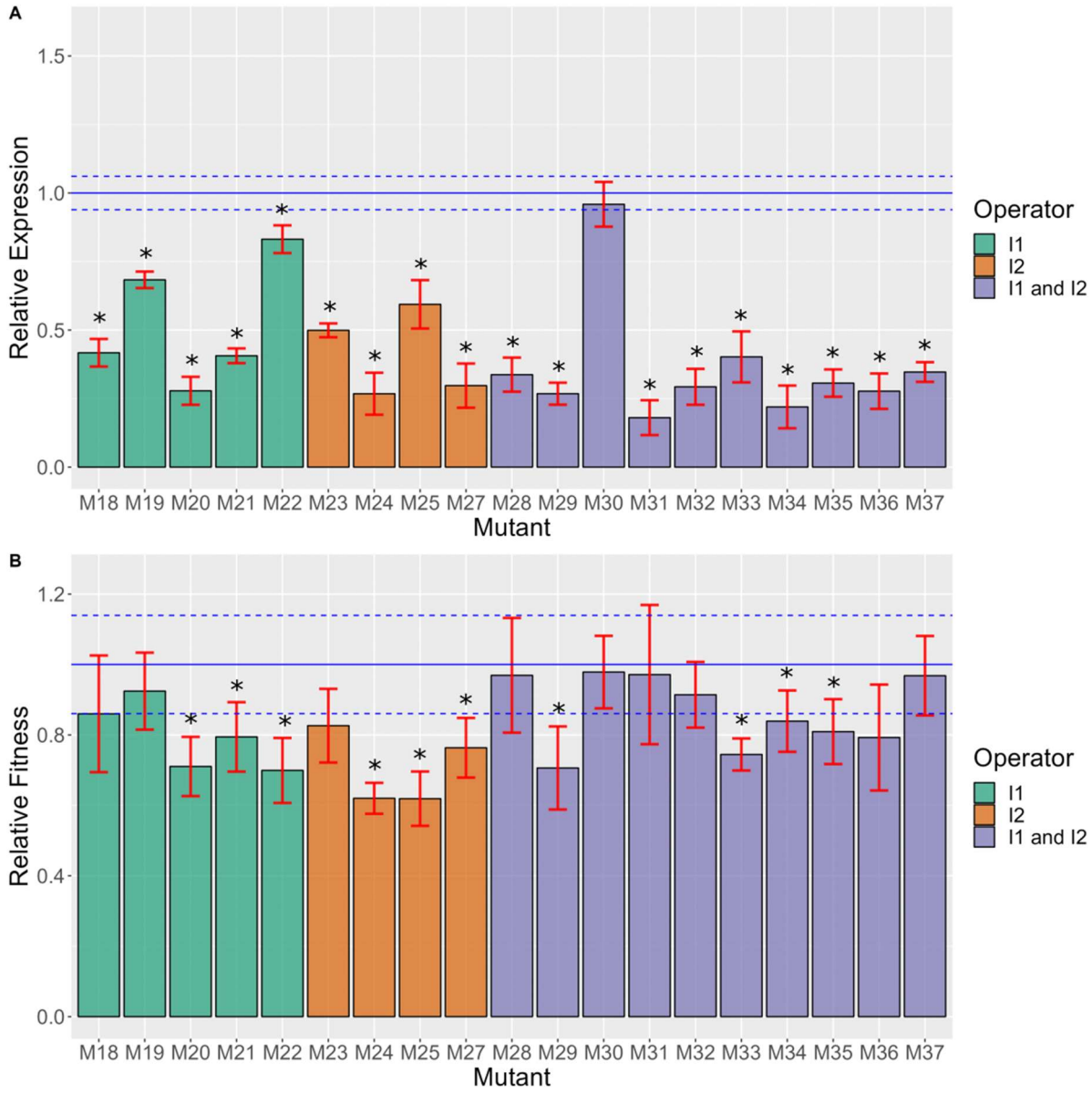


***Figure 3.6 Relative fitness of double mutants**. A) The relative expression of pBAD CRE double mutants compared to the wild type strain indicated with the blue line. Error bars represent standard deviation. B) Relative fitness of pBAD CRE double mutants compared to the wild type indicated with the blue line. Error bars represent standard deviation. Asterisks signify significant difference from wild type value ($p < 0.05$). Methods and media outlined in (Section 2.1.4.).*

Although the fitness values of double mutants seem to differ less from the expression values than in the case of single mutants, there are still some which are notably different. For example, mutant 31 shows a dramatic difference between fitness and gene expression and indicates the magnitude of the differences that can be observed when comparing fluorescence to growth rate (**Figure 3.6**). On the contrary, mutant 30 showed almost no difference relative to the wild type in both studies. While 18 mutants were significantly different from the wild type when measuring expression, it is important to note that eight of these show no significant difference when measuring growth rates. This means several mutants are no longer significantly different from the wild type when measuring growth rate, further emphasising the difference in the effects of mutations on fitness versus expression.

3.2.3. Positional Effects

One of the interesting factors to consider with the above data is whether the location of the mutations influence the differences we see between measuring expression versus fitness. Considering whether *araI*$_1$ or *araI*$_2$ have stronger mutational effects, or indeed whether mutants containing mutations in both sites show larger changes, is important for understanding the epistatic landscape of the *pBAD* promoter. Fluorescence values were plotted against growth rate fitness and a Pearson correlation test produced a correlation of 0.49 (p=0.002) (**Figure 3.7**).

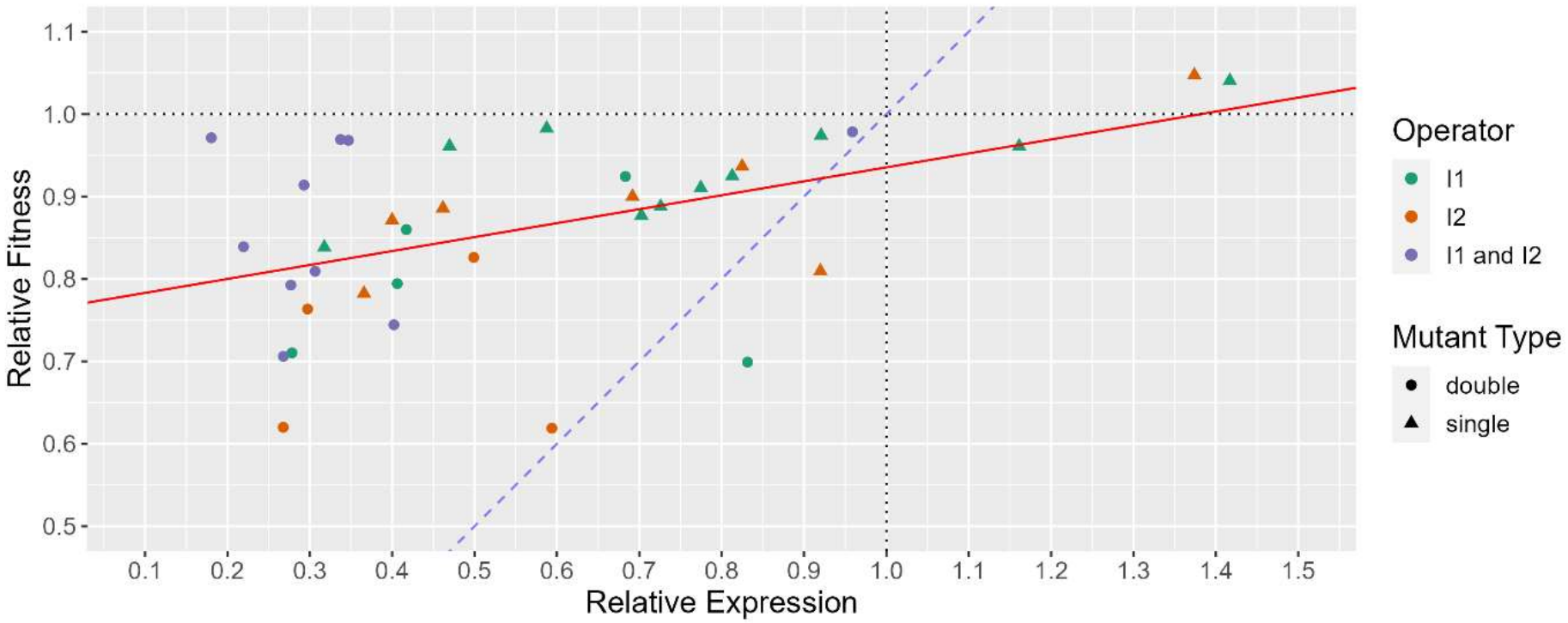


Figure 3.7 Fitness of mutants is weakly correlated with expression values. *The blue dashed line represents a theoretical correlation of 1. The black dotted lines represent the respective wild type values for each measurement. Individual data points represent mean values for mutants and are coloured based on their operator location. Shapes represent single or double mutants. Red line shows the correlation of the data points (Pearson correlation = 0.49). Expression data from (Lagator et al., 2016). Fitness data from (Section 2.1.4.).*

Figure 3.7 plots all 37 mutants from **Table 3.1** and shows that double mutants with mutations in both operators (purple dots) were clustered together except for a single mutant (M30). These mutants are seen to have lower expression with respect to their fitness. This trend is also echoed in **Figure 3.6**. Another point of interest is that out of the five mutants whose expression values are greater than their fitness (below the blue dashed line), four are single mutants with two each in operator $araI_1$ and $araI_2$.

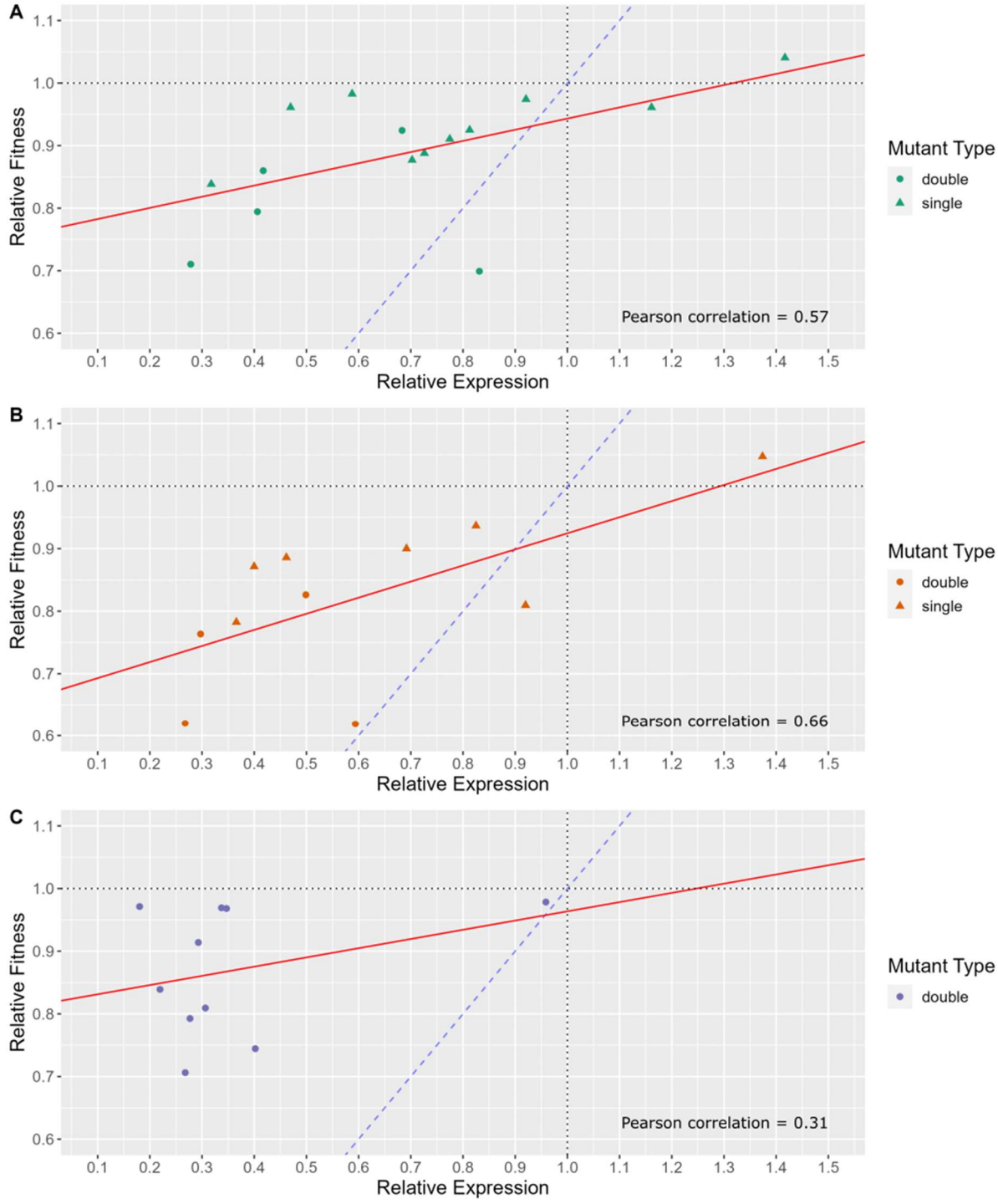


Figure 3.8 Fitness of mutants with mutations in both operator sites have a weaker positive correlation with expression than mutants located within a single site. *A) Operator araI1. B) Operator araI2. C) Operator araI1 and araI2. Blue dotted lines represent theoretical correlations of 1. Red lines represent linear regression of data points. (Pearson correlation = A) r = 0.57, B) r = 0.66; and, C) r = 0.31).*

Figure 3.8 plots mutants on separate graphs based on their operator locations outlined in **Table 3.1**. These figures show that mutants within either $araI_1$ or $araI_2$ had a positive pearson correlation (any value greater than 0) between expression and fitness with similar values of 0.57 and 0.66 respectively. However, mutants with mutations in both operators had a weaker positive correlation of 0.31. This suggests that the mutants spanning both operators show a greater disparity between expression and fitness values.

Table 3.2 Mutants with mutation in both operator sites have a larger proportion of significant differences between their expression and fitness.

Operator	Significant Changes	Non-significant changes
araI1	7	8
araI2	7	4
araI1 **and** ***araI2***	9	1

A Fisher's exact test was performed on the number of mutants with significant differences between expression and fitness in each of the operators and those within both (**Table 3.2**). It was observed that the location of mutations does not significantly affect the likelihood of having significant differences between expression and fitness measurements ($p > 0.05$).

3.2.4. Epistasis

Epistasis values were calculated (**Section 2.1.6.**) from fitness data and compared against epistasis values from expression data from (Lagator *et al.*, 2016) (**Figure 3.9**) to determine if epistasis was consistent between these measurements. A noticeable trend observed was the error values for growth rate epistasis which were larger on average than the expression phenotype epistasis values. Consequently, there were fewer significant epistasis values from the growth rate data compared to data from Lagator *et al.* (2016). As epistasis only results from double mutations, mutants 18-37 (**Table 3.1**) were analysed (**Figure 3.9**).

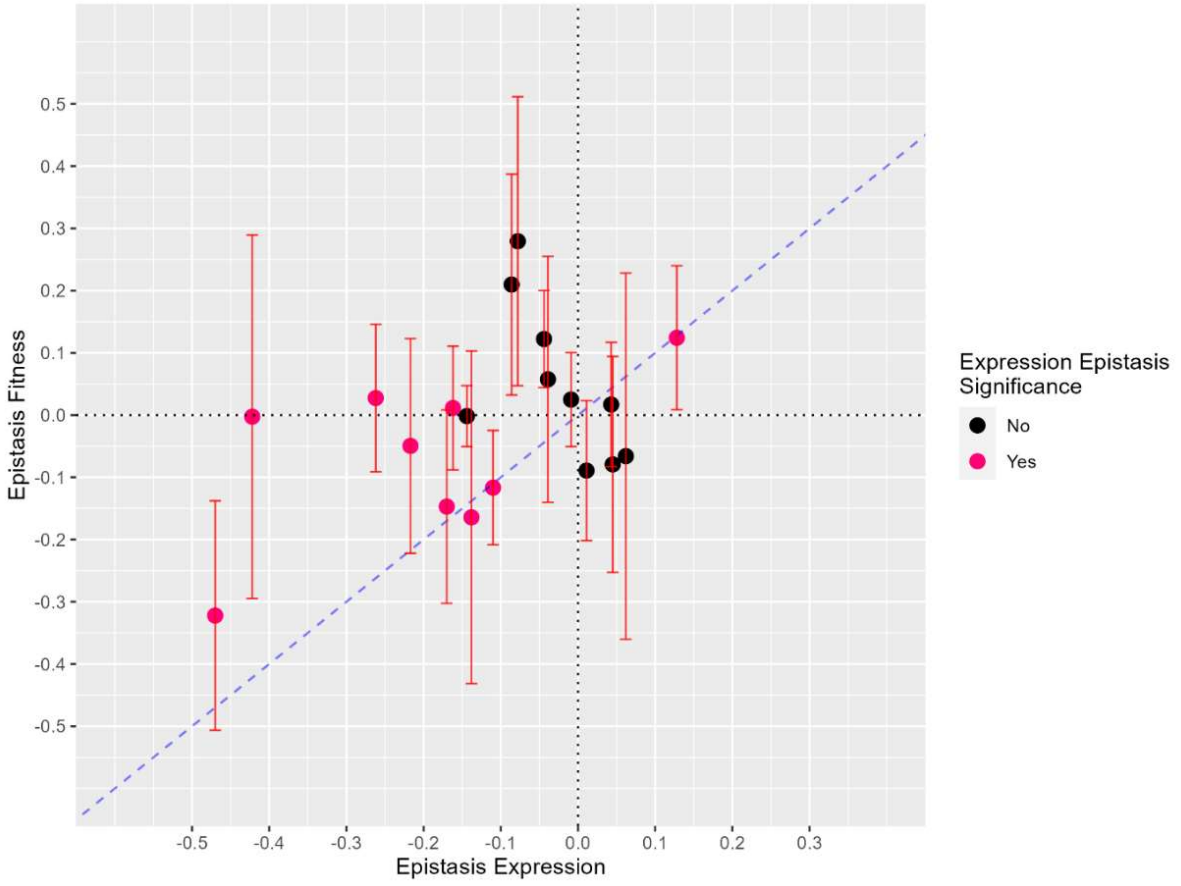


Figure 3.9 Fitness data does not show the same pattern of epistasis as gene expression data. *Points represent mean epistasis values of double mutants (18-37). Epistasis values were calculated from each replicate plate and then averaged. Error bars are standard deviation. Dotted lines represent an epistasis value of zero on the respective axes. Blue dashed line indicates correlation of 1, signifying if epistasis is consistent between expression and fitness. Formula for epistasis calculation is* $\varepsilon = \omega_{m12} - \omega_{m1} \times \omega_{m2}$ *where* ω_{m12} *is the fitness value of the double mutant and* ω_{m1}/ω_{m2} *are the fitness values of the respective constituent single mutants.*

Epistasis values from fitness data were not found to be significantly different from zero meaning no epistasis was detected using fitness data. This suggests the approach used to measure growth rates may not have been sensitive enough to detect the changes in fitness required to calculate epistasis.

3.3. Discussion

In this study, I aimed to build upon the work from Lagator *et al.* (2016) which decoupled expression from post-translational fitness effects to ascertain the epistatic landscape of the expression levels of the *pBAD* promoter. For this, the arabinose operon genes were cloned into the construct under the control of the mutant promoters, in order to determine the extent of differences in epistasis between expression and more general fitness. I found that, whilst many significant epistatic interactions were seen when measuring expression alone, those same interactions were not preserved when measuring fitness using growth rates. These findings suggest that epistatic effects on phenotypes such as expression do not translate into effects on organismal fitness (**Figure 3.9)** when using growth rate to assess fitness.

3.3.1. Fitness Differences

3.3.1.1. Single Mutants

One of the most striking observations when measuring fitness compared to expression was that out of the 17 single mutants, only one was found to have significantly different fitness than the wild type, whereas 15 mutants had significantly different expression than the wild type. This indicates that single mutation effects are heavily masked by external factors beyond expression levels as the relative fitness of the cell is not compromised as much as would be expected. The fact that the expression levels of mutants showed significant differences from the wild type suggests that the difference does not lie at the transcriptional level, but the difference may lie at the translational or post-translational level, where higher or lower expression of a protein does not necessarily translate to changes in metabolic rate (Liu *et al.*, 2016).

It has been shown that expression levels are subject to strong stabilising selection in *Caenorhabditis elegans* (Denver *et al.*, 2005) which suggests natural selection favours expression levels surrounding the average level rather than those at extremes. This phenomenon could explain why there are no mutants with significantly positive fitness as, if the high level of expression does not result in a marked increase in fitness, it would be seen as a waste of energy and resources and therefore be selected against. One possible reason for increased expression not resulting in an equivalent increase in fitness is the heterogeneity of metabolism at the population level. It has been shown that at certain concentrations of inducer (arabinose in this case) some cells are induced to metabolise a carbon

source whilst others do not reach a threshold for induction. This is due to the exponential nature of positive feedback and has been coined 'multistationarity' (Novick & Weiner, 1957; Smits *et al.*, 2006). If not all the individuals in a population are induced to metabolise a carbon source, then the increase in expression levels may not be proportional to an increase in growth rate or 'fitness'.

Another possible contributing factor is that a major constraint on metabolism can be intracellular crowding, where cytoplasmic space limits maximum flux attainable by the cell, which could suggest that there is a post transcriptional bottleneck in the ability of the cell to convert a carbon source into growth potential (Beg *et al.*, 2007). Another example of a possible bottleneck in the system is the rate of uptake of a carbon source from the environment, regulated by the *araFGH* operon (Kolodrubetz & Schleif, 1981). Higher expression levels will not result in increased growth if there is an excess of metabolic enzyme and not enough substrate to saturate the system.

3.3.1.2. Double Mutants

Out of the double mutants, 10 retained a significant difference from the wild type in the fitness data compared to 18 in the expression data. This is a marked increase from the single mutants, suggesting that the combinatoric effects on expression of mutations in the *pBAD* CRE are strong enough to be present in the measurement of fitness, although to a significantly lower effect (**Figure 3.5**). A plausible explanation for this could be that the masking effects on single mutations are similarly affecting the double mutants but because they had stronger effects on expression to begin with, they still differ significantly from the wild type. This aligns with the fact that Lagator *et al.* (2016) found strong negative epistatic effects between most mutations and so the non-additive effects of double mutants are more resistant to masking than the single mutants.

3.3.2. Epistatic Effects

Statistical tests showed that there was no significant epistasis found in any of the double mutants in the fitness data and this may be a limitation of the study (**Section 3.4.**). This is an interesting finding as it suggests that epistatic interactions between mutations are lost when measuring fitness versus expression which implies the fitness landscape of the *pBAD* CRE is smoother than expected when considering what natural selection can act upon. Something to consider is the fact that all mutations having fewer extreme effects in the growth rate data likely means that the combinatoric effects would also be milder and so harder to detect statistically. This finding is important when considering what evolutionary pathways are available to promoters as mutations that would initially seem inaccessible when considering expression data may be accessible in the wild as the landscape is made less restrictive by incomplete phenotypic penetrance.

One concept that could explain the differences we see between expression data and fitness data is that, when measuring expression, the epistatic interactions are limited to those of the two interacting mutations. That is to say, no external factors are affecting the phenotype in question; *yfp* fluorescence. However, when measuring growth rate, we introduce the potential of wider physiological effects as lower expression of metabolic enzymes does not necessarily mean decreased sugar catabolism. As mentioned previously, uptake of arabinose was not changed in the system and so if the uptake operon *araFGH* and unlinked *araE* gene (Macpherson *et al.*, 1981) were already a bottleneck to the metabolic flux then reducing the amount of *araBAD* enzymes may not affect total flux as drastically as expected. Metabolic enzymes are also heavily regulated to manage sugar catabolism (Reid & Abratt, 2005) and thus, if the *araBAD* enzymes were regulated to limit flux previously then lowering the expression of the enzymes may not have an effect if regulation is reduced to ameliorate the effect.

3.4. Limitations

Although this study was designed to be as robust as possible there are still some limitations to the study design. Firstly, although five repeat growth curves were assessed for each mutant, due to the natural variability of growth curves the standard deviation of the means was particularly large (Hall *et al.*, 2013). This unfortunately limited statistical power when assessing whether the mutants' relative growth rate was different from the wild type growth rate. This then had consequences for epistasis calculations as there were also large standard deviations which resulted in most epistasis values being non-significant. Performing several large-scale assays to increase the sample number would potentially reduce these standard deviations and allow for better statistical power. There is also the possibility that the approach of using growth rates to assess fitness was not sensitive enough to detect the changes in fitness and using a competition assay approach may yield more significant results.

Another point to consider is single time point fluorescence measurements and average growth rate measurements are not interchangeable metrics. The measuring of a single time point will not account for differences in lag phase duration and so may not be measuring the same relative point in the growth curve of a given strain. Following the previous point, measuring average growth rate accounts for any differences in lag phase or stationary phase but these data are informative in other ways (Rolfe *et al.*, 2012) which is not picked up when measuring average growth rate. Both methods have pros and cons and it is important to keep in mind that neither incorporates the full picture of microbial expression or growth, respectively.

The system that was designed for the experiment also had its own limitations. Due to the operon being present on a plasmid background, multiple copies of the operon would be present (3-4) (Lutz & Bujard, 1997) whereas on the native chromosome there would only be a single copy of the operon. There are

also differences between chromosomal expression and plasmid expression which are detailed in this review (Mairhofer *et al.*, 2013).

The use of ANOVA was also limited as the ANOVA analysed all possible comparisons within the dataset. This resulted in non-meaningful comparisons such as comparing the fluorescence of M1 to the growth rate of M2. This resulted in a significant 'difference' but the comparison is not scientifically meaningful.

3.5. Conclusions

Overall, these findings point to the fact that, when considering evolutionary pathways and fitness landscapes, there can be drastic differences when observing the difference between expression and fitness. Whilst expression epistatic landscapes may be less smooth with strong epistatic effects, fitness landscapes may be a lot smoother and therefore allow certain mutations to occur without the expected detriment to the cell. For this system, it implies that epistatic effects can be ameliorated via buffering in other levels of cell metabolism. To determine which levels contribute to this buffering would be interesting for further work in this area.

These findings emphasize the fact that when thinking about evolution of microbes we must consider the contributions of all factors that affect cell metabolism and fitness and must not place too much emphasis on one contributing factor such as expression levels. Although it should be noted that this system may differ from other promoters and metabolic pathways so researching the landscapes of other regulatory systems would be an interesting way to build on this work. However, the key findings may reflect technical limitations associated with growth rate measurements and the equipment used. This important issue is addressed in **Chapter 4** and in the general discussion (**Chapter 6**).

CHAPTER 4

4.1. Introduction

Organisms evolving in the wild will likely be exposed to a multitude of environments across their lifespan. Whether this be changes in temperature, nutrient availability, or presence of competitors to name a few, the 'goalposts' of evolution are constantly being moved by a changing environment Lahti *et al.*, 2009; Siepielski *et al.*, 2009). Microbes such as *E. coli* experience this in their natural environment which can affect the molecular evolution of their genome (Blount *et al.*, 2020). It could therefore be expected that changing environments affect the evolution of regulatory sequences due to their significant impact on evolution (Wray, 2007). Lagator *et al.* (2016) suggested that the arabinose operon CRE experiences competing selective forces for higher expression in the presence of arabinose alongside tighter regulation in its absence (Lagator *et al.*, 2016). To understand the full extent of regulatory sequence evolution we must study these sequences under different environments to elucidate the evolutionary forces shaping them.

As the fitness effects of mutations are environment dependent, so may be the interactions between these fitness effects. Whilst it is widely known that the effects of mutations on the fitness of an organism are dependent on the environment the organism finds itself in, less thought has been given to the effects of environment on the epistatic interactions between mutations. Phenotype has long been studied as a product of gene x environment (G x E) interactions and gene x gene (G x G) interactions (Epistasis), but less attention has been given to the concept of gene x gene x environment (G x G x E) interactions (Domingo *et al.*, 2019). Given that fitness landscapes are used to visualise fitness of genotypes within the sequence space, and that fitness is a direct product of environment, there is evidence that fitness landscapes can vary between environments (Anderson *et al.*, 2021); yet most studies have only characterised fitness landscapes under single environments due to technical limitations (Li & Zhang, 2018). Therefore, to fully understand the fitness landscapes of genes and the potential evolutionary trajectories of sequences, epistasis interactions must be studied in multiple environments to reflect the ever-changing environment organisms find themselves in. One significant consequence of fitness landscapes varying between environments can be allowing populations stranded on sub-optimal local fitness peaks to migrate through fitness valleys due to a change in environment (de Vos *et al.*, 2015; Steinberg & Ostermeier, 2016). Recent studies have begun to explore how environments interact with epistasis (Lagator, Paixão, *et al.*, 2017; Lagator, Sarikas, *et al.*, 2017; Li & Zhang, 2018).

Previous studies have found environmental effects on epistasis within phage (You & Yin, 2002), yeast (Harrison *et al.*, 2007), *Arabidopsis Thaliana* (Kerwin *et al.*, 2017) and even insects (Arnqvist *et al.*, 2010). Environmental epistasis has also been evidenced in *E. coli* (Remold & Lenski, 2004) showing that mutation's effects on fitness were not only dependent on genetic background but also on which environment (maltose or glucose) the cells were grown in. One study also showed that the mutational pathways available to a metalloenzyme can vary depending on which metal ions are present in the environment. This had significant effects on the evolutionary pathways available to the enzyme causing it to become stranded on a suboptimal peak in certain environments, demonstrating the importance of environment on epistatic interactions (Anderson *et al.*, 2021).

The aims of this experiment were to test whether the epistatic interactions and 'landscape' of the *araBAD* CRE determined in **Chapter 3** would be affected by environmental change. *E. coli* is commonly found in the microbiome of mammals (Hartl & Dykhuizen, 1984) where environmental traits such as temperature are consistent. However, gut microbes are frequently excreted into the environment and must adapt to the external environment to last long enough to be ingested again (van Elsas *et al.*, 2011). The natural environment fluctuates drastically, changing conditions such as temperature and nutrient availability (Savageau, 1983), therefore experimental conditions were chosen to best represent the conditions that may fluctuate in the external environment of *E. coli*. Arabinose is a sugar present in many natural environments and has been identified as a key carbon source affecting microbial metabolism (Wang *et al.*, 2021). This means *E. coli* could encounter arabinose in many environments but most likely in soil around the rhizosphere (Habteselassie *et al.*, 2010). The environmental factors selected were therefore arabinose concentration and temperature, two conditions likely to fluctuate in the natural environment. To determine the influence of temperature on epistasis, experiments were done at both 30^{o}C and 37^{o}C, previous studies have shown jumps in temperature from 23°C and 37°C to be physiologically relevant (Kanegusuku *et al.*, 2021), but due to technical limitations, 30°C was the lowest temperature that could be achieved reproducibly.

4.2. Results

4.2.1. Single Mutants

To understand how environment affects epistasis it is important to first understand how environment affects fitness. To determine this, the fitness of single mutants was measured in four different environments, either varying in arabinose concentration, to simulate alterations in nutrient abundance, or temperature. The four environments were 37°C 0.1% arabinose (the 'base' environment), 37°C 0.25% arabinose, 37°C 0.5% arabinose and 30°C 0.1% arabinose. Growth rates were measured for all mutants and then normalised against the wild type growth rate within each environment (**Figure 4.1**).

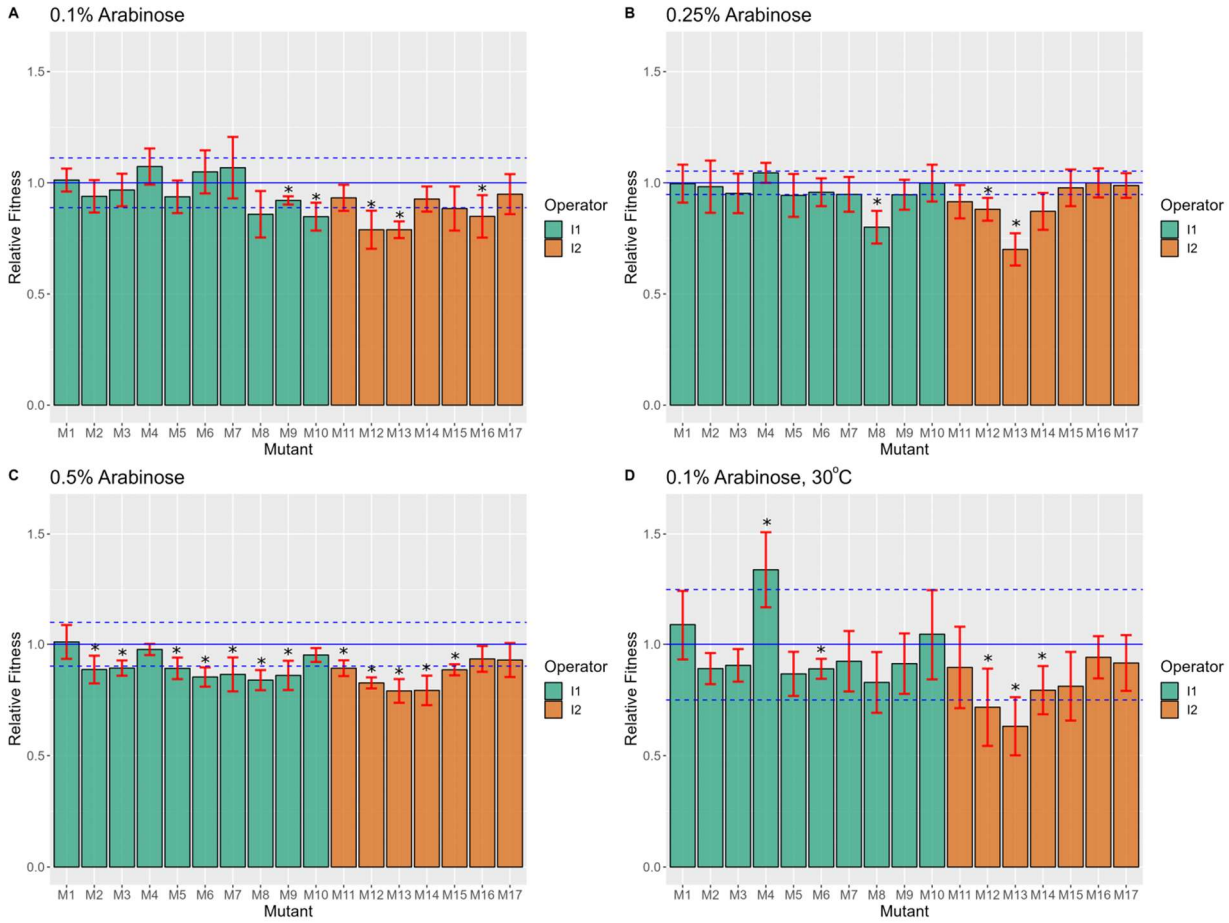


Figure 4.1 Relative fitness of single mutants in different environments. *To determine whether increased concentrations of arabinose influenced epistasis, the base concentration of 0.1% (Section 2.1.4.) was compared with 0.25% and 0.5%. Environments are A) 0.1% arabinose - 37°C B) 0.25% arabinose - 37°C C) 0.5% arabinose - 37°C D) 0.1% arabinose - 30°C. Fitness values are mean growth rates relative to wild type (blue line) in the corresponding environment. Error bars are standard deviation. Asterisks indicate significant difference from 1. Media and growth conditions described in Section 2.1.4.*

Figure 4.1 shows the fitness distribution of single mutants across four environments. In the 0.1% arabinose and 0.25% arabinose environments there were five and three mutants respectively that significantly differed from the wild type. In terms of fitness, all had lower fitness values than the wild type, with no mutant having increased fitness in either the 0.1% arabinose or 0.25% arabinose environment. In 0.5% arabinose there are 13/17 mutants with significantly lower fitness than the wild type suggesting that mutations in the *pBAD* CRE were more likely to have a negative effect on fitness in this environment. At 30°C there were four mutants with significantly negative fitness and a single mutant with significantly positive fitness. Across all environments mutants 12 and 13 had consistently lower fitness than the wild type. Mutants 6, 8, 9, 10 and 14 were negative in two environments although the specific environments varied. Several mutants were only significantly different from the wild type in the 0.5% arabinose environment.

4.2.2. Double Mutants

To understand how environment affects epistasis, it was important to explore the effects of environment on the fitness of double mutants. The relative fitness of double mutants within the same environments (**Section 2.1.4.**) was measured and tested to see if the double mutants significantly differed from the wild type value (**Figure 4.2**).

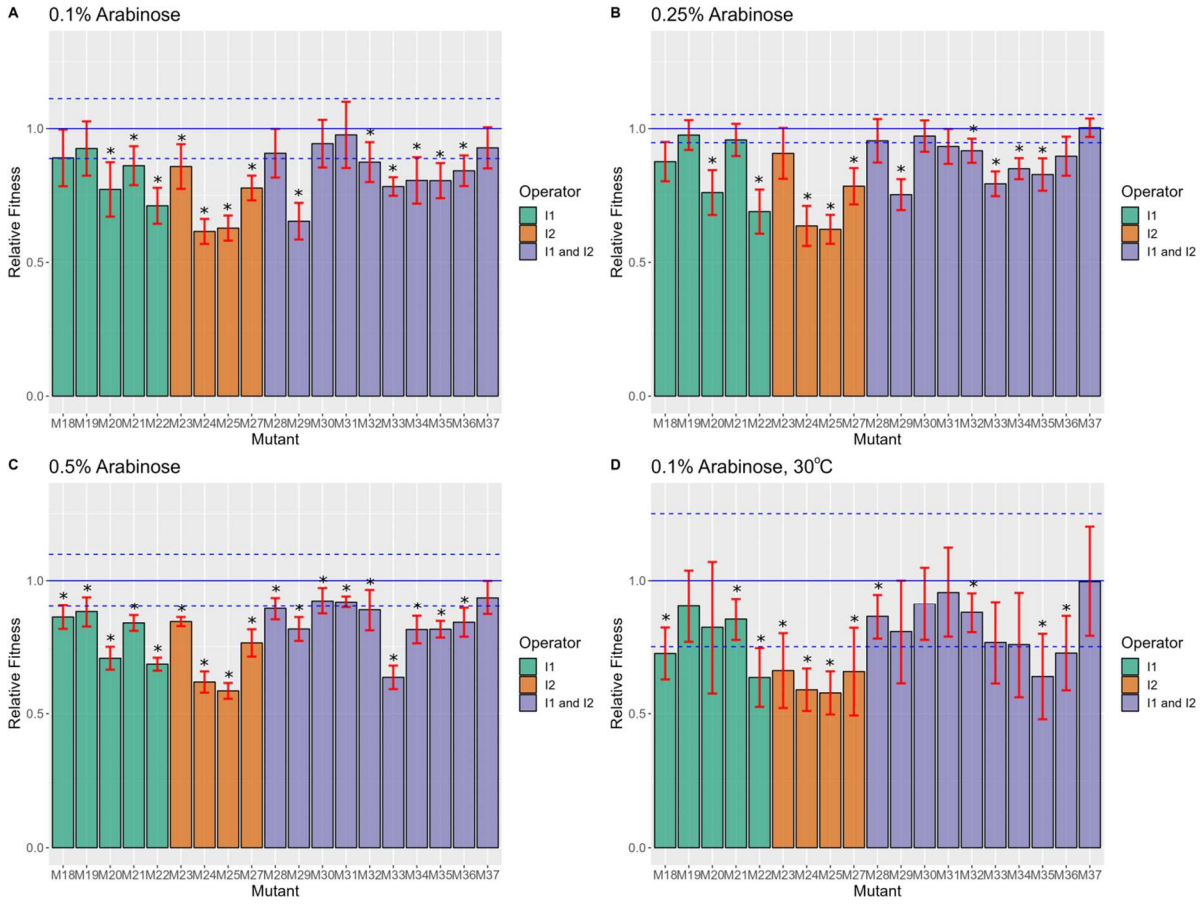


Figure 4.2 Relative fitness of double mutants in different environments. *Environments were A) 0.1% arabinose B) 0.25% arabinose C) 0.5% arabinose D) 30°C. Fitness values represent mean growth rates relative to wild type (blue line) in the corresponding environment. Error bars represent standard deviation. Asterisks indicate significant difference from 1. Media and growth conditions described in Section 2.1.4.*

Figure 4.2 shows that more than half of double mutants in all environments had fitness lower than the wild type. By contrast, in only the 0.5% arabinose - 37°C environment did more than half of the single mutants have significantly lower fitness than the wild type. An interesting observation is that in the 30°C environment there was a single mutant with increased fitness (M4) (**Figure 4.1 (d)**). However, there were no double mutants with increased fitness, indicating epistatic effects.

4.2.3. Epistasis

To understand if environment affects epistasis, the distribution of epistatic effects was compared for each environment. Epistasis was calculated for all double mutants in each environment and checked for a significant difference from zero. 0.1% arabinose 37°C was used as the base environment as this was the environment used in the previous study (**Chapter 3**) and serves as the comparator for the other environments against (**Figure 4.3**). An ANOVA test showed that epistasis values significantly differed between mutant strains (p-value <0.001), and between environments (p-value <0.05). The ANOVA additionally showed that the effect of environment on epistasis significantly varied between each mutant strain (p-value <0.001).

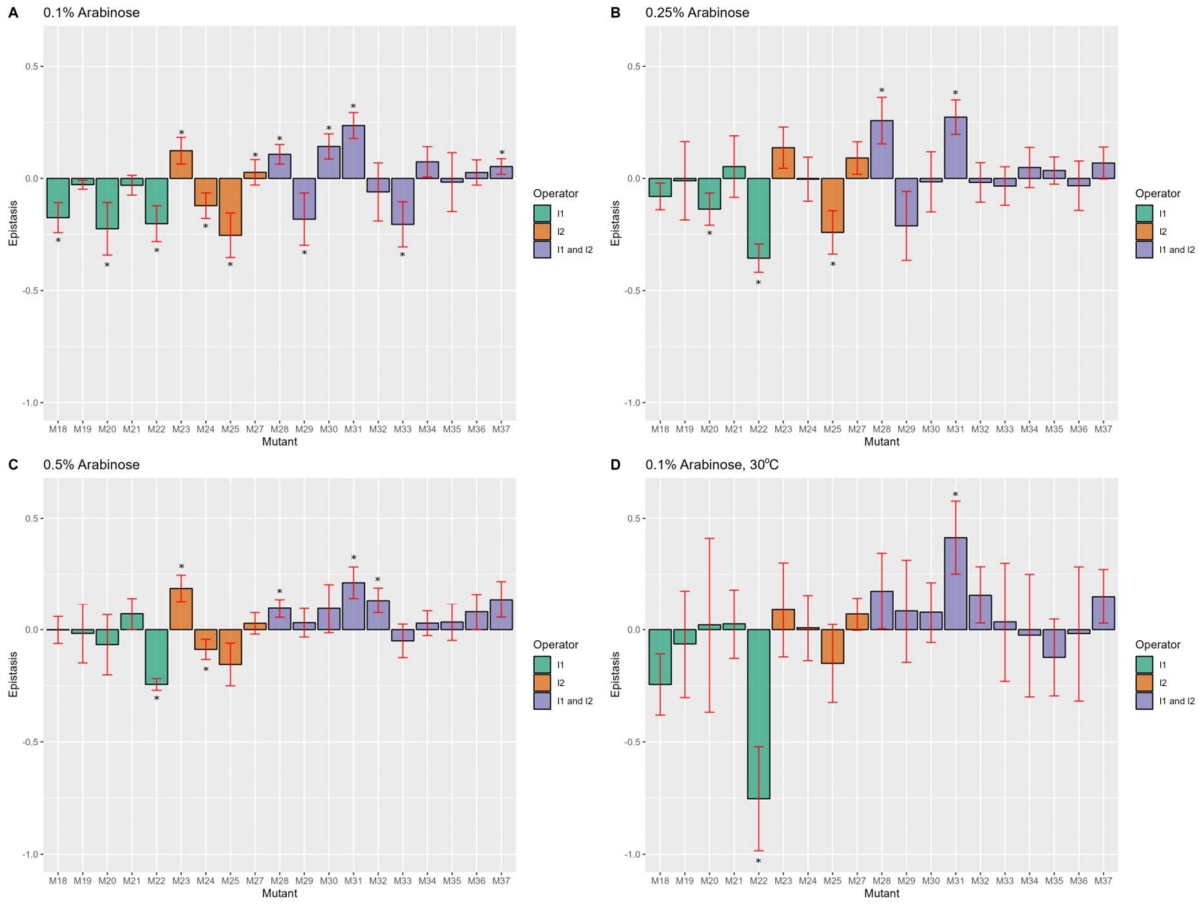


Figure 4.3 Distribution of epistatic effects across four environments. *Bars represent the mean epistasis value calculated for the respective double mutant. Colours of bars indicated operator location of the double mutant. Error bars are standard deviation. Asterisks indicate significant difference from zero. Four different environments were used; A) 0.1% arabinose 37°C, B) 0.25% arabinose 37°C, C) 0.5% arabinose 37°C and D) 0.1% arabinose 30°C.*

Figure 4.3 showed contrasting distributions of epistatic effects between environments. The base environment of 0.1% arabinose 37°C showed seven double mutants with negative epistasis and five with positive epistasis. There were less epistatic interactions overall at 0.25% arabinose 37°C in which three mutants showed negative epistasis and two showed positive. In 0.5% arabinose there were more mutants with positive epistasis (four) than with negative (two) and at 30°C there was far less epistasis present, with only one mutant each showing positive and negative epistasis. An interesting observation is that mutant 22 consistently showed negative epistasis between environments and mutant 31 consistently showed positive epistasis (**Figure 4.3**). Mutants 22 and 31 having the largest epistasis values across all environments could indicate that these epistatic interactions are resistant to changes in environment.

4.3. Discussion

For this study I aimed to elucidate the effects of environment on epistasis. To do this I used the mutant library from **Chapter 3** and measured the growth rates of mutants under different conditions before calculating any epistatic effects. The conditions were chosen to represent variable environments that *E. coli* might experience in the wild (Savageau, 1983) and that were predicted to influence growth rate on arabinose (Ammar *et al.*, 2018). It was hypothesised that mutants would have different fitness values in different environments and that epistatic interactions may differ between environments.

4.3.1. Fitness Effects

4.3.1.1. Single mutants

Different distributions of mutant fitness values were observed across the four environments (**Figure 4.1**). In the 'base' environment (0.1% arabinose, 37°C) five mutants significantly differed from the wild type, all having lower fitness than the wild type. This trend aligns with predictions from previous studies (Niland *et al.*, 1996) that mutations in the *pBAD* promoter decrease AraC binding significantly. Decreased binding of AraC limits the ability of the cell to process arabinose which can, in turn, reduce growth rate. When tested in 0.25% arabinose - 37°C, there were fewer mutants that significantly differed from the wild type than in the base environment of 0.1% arabinose - 37°C. Mutant 8 had lower fitness which was not observed at 0.1% arabinose and mutants 12 and 13 retained their significantly lower fitness.

In contrast to the other environments, at 0.5% arabinose there was a marked increase in the number of mutants showing significant differences in fitness. 13/17 mutants showed reduced fitness relative to the wild type which suggests that the mutations in the AraC binding region have more negative effects on fitness in this environment. One potential mechanism for this could be that sugar concentration was a limiting factor of growth in 0.1% and 0.25% arabinose concentrations but at 0.5%

the ability to process arabinose became the limiting factor. Consequently, the wild type strain could take full advantage of the extra arabinose available whereas the mutants were not able to capitalise, so the disparity became larger.

When investigating the effects of temperature on the single mutant fitness, I found that the 30°C – 0.1% arabinose environment also had five mutants with significant differences in fitness, similar to the base environment. However, only mutants 12 and 13 had similar fitness in both the 30°C – 0.1% arabinose and 37°C – 0.1% arabinose environments, the remaining three mutants differed between the base and 30°C – 0.1% arabinose environments. An interesting observation is M4 has significantly positive fitness in the 30°C environment which is the only case of significantly positive fitness across all environments. Data from Niland *et al.* suggested that mutations at the position of M4 had negligible effects on AraC binding so whilst it is not surprising that it has a higher fitness than other single mutants, it is surprising that it is higher than the wild type promoter, this suggests that at 30°C this position may enhance AraC binding. Pervasive G x E interactions (as defined in **Section 4.1.**) have been found in tRNA genes of yeast showing that single mutations can have drastically different fitness effects in different environments (Li & Zhang, 2018). The data shown in **Figure 4.1** also reflects these findings.

4.3.1.2. Double Mutants

When observing the effects on fitness of double mutants across the different environments, there is more consistency between environments than with the single mutants. The lower sugar concentrations of 0.1% and 0.25% had 13/19 and 10/19 significantly different fitness values respectively which is many more significant differences than when compared to the single mutants in the same conditions (**Figure 4.2**). The 0.5% sugar environment resulted in 18/19 mutants significantly differing from the wild type which was consistent with the trend seen in the single mutants where 0.5% arabinose had the highest number of significant differences, however, there were still more significantly different double mutants than single mutants in this environment. The 30°C environment had 11/19 significant differences (**Figure 4.2**), which was a higher proportion than for single mutants in the same environment (**Figure 4.1**). The consistent trend seen when measuring double mutant fitness is that a higher proportion of mutants were significantly different from the wild type in all environments. This observation would support the additive expectation of double mutations where two mutations interact additively in the phenotypic outcome (Mani *et al.*, 2008), however, even if epistasis were occurring, synergistic epistasis or weak antagonistic epistasis would still result in a greater reduction in fitness than the individual mutations alone (Phillips, 2008) so this could explain

why more significant differences are observed. The findings from Lagator *et al.* (2016) also found that the majority of double mutants showed negative epistasis which is reflected in the results of this study.

4.3.2. Epistasis

When comparing epistasis between the four environments, there were some double mutants which showed consistent epistasis and others which vary drastically. This suggests that some epistasis is resistant to changes in environment whereas some epistasis is environment dependent. Other studies have found similar interactions of epistasis with environment, where epitasis interacts with environment in some but not all cases (Chen *et al.*, 2022; Kerwin *et al.*, 2017). There were two notable mutants which displayed consistent epistatic interactions between environments: mutant 22 showing negative epistasis and mutant 31 showing positive epistasis in all four growth conditions. These mutants consistently had the largest values of negative and positive epistasis, respectively, within each environment, with the exception of mutant 22 in the base environment (**Figure 4.3**). This observation is likely explained by the large variation in the data collected. The fact that these mutants often displayed the greatest values of epistasis could suggest that the interaction is sufficiently strong enough to resist changes in the environment, as the fitness is so drastically affected that the epistatic effects persist between environments.

To put these findings in the context of relevant literature, Knijnenburg *et al.* (2009) found that the transcriptome of yeast cells was affected by 'environmental' epistasis meaning the non-additive effects of multiple environmental stimuli. Samir *et al.* (2015) then expanded on the idea of environmental epistasis, claiming that environmental factors can be treated as 'analogous' to genetic factors when affecting fitness (Knijnenburg *et al.*, 2009; Samir *et al.*, 2015). Although controversial, this claim would explain why the distribution of epistatic effects varies between environments as the given environment is interacting non-additively with the mutations to determine fitness. Temperature interacting analogously to mutations to determine fitness may explain why the 30°C – 0.1% arabinose environment showed less epistatic effects overall.

4.3.3. Limitations

The limitations to this study include the use of relative growth rates to calculate fitness rather than using competition assays. It is often reported that competition assays allow for greater sensitivity in the detection of fitness differences (Hibbing *et al.*, 2010; Ram *et al.*, 2019) and this could account for why several non-significant fitness and epistatic effects were observed **(Sections 4.2.)**. Another limitation of using a growth rates-based approach, even when using five replicates per mutant in each environment, there are still large amounts of variation in the data which reduces the statistical power of the analyses.

The use of ANOVA was limited as ANOVA compared all available data. This resulted in comparing different mutants between different environments such as M1 at 0.1% arabinose with M2 at 0.5% arabinose. This is not meaningful for this study.

4.3.4. Conclusions

This work set out to determine whether patterns of epistasis for double mutants were consistent between environments to further understand the evolutionary forces influencing the *pBAD* promoter. The data shows that epistatic effects vary between environments even when only one environmental factor is altered. This demonstrates how complex G x G x E interactions (as defined in **Section 4.1.**) can become and supports the claim that environment can be treated as analogous to a genetic factor when determining fitness (Samir *et al.*, 2015). One recent study showed that, when in a lactose rich environment, epistasis altered the evolutionary trajectory of the lac promoter, causing a different sequence of mutations to be necessary to reach the maximum fitness, due to the presence of a single mutation (Karkare *et al.*, 2021). Although it was beyond the scope of this study to produce fitness landscapes, it would be interesting to precisely map the fitness landscape within each environment to provide a higher resolution of data on how environment affects mutational interactions within the AraC binding site. Similar approaches have been taken in other studies (Chen *et al.*, 2022; Li & Zhang, 2018).

CHAPTER 5

5.1. Introduction

Horizontal gene transfer (HGT) has long been known to be a driving force in bacterial evolution. The acquisition of potentially large, novel fragments of DNA can provide a seemingly unparalleled substrate for evolutionary processes. A significant portion of prokaryotic genomes have been subject to horizontal transfer (Koonin *et al.*, 2001; Sevillya *et al.*, 2020) which highlights the impact HGT has on shaping the genomic landscape of bacteria. It has been suggested that HGT is so prevalent in microbes that the metaphor 'tree of life' should be changed to 'web of life' or rather comically the 'potato of life', as that more accurately reflected the evolution of prokaryotic species especially in early stages of cellular life (Olendzenski & Gogarten, 2009).

Due to its influence on the bacterial genomic landscape, HGT has been implicated in the transfer and creation of bacterial operons (Bundalovic-Torma *et al.*, 2020; Omelchenko *et al.*, 2003). Operons were initially defined as 'coordinated units of expression' (Jacob *et al.*, 1960, 2005) and the definition has since been expanded to include 'clusters of co-regulated genes with related functions' (Osbourn & Field, 2009) and 'any group of adjacent genes that are transcribed from a promoter into a polycistronic mRNA' (Fondi *et al.*, 2009). Operons are widespread amongst bacteria and archaea and are the most common form of gene organisation in prokaryotes (Koonin, 2009). Most operons are poorly conserved with a few significant exceptions including the ribosomal superoperon and proton ATPases which often encode proteins that physically interact (Itoh *et al.*, 1999; Wolf *et al.*, 2001). There are several theories attempting to elucidate the evolution and formation of operons and a significant contributor is the 'selfish operon model' described by Jeffrey Lawrence in 1996 (Lawrence & Roth, 1996). The selfish operon model asserts that operons are formed due to the clustering of 'non-essential' genes which conveys an evolutionary advantage in the form of gene linkage, consequently these genes are more likely to be horizontally transferred together and avoid extinction within a population. Opposing theories postulate that co-transcription or coadaptation may be plausible explanations for the presence of operons, however, these theories struggle to explain why almost all genes involved in central metabolic processes are found outside of operons (Lawrence, 1997). There is, however, evidence against the selfish operon theory that posits that the selfish operon does not explain the mechanism of gene clustering, nor does it account for the fact that essential genes are more commonly found within operons than non-essential genes (Pál & Hurst, 2004). One of the main arguments for operon existence is that of co-transcription, it is asserted that co-transcription allows

for precise control over stoichiometry as well as eliminating translational noise. Transcribing all genes as one unit allows for each gene to be expressed in equal amounts, however, the optimal level of expression for each gene within an operon is often different (Rocha, 2008) which somewhat refutes this argument.

Evidence suggests that some orthologous operons have significantly diverged from one another while still retaining function (Buvinger & Riley, 1985; Leonard *et al.*, 2015). According to Dover & Flavell (1984) and Lovell & Robertson (2010), this shows that the genes within the operons may be coevolving together to retain function, which can result in epistasis. Epistasis is the phenomenon where the fitness effects of a given mutation are dependent on the genetic background on which it appears (Phillips, 2008). Epistasis can shape the fitness landscape of operons rendering them either 'smooth' or 'rough' based on the type of epistasis present (Poelwijk *et al.*, 2011), a smooth landscape would suggest operon genes can move promiscuously between species whereas a rough landscape would suggest only horizontal transfer of the whole operon is possible as the constituent genes rely on each other to confer a fitness advantage. Evidence of HGT can indicate that negative epistasis is present as the movement of genes increases the chances of favourable combinations of genes whereas a lack of HGT can indicate positive epistasis is present as there is selective pressure against the separation of genes in the operon (Muñoz *et al.*, 2008). There is evidence for horizontal transfer of whole operons resulting in operon duplication which can act as an evolutionary substrate (Bundalovic-Torma *et al.*, 2020).

Firstly, to understand whether within-operon HGT was prevalent I looked at the gene neighbourhoods of the arabinose operon. If the arabinose operon synteny and neighbourhood are conserved across multiple species this indicates there are low levels of horizontal transfer as a successful transfer event would need to replace the native orthologue of the gene being transferred which would involve insertion next to, and subsequent loss of the native gene; an unlikely scenario (Cornet *et al.*, 2021; Rolland *et al.*, 2009; Snir, 2016).

I also looked at the clustering of genes within the operon across a wide range of species to understand if the genes are often clustered together, indicating linkage disequilibrium (Ramakrishnan, 2013). Linkage disequilibrium can be the result of epistasis acting on pairs or groups of genes (Pedruzzi *et al.*, 2018). Epistasis acting upon genes within the operon would limit potential HGT events as the operon genes would depend on the presence of each other to bring a selective advantage.

When looking to identify specific HGT events, different approaches can be taken. The phylogenetic and parametric approaches are the two most common (Sevillya *et al.*, 2020). I used the phylogenetic approach to compare the phylogenetic trees of the arabinose operon genes (*araA, araB, araC and*

araD) from a range of species in the *Enterobacteriaceae* family with a core genome tree of the same species to identify any possible HGT events.

5.2. Results

To investigate whether the arabinose operon has been exchanged between bacterial species via horizontal gene transfer, it was important to use a whole genome-based approach and to consider several factors. Gene neighbourhoods can inform whether operon clusters occur in similar genomic contexts and can reveal whether co-linear grouping of genes may be regularly transferred or not (Cornet *et al.*, 2021; Rolland *et al.*, 2009; Snir, 2016). The CBlaster tool investigates gene clustering to provide information on whether the operon is always present as a whole unit or if individual genes can appear in other combinations, suggesting whether genes are affected by epistasis. Finally, comparing arabinose gene-based phylogenies with a core genome phylogeny enabled the detection of discrepancies that could signal whether a particular gene has undergone horizontal transfer (Sevillya *et al.*, 2020), and if individual genes can be transferred independently of one another or not.

5.2.1. Gene Neighbourhoods

To start understanding the evolution of the arabinose operon between bacterial species, it was important to understand the genomic contexts in which the operon is found and whether consistent patterns were observed between species. Using GeCoViz the genomic neighbourhood of the *araBAD* operon was visualised at two different taxonomic levels.

Figure 5.1 Conservation of Arabinose operon gene neighbourhood is within the Enterobacteriaceae. *The gene neighbourhood of araB within the Enterobacteriaceae family was investigated with GeCoViz (Botas et al., 2022). Colours denote orthologous genes. Gene names are shown in white text. Bacterial species names are listed in the left-hand column. Grey genes represent genes which were not present in at least 20% of genomes. Genome order was determined by GeCoViz software. Methods described in* ***Section 2.3.1.***

The gene neighbourhood of the *araB* gene shows the conservation of gene order within the arabinose operon and reveals that the genomic context is broadly consistent amongst members of the Enterobacteriaceae. Conserved neighbourhoods imply that horizontal transfer of the *araC, araB, araA* or *araD* genes is not prevalent within this taxon. This observation is consistent with the existence of epistatic interactions between genes of the arabinose operon (**Section 5.1.**).

Figure 5.2 Reduced conservation of arabinose operon gene neighbourhood within the Gammaproteobacteria class than Enterobacteriaceae family. *Gene neighbourhood of araB within Gammaproteobacteria class using GeCoViz (Botas et al., 2022). Colours denote orthologous genes. Gene names are shown in white text. Species names are listed in the left-hand column. Grey genes represent genes which were not present in at least 20% of genomes. Genome order was determined by GeCoViz software. Methods described in* ***Section 2.3.1.***

The gene neighbourhood of the *araB* gene within the Gammaproteobacteria class (**Figure 5.2**) is less conserved than within the Enterobacteriaceae (**Figure 5.1**). It is apparent that the four constituent genes of the arabinose operon (*araC, araB, araA* and *araD*) are not always syntenic or even in close proximity. Most species carry co-linear *araA* and *araB* genes which could suggest these genes are in linkage disequilibrium caused by epistasis. This phenomenon reflects the fact that epistatic interactions impose evolutionary constraints on the rearrangement of genes, and so strong linkage between genes can indicate epistatic forces at play (Pedruzzi *et al.*, 2018). The presence of *araD* is sporadic between species, suggesting that *araD* may be transferred as a singleton gene between species as the other operon genes are not mutually dependent on *araD*. This observation suggests there may not be epistatic effects between *araD* and other genes within the operon as epistasis would act to prevent reassortment of the operon genes (Pedruzzi *et al.*, 2018).

5.2.2. Gene Clustering

Whilst gene neighbourhood analysis can inform whether the whole operon is often found in a similar genomic location, indicating whether it has experienced horizontal transfer, clustering of the operon genes can indicate whether the genes are in linkage disequilibrium. Whether the genes are consistently found in proximity to each other can reveal information about the evolutionary forces acting on this gene cluster (Pedruzzi *et al.*, 2018). Using cBlaster the arabinose operon of *E. coli* K12-MG1655 was compared with the database created in **Section 2.3.3.** 103 results were returned that had at least two arabinose genes in a cluster and these results were fed into the clinker software tool (**Appendix 1**). These results were used in conjunction with the Panaroo output from **Section 2.3.4.** to analyse the gene clustering in the context of the core genome phylogeny (**Figure 5.3**).

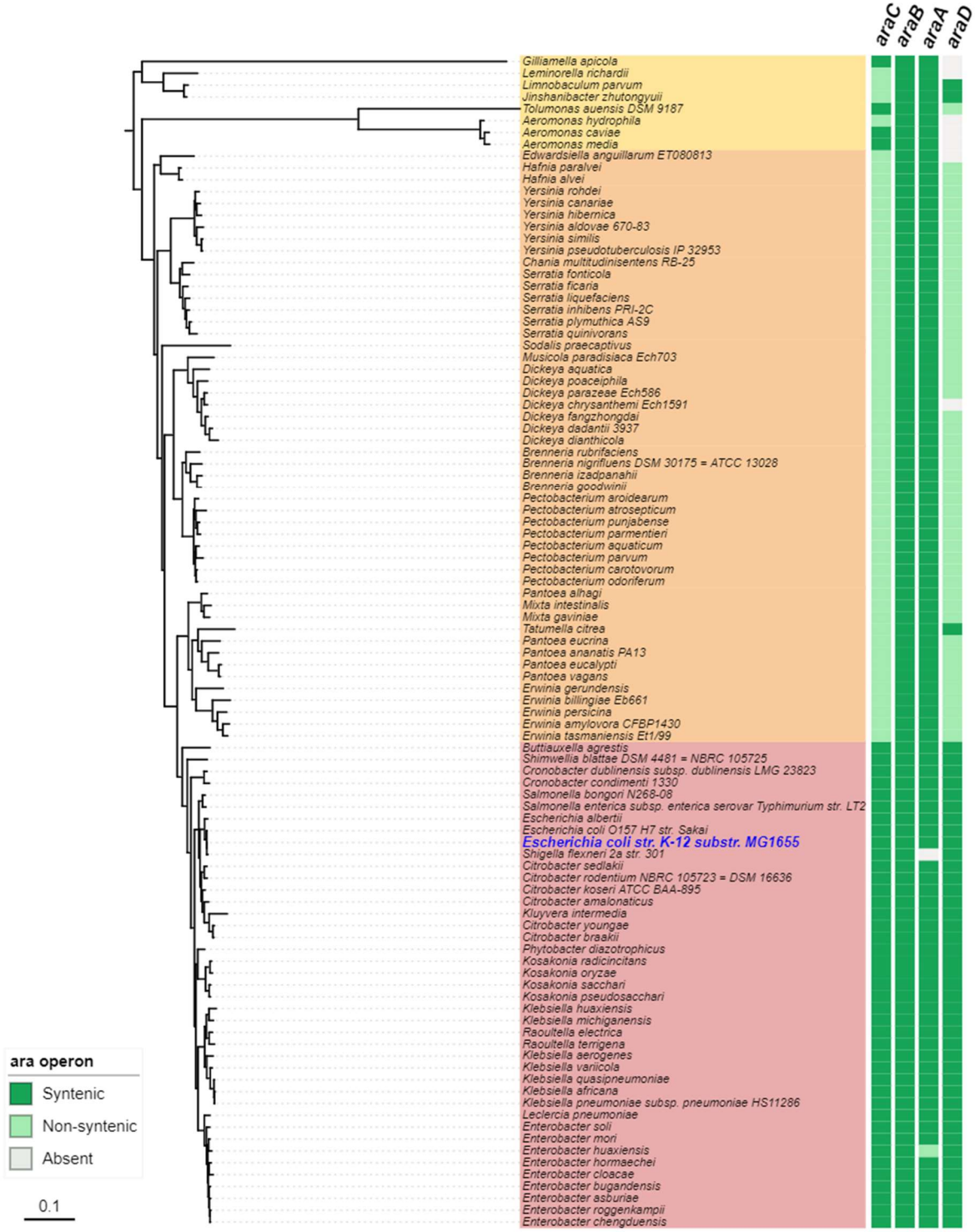


Figure 5.3 Arabinose operon genes show variable syntenic conservation across different clades. *Core genome phylogeny showing presence (green) and absence (white) of arabinose genes as well as whether they are syntenic (dark green) or non-syntenic (light green). Yellow, orange and red phylogenetic groups show common patterns of synteny. 99 core genomes produced using Panaroo as outlined in* ***Section 4.2.4****. Gene presence and absence detected using Panaroo, synteny determined as described in* ***Section 4.2.3****. Tree scale is mean number of substitutions per base of the core SNP alignment.*

The gene clustering in **Figure 5.3** shows two consistent trends, defined in red and orange. First, operons with all four genes present as a co-linear cluster that retains synteny (red group). Second, operons with only *araB* and *araA* in synteny whilst *araC* and *araD* are present but not co-linear (orange group). The data show that the synteny of *araB* and *araA* is highly conserved and the two genes experience strong linkage disequilibrium, meaning they do not segregate randomly. In contrast, *araC* and *araD* are less highly conserved and can be absent from the operon structure. There is one phylogenetic group, defined in yellow, which consistently deviate from the two common patterns of synteny found in other groups (orange and red) on the tree. Previous studies have found that there are phylogenetic groups that can be defined by operon organisation (Bundalovic-Torma *et al.*, 2020) and these findings suggest this is the case for the arabinose operon as well.

5.2.3. Phylogenetic Analysis

To further investigate the evolutionary history of the arabinose operon, gene phylogenies of each of the *araCBAD* genes were generated (**Appendix 2 (a)-(d)**) as well as a core genome phylogeny (**Appendix 2 (e)**) for all 103 species analysed (**Section 2.3.4.**).

The gene phylogenies were compared against the core phylogeny to obtain branch score values (Kuhner & Felsenstein, 1994)(**Table 5.1**) which are defined as the square root of the sum of squared differences between the branches of two trees. Branch score values allow the topologies of the four phylogenetic trees to be compared.

Table 5.1 Branch scores for araCBAD gene phylogenies compared to the core phylogeny.

	araA	*araB*	*araC*	*araD*
Core Phylogeny	1.54	0.97	0.38	1.38

The branch scores for each gene were different, with *araC* having the lowest value of 0.38 and *araA* having the highest of 1.54. This indicates the *araA* phylogeny is the least similar to the core phylogeny followed by *araD*. Having a higher branch score can indicate these genes have experienced more horizontal transfer events causing their phylogenies to differ more from the core genome than those genes with lose branch score values (Carruthers *et al.*, 2022; Planet, 2006).

5.3. Discussion

In this study I aimed to discover whether horizontal gene transfer of arabinose operon genes had occurred by looking at several traits of the operon including the gene neighbourhood, clustering of the operon genes and the individual gene phylogenies compared to a core species phylogeny.

5.3.1. Gene Neighbourhoods

The gene neighbourhood analysis shown in **Figure 5.1** and **Figure 5.2** showed that, in the Enterobacteriaceae family, there was very strong conservation of the operon with all 4 operon genes being present in the majority of species and, in most cases, being found in the same genomic contexts. This is evidence of vertical transmission and suggests that very low levels of HGT occur in this family (De & Babu, 2010). This high level of conservation can be indicative of 'colinear syntenic blocks' which are suggested to be essential to gene regulation as genes present in these blocks are strongly associated with co-transcription which aids in regulation of multiple genes associated with the same pathways (Junier & Rivoire, 2016; Svetlitsky *et al.*, 2020).

When the analysis was widened to include the class of Gammaproteobacteria, the *araBAD* operon neighbourhood showed greater divergence than within the Enterobacteriaceae family alone. There were several examples of species where the full operon was not present or where the genes were in a different order from the canonical *E. coli* operon (Schleif, 2000). The two genes that were consistently present and colinear were *araA* and *araB,* consistent with these genes being functionally dependent on each other to provide a fitness benefit, whereas *araC* and *araD* do not seem to be required. The fact that *araD* was not always present aligns with the fact that L-ribulose-5-phosphate 4-epimerase is not a necessary component of the arabinose pathway for the utilisation of arabinose as a carbon source (Boulanger *et al.*, 2021). AraA is L-arabinose isomerase which converts L-arabinose into L-ribulose which can only be metabolised via AraB. L-ribulose is converted to L-ribulose-5-phosphate by AraB; Ribulokinase. *araD* encodes L-ribulose-phosphate-4-epimerase which converts L-ribulose-5-phosphate into D-xylulose-5-phosphate. However, this step is not necessary as L-ribulose-5-phosphate can be utilised in other metabolic pathways and so can still be utilised as a carbon source (Boulanger *et al.*, 2021). This alternative pathway for L-ribulose-5-phosphate could explain why *araB* is necessary when *araA* is present but *araD* is not necessary.

5.3.2. Cluster Analysis

When analysing the clustering of the arabinose genes across 103 species, clear patterns were observed. As in **Section 5.3.1.** , *araA* and *araB* were co-linear in all but two of the genomes, suggesting a strong dependence on each other. *araC* and *araD* were syntenic in roughly half of the genomes and were often co-linear together completing the full complement of genes found in the canonical *E. coli* operon (Schleif, 2000). This pattern of synteny indicates that vertical inheritance may be responsible for these clusters as there was no evidence the operon genes were being transferred promiscuously between species.

5.3.3. Phylogenetic Analysis

When comparing the phylogenies of the arabinose genes to the core genome phylogeny of the 103 species (**Section 2.3.3.**) the branch scores varied between the genes. *araA* and *araD* had the highest scores of 1.54 and 1.38 respectively (**Table 5.1**) meaning these trees were the most different from the core tree. A higher branch score can suggest these genes have experienced more horizontal transfer events than *araC* and *araB*. It is unexpected that *araA* had a higher score than the other genes as the gene neighbourhood and gene clustering analysis suggested it was highly conserved and commonly occurred in the presence of *araB. araB,* by contrast, had a lower branch score than *araA,* indicating less horizontal transfer. Due to the limitations of the trees themselves, the branch score can only provide limited insight into the discordance of the trees.

5.3.4. Limitations

The analysis in this study was mostly investigative and was not designed to provide conclusive evidence of horizontal transfer and, consequently, epistasis within the arabinose operon. Nevertheless, there were several limitations to the approaches used, including the pool of genomes that were analysed as part of the GeCoViz software. The limited selection of genomes that are available to use in GeCoViz prevents the inclusion of potentially relevant genomes. As a result, not all possible species were analysed and some meaningful variations of the arabinose operon may have been missed, somewhat biasing the data.

The clustering analysis performed by Cblaster provided information about genes in close proximity (**Appendix 1**). However, Cblaster did not provide information on whether the missing genes were found elsewhere in the genome. The genes of the operon were not always syntenic in some species but analysis via Panaroo indicated that they were still present in the genome (**Figure 5.3**). Limited conclusions can be drawn about whether the genes are experiencing epistasis, which can occur without collinearity, but is often associated with linkage disequilibrium.

The phylogenetic analysis provides an indication of the similarity of the gene trees to the core phylogeny. However, upon visual inspection, some of the gene trees show such a high level of incongruence that it is possible they are not an accurate reflection of the true gene phylogeny. This could be due to the genes themselves not providing enough resolution to create a reliable phylogeny (V'yugin *et al.*, 2003) as some of them have very short sequence lengths and could also be due to the difficulty of comparing closely related species (Adato *et al.*, 2015).

5.3.5. Conclusions

The findings in this chapter provide a preliminary exploration of the conservation of the arabinose operon, and the evolutionary forces that may be at play. Both the gene neighbourhood and gene clustering analyses indicated a strong co-occurrence of *araA* and *araB* which can be explained by their roles in the arabinose metabolism pathway. This conserved synteny suggests there is positive epistasis between *araA* and *araB* as the two gene products cannot fulfil the function of utilising arabinose as a carbon source without one another. AraC is a regulatory protein and is known to have many homologues (Schleif, 2010) and so it not necessary for the expression of *araA* and *araB* which could explain why it is not always present. AraD is also not required for the metabolism of arabinose and therefore is most likely not linked via epistasis to the other operon genes.

The phylogenetic analysis, although limited, provided insight into the incongruence of the arabinose gene trees compared to the core phylogeny. This analysis indicated that *araA* phylogeny deviates more from the core tree than *araB* which is unusual if the two genes are linked via epistasis as this should mean they show similar evolutionary histories. However, more detailed analysis would need to be undertaken to truly elucidate the evolutionary histories of these genes. Possible further methods could involve reconciliation of gene trees with the species tree or using software such as GATC to infer gene trees (Noutahi & El-Mabrouk, 2018).

CHAPTER 6

General Discussion

6.1. A Brief Introduction

Genetic interactions are an important force influencing the evolution of both coding and non-coding sequences, termed Epistais (Domingo *et al.*, 2019; Lagator *et al.*, 2016; Lehner, 2011). With the advances of genetic techniques over the last few decades it has become easier to study epistasis as high throughput screening of double mutants has become standard (Anderson *et al.*, 2021; Lagator, Sarikas, *et al.*, 2017; Nghe *et al.*, 2018). Recently, scientists have begun to investigate epistasis in greater detail, evaluating all permutations of specific sequences to generate fitness landscapes (Chen *et al.*, 2022). This provides insight into the evolutionary potential of genetic systems and metabolic networks. In this thesis I set out to build upon this knowledge using the arabinose operon and associated regulatory regions as a model system.

6.2. Epistatic effects on expression versus fitness

In **Chapter 3**, I investigated whether epistatic effects have a consistent impact upon gene expression values versus bacterial fitness. Using gene expression data and epistasis values from (Lagator *et al.*, 2016), I then set up a system to convert (Lagator *et al.*, 2016) mutant library from a system that expressed the fluorescent protein Venus-YFP into a system that expressed the metabolic genes of the arabinose operon. My approach was designed to allow the mutants growth on arabinose to be measured as a proxy for fitness. The data were used to calculate epistasis values to compare with those of (Lagator *et al.*, 2016). Significant difficulties were encountered in measuring epistasis using the equipment selected and large variances in the data were observed. Consequently, no statistically significant genetic interactions were detected meaning that a link between the impact of epistasis upon gene expression and fitness could not be determined. Although each measurement of growth rate (as a proxy for fitness) was measured five times, this was not sufficient to reduce the experimental variation. While considering whether there may be a fundamental problem with trying to detect epistasis via growth rate measurement, I noted that others have successfully used growth rate to detect epistasis in viruses and yeast (Poyatos, 2020; Sackman & Rokyta, 2018).

An observation regarding the system used in this chapter is that, due to the plasmid being present in multiple copies, even at a very low copy number (3-4)(Lutz & Bujard, 1997), multiple copies of the *araC* gene could lead to more AraC protein being made by the cell. Because AraC autoregulates its own transcription (Schleif, 2010) and that of the *araBAD* genes, a greater amount of AraC in the cell

could lead to higher concentrations of arabinose being needed to bind all AraC molecules and cause a conformational change, resulting in expression of the *araBAD* genes. Therefore, a plasmid-based system could react to arabinose levels in a manner distinct from that of a chromosomally-based system.

6.3. Environmental effects on epistasis

In **Chapter 4**, I used an alternative technical approach to investigate whether the distribution of epistatic effects was affected by environment. I used the mutant library created in **Chapter 3** to help me answer this question. I grew the mutant library in different conditions, varying in either sugar concentration or temperature, and calculated epistasis values to compare to a 'base environment' used in (Lagator *et al.*, 2016) namely 0.1% arabinose and 37°C. I discovered that certain epistatic effects remained consistent between environments whilst others varied dramatically. This implies some genetic interactions are strong enough to resist environmental perturbation whilst others are more susceptible to changes in environment. Other studies have shown that epistasis can be influenced by environment in bacteria and yeast (Anderson *et al.*, 2021; Baier *et al.*, 2022; Chen *et al.*, 2022; Li & Zhang, 2018) and so these findings are consistent with the literature. One study in particular found that, when in a lactose rich environment, epistasis created two evolutionary pathways to reach a maximum fitness, dependent on the presence of a single mutation (Karkare *et al.*, 2021). I speculate that because certain epistatic interactions are not environmentally sensitive, these interactions could have an evolutionary impact, regardless of the environmental niche of the organism in question. Another factor affecting whether beneficial mutations such as M4 may become fixed in a population are the opposing forces of stimulation of the operon in the presence of arabinose and repression in the absence of arabinose. Whilst a mutation may be beneficial in one scenario, it may be detrimental in the other. Therefore, due to changing environments in nature, particular mutations may not become fixed in a population. This idea is also discussed by Lagator *et al.* (2016).

Whilst environment may affect epistasis, it is also important to note that environmental changes affect the basal growth rate of any cell regardless of genetic makeup. Detailed in **Table 6.1** are the relative growth rates observed for the wild type plasmid in each of the environments tested. It is clear that the 0.5% arabinose provides a small growth benefit, whilst the 30°C environment results in a much slower growth rate.

Table 6.1 Growth rates of the wild-type strain in the four environmental conditions tested (Section 2.2.2.).

Environmental Condition	Percentage Growth Rate of Base Environment
37°C 0.1% arabinose (the 'base' environment)	100%
37°C 0.25% arabinose	93%
37°C 0.5% arabinose	107%
30°C 0.1% arabinose	43%

6.4. Epistasis between genes of the arabinose operon

In **Chapter 5**, to look for evidence of epistasis between the genes of the arabinose operon, it was important to look at several factors including horizontal gene transfer, linkage disequilibrium and syntenic conservation to understand whether genes within the arabinose operon may have been under the effects of epistasis. The findings could then shed light on the evolution of the arabinose operon and whether the gene cluster is largely inherited vertically or undergoes dynamic changes due to gene transfer and rearrangements. I found the operon was largely conserved in certain phylogenetic taxa such as Enterobacteriaceae and that the gene order of the operon fell largely into two, group specific patterns. The synteny of *araA* and *araB* was largely conserved across all species, indicating there could be genetic interactions between these genes, causing them to be co-dependent. On the other hand, *araC* and *araD* were only syntenic in one clade. In the other clade, *araC* and *araD* were not co-located on the chromosome. It is therefore unlikely that epistasis affects these genes. The clade in which *araC* and *araD* were syntenic were all gut residing bacteria which encounter arabinose through food ingested by the host, whilst the clade lacking *araC* and *araD* included various environmental bacteria, including plant pathogens and soil microbes. The two aforementioned clades are exposed to completely different environments which could reflect the evolution of the arabinose operon in these species. A further area of investigation could be to study the relationship between the arabinose metabolism of these organisms and their natural environment to shed light on the importance of various arabinose genes within species occurring in different environments.

There was no significant evidence for horizontal transfer of individual genes in the Enterobacteriaceae, although the techniques used were limited. I conclude that epistatic forces are likely acting upon genes in the operon. This is because epistasis causes genes to display linkage disequilibrium (Pedruzzi *et al.*, 2018) and horizontal transfer of individual genes would act against this disequilibrium. If genes are co-dependent, then the transfer of individual genes would be selected against. The lack of evidence for HGT of the arabinose operon genes is consistent with the existence of significant epistatic interactions.

The findings from **Chapter 4**, and **Chapter 5** support the fact that epistasis affects the evolution of the arabinose operon, both in its encoding and regulatory sequences. Genetic interactions are therefore an important mechanism which affects how organism may acquire and adapt, operon encoded traits.

6.5. Techniques for measuring epistasis

A common problem faced in several of the studies in this thesis was the suitability of the techniques used to detect epistasis to a sufficient level to make robust conclusions. In **Chapter 3**, the Tecan Infinite® 200 PRO plate reader was used to measure growth rates. Drastically different values were measured between repeat measurements of the same mutants in identical growth conditions (Section 2.2.4). There is some evidence of successful optical density measurements being performed with this instrument (Dlugaszewska *et al.*, 2016; Kuznetsova *et al.*, 2013). However, the Tecan Infinite® 200 PRO is often used for fluorescence assays (Brigo *et al.*, 2023; Correa *et al.*, 2020) and no literature that demonstrated successful measurement of growth rates using this instrument could be found.

Upon the commencement of the environment x epistasis assays in **Chapter 4**, the instrument was switched to the Growth Profiler 960 platform (Enzyscreen) which is an instrument tailored to performing multiple growth curves simultaneously. This instrument was installed in the University of Liverpool microbiology labs in the final year of my PhD. The Growth Profiler 960 platform (Enzyscreen) provided highly reproducible growth rate data (**Section 2.2.2.**) which allowed statistically significant differences in fitness-based epistasis values to be detected (**Section 2.2.2.**). My findings highlight the importance of accurate growth rate measurements for the detection of epistasis. The determination of epistasis (**Section 2.1.6.**) involves multiplying the single mutant growth rates together to determine the additive expectation of fitness. Therefore, if the growth rates were inaccurate the variance was amplified in the additive value resulting in drastically inaccurate epistasis values being calculated.

In **Chapter 5** an approach was taken that did not involve experimental measurements, relying upon a comparative genomic investigation of 'footprints' of epistasis. The arabinose operon was analysed by studying conservation and gene synteny within a range of genomes from the Enterobacteriaceae. A phylogenetic analysis was done in parallel. The goal of these techniques was to identify any potential signs of horizontal transfer occurring within the arabinose operon, as HGT can indicate a lack of epistasis due to HGT reducing linkage disequilibrium which often correlates with epistasis (Cornet *et al.*, 2021; Pedruzzi *et al.*, 2018). Other studies that have successfully used synteny to infer HGT (Rolland *et al.*, 2009; Snir, 2016) formed a starting point for the analysis.

Whilst the results indicated significant syntenic conservation, especially between closely related species, this did not disprove the occurrence of HGT events. Therefore, a phylogenetic analysis was performed because discordance between phylogenetic trees can indicate HGT (Carruthers *et al.*, 2022; Planet, 2006). Several issues arose with this method, the first being that visual determination of tree incongruence by creating 'tanglegrams' resulted in extremely levels of incongruence with almost no branches showing congruence between the core genome phylogeny and the arabinose gene phylogenies.

A quantitative method, using 'branch score values' (Kuhner & Felsenstein, 1994) was selected to assess the differences in the trees. This approach returned values which conflicted with the synteny analysis, suggesting that *araA* was the most incongruent gene tree even though it was one of the genes found to have the highest level of syntenic conservation.

There is not extensive literature demonstrating the use of 'branch score values' for the comparison of gene-based phylogenetic trees. The approach of using tree incongruence to identify evolutionary events can be unreliable due to the innate level of incongruence generated when using different tree building methods as well as natural variations in incongruence between different genomic regions (Som, 2015). Although tree incongruence techniques have been used to identify HGT (Anselmetti *et al.*, 2021; Kim *et al.*, 2018; Sutherland *et al.*, 2021), there may not have been enough diversity between the arabinose gene sequences to produce reliable phylogenies which could provide accurate phylogenies for incongruence analysis (Xi *et al.*, 2015).

6.6. Concluding statements

The studies in this thesis shed light on epistasis and its prevalence in the arabinose operon and the *pBAD* CRE regulatory region. It shows that epistatic interactions between mutations in the *pBAD* CRE region can be strong enough to resist changes in the environment which raises the possibility that these interactions could shape the evolution of this region more broadly. The identification of certain environmentally sensitive epistatic interactions indicates these interactions are likely to influence the evolution of the *pBAD* CRE dynamically depending on the environmental conditions affecting the cell. It is worth noting that in the presence of an *araC* mutant, in which AraC cannot bind $araI_1$ or $araI_2$, the epistatic interactions within the *pBAD* promoter would cease to exist as they depend on AraC binding to exert an effect on the cell. Evidence of strong syntenic conservation of some genes within the arabinose operon was found, indicating that there could be epistatic forces affecting specific genes within the operon but not necessarily others. These findings provide insight into how epistasis shapes the evolution of operons through not only gene interactions but also at the sequence level, determining mutational limitations affecting regulatory regions.

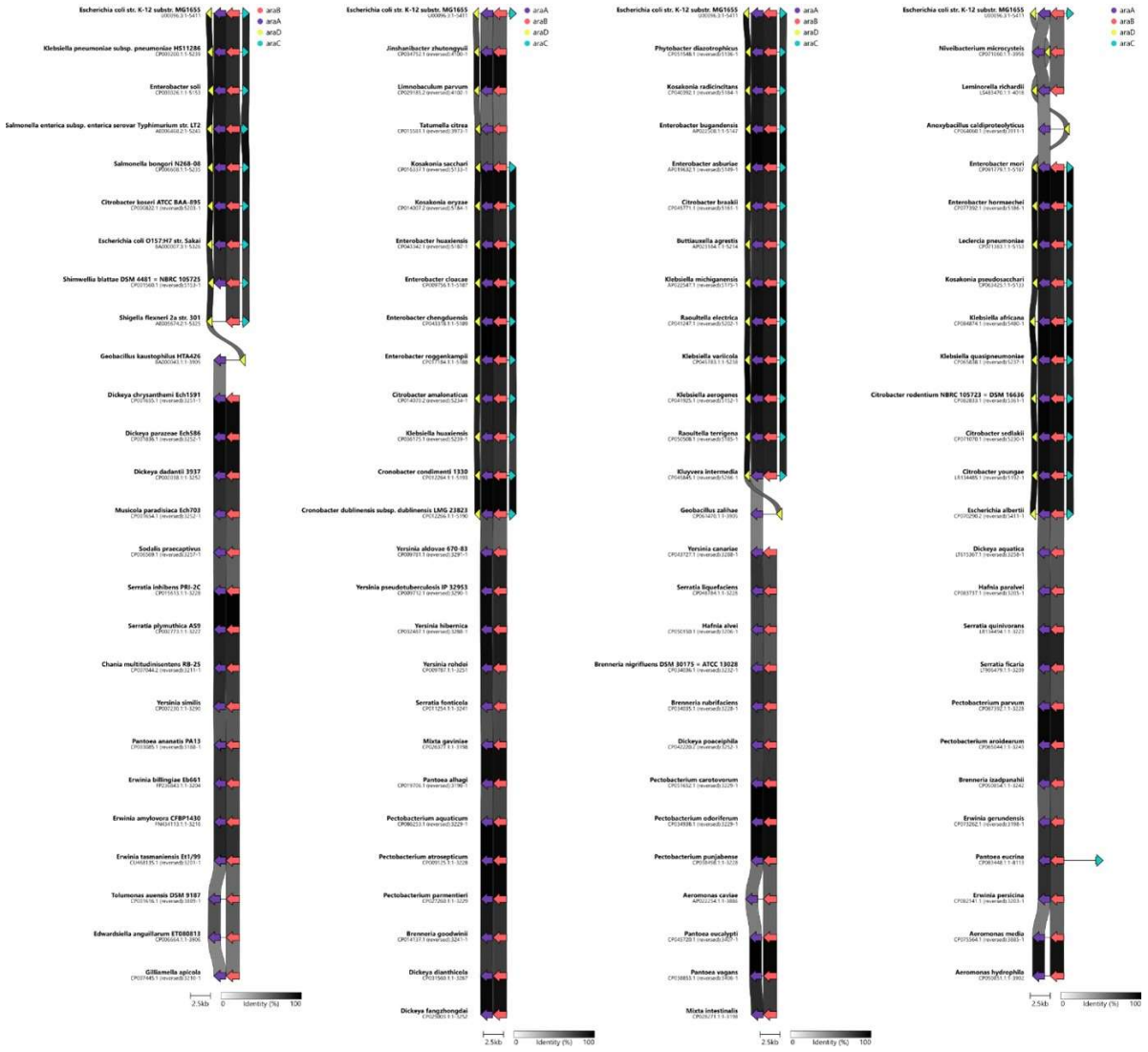
The studies also demonstrate how detecting epistasis is not a straightforward process and can be sensitive to the techniques used as well as the quality of data obtained. Studies are increasingly using high throughput methods to assess epistatic interactions across all possible mutational combinations and generating fitness landscapes of metallo-enzymes, β-lactamases and fluorescent proteins (Anderson *et al.*, 2021; Chen *et al.*, 2022; Sarkisyan *et al.*, 2016) which gives the deepest insight into the evolutionary potential of both coding and non-coding sequences.

Although generating an extensive fitness landscape was beyond the scope of the work in this thesis, this work provides the basis for understanding the evolutionary dynamics of operons and their regulatory sequences which in turn can help to understand the evolution of bacteria and the acquisition and evolution of traits such as antibiotic resistance and hydrocarbon degradation (Kunonga *et al.*, 2000; Zylstra *et al.*, 1988).

Supplementary Materials

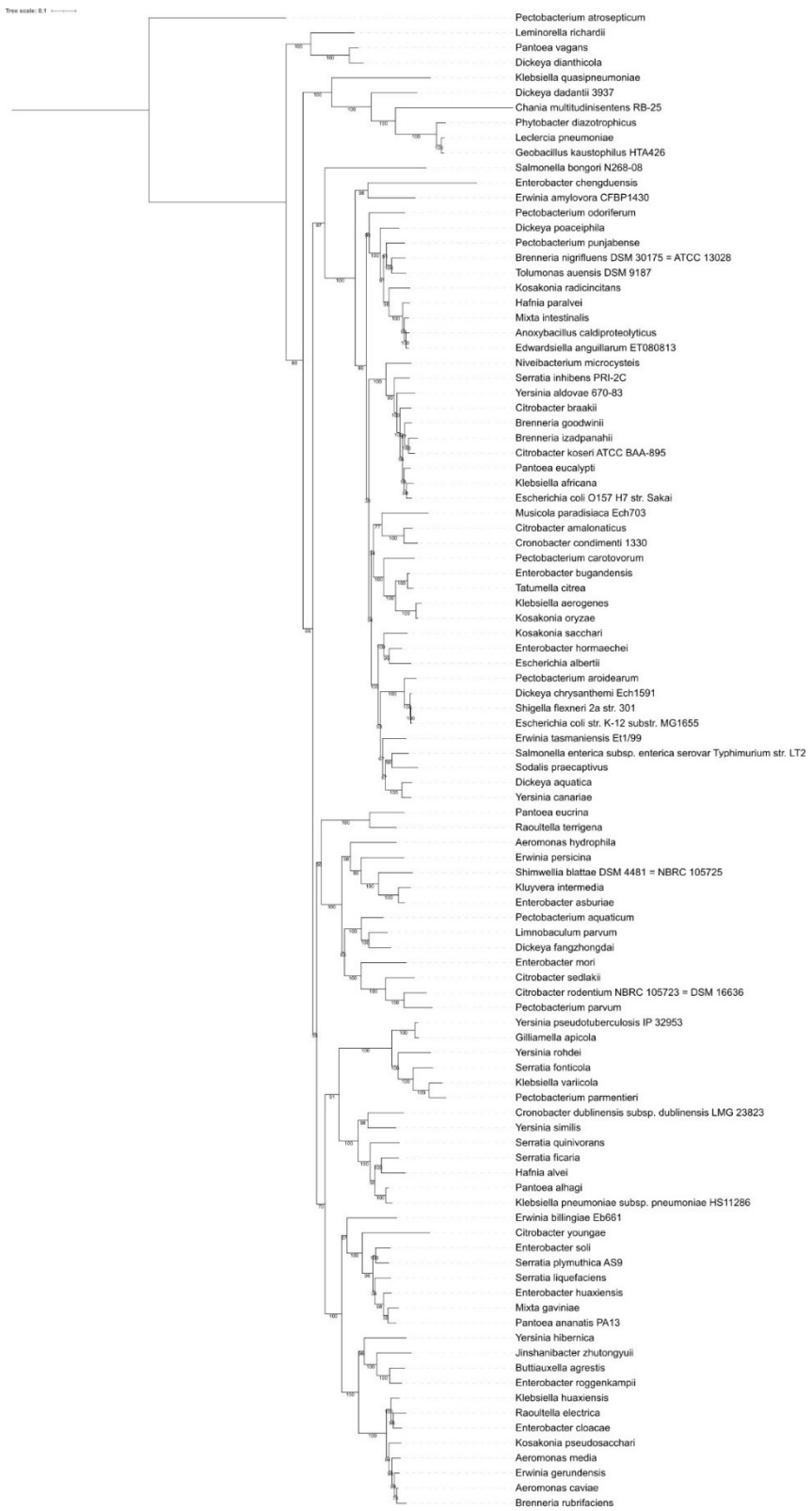
Table of Appendices

Appendix 1: Arabinose gene clusters show variable patterns of synteny and presence across 103 species analysed. *Gene clusters found through Cblaster (5.2.4) visualised in clinker (5.2.5). Species names given to the left of each cluster. Genes are coloured to show orthologs.*

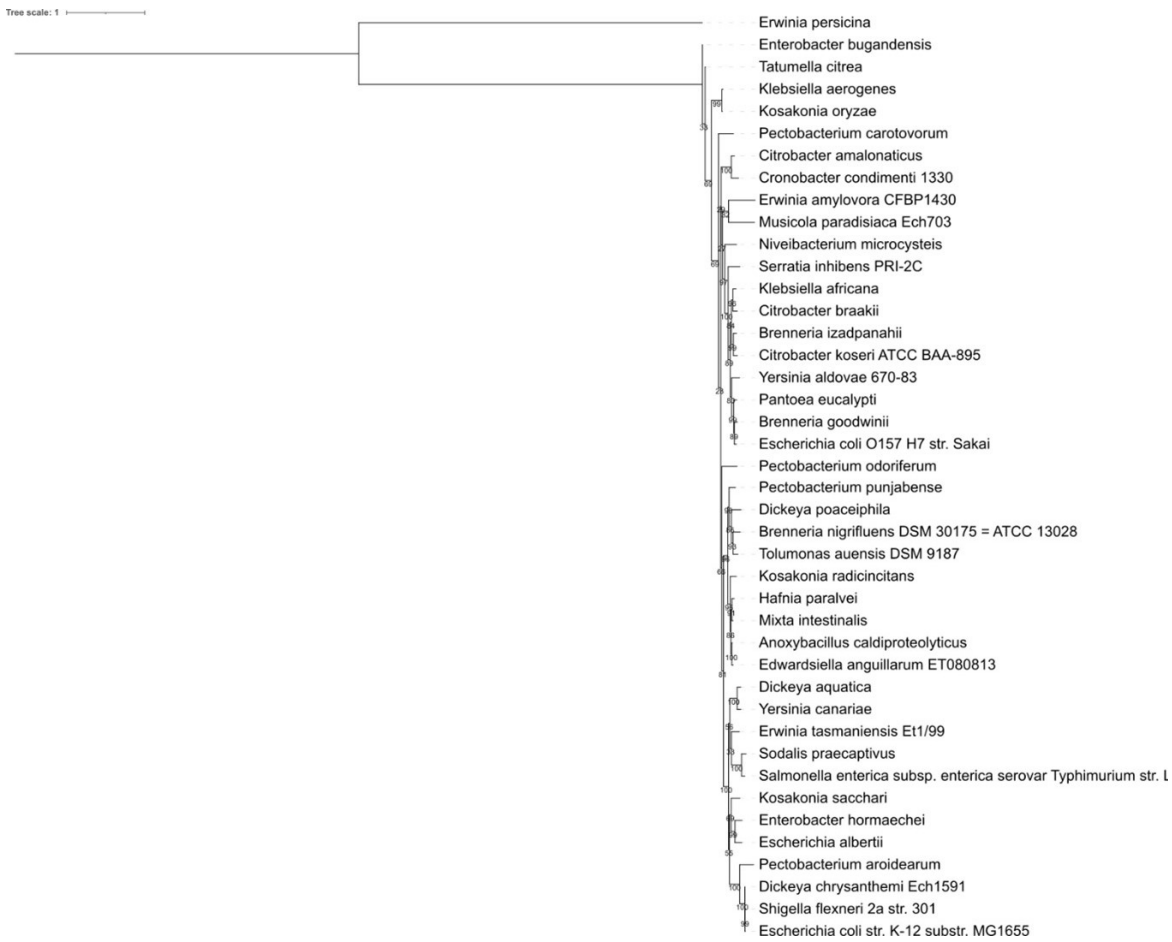


Appendix 2: Phylogenetic trees showing all species containing the araA (a), araB (b), araC (c), and araD (d) genes, in addition to the core genome (e).

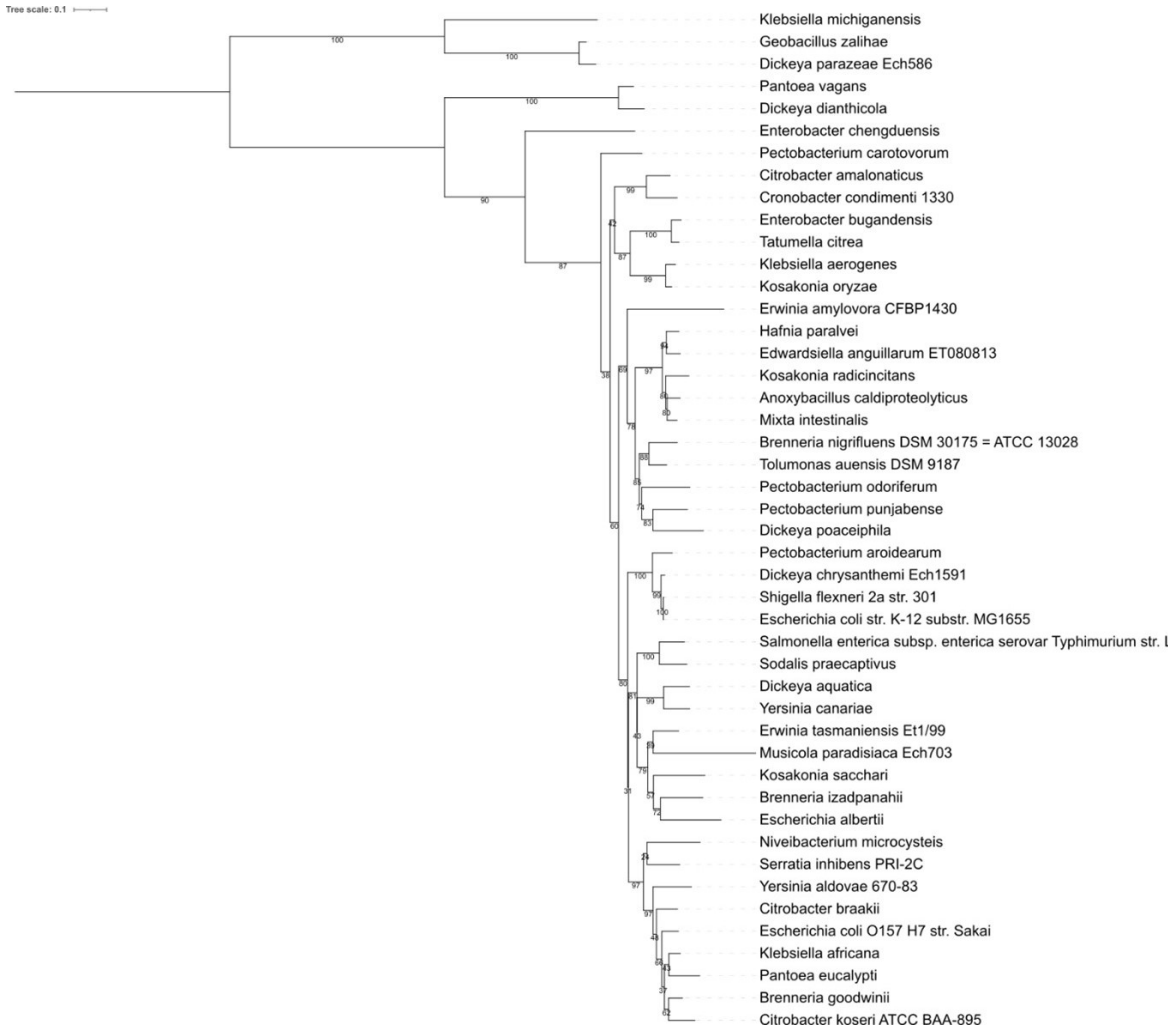
(a) araA Gene Phylogeny*. Phylogenetic tree showing all species that contained the araA gene. Produced using IQTREE. Node labels are bootstrap values for 1000 bootstraps. Tree scale is mean number of substitutions per base of the core SNP alignment.*



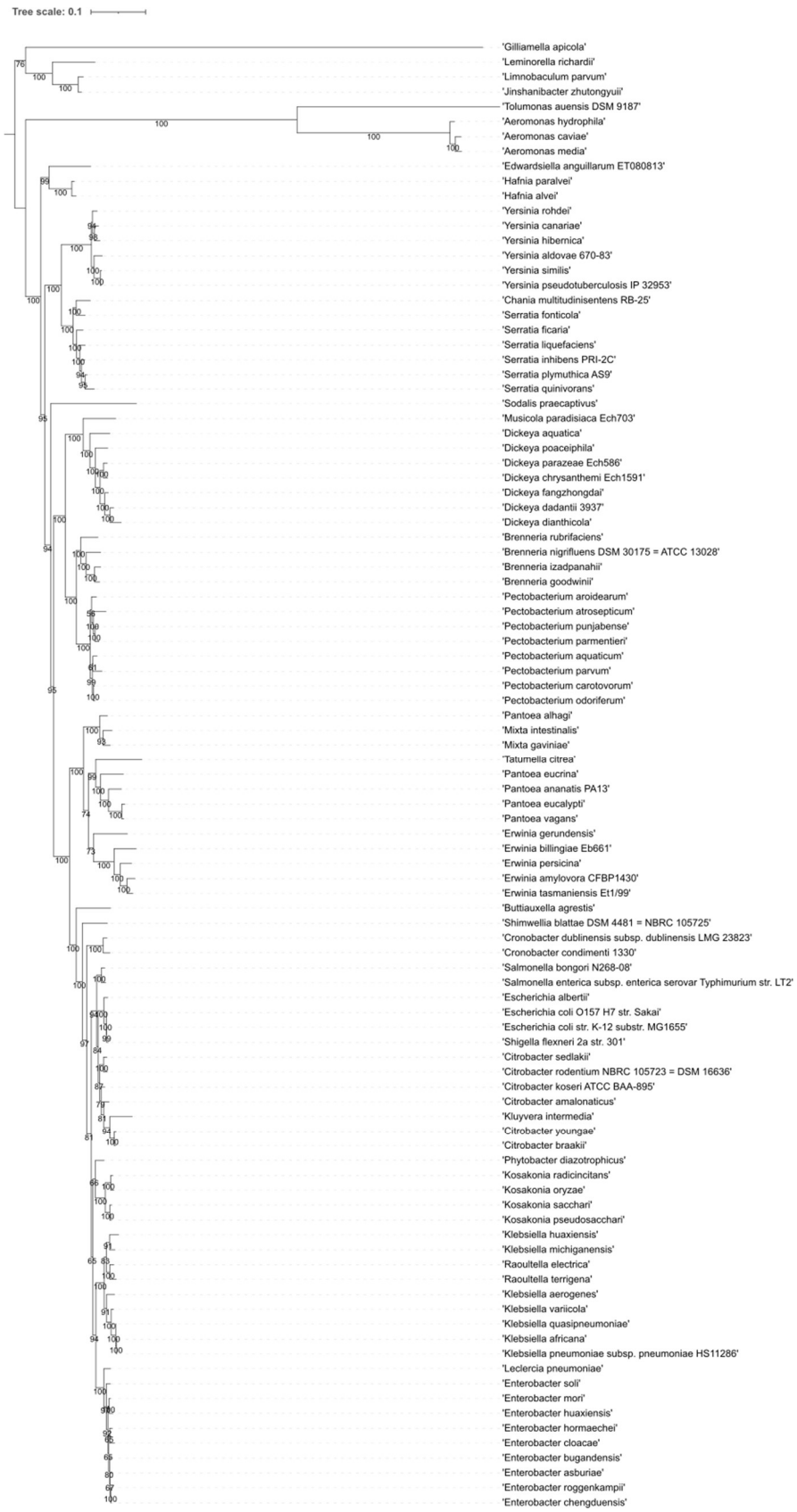
(b) araB Gene Phylogeny*. Phylogenetic tree showing all species that contained the araA gene. Produced using IQTREE. Node labels are bootstrap values for 1000 bootstraps. Tree scale is mean number of substitutions per base of the core SNP alignment.*



(c) araC Gene Phylogeny*. Phylogenetic tree showing all species that contained the araA gene. Produced using IQTREE. Node labels are bootstrap values for 1000 bootstraps. Tree scale is mean number of substitutions per base of the core SNP alignment.*



(d) araD Gene Phylogeny*. Phylogenetic tree showing all species that contained the araD gene. Produced using IQTREE. Node labels are bootstrap values for 1000 bootstraps. Tree scale is mean number of substitutions per base of the core SNP alignment.*



(e) Core Genome phylogeny of 103 species. *Phylogenetic tree produced using IQTREE using the 103 species described in section 5.2.3.Node labels are bootstrap values for 1000 bootstraps. Tree scale is mean number of substitutions per base of the core SNP alignment.*

BIBLIOGRAPHY

Adato, O., Ninyo, N., Gophna, U., & Snir, S. (2015). Detecting Horizontal Gene Transfer between Closely Related Taxa. *PLOS Computational Biology*, *11*(10), 1–23. https://doi.org/10.1371/journal.pcbi.1004408

Afroz, T., Biliouris, K., Kaznessis, Y., & Beisel, C. L. (2014). Bacterial sugar utilization gives rise to distinct single-cell behaviours. *Molecular Microbiology*, *93*(6), 1093–1103. https://doi.org/10.1111/mmi.12695

Alam, Y. H., Kim, R., & Jang, C. (2022). Metabolism and Health Impacts of Dietary Sugars. *Journal of Lipid and Atherosclerosis*, *11*(1), 20. https://doi.org/10.12997/jla.2022.11.1.20

Alizon, S., & Michalakis, Y. (2015). Adaptive virulence evolution: the good old fitness-based approach. *Trends in Ecology & Evolution*, *30*(5), 248–254. https://doi.org/10.1016/J.TREE.2015.02.009

Ammar, E. M., Wang, X., & Rao, C. V. (2018). Regulation of metabolism in Escherichia coli during growth on mixtures of the non-glucose sugars: arabinose, lactose, and xylose. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-017-18704-0

Anderson, D. W., Baier, F., Yang, G., & Tokuriki, N. (2021). The adaptive landscape of a metallo-enzyme is shaped by environment-dependent epistasis. *Nature Communications*, *12*(3867). https://doi.org/10.1038/s41467-021-23943-x

Anselmetti, Y., El-Mabrouk, N., Lafond, M., & Ouangraoua, A. (2021). Gene tree and species tree reconciliation with endosymbiotic gene transfer. *Bioinformatics*, *37*, i120–i132. https://doi.org/10.1093/bioinformatics/btab328

Arnqvist, G., Dowling, D. K., Eady, P., Gay, L., Tregenza, T., Tuda, M., & Hosken, D. J. (2010). Genetic architecture of metabolic rate: environment specific epistasis between mitochondrial and nuclear genes in an insect. *Evolution*, *64*(12), 3354–3363. https://doi.org/10.1111/j.1558-5646.2010.01135.x

Arrigo, K. R. (2014). Sea Ice Ecosystems. *Annual Review of Marine Science*, *6*, 439–467. https://doi.org/10.1146/annurev-marine-010213-135103

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., & Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout

mutants: the Keio collection. *Molecular Systems Biology*, *2006*(0008). https://doi.org/10.1038/msb4100050

Babu, M. M., & Teichmann, S. A. (2003). Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Research*, *31*(4), 1234–1244. https://doi.org/10.1093/NAR/GKG210

Baier, F., Gauye, F., Perez-Carrasco, R., Payne, J. L., & Schaerli, Y. (2022). Environment-dependent epistasis increases phenotypic diversity in gene regulatory networks. *BioRxiv*. https://doi.org/10.1101/2022.09.18.508240

Beg, Q. K., Vazquez, A., Ernst, J., de Menezes, M. A., Bar-Joseph, Z., Barabási, A.-L., & Oltvai, Z. N. (2007). Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. *Proceedings of the National Academy of Sciences*, *104*(31), 12663–12668. https://doi.org/10.1073/pnas.0609845104

Benes, V., Kilger, C., Voss, H., Pääbo, S., & Ansorge, W. (1997). Direct Primer Walking on P1 Plasmid DNA. *BioTechniques*, *23*(1), 98–100. https://doi.org/10.2144/97231bm21

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., & Tawfik, D. S. (2006). Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, *444*(7121), 929–932. https://doi.org/10.1038/nature05385

Blount, Z. D., Maddamsetti, R., Grant, N. A., Ahmed, S. T., Jagdish, T., Baxter, J. A., Sommerfeld, B. A., Tillman, A., Moore, J., Slonczewski, J. L., Barrick, J. E., & Lenski, R. E. (2020). Genomic and phenotypic evolution of Escherichia coli in a novel citrate-only resource environment. *ELife*, *9*(e55414), 1–64. https://doi.org/10.7554/eLife.55414

Botas, J., Rodríguez del Río, Á., Giner-Lamia, J., & Huerta-Cepas, J. (2022). GeCoViz: genomic context visualisation of prokaryotic genes from a functional and evolutionary perspective. *Nucleic Acids Research*, *50*(W1), W352–W357. https://doi.org/10.1093/nar/gkac367

Boulanger, E. F., Sabag-Daigle, A., Thirugnanasambantham, P., Gopalan, V., & Ahmer, B. M. M. (2021). Sugar-Phosphate Toxicities. *Microbiology and Molecular Biology Reviews*, *85*(4). https://doi.org/10.1128/MMBR.00123-21

Brandis, G. (2021). Reconstructing the Evolutionary History of a Highly Conserved Operon Cluster in Gammaproteobacteria and Bacilli. *Genome Biology and Evolution*, *13*(4). https://doi.org/10.1093/gbe/evab041

Brewster, R. C., Jones, D. L., & Phillips, R. (2012). Tuning Promoter Strength through RNA Polymerase Binding Site Design in Escherichia coli. *PLoS Computational Biology*, *8*(12), e1002811. https://doi.org/10.1371/journal.pcbi.1002811

Brigo, N., Grubwieser, P., Theurl, I., Nairz, M., Weiss, G., & Pfeifhofer-Obermair, C. (2023). Continuous Measurement of Reactive Oxygen Species Formation in Bacteria-infected Bone Marrow–derived Macrophages Using a Fluorescence Plate Reader. *BIO-PROTOCOL*, *13*(3). https://doi.org/10.21769/BioProtoc.4604

Brown, C. E., & Hogg, R. W. (1972). A Second Transport System for l -Arabinose in Escherichia coli Controlled by the araC Gene. *Journal of Bacteriology*, *111*(2), 606–613. https://doi.org/10.1128/jb.111.2.606-613.1972

Brown, R. P., & Feder, M. E. (2005). Reverse transcriptional profiling: non-correspondence of transcript level variation and proximal promoter polymorphism. *BMC Genomics*, *6*(110). https://doi.org/10.1186/1471-2164-6-110

Browning, D. F., & Busby, S. J. W. (2004). The regulation of bacterial transcription initiation. *Nature Reviews Microbiology*, *2*, 57–65. https://doi.org/10.1038/nrmicro787

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*, 59–60. https://doi.org/10.1038/nmeth.3176

Bundalovic-Torma, C., Whitfield, G. B., Marmont, L. S., Howell, P. L., & Parkinson, J. (2020). A systematic pipeline for classifying bacterial operons reveals the evolutionary landscape of biofilm machineries. *PLoS Computational Biology*, *16*(4), e1007721. https://doi.org/10.1371/journal.pcbi.1007721

Buvinger, W. E., & Riley, M. (1985). Nucleotide Sequence of Klebsiella pneumoniae lac Genes. *Journal of Bacteriology*, *163*(3), 850–857. http://jb.asm.org/content/163/3/850.full.pdf

Carroll, S. B. (2008). Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*, *134*, 25–36. https://doi.org/10.1016/j.cell.2008.06.030

Carruthers, T., Sun, M., Baker, W. J., Smith, S. A., de Vos, J. M., & Eiserhardt, W. L. (2022). The Implications of Incongruence between Gene Tree and Species Tree Topologies for Divergence Time Estimation. *Systematic Biology*, *71*(5), 1124–1146. https://doi.org/10.1093/sysbio/syac012

Casadevall, A., & Pirofski, L.-A. (1999). Host-Pathogen Interactions: Redefining the Basic Concepts of Virulence and Pathogenicity. *Infection and Immunity*, *67*(8), 3703–3713. https://doi.org/10.1128/iai.67.8.3703-3713.1999

Chen, J. Z., Fowler, D. M., & Tokuriki, N. (2022). Environmental selection and epistasis in an empirical phenotype–environment–fitness landscape. *Nature Ecology & Evolution*, *6*(4), 427–438. https://doi.org/10.1038/s41559-022-01675-5

Chernomor, O., von Haeseler, A., & Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*, *65*(6), 997–1008. https://doi.org/10.1093/sysbio/syw037

Chien, A., Edgar, D. B., & Trela, J. M. (1976). Deoxyribonucleic Acid Polymerase from the extreme thermophile Thermus aquaticus. *Journal of Bacteriology*, *127*(3), 1550–1557. https://doi.org/10.1128/JB.127.3.1550-1557.1976

Christaki, E., Marcou, M., & Tofarides, A. (2020). Antimicrobial Resistance in Bacteria: Mechanisms, Evolution, and Persistence. *Journal of Molecular Evolution*, *88*(1), 26–40. https://doi.org/10.1007/s00239-019-09914-3

Cornet, L., Magain, N., Baurain, D., & Lutzoni, F. (2021). Exploring syntenic conservation across genomes for phylogenetic studies of organisms subjected to horizontal gene transfers: A case study with Cyanobacteria and cyanolichens. *Molecular Phylogenetics and Evolution*, *162*, 107100. https://doi.org/10.1016/j.ympev.2021.107100

Correa, G. G., da Costa Ribeiro Lins, M. R., Silva, B. F., de Paiva, G. B., Zocca, V. F. B., Ribeiro, N. V., Picheli, F. P., Mack, M., & Pedrolli, D. B. (2020). Dataset for supporting a modular autoinduction device for control of gene expression in Bacillus subtilis. *Data in Brief*, *31*. https://doi.org/10.1016/j.dib.2020.105736

Cribbs, R., & Englesberg, E. (1964). L-arabinose negative mutants of the l-ribulokinase structural gene affecting the levels of l-arabinose isomerase in Escherichia coli. *Genetics*, *49*(1), 95–108. https://doi.org/10.1093/genetics/49.1.95

Davenport, E. R., Sanders, J. G., Song, S. J., Amato, K. R., Clark, A. G., & Knight, R. (2017). The human microbiome in evolution. *BMC Biology*, *15*(127). https://doi.org/10.1186/s12915-017-0454-7

Dawid, A., Kiviet, D. J., Kogenaru, M., de Vos, M., & Tans, S. J. (2010). Multiple peaks and reciprocal sign epistasis in an empirically determined genotype-phenotype landscape. *Chaos*, *20*(2). https://doi.org/10.1063/1.3453602

De, S., & Babu, M. M. (2010). Genomic neighbourhood and the regulation of gene expression. *Current Opinion in Cell Biology*, *22*(3), 326–333. https://doi.org/10.1016/j.ceb.2010.04.004

de Visser, J. A. G. M., & Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, *15*(7), 480–490. https://doi.org/10.1038/nrg3744

de Vos, M. G. J., Dawid, A., Sunderlikova, V., & Tans, S. J. (2015). Breaking evolutionary constraint with a tradeoff ratchet. *Proceedings of the National Academy of Sciences*, *112*(48), 14906–14911. https://doi.org/10.1073/pnas.1510282112

Denver, D. R., Morris, K., Streelman, J. T., Kim, S. K., Lynch, M., & Thomas, W. K. (2005). The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. *Nature Genetics*, *37*(5), 544–548. https://doi.org/10.1038/ng1554

Diard, M., & Hardt, W.-D. (2017). Evolution of bacterial virulence. *FEMS Microbiology Reviews*, *41*(5), 679–697. https://doi.org/10.1093/femsre/fux023

Ding, D., Green, A. G., Wang, B., Lite, T.-L. V., Weinstein, E. N., Marks, D. S., & Laub, M. T. (2022). Coevolution of interacting proteins through non-contacting and non-specific mutations. *Nature Ecology & Evolution*, *6*(5), 590–603. https://doi.org/10.1038/S41559-022-01688-0

Dlugaszewska, J., Leszczynska, M., Lenkowski, M., Tatarska, A., Pastusiak, T., & Szyfter, W. (2016). The pathophysiological role of bacterial biofilms in chronic sinusitis. *European Archives of Oto-Rhino-Laryngology*, *273*(8), 1989–1994. https://doi.org/10.1007/s00405-015-3650-5

Domingo, J., Baeza-Centurion, P., & Lehner, B. (2019). The Causes and Consequences of Genetic Interactions (Epistasis). *Annual Review of Genomics and Human Genetics*, *20*(1), 17.1-17.28. https://doi.org/10.1146/annurev-genom-083118-014857

Dover, G. A., & Flavell, R. B. (1984). Molecular coevolution: DNA divergence and the maintenance of function. *Cell*, *38*(3), 622–623. https://doi.org/10.1016/0092-8674(84)90255-1

Dunn, T. M., Hahn, S., Ogden, S., & Schleif, R. F. (1984). An operator at -280 base pairs that is required for repression of araBAD operon promoter: Addition of DNA helical turns between the operator and promoter cyclically hinders repression. *Proceedings of the National Academy of Sciences*, *81*(16), 5017–5020. https://doi.org/10.1073/pnas.81.16.5017

Englesberg, E., Irr, J., Power, J., & Lee, N. (1965). Positive control of enzyme synthesis by gene C in the L-arabinose system. *Journal of Bacteriology*, *90*(4), 946–957.

Ermolaeva, M. D., White, O., & Salzberg, S. L. (2001). Prediction of operons in microbial genomes. *Nucleic Acids Research*, *29*(5), 1216–1221. https://doi.org/10.1093/nar/29.5.1216

Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, *52*(2), 399–433. https://doi.org/10.1017/S0080456800012163

Flint, J., & Mackay, T. F. C. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Research*, *19*(5), 723. https://doi.org/10.1101/GR.086660.108

Fondi, M., Emiliani, G., & Fani, R. (2009). Origin and evolution of operons and metabolic pathways. *Research in Microbiology*, *160*(7), 502–512. https://doi.org/10.1016/j.resmic.2009.05.001

Fragata, I., Blanckaert, A., Dias Louro, M. A., Liberles, D. A., & Bank, C. (2019). Evolution in the light of fitness landscape theory. *Trends in Ecology & Evolution*, *34*(1), 69–82. https://doi.org/10.1016/j.tree.2018.10.009

Friedensohn, S., & Sawarkar, R. (2014). Cis-regulatory variation: significance in biomedicine and evolution. *Cell and Tissue Research*, *356*(3), 495–505. https://doi.org/10.1007/s00441-014-1855-3

Gant Kanegusuku, A., Stankovic, I. N., Cote-Hammarlof, P. A., Yong, P. H., & White-Ziegler, C. A. (2021). A Shift to Human Body Temperature (37°C) Rapidly Reprograms Multiple Adaptive Responses in Escherichia coli That Would Facilitate Niche Survival and Colonization. *Journal of Bacteriology*, *203*(22). https://doi.org/10.1128/JB.00363-21

Geertz, M., Shore, D., & Maerkl, S. J. (2012). Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. *Proceedings of the National Academy of Sciences*, *109*(41), 16540–16545. https://doi.org/10.1073/pnas.1206011109

Gilchrist, C. L. M., Booth, T. J., van Wersch, B., van Grieken, L., Medema, M. H., & Chooi, Y.-H. (2021). cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters. *Bioinformatics Advances*, *1*(1). https://doi.org/10.1093/bioadv/vbab016

Gilchrist, C. L. M., & Chooi, Y.-H. (2021). Clinker & clustermap.js: Automatic generation of gene cluster comparison figures. *Bioinformatics (Oxford, England)*, *37*(16), 2473–2475. https://doi.org/10.1093/BIOINFORMATICS/BTAB007

Gourse, R. L., Gaal, T., Bartlett, M. S., Appleman, J. A., & Ross, W. (1996). rRNA transcription and growth rate-dependent regulation of ribosome synthesis in Escherichia coli. *Annual Review of Microbiology*, *50*(1), 645–677. https://doi.org/10.1146/annurev.micro.50.1.645

Habteselassie, M. Y., Bischoff, M., Applegate, B., Reuhs, B., & Turco, R. F. (2010). Understanding the Role of Agricultural Practices in the Potential Colonization and Contamination by Escherichia coli in the Rhizospheres of Fresh Produce. *Journal of Food Protection*, *73*(11), 2001–2009. https://doi.org/10.4315/0362-028X-73.11.2001

Hahn, S. (2014). Ellis Englesberg and the Discovery of Positive Control in Gene Regulation. *Genetics*, *198*(2), 455–460. https://doi.org/10.1534/genetics.114.167361

Hall, B. G., Acar, H., Nandipati, A., & Barlow, M. (2013). Growth Rates Made Easy. *Molecular Biology and Evolution*, *31*(1), 232–238. https://doi.org/10.1093/molbev/mst187

Hall, B. K., Hallgrímsson, B., & Strickberger, M. W. (2014). *Strickberger's Evolution* (5th ed.). Jones & Bartlett Publishers. https://books.google.co.uk/books?id=jrDD3cyA09kC&pg=PA4&redir_esc=y#v=onepage&q&f=false

Hammarlöf, D. L., Kröger, C., Owen, S. V., Canals, R., Lacharme-Lora, L., Wenner, N., Schager, A. E., Wells, T. J., Henderson, I. R., Wigley, P., Hokamp, K., Feasey, N. A., Gordon, M. A., & Hinton, J. C. D. (2018). Role of a single noncoding nucleotide in the evolution of an epidemic African clade of Salmonella. *Proceedings of the National Academy of Sciences*, *115*(11). https://doi.org/10.1073/pnas.1714718115

Harrison, R., Papp, B., Pál, C., Oliver, S. G., & Delneri, D. (2007). Plasticity of genetic interactions in metabolic networks of yeast. *Proceedings of the National Academy of Sciences*, *104*(7), 2307–2312. https://doi.org/10.1073/pnas.0607153104

Hartl, D. L., & Dykhuizen, D. E. (1984). The population genetics of Escherichia coli. *Annual Review of Genetics*, *18*(1), 31–68. https://doi.org/10.1146/annurev.ge.18.120184.000335

Hazkani-Covo, E., & Graur, D. (2005). Evolutionary conservation of bacterial operons: does transcriptional connectivity matter? *Genetica*, *124*, 145–166. https://doi.org/10.1007/s10709-005-0950-5

Hibbing, M. E., Fuqua, C., Parsek, M. R., & Peterson, S. B. (2010). Bacterial competition: surviving and thriving in the microbial jungle. *Nature Reviews Microbiology*, *8*(1), 15–25. https://doi.org/10.1038/nrmicro2259

Hogg, R. W. (1977). L-arabinose transport and the L-arabinose binding protein of escherichia coli. *Journal of Supramolecular Structure*, *6*(3), 411–417. https://doi.org/10.1002/jss.400060314

Holtzapple, M. T. (2003). HEMICELLULOSES. In *Encyclopedia of Food Sciences and Nutrition* (pp. 3060–3071). Elsevier. https://doi.org/10.1016/B0-12-227055-X/00589-7

Horazdovsky, B. F., & Hogg, R. W. (1987). High-affinity l-arabinose transport operon. *Journal of Molecular Biology*, *197*(1), 27–35. https://doi.org/10.1016/0022-2836(87)90606-1

Islam, Md. S., Pallen, M. J., & Busby, S. J. W. (2011). A cryptic promoter in the LEE1 regulatory region of enterohaemorrhagic Escherichia coli: promoter specificity in AT-rich gene regulatory regions. *Biochemical Journal*, *436*(3), 681–686. https://doi.org/10.1042/BJ20110260

Itoh, T., Takemoto, K., Mori, H., & Gojobort, T. (1999). Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Molecular Biology and Evolution*, *16*(3), 332–346. https://doi.org/10.1093/oxfordjournals.molbev.a026114

Jacob, F., & Monod, J. (1961). Gene Regulatory Mechanisms in Protein Synthesis. *Journal of Molecular Biology*, *3*(3), 318–356.

Jacob, F., Perrin, D., Sanchez, C., & Monod, J. (1960). [Operon: a group of genes with the expression coordinated by an operator]. *Comptes Rendus Hebdomadaires Des Seances de l'Academie Des Sciences*, *250*, 1727–1729. https://pubmed.ncbi.nlm.nih.gov/14406329/

Jacob, F., Perrin, D., Sánchez, C., & Monod, J. (2005). L'opéron : groupe de gènes à expression coordonnée par un opérateur [C. R. Acad. Sci. Paris 250 (1960) 1727–1729]. *Comptes Rendus Biologies*, *328*(6), 514–520. https://doi.org/10.1016/j.crvi.2005.04.005

Jiang, X., Hall, A. B., Arthur, T. D., Plichta, D. R., Covington, C. T., Poyet, M., Crothers, J., Moses, P. L., Tolonen, A. C., Vlamakis, H., Alm, E. J., & Xavier, R. J. (2019). Invertible promoters mediate bacterial phase variation, antibiotic resistance, and host adaptation in the gut. *Science*, *363*(6423), 181–187. https://doi.org/10.1126/science.aau5238

Joshi, M., Kapopoulou, A., & Laurent, S. (2021). Impact of Genetic Variation in Gene Regulatory Sequences: A Population Genomics Perspective. *Frontiers in Genetics*, *12*, 660899. https://doi.org/10.3389/fgene.2021.660899

Junier, I., & Rivoire, O. (2016). Conserved Units of Co-Expression in Bacterial Genomes: An Evolutionary Insight into Transcriptional Regulation. *PLOS ONE*, *11*(5), e0155740. https://doi.org/10.1371/journal.pone.0155740

Kacser, H., & Burns, J. A. (1981). The Molecular Basis of Dominance. *Genetics*, *97*(3–4), 639–666. https://doi.org/10.1093/genetics/97.3-4.639

Kambourova, M. (2018). Thermostable enzymes and polysaccharides produced by thermophilic bacteria isolated from Bulgarian hot springs. *Engineering in Life Sciences*, *18*(11), 758–767. https://doi.org/10.1002/elsc.201800022

Karkare, K., Lai, H. Y., Azevedo, R. B. R., & Cooper, T. F. (2021). Historical Contingency Causes Divergence in Adaptive Expression of the lac Operon. *Molecular Biology and Evolution*, *38*(7), 2869–2879. https://doi.org/10.1093/MOLBEV/MSAB077

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Kazan, D., Çamurdan, A., & Hortaçsu, A. (1995). The effect of glucose concentration on the growth rate and some intracellular components of a recombinant E. coli culture. *Process Biochemistry*, *30*(3), 269–273. https://doi.org/10.1016/0032-9592(95)85008-2

Kerwin, R. E., Feusier, J., Muok, A., Lin, C., Larson, B., Copeland, D., Corwin, J. A., Rubin, M. J., Francisco, M., Li, B., Joseph, B., Weinig, C., & Kliebenstein, D. J. (2017). Epistasis × environment interactions among Arabidopsis thaliana glucosinolate genes impact complex traits and fitness in the field. *New Phytologist*, *215*(3), 1249–1263. https://doi.org/10.1111/nph.14646

Kim, H., Kwak, W., Yoon, S. H., Kang, D.-K., & Kim, H. (2018). Horizontal gene transfer of Chlamydia: Novel insights from tree reconciliation. *PLOS ONE*, *13*(4), e0195139. https://doi.org/10.1371/journal.pone.0195139

King, M.-C., & Wilson, A. C. (1975). Evolution at Two Levels in Humans and Chimpanzees. *Science*, *188*(4184), 107–116. https://doi.org/10.1126/science.1090005

Kinney, J. B., Murugan, A., Callan, C. G., & Cox, E. C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*, *107*(20), 9158–9163. https://doi.org/10.1073/pnas.1004290107

Knijnenburg, T. A., Daran, J. M. G., van den Broek, M. A., Daran-Lapujade, P. A. S., de Winde, J. H., Pronk, J. T., Reinders, M. J. T., & Wessels, L. F. A. (2009). Combinatorial effects of environmental parameters on transcriptional regulation in Saccharomyces cerevisiae: A quantitative analysis of a compendium of chemostat-based transcriptome data. *BMC Genomics*, *10*(53). https://doi.org/10.1186/1471-2164-10-53

Kolodrubetz, D., & Schleif, R. (1981). L-arabinose transport systems in Escherichia coli K-12. *Journal of Bacteriology*, *148*(2), 472–479. https://doi.org/10.1128/JB.148.2.472-479.1981

Koonin, E. V. (2009). Evolution of Genome Architecture. *International Journal of Biochemistry and Cell Biology*, *41*(2), 298–306. https://doi.org/10.1016/j.biocel.2008.09.015

Koonin, E. V., Makarova, K. S., & Aravind, L. (2001). Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. *Annual Review of Microbiology*, *55*(1), 709–742. https://doi.org/10.1146/annurev.micro.55.1.709

Kuhner, M. K., & Felsenstein, J. (1994). A simulation Comparison of Phylogeny Algorithms under Equal and Unequal Evolutionary Rates. *Molecular Biology and Evolution*, *11*(3), 459–468. https://doi.org/10.1093/oxfordjournals.molbev.a040126

Kunonga, N. I., Sobieski, R. J., & Crupper, S. S. (2000). Prevalence of the multiple antibiotic resistance operon (marRAB) in the genus Salmonella. *FEMS Microbiology Letters*, *187*(2), 155–160. https://doi.org/10.1111/J.1574-6968.2000.TB09153.X

Kuznetsova, M. V, Maslennikova, I. L., Karpunina, T. I., Nesterova, L. Y., & Demakov, V. A. (2013). Interactions of Pseudomonas aeruginosa in predominant biofilm or planktonic forms of existence in mixed culture with Escherichia coli in vitro. *Canadian Journal of Microbiology*, *59*(9), 604–610. https://doi.org/10.1139/cjm-2013-0168

Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C., & Cohen, B. A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proceedings of the National Academy of Sciences*, *109*(47), 19498–19503. https://doi.org/10.1073/pnas.1210678109

Lagator, M., Igler, C., Moreno, A. B., Guet, C. C., & Bollback, J. P. (2016). Epistatic Interactions in the Arabinose Cis-Regulatory Element. *Molecular Biology and Evolution*, *33*(3), 761–769. https://doi.org/10.1093/molbev/msv269

Lagator, M., Paixão, T., Barton, N. H., Bollback, J. P., & Guet, C. C. (2017). On the mechanistic nature of epistasis in a canonical cis-regulatory element. *ELife*, *6*(e25192). https://doi.org/10.7554/eLife.25192

Lagator, M., Sarikas, S., Acar, H., Bollback, J. P., & Guet, C. C. (2017). Regulatory network structure determines patterns of intermolecular epistasis. *ELife*, *6*(e28921). https://doi.org/10.7554/eLife.28921

Lahti, D. C., Johnson, N. A., Ajie, B. C., Otto, S. P., Hendry, A. P., Blumstein, D. T., Coss, R. G., Donohue, K., & Foster, S. A. (2009). Relaxed selection in the wild. *Trends in Ecology & Evolution*, *24*(9), 487–496. https://doi.org/10.1016/j.tree.2009.03.010

Lawrence, J. G. (1997). Selfish operons and speciation by gene transfer. In *Trends in Microbiology* (Vol. 5, Issue 9, pp. 355–359). https://doi.org/10.1016/S0966-842X(97)01110-4

Lawrence, J. G., & Roth, J. R. (1996). Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters. *Genetics*, *143*(4), 1843–1860. http://www.ncbi.nlm.nih.gov/pubmed/8844169

Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*, *27*(8), 323–331. https://doi.org/10.1016/j.tig.2011.05.007

Leonard, S. R., Lacher, D. W., & Lampel, K. A. (2015). Acquisition of the lac operon by Salmonella enterica. *BMC Microbiology*, *15*(173). https://doi.org/10.1186/s12866-015-0511-8

Lewis, M. W., Li, S., & Franco, H. L. (2019). Transcriptional control by enhancers and enhancer RNAs. *Transcription*, *10*(4–5), 171–186. https://doi.org/10.1080/21541264.2019.1695492

Li, C., & Zhang, J. (2018). Multi-environment fitness landscapes of a tRNA gene. *Nature Ecology & Evolution*, *2*(6), 1025–1032. https://doi.org/10.1038/s41559-018-0549-8

Liu, Y., Beyer, A., & Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, *165*(3), 535–550. https://doi.org/10.1016/j.cell.2016.03.014

Lovell, S. C., & Robertson, D. L. (2010). An Integrated View of Molecular Coevolution in Protein-Protein Interactions. *Molecular Biology and Evolution*, *27*(11), 2567–2575. https://doi.org/10.1093/molbev/msq144

Luo, Y., Zhang, T., & Wu, H. (2014). The transport and mediation mechanisms of the common sugars in Escherichia coli. *Biotechnology Advances*, *32*(5), 905–919. https://doi.org/10.1016/j.biotechadv.2014.04.009

Lutz, R., & Bujard, H. (1997). Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research*, *25*(6), 1203–1210. https://doi.org/10.1093/nar/25.6.1203

Macpherson, A.J.S., Jones-Mortimer, M. C., & Henderson, P. J. F. (1981). Identification of the AraE transport protein of Escherichia coli. *The Biochemical Journal*, *196*(1), 269–283. https://doi.org/10.1042/BJ1960269

Madigan, M. T., Martinko, J. M., Stahl, D. A., & Clark, D. P. (2012). Brock Biology of Microorganisms. In *Notes and Queries* (13th ed, Vols s3-XII, Issue 310). Pearson. https://doi.org/10.1093/nq/s3-XII.310.469-a

Maerkl, S. J., & Quake, S. R. (2007). A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. *Science*, *315*(5809), 233–237. https://doi.org/10.1126/science.1131007

Mairhofer, J., Scharl, T., Marisch, K., Cserjan-Puschmann, M., & Striedner, G. (2013). Comparative transcription profiling and in-depth characterization of plasmid-based and plasmid-free Escherichia coli expression systems under production conditions. *Applied and Environmental Microbiology*, *79*(12), 3802–3812. https://doi.org/10.1128/AEM.00365-13

Maisnier-Patin, S., Paulander, W., Pennhag, A., & Andersson, D. I. (2007). Compensatory Evolution Reveals Functional Interactions between Ribosomal Proteins S12, L14 and L19. *Journal of Molecular Biology*, *366*(1), 207–215. https://doi.org/10.1016/J.JMB.2006.11.047

Mani, R., St. Onge, R. P., Hartman IV, J. L., Giaever, G., & Roth, F. P. (2008). Defining genetic interaction. *Proceedings of the National Academy of Sciences*, *105*(9), 3461. https://doi.org/10.1073/PNAS.0712255105

Marsh, C. L., & Larsen, D. H. (1953). Characterization of Some Thermophilic Bacteria From The Hot Springs Of Yellowstone National Park. *Journal of Bacteriology*, *65*(2), 193–197. https://doi.org/10.1128/jb.65.2.193-197.1953

Mavrommati, M., Daskalaki, A., Papanikolaou, S., & Aggelis, G. (2022). Adaptive laboratory evolution principles and applications in industrial biotechnology. *Biotechnology Advances*, *54*, 107795. https://doi.org/10.1016/j.biotechadv.2021.107795

McLoughlin, S. Y., & Copley, S. D. (2008). A compromise required by gene sharing enables survival: Implications for evolution of new enzyme activities. *Proceedings of the National Academy of Sciences*, *105*(36), 13497–13502. https://doi.org/10.1073/pnas.0804804105

Muñoz, E., Park, J.-M., & Deem, M. W. (2008). Quasispecies theory for Horizontal Gene Transfer and Recombination. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, *78*(6 0 1), 061921. https://doi.org/10.1103/PHYSREVE.78.061921

Nagai, T., Ibata, K., Park, E. S., Kubota, M., Mikoshiba, K., & Miyawaki, A. (2002). A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nature Biotechnology*, *20*(1), 87–90. https://doi.org/10.1038/NBT0102-87

Naidoo, T., Sjödin, P., Schlebusch, C., & Jakobsson, M. (2018). Patterns of variation in cis-regulatory regions: examining evidence of purifying selection. *BMC Genomics*, *19*(95). https://doi.org/10.1186/s12864-017-4422-y

Nghe, P., Kogenaru, M., & Tans, S. J. (2018). Sign epistasis caused by hierarchy within signalling cascades. *Nature Communications*, *9*(1451). https://doi.org/10.1038/s41467-018-03644-8

Niland, P., Hühne, R., & Müller-Hill, B. (1996). How AraC interacts specifically with its target DNAs. *Journal of Molecular Biology*, *264*(4), 667–674. https://doi.org/10.1006/jmbi.1996.0668

Nomura, M. (1999). Regulation of ribosome biosynthesis in Escherichia coli and Saccharomyces cerevisiae: diversity and common principles. *Journal of Bacteriology*, *181*(22), 6857–6864. https://doi.org/10.1128/JB.181.22.6857-6864.1999

Noutahi, E., & El-Mabrouk, N. (2018). GATC: a genetic algorithm for gene tree construction under the Duplication-Transfer-Loss model of evolution. *BMC Genomics*, *19*(Suppl 2), 102. https://doi.org/10.1186/s12864-018-4455-x

Novick, A., & Weiner, M. (1957). Enzyme Induction as an All-or-None Phenomenon. *Proceedings of the National Academy of Sciences*, *43*(7), 553–566. https://doi.org/10.1073/pnas.43.7.553

Olendzenski, L., & Gogarten, J. P. (2009). Evolution of Genes and Organisms. *Annals of the New York Academy of Sciences*, *1178*(1), 137–145. https://doi.org/10.1111/j.1749-6632.2009.04998.x

Omelchenko, M. V, Makarova, K. S., Wolf, Y. I., Rogozin, I. B., & Koonin, E. V. (2003). Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biology*, *4*(9), R55. https://doi.org/10.1186/gb-2003-4-9-r55

Ortlund, E. A., Bridgham, J. T., Redinbo, M. R., & Thornton, J. W. (2007). Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. *Science*, *317*(5844), 1544–1548. https://doi.org/10.1126/science.1142819

Osada, N., Miyagi, R., & Takahashi, A. (2017). Cis - and Trans-regulatory Effects on Gene Expression in a Natural Population of Drosophila melanogaster. *Genetics*, *206*(4), 2139–2148. https://doi.org/10.1534/genetics.117.201459

Osbourn, A. E., & Field, B. (2009). Operons. *Cellular and Molecular Life Sciences*, *66*(23), 3755–3775. https://doi.org/10.1007/s00018-009-0114-3

Pál, C., & Hurst, L. D. (2004). Evidence against the selfish operon theory. In *Trends in Genetics* (Vol. 20, Issue 6, pp. 232–234). https://doi.org/10.1016/j.tig.2004.04.001

Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'Er, D., & Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnology*, *27*(12), 1173–1175. https://doi.org/10.1038/nbt.1589

Paulander, W., Andersson, D. I., & Maisnier-Patin, S. (2010). Amplification of the Gene for Isoleucyl–tRNA Synthetase Facilitates Adaptation to the Fitness Cost of Mupirocin Resistance in Salmonella enterica. *Genetics*, *185*(1), 305–312. https://doi.org/10.1534/genetics.109.113514

Pedruzzi, G., Barlukova, A., & Rouzine, I. M. (2018). Evolutionary footprint of epistasis. *PLoS Computational Biology*, *14*(9). https://doi.org/10.1371/JOURNAL.PCBI.1006426

Phillips, P. C. (2008). Epistasis - The essential role of gene interactions in the structure and evolution of genetic systems. In *Nature Reviews Genetics* (Vol. 9, Issue 11, pp. 855–867). https://doi.org/10.1038/nrg2452

Planet, P. J. (2006). Tree disagreement: Measuring and testing incongruence in phylogenies. *Journal of Biomedical Informatics*, *39*(1), 86–102. https://doi.org/10.1016/j.jbi.2005.08.008

Poelwijk, F. J., Tănase-Nicola, S., Kiviet, D. J., & Tans, S. J. (2011). Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *Journal of Theoretical Biology*, *272*(1), 141–144. https://doi.org/10.1016/j.jtbi.2010.12.015

Poon, A., & Chao, L. (2005). The Rate of Compensatory Mutation in the DNA Bacteriophage ϕX174. *Genetics*, *170*(3), 989–999. https://doi.org/10.1534/genetics.104.039438

Poyatos, J. F. (2020). Genetic buffering and potentiation in metabolism. *PLOS Computational Biology*, *16*(9), e1008185. https://doi.org/10.1371/journal.pcbi.1008185

Price, M. N., Arkin, A. P., & Alm, E. J. (2006). The Life-Cycle of Operons. *PLoS Genetics*, *2*(6), e96. https://doi.org/10.1371/journal.pgen.0020096

Price, M. N., Huang, K. H., Alm, E. J., & Arkin, A. P. (2005). A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Research*, *33*(3), 880–892. https://doi.org/10.1093/nar/gki232

Price, M. N., Huang, K. H., Arkin, A. P., & Alm, E. J. (2005). Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Research*, *15*(6), 809–819. https://doi.org/10.1101/gr.3368805

Ram, Y., Dellus-Gur, E., Bibi, M., Karkare, K., Obolski, U., Feldman, M. W., Cooper, T. F., Berman, J., & Hadany, L. (2019). Predicting microbial growth in a mixed culture from growth curve data.

Proceedings of the National Academy of Sciences, *116*(29), 14698–14707. https://doi.org/10.1073/PNAS.1902217116/SUPPL_FILE/PNAS.1902217116.SAPP.PDF

Ramakrishnan, A. P. (2013). Linkage Disequilibrium. In *Brenner's Encyclopedia of Genetics: Second Edition* (Second Edition). Academic Press. https://doi.org/10.1016/B978-0-12-374984-0.00870-6

Reeve, M. A., & Fuller, C. W. (1995). A novel thermostable polymerase for DNA sequencing. *Nature*, *376*(6543), 796–797. https://doi.org/10.1038/376796a0

Reid, S. J., & Abratt, V. R. (2005). Sucrose utilisation in bacteria: genetic organisation and regulation. *Applied Microbiology and Biotechnology*, *67*(3), 312–321. https://doi.org/10.1007/s00253-004-1885-y

Remold, S. K., & Lenski, R. E. (2004). Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli. *Nature Genetics*, *36*(4), 423–426. https://doi.org/10.1038/ng1324

Rocha, E. P. C. (2006). Inference and Analysis of the Relative Stability of Bacterial Chromosomes. *Molecular Biology and Evolution*, *23*(3), 513–522. https://doi.org/10.1093/molbev/msj052

Rocha, E. P. C. (2008). The Organization of the Bacterial Genome. *Annual Review of Genetics*, *42*(1), 211–233. https://doi.org/10.1146/annurev.genet.42.110807.091653

Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., Szekely, L. A., & Koonin, E. V. (2002). Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Research*, *30*(10), 2212–2223. https://doi.org/10.1093/nar/30.10.2212

Rolfe, M. D., Rice, C. J., Lucchini, S., Pin, C., Thompson, A., Cameron, A. D. S., Alston, M., Stringer, M. F., Betts, R. P., Baranyi, J., Peck, M. W., & Hinton, J. C. D. (2012). Lag phase is a distinct growth phase that prepares bacteria for exponential growth and involves transient metal accumulation. *Journal of Bacteriology*, *194*(3), 686–701. https://doi.org/10.1128/JB.06112-11

Rolland, T., Neuvéglise, C., Sacerdot, C., & Dujon, B. (2009). Insertion of Horizontally Transferred Genes within Conserved Syntenic Regions of Yeast Genomes. *PLoS ONE*, *4*(8), e6515. https://doi.org/10.1371/journal.pone.0006515

Romeo, T., Vakulskas, C. A., & Babitzke, P. (2013). Post-transcriptional regulation on a global scale: form and function of Csr/Rsm systems. *Environmental Microbiology*, *15*(2), 313–324. https://doi.org/10.1111/j.1462-2920.2012.02794.x

Romero, P. A., & Arnold, F. H. (2009). Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, *10*(12), 866–876. https://doi.org/10.1038/nrm2805

Ruddle, S. J., Massis, L. M., Cutter, A. C., & Monack, D. M. (2023). Salmonella-liberated dietary L-arabinose promotes expansion in superspreaders. *Cell Host & Microbe*, *31*(3), 405-417.e5. https://doi.org/10.1016/j.chom.2023.01.017

Sackman, A. M., & Rokyta, D. R. (2018). Additive Phenotypes Underlie Epistasis of Fitness Effects. *Genetics*, *208*(1), 339–348. https://doi.org/10.1534/genetics.117.300451

Salgado, H., Moreno-Hagelsieb, G., Smith, T. F., & Collado-Vides, J. (2000). Operons in Escherichia coli: genomic analyses and predictions. *Proceedings of the National Academy of Sciences*, *97*(12), 6652–6657. https://doi.org/10.1073/pnas.110147297

Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millán-Zárate, D., Díaz-Peredo, E., Sánchez-Solano, F., Pérez-Rueda, E., Bonavides-Martínez, C., & Collado-Vides, J. (2001). RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Research*, *29*(1), 72–74. https://doi.org/10.1093/NAR/29.1.72

Samir, P., Rahul, Slaughter, J. C., & Link, A. J. (2015). Environmental Interactions and Epistasis Are Revealed in the Proteomic Responses to Complex Stimuli. *PLOS ONE*, *10*(8), e0134099. https://doi.org/10.1371/journal.pone.0134099

Sarkisyan, K. S., Bolotin, D. A., Meer, M. V, Usmanova, D. R., Mishin, A. S., Sharonov, G. V, Ivankov, D. N., Bozhanova, N. G., Baranov, M. S., Soylemez, O., Bogatyreva, N. S., Vlasov, P. K., Egorov, E. S., Logacheva, M. D., Kondrashov, A. S., Chudakov, D. M., Putintseva, E. V, Mamedov, I. Z., Tawfik, D. S., … Kondrashov, F. A. (2016). Local fitness landscape of the green fluorescent protein. *Nature*, *533*(7603), 397–401. https://doi.org/10.1038/nature17995

Savageau, M. A. (1983). Escherichia coli Habitats, Cell Types, and Molecular Mechanisms of Gene Control. *Https://Doi.Org/10.1086/284168*, *122*(6), 732–744. https://doi.org/10.1086/284168

Saviola, B., Seabold, R., & Schleif, R. F. (1998). Arm-Domain Interactions in AraC. *Journal of Molecular Biology*, *278*(3), 539–548. https://doi.org/10.1006/jmbi.1998.1712

Schaechter, M., MaalØe, O., & Kjeldgaard, N. O. (1958). Dependency on Medium and Temperature of Cell Size and Chemical Composition during Balanced Growth of Salmonella typhimurium. *Journal of General Microbiology*, *19*(3), 592–606. https://doi.org/10.1099/00221287-19-3-592

Schleif, R. (2000). Regulation of the L-arabinose operon of Escherichia coli. *Trends in Genetics*, *16*(12), 559–565. https://doi.org/10.1016/S0168-9525(00)02153-3

Schleif, R. (2010). AraC protein, regulation of the L-arabinose operon in Escherichia coli , and the light switch mechanism of AraC action. *FEMS Microbiology Reviews*, *34*(5), 779–796. https://doi.org/10.1111/j.1574-6976.2010.00226.x

Schleif, R. (2022). A Career's Work, the l-Arabinose Operon: How It Functions and How We Learned It. *EcoSal Plus*, *10*(1), eESP00122021. https://doi.org/10.1128/ecosalplus.ESP-0012-2021

Sevillya, G., Adato, O., & Snir, S. (2020). Detecting horizontal gene transfer: a probabilistic approach. *BMC Genomics*, *21*(S1), 106. https://doi.org/10.1186/s12864-019-6395-5

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., & Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology*, *30*(6), 521–530. https://doi.org/10.1038/nbt.2205

Shortle, D., & Lin, B. (1985). Genetic Analysis of Staphylococcal Nuclease: Identification of Three Intragenic 'Global' Suppressors of Nuclease-Minus Mutations. *Genetics*, *110*(4), 539–555. https://doi.org/10.1093/genetics/110.4.539

Siepielski, A. M., DiBattista, J. D., & Carlson, S. M. (2009). It's about time: the temporal dynamics of phenotypic selection in the wild. *Ecology Letters*, *12*(11), 1261–1276. https://doi.org/10.1111/j.1461-0248.2009.01381.x

Signor, S. A., & Nuzhdin, S. V. (2018). The Evolution of Gene Expression in cis and trans. *Trends in Genetics*, *34*(7), 532–544. https://doi.org/10.1016/j.tig.2018.03.007

Smits, W. K., Kuipers, O. P., & Veening, J.-W. (2006). Phenotypic variation in bacteria: the role of feedback regulation. *Nature Reviews Microbiology*, *4*(4), 259–271. https://doi.org/10.1038/nrmicro1381

Snir, S. (2016). Ordered orthology as a tool in prokaryotic evolutionary inference. *Mobile Genetic Elements*, *6*(6), e1120576. https://doi.org/10.1080/2159256X.2015.1120576

Som, A. (2015). Causes, consequences and solutions of phylogenetic incongruence. *Briefings in Bioinformatics*, *16*(3), 536–548. https://doi.org/10.1093/BIB/BBU015

Srinivasan, R., Chandraprakash, D., Krishnamurthi, R., Singh, P., Scolari, V. F., Krishna, S., & Seshasayee, A. S. N. (2013). Genomic analysis reveals epistatic silencing of 'expensive' genes in Escherichia coli K-12. *Molecular BioSystems*, *9*(8), 2021–2033. https://doi.org/10.1039/c3mb70035f

Steinberg, B., & Ostermeier, M. (2016). Environmental changes bridge evolutionary valleys. *Science Advances*, *2*(1). https://doi.org/10.1126/sciadv.1500921

Stern, D. L., & Orgogozo, V. (2008). The Loci of Evolution: How Predictable is Genetic Evolution? *Evolution*, *62*(9), 2155–2177. https://doi.org/10.1111/j.1558-5646.2008.00450.x

Sutherland, K. M., Ward, L. M., Colombero, C. -R., & Johnston, D. T. (2021). Inter-domain horizontal gene transfer of nickel-binding superoxide dismutase. *Geobiology*, *19*(5), 450–459. https://doi.org/10.1111/gbi.12448

Svetlitsky, D., Dagan, T., & Ziv-Ukelson, M. (2020). Discovery of multi-operon colinear syntenic blocks in microbial genomes. *Bioinformatics*, *36*, i21–i29. https://doi.org/10.1093/bioinformatics/btaa503

Takai, K., Nakamura, K., Toki, T., Tsunogai, U., Miyazaki, M., Miyazaki, J., Hirayama, H., Nakagawa, S., Nunoura, T., & Horikoshi, K. (2008). Cell proliferation at 122°C and isotopically heavy CH4 production by a hyperthermophilic methanogen under high-pressure cultivation. *Proceedings of the National Academy of Sciences*, *105*(31), 10949–10954. https://doi.org/10.1073/PNAS.0712334105

Thoday, J. M. (1953). Components of fitness. *In Symposia of the Society for Experimental Biology*, *7*, 96–112.

Tomioka, S., Seki, N., Sugiura, Y., Akiyama, M., Uchiyama, J., Yamaguchi, G., Yakabe, K., Ejima, R., Hattori, K., Kimizuka, T., Fujimura, Y., Sato, H., Gondo, M., Ozaki, S., Honme, Y., Suematsu, M., Kimura, I., Inohara, N., Núñez, G., … Kim, Y.-G. (2022). Cooperative action of gut-microbiota-accessible carbohydrates improves host metabolic function. *Cell Reports*, *40*(3), 111087. https://doi.org/10.1016/j.celrep.2022.111087

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D. W., Corander, J., Bentley, S. D., & Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, *21*(180), 1–21. https://doi.org/10.1186/S13059-020-02090-4/FIGURES/7

Trivedi, P., Leach, J. E., Tringe, S. G., Sa, T., & Singh, B. K. (2020). Plant–microbiome interactions: from community assembly to plant health. *Nature Reviews Microbiology*, *18*(11), 607–621. https://doi.org/10.1038/s41579-020-0412-1

Unterholzner, S. J., Poppenberger, B., & Rozhon, W. (2013). Toxin–antitoxin systems. *Mobile Genetic Elements*, *3*(5), e26219. https://doi.org/10.4161/mge.26219

van Elsas, J. D., Semenov, A. V, Costa, R., & Trevors, J. T. (2011). Survival of Escherichia coli in the environment: fundamental and public health aspects. *The ISME Journal*, *5*(2), 173–183. https://doi.org/10.1038/ismej.2010.80

V'yugin, V. V., Gelfand, M. S., & Lyubetsky, V. A. (2003). Identification of Horizontal Gene Transfer from Phylogenetic Gene Trees. *Molecular Biology 2003 37:4*, *37*(4), 571–584. https://doi.org/10.1023/A:1025191411933

Waclaw, B. (2016). Evolution of Drug Resistance in Bacteria. In *Advances in Experimental Medicine and Biology* (Vol. 915, pp. 49–67). Adv Exp Med Biol. https://doi.org/10.1007/978-3-319-32189-9_5

Wang, X., Xia, K., Yang, X., & Tang, C. (2019). Growth strategy of microbes on mixed carbon sources. *Nature Communications*, *10*(1), 1279. https://doi.org/10.1038/s41467-019-09261-3

Wang, Z., Ji, X., Wang, S., Wu, Q., & Xu, Y. (2021). Sugar profile regulates the microbial metabolic diversity in Chinese Baijiu fermentation. *International Journal of Food Microbiology*, *359*, 109426. https://doi.org/10.1016/j.ijfoodmicro.2021.109426

Weinreich, D. M., Lan, Y., Jaffe, J., & Heckendorn, R. B. (2018). The Influence of Higher-Order Epistasis on Biological Fitness Landscape Topography. *Journal of Statistical Physics*, *172*(1), 208–225. https://doi.org/10.1007/s10955-018-1975-3

Weinreich, D. M., Watson, R. A., & Chao, L. (2005). Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution; International Journal of Organic Evolution*, *59*(6), 1165–1174. https://doi.org/10.1554/04-272

Wen, Z., Liu, Y., Qu, F., & Zhang, J.-R. (2016). Allelic Variation of the Capsule Promoter Diversifies Encapsulation and Virulence In Streptococcus pneumoniae. *Scientific Reports*, *6*(1), 30176. https://doi.org/10.1038/srep30176

Wiser, M. J., & Lenski, R. E. (2015). A Comparison of Methods to Measure Fitness in Escherichia coli. *PLOS ONE*, *10*(5), e0126210. https://doi.org/10.1371/journal.pone.0126210

Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, *13*(1), 59–69. https://doi.org/10.1038/nrg3095

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews*, *51*(2), 221–271. https://doi.org/10.1128/mr.51.2.221-271.1987

Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S., & Koonin, E. V. (2001). Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Research*, *11*(3), 356–372. https://doi.org/10.1101/gr.GR-1619R

Wong, A. (2017). Epistasis and the Evolution of Antimicrobial Resistance. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.00246

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, *8*(3), 206–216. https://doi.org/10.1038/nrg2063

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics*, *1*, 356–366.

Xi, Z., Liu, L., & Davis, C. C. (2015). Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. *Molecular Phylogenetics and Evolution*, *92*, 63–71. https://doi.org/10.1016/j.ympev.2015.06.009

Yang, T.-Y., Sung, Y.-M., Lei, G.-S., Romeo, T., & Chak, K.-F. (2010). Posttranscriptional repression of the cel gene of the ColE7 operon by the RNA-binding protein CsrA of Escherichia coli. *Nucleic Acids Research*, *38*(12), 3936–3951. https://doi.org/10.1093/nar/gkq177

You, L., & Yin, J. (2002). Dependence of Epistasis on Environment and Mutation Severity as Revealed by in Silico Mutagenesis of Phage T7. *Genetics*, *160*(4), 1273–1281. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1462038/pdf/11973286.pdf

Young, R. S., Talmane, L., Marion de Procé, S., & Taylor, M. S. (2022). The contribution of evolutionarily volatile promoters to molecular phenotypes and human trait variation. *Genome Biology*, *23*(1), 89. https://doi.org/10.1186/s13059-022-02634-w

Zhang, H., Yu, T., Wang, Y., Li, J., Wang, G., Ma, Y., & Liu, Y. (2018). 4-Chlorophenol Oxidation Depends on the Activation of an AraC-Type Transcriptional Regulator, CphR, in Rhodococcus sp. Strain YH-5B. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.02481

Zheng, X., Zhu, K., Sun, Q., Zhang, W., Wang, X., Cao, H., Tan, M., Xie, Z., Zeng, Y., Ye, J., Chai, L., Xu, Q., Pan, Z., Xiao, S., Fraser, P. D., & Deng, X. (2019). Natural Variation in CCD4 Promoter Underpins Species-Specific Evolution of Red Coloration in Citrus Peel. *Molecular Plant*, *12*(9), 1294–1307. https://doi.org/10.1016/j.molp.2019.04.014

Zylstra, G. J., McCombie, W. R., Gibson, D. T., & Finette, B. A. (1988). Toluene degradation by Pseudomonas putida F1: genetic organization of the tod operon. *Applied and Environmental Microbiology*, *54*(6), 1498–1503. https://doi.org/10.1128/AEM.54.6.1498-1503.1988