PHD

The causes and consequences of a mutational hotspot in Pseudomonas fluorescens (Alternative Format Thesis)

Horton, James

Award date:
2021

Awarding institution:
University of Bath

PHD

The causes and consequences of a mutational hotspot in Pseudomonas fluorescens (Alternative Format Thesis)

Horton, James

Award date:
2021

Awarding institution:
University of Bath

[Link to publication](#)

The causes and consequences of a mutational hotspot in *Pseudomonas fluorescens*

Volume 1 of 1

James Simon Horton

A thesis submitted for the degree of Doctor of Philosophy

University of Bath

Department of Biology & Biochemistry

June 2021

Declaration of any previous submission of the work

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

Declaration of Authorship

This thesis includes an introductory chapter (1), a discussion chapter (5), and three results chapters (2-4). I am the sole author of each chapter aside from chapter 3, which alongside chapter 2 is presented in its manuscript form as intended for publication. As such this chapter includes sections that were primarily penned by my co-author – the details of these contributions are outlined at the beginning of the chapter. In addition, a portion of the experimental work included in results chapters 2 and 3 include those completed by my collaborators. These contributions are likewise accredited in detail at the beginning of each relevant results chapter. All experimental work presented in results chapter 4 was completed by the submitting candidate.

Foreword

Science is an ironic profession. At its origin, science was undertaken to expel the notion of magic, but so much of what we do as scientists possesses a magical quality. There are precious moments during a scientist’s work that occur when, after hundreds if not thousands of hours of experiments, they’ve finally made a major breakthrough – the result is there before their eyes. For just a moment in time, however brief it may be, they are the first and only person in the world to know something fascinating. The internal magic of that moment is as real as any magician’s spell.

I have enjoyed the privilege of experiencing that moment. After two years of attempting to solve a puzzle that I wasn’t sure had a solution, I came across an innocuous comment in a paper that I was lazily reading. In theory this comment provided an answer to my puzzle, but I had to prove it. I spent the next several weeks frantically constructing the tools I needed to test the idea. And when I had finished constructing them I did indeed test it, barely hoping to pray that it was correct. But it worked. That moment remains one of the most satisfying of my life, and multiple years of challenge and strife in academic research was paid for in kind by that feeling.

This experience was not the culmination of my scientific endeavours but its catalyst, as the finding posed as many fascinating questions as it answered (a typical trait of science). In this thesis I will present the data of this moment, and my subsequent research exploring its implications. Our hope as scientists is that at least a mote of the exhilaration we feel when we’ve made a discovery survives transcription onto the page. I hope that if any choose to read this work, these findings at least stir their scientific minds. And I hope that if you are stuck on a puzzle of your own, then an innocuous comment made in these pages inspires you to find the solution you are searching for.

James S. Horton

Dedications

I dedicate this work to the following individuals:

My grandfather, Elihu Ivor Horton.

My grandmother, Brenda Edith Horton.

My mother, Jeannette Cindy Horton (nee Bateman).

My father, Simon Horton. And finally my siblings Jade and Josh, without whom I am nothing at all.

As a young boy, my grandad was the only close family member who was as passionate about science as I was. I remember being handed a several-page long essay by him on the difference between fission and fusion when I was about 13. I regret to say that I never read it, and as my grandad passed away when I was in my mid-teens and I didn't start enjoying physics until I was in my 20's, we never got a chance to discuss it. But we did enjoy many other conversations, and I think of him often when I'm slipping on my lab coat or have made a new discovery, and I wonder what he would have thought of it.

So I thank you grandad for inspiring me, of all the mountains of discovery I was interested in climbing, to take my first step onto the scientific path.

I was not always set on being an academic, having spent many years riddled with insecurity about my talent. But when I was 23 and working on a brewery factory floor, I was struck with renewed resolve to change my stars. My nan threw her support behind me and did everything she could so that I could chase my dreams.

So I thank you nan, for cheering me on as I continued to climb.

Nothing I have accomplished throughout my life would be possible without my mother. She is not a scientist, but her superpower has always been to love and support her children no matter their goals. I could have set my heart of becoming a fisherman in Alaska, a video game developer in Japan, a historian in Egypt, and she would have backed me all the way. As it turns out my dream took me to a lab bench in the south of England, and she was behind that too.

So thank you mum for mopping my brow when I tire from the climb, for encouraging me to look in other directions once in a while, and for giving me a parachute for the times I slip and fall.

And so I dedicate this work to two people who cannot read it, and others who likely never will. But if my name is to forever be immortalised in print in a physical and virtual archive – even though it will inevitably acquire digital and real dust bunnies – it is only fitting that the people who made it possible forever sit alongside me. I think my nan, in an email to me in 2018 (toward the beginning of my doctorate), summarised the experience of a PhD the best:

"I agree with you that in time things will become clearer and **suddenly you will shout Eureka**

because it has ***all fallen into place****, and it will all be worth it,*

even though at times you wonder *'why did I ever bother'."*

Table of contents

Table of figures

Chapter I……………………………………………………………………………………….

Chapter II……………………………………………………………………………………….

Chapter III……………………………………………………………………………………..

Chapter IV…………………………………………………………………………………

Chapter V……………………………………………………………………………………….

Supplementary figures and tables

Chapter I…………………………………………………………………………………….

Chapter II……………………………………………………………………………………

Chapter III…………………………………………………………………………………..

Chapter IV…………………………………………………………………………………...

Chapter V……………………………………………………………………………………

Acknowledgements

I must firstly thank my primary supervisor, Dr. Tiffany Taylor, for allowing me to pursue this work and run in a cascading fashion into the wilderness of exploration. Tiffany is a friend, a mentor, a role model. She has been an arbiter of sound scientific guidance and a dispenser of scientific passions throughout the time I have known her. Dr. Taylor is the reason that I study evolution, and is the main reason that I am so excited to continue to do so. I couldn't have asked for a better mentor, and she continues to inspire me not just in her work but also in her approach to life.

Secondly, I would like to thank Dr. Renos Savva, who taught me as a fledgling master's student. Dr. Savva did not so much polish the edges of this diamond in the rough, but more-so smelted a simple piece of ore into something of more value. I left his lab as not only a competent molecular microbiologist, a far-cry from how I'd began, but I also left with an instillment of belief that I could achieve. I wouldn't have realised my love of writing without his mentorship, nor would I have had the confidence to apply for what turned out to be one of the most fascinating experiences of my life.

As the founding member of Dr. Taylor's lab group at the University of Bath, I would have been resolutely lost during my first year without the support and guidance of my lab mates. I learned much from Dr. Susanne Gebhard during this time, not so much directly but through the conduits of her fantastic lab members. Dr. Carolin Kobras, Dr. Majorie Gibbon and Dr. Bianca Reeksting were invaluable figures at this early stage, and Dr. Reeksting continues to be an invaluable figure in my life today. I am also grateful for the support by Dr. Nicholas Priest throughout this time, and his efforts to introduce me to new collaborators and new opportunities.

I would also like to thank Dr. Jonathon Abrahams and (as of this writing) final year PhD student Sali Morris. Academic research is a peculiar life to those outside our bubble, and so I am hugely grateful to have been trapped inside the metaphorical bubble with two such fantastic people as them (although they will probably advocate that it is them who have been trapped in the bubble with me).

This work would also not have been possible without the two members of the Taylor Lab who have been with me for the past couple of years. PhD student Louise Flanagan's presence alone is enough to make her an asset to any team, but it was only through conversations with her that I was finally inspired to attempt my crowning experiment. And PhD student Matthew Shepherd has been my experimental brother-in-arms. Together we have breached the battlements of frustratingly stubborn experimental protocols, all the while passing our days by indulging in our shared love of history. A huge thank you to them both. Finally I would like to thank Prof. Laurence Hurst for his advice. Academic study is a marriage of innovative research and a compelling narrative, and Prof. Hurst is a true master of both.

Scientists may be soloists, but we are nothing without our orchestra. It has been a privilege to perform in my friend's and colleague's ensemble, and I thank them for supporting me by acting as mine.

Abstract

Mutational hotspots describe areas of the genome where genetic change is more likely to occur due to underlying biases. These biases are often a consequence of interacting genetic features, which raise the likelihood of DNA damage and error. When found at genomic positions under selection, such hotspots can be powerful determinants of evolutionary outcomes, as they can facilitate highly repeatable genetic evolution from the stochastic process of mutation. However our current lack of understanding means such hotspots can be difficult to identify and characterise, and as such their role in evolutionary processes is largely under-appreciated. In this thesis, I reveal the genetic causes and evolutionary consequences of a powerful mutational hotspot in the soil bacterium *Pseudomonas fluorescens*. Engineered immotile lines of this microbe have been found in previous work to rapidly re-evolve motility though a one-step *de novo* mutation. I find a prominent role for mutation bias, which is predicated on silent genetic changes, in facilitating a highly repeatable genetic outcome during the evolution of the motility phenotype. My work also reveals that the local tract of DNA that enables repeatable evolution works alongside several other genetic features, namely strand orientation, genomic position, and mismatch repair proteins to generate this nearly deterministic adaptive outcome. Finally, I examine the evolutionary history of this mutational hotspot, and suggest that the hotspot is not preserved but suppressed by selection, and so hotspots of this type may well appear transiently and throughout the genome. This model system establishes a novel framework with which we can empirically determine the nuances of hotspots and define their role in evolution. The collective work in this thesis therefore moves us closer to comprehensively understanding the mutational drivers that underpin repeatable evolution, and as such acts as a key stepping stone for future forecasts of evolution.

Chapter I

1. Introduction

1.1. Mutational hotspots: The harbingers of repeatable genetic evolution

"Wind back the tape of life to the early days of the Burgess Shale; let it play again from an identical starting point, and the chance becomes vanishingly small that anything like human intelligence would grace the replay." – Stephen Jay Gould, Wonderful Life

1.1.1. The historical contingency

What defines how organisms evolve? If you were to reverse the flow of time and visit the coastline of Uruguay in 1832, you may find a young Englishman who could lay the foundation of an answer. Not yet twenty-five years old, with his face blackened from dirt, his waist loaded with pistols, and his hands encumbered with geological hammers, this young man, Charles Darwin, would have likely offered lengthy lectures to the visiting inquisitor (Darwin 1832). He'd perhaps share his theories on many things he'd observed throughout his travels, including on the geological processes that shaped the earth, and on the living organisms that called it home. Although three years would pass from this moment to his arrival at the Galapagos Islands, the idea of natural selection may well have already been swirling around and fermenting in his mind. Over a decade later, then back in Britain, Darwin would expand on the idea of selection, penning an iconic work that would reframe our view of the natural world (Darwin 1859). His theory encapsulated how populations of a species change slowly over time to better suit the environment around them. Those in a population who acquire a novel trait that allows them to thrive subsequently have a better chance of producing more offspring, and so eventually this trait grows to dominate. The theory of evolution by natural selection not only describes how life becomes specialised to suit certain niches, but also explains speciation events that gave rise to the gargantuan tree of life. But a question remained: how was the information of a new trait passed from parents to offspring?

Darwin espoused the idea of pangenesis in the 1860's, which described how each cell of the body could shed and share 'gemmules' that coalesced into bundles of information and seeded their offspring (Zou 2014). As each cell of a fertile organism was able to commit gemmules to the next generation, this theory championed the idea that multicellular organisms could inherit characteristics acquired over a parent's lifetime. Acquired inheritance was popular amongst scholars at the time but would fall from fashion in the coming decades. To find the origins of why this would occur, we migrate not through time but through space, to visit a quiet contemporary of Darwin's in a monastery in Northern Moravia.

In Brno – situated in modern day Czech Republic – in the 1850's, we arrive within a botanical garden worked by one Gregor Johann Mendel. Mendel, like Darwin, was an avid scholar, but unlike the Englishman lacked the financial means to pursue his passions through his father's purse. Instead he had been admitted as a monk, which allowed him to teach and continue his intellectual pursuits. Mendel was fastidious, and spent many years busily growing and studying variants of peas at his monastery.

Although it wasn't his intention at the time, this tranquil setting facilitated his wonderful mind to establish the foundation of genetics. By breeding and recording the traits of his peas, Mendel demonstrated that traits of offspring weren't simply an amalgamated blend of their parent's traits – they were compartmentalised, inherited distinctly from one another, and (in heterozygous cells) could have dominant and recessive characteristics. In 1865 Mendel published his findings under the title "Experiments on Plant Hybrids", the contents of which birthed the idea of inheritance (Abbott and Fairbanks 2016). But these ideas would require a long gestation, and by the time Mendel died in 1884 his work had not yet found prominence.

Around the same time, a reverend and scientist by the name of William Dallinger conducted perhaps the world's first experimental evolution experiment. Dallinger's work had already garnered many plaudits, including from Charles Darwin, and throughout the 1880's he performed a seven-year experiment aimed to test Darwin's theories in real time. The reverend took unicellular microbes and placed them under selection for increased temperature resistance by periodically increasing the thermal stress of their environment over an estimated half a million generations. In the first year of his experiment many of his microbes suffered or died at >23°C, but by the end of the experiment they could grow at 70°C. And when placed back in their ancestral condition of ~18.3°C, many of the adapted lines could no longer survive (Haas 2000). Thus Dallinger exemplified that experimental evolution of microbes allows evolution to happen before our eyes.

At the turn of the 20th Century, a small enclave of budding geneticists starting publishing similar findings to Mendel and in doing so (perhaps inadvertently) shined light on Mendel's findings. In 1901 Hugo de Vries coined the term 'mutation' to describe spontaneous change in an inherited trait (Theunissen 1994). And in 1909, a fellow botanist named Wilhelm Johannsen first coined the term 'gene' to describe the inherited units first described by Mendel. Johannsen also used the terms 'genotype' and 'phenotype' to differentiate between an organism's genetic makeup and its physical traits. The laws of inheritance had finally blossomed, but this advancement in genetics was met by some resistance, as it posed the question as to how this married with Darwin's theory of natural selection. This prompted another era of advancement, but rather than observations of the natural world, these innovations were products of great theoretical minds, of which Ronald Fisher was at the forefront. Fisher, like Mendel before him, was a disciple of mathematics, and would grow famous throughout his lifetime for his innovations in statistics. But Fisher also turned to his formidable mind to genetics and fused it mathematically with Darwin's principles of evolutionary dynamics. This was the genesis of what is now known as the 'modern synthesis' (a term typically used interchangeably with 'neo-Darwinism'; Noble 2011).

While the fate of mutations and genes under selection was becoming clearer, the events that drove mutation itself remained largely unknown. Until the 1940's, the prevailing thought was that genetic information must be transferred via proteins. This began to change (albeit slowly) from 1944 following a publication by Oswald Avery and colleagues. Just as Dallinger had utilised microbiology to demonstrate evolution, so too did Avery utilise microbiology to demonstrate the keystone of genetics. Using a pneumococcal species of bacteria, Avery found that only through 'sodium deoxyribonucleate' (DNA) could encoding information be transferred to new cells, information that the recipient cell could utilise to form a capsule – a novel phenotype (Avery et al. 1944; Cobb 2014). Then in the 1950's, while many scientists around the globe were engrossed in designing and refining atomic bombs, in the UK future Nobel Prize laureates Watson and Crick, with considerable assistance from Rosalind Franklin and Maurice Wilkins (Klug 1968), dedicated their time to describing the helical structure of DNA for the first time (Watson and Crick 1953). What these four people revealed was that DNA was so simple, yet so complex. So stable, yet so vulnerable. As the years raced forward and scientific research began to proliferate, laboratories across the globe uncovered more and more about how these vulnerabilities led to mutations – the genetic changes that underpins the evolution of all life.

The act of mutation leads to genetic variation, but this process is a double-edged sword for an organism. While increasing genetic diversity facilitates evolution, mutations can also often be harmful. This is especially true when an organism is living in an environment to which it is already well adapted. At this point most possible mutations will be deleterious i.e., harmful to the organism's fitness, an idea that was first explored by Japanese scientist Motoo Kimura in 1967 (Kimura 1967). The following year Kimura expanded on this theory, and argued that i) As adaptive mutations are rare, and ii) Heavily deleterious mutations often wouldn't persist, most mutations that persisted over evolutionary time would be mildly deleterious, or neutral i.e. neither harmful nor helpful to an organism's fitness (Kimura 1968). This led scientists to wonder what balance had been struck within genomes – did natural selection work to drive mutation rates down, and if so, why did mutations occur at all?

In 1991, geneticist and microbiologist John W. Drake reported a striking pattern in mutation rates that existed across diverse microbial taxa. He observed that the number of mutations per nucleotide base pair was commonly inversely proportional to the size of the microbial genome (Drake 1991), meaning that mutation rate was almost equal across all DNA-based microbes. This convergence suggested that mutation rates were not defined at random but were subject to natural selection. In 2010, geneticist Michael Lynch suggested that mutation rates were low, but not entirely absent, because successive adaptive mutations lowering mutation rates would offer diminishing returns. As a result natural selection could only restrict mutagenic mechanisms so far before the power of genetic drift (i.e. random events that cause mutations to rise and fall from prominence) prevented mutation rates from falling further (Lynch 2010).

By the 2010's Darwinian theory and genetics had been reconciled, and a cavalcade of papers had outlined the mechanisms and rates by which mutations can occur. We knew then, in principle, the features that defined how organisms evolve. Yet one of these features – that of mutation – possessed an inherent randomness absent in the other features. Global rates of mutation may well be suppressed by natural selection, but the individual mutations that drove genetic change could not be so tightly dictated. What did this mean for the passage of evolution? In a book released in 1989, palaeontologist Stephen Jay Gould had argued that unpredictable events introduce changes that send evolving organisms down particular evolutionary paths – a chain of events he termed as historical contingency (Gould 1989). Therefore, Gould argued that if we were to reverse the flow of time even further back than the Uruguayan summer of 1832 – when Darwin was busy storming a rebel-held fortification with pistols and cutlass in-hand (Darwin 1832) – if we were to go back long before humans or even mammals came into existence, then the stochastic principles and processes of evolution would mean our species would not appear again.

What many took from Gould's arguments, therefore, was that evolution was inherently unpredictable. And although his thought experiment was provoking, we lacked the means to test it on a planet-wide scale; most readily observable natural evolutionary events only happen once, and so we cannot know just how likely they were to occur. However, with work built on the principles of Darwin, on the foundation of Dallinger, on the mantle of Avery, modern microbiologists have found a way to perform Gould's grand thought experiment on a micro-scale (Blount et al. 2018). By working with clonal organisms evolved in real time under identical conditions, microbiologists can rewind the "tape of life" and play it again and again (Buckling, Brockhurst and Colegrave 2009), and have shown that sometimes evolution is, in fact, repeatable, and therefore predictable. This tells us that the defining features of evolution are not always slaves to chaos, but can align to drive evolution to realise a single outcome repeatedly.

It has been shown that predictable evolution is possible because of selection, but also because of the act of mutation itself. This is because the processes that cause mutations do so unevenly across the genome, and as a result certain positions within DNA can become "mutational hotspots" that mutate more readily than at other positions. Mutational hotspots are powerful evolutionary determinants of repeatable evolution, because they bias the very process that creates genetic variation. This means that selection has more opportunity to act on these sites, which can lead to the enrichment of the frequently mutated alleles in a population. Thus from the unpredictable chance event of mutation we can achieve predictable evolutionary outcomes.

1.1.2. Aims of this study

In this work, I utilise molecular microbiology to investigate the role played by mutation on the repeatability of evolutionary outcomes, and in doing establish a novel model system for the study of features that define predictability in evolution. In chapter 2, I discover that a mutational hotspot is responsible for a highly repeatable evolutionary event in the soil bacterium *Pseudomonas fluorescens*.

I show that evolutionary forces such as mutational accessibility, the environment, and clonal interference cannot explain the observation of an identical single nucleotide polymorphism occurring repeatedly across independent lines. Instead I uncover a critical role for mutation bias by showing that a handful of synonymous genetic changes around the hotspot site drastically alter the likelihood of repeatable evolution across two *P. fluorescens* strains. I complement these experimental results with *in silico* models that show that silent variation may well impact mutation bias by altering the formation of stem-loops. These are alternate DNA secondary structures formed when the DNA helix is unwound into single strands, which allows repeats of complementary nucleotides on the same strand to bind to one another. These findings therefore demonstrate how powerful mutation bias can be in determining evolution, and reveal a powerful evolutionary role for silent genetic variation in driving adaptive divergence or convergence across homologous genetic backgrounds.

These conclusions present several fundamental questions: are repeat sequences (and DNA secondary structure) enough on their own to build a frequently mutated mutational hotspot? Or do other genetic features intertwine to enable highly predictable evolution? And just how did this hotspot come to be? Is it a product of genetic drift, or did selection for a mutable locus establish and preserve the hotspot in the organism's evolutionary history?

In chapter 3, I elucidate the other key genetic features that work in conjunction to enforce highly repeatable evolution. Myself and my colleague Matthew Shepherd utilised genetic engineering to augment genome position, synonymous sequence, strand orientation and mismatch repair function to demonstrate that each component is essential for realising highly predictable adaptive outcomes. By doing so I am able to form a theoretical framework that describes the genetic context needed for repeatable evolution in our model organism.

In chapter 4, I explore the evolutionary origins of the *P. fluorescens* strain SBW25 mutational hotspot. By comparing the synonymous variations around the hotspot site through comparative sequence analysis, I find a suggestion that the hotspot was subject to selection in its evolutionary history. I next therefore examine if positive selection could be operating on the hotspot by experimentally assessing the hotspot's propensity to remain mutable under fluctuating environmental regimes. I find no evidence for positive selection enforcing the mutability of the hotspot, and therefore this chapter concludes with an argument that mutable hotspots may arise by genetic drift and be subsequently removed by purifying selection.

In chapter 5, I conclude this work by outlining how these findings both reveal that mutational hotspots can drive highly repeatable evolutionary outcomes, and describe how our forensic understanding on their mechanistic drivers will facilitate future predictions of evolution. I finish by outlining future work that can expand on the findings within this thesis, and bring us closer to making accurate evolutionary forecasts that exploit predictable evolutionary actors.

1.2. Biased Beginnings: Genetic agents of mutation rate heterogeneity

"Truth is... the game was rigged from the start." – Benny (Matthew Perry), Fallout New Vegas

Mutation is a game of chance. An unwanted mutation can consign an organism to oblivion. A desirable mutation can lead to its proliferation. But while mutation is a chance event, not all mutations share the same likelihood of occurring. Instead, the vast spectrum of possible mutations – the various types and locations they can occur – are each a face on a loaded die. The weighted nature of the die, the element that skews which faces will appear, is decided by a slew of factors. However, these multitudes can be summarised into two broad themes: the genetic architecture of DNA, and the molecular and environmental components that act upon it. As such the mutability of DNA is linked to, but certainly distinct from, the protein-coding structure of genetic code, as mutation rates will not be uniformly correlated with protein expression and essentiality. In this section we will consider DNA from the standpoint of mutability, and by doing so will discover the genetic and environmental features that bias the rolls of the dice.

Mutational biases have been shown to play a key role in evolution and adaptation across all domains of life including bacteria, archaea, retroviruses, yeast and even in higher eukaryotes – including humans (for review see Buisson et al. 2019). However, this review will primarily focus on the identified mutational biases operating in bacteria, of which the major contributors are highlighted in Fig. 1.1. Literature using other such model organisms will be discussed, however I will not directly discuss features that do not apply to bacteria, such as the heightened mutability of single-stranded RNA genomes found in some viruses (Carrasco-Hernandez et al. 2017). Likewise as the focus will be on the core chromosome accessory genomic elements such as plasmids, which introduce genetic variation through horizontal gene transfer, will not be discussed in detail. The implications of harbouring horizontally transferred genetic material on a plasmid's own evolutionary persistence (Carroll and Wong 2018) and the evolution of their microbial hosts (Rodríguez-Beltrán et al. 2021) has been reviewed elsewhere. Microbes from other domains of life additionally possess fundamental differences in genome organisation, such as the multiple origins of replication commonly found in archaea (Wu et al. 2014). This consequentially negates some of the key drivers of mutation bias witnessed in bacteria (see review by Rocha 2004). As such commonalities of biases shared by bacteria and other domains of life will be cited in the text.

1.2.1.1. Drivers of mutation bias: Distance from the replication origin

We can glean many insights into the mutability of genomic regions purely from possessing an assembled genome, because these genome-wide strings provide us with positional insight. Mutability can operate on an extremely specific scale, but biases also fluctuate over broad genomic regions. A key marker for identifying a region with high replication fidelity, for example, are those that lie close to the *ori* – the origin of replication, (Hudson et al. 2002; Long et al. 2014). Using a mismatch repair-defective strain of *P. fluorescens* SBW25, Long et al., performed full genome sequencing on a number of evolved lines and observed that mutation rate increased sharply within a megabase of the origin, then reached a plateau that was maintained to the replication terminus (*ter*) region. Zhang and colleagues performed duplex sequencing of *E. coli* and observed higher mutation rates at replication fork stopping points (Zhang et al. 2018). Hudson et al., measured the mutability of a locus inserted in multiple locations around the *Salmonella enterica* genome, and found that replication fidelity was higher at both the origin and the replication terminus than it was at intermediate distances (Hudson et al. 2002). A similar 'bulge' in mutation rate was observed by Dillon and colleagues who performed whole genome sequencing on mismatch repair mutant lines of *Vibrio cholerae* and *V. fischeri* (Dillon et al. 2018).

These findings suggest that fluctuations in mutation rate vary across bacterial species but share commonality in that fidelity is higher close to the replication origin. This correlation may not be determined in absolute terms by genetic distance but instead by replication timing (Dillon et al. 2018), which has also been incriminated as a driver of mutation bias in human cancer cells lines (Tomkova et al. 2018). Replication timing coincides with genomic position, as DNA polymerase begins at a defined position (*ori*) and the replication forks progress in a defined manner toward the terminus. As such the distance from the replication origin enjoys a tightly correlated relationship with the time span until that region is replicated. Therefore, rather than the genomic position itself defining mutation rate, the molecular components that are produced at the start of replication and change in concert with time since replication began may be the causative agents of mutagenicity. One mechanism that supports the hypothesis that molecular components are responsible for varying replication fidelity across genomic regions are the errors induced by the accumulation of deoxyribonucleotides (dNTPs), (Dillon et al. 2018). An excess of dNTPs have been shown to increase mutation rate in *Saccharomyces cerevisiae* (Chabes et al. 2003) and *E. coli* (Gon et al. 2006) which may be owed to an increased chance of mismatches when these molecules are abundant (Dillon et al. 2018). Furthermore, the relative abundances of dNTP types can also invoke biases in mismatch mutations (Watt et al. 2016). In contrast, a lack of dNTPs may likewise increase mutation rate through an alternate mechanism of causing the replication fork to stall (Gon et al. 2006). dNTP production is triggered by the initiation of replication (Gon et al. 2006) and the components then subsequently spike then decline, an effect that can be exacerbated during exponential growth when rounds of replication overlap (Dillon et al. 2018). This therefore may cause mutation rate to 'bulge' at intermediate genome locations (see review by Kivisaar 2020 for examination of this hypothesis).

1.2.1.2. DNA topology

DNA polymerase complexes are not the only proteins to bind DNA during replication. While replication timing and the resultant change in molecular components affect polymerase fidelity, nucleoid associated proteins likewise alter DNA topology both by affecting coiling and remaining bound to DNA for extended periods. As such it is likely that multiple factors exert an influence on genomic location-based mutagenesis. The MatP protein in *E. coli* binds across the *ter* macrodomain where it coordinates segregation of the region (Crozat et al. 2020) but a strain deficient in this protein was found to remove the mutation rate 'bulge' at intermediate genomic distance (Niccum et al. 2019). Niccum and colleagues additionally reported that the deletion of two other nucleoid-associated proteins involved in augmenting DNA superhelical structure, HU and Fis, also changed the positional bias of base-pair substitutions. Deletion of HU subunit HUα decreased substitution mutations substantially in the *ter* region, whereas Fis-deficient strains only exhibited decreased mutations near the *ori* (Niccum et al. 2019).

1.2.1.3. Substitution biases

Having access to an assembled genome offers considerable insight into anticipated mutagenicity across genomic regions, but performing mutation accumulation experiments (MA) with these lines and sequencing their evolved descendants also provides insight into biases that operate on each base variant. MA experiments often highlight genomic mutational biases toward transitions – where a purine (A and G) changes to a purine, and pyrimidines (C and T) change to pyrimidines, or transversions – where purines change to pyrimidines and vice versa. When transitional biases preferentially move in one direction, we can also observe a bias from A:T → C:G or C:G → A:T genome enrichment. To highlight a few examples, a G:C → A:T bias has been observed in MA wild type lines of the gram-negative bacteria *E. coli* (Lee et al. 2012) and *Pseudomonas aeruginosa* (Dettman et al. 2016), and in mis-match repair deficient lines of *S. enterica* (Hudson et al. 2003). Transitional biases have additionally been demonstrated in *Mycobacterium tuberculosis* (Payne et al. 2019). Such biases are noteworthy, as they have been shown to affect instances of parallel evolution (Stoltzfus and McCandlish 2017), meaning that the same genetic changes are repeatedly observed. As well as each of the four biases enjoying its own mutational bias, MA experiments have also highlighted that a given nucleotide's immediate flanking bases can also greatly impact its mutation rate. Therefore focal nucleotides are considerably more, or less, mutable depending on which nucleotides they are sandwiched between – e.g. a focal A has been found to be highly mutable when flanked by two C's (C$\underline{\text{A}}$C) in multiple bacteria (Long et al. 2014; Dettman et al. 2016; Schroeder et al. 2016).

1.2.1.4. Homopolymeric tracts and tandem repeats

As highlighted by their impact on changing DNA topology, proteins that are involved in the replication, regulation and maintenance of DNA are of mammoth importance to biased mutagenesis. Microbes are equipped with an arsenal of enzymes involved in replicating, proof-checking, and repairing broken or mismatched DNA, all of which have a bearing on mutation rate (reviewed in Ganai and Johansson 2016). Certain genetic features are therefore conducive to mutation owing to the molecular apparatus that interacts with them. Notable amongst these are homopolymeric tracts and repetitive stretches of nucleotides that are affiliated with polymerase-slippage and resultant mispairing and indel mutations (Moxon et al. 2006), and recombination events (Zhou et al. 2014). For example, indels in mismatch-repair deficient mutants in *P. aeruginosa* were enriched at homopolymeric tracts of C's and G's numbering 5-base pairs or longer (Dettman et al. 2016). These frameshift mutations, which occur as a result of this polymerase-DNA substrate relationship, can happen at such high frequencies that they can form a "contingency locus". Such loci allow evolving populations to rapidly switch phenotypes when frameshift mutations occur in either coding or promoter regions, utilising mutation as a primary means to alter protein activity following environmental change (Koch 2004; Moxon et al. 2006).

1.2.1.5. Replicative strand

The polymerase-slippage induced frameshift mutations introduced above can also operate in a strand-dependent manner, as polymeric tracts composed of pyrimidines have been demonstrated as more susceptible to slippage mutagenesis in *P. Putida* (Juurik et al. 2012). Strand-affiliated mutational biases differentially impact the leading strand, which is replicated near-continuously, and the lagging strand – which is replicated in fragments – and play prominent roles in biasing mutation types as well as frameshifts. Strand bias is widespread in bacteria (Rocha 2004) and is reliant on varying mechanisms, which operate differently depending on the strand owing to the nature of the replication fork. For example, replication fidelity of substitution mutations have been documented to be higher on the lagging strand in *E. coli* (Fijalkowska et al. 1998; Maslowska et al. 2018), which is possibly owed to an increased disassociation ability of the lagging strand polymerase (Maslowska et al. 2018). In contrast, higher mutation rates have been observed on the lagging strand owing to the formation of hairpin structures (Trinh and Sinden 1991; Leach 1994; Langenbucher et al. 2021).

1.2.1.6. Single-stranded DNA hairpin formation between inverse repeats

The formation of hairpin structures is facilitated by perfect or imperfect inverse-repeat regions of nucleotides (also known as palindromes and quasipalindromes). When DNA is single-stranded, neighbouring segments of inverse-repeats on the same strand – which are complimentary – can pair and form stem-loop secondary DNA structures (De Boer and Ripley 1984), a type of hairpin. If both strands of DNA form symmetrical intra-strand DNA structures then they form a cruciform, a secondary structure known to play a role in the coiling of DNA (White and Bauer 1987) and in translocation in human genomes (Kato et al. 2011).

Inverse-repeat regions that form DNA hairpins have been documented to increase the rate of DNA polymerase slippage in both bacteria (Leach 1994) and archaea (Castillo-lizardo et al. 2014). The mechanism incriminated for this is replication stalling (Voineagu et al. 2008), which often occurs when a hairpin forms on the Okazaki fragment of the lagging strand as it remains single-stranded for protracted periods (Bikard et al. 2010). Replication stalling as a consequence of inverse repeats has been demonstrated by Voineagu and colleagues in bacteria, yeast and primate cells (Voineagu et al. 2008) and so the related mutagenesis may be ubiquitous throughout life. Both deletions and tandem duplications are possible outcomes of slippage events, but cleavage of hairpin structures by nucleases can in some contexts bias inverse repeat regions to undergo deletions (Darmon and Leach 2014).

Owing to their propensity to form secondary structures, inverse repeats can illicit highly targeted mutational biases at limited genomic sites. In some instances, these biases can become yet more refined so that an exact mutational event occurs at orders of magnitude more often than alternative mutational events (Dutra and Lovett 2006; Buisson et al. 2019), making these sites hotbeds for evolution. Such 'mutational hotspots' can occur when the hairpin acts as a substrate for a mutagenic protein (Langenbucher et al. 2021), or when hairpins formed of imperfect inverse-repeats engage in template switching (Lavi et al. 2018). In this latter case, a polymerase will use one arm of the stem as a template in place of the other, converting the two arms of the stem into perfect compliments through either substitution or indel mutations (Dutra and Lovett 2006; Klaric et al. 2020). As such these hairpins can bias mutation to such a degree that even when relying on the chance event of mutation to drive adaptation, an identical adaptive genotype will repeatedly appear (Dutra and Lovett 2006).

Other than DNA hairpins and cruciforms, inverse repeats can additionally form alternative DNA configurations (see reviews by Mirkin and Mirkin 2007; Wang and Vasquez 2017; Brazda et al. 2020). These alternate types, such as triplex structures formed between three strands of simple repeats, can introduce their own mutational biases by increasing genome rearrangements (Holder et al. 2015). This reinforces repeat regions (inverse and otherwise for alternative forms of DNA structure) as hotbeds for mutation, owing to an array of DNA structural alterations and the specific and repeatable mutational events that they can invoke.

1.2.1.7. Head-on collisions

DNA exists in a state of flux – as we have discussed above it is often being wound and bound by a slew of proteins involved in replication, maintenance, and repair. However during this time DNA is simultaneously performing its role as an instructional instrument, where RNA polymerases and their affiliated proteins busily transcribe genetic information into mRNA. Genes are transcribed in two orientations, either 'forward' – which moves in the same direction of the replication fork, or 'reverse' – which moves toward the replication fork. As a cell does not cease its protein manufacture during periods of replication, collisions between proteins are inevitable. Genes that are transcribed in the forward direction are susceptible to collisions as the replication fork moves faster than transcription machinery, but polymerases transcribing in the reverse orientation more readily undergo head-on collisions with the replication fork (Merrikh 2018). Thus reverse-transcribed genes tend to have higher mutation rates than their forward counterparts, which has been exemplified by multiple studies (Paul et al. 2013; Sabari Sankar et al. 2016; Lang et al. 2017) including by Juurik and colleagues who flipped a gene's strand orientation and observed a notable change in mutation rate (Juurik et al. 2012). In *Bacillus subtilis*, the increased mutagenicity following head-on collisions has been attributed to the formation of R-loops, which are hybrid nucleic acid structures comprised of both DNA and RNA (Lang et al. 2017). The same bacteria have been utilised to show that mutation types are also biased depending on where the collision occurs. Collisions at transcription initiation sites have been seen to invoke adenine deamination leading to substitution mutations, whereas collisions with the transcription elongation complex invokes replication stalling and subsequent indel mutation (Sabari Sankar et al. 2016).

1.2.1.8. Transcriptional mutagenesis

While many of the discussed mechanisms of mutation involve the process of replication, the process of transcription itself can also increase the likelihood of genetic change, as genes that are expressed more highly have been noted to mutate at higher rates in *P. putida* (Juurik et al. 2012) and yeast (Park et al. 2012). This is however not a universal trait, as highly expressed genes in *E. coli* cells have been documented to have normal (Zhang et al. 2018) and even lower (Martincorena et al. 2012) mutation rates than other genes. Increased mutability can be explained in part by increased opportunity for collision (Paul et al. 2013), but as with replication part of the mutability of transcription can be a consequence of DNA being separated into single strands, which renders nucleotides more susceptible to damage (Chan et al. 2012). Deamination damage has been demonstrated to occur during replication stalling in aphids (Klasson and Andersson 2006) and this damage also occurs during transcription. Cytosine deamination is biased toward non-transcribed strands, which remain exposed as unbound single strands for longer than their counterpart (Davis 1989; Morreall et al. 2015). Such situations do offer an alternative functionality for hairpin formation, however. Although seemingly counter-intuitive, the formation of hairpins in such cases has been noted to lower rather than raise mutation rate, as inverse repeats allow single-stranded DNA to form intrastrand bonds and thus become less exposed and susceptible to damage (Hoede et al. 2006). This highlights the difficulty in determining mutability from single genetic features alone, as we must often also consider the mechanism in which they're involved i.e. the context of the protein-DNA substrate interaction.

1.2.1.9. Homologous recombination

We have so far discussed mutations that affect very small areas of the genome, such as single nucleotide substitutions and small indels. But large-scale genomic rearrangements can also occur through one mutational event of homologous recombination. This mechanism facilitates deletions, duplications, inversions, and translocations of DNA segments that can span multiple coding regions and longer (Roth et al. 1996; Darmon and Leach 2014). Rearrangements following recombination are mostly the product of mobile genetic elements (MGE's) such as transposons (Darmon and Leach 2014) and smaller insertion-sequence (IS) elements (Lee et al. 2012), both of which are flanked by direct and inverse repeat regions. MGE's migrate via a number of mechanisms (see Roth et al. 1996), but many are facilitated by certain genetic sequences and as such are biased to rearrange at these positions, including homologous recombination between repeat regions (Naito and Pawlowska 2016). As MGE's can leave relics – in the form of repeat regions – behind following a rearrangement, genomes can become enriched with IS elements and recombination sites (Naito and Pawlowska 2016). This process has aided in rearrangements becoming one of the primary means of creating genetic variation in many bacterial species that colonise human hosts, including *S. enterica* (Matthews et al. 2011), *M. tuberculosis* (Chen et al. 2018), and *Bordetella pertussis* (Weigand et al. 2017). There is also bias within rearrangement events, as deletions are often found to be much more common than insertions (e.g. Raeside et al. 2014) which has been suggested to be a consequence of deletion mutational bias (Mira et al. 2001). Genome rearrangements are therefore more likely to occur when MGE's increase in abundance and are biased to occur at repeat regions, whereafter deletions are expected more than insertions. However, rearrangements can also facilitate subsequent mutagenic mechanisms. For example, inversions will swap the leading and lagging strands within the rearranged segment of DNA, and the orientation of any enclosed reading frames will also be reversed. Similarly, translocations can migrate segments closer or further from the replication origin, therefore impacting their future mutability.

1.2.2.1. Mutagenicity across time and space: Growth cycle

The relative importance of mutagenic actors related to either replication or transcription will fluctuate throughout the bacterial growth cycle. Mutagenic mechanisms that operate during replication, and head-on collisions between RNA polymerase and the replication fork, will naturally perform more prominent roles during exponential growth when cells are dividing rapidly. In contrast, mutagenic mechanisms involved in transcription can predominate during stationary phase when cells are not replicating as frequently (Davis 1989), and because reactive oxygen species accumulate during growth resulting in increased amounts of 8-oxo-guanine (Alhama et al. 1998), which continues to cause mispairing in non-dividing cells as transcription continues (Sekowska et al. 2016). In addition, chemical agents produced in the stationary phase and later during starvation can introduce their own mutational biases, such as GC → AT transition mutations observed in *E. coli* following DNA damage from an alkylating metabolite (Taverna and Sedgwick 1996). Furthermore, nucleoid-associated proteins linked with biasing mutation rate have been shown capable of both lowering and raising mutation rate depending on the growth phase (Warnecke et al. 2012). Warnecke and colleagues observed that regions bound by nucleoid-associated proteins in *E. coli* typically displayed lower mutation rates, but this effect could be reversed later in the bacterial growth cycle, possibly due to the nucleoid-associated proteins interfering with DNA repair machinery (Warnecke et al. 2012).

1.2.2.2. Environment

While growth phase-affiliated mutagenesis means that mutational biases diverge across time, a cell's local environment means that mutational bias also diverges across space. Environment and growth cycle are of course entwined e.g. high energy in the environment improves the DnaA^{ATP}/DnaA^{ADP} ratio, which helps trigger replication initiation (Kurokawa et al. 1999). However changing environments also alters the molecular elements that interact with genetic apparatus, thereby introducing biases directly and indirectly by affecting gene expression. The model organism *E. coli* has been utilised to show the impact of environment on mutational outcomes. Foster and colleagues used a mismatch repair deficient mutant to show that by changing temperature and the nutrient condition, they could lower the maximum growth rate and decrease an A:T transition bias (Foster et al. 2018). Klaric and colleagues adopted a targeted approach and assessed the impact of a suite of small molecules on template-switching mutations following hairpin formation, and found that several molecules stimulate these mutation types (Klaric et al. 2020).

1.2.2.3. Combinatorial mutagenesis

Finally, it should be noted that the genetic features which introduce local mutational biases are often interlinked (e.g. DNA topology and distance from the replication origin, or template-switching and environmental triggers). This applies to many of the mechanisms outlined above; for example, gene orientation and distance from the origin combine to produce stand-specific mutational biases in yeast (Pavlov et al. 2002). Likewise, transition substitution bias has been documented in *B. subtilis* on replication fork-facing genes, but only if the local nucleotide composition allows (Schroeder et al. 2016). Transition bias has also been noted to change with levels of transcription in *E. coli* (Hudson et al. 2003). In some instances mutational biases are enforced or counteracted by DNA repair machinery. In *P. aeruginosa*, mismatch repair enzymes were implicated to more readily correct transition mutation errors (Dettman et al. 2016). However the *B. subtilis* polymerase PolY1, which helps replication machinery circumvent obstructing genetic features, is error-prone and increases the mutation rate of genes orientated toward the replication fork (Million-weaver et al. 2015). DNA winding can also be a mechanism for gene regulation (Dorman and Dorman 2016), and as such changing DNA topology can potentially exert an influence on transcription-related mutagenesis. As such many of these mutagenic mechanisms do not operate in a vacuum but rather in a combinatorial fashion.

1.2.3.1. Conclusion

In this introduction, I have highlighted a number of key features that drive mutational biases to fluctuate significantly throughout the bacterial genome. While this demonstrates our detailed understanding at the molecular scale of the many factors that influence and dictate mutation biases, it also highlights the challenges we face when trying to understand these factors in the context of evolutionary processes.

When analysing evolutionary data it can be challenging to appreciate the effect of these features, as it is often difficult to disentangle the role played by mutational biases from the role played by selection. This is because an emergent mutation could be the result of higher fitness, which allows the genotype to outperform its ancestor and grow in frequency, or the result of mutational bias that introduces the genotype into the population early and often. It has been championed that when assessing evolutionary outcomes, it is appropriate to start with a null hypothesis of neutral evolution (Duret 2008). Yet oftentimes in experimental lab settings, where populations evolve under strong selection, we instead assume selection as the null hypothesis (as often reported when a novel genotype reaches high frequency in a population and its relative fitness has not been described). The two forces are by no means mutually exclusive, as will be discussed in the next section. However, as selection is treated as the dominant force, mutational biases tend to play an underappreciated role in determining evolution.

Mutational bias can masquerade as selection, but so too can selection masquerade as mutation bias, and it is important to be aware of such forces. One notable example of this is retromutagenesis (outlined in Supplementary Fig. 1.1). As discussed above replication is a key process in generating mutation, but it is also responsible for mutation immortalisation. Mutations affecting one strand of the parent cell, such as in mismatches owed to oxidative damage (Hogg et al. 2005), occur in the parent cell but are only immortalised in the daughter cell when both strands are changed. The daughter cell is therefore the first to exhibit the phenotype from this genetic change, meaning that mutation precedes selection as in the foundation of the neo-Darwinian theory (summarised in Basener and Sanford 2018). Retromutagenesis describes an exemption to this rule: In the stationary phase of the bacterial growth cycle, where replication can cease but transcription continues, selection can limit viable mutational avenues to the template strand of transcribed regions. As the template strand alone will be transcribed and translated into a protein, mutations here can have an immediate effect on fitness. As such in some instances the selective advantage enjoyed by the augmented protein product can allow the cell to divide (Morreall et al. 2015), and then it is a matter of chance if the mutation is immortalised in the daughter cell. Instances of retromutagenesis can therefore appear to be owed to mutation bias, as evolution is funnelled to transcribed areas on a single strand, but selection is nonetheless influencing the outcome. Prying apart mutational biases and selection continues to be a challenge, but it is imperative that we are able to do so if we are to ever to comprehensively understand mutational outcomes.

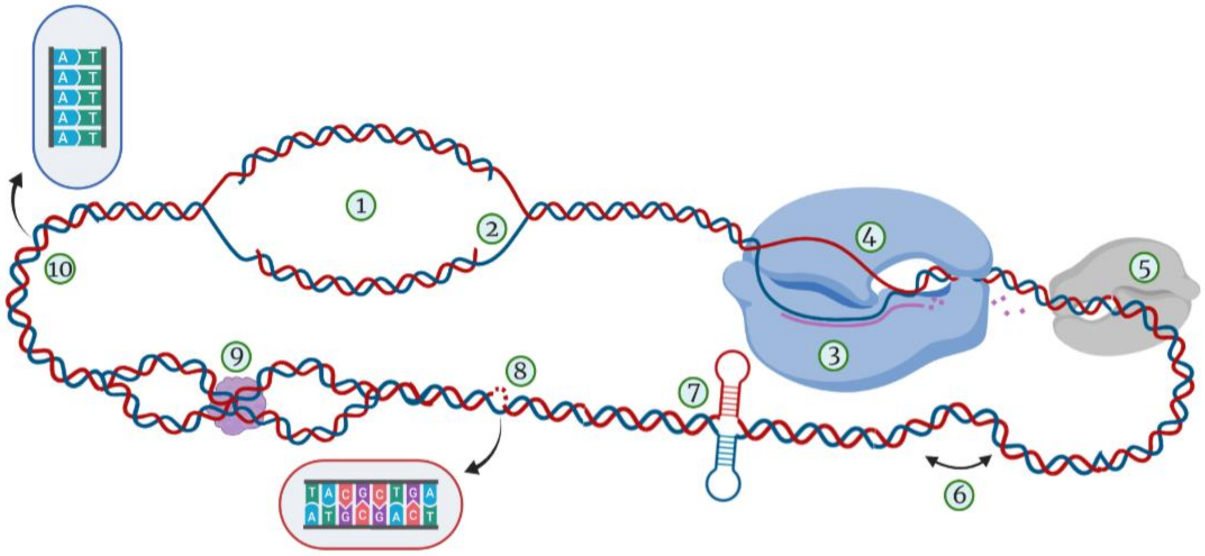


Figure 1.1. Drivers of mutation rate heterogeneity across the bacterial genome. An illustrative overview of the key genetic features that impact rates of mutation: **1)** Distance from the replication origin: replication fidelity is typically higher closer to the origin and can drop to its lowest at intermediate distances or toward the replication terminus. **2)** Leading/lagging strand position during replication: replication fidelity can be higher on the lagging strand, but this strand has a greater opportunity to form secondary structures that can cause replication fork stalling. **3)** Rate of transcription: highly expressed genes can offer more opportunities for mutation due to variables 4 and 5. **4)** Single-stranded DNA exposure on non-transcribed strand: single-stranded DNA is susceptible to deamination damage. **5)** Head-on collisions between RNA polymerase complex and the replication fork: collisions often result in mutation at the site of impact. **6)** Homologous recombination between repeat regions: genome rearrangements are facilitated by repeat regions and can result in deletions, duplications, inversions, and transitions. **7)** Alternative DNA secondary structure formations such as hairpins (stem-loops) and cruciforms: secondary structures can cause stalling of the replication fork, and stem-loop structures formed between inverse repeats can cause template switching . **8)** Local nucleotide neighbourhood: a focal nucleotide's mutability can be heavily influenced by the nucleotides immediately flanking it. **9)** Nucleoid-associated proteins and resultant DNA topology: binding DNA and changing the state of winding can either heighten or decrease the mutability of affected regions. **10)** Homopolymeric tracts and tandem repeats: these low-complexity regions cause polymerase strand-slippage that introduces mutations within the tracts. Full descriptions and citations for each of these features can be found within the main text.

1.3. Domination, Dither or Doom: Examining the fates of mutagenic mechanisms under selection

"One general law, leading to the advancement of all organic beings, namely… let the strongest live and the weakest die." — Charles Darwin, On The Origin of Species

Mutagenic mechanisms may operate throughout the genome, but their persistence within DNA across generations is by no means assured. Detailed insight into the structure of genetic code that enables all life: how it works, and how it can go wrong (causing mutations), has been an emergent area of science for around the past half-century. However the force of natural selection, which reshaped how we viewed biology, has been known for much longer (Darwin 1859). The modern synthesis fused Mendelian genetics and natural selection (outlined by Basener and Sanford 2018) and is still utilised by many evolutionary biologists. With this reconciliation it is accepted that mutations are responsible for generating the genetic diversity that natural selection subsequently acts upon. As such mutations are the raw material – the clay – to the shaping hands of natural selection. Sometimes the hands of natural selection are firm (as when under strong directional selection) and sometimes they are loose (when under relaxed selection), allowing the clay to form all manner of shapes. But the hands can only shape what they are given, and so the genetic material of life is of undeniable importance to the final forms we see following adaptation. However, if mutations persistently appear that are unfavoured by natural selection, then the hands can work to suppress their appearance by eliminating the mechanisms that enable them. In this section we will discuss the interplay between the mechanisms facilitating biased mutation rates and the power of selection, and ask when the potter acts to control the composition of their clay.

There are clear signatures within most genomes that an organism's fitness is often not aligned with runaway mutation. DNA error repair proteins, encoded by the core genome, are ubiquitous within bacteria and throughout the domains of life (Morita et al. 2010). With the prevalence and persistence of these repair enzymes we have a clear clue that the evolutionary history of life has driven genomes to limit mutation rates. However it would not be in a cell's interest to eradicate mutation entirely, as doing so would erase their ability to evolve. Therefore, we are posed with questions of how much tolerance do bacterial populations have for mutational load? And at what point does the cost of mutation driving genetic change become a price too heavy to pay? These questions are often answered by natural selection.

Under stable, natural environments, populations are often in positions where most genetic changes will be neutral or deleterious (Kimura 1968). The chance of a mildly deleterious, neutral or advantageous mutation reaching fixation under such conditions will depend on population size and its rate of change (Otto and Whitlock 1997), but in general there is limited adaptive potential for new mutations. As such microbes with high mutation rates will find themselves subject to purifying selection, as by chance mutations will hamper, not improve, their fitness. The consequence of this is that global mutation rate will selectively be driven down, such as by encouraging the acquisition and persistence of DNA repair proteins and evolving to remove large, disruptive secondary structures throughout the chromosome (Leach 1994).

Yet microbes do not perpetually persist in a state of bliss. When introduced to a new environment, especially a stressful one, the ability to rapidly adapt can ascend to paramount importance. In these instances, the mutability pendulum swings toward mutator lines that can adapt more rapidly (Swings et al. 2017). Mutator lines can be a boon to a population's fitness even if present in low frequency (Taddei et al. 1997), meaning that if but a few lines evolve to break their DNA repair mechanisms they can exploit the opportunity of heightened mutation rates. An example of this can be found in the pathogenic bacteria *P. aeruginosa*, which readily infects human lungs affected by cystic fibrosis (Bhagirath et al. 2016). Sequenced strains of *P. aeruginosa* often have functional mismatch repair, but foundling cells entering a host will go through a bottleneck that reduces population size. Once inside, however, they are met with a harsh environment which can select for hyper-mutability (Oliver 2010) sometimes through the removal of mismatch repair (Dettman et al. 2016).

Engaging a mutagenic actor such as non-functional mismatch repair proteins works to increase mutation rate nearly genome-wide (Long et al. 2014; Dettman et al. 2016). However, as elevating global mutation rates also raise the likelihood of deleterious mutations (Lynch et al. 1995) , bacteria may instead find more nuanced solutions. Rather than impacting mutation rate on a global scale, certain genomic areas may have suppressed or elevated mutation rates according to their selective advantage for the cell. In *E. coli*, for example, it has been observed that mutational hotspots are found in repeat and non-functional regions of the genome (Zhang et al. 2018). This suggests that crucial genes may be under stronger selection for limited mutation rates. A study by Martincorena and colleagues found evidence to this effect, with "mutational cold spots" found in highly expressed genes and genes under strong purifying selection (Martincorena et al. 2012). Hoede and colleagues offered a mechanism for limiting mutations in highly expressed genes, as selectively driven hairpin formation can allow exposed single-stranded DNA (which is susceptible to mutation, see previous section 1.2.1.8) to become double-stranded during transcription (Hoede et al. 2006). Similarly, essential genes have been found to be much more common on the leading strand (Rocha et al. 2003) and transcribed co-orientationally with the replication fork (Srivatsan et al. 2010), reducing the influence of mutagenic actors on these genes. As genome rearrangements facilitate gene orientation and leading-lagging strand order through gene inversions, it has synergistically been suggested that selection suppresses recombination mutants from reaching fixation (Roth et al. 1996).

If selection can supress locally acting mutagenic mechanisms, it engenders the question of whether the reverse is also true. I.e., do any mutable features of the genome coincide with areas where a higher mutation rate would be favourable? It has long been a fierce source of debate that microbial stress may increase rates of mutation (Roth et al. 2006), which would provide a neat explanation that mutagenicity occurs when it is needed as a consequence of cell biology (Katsnelson et al. 2019). However, another nuanced argument derived from a similar vein proposes that while in stationary phase, genes that are continually transcribed are both under selection – because they're still being expressed – and will mutate at higher rates because transcription often increases the likelihood of mutation (Davis 1989). This proposes that mutagenicity occurs consequentially of standard cell behaviour, but that it is tethered to regions where mutations could be adaptive. However rather than being situationally driven, it may be that highly mutable genetic features are always operating yet are preserved due to their assistance to cell evolvability.

One notable genetic feature that has proven adept at evading purifying selection is homopolymeric tracts and tandem repeats (see section 1.2.1.4). These remain pervasive in many microbial genomes, and the frameshift mutations they facilitate are selected for under certain circumstances (Wernegreen et al. 2010). When homopolymeric tracts or tandem repeats are enriched within certain coding regions or promoter regions they allow genes to be rapidly switched on and off through mutation (Moxon et al. 2006). Bacteria have evolved sophisticated regulatory hierarchies that allow them to augment their gene expression to environmental triggers and as such cope with environmental change (Shis et al. 2018). Some bacterial species utilise regulatory pathways to drastically alter their activity and endure unfavourable environments through the process of sporulation (de Hoon et al. 2010). However the 'phenotypic state-switching' facilitated by mutable homopolymeric tracts and tandem repeats allows certain genes to respond to environmental perturbations using an alternative option of mutation (Moxon et al. 1994).

Moxon and colleagues championed the idea that these so-dubbed "contingency loci" could be selectively advantageous when they stated the following: "*We conclude by restating our thesis as a general hypothesis: mutation rates vary among sites in a genome, and this variation is adaptive because it promotes evolutionary flexibility in the face of environmental change, without necessarily increasing the overall load of deleterious mutations*" (Moxon et al. 1994). This reasoning suggests that mutation, like gene regulatory hierarchies that respond to environmental triggers, has evolved to make bacteria robust against changing conditions. The mutagenicity at these sites is expected to be high enough that variations at contingency loci will be consistently appearing and reverting. As such when the microbes are challenged with a new environment, the standing genetic variation in the population will allow at least some cells to cope with the novel phenotypic requirements and proliferate, via a process known as a soft-selective sweep (Hermisson and Pennings 2005). Evidence for standing genetic variation at contingency loci facilitating soft-selective sweeps has been found in subsequent research by Jerome and colleagues, who studied the rapid adaptation of *Campylobacter jejuni* to a new host (Jerome et al. 2011).

While homopolymeric tracts and tandem repeats allow populations to ‘flip-flop’ gene activity between environmental conditions and so often provide a selective benefit, not all mutagenic mechanisms are primed to do this. Evolving populations often adopt a ‘no going back’ approach, as once a mutation has appeared in the population it is more likely that a compensatory mutation, rather than genetic reversion, will occur when the mutation is no longer selectively advantageous. This is because the pool of mutational targets for compensatory mutations are often much larger than the single mutation that leads to genetic reversion (Poon and Chao 2005). Contingency loci that are built on highly mutable strand-slippage events circumvent this problem, however, as both the initial mutation and the genetic reversion are highly mutable. I.e. the same mutagenic mechanism is responsible for hyper-mutability in both mutational directions at the same site (Zhou et al. 2014).

In this sense, strand-slippage mutagenesis may occupy a unique zone where the mutagenic mechanism can exist stably in a population. This is not the case for other mutagenic mechanisms. For example, template-switching mutations that occur as a result of hairpin formation drive imperfect repeat sequences into perfect repeats (De Boer and Ripley 1984; Dutra and Lovett 2006; Lavi et al. 2018). The mechanism driving this therefore only works in one direction; once the quasipalindrome has become a perfect palindrome, the same mechanism will not revert the sequence back into its previous quasipalindromic form. The result is that such mutations irrevocably lose information – the gene’s sequence has changed, and the mutagenic hotspot has been lost. Unless another mutagenic mechanism is driving genetic reversion at the same position, hotspots will be readily lost once mutation has occurred. Compensatory mutations will then likely drive subsequent adaptation elsewhere in the genome. Mechanisms such as head-on collisions from antagonistically oriented genes occupy a niche between these two extremes, whereby the mutational hotspot is not lost following mutation, but compensatory mutations elsewhere in the mutable locus may occur rather than genetic reversion. It is also worth noting that hotspots can also be generated through largely neutral processes i.e., synonymous variation in the case of repeat regions. Therefore, selection may suppress many mutational hotspots over multi-generations and enforce the stability of some others, but it cannot entirely prevent the genesis of new highly mutable sites.

1.4. A Model Microbe: *Pseudomonas fluorescens*

"Essentially, all life depends upon the soil… There can be no life without soil and no soil without life; they have evolved together." - Charles E Kellogg, USDA Yearbook of Agriculture (1938)

The *Pseudomonas* genus was first characterised, albeit very briefly, in the late nineteenth century by German Professor Walter Migula (Henry 2012). The elements *pseudo* and *monas* are Greek derivations meaning "false" and "unit" respectively. Although Migula did not justify his reasoning for using these derivations, the praenomen title of *pseudo* would turn out to be almost prescient. The genus soon collected a bloated repertoire of bacterial species, many of which would be later to reassigned to more appropriate genera (Palleroni 2010). Many species do however remain in the genus, and all match Migula's brief description – the opening statement of which reads that Pseudomonads are: "*Cells with polar organs of motility*" (Henry 2012). These polar flagella, as they are known today, form the phenotypic centrepiece of the model system described throughout this work.

Pseudomonas fluorescens, the 'fluorescent' Pseudomonad, is one species of *Pseudomonas* to enjoy a descriptive phenotypic title. *P. fluorescens* is a gram-negative, rod-shaped, polar-flagellated, mostly obligately aerobic and non-pathogenic saprophytic species that predominantly colonizes soil, water and plant surfaces including roots and leaves (Ganeshan and Kumar 2007). *P. fluorescens* strains have been shown to exhibit numerous functions in these diverse ecosystems (Silby et al. 2011) and the species is trans-continental, with strains being found across Europe from Spain to Poland and wider still, in Taiwan, Australia, and North America (Ganeshan and Kumar 2007). Many strains aid in the growth of an array of crop types (Ganeshan and Kumar 2007), which is partly owed to their antimicrobial output (O'Sullivan and O'Gara 1992) that counteracts fungi and oomycetes (Haas and Défago 2005).

P. fluorescens are rarely affiliated with human hosts, owed in part to their limited pathogenicity (a trait first discovered by scientists Baader and Garre who opted to swallow *P. fluorescens* to assess its ability to harm the gastro-intestinal tract (Scales et al. 2014)). However more recent work has revealed that *P. fluorescens* is associated with certain human diseases, including Chron's disease (Scales et al. 2014), an affiliation that highlights just one of the diverse ecosystems capable of colonisation by this bacterial species. Combining this range with the diversity of the strains genetically (Silby et al. 2009), and in their geographic distribution, has led to the species being described more-so as a broader "species complex" rather than a species (Silby et al. 2011; Scales et al. 2014).

Two strains of *P. fluorescens* are utilised as ancestral genetic backgrounds throughout this work. SBW25, which was first isolated from a leaf of a sugar beet plant in Oxford, UK, 1989 (Rainey and Bailey 1996); and Pf0-1, which was first isolated from loam soil in Massachusetts, USA, in 1987 (Compeau et al. 1988). Comparative genomics performed by Silby and colleagues revealed that these two strains differ fairly substantially in genome size, gene number and the proportion of shared genes (Silby et al. 2009). SBW25 boasts a genome of approximately 6.7 Mb harbouring 6009 coding sequences, and Pf0-1 a genome of approximately 6.4 Mb and 5,742 coding sequences. Of these, only 4109 genes are shared across the two strains – representing 71.6% of Pf0-1's genome and 68.4% of SBW25's genome – meaning that just less than one third of the genome is divergent between the two (Silby et al. 2009). Amino acid identities between the strains additionally placed their homology between the species and genus boundary (Silby et al. 2009). Furthermore, Pf0-1 has also been identified as a natural *gacA* mutant, which debilitates the GacS/GacA two-component system associated with antifungal activity, biofilm formation and motility (Seaton et al. 2013). The GacS/GacA system, in contrast, is active in wild type strains of SBW25 (Cheng et al. 2016).

The two strains are not without their similarities, however. They share considerable synteny toward the origin of replication, and each strain has a large complement of over 100 genes related to chemotaxis and motility (Silby et al. 2009). They both also possess many regulatory elements (Silby et al. 2009), of which they share close homology and network architecture of the nitrogen gene regulatory pathway (Taylor et al. 2015). Therefore, selective pressures that focus adaptation on genes comprising motility and nitrogen regulation will operate similarly across the two strains. This offers an opportunity to observe a similar spectrum of mutational targets when both strains are placed under directional selection, but also allows for an investigation of the role played by diverse genetic backgrounds in driving evolutionary outcomes.

Previous work by Alsohim and colleagues utilised strain SBW25 in their investigations into *P. fluorescens* motility. Firstly, they observed that a functional deletion of the gene encoding the master regulator of flagella synthesis, FleQ, denied SBW25 strains access to flagella-mediated motility (Alsohim et al. 2014). FleQ is an enhancer-binding protein that is part of the NtrC-like transcription factor family (Jyot et al. 2002). Structurally, the FleQ protein performs its role of regulating gene expression through three major domains. These consist of a receiver (REC) domain which can respond to environmental signals, a σ^{54} (RpoN) interaction domain, and a helix-turn-helix DNA binding domain (Bush and Dixon 2012). Blanco-Romero and colleagues implicated a plethora of genes that are under control of FleQ in *P. fluorescens* F113 and *P. putida* KT2440 (Blanco-romero et al. 2018). These included genes involved in iron chelation, secretion systems, and a suite of genes involved in flagella biosynthesis and flagellar basal-body proteins (Blanco-romero et al. 2018). This wide influence on flagella genes is owed to FleQ's core position at the top of a tiered regulatory hierarchy (Dasgupta et al. 2003) which includes the two-component system *fleSR* and proteins involved in flagella export (Jyot et al. 2002). Thus a debilitating mutation in the *fleQ* locus will exert downstream consequences on a battery of genes (Dasgupta et al. 2003) and erase the flagella motility phenotype (Hickman and Harwood 2008).

In *P. aeruginosa* evolving to become non-motile can be advantageous e.g. through *fleQ* mutation, as losing flagellin can be a key means to evade a host's immune response (Hayashi et al. 2001; Faure et al. 2018). However, in *P. fluorescens* motility is often aligned with high fitness. For example, the phenotype has been shown to improve both the attachment and colonisation of wheat roots in soil over non-motile lines (Turnbull et al. 2001). In the event of FleQ removal, Alsohim and colleagues observed that strain SBW25 had a phenotypic contingency in the form of viscosin-mediated biosurfactant, which allowed for an alternate form of sliding motility (Alsohim et al. 2014). However in a follow-up study, Taylor and colleagues found that when this form of sliding motility was additionally removed, flagella-mediated motility would be swiftly recovered within a matter of days (Taylor et al. 2015). The authors repeated this experiment using a *fleQ* defective mutant of Pf0-1, and observed that this strain too was able to rapidly re-evolve flagella motility (Taylor et al. 2015). Both strains resurrected motility through a one-step *de novo* mutation.

In strain SBW25 independent lines fixed mutations persistently within the *ntrB* locus, which encodes for the histidine kinase belonging to the two-component system of the nitrogen regulatory network. In strain Pf0-1 the observed mutations were more variable but still confined to loci belonging to the same regulatory network (Taylor et al. 2015). Mutations were found in *ntrB* but also in *glnK*, which encodes for a NtrB's binding partner that determines its phosphatase activity. Mutations were also observed in *glnA*, which encodes for glutamine synthetase (GlnA), a gene under the control of the response regulator of the nitrogen pathway's two-component system, NtrC. Although the observed mutations differed in their mutational target, each produced the same output – the hyper-phosphorylation of NtrC. As introduced above, this protein belongs to the same transcription factor family as FleQ. However Taylor and colleagues revealed the two proteins had retained sufficient homology that a hyper-active NtrC can drive the expression of FleQ-dependent genes in its absence and recover flagella-mediated motility (Taylor et al. 2015).

A visual overview of the nitrogen regulatory pathway in *P. fluorescens* is provided in Fig. 1.2. At the nexus of this network lies the *ntrB/ntrC* two-component system. When phosphorylated by NtrB, NtrC is responsible for driving the expression of core genes within the network. This includes its own operon (Taylor et al. 2015), NtrB's binding partner GlnK (Hervas et al. 2009), the ammonia transport membrane protein AmtB (Hervás et al. 2008), and glutamine synthetase (M. J. Merrick and Edwards 1995), which catalyses the synthesis of glutamine by fusing glutamate and ammonia (Fig. 1.2). The availability of cellular nitrogen is fed back into the sensor of the network, UTase, using ratios of available glutamine and 2-ketoglutarate in the cell (M. J. Merrick and Edwards 1995). As such when glutamine synthetase has driven the production of sufficient glutamine, UTase will cease to catalyse the addition of uridine monophosphate (UMP) onto GlnK (M. J. Merrick and Edwards 1995). This consequentially allows GlnK to associate with NtrB (Hervas et al. 2009), which stops the histidine kinase's phosphorylation activity of NtrC. The nitrogen regulatory network, therefore, provides a neat feedback loop that modulates the phosphorylation and activity of NtrC. However, during an evolutionary event wherein high NtrC activity is needed to drive the expression of flagella genes, each of these loci presents itself as a potential mutational target.

Strong directional selection often results in the acquisition of loss-of-function mutations (Kimura 1968; Lind et al. 2015). The rapid evolution of the nitrogen pathway to facilitate hyper-phosphorylation of NtrC is no different. The observed mutations in *ntrB* were each missense mutations, which likely prevent the kinase's ability to associate with GlnK and so de-sensitise it from negative feedback from the network (Taylor et al. 2015). Likewise mutations in *glnK*, which can be catastrophic frameshift mutations (Taylor et al. 2015), act to cripple GlnK activity and prevent its association with NtrB, once again de-sensitising the kinase from negative feedback. Mutations in *glnA* instead offer an indirect physiological route to increased NtrC-phosphorylation (Fig. 1.2; Taylor et al. 2015). It is likely that mutations in this locus hamper glutamine synthesis, meaning the ratio of glutamine to 2-ketoglutarate would not switch sufficiently to prevent UTase catalysis of UMP to GlnK (Fig. 1.2). Yet although *glnK* and *glnA* mutations were readily observable in Pf0-1, they were curiously absent in SBW25. This asked whether the broad differences in genetic background meant these two genes were not as beneficial to phenotype in one strain, or whether the discrepancy was explained by genetic differences that differentially affected mutability of the *ntrB* locus.

1.4.1. A pathway of questions: Uncovering a mutational hotspot at the nexus of a regulatory network

At the outset of this work, we sought to determine the key evolutionary forces that drove the evolution of a key gene regulatory network, allowing it to re-wire to control a secondary major pathway (Taylor et al. 2015). We endeavoured to investigate the role played by the environment; as the presence of amino acids glutamate and glutamine, as well as ammonia, are critical molecules in regulating the pathway (Fig. 1.2) and their abundance may have altered which of the genes remained viable mutational targets. We also endeavoured to assess the role played by pleiotropy, epistasis and redundancy facilitated by genes outside the core network. Lastly, we searched for genetic features that may be biasing the mutational re-wiring event which allowed for the recovery of a core phenotype. The two strains of *P. fluorescens* SBW25 and Pf0-1 provided us with natural homologs to investigate the role of broad genetic background and genetic sequence divergence on realising these evolutionary outcomes. We were able to ask why the two strains targeted different genes during the evolution of motility. I.e., what elements of their genomes are driving this change – is it the broad divergences of loci between them, or is it something intrinsic to the loci of the network itself? Then, once this was resolved, we could ask what implications do these different mutational outcomes have for the adaptive potential of these two strains?

In answering these questions, this work uncovers a prominent role for mutation bias which is predicated on innocuous-appearing genetic variation. It subsequently reveals that the mutational hotspots which are so integral to driving evolution are both readily acquired through synonymous variation and readily lost following evolution. Thus *P. fluorescens* is utilised as a key model organism that shows powerful evolutionary outcomes are not merely the product of the 'coding genome', but also the 'mutable genome' that lays underneath. It reveals that this mutable genome is transient, constantly in a state of flux following the guiding hands of natural selection and genetic drift. As we turn toward predicting evolutionary outcomes in future work, these core principles teach us of the power of the mutable genome, and that it must be understood if we are to ever make accurate evolutionary forecasts.

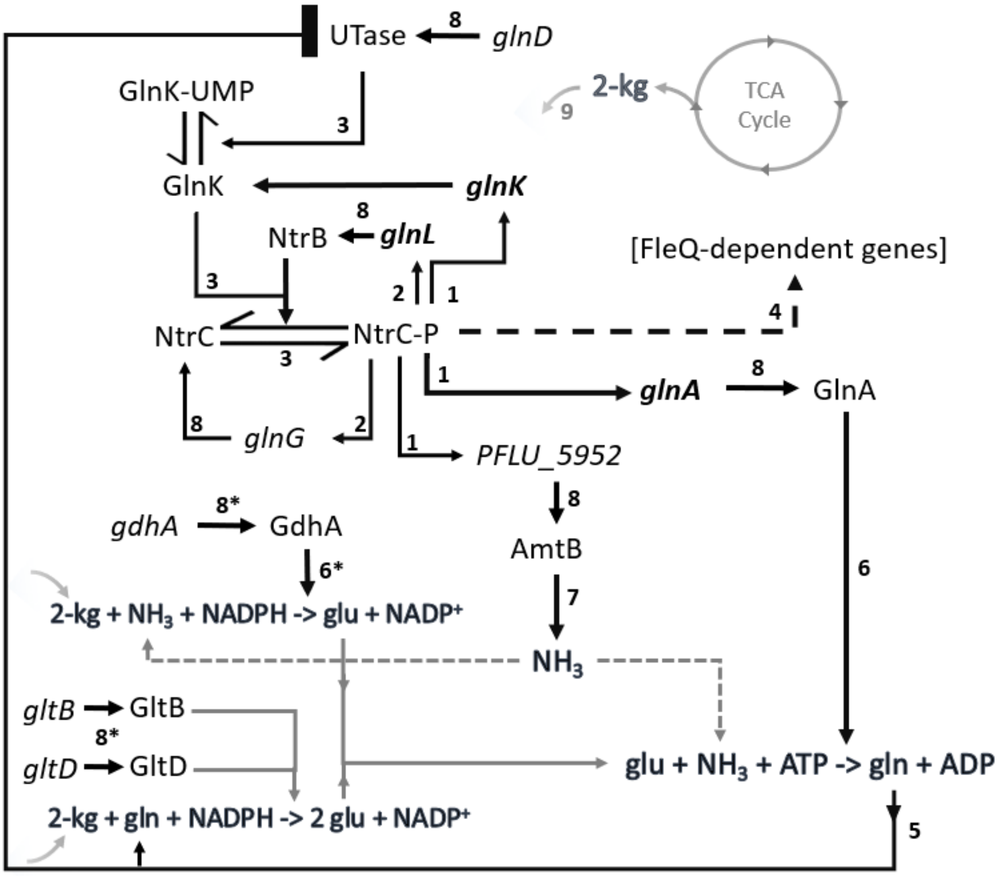
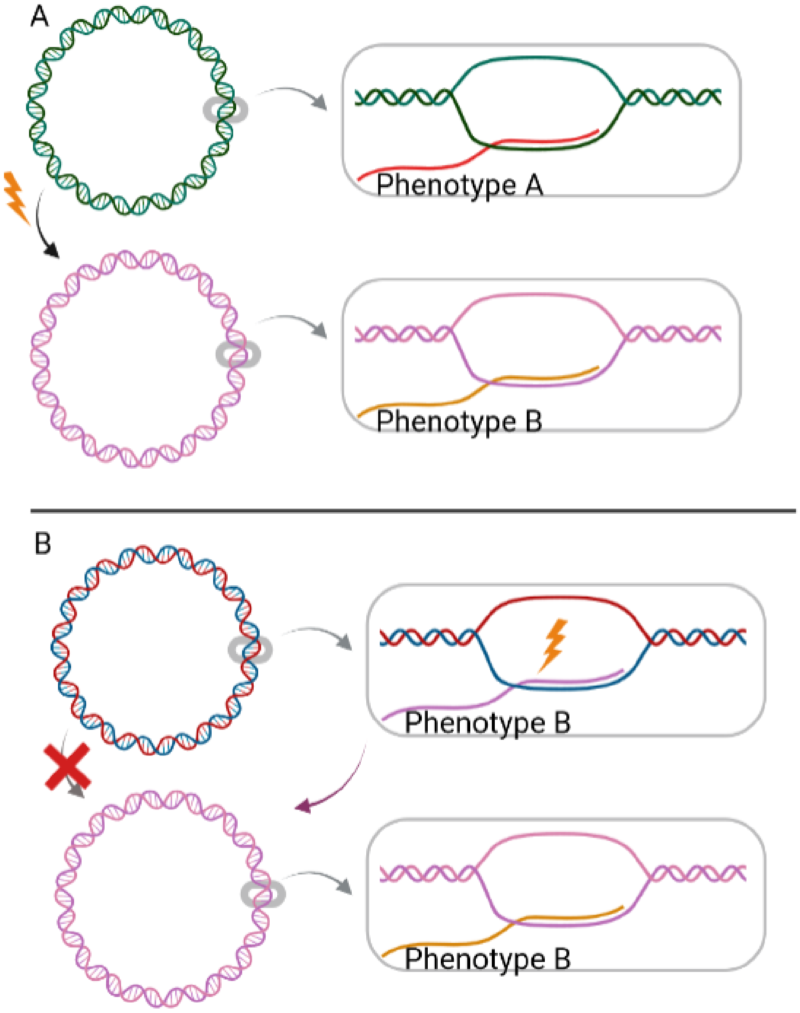


Figure 1.2. The nitrogen regulatory circuit in *Pseudomonas fluorescens*. Arrow key (citations that demonstrate network connections are included in the figure legend): **1)** Direct regulation of expression by a transcription factor (M. J. Merrick and Edwards 1995; Hervas et al. 2009) **2)** Suspected auto-regulation by a transcription factor (Taylor et al. 2015) **3)** Direct regulation of protein activity through post-translational modifications (M. J. Merrick and Edwards 1995; Hervas et al. 2009). **4)** Transient regulation via spurious transcription factor binding (Taylor et al. 2015) **5)** Indirect regulation through intermediate molecular actors. When glutamine (gln) >> 2-ketogluric acid (2-kg) we observe UTase repression, when gln << 2-kg we observe UTase activity (M. J. Merrick and Edwards 1995) **6)** Amino acid synthesis (M. J. Merrick and Edwards 1995) **7)** Active molecule transport into the cytoplasm (Zheng et al. 2004) **8)** Translation of mRNA transcripts. **9)** 2-ketoglutaric acid (2-kg) is produced as an intermediate of the tricarboxylic acid (TCA) cycle and plays a role in both nitrogen and carbon regulatory networks (Li and Lu 2007; Huergo and Dixon 2015). *****Secondary pathways: glutamate dehydrogenase (gdh) and glutamate synthase (glt) (M. J. Merrick and Edwards 1995). Genes are *italicised*. Observed mutational targets during the evolution of motility are highlighted in **bold**.

1.5. Supplementary materials



Supplementary Figure 1.1. Mutation immortalisation typically precedes selection, but this is not the case for retromutagenesis. (**A**) An integral gene within a parent cell causes phenotype A. During replication a mutation within the gene is immortalised on both strands, resulting in phenotype B when the gene is transcribed and translated in the daughter cell. As such mutation is immortalised during replication and the change in phenotype is subsequently acted upon by selection in the daughter cell. (**B**) In some instances, such as when a cell is starving and its current genetic arsenal is unable to metabolise a nutrient in the environment, replication will cease to occur. However, as transcription in the cell will continue, a mutation on the template strand which is transcribed and translated produces an immediate change in phenotype in the parent cell. If such a change allowed for the cell to replicate e.g. it allowed for the previously unusable nutrient in the environment to be metabolised, then the genome will divide and the mutation has a chance of being immortalised. If it is immortalised, then the daughter cell will retain the amended phenotype which evolved in the parent. In these cases selection precedes mutation immortalisation, and as viable mutational routes are limited to the template strands of certain coding regions, this mechanism may incorrectly infer mutation bias.

Chapter II

2.1. Appendix 6B: Statement of Authorship

This declaration concerns the article entitled:	
A mutational hotspot that determines highly repeatable evolution can be built and broken by silent genetic changes	
Publication status (tick one) **Draft manuscript** □ **Submitted** □ **In review** □ **Accepted** □ **Published** ☒	
Publication details (reference)	A revised version of the included manuscript was published in *Nature Communications* on 19.10.2021 under Creative Commons license: Attribution 4.0 International (CC BY 4.0). The published manuscript can be accessed here: https://doi.org/10.1038/s41467-021-26286-9.
Copyright status (tick the appropriate statement) I hold the copyright for this material ☒ Copyright is retained by the publisher, but I have been given permission to replicate the material here □ **Permission to replicate the material is in process. In the case of difficulties, a PGR7 form will be submitted.*	
Candidate's contribution to the paper (provide details, and also indicate as a percentage)	The candidate contributed to / considerably contributed to / predominantly executed the… Formulation of ideas: **90%** The experimental design, interpretation of data and research niche explored by this manuscript was predominantly the work of the first/co-corresponding author (the candidate). Supervision and support was provided by co-corresponding author T. B. Taylor (regular meetings to provide feedback and share ideas with the first author, advice on statistical measures), with additional advice from co-author R. W. Jackson (expert insight into the employed model system). Design of methodology: **100%** The first/co-corresponding author designed all the novel experiments included in the manuscript. Experimental work: **95%** All experimental work was completed by the first/co-corresponding author aside from the directed evolution and sequencing work for the ancestral Pf0-2x lines, which was completed by co-author L. M. Flanagan. These data are included in Fig. 2.5B.

	Presentation of data in journal format: **90%** The original manuscript draft, and the major revisions written into the current version of the manuscript were completed by the first/co-corresponding author. Interim draft sections were penned by, and rounds of editing/feedback were provided by, co-authors T.B. Taylor, N. K. Priest and R. W. Jackson. Data visualisation assistance and advice was provided by co-corresponding author T. B. Taylor.		
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.		
Signed	[redacted]	**Date**	**07.11.2021**

2.2. Title page: A mutational hotspot that determines highly repeatable evolution can be built and broken by silent genetic changes

Authors: James S. Horton[1,*], Louise M. Flanagan[1], Robert W. Jackson[2], Nicholas K. Priest[1,a], and Tiffany B. Taylor[1,*,a]

[1]Milner Centre for Evolution, Department of Biology & Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

[2]School of Biosciences and Birmingham Institute of Forest Research (BIFoR), University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

[a] These authors share senior authorship

***Corresponding Authors:** James S. Horton, Milner Centre for Evolution, Department of Biology & Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK, +44 (0)1225 385116, j.s.horton@bath.ac.uk.

Tiffany B. Taylor, Milner Centre for Evolution, Department of Biology & Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK, +44 (0)1225 384398, t.b.taylor@bath.ac.uk

Author Contributions: Criteria taken from McNutt et al., 2018. PNAS. (1) Substantial contributions to the conception or design of the work; (2) the acquisition, analysis, or interpretation of data; (3) drafted the work or substantively revised it; (n/a) the creation of new software used in the work: James S. Horton[1,2,3], Louise M. Flanagan[2], Robert W. Jackson[1,3], Nicholas K. Priest[3] and Tiffany B. Taylor[1,3]

Competing Interest Statement: The authors declare no competing interests.

Preprint Available At: https://doi.org/10.1101/2021.01.04.425178 [Earlier version]

Keywords: Repeatable evolution, Mutational hotspot, Silent mutation, Mutation rate heterogeneity, Experimental evolution

2.3. Abstract

Mutational hotspots can determine evolutionary outcomes and make evolution repeatable. Hotspots are products of multiple evolutionary forces including mutation rate heterogeneity, but this variable is often hard to identify. In this work we reveal that a powerfully deterministic genetic hotspot can be built and broken by a handful of silent mutations. We observed this when studying homologous immotile variants of the bacteria Pseudomonas fluorescens, AR2 and Pf0-2x. AR2 resurrects motility through highly repeatable de novo mutation of the same nucleotide in >95% lines in minimal media (ntrB A289C). Pf0-2x, however, evolves via a number of mutations meaning the two strains diverge significantly during adaptation. We determined that this evolutionary disparity was owed to just 6 synonymous variations within the ntrB locus, which we demonstrated by swapping the sites and observing that we were able to both break (>95% to 0%) and build (0% to 80%) a powerfully deterministic mutational hotspot. Our work reveals a fundamental role for silent genetic variation in determining adaptive outcomes.

2.4. Introduction

Mutational hotspots, which describe instances where independent cell lines persistently fix mutations at the same genomic sites, can make evolution remarkably repeatable. Such hotspots are of immense importance as they have been observed to drive evolution across the domains of life, from viruses (including SARS-CoV-2; Weber et al. 2020), to bacteria (including MRSA; Sekowska et al. 2016), to higher eukaryotic cell lines including those in avian species (Galen et al. 2015) and human cancers (Trevino 2020). Our understanding of evolutionary dynamics (e.g. competitive selection and clonal interference) can sometimes explain the appearance of hotspots, but genetic features that build hotspots by biasing mutation rates are much less understood.

There have been many examples of experimental systems evolving via repeatable evolution. Microbes evolving under strong selection often rapidly adopt similar novel phenotypes (Fong et al. 2005; Ostrowski et al. 2008). Furthermore, these phenotypes are often underpinned by mutation hotspots, which come in the form of clustered genetic changes within the same region of the genome (Riehle et al. 2001; Fraebel et al. 2017), or within limited pockets of loci (Bull et al. 1997; Wichman et al. 1999; Herron and Doebeli 2013; Kram et al. 2017). Sometimes realised mutations are found only in genes from a single regulatory pathway (Notley-McRobb and Ferenci 1999; Miller et al. 2013) or a single protein complex (Avrani et al. 2017). In extreme cases, evolutionary events can be seen to repeatedly target just a handful of sites within a single locus (Meyer et al. 2012; Van Ditmarsch et al. 2013). Repeatable evolution allows lines to evolve in parallel, and the degree of parallelism typically becomes less common as it descends from broader genomic regions to the nucleotide (Tenaillon et al. 2012; Bailey et al. 2015). However, despite frequent descriptions of repeatable evolutionary events, a detailed understanding of the hotspots that ensure their occurrence is often lacking.

There are three primary facilitators of mutational hotspots that drive repeatable evolution: (*i*) Fixation bias, which skews evolution toward mutations that enjoy a higher likelihood of dominating the population pool. Not all facilitators of fixation bias are considered adaptively advantageous (e.g. homologous recombination events in mammalian genomes can bias gene conversion toward certain alleles; Eyre-Walker and Hurst 2001). But in instances where we observe rapid and highly parallel sweeps fixation bias will likely take the form of selection, which drives the fittest competing genotypes in the population to fixation (Wood et al. 2005; Woods et al. 2006). (*ii*) Mutational accessibility, as there may be only a small number of readily accessible mutations a genotype can undergo to improve fitness (Weinreich et al. 2006). And, (*iii*) Mutation bias, where genetic and molecular features scattered throughout the genome cause sites to mutate at different frequencies and toward certain mutation types (for example, A:T → G:C), constraining the mutational spectrum to favour particular outcomes (Bailey et al. 2017). Previous research shows that mutation rate heterogeneity can be influenced by the arrangement of nucleotides surrounding a particular site (Long et al. 2014), and genetic features such as the secondary structure of DNA (Duan et al. 2018) including the formation of single-stranded DNA hairpins (De Boer and Ripley 1984). Nevertheless, the prominence of genetic sequence in driving parallel evolutionary outcomes remains unknown.

To establish which mechanisms are at play, it is important to consider whether parallel outcomes are robust to experimental conditions such as environment (Turner et al. 2018) and to account for clonal interference, which can alter the chance of observing parallel evolution (Bailey et al. 2017; Lässig et al. 2017). Clonal interference can occur either due to standing genetic variation in the founder population which yields multiple adaptive genotypes in a novel environment (i.e. a soft selective sweep; Hermisson and Pennings 2005) or when mutation rate is high relative to the selective coefficient (Barrett et al. 2006). Clonal interference does not often play an important role when founding experimental lines with clonal samples, performing experimental procedures over short timescales, and ensuring rapid fixation of adaptive mutants e.g. through spatial separation and/or introducing an artificial bottleneck. However, under such conditions the primary influence of selection will manifest as clonal interference, as a large starting populating may give rise to multiple adaptive genotypes which compete for fixation throughout the course of the experiment (Jerison and Desai 2015).

In this work, we have utilised an ideal system for identifying the key features that build mutational hotspots. We have employed two engineered non-flagellate and biosurfactant-deficient strains of the soil bacteria *P. fluorescens*: AR2, derived from SBW25, and Pf0-2x, derived from Pf0-1 (see materials and methods). The strains share homologous genetic backgrounds, including highly similar gene regulatory architectures and translated protein products, yet they evolve divergently. Both engineered strains lack function of the master regulator of flagella-dependent motility, FleQ, and both AR2 and Pf0-2x rapidly re-evolved flagella-mediated motility under strong directional selection (Taylor et al. 2015). In AR2, this phenotype was achieved in independent lineages via repeatable *de novo* mutation in the *ntrB* locus of the nitrogen regulatory (ntr) pathway. The parallel evolution of *ntrB* mutants was noteworthy as the locus was consistently targeted, whereas Pf0-2x lines evolved motility via mutations across the ntr regulatory hierarchy (Taylor et al. 2015). As such parallel evolution between these homologs varied across scale; both were parallel to the phenotype and targeted gene regulatory network, but only one possessed a mutational hotspot that concentrated mutations at a single nucleotide site within a single locus. We conducted a series of experiments to find out why.

Here we show that motility evolves in AR2 in an extremely repeatable manner, which is absent in Pf0-2x due to a genetic feature predicated on synonymous variation. The evolution of flagella motility in AR2 was found to target the same nucleotide substitution in over 95% of cases in minimal medium (M9). This outcome was found to be robust across multiple nutrient regimes both in the immotile SBW25 variant (AR2) and another SBW25 variant that was able to access biosurfactant-mediated motility prior to evolution (SBW25 Δ*fleQ*). The role of selection and the number of viable mutational routes in ensuring the parallel outcome were found to provide some explanation for parallel evolution to the level of the *ntrB* locus, but not the nucleotide. This therefore implied that intra-locus mutational biases were playing a critical role. We then genetically augmented the *ntrB* locus to indirectly incriminate mutation bias and revealed a key underlying genetic driver of parallel evolution. Six silent nucleotide changes were introduced within the local region around the frequently targeted site to make AR2’s genetic sequence match Pf0-2x, but without altering the protein product. These changes were found to reduce parallel evolution at the mutational hotspot from >95% to 0%. In a reciprocal experiment, silent changes introduced to the homologous strain Pf0-2x to match AR2’s local native sequence raised parallel evolution at this site from 0% to 80%. These results reveal that synonymous genetic sequence can play a dominant role in ensuring parallel evolutionary outcomes, and shines a spotlight on the overlooked mechanistic drivers behind mutational hotspots.

2.5. Materials and Methods

2.5.1. Model System

Our model system employs strains of the soil microbe *P. fluorescens* SBW25 and Pf0-1 that lack motility through partial gene deletion or disruption of *fleQ*, the master regulator of flagellar motility (Robleto et al. 2003; Alsohim et al. 2014). Motility can be recovered in the absence of *fleQ* following *de novo* mutation that allows for the recruitment of a homologous response regulator, of which the most readily targeted is *ntrC* of the nitrogen regulatory pathway. The initial mutation that facilitates *ntrC* recruitment occurs in other loci in the nitrogen pathway, resulting in the hyper-phosphorylation of *ntrC* (Taylor et al. 2015). Two SBW25-derived strains were used as ancestors in this study: SBW25 Δ*fleQ* (hereafter Δ*fleQ*) and a Δ*fleQ* variant with a functional *viscB* knockout isolated from a transposon library (SBW25Δ*fleQ* IS-ΩKm-hah: PFLU2552, hereafter AR2; Alsohim et al. 2014). Δ*fleQ* can migrate on soft agar (0.25%) prior to mutation via a form of sliding motility, which is owed to the strain's ability to produce viscosin. AR2 cannot produce viscosin and is thus rendered completely immotile prior to mutation. Pf0-1 is a native *gacA* mutant (Seaton et al. 2013) thus does not make viscosin, therefore its Δ*fleQ* variant, Pf0-2x, is rendered completely immotile following disruption of *fleQ*. All cells were grown at 27°C and all strains used throughout the study (ancestral, evolved and engineered) were stored at -80°C in 20% glycerol. The nutrient conditions used throughout the work were lysogeny broth (LB) and M9 minimal media containing glucose and 7.5 mM NH_4. The minimal media was used in isolation or supplemented with either glutamate (M9+glu) or glutamine (M9+gln) at a final supplement concentration of 8 mM unless stated otherwise.

2.5.2. Motility Selection Experiment

Immotile variants were placed under selection for flagella-mediated motility using LB and M9 soft agar (0.25%) motility plates. Details of agar preparation are described in Alsohim et al. 2014. Supplemented concentrations of glutamate (glu)/glutamine (gln) in M9 soft agar were expanded to include final concentrations at 4 mM, 8 mM and 16 mM, as it was observed that biosurfactant-mediated dendritic motility in Δ*fleQ* lines was enhanced at higher supplement concentrations, which masked any emergent blebs (data not shown). Lowering the gln supplement concentration improved the likelihood of observing an emergent flagella bleb in M9+gln motility plates (16 mM: 4/12, 8 mM: 9/20, 4 mM: 7/12 independent lines). However, dendritic motility remained high on all supplements of M9+glu and persistently masked blebbing (16 mM: 2/12, 8 mM: 3/20, 4 mM: 2/11 independent lines). Although gln/glu supplementation had no bearing on motility in AR2 lines, supplement conditions across both gln/glu were expanded for consistency. Each motility plate was seeded with a single clonal colony derived from a streak plate prepared from clonal cryogenic stock. Initiating the assay with a colony minimised the number of generations from the clonal cryogenic ancestor to the initiation of the assay, helping to ensure a clonal starting population. A single colony was inoculated into the centre of the agar using a sterile pipette tip and monitored daily until emergence of motile bleb zones (as visualised in Fig. 2.1A). Samples were isolated from the leading edge, selecting for the strongest motility phenotype on the plate, within 24 h of emergence and streaked onto LB agar (1.5%) to obtain a clonal sample. As Δ*fleQ* lines were motile via dendritic movement prior to re-evolving flagella motility and could visually mask flagella-mediated motile zones, samples were left for 120 h prior to sampling from the leading edge of the growth. An exception was made in instances where blebbing motile zones were observed solely further within the growth area, in which case this area was preferentially sampled.

2.5.3. Sequencing

Motility-facilitating changes were determined through PCR amplification and sequencing of *ntrB, glnK* and *glnA* genes (Supplementary Table 2.1). Polymerase chain reaction (PCR) products and plasmids were purified using Monarch® PCR & DNA Cleanup Kit (New England Biolabs) and Sanger sequencing was performed by Eurofins Genomics. A subset of AR2 samples evolved on different nutritional backgrounds was additionally screened through Illumina Whole-Genome Sequencing (WGS) by the Milner Genomics Centre and MicrobesNG (LB: n = 5, M9: n = 6, M9+gln: n = 6, M9+glu: n = 7). This allowed us to screen for potential secondary mutations and to identify rare changes in motile strains with wildtype *ntrB* sequences. We observed no adaptive secondary mutations within motile lines that underwent WGS, however all AR2-derived strains shared variations from the SBW25 assembly genome at the same 5 positions: 45877 A → AG, 985332 G → GC, 1786536 A → G, 3447980 TCC → T, and 3694384 A → G. The commonality of these mutations strongly indicates that the background AR2 line differs from the reference genome at these positions. *P. fluorescens* SBW25 genome was used as an assembly template (NCBI Assembly: ASM922v1, GenBank sequence: AM181176.4) and single nucleotide polymorphisms were called using Snippy with default parameters (Seemann 2015) through the Cloud Infrastructure for Microbial Bioinformatics (CLIMB; Connor et al. 2016). In instances where coverage at the called site was low (≤10x), called changes were confirmed by Sanger sequencing.

2.5.4. Soft Agar Motility Assay

Cryopreserved samples of AR2 and derived *ntrB* mutants were streaked and grown for 48 h on LB agar (1.5%). Three colonies were then picked, inoculated in LB broth and grown overnight at an agitation of 180 rpm to create biological triplicates for each sample. Overnight cultures were pelleted via centrifugation, their supernatant withdrawn and the cell pellets re-suspended in phosphate buffer saline (PBS) to a final concentration of OD1 cells/ml. 1 µl of each replicate was inoculated into soft-agar by piercing the top of the agar with the pipette tip and ejecting the culture into the cavity as the tip was withdrawn. Plates were incubated for 48 h and photographed. Diameters of concentric circle growths were calculated laterally and longitudinally, allowing us to calculate an averaged total surface area using $A= \pi r^2$. This process was repeated as several independent lines underwent a second-step mutation

(Taylor et al. 2015) within the 48 h assay. This phenotype was readily observable as a blebbing that appeared at the leading edge along a segment of the circumference, distorting the expected concentric circle of a clonal migrating population. As such these plates were discarded from the study. By completing additional sets of biological triplicates, we ensured that each sample had at least three biological replicates for analysis.

2.5.5. Invasion Assay

OD-corrected biological triplicates of *ntrB* mutant lines were prepared as outlined above. For each of the biological triplicates, 1 OD unit of *ntrB* Δ406-417 and *ntrB* A683C were mixed at equal cell densities (giving 2 OD units in total), and 1 OD unit of *ntrB A289C* was pelleted and re-suspended in the same volume as the mixed culture. 1ul of each biological replicate of the re-suspended mixed culture was used to inoculate four soft agar plates as outlined above and incubated, followed by *ntrB* A289C's inoculation into the same cavity after the allotted time had elapsed (0 h, 3 h, and 6 h). When inoculated at 0 h, biological replicates of *ntrB* A289C were added to the plate immediately after *ntrB* Δ406-417 and A683C, with each replicate seeding four soft agar plates. In instances where *ntrB* A289C was added to the plate 3h or 6 h after *ntrB* Δ406-417 and A683C, overgrowth of culture was avoided by incubating *ntrB* A289C cultures at 22°C at 0 h until cell pelleting and re-suspension approximately 1 h prior to inoculation. The same 'angle of attack' was used for both instances of inoculation (i.e. the side of the plate that the pipette tip travelled over on its way to the centre), as small volumes of fluid falling from the tip onto the plate could cause local satellite growth. To avoid the risk of satellite growths affecting results, isolated samples were collected from the leading edge 180° from the angle of attack after a period of 24 h. The *ntrB* locus of one sample per replicate was determined by Sanger sequencing to establish the dominant genotype at the growth frontier.

2.5.6. Genetic engineering

A pTS1 plasmid containing *ntrB* A683C was assembled using overlap extension PCR (oePCR) cloning (for detailed protocol see, Bryskin and Matsumura 2010) using vector pTS1 as a template. The *ntrB* synonymous mutants (AR2-sm and Pf0-2x-sm6) and AR2-sm *ntrB* A289C pTS1 plasmids were constructed using oePCR to assemble the insert sequence for allelic exchange, followed by amplification using nested primers and annealed into a pTS1 vector through restriction-ligation (for full primer list see Supplementary Table. 2.1). pTS1 is a suicide vector, able to replicate in *E. coli* but not *Pseudomonas*, and contains a tetracycline resistance cassette as well as an open reading frame encoding SacB. Cloned plasmids were introduced to *P. fluorescens* SBW25 strains via puddle mating conjugation with an auxotrophic *E. coli* donor strain ST18. Mutations were incorporated into the genome through two-step allelic exchange, using a method outline by Hmelo et al. 2015, with the following adjustments: (*i*) *P. fluorescens* cells were grown at 27^{o}C. (*ii*) An additional passage step was introduced prior to merodiploid selection, whereby colonies consisting of *P. fluorescens* cells that had incorporated the plasmid (merodiploids) were allowed to grow overnight in LB broth free from selection, granting extra generational time for expulsion of the plasmid from the genome. (*iii*) The overnight cultures were subsequently serially diluted and spot plated onto NSLB agar + 15% (wt/vol) sucrose for AR2 strains and NSLB agar + 5% (wt/vol) sucrose for the Pf0-2x strain. Positive mutant strains were identified through targeted Sanger sequencing of the *ntrB* locus. Merodiploids, which have gone through just one recombination event, will possess both mutant and wild type alleles of the target locus, as well as the *sacB* locus and a tetracycline resistance cassette. However, the wild type allele, *sacB* and tetracycline resistance will be subsequently lost following successful two-step recombination. We therefore also screened these mutant strains for counter-selection escape through PCR-amplification and sequencing of the *sacB* locus and growth on tetracycline. Mutants were only considered successful if there was no product on an agarose gel following amplification of *sacB* alongside appropriate controls, the lines were sensitive to tetracycline, and PCR results of the target locus reported expected changes at the targeted sites.

2.5.7. Statistics

All statistical tests and figures were produced in R (R Core Team 2014). Figures were created using the *ggplot* package (Wickham 2016). Simulated datasets were produced for the Bootstrap tests by randomly drawing from a pool of *n* values with equal weights *x* times for 1 million iterations. Note that for the test examining the mutational spectrum when discussing mutational accessibility, the simulated dataset drew from a pool of 3 values, and as such encodes that no other mutational routes are possible aside from the observed 3. Therefore the derived statistic is an underestimate, with additional routes at any weight lowering the likelihood of repeat observations of a single value. All other tests were completed using functions in base-R aside from the Dunn test, which was performed using the *FSA* package (Ogle et al. 2020). Along with the Bootstrap tests, the statistical tests used throughout the study were: Kruskal-Wallis chi-squared tests, Kruskal-Wallis post-hoc Dunn test, and Wilcoxon rank sum tests with continuity correction.

2.5.8. Data availability

All raw data used for generation of this manuscript is publicly available and can be accessed at https://github.com/J-S-Horton/Syn-sequence-parallel-evolution.

2.6. Results

2.6.1. SBW25-derived immotile strains evolve motility via highly repeatable evolution

To quantify the degree of parallel evolution of flagellar motility within the immotile SBW25 model system, we placed 24 independent replicates of AR2 under strong directional selection in a minimal medium environment (M9). Motile mutants were readily identified through emergent motile zones that migrated outward in a concentric circle (Fig. 2.1A). Clonal samples were isolated from the zone's leading edge within 24 h of emergence and their genotypes analysed through either whole-genome or targeted Sanger sequencing of the *ntrB* locus. Motile strains evolved rapidly (Fig. 2.1B) and each independent line was found to be a product of a one-step *de novo* mutation. All 24 lines had evolved in parallel at the locus level: each had acquired a single, motility-restoring mutation within *ntrB* (Fig. 2.1C). More surprising however, was the level of parallel evolution within the locus. 23/24 replicates had acquired a single nucleotide polymorphism at site 289, resulting in a transversion mutation from A to C (hereafter referred to as *ntrB* A289C). This resulted in a T97P missense mutation within NtrB's PAS domain. The remaining sample had acquired a 12-base-pair deletion from nucleotide sites 406-417 (Δ406-417), resulting in an in-frame deletion of residues 136-139 (ΔLVRG) within NtrB's phospho-acceptor domain.

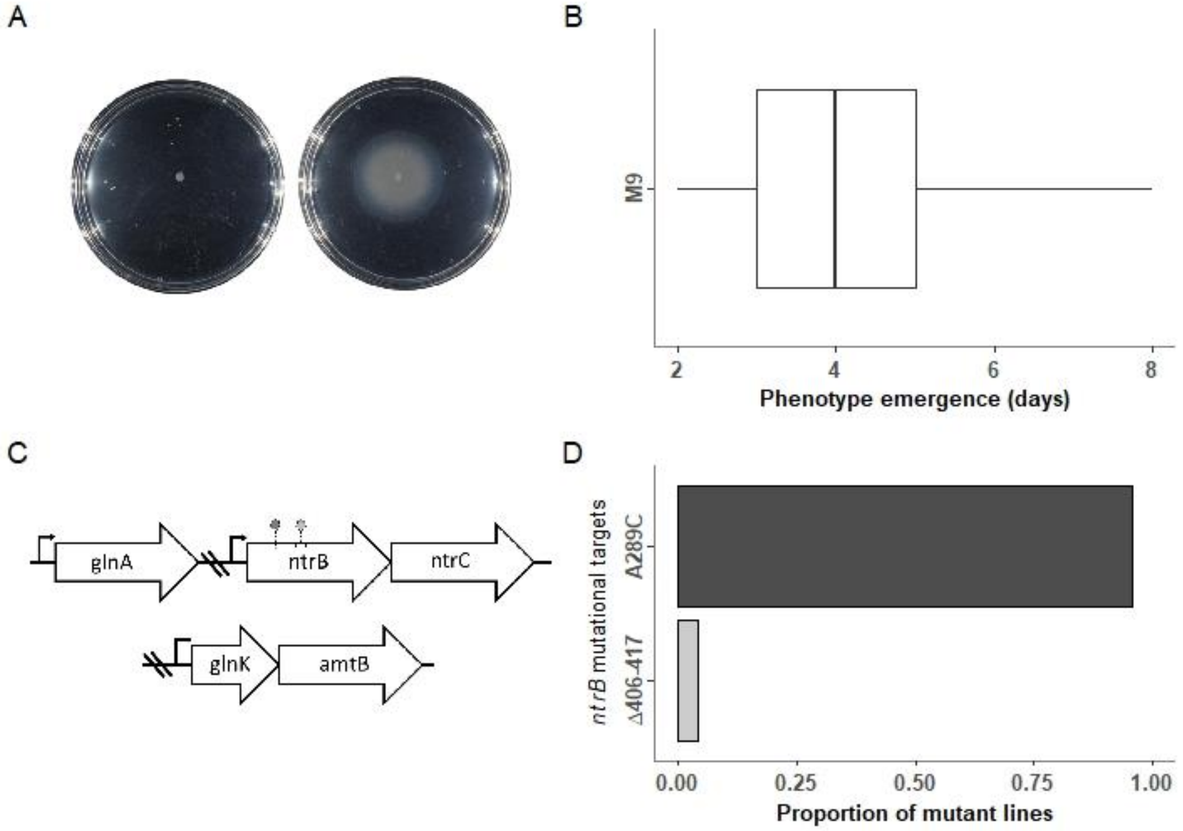


Figure 2.1. Highly repeatable evolution of flagella-mediated motility in immotile variants of *P. fluorescens* SBW25 (AR2). (A) Immotile populations evolved on soft agar (left) re-evolved flagella-mediated motility through one-step *de novo* mutation (right). (B) Phenotype emergence appeared rapidly, typically within 3-5 days following inoculation. Box edges represent the 25th and 75th percentiles and the whiskers show the observed range, individual data points are also plotted. (C) The underlying genetic changes were highly parallel, with all independent lines targeting one of two sites (left circle, A289C and right circle Δ406-417) within the *ntrB* locus at the expense of other sites within the nitrogen (ntr) pathway. (D) A single transversion mutation, A289C, was the most common mutational route, appearing in over 95% of independent lines (23/24).

2.6.2. Repeatable evolution is robust to nutritional environment

Repeatable evolution could be robust or highly context-dependent, especially when it occurs via *de novo* mutations with antagonistic pleiotropic effects (McGrath et al. 2011; Mcgee et al. 2016; Sackman et al. 2017). However, we found that the repeatability of the *ntrB* A289C mutation was robust across all tested conditions, despite evidence of antagonistic pleiotropic effects on growth. We tested for environment-specific antagonistic pleiotropy by measuring relative growth of the ancestral line and both evolved *ntrB* mutants on rich lysogeny broth and minimal medium containing either ammonia as the sole nitrogen source or supplemented with either glutamate (M9+glu) or glutamine (M9+gln), both of which are naturally assimilated and metabolised by the ntr system. Though large fitness costs were evident in M9 minimal medium, supplementing M9 with glu or gln reduced levels of antagonistic pleiotropy for both the *ntrB* A289C and the Δ406-417 mutants (Supplementary Fig. 2.1). Indeed, the antagonistic pleiotropy of impaired metabolism was sufficiently low in M9 supplemented with the amino acid glutamine (M9+gln) that motile mutants had increased fitness over the ancestral line in static broth, which was significant in *ntrB* A289C (P = 0.0361, Supplementary Fig. 2.1). These findings show that antagonistic pleiotropy is harsh in M9 and alleviated substantially in other nutritional environments, and therefore evolution in minimal media may have been limiting the viable number of adaptive routes.

We then tested whether repeatable evolution was robust to varying levels of antagonistic pleiotropy in our model system. Our expectation was that supplemented nutrient regimes would lower pleiotropic costs and thus unlock alternative routes of adaptation. We additionally hypothesised that a strain which is able to migrate prior to mutation would also ease starvation-induced selection pressures and could facilitate yet more mutational routes. For this experiment we therefore utilised an additional immotile variant of SBW25, which unlike AR2 did not have a transposon inserted into *viscB* (see materials and methods) and thus could migrate via a form of sliding motility prior to mutation (SBW25-Δ*fleQ* (herafter Δ*fleQ*); Alsohim et al. 2014). We observed a 'blebbing' phenotype (Fig. 2.1A) in Δ*fleQ* lines despite their ability to migrate in a dendritic fashion; however, we also found blebbing was less frequent under richer nutrient regimes (where populations migrated more rapidly utilising viscosin, see materials and methods). Overall, there was no evidence that the prevalence of the mutational hotspot *ntrB* A289C changed with nutrient condition (Gene-by-environment interaction: χ^2= 0.9375, df = 7, P = 0.9958, see Fig. 2.2). Instead, we observed that the *ntrB* A289C mutation was robust across all tested conditions, featuring in 90-100% of the Δ*fleQ* strains and 80-100% of AR2 strains (Fig. 2.2).

2.6.3. Repeatable evolution occurs despite motility being accessible via several mutational routes

Our evolution experiments across nutrient regimes uncovered three novel mutational routes that were observed in a small number of mutants (Fig. 2.2), revealing that mutational accessibility could not explain the level of observed parallel evolution. Most notably was a non-synonymous A-C transversion mutation at site 683 (*ntrB* A683C) in a Δ*fleQ* line evolved on M9+gln, resulting in a missense mutation within the NtrB histidine kinase domain. As a single A-C transversion within the same locus, we may expect A683C to mutate at a similar rate to A289C. We also observed a 12 base-pair deletion from sites 410-421 (*ntrB* Δ410-421) in an AR2 line evolved on M9+gln. Furthermore, we discovered a double mutant in an AR2 line evolved on M9+glu: one mutation was a single nucleotide deletion at site 84 within *glnK*, and the second was another A to C transversion at site 688 resulting in a T230P missense mutation within RNA polymerase sigma factor 54.

GlnK is NtrB's native regulatory binding partner and repressor in the ntr pathway, meaning the frameshift mutation alone likely explains the observed motility phenotype. However, as this mutant underwent two independent mutations we will not consider it for the following analysis. In addition, *ntrB* Δ410-421 and *ntrB* Δ406-417, despite targeting different nucleotides, translate into identical protein products (both compress residues LVRGL at positions 136-140 to a single L at position 136). Therefore, we will also group them for the following analysis. Under the assumptions that the three remaining one-step observed mutational routes to novel proteins are (*i*) equally likely to appear in the population and (*ii*) equally likely to reach fixation, the original observation of *ntrB* A289C appearing in 23/24 cases becomes exceptional (Bootstrap test: n = 1000000, $P < 1 \times 10^{-6}$). The likelihood of our observing this by chance, therefore, is highly unlikely. This means that one or both assumptions are almost certainly incorrect. Either the motility phenotype facilitated by the mutations may be unequal, enabling clonal interference to enforce a repeatable outcome; or the spectrum of adaptive mutations may appear in the population at different rates, resulting in mutation bias. One or both of these elements must be skewing evolution to such a degree that parallel evolution to nucleotide resolution becomes highly predictable.

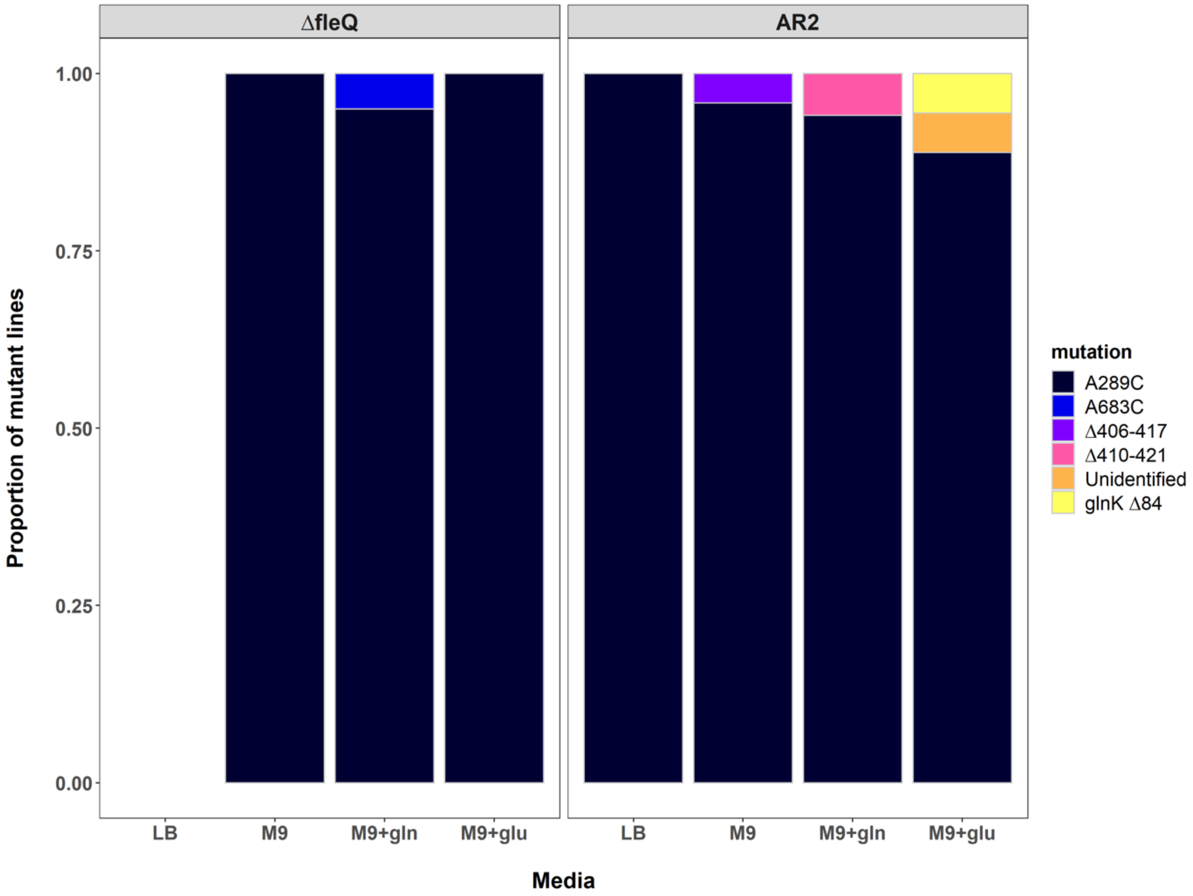


Figure 2.2. Repeatability of the A289C *ntrB* mutation across genetic background and nutrient environment (total *N* = 116). The proportion of each observed mutation is shown on the y axis. *ntrB* mutation A289C was robust across both strain backgrounds (SBW25Δ*fleQ* - shown as Δ*fleQ*, and AR2) and the four tested nutritional environments, remaining the primary target of mutation in all cases (>87%). Lines were evolved using 4mM, 8mM and 16mM of amino acid supplement (see materials and methods). No significant relationship between supplement concentration and evolutionary target was observed (Kruskal-Wallis chi-squared tests: AR2 M9+glu, df = 2, *P* > 0.2; AR2 M9+gln, df = 1, *P* > 0.23; Δ*fleQ* M9+gln, df = 1, *P* > 0.3), as such they are treated as independent treatments for statistical analysis but visually grouped here for convenience. Δ*fleQ* lines evolved on LB were able to migrate rapidly through sliding motility alone, masking any potential emergent flagellate blebs (see Alsohim et al. 2014). Sample sizes (*N*) for other categorical variables: Δ*fleQ* – M9: 25, M9+gln: 20, M9+glu: 7; AR2 - LB: 5, M9: 24, M9+gln: 17, M9+glu: 18.

2.6.4. Clonal interference cannot explain repeatability to nucleotide resolution

The adaptationist explanation for parallel evolution is that the observed mutational path is outcompeting all others on their way to fixation. For the purposes of our experiments, we define fixation as establishment on the frontier of the motile zone by the time of sampling. If selection via clonal interference alone was driving repeatable evolutionary outcomes, the superior fitness of the *ntrB* A289C genotype should have allowed it to out-migrate other motile genotypes co-existing in the population. To test if the *ntrB* A289C mutation granted the fittest motility phenotype, we allowed the evolved genotypes (A289C, Δ406-417, A683C and *glnK* Δ84) to migrate independently on the four nutritional backgrounds and measured their migration area after 48 h. To allow direct comparison, we first engineered the *ntrB* A683C mutation, which originally evolved in the Δ*fleQ* background, into an AR2 strain. We observed that the non-*ntrB* double mutant, *glnK* Δ84, migrated significantly more slowly than *ntrB* A289C in all four nutrient backgrounds (M9: $P = 0.00153$, M9+gln: $P = 0.0229$, M9+glu: $P = 0.00460$, LB: $P = 0.00476$, Fig. 2.3). However, *ntrB* A289C did not significantly outperform either of the alternative *ntrB* mutant lines in any environmental condition (*P* value range = 0.0567 – 0.878 Fig. 2.3). This suggests that selection may have played a role in driving parallel evolution to the level of the *ntrB* locus, but it cannot explain why nucleotide site 289 was so frequently mutated.

To determine if this result remained true when mutant lines were given the opportunity for clonal interference, we directly competed *ntrB* A289C against the alternative *ntrB* mutant lines, Δ406-417 and A683C, on M9 minimal medium. In brief, we co-inoculated the three mutant lines on the same soft agar surface at equal concentrations and allowed them to competitively migrate before sampling from the leading edge after 24 h of competition. Across 15 independent replicates, we observed no significant bias for any *ntrB* mutation at the growth's frontier (*ntrB* A289C = 4/15, *ntrB* Δ406-417 = 8/15, *ntrB* A683 = 3/15; Bootstrap test: n = 1000000, $P > 0.26$). We next emulated *ntrB* A289C appearing in the population within a handful of generations after the alternative mutations, and observed that the common genotype is significantly outperformed when inoculated both 6 h and 3 h after the alternative mutant lines (*ntrB* A289C establishment at frontier = 0/16 independent replicates (3 h and 6 h); Bootstrap test: n = 1000000, $P < 0.005$). These results highlight that if motile lines were to appear in the population simultaneously in minimal medium, *ntrB* A289C exhibits no evidence of clonal interference. Furthermore, if the common genotype appears in the population after just a handful of generations of its competitors it fails to establish itself at the frontier. Additionally, given that the range in time before a motility phenotype was observed could vary considerably between independent lines (Fig. 1B), our data do not support the hypothesis that global mutation rate could be high enough to allow multiple phenotype-granting mutations to appear in the population almost simultaneously, as was explored in this assay. As such our evidence suggests that opportunity for clonal interference during the short course of the experiment would be minimal, and if it were to occur there is no evidence to support it as the causative agent of our repeatable observations of *ntrB* A289C. More likely is that each independent line adhered to the “early bird gets the worm” maxim, i.e. the *ntrB* mutant which was the first to appear in the population was the genotype subsequently sampled. This therefore suggests that the reason *ntrB* A289C is so frequently collected when sampling is owed at least in part to an evolutionary force other than selection and mutational accessibility.

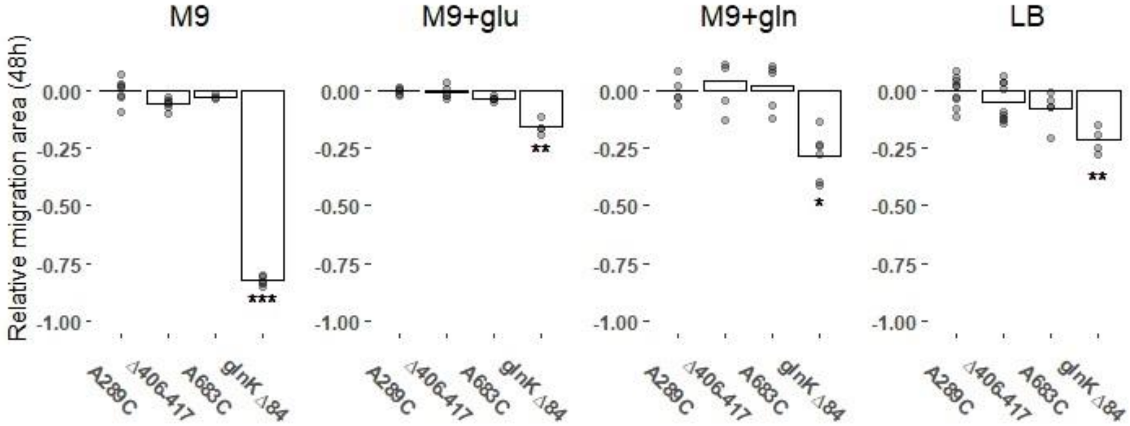


Figure 2.3. *ntrB* mutants possess comparable motility phenotypes. Surface area of motile zones within an AR2 genetic background following 48h of growth across four environmental conditions. Individual data points from biological replicates are plotted and each migration area has been standardised against the surface area of an AR2 *ntrB* A289C mutant grown in the same environment (*ntrB* A289C growth mean = 0). Significance values: * = $P < 0.05$, ** = $P < 0.01$ (Kruskal-Wallis post-hoc Dunn test).

2.6.5. Silent genetic variation can break a mutational hotspot

Local mutational biases can play a key role in evolution (Bailey et al. 2017; Lind et al. 2019). Such biases can be introduced by changing DNA curvature (Duan et al. 2018) or through neighbouring tracts of reverse-complement repeats (palindromes and quasi-palindromes), which have been shown to invoke local mutation biases by facilitating the formation of single-stranded DNA hairpins (De Boer and Ripley 1984). Therefore we next searched for a local mutation bias at *ntrB* site 289. Previously, we re-evolved motility in two engineered immotile strains of *P. fluorescens*, AR2 (derived from SBW25) and Pf0-2x (derived from Pf0-1; Taylor et al. 2015). Although evolved lines in AR2 frequently targeted *ntrB*, Pf0-2x lines fixed mutations across the ntr regulatory pathway. Furthermore, although Pf0-2x did acquire *ntrB* mutations in multiple independent lines, we observed no evidence of *ntrB* site 289 being targeted (Taylor et al. 2015). The NtrB proteins of SBW25 and Pf0-1 are highly homologous (95.57% identity) but share less identity at the genetic level (88.88% identity). A considerable portion of this genetic variation is explained by synonymous genetic variation (8.34%) rather than non-synonymous variance (2.76%). Synonymous mutations can play a role in altering local mutation bias. This may occur by altering the nucleotide-triplet to one with a higher mutation rate (Long et al. 2014) or by altering the secondary structure of longer DNA tracts via the mechanisms outlined above. Nucleotides that remain unpaired when their neighbouring nucleotides form hairpins with nearby reverse-complement tracts have been observed to exhibit increased mutation rates (Wright et al. 2003). Both SBW25 and Pf0-1 were found to have short reverse-complement tracts that flanked site 289, however the called hairpins were not entirely identical in their composition owing to synonymous variance (Supplementary Fig. 2.2). Overall, there are 6 synonymous nucleotide substitutions ± 5 codons flanking site 289 (C276G, C279T, C285G, C291G, T294G and G300C), which may have been affecting such hairpin formations and impacting local mutation rate.

To test if synonymous sequence was biasing evolutionary outcomes, we replaced the 6 synonymous sites in an AR2 strain with those from a Pf0-1 background (hereafter AR2-sm). Not all these sites formed part of a theoretically predicted stem that overlapped with site 289 but all were targeted due to their close proximity with the site. This ensured that the changes captured any secondary structures forming in the local region around nucleotide position 289. AR2-sm lines were placed under selection for motility and we observed that these lines evolved motility significantly more slowly (Fig. 2.4A), both in M9 minimal medium and LB (Wilcoxon rank sum tests with continuity correction: M9, W = 44.5, $P < 0.001$; LB, W = 22, $P < 0.001$). Evolved AR2-sm lines that re-evolved motility within 8 days were sampled and their *ntrB* locus analysed by Sanger sequencing (Fig. 2.4B). We observed some similar *ntrB* mutations to those identified previously: the *ntrB* A683C mutation was observed in one independent line evolved on LB, and *ntrB* Δ406-417 was also observed in both strain backgrounds. However, the most common genotype of *ntrB* A289C fell from being observed in over 95% of independent lines in M9 to 0%. Furthermore, we observed multiple previously unseen *ntrB* mutations, while a considerable number of lines reported wildtype *ntrB* sequences, instead either targeting another gene of the ntr pathway (*glnK*) or unidentified targets that may lay outside of the network (Fig. 2.4B).

To test that the A289C transversion remained a viable mutational target in the AR2-sm genetic background, we subsequently engineered the AR2-sm strain with this motility-enabling mutation. We observed that AR2-sm *ntrB* A289C was motile and comparable in phenotype to a *ntrB* A289C mutant that had evolved in the ancestral AR2 genetic background (Supplementary Fig. 2.3). We additionally found that AR2-sm *ntrB* A289C retained comparable motility to the other *ntrB* mutants evolved from AR2-sm (Supplementary Fig. 2.3). Therefore, we can determine that the AR2-sm genetic background would not prevent motility following mutation at *ntrB* site 289, nor does it render such a mutation uncompetitive. This therefore infers that the sole variable altered between the two strains (the 6 synonymous changes) are precluding mutation at site 289. Taken together these results strongly suggest that the synonymous sequence immediately surrounding *ntrB* site 289 facilitates its position as a local mutational hotspot, and that local mutational bias is imperative for realising extreme parallel evolution in our model system.

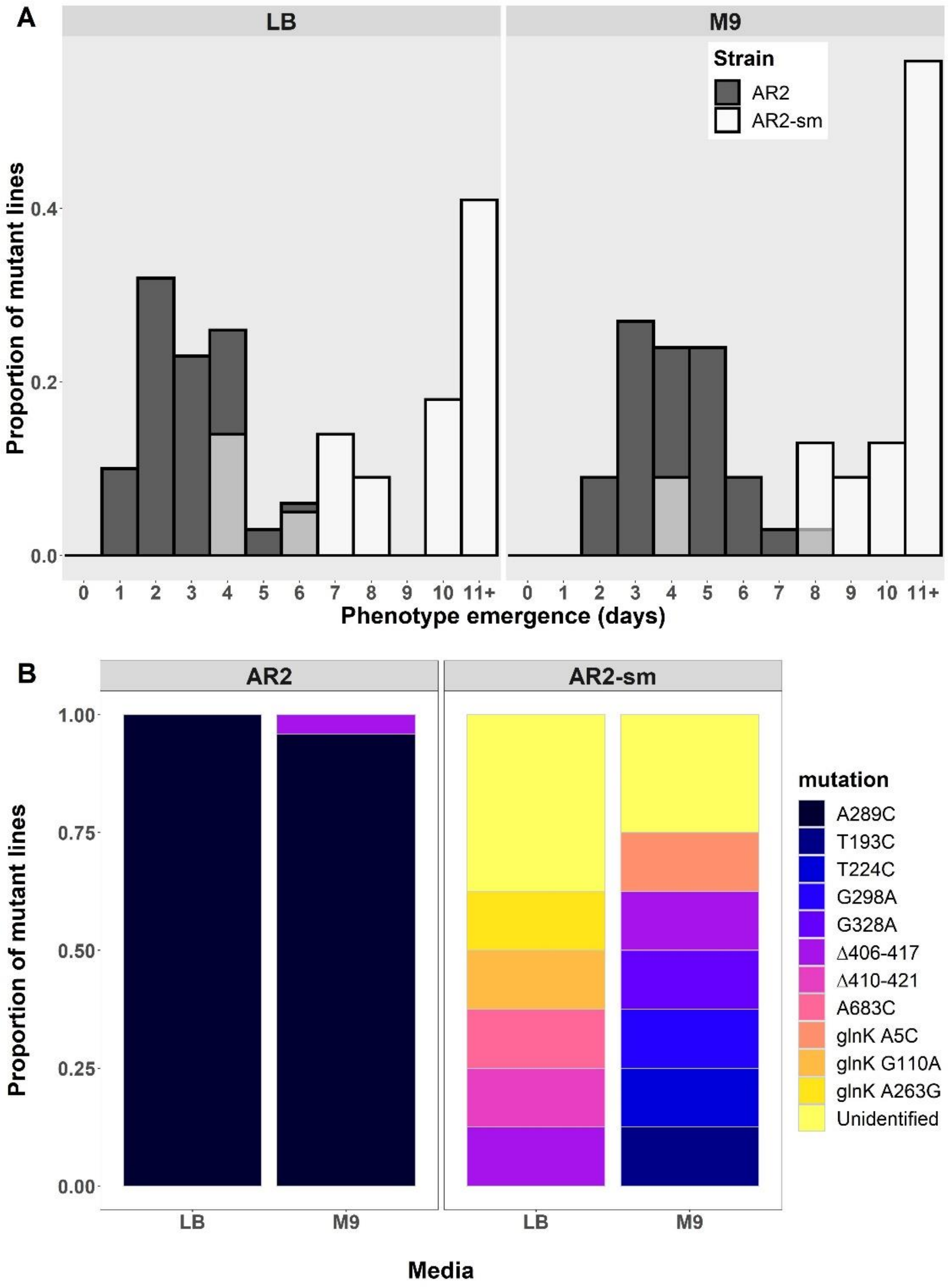


Figure 2.4. Loss of repeatable evolution conferred by a synonymous sequence mutant (AR2-sm). (A) Histogram of motility phenotype emergence times across independent replicates of immotile SBW25 (AR2) and an AR2 strain with 6 synonymous substitutions in the *ntrB* locus (AR2-sm) in two nutrient conditions. (B) Observed mutational targets across two environments (AR2: LB N = 5, M9 N = 24; AR2-sm: LB N = 8, M9 N = 8). Note that characterised genotypes were sampled within 8 days of experiment start date. Unidentified mutations could not be distinguished from wild type sequences of genes belonging to the nitrogen regulatory pathway *(ntrB, glnK* and *glnA*) which were analysed by Sanger sequencing (Supplementary Table 2.1). *ntrB* Δ406-417 was the only mutational target shared by both lines within the same nutritional environment.

2.6.6. Silent variation can build a mutational hotspot

As the previous result exemplified the power of synonymous variation in breaking mutational hotspots, we next hypothesised that the same amount of variation could just as readily build a mutational hotspot. To achieve this we engineered a synonymous variant of the immotile Pf0-2x strain (Pf0-2x-sm6). This strain was a reciprocal mutant of AR2-sm, in that it had synonymous variations at the same six sites within *ntrB* but substituted so that they matched AR2's native sequence (G276C, T279C, G285C, G291C, G294T and C300G). We placed both Pf0-2x and Pf0-2x-sm6 under directional selection for motility and observed that Pf0-2x evolved motility slower than Pf0-2x-sm6 (Fig. 2.5A) and targeted a multitude of sites across multiple loci (Fig. 2.5B). In stark contrast, Pf0-2x-sm6 evolved both more quickly (Fig. 5A; Wilcoxon rank sum tests with continuity correction: M9, W = 239.5, *P* < 0.001; LB, W = 461.5, *P* < 0.001) and massively more parallel than its native counterpart. Pf0-2x-sm6 fixed *ntrB* A289C in 80% of instances in M9 (8/10 independent lines), despite this *de novo* mutation not appearing once in a Pf0-2x evolved line (0/22 independent lines, Fig. 2.5B). The striking differences between the two strains from a Pf0-2x genetic background (Fig. 2.5) clearly mirror the results observed in the AR2 genetic background (Fig. 2.4). This reveals that a small number of synonymous variations can heavily bias mutational outcomes across genetic backgrounds and between homologous strains.

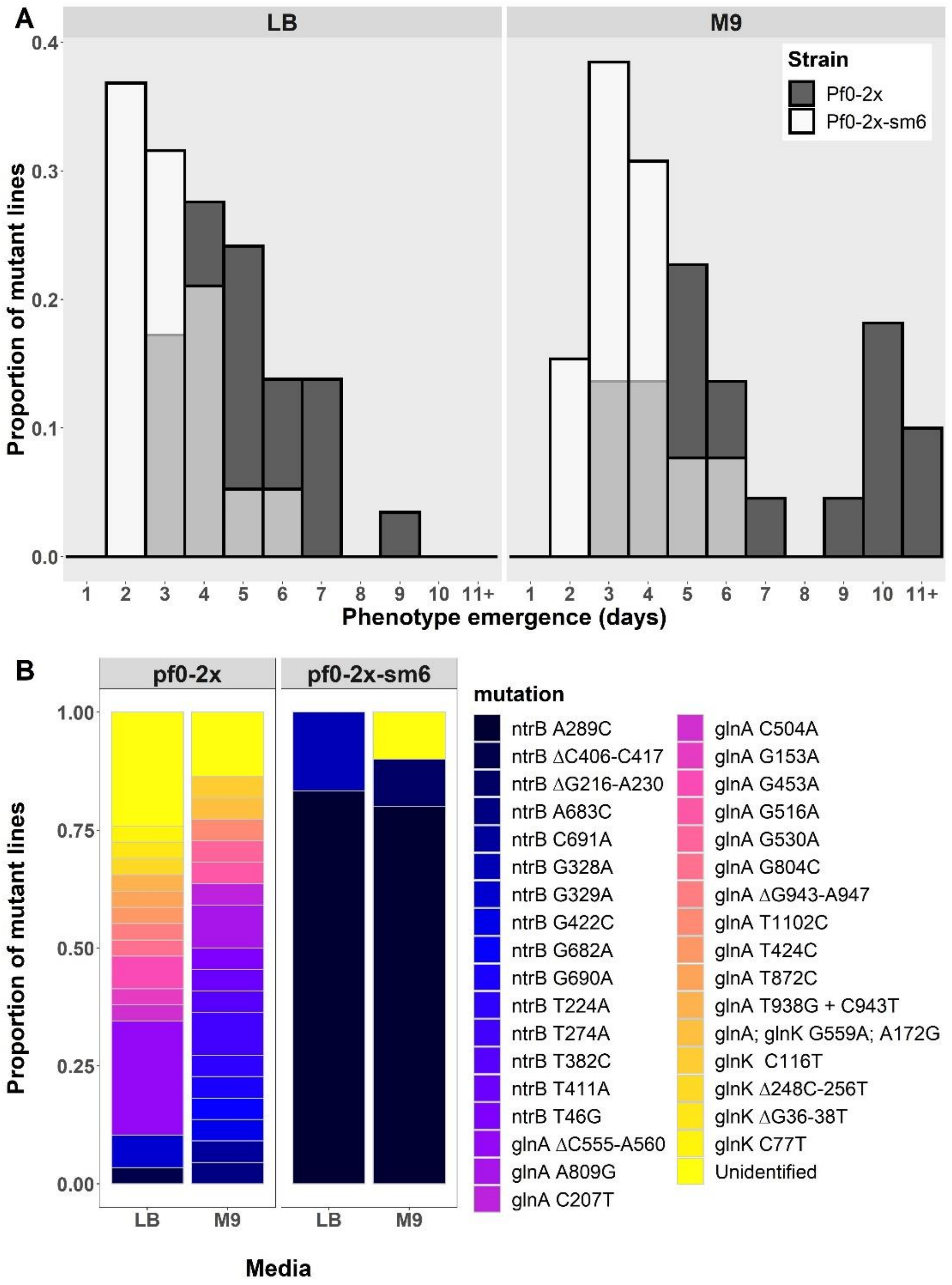


Figure 2.5. Gain of repeatable evolution conferred by a synonymous sequence mutant (Pf0-2x-sm). (A) Histogram of motility phenotype emergence times across independent replicates of an immotile variant of *P. fluorescens* strain Pf0-1 (Pf0-2x; Taylor et al. 2015) and a Pf0-2x strain with 6 synonymous substitutions in the *ntrB* locus (Pf0-2x-sm) in two nutrient conditions. (B) Observed mutational targets across two environments (Pf0-2x: LB *N* = 29, M9 *N* = 22; Pf0-2x-sm: LB *N* = 6, M9 *N* = 10). Unidentified mutations could not be distinguished from wild type sequences of genes belonging to the nitrogen regulatory pathway *(ntrB*, *glnK* and *glnA*) which were analysed by Sanger sequencing (Supplementary Table 2.1). Mutation *ntrB A289C* was not observed in a single instance in evolved Pf0-2x lines but became the strongly preferred target following synonymous substitution.

2.7. Discussion

Understanding the evolutionary forces that forge mutational hotspots and repeatedly drive certain mutations to fixation remains an immense challenge. This is true even in simple systems such as the one employed in this study, where clonal bacterial populations were evolved under strong directional selection for very few phenotypes, namely motility and nitrogen metabolism. Here we took immotile variants of *P. fluorescens* SBW25 (AR2) and Pf0-1 (Pf0-2x) that had been observed to repeatedly target the same gene regulatory pathway during the re-evolution of motility (Taylor et al. 2015). We found that evolving populations of AR2 adapted via *de novo* substitution mutation in the same locus (*ntrB*) and at the same nucleotide site (A289C) in over 95% of cases in M9 minimal medium. AR2 populations were constrained in which genetic avenues they could take to access the phenotype under selection, but mutational accessibility and clonal interference alone could not explain such a high degree of parallel evolution. Pf0-2x was distinct in that it did not evolve in parallel to nucleotide nor locus resolution. We observed that by introducing synonymous changes around the mutational hotspot (*ntrB* site 289) in both AR2 and Pf0-2x so that their local genetic sequences were swapped, we could push evolving AR2 populations away from the parallel path and pull Pf0-2x lines onto the parallel path. This work reveals that synonymous sequence is an integral factor toward realising highly repeatable evolution and building a mutational hotspot in our system.

More recent studies have revealed that synonymous changes have an underestimated effect on fitness through their perturbances before and during translation. Synonymous sequence variance can impact fitness by changing the stability of mRNA (Kudla et al. 2009; Kristofich et al. 2018; Lebeuf-Taylor et al. 2019) and altering codons to perturb or better match the codon-anticodon ratio (Frumkin et al. 2018). To our knowledge, we have shown here for the first time that synonymous sequence can also be essential for ensuring parallel evolutionary outcomes across genetic backgrounds. Our results strongly infer that this is due to its impact on local mutational biases, which mechanistically may be owed to the formation of single-stranded hairpins that form between short inverted repeats on the same DNA strand (De Boer and Ripley 1984; Fieldhouse and Golding 1991). The formation of these secondary DNA structures provides a mechanism for intra-locus mutation bias that can operate with extremely local impact and is contingent on DNA sequence variation, as introducing synonymous changes could readily perturb the complementarity of neighbouring inverse repeats (e.g. Supplementary Fig. 2.2). Furthermore, the finding of just six synonymous mutations having a significant impact on DNA structure would not represent a surprising result, as secondary structures can be altered by single mutations (Dong et al. 2001).

We can confidently assert that the altered mutational bias is owed to an intra-locus effect, owing to the six synonymous sites all residing within 14 bases at either flank of site 289. However, the full elucidation of the secondary structure and genetic mechanistic features enabling this powerful mutation bias awaits further study. We know that at least a portion of the 6 substituted nucleotide sites are imperative for parallel genetic outcomes, but we do not yet know if other nucleotide features in the local neighbourhood or more broadly e.g. strand orientation (Merrikh and Merrikh 2018) or distance from the origin of replication (Long et al. 2014) may be combining with local sequence to enforce mutational biases. Interestingly, our data suggest that the mutational hotspot typically mutates so quickly as to mask mutations appearing elsewhere and outside of the nitrogen regulatory pathway, which only appear when the hotspot is perturbed (Figs. 2.4 and 2.5). This therefore presents the opportunity to additionally quantify the difference in mutation bias owed to secondary structure.

Our findings show that the presence of a mutational hotspot was a stronger deterministic evolutionary force in our system than other variables such as nutrient regime, starvation-induced selection and genetic background. We expected the selective environments to hold some influence over evolutionary outcomes (Bailey et al. 2015) mostly owing to varying levels of antagonistic pleiotropy, which has been found to be a key driver in similar motility studies (Fraebel et al. 2017). Similarly, while parallel evolution can sometimes be impressively robust across genetic backgrounds (Vogwill et al. 2014), some innovations are strongly determined by an organism’s evolutionary history (Blount et al. 2012). Genomic variation also typically combines with environmental differences to drive populations down diverse paths (Spor et al. 2014). However in our experiments, the strains that share the same 6 synonymous sites evolve more similarly than those that share the same broader genetic background (Figs. 2.4B and 2.5B). These results show that strains can share not only high global homology but also similar genomic architecture – including translated protein structures and gene regulatory network organisation – and yet can have strikingly different mutational outcomes when under selection for the exact same traits owing to synonymous variation. This presents intriguing questions as to whether neutral changes could facilitate the dominance of a genotype during adaptation because of a previously acquired mutational hotspot, and asks whether these mutational hotspots can be selectively enforced.

Models looking to describe drivers of adaptive evolution often place precedence on fitness and the number of accessible adaptive routes (Orr 2005; Zagorski et al. 2016) yet pay little attention to local mutational biases (however see, Sackman et al. 2017). However, heterogeneity in mutation becomes of paramount importance when systems adhere to the Strong Selection Weak Mutation model (SSWM), which describes instances when an advantageous mutation undergoes a hard sweep to fixation before another beneficial mutation appears (Gillespie 1984). In such cases relative fitness values between adaptive genotypes are relegated to secondary importance behind the likelihood of an adaptive genotype appearing in the population. Indeed, experimental systems that adhere to the SSWM maxim have been observed to evolve in parallel despite the option of multiple mutational routes to improved fitness (Vogwill et al. 2014). This suggests that uneven mutational biases can be a key driver in forming mutational hotspots and realising parallel evolution, a conclusion which has been reinforced theoretically (Bailey et al. 2017) although empirical data is still lacking. Understanding the mechanistic causes of mutation rate heterogeneity, therefore, will be essential if we are to determine the presence of mutational hotspots that allow for accurate predictions of evolution (Bailey et al. 2018; Lind et al. 2019). The challenge remains in identifying what these mechanistic quirks may be, where they may be found, and determining how they impact evolutionary outcomes.

Our work sheds light on the ability of silent genetic variation to build a mutational hotspot with functionally significant evolutionary outcomes. This hotspot is built by an adaptive site under strong directional selection that enjoys biased mutation, facilitating highly repeatable evolution when mutation bias and selection align. Mutation is inherently a random process, but not all sites in the genome possess equal fixation potential. Most changes will not improve a phenotype under selection, and those that do will not necessarily mutate at the same rates. Therefore, we can increase our ability to anticipate the location of a mutational hotspot dramatically, permitting we have a detailed understanding of the evolutionary variables at play. Considerable inroads have already been made toward realising this goal. When searching for adaptive targets, it has been highlighted that loss-of-function mutations are the most frequently observed mutational type under selection (Kimura 1968; Lind et al. 2015) and that a gene's wider position within its regulatory network determines its propensity in delivering phenotypic change (McDonald et al. 2009). When searching for mutational biases, it has been shown that parallel evolution at the level of the locus is partially determined by gene length (Bailey et al. 2018) and that molecular apparatus involved in replication and repair can strongly influence the likelihood of a given nucleotide substitution (Lind and Andersson 2008; Stoltzfus and McCandlish 2017). Here, we show that synonymous sequence warrants consideration alongside these other variables by highlighting its impact on the realisation of highly repeatable evolution.

2.8.1. Acknowledgments and funding information

We thank Laurence Hurst for comments on earlier versions of this manuscript. In addition we thank member of the Taylor lab Matthew Shepherd for insightful comments and discussion, and Mark Silby for contributing the ancestral *P. fluorescens* Pf0-2x strain used in the study. This work was supported by the University of Bath University Research Studentship Account (URSA) awarded to TBT and NKP; a Royal Society Dorothy Hodgkin Research Fellowship awarded to TBT (DH150169); and the JABBS Foundation for RWJ. Bioinformatics analysis of the paper was carried out using the Medical Research Council's (MRC) Cloud Infrastructure for Microbial Bioinformatics (CLIMB), and Illumina Whole-Genome Sequencing by the Milner Genomics Centre, Bath, UK and MicrobesNG, Birmingham, UK.

2.8.2. References

1. Alsohim AS, Taylor TB, Barrett GA, Gallie J, Zhang X, Altamirano-Junqueira AE, Johnson LJ, Rainey PB, Jackson RW. 2014. The biosurfactant viscosin produced by Pseudomonas fluorescens SBW25 aids spreading motility and plant growth promotion. Environ. Microbiol. 16:2267–2281.
2. Avrani S, Bolotin E, Katz S, Hershberg R. 2017. Rapid Genetic Adaptation during the First Four Months of Survival under Resource Exhaustion. Mol. Biol. Evol. 34:1758–1769.
3. Bailey SF, Blanquart F, Bataillon T, Kassen R. 2017. What drives parallel evolution?: How population size and mutational variation contribute to repeated evolution. BioEssays 39:1–9.
4. Bailey SF, Guo Q, Bataillon T. 2018. Identifying drivers of parallel evolution: A regression model approach. Genome Biol. Evol. 10:2801–2812.
5. Bailey SF, Rodrigue N, Kassen R. 2015. The effect of selection environment on the probability of parallel evolution. Mol. Biol. Evol. 32:1436–1448.
6. Barrett RDH, M'Gonigle LK, Otto SP. 2006. The distribution of beneficial mutant effects under strong selection. Genetics 174:2071–2079.
7. Blount ZD, Barrick JE, Davidson CJ, Lenski RE. 2012. Genomic analysis of a key innovation in an experimental Escherichia coli population. Nature 489:513–518.
8. De Boer JG, Ripley LS. 1984. Demonstration of the production of frameshift and base-substitution mutations by quasipalindromic DNA sequences.
9. Bryksin A V, Matsumura I. 2010. Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. Biotechniques 48:463–465.
10. Bull JJ, Badgett MR, Wichman HA, Huehenbeck JP, Hillis DM, Gulati A, Ho C, Molineux IJ. 1997. Exceptional Convergent Evolution in a Virus. Genetics 147:1497–1507.
11. Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, Bull MJ, Richardson E, Ismail M, Elwood-Thompson S, et al. 2016. CLIMB (the Cloud Infrastructure for Microbial

Bioinformatics): an online resource for the medical microbiology community. Microb. Genomics 2:6.

12. Van Ditmarsch D, Boyle KE, Sakhtah H, Oyler JE, Carey D, Déziel É, Dietrich LEP, Xavier JB. 2013. Convergent Evolution of Hyperswarming Leads to Impaired Biofilm Formation in Pathogenic Bacteria. Cell Rep 4:697–708.
13. Dong F, Allawi HT, Anderson T, Neri BP, Lyamichev VI. 2001. Secondary structure prediction and structure-specific sequence analysis of single-stranded DNA. Nucleic Acids Res. 29:3248–3257.
14. Duan C, Huan Q, Chen X, Wu S, Carey LB, He X, Qian W. 2018. Reduced intrinsic DNA curvature leads to increased mutation rate. Genome Biol. 19:1–12.
15. Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. Nat. Rev. Genet. 2:549–555.
16. Fieldhouse D, Golding B. 1991. A source of small repeats in genomic DNA. Genetics 129:563–572.
17. Fong SS, Joyce AR, Palsson BØ. 2005. Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res.:1365–1372.
18. Fraebel DT, Mickalide H, Schnitkey D, Merritt J, Kuhlman TE, Kuehn S. 2017. Environment determines evolutionary trajectory in a constrained phenotypic space. Elife 6:e24669.
19. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. 2018. Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proc. Natl. Acad. Sci. U. S. A. 115:E4940–E4949.
20. Galen SC, Natarajan C, Moriyama H, Weber RE, Fago A, Benham PM, Chavez AN, Cheviron ZA, Storz JF, Witt CC. 2015. Contribution of a mutational hot spot to hemoglobin adaptation in high-Altitude Andean house wrens. Proc. Natl. Acad. Sci. U. S. A. 112:13958–13963.
21. Gillespie JH. 1984. Molecular Evolution Over the Mutational Landscape. Evolution (N. Y). 38:1116.
22. Hermisson J, Pennings PS. 2005. Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335–2352.
23. Herron MD, Doebeli M. 2013. Parallel Evolutionary Dynamics of Adaptive Diversification in Escherichia coli. PLoS Biol. 11:e1001490.
24. Hmelo LR, Borlee BR, Almblad H, Love ME, Randall TE, Tseng BS, Lin CY, Irie Y, Storek KM, Yang JJ, et al. 2015. Precision-engineering the Pseudomonas aeruginosa genome with two-step allelic exchange. Nat. Protoc. 10:1820–1841.
25. Huang S, Pang L. 2012. Comparing statistical methods for quantifying drug sensitivity based on in vitro dose-response assays. Assay Drug Dev. Technol. 10:88–96.

26. Jerison ER, Desai MM. 2015. Genomic investigations of evolutionary dynamics and epistasis in microbial evolution experiments. Curr. Opin. Genet. Dev. [Internet] 35:33–39. Available from: http://dx.doi.org/10.1016/j.gde.2015.08.008
27. Kram KE, Geiger C, Ismail WM, Lee H, Tang H, Foster PL, Finkel SE. 2017. Adaptation of Escherichia coli to Long-Term Serial Passage in Complex Medium: Evidence of Parallel Evolution. mSystems 2:1–12.
28. Kristofich J, Morgenthaler AB, Kinney WR, Ebmeier CC, Snyder DJ, Old WM, Cooper VS, Copley SD. 2018. Synonymous mutations make dramatic contributions to fitness when growth is limited by a weak-link enzyme.Matic I, editor. PLOS Genet. [Internet] 14:e1007615. Available from: https://dx.plos.org/10.1371/journal.pgen.1007615
29. Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science (80-.). [Internet] 324:255–258. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf
30. Lässig M, Mustonen V, Walczak AM. 2017. Predicting evolution. Nat. Ecol. Evol. [Internet] 1:1–9. Available from: http://dx.doi.org/10.1038/s41559-017-0077
31. Lebeuf-Taylor E, McCloskey N, Bailey SF, Hinz A, Kassen R. 2019. The distribution of fitness effects among synonymous mutations in a gene under selection. Elife [Internet]:e45952. Available from: https://doi.org/10.7554/eLife.45952.001
32. Lind PA, Andersson DI. 2008. Whole-genome mutational biases in bacteria. Proc. Natl. Acad. Sci. U. S. A. [Internet] 105:17878–17883. Available from: www.pnas.org/cgi/content/full/
33. Lind PA, Farr AD, Rainey PB. 2015. Experimental evolution reveals hidden diversity in evolutionary pathways. Elife 4.
34. Lind PA, Libby E, Herzog J, Rainey PB. 2019. Predicting mutational routes to new adaptive phenotypes. Elife [Internet] 8:e38822. Available from: https://www.ncbi.nlm.nih.gov/pubmed/30616716
35. Long H, Sung W, Miller SF, Ackerman MS, Doak TG, Lynch M. 2014. Mutation rate, spectrum, topology, and context-dependency in the DNA mismatch repair-deficient Pseudomonas fluorescens ATCC948. Genome Biol. Evol. 7:262–271.
36. M. Kimura. 1968. Evolutionary Rate at the Molecular Level. Nature [Internet] 217:624–626. Available from: https://www-nature-com.remote.library.osaka-u.ac.jp:8443/articles/217624a0.pdf
37. Markham NR, Zuker M. 2005. DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res. 33:577–581.
38. McDonald MJ, Gehrig SM, Meintjes PL, Zhang XX, Rainey PB. 2009. Adaptive divergence in experimental populations of Pseudomonas fluorescens. IV. Genetic constraints guide evolutionary trajectories in a parallel adaptive radiation. Genetics 183:1041–1053.

39. Mcgee LW, Sackman AM, Morrison AJ, Pierce J, Anisman J, Rokyta DR. 2016. Synergistic Pleiotropy Overrides the Costs of Complexity in Viral Adaptation. Genetics 202:285–295.

40. McGrath PT, Xu Y, Ailion M, Garrison JL, Butcher RA, Bargmann CI. 2011. Parallel evolution of domesticated Caenorhabditis species targets pheromone receptor genes. Nature 477:321–325.

41. Merrikh CN, Merrikh H. 2018. Gene inversion potentiates bacterial evolvability and virulence. Nat. Commun. 9:10.

42. Meyer JR, Dobias DT, Weitz JS, Barrick JE, Quick RT, Lenski RE. 2012. Repeatability and contingency in the evolution of a key innovation in phage lambda. Science (80-.). [Internet] 335:428–432. Available from: http://science.sciencemag.org/

43. Miller C, Kong J, Tran TT, Arias CA, Saxer G, Shamoo Y. 2013. Adaptation of Enterococcus faecalis to daptomycin reveals an ordered progression to resistance. Antimicrob. Agents Chemother. 57:5373–5383.

44. Notley-McRobb L, Ferenci T. 1999. Adaptive mgl-regulatory mutations and genetic diversity evolving in glucose-limited Escherichia coli populations. Environ. Microbiol. 1:33–43.

45. Ogle DH, Wheeler P, Dinno A. 2020. FSA: Fisheries Stock Analysis. Available from: https://github.com/droglenc/FSA

46. Orr HA. 2005. THE PROBABILITY OF PARALLEL EVOLUTION. Evolution (N. Y). 59:216.

47. Ostrowski EA, Woods RJ, Lenski RE. 2008. The genetic basis of parallel and divergent phenotypic responses in evolving populations of Escherichia coli. Proc. R. Soc. B Biol. Sci. 275:277–284.

48. R Core Team. 2014. R: A language and environment for statistical computing. R Found. Stat. Comput. Vienna, Austria [Internet]. Available from: http://www.r-project.org/.

49. Riehle MM, Bennett AF, Long AD. 2001. Genetic architecture of thermal adaptation in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. [Internet] 98:525–530. Available from: www.pnas.orgcgidoi10.1073pnas.021448998

50. Ripley LS. 1982. Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. Proc. Natl. Acad. Sci. U. S. A. 79:4128–4132.

51. Robleto EA, López-Hernández I, Silby MW, Levy SB. 2003. Genetic analysis of the AdnA regulon in Pseudomonas fluorescens: Nonessential role of flagella in adhesion to sand and biofilm formation. J. Bacteriol. 185:453–460.

52. Sackman AM, McGee LW, Morrison AJ, Pierce J, Anisman J, Hamilton H, Sanderbeck S, Newman C, Rokyta DR. 2017. Mutation-driven parallel evolution during viral adaptation. Mol. Biol. Evol. 34:3243–3253.

53. Seaton SC, Silby MW, Levy SB. 2013. Pleiotropic effects of gaca on pseudomonas fluorescens pf0-1 in vitro and in soil. Appl. Environ. Microbiol. 79:5405–5410.

54. Seemann T. 2015. Snippy: fast bacterial variant calling from NGS reads. Available from: https://github.com/tseemann/snippy
55. Sekowska A, Wendel S, Fischer EC, Nørholm MHH. 2016. Generation of mutation hotspots in ageing bacterial colonies. Sci. Rep. 6:4–10.
56. Spor A, Kvitek DJ, Nidelet T, Martin J, Legrand J, Dillmann C, Bourgais A, De Vienne D, Sherlock G, Sicard D. 2014. Phenotypic and genotypic convergences are influenced by historical contingency and environment in yeast. Evolution (N. Y). 68:772–790.
57. Stoltzfus A, McCandlish DM. 2017. Mutational biases influence parallel adaptation. Mol. Biol. Evol. 34:2163–2172.
58. Taylor TB, Mulley G, Dills AH, Alsohim AS, McGuffin LJ, Studholme DJ, Silby MW, Brockhurst MA, Johnson LJ, Jackson RW. 2015. Evolutionary resurrection of flagellar motility via rewiring of the nitrogen regulation system. Science (80-.). 347:1014–1017.
59. Tenaillon O, Rodríguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS. 2012. The molecular diversity of adaptive convergence. Science (80-.). 335:457–461.
60. Trevino V. 2020. HotSpotAnnotations — a database for hotspot mutations and annotations in cancer. Database:1–8.
61. Turner CB, Marshall CW, Cooper VS. 2018. Parallel genetic adaptation across environments differing in mode of growth or resource availability. Evol. Lett. 2:355–367.
62. Vogwill T, Kojadinovic M, Furió V, Maclean RC. 2014. Testing the role of genetic background in parallel evolution using the comparative experimental evolution of antibiotic resistance. Mol. Biol. Evol. 31:3314–3323.
63. Weber S, Ramirez C, Doerfler W. 2020. Signal hotspot mutations in SARS-CoV-2 genomes evolve as the virus spreads and actively replicates in different parts of the world. Virus Res. [Internet] 289:198170. Available from: https://doi.org/10.1016/j.virusres.2020.198170
64. Weinreich DM, Delaney NF, De Pristo MA, Hartl DL. 2006. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. Science (80-.). 312.
65. Wichman HA, Badgett MR, Scott LA, Boulianne CM, Bull JJ. 1999. Different trajectories of parallel evolution during viral adaptation. Science (80-.). 285:422–424.
66. Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. :ISBN 978-3-319-24277-4. Available from: https://ggplot2.tidyverse.org
67. Wood TE, Burke JM, Rieseberg LH. 2005. Parallel genotypic adaptation: When evolution repeats itself. Genetica 123:157–170.
68. Woods R, Schneider D, Winkworth CL, Riley MA, Lenski RE. 2006. Tests of parallel molecular evolution in a long-term experiment with Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 103:9107–9112.

69. Wright BE, Reschke DK, Schmidt KH, Reimers JM, Knight W. 2003. Predicting mutation frequencies in stem-loop structures of derepressed genes: Implications for evolution. Mol. Microbiol. 48:429–441.

70. Zagorski M, Burda Z, Waclaw B. 2016. Beyond the Hypercube: Evolutionary Accessibility of Fitness Landscapes with Realistic Mutational Networks. PLoS Comput. Biol. 12:1–18.

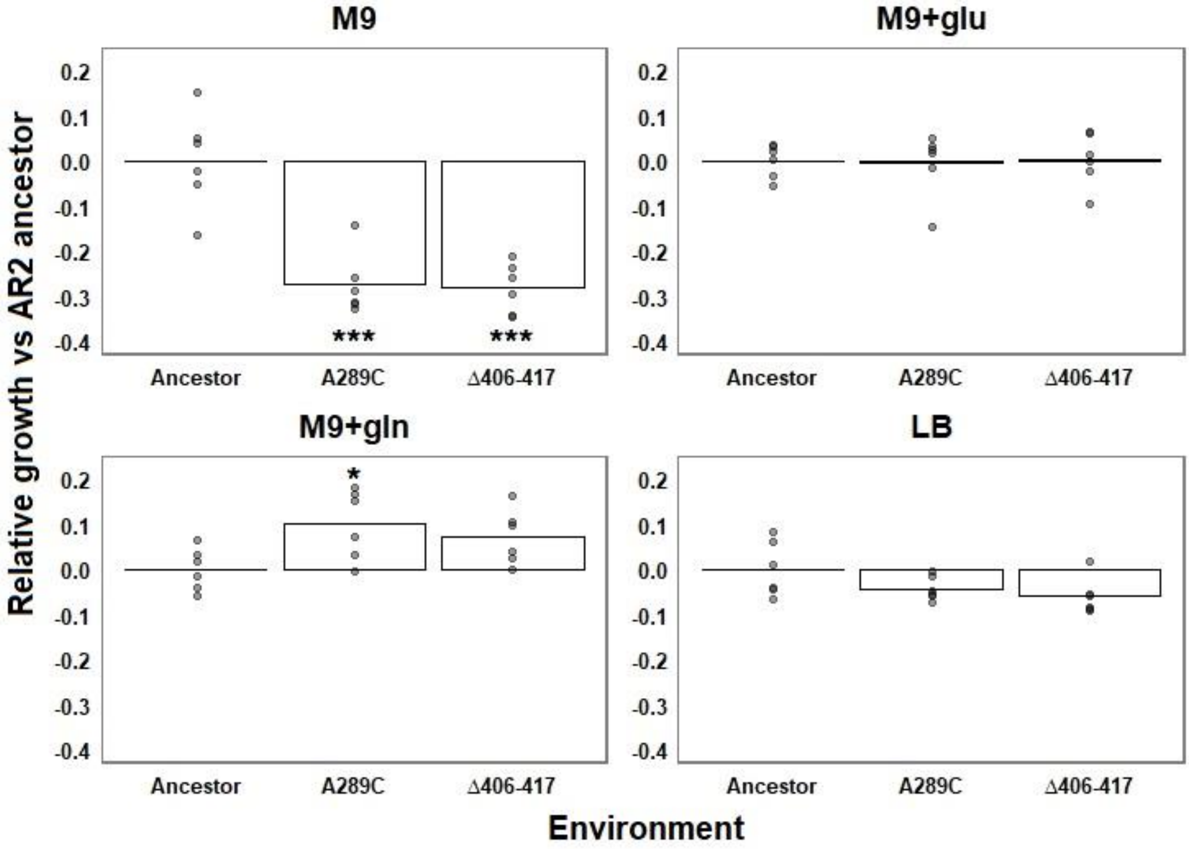
2.9. Supplementary materials

2.9.1. Assessing Pleiotropy via Growth Rate

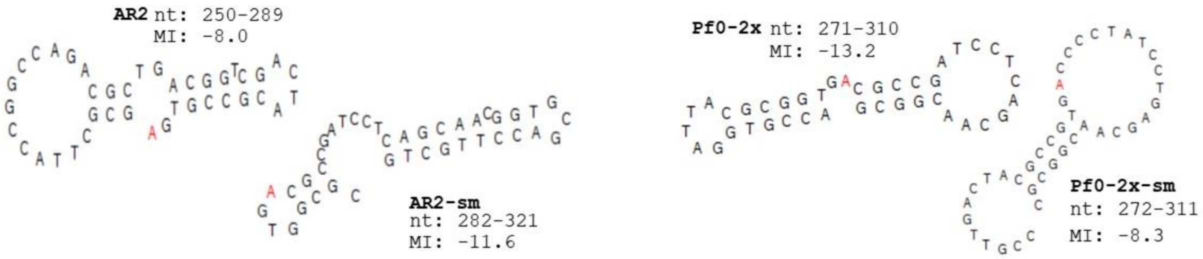
Cryopreserved samples of AR2 and derived *ntrB* mutants were streaked and grown for 48 h on LB agar (1.5%). Three colonies were then picked, inoculated in LB broth and grown overnight at an agitation of 180 rpm to create biological triplicates for each sample. This process was repeated with an independent batch of biological triplicates on a separate day to produce a total of 6 biological replicates for each sample. Overnight cultures were pelleted via centrifugation, their supernatant withdrawn and the cell pellets re-suspended in phosphate buffer saline (PBS) to a final concentration of OD1 cells/ml. The resuspension was subsequently diluted 100-fold into a 96-well plate (Costar®) containing nutrient broth. The plates were analysed in a Multiskan™ FC Microplate Photometer (Thermo Fisher Scientific) for 24h, with autonomous OD readings every 10 min without agitation. Growth values were determined by calculating area under the curve using the trapezoidal rule (approach outlined in, Huang and Pang 2012). This allowed us to incorporate elements of the pleiotropic consequences to metabolism as well as the benefits of the motile swimming phenotype, including prolonged lag phases, steeper exponential phases and differing eventual yields achieved by mutant populations relative to the ancestral strain (growth curves not shown).

2.9.2. *ntrB* loci analysis

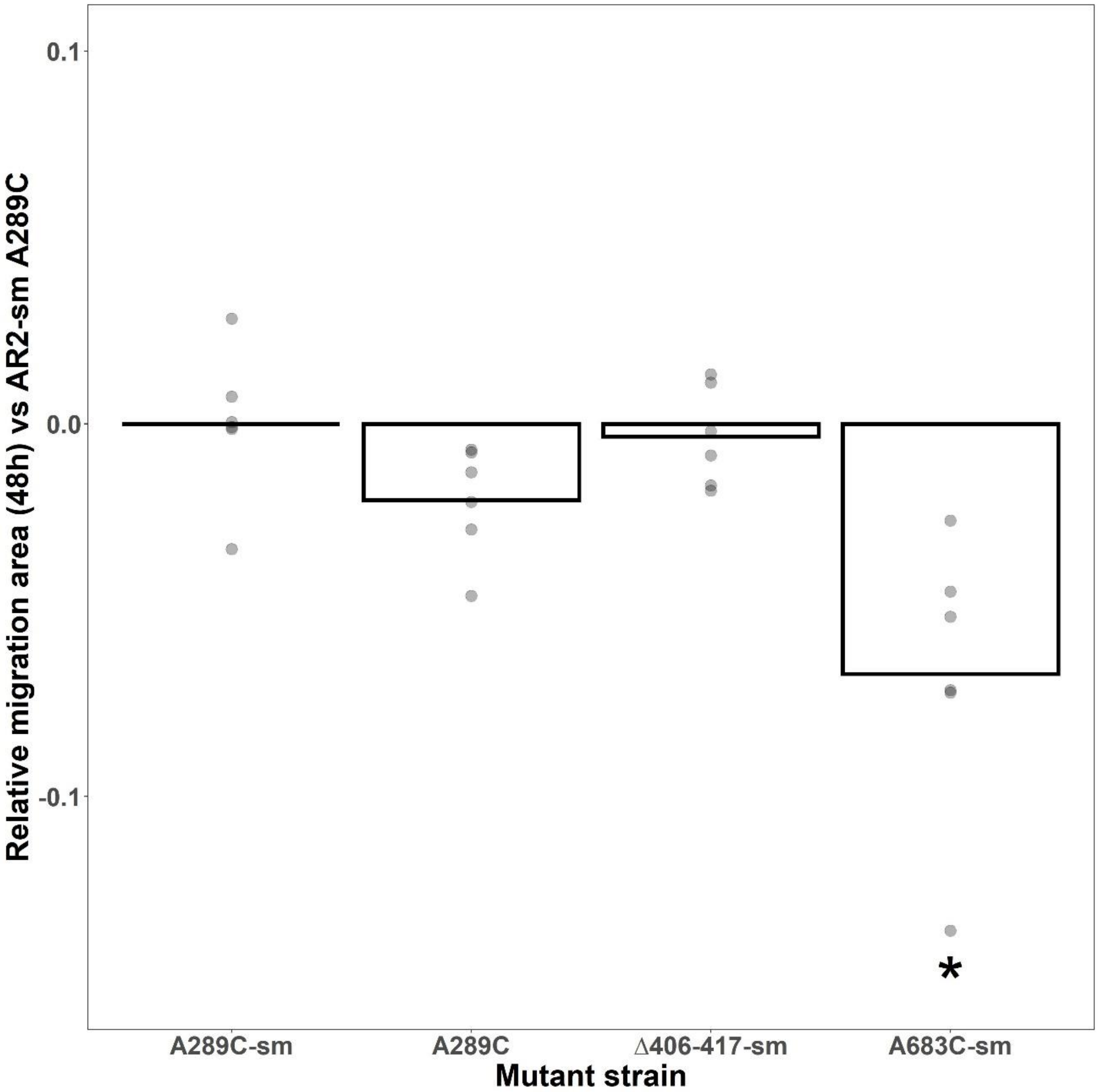
Theoretical hairpin stem-loop structures were generated using the *mfg* tool and methodology developed by Wright et al. 2003. The *mfg* tool is used in conjunction with the Quikfold tool on the DINAMelt Web Server (Markham and Zuker 2005). Default parameters were used for Quikfold with the exception of temperature, which was amended to 27^{0}C. The first 400 nucleotides of the open reading frames of *P. fluorescens* SBW25 *ntrB* and Pf0-1 *ntrB* were used as input sequences, and AR2-sm's and Pf0-2x-sm's input sequences were created by manually editing SBW25's and Pf0-1's *ntrB* sequence. The *mfg* application generates the most stable stem-loop structure for each base in which the selected base remains unpaired and so is at a higher likelihood of mutation. The window size of neighbouring nucleotides that are used to form the stem-loop structure can be adjusted, and a window length of 40 nucleotides was used for the analysis in this study.



Supplementary Fig. 2.1. Growth kinetics of mutant AR2 lines in static liquid culture over 24h. Nutrient environments: M9 = M9 minimal media supplemented with NH_4 at 7.5 mM. M9+glu = additional glutamate added at 8 mM. M9+gln = additional glutamine added at 8 mM. LB = lysogeny broth. Growth yield was determined using area under the curve, and each yield has been standardised against the yield of the AR2 ancestral strain grown in the same environment (AR2 ancestor growth mean = 0). Individual data points from biological replicates are plotted, and ranges around the mean growth of the ancestral strain are shown in column one of each frame. Plots are the means of six biological replicates. Significance values: * = *P* < 0.05, ** = *P* < 0.01, *** = *P* < 0.001 (one-way ANOVA post-hoc Tukey HSD test).



Supplementary Fig. 2.2. Quasi-palindromic sequences flank *ntrB* site 289 in both *P. fluorescens* SBW25 and Pf0-1 derived strains. Theoretical hairpin formations were generated using the *mfg* program(Wright et al. 2003). This software calculates the most stable hairpin formed between neighbouring tracts (± 40 nucleotides from site 289) in which the site of interest (in this case site 289, highlighted in red) remains unpaired. In these examples the nucleotides are forced into stem-loop structures that have been documented to comprise hairpins (Ripley 1982). The stability, structure and included nucleotide tracts of the predicted hairpins differ between strains and determine the mutated nucleotide site's Mutational Index (MI), which is a multiplication of the secondary structure's maximum energy (ΔG) and the percentage of alternative DNA folds in which the base of interest is unpaired: AR2 = -8.0. AR2-sm = -11.6, Pf0-2x = -13.2, Pf0-2x-sm = -8.3.These differences are partially owed to synonymous sequence variation as highlighted by the altered hairpin formation exhibited by AR2-sm and Pf0-2x-sm, which differ from their ancestors by 6 synonymous substitutions. AR2 and Pf0-2x-sm, the two strains that evolve in a highly parallel manner, share similar features that are absent in the other two strains. Namely their MI's are similar (-8.0 and -8.3) and the frequently mutated 'A' is located two nucleotides away from the base of a singular long, stable stem. As the *mfg* program only calls the most stable hairpin configuration it may miss alternative structures that temporarily form and introduce mutation bias, however the tool exemplifies the power of synonymous variance in altering hairpin stability.



Supplementary Fig. 2.3. *ntrB* A289C in AR2-sm retains comparative fitness to its ancestral counterpart. The motility phenotype of AR2 *ntrB* A289C and alternative AR2-sm *ntrB* mutants (Δ406-417-sm and A683-sm) were measured against an engineered AR2-sm *ntrB* A289C mutant (A289C-sm) in minimal medium. A289C-sm was not significantly outperformed by any strain, instead showing a significantly superior motility phenotype to A683-sm in M9. Although the two motile lines displayed comparable motility in an AR2 background (Fig. 2.3), the inferior phenotype observed here may be owed to an uncharacterised secondary mutation. Individual data points from biological replicates are plotted and each migration area has been standardised against the surface area of a *ntrB* A289C-sm mutant grown in the same environment (*ntrB* A289C-sm growth mean = 0). Significance values: * = $P < 0.05$, Kruskal-Wallis post-hoc Dunn test).

Supplementary Table 2.1. List of primers used throughout the study.

For use in:	Primer description:	Sequence:
Sanger sequencing of ntr pathway / Invasion assay	SBW25 *ntrB* locus (forward)	5'- GAGGTCCCAATGACCATCAG -3'
	SBW25 *ntrB* locus (reverse)	5'- GACGATCCAGACGGTTTCAC -3'
	SBW25 *glnK* locus (forward)	5'-GTGGGCAAAGGACTGATTTC-3'
	SBW25 *glnK* locus (reverse)	5'-GATGATGGCGAAGGTCATCT-3'
	SBW25 *glnA* locus (forward)	5'-CGGAAATCGCTCAAGGTTTA-3'
	SBW25 *glnA* locus (reverse)	5'-CTGATAATCCCCAGGCAAAA-3'
AR2 *ntrB* A683C integration into pTS1 backbone (allelic exchange)	Upstream fragment (forward)	5'- GAAATTAATAGGTTGTATTGATGTTGATGACCATCAGCGATGCACTG -3'
	Upstream fragment (reverse)	5'- GAATGCTCGGGGCGTAGTCGC -3'
	Downstream fragment (forward)	5'- GCGACTACGCCCCGAGCATTC -3'
	Downstream fragment (reverse)	5'- GCCGTTTCTGTAATGAAGGAGAAAACTCATGTCGATGGGGCTCCTTG -3'
AR2 *ntrB* synonymous substitution sequence integration into pTS1 backbone (allelic exchange)	Upstream fragment (forward)	5'- GAAATTAATAGGTTGTATTGATGTTGTGCCAAATGCCGCCTACATC -3'
	Upstream fragment (reverse)	5'- CGTTGCTGAGGATCGGCGTCACCGCGTAATCCACCGTCAG -3'
	Downstream fragment (forward)	5'- CTGACGGTGGATTACGCGGTGACGCCGATCCTCAGCAACG -3'
	Downstream fragment (reverse)	5'- GCCGTTTCTGTAATGAAGGAGAAAACGTTGATCAGCACGGTGATGT -3'
	SBW25 *ntrB* nested primer (forward)	5'- AATTTGGATCCATGACCATCAGCGATGCACTG -3'
	SBW25 *ntrB* nested primer (reverse)	5'- AATTTAAGCTTGATCCAGACGGTTTCACTACG -3'
AR2 *ntrB* synonymous substitution sequence with A289C	Upstream fragment (reverse)	5'- CGTTGCTGAGGATCGGCGGCACCGCGTAATCCACCGTCAG -3'
	Downstream fragment (forward)	5'- CTGACGGTGGATTACGCGGTGCCGCCGATCCTCAGCAACG -3'
Pf0-2x *ntrB* synonymous substitution sequence integration into pTS1 backbone (allelic exchange)	Upstream fragment (forward)	5'-TATCGCCTGCTGCTGGATGG-3'
	Upstream fragment (reverse)	5'- CGTTGCTCAGGATAGGGGTCACGGCGTAGTCGACGGTCAG -3'
	Downstream fragment (forward)	5'- CTGACCGTCGACTACGCCGTGACCCCTATCCTGAGCAACG -3'
	Downstream fragment (reverse)	5'-TCCACACGGTTTCACTACGG-3'
	Pf0-1 *ntrB* nested primer (forward)	5'-AATTTGGATCCAGCGTCAGGTCAAACCGTGT-3'
	Pf0-1 *ntrB* nested primer (reverse)	5'-AATTTAAGCTTTGGTGCTGGCTGATGATGTT-3'
Screening engineered lines for counter-selection escape	*sacB* check (Forward)	5'-TCAATCATACCGAGAGCGCC-3'
	sacB check (Reverse)	5'-TGTCGCAAACTATCACGGCT-3'

Chapter III

3.1. Appendix 6B: Statement of Authorship

This declaration concerns the article entitled:
How to ensure repeatable evolution: Uncovering the genetic features that build a mutational hotspot in *Pseudomonas fluorescens*

Publication status (tick one)

Draft manuscript	☒	**Submitted**	☐	**In review**	☐	**Accepted**	☐	**Published**	☐

Publication details (reference)	

Copyright status (tick the appropriate statement)

I hold the copyright for this material	☒	Copyright is retained by the publisher, but I have been given permission to replicate the material here	☐

Candidate's contribution to the paper (provide details, and also indicate as a percentage)	The candidate contributed to / considerably contributed to / predominantly executed the… Formulation of ideas: **95%** The co-first/co-corresponding author (the candidate) was predominantly responsible for the broad experimental design and hypotheses driving the research in this manuscript. Co-first author M. J. Shepherd inspired the project after developing the mini-Tn7 transposon system, and provided support through regular discussions regarding the acquired data. Design of methodology: **50%** The co-first/co-corresponding author (JSH) designed the methodology with co-first author M. J. Shepherd. MJS developed the protocol for the mini-Tn7 transposon system which was used as the basis for much of the experimental work in this manuscript. JSH expanded the scope of the methodology to invert strand orientation and use synonymous mutant constructs as templates, allowing the authors to maximise data output from the transposon system. JSH also designed the plasmid construct used for the *mutS* deletion strain. Experimental work: **30%** The co-first/co-corresponding author (JSH) considerably contributed to the experimental work by engineering the *mutS* deletion strain and collecting directed evolution data for this strain and the ancestral AR2 strain. These data are found in

	Fig. 3.2 and Fig 3.3 in the included manuscript. JSH also provided support to co-first author M. J. Shepherd by aiding in the preparation of evolved transposon-engineered strains for Sanger sequencing. Synonymous variant strains engineered by JSH were additionally used as sequence templates for transposon variants in this work, included in Fig. 3.2. The remaining experimental data was collected by co-first author M. J. Shepherd. Presentation of data in journal format: **60%** The co-first/co-corresponding author (JSH) was the primary author of the manuscript's introduction, discussion and abstract. Co-first author M. J. Shepherd was the primary author of the results, and materials and methods, aside from sections relating to Fig 3.3 which were authored by JSH.		
Statement from Candidate	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature.		
Signed	[redacted]	**Date**	**28.06.2021**

3.2. Title page: How to ensure repeatable evolution: Uncovering the genetic features that build a mutational hotspot in *Pseudomonas fluorescens*

Authors: *M. J. Shepherd*,* ***J. S. Horton*[#],*** *and T. B. Taylor[#]*

Milner Centre for Evolution, Department of Biology & Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK

*****These authors contributed to this work equally.

[#]**Corresponding Authors:** James S. Horton, Milner Centre for Evolution, Department of Biology & Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK, +44 (0)1225 385116, j.s.horton@bath.ac.uk.

Tiffany B. Taylor, Milner Centre for Evolution, Department of Biology & Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK, +44 (0)1225 384398, t.b.taylor@bath.ac.uk

Key words: Experimental genetics, Experimental evolution, Repeatable evolution, Mutation bias, Forecasting evolution

3.3. Abstract

Mutation is a game of chance, but the rules are far from fair. Depending on the genetic and environmental context, certain mutational types and sites can evolve at drastically different rates than their neighbours. When a mutable site coincides with a position under directional selection, it can forge a mutational hotspot that facilitates highly repeatable - and by extension reliably predictable - evolutionary outcomes. However, it is often unclear which features of the genetic context are essential for ensuring repeatable evolution. Here we reveal that genome location, strand orientation, DNA secondary structure and mismatch repair proteins all operate in concert to forge a mutational hotspot in *Pseudomonas fluorescens*. We have previously shown that a non-motile variant of *P. fluorescens* SBW25 repeatedly re-evolves motility via an identical *de novo* single nucleotide polymorphism (*ntrB* A289C) and that this evolution is sensitive to local genetic sequence, which we proposed was owed to DNA secondary structure. In this work we translocate, alter local synonymous sequence, and flip the orientation of the highly mutable locus to show that these features respectively lower the reliability, resolution and eradicate the observed mutational hotspot. We complement this finding by revealing that repeatable evolution is achieved by successfully overcoming mismatch repair proteins that suppress alternative adaptive transition mutations. We complete this study by presenting a theoretical framework that describes the necessary components for highly repeatable and predictable evolution. Our study moves us closer to forecasting evolution by showing how genetic features bend the rules of chance to ensure repeatable outcomes.

3.4. Introduction

Mutation represents the genesis of genetic variation, and as such acts as the keystone for the evolution of all life. By its nature mutation is an agent of chaos, as genetic change is a product of chance mistakes. Therefore it may seem that the generation of genetic material which is subsequently worked upon by natural selection and swept aside by genetic drift is inherently unpredictable. But this is not the case. Mutation may be a slave to probability, but the weights of these probabilities are by no means the same. Instead each position across the genome evolves with its own mutation rate and biases, which can fluctuate depending on the environment (Klaric et al. 2020), the genetic material that surrounds it (Long et al. 2014), and the molecular apparatus that interacts with it (Dettman et al. 2016). If a site that enjoys biased mutation proves adaptive, it can synergise with natural selection and allow evolution to realise a highly repeatable evolutionary outcome. However, the genetic features and molecular apparatus that ensure these near-deterministic outcomes are rarely demonstrated.

The field of experimental genetics has greatly enriched our understanding of mutagenic actors within the genome. Mutation accumulation experiments have provided clear evidence that mutation rates differ depending on genomic position (Hudson et al. 2002), and that that DNA mismatch repair proteins can introduce or remove biases via the mutation types they preferentially repair (Long et al. 2018). On a more local level, gene strandedness has been shown to incur its own mutation bias. Antagonistically facing operons (those that are transcribed in the opposite direction to the directionality of the replication fork) show an elevated mutation rate both in comparative genomic studies (Paul et al. 2013) and in engineered experimental systems (Juurik et al. 2012). More locally still, intra-locus mutation bias can be caused by short tracts of nucleotides, such as homopolymeric tracts or strings of neighbouring repeats, that raise the likelihood of small indels (Moxon et al. 2006) and specific nucleotide polymorphisms (De Boer and Ripley 1984) respectively. The genetic features scattered about the genome can therefore greatly bias the mutational spectrum, generating "mutational hotspots". If mutation at a hotspot were to prove strongly adaptive in a given environment, then directed evolution can result in highly repeatable genetic evolution. But are mutational hotspots generated by a single defining genetic feature, or the interplay of many?

In previous work, we identified a mutational hotspot that drove repeatable evolutionary outcomes in immotile variants of *Pseudomonas fluorescens*, as an identical single nucleotide polymorphism (*ntrB* A289C) was realised in >95% of independent replicates in minimal media (Fig. 2.1). We observed that this mutational hotspot was predicated on silent genetic changes, with just 6 synonymous mutations defining whether strong mutation bias operated at the hotspot site (Fig. 2.4). We predicted this was owed to the formation of single-stranded secondary DNA structures during either transcription or replication, consisting of stem-loops formed from bound complementary repeat nucleotide tracts (principle outlined in Dutra and Lovett 2006). We therefore exemplified the power of silent variation to define adaptive evolution, but it was unclear as to what role the broader genetic context played in supporting these outcomes. If we are to discover similar hotspots capable of facilitating repeatable evolutionary events, we must first be able to identify which genetic features are important and quantify their respective impacts on parallel evolution.

In this work, we demonstrate the respective roles of key genetic features on ensuring highly repeatable evolution by augmenting them piece-meal and measuring their effect on parallel outcomes. We find that moving genomic position lowers parallel evolution at the level of the nucleotide, but evolution at the locus remains high. We next find that disrupting DNA secondary structures through silent changes affects both repeatability at the level of the nucleotide and the locus. We then find that flipping the orientation of the operon by inverting DNA strandedness eradicates the hotspot across the locus and considerably lowers evolvability of the phenotype under selection. Finally, we make a functional deletion of a key mismatch repair protein (MutS) to demonstrate that genetic evolution is constrained by DNA repair complexes that suppress rare transitional mutational events, which aids in enforcing a repeatable outcome by stifling alternative adaptations. We visualise these insights by presenting a graphical framework that displays the mechanistic interplay of these features, which together ensure highly repeatable evolutionary outcomes.

3.5. Materials and Methods

3.5.1. Strains and culture conditions

Pseudomonas fluorescens AR2 (SBW25Δ*fleQ* IS-ΩKm-hah: PFLU2552) was the ancestral strain used for this study. This strain was constructed from *P. fluorescens* SBW25, lacks the flagellum master regulator FleQ and has a transposon-insertional disruption of the gene *viscB* (PFLU2552). This prevents production of the flagellum and the biosurfactant viscosin, abolishing both flagellar and surface spreading motility respectively (Alsohim et al. 2014). AR2 and its derivative strains subsequently detailed were cultured on LB (Miller) media. *Escherichia coli* strains were cultured on LB media at 37°C with 230 rpm shaking for liquid cultures. Media supplements are detailed in Supplementary Table 3.1.

3.5.2. Strain construction by two-step allelic exchange and miniTn7 transposon insertion

Experimental manipulation of the chromosomal locus and strandedness of the *ntrBC* operon was conducted in an AR2Δ*ntrBC* background constructed by two-step allelic exchange. We followed the protocol of Hmelo et al., 2015 with some alterations. In brief, 400bp flanking regions up- and down-stream of *ntrBC* were amplified and joined by strand overlap extension (SOE) PCR to create a knockout allele. This was inserted into the allelic-exchange suicide vector pTS1 (Contains *sacB* and *tetR*, and incapable of independent replication in *Pseudomonas*) by SOE-cloning (Bryksin and Matsumura 2010) and transformed into the conjugal *E. coli* strain ST18 by chemical-competence heat-shock. The plasmid pTS1-Δ*ntrBC* was transferred to AR2 by two-parent puddle-mating with E. coli ST18 pTS1-Δ*ntrBC*, and *ntrBC* merodiploids selected for on LB supplemented with Kanamycin sulphate and tetracycline hydrochloride but lacking 5-ALA supplement required for growth of the *E. coli* ST18 auxotrophic Δ*hemA* mutant. Following isolation, merodiploids were cultured overnight in LB broth lacking tetracycline selection, and then diluted before spread plating onto NSLB media supplemented with 15% w/v sucrose for *sacB* levansucrase-mediated counterselection of the pTS1 plasmid backbone. Sucrose-resistant colonies were isolated and screened for tetracycline sensitivity. Chromosomal presence of the Δ*ntrBC* allele and absence of *ntrBC* coding sequences was confirmed by colony PCR.

Chromosomal reintroduction of *ntrBC* alleles was performed using the miniTn7 transposon insertion system closely following the protocol of Choi & Schweizer, 2006. This system reliably introduces miniTn7 downstream of the *glmS* gene in *Pseudomonads* (PFLU6114 in *P. fluorescens* SBW25). Colony PCR was used to generate inserts containing *ntrBC*, and *ntrBC-sm* which includes the DNA-hairpin abolishing *ntrB* synonymous mutations detailed previously (see section 2.5). These inserts included 293bp upstream and 167bp downstream of *ntrBC* to maintain the native promoter, enhancer and terminator. The miniTn7 vector pJM220 was digested with SacI and HindIII to remove *rhaSR* and p*rhaBAD* (Meisner and Goldberg 2016). Insertion of *ntrBC* constructs was performed by ligation with

T4 DNA ligase overnight at 4°C to generate plasmids pMS-*ntrBC*-Lag and pMS-*ntrBC*-sm-Lag, and transformed into *E. coli* DH5α by chemical-competence heat-shock. Transposon insertion was conducted by four-parent puddle-mating of the relevant *E. coli* DH5α miniTn7 plasmid-containing donor with the recipient AR2Δ*ntrBC*, transposition helper *E. coli* SM10 λpir pTNS2 and conjugation helper *E. coli* SP50 pRK2073. Transposon insertions were selected for on LB supplemented with Gentamicin sulphate and Kanamycin sulphate, which restricts growth of the donor and helper *E. coli* whilst selecting for AR2Δ*ntrBC* containing the miniTn7 transposon. Chromosomal Transposon insertion downstream of *glmS* was confirmed by colony PCR.

Manipulation of the *ntrBC* or *ntrBC*-sm strandedness was conducting by taking advantage of the reliable insertional orientation of the miniTn7 transposon in the SBW25 chromosome (Choi and Schweizer 2006; Liu et al. 2014). Introduction of a HindIII restriction site upstream of *ntrBC* and a SacI restriction site downstream of *ntrBC* allowed cloning of the insert into the miniTn7 transposon plasmid with the Tn7R site downstream of the *ntrBC* genes. As the Tn7R site is always situated directly downstream of the *glmS* upon this reliably introducing the ntrBC open reading frame (ORF) onto the opposite strand to *glmS*, which is the lagging strand due to its location in relation to the origin of replication on the chromosome (Fig. 3.1A). To switch *ntrBC* insertion onto the leading strand the positions of the HindIII and SacI restriction sites each side of *ntrBC* were switched, meaning the Tn7R site sat upstream of the *ntrBC* genes and were therefore introduced onto the chromosome with their ORF on the same strand as *glmS* which is the leading strand with respect to the origin of replication (Fig. 3.1A). The orientation of the *ntrBC* ORF sequence relative to the *glmS* ORF was confirmed by PCR (Fig. 3.1B).

3.5.3. Motility evolution experiments

AR2, and AR2 miniTn7[ntrBC] variants were challenged to rescue motility in the absence of the FleQ master flagellar regulator on soft agar, as described previously (Taylor et al. 2015; section 2.5). Pure colonies were picked and inoculated into 0.25% agar LB plates made as described in Alsohim et al., 2014, and incubated at 27°C. At least 20 replicates were performed for each condition. Plates were checked a minimum of twice daily for motility, recording time to emergence. Motile zones were sampled immediately and always from the leading edge. Motile isolates were streaked on LB agar, and a pure colony picked and stored at -80°C as glycerol stocks of LB overnight cultures. All subsequent analysis was conducted on these pure motile isolates. Experiment was run for six weeks and any replicates without motility after this cut-off recorded as having not evolved.

3.5.4. Colony PCR and Sanger sequencing to identify mutations

Mutations conferring motility were identified by colony PCR amplification and subsequent Sanger sequencing service provided by Eurofins Genomics. PCR amplicons were purified for sequencing using the Monarch® PCR & DNA Cleanup Kit (New England Biolabs). The genes screened by sequencing were *ntrB, glnK, glnA,* and PFLU1131, which were selected based on being known mutational targets previously known to rescue motility in the AR2 background (section 2.6; Taylor et al., 2015; Shepherd et al., unpublished data). Mutations were identified by alignment of the returned sequence against the *P. fluorescens* SBW25 reference genome (Silby et al. 2009) using NCBI BLAST. All motile isolates were checked for presence or absence of *ntrB* mutation, and then isolates without an *ntrB* mutation were checked for *glnK, glnA* and PFLU1131 mutation in that order until a mutation was identified. Although listed as unidentified in Fig. 3.2, *glnK* PCR amplifications that failed to produce bands following multiple attempts are briefly discussed in the main text as pertaining to large deletions, which are believed to have deleted at least one of the primer binding sites.

3.5.5. MutS strain assay methods

AR2 Δ*mutS* strains were assembled via two-step allelic exchange as outlined above. The primer sets used (provided in Supplementary Table 3.3) produced a knockout construct that deleted nucleotides 760-2554 in the coding region of *mutS*, which removed the core domain including the clamp structure responsible for binding DNA in the translated protein product. Following allelic exchange, we isolated and screened 3 individual Δ*mutS* mutants and utilised all 3 for future tests and experiments. This was decided as several generations elapsed during the allelic exchange process between mutant construction, identification, and storage at -80°C. As such several secondary mutations may have accumulated in each of the strains relative to the ancestral strain background, and these may have had a subsequent effect on a strain's ability to evolve motility. To mitigate the risk of the role played by secondary mutations therefore, all 3 strains were used and their data was collated. To test if the knockout constructs were mismatch repair deficient, we performed a fluctuation assay using rifampicin to estimate relative mutation rates between *mutS* mutants and ancestral AR2 lines. This approach is outlined in (Vogwill et al. 2014). In brief, we grew biological triplicates of each Δ*mutS* line and biological triplicates of AR2 in 10 ml LB liquid culture agitated at 180 rpm for 48 h free from antibiotic selection. We then spread 100 μl of the culture (1% total volume) onto LB agar plates containing 30 ng/ml rifampicin, listed as the minimum inhibitory concentration for *P. fluorescens* SBW25 (Vogwill et al. 2014). As resistance to rifampicin only requires a single nucleotide polymorphism in the *rpoB* gene encoding an RNA polymerase subunit, we could measure mutation rate by using spontaneous mutations within this locus as a proxy for rates across the genome (Krašovec et al. 2019). The three replicates of AR2 reported 7, 10 and 64 resistant colonies per 100 μl culture following 48 h incubation. All biological replicates of Δ*mutS* lines each reported ≥400 colonies per 100 μl culture following 48 h incubation. Therefore we can conservatively report an elevated relative mutation rate (Kruskal-Wallis chi-squared = 4.3548, df = 1, $P < 0.037$) for all three constructed lines. We subsequently evolved motility from the mutator strains and screened their genotypes via the methods outlined above. However our Sanger sequences of *glnA* captured only ~800bp from the coding regions' N terminus, and as such some or all of the five 5 unidentified mutations may have been found in the C terminus of *glnA*.

3.5.6. Phenotypic assays and analysis

The motility phenotype of AR2 miniTn7[ntrB'C] strains was assayed by measuring distance moved after 24 h of incubation in 0.25% agar LB plates. Six biological replicates of each strain were grown as separate overnight cultures. Cultures were adjusted to an OD_{595} = 1 and resuspended in PBS. Soft agar plates were inoculated with 1 µL of these suspensions, by piercing the surface of the plate with the pipette tip, and then effusing the sample into the gap left by the tip. Plates were incubated for 24 h at 27°C, and photographs taken of motile zones. Surface area moved was then calculated from the radius of the concentric motile zone measured from these images ($A = \pi r^2$). Values were square root transformed before plotting.

Growth phenotype in shaking LB broth was measured by inoculating 99 µL of sterile LB broth with 1 µL of the OD_{595}=1 PBS cell suspensions for each replicate in a 96-well plate. Plates were incubated at 27°C with 180 rpm shaking in a plate reader, recording OD_{595} every ten minutes. Area under the bacterial growth curve was calculated using the growthcurver package in R and plotted (Sprouffske and Wagner 2016). Area under the bacterial growth curve provides a metric for fitness, as it accounts for the characteristics of lag- and log-phase growth, as well as the final carrying capacity of the population.

3.5.7. RNA extraction, cDNA preparation and RT-qPCR

RNA was extracted from 20 OD units of *P. fluorescens* cultures in mid-log phase growth (OD_{595} ~1.5). Cultures were incubated in LB broth at 27°C and 180 rpm shaking. Extractions were performed for biological triplicates of each strain of interest. Upon reaching the desired OD, growth and RNA expression was halted by addition of a ½ culture volume of ice-cold killing buffer (20 mM NaN_3, 20 mM Tris-HCl, 5 mM $MgCl_2$). Cells were pelleted, and the killer buffer removed. A lysis buffer of β-mercaptoethanol in buffer RLT from the Qiagen RNeasy extraction kit was used to resuspend pellets, which were then lysed by bead-milling at 4500 rpm for 45 s with lysing matrix B. Lysates were spun through columns from the Qiagen RNeasy extraction kit, and the extraction completed following the RNeasy kit protocol. A DNase I treatment step was included between washes with buffer RW1, by adding DNase I directly to the column from the RNase-free DNase kit (Qiagen) following kit protocol. Samples were eluted in nuclease-free water, and subsequently treated with TURBO DNase from the Turbo DNA-free kit (Invitrogen) following kit protocols.

Purified RNA concentration was measured by Qubit RNA BR assay (Thermo-scientific), RNA quality by nanodrop spectrophotometry, and RNA integrity and agarose gel electrophoresis. Production of cDNA for subsequent qPCR was performed from extracted RNA using the Protoscript-II First strand cDNA synthesis kit (New England Biolabs) with random hexamer priming following kit protocols, and including no reverse-transcriptase controls to allow detection of any prior DNA contamination by PCR. RT-qPCR was used to measure gene expression of ntrB and ntrC by the comparative Ct (ΔΔCt) method with *gyrB* as an endogenous reference. Reaction plates were set up using SYBR green PCR master mix (applied biosystems), with cDNA template preps diluted to 10^{-2} in nuclease-free water.

3.5.8. Statistical analyses

All statistical analysis and data handling was performed using R core statistical packages, aside from the Dunn test and corrections which was performed using the Dunn.test package. Shapiro-Wilks normality tests were performed to confirm non-normality of datasets. To test for differences between group medians, a Kruskal-Wallis test with post-hoc Dunn test and Benjamini-Hochberg correction was performed. When comparing just two groups, a Kruskal-Wallis chi-squared test and a Pearson's Chi-squared test with Yates' continuity correction was used. $P \leq 0.025$ was taken to indicate significance when tests involved corrections, otherwise $P \leq 0.05$ taken to indicate significance.

3.5.9. Data availability

All raw data used for generation of this manuscript can be accessed at https://github.com/J-S-Horton/Genetic-actors-for-repeatable-evo. At the time of writing this repository is currently locked, but will made publicly accessible following manuscript submission. In the interim, please contact the corresponding authors to request access.

3.6. Results

3.6.1. Translocation results in a reduction of parallel evolution at the *ntrB* mutational hotspot

To test the impact of genetic context on parallel evolutionary outcomes we utilised transposon-mediated mutagenesis followed by directed evolution, starting with AR2 miniTn7[*ntrBC*-Lag], which constitutes a chromosomal translocation of the *ntrBC* operon. The miniTn7 transposon inserts downstream of the *glmS* gene, which sits ~28kbp from the origin of replication (*oriC*) of the *P. fluorescens* SBW25 chromosome (Fig. 3.1A). This moves *ntrBC* ~348kbp closer to *oriC* than the native *ntrBC* operon (~376kbp from *oriC*). No significant difference in gene expression of the *ntrB* and *ntrC* genes was detected by RT-qPCR between AR2 and AR2 miniTn7[*ntrBC*-Lag] (Supplementary Table 3.2).

When challenged to rescue flagellar motility, AR2 miniTn7[*ntrBC*-Lag] replicates displayed several differences in the rate and parallelism of evolution. Median time to emergence of motility increased from 3.83 days for AR2 to 5.35 days (Fig. 3.2A), although this was not found to be significant when adjusting for the multiplicative effects of post-hoc analyses (*P* = 0.028, Dunn test). AR2 miniTn7[*ntrBC*-Lag] also demonstrated a reduction in the percentage of independent lines evolving motility within 6 weeks, dropping from 100% for AR2 to 95.65%. The observed mutational spectrum was also significantly altered in AR2 miniTn7[*ntrBC*-Lag] when compared to AR2 (Fig. 3.2B). The frequently mutated A to C transversion at site 289 in *ntrB* dropped from 95% to 50% of observed mutations. Alternative *ntrB* mutations occurred at increased frequencies, including the 12bp deletion Δ410-421 (22.7%) which has been previously observed (Fig. 2.2), and the SNP G682A (9%) that results in the amino acid change D228N similar to the previously seen D228A (Taylor et al. 2015) and mirroring the *ntrB* activating mutation D227A reported in *P. aeruginosa* (Li and Lu 2007) . Along with other rarer *ntrB* mutations, the remaining 9% of AR2 miniTn7[*ntrBC*-Lag] mutations were in *glnK*, a gene encoding an NtrB-repressor (Hervás et al., 2009) that has also been observed to permit *ntrBC*-mediated rescue of motility (Fig. 2.3). This result constitutes a significant lowering of mutational parallelism for the *ntrB* mutational hotspot upon translocation to an alternative chromosomal locus without alteration to *ntrBC* sequence identity, gene expression or strand topology.

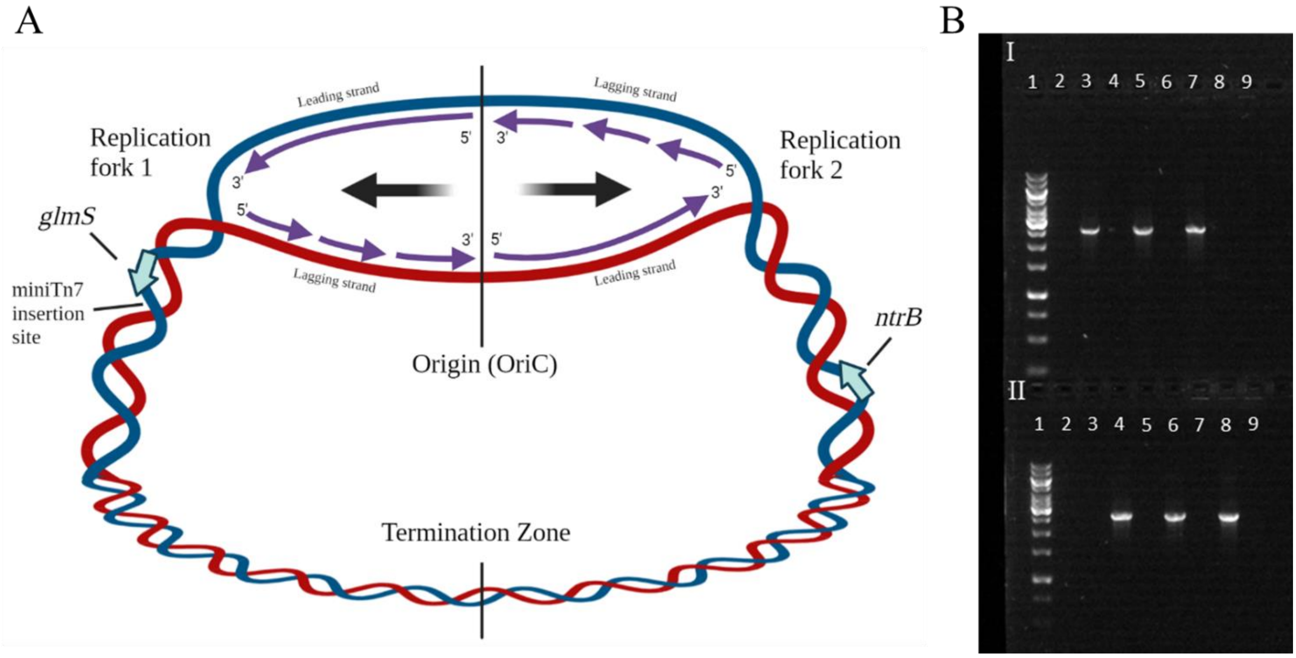


Figure 3.1. Manipulation of the chromosomal locus and DNA strandedness of *ntrBC* using the miniTn7 system. **A)** Diagram of the positioning and DNA strandedness of *ntrB*, *glmS*, and the miniTn7 insertion site. The circular bacterial chromosome is shown undergoing theta-replication, with two replication forks moving out from the origin of replication (*oriC*). Synthesis of leading and lagging strands are shown by purple arrows. Black arrows indicate direction of movement of the two replication forks. **B)** Agarose gel demonstrating difference in strand orientation between miniTn7[*ntrBC*-Lag] and miniTn7[*ntrBC*-Lead] engineered strains. Primer pair 'LAG' (SBW25-glmS + ntrC-Down-F) will only amplify if *ntrBC* are downstream of *glmS* and coding on the lagging strand. Primer pair 'LEAD' (SBW25-glmS + ntrBC-KO-up-R) will only amplify if *ntrBC* are downstream of *glmS* and coding on the leading strand. **Well contents: Row 'I'**: 1 – 1Kb Generuler DNA ladder. 2 – LEAD negative control. 3 – LAG AR2 miniTn7[*ntrBC*-Lag]. 4 – LEAD AR2 miniTn7[*ntrBC*-Lag]. 5 – LAG AR2 miniTn7[ntrBC-sm-Lag]. 6 – LEAD AR2miniTn7[*ntrBC*-sm-Lag]. 7 - LAG AR2 miniTn7[*ntrB'C*-Lag]. 8 – LEAD AR2miniTn7[*ntrB'C*-Lag]. 9 – LEAD AR2. 10 to 15 – empty. **Row 'II':** 1 – 1Kb Generuler DNA ladder. 2 – LAG negative control. 3 – LAG AR2 miniTn7[*ntrBC*-Lead]. 4 – LEAD AR2 miniTn7[*ntrBC*-Lead]. 5 – LAG AR2 miniTn7[*ntrBC*-sm-Lead]. 6 – LEAD AR2 miniTn7[*ntrBC*-sm-Lead]. 7 - LAG AR2miniTn7[*ntrB'C*-Lead]. 8 – LEAD AR2 miniTn7[*ntrB'C*-Lead]. 9 – LAG AR2. 10 to 15 – empty.

3.6.2. Inverting the DNA strandedness of *ntrBC* abolishes the mutational hotspot effect

Having established that genomic position had a considerable bearing on evolutionary parallelism, we next swapped the orientation of the translocated operon to measure this additional variable's effect on repeatable outcomes. This was achieved by switching the 5'-3' directionality of the *ntrBC* ORF from the lagging strand to the leading strand at the miniTn7 insertion site, and was observed to cause significant changes to the activity of the *ntrB* mutational hotspot (Fig. 3.2). The median time to emergence increased from 5.35 days for AR2 miniTn7[ntrBC-Lag] to 8.75 days for AR2 miniTn7[*ntrBC*-Lead] (*P* = 0.0013, Dunn test). There was also a significant drop in the percentage of replicate populations evolving within 6 weeks, from 95.65% for AR2 miniTn7[*ntrBC*-Lag] to 74.26% for AR2 miniTn7[*ntrBC*-Lead]. AR2 miniTn7[*ntrBC*-Lead] replicates displayed a drastically different mutational spectrum when challenged to rescue flagellar motility. The *ntrB* SNP A289C was not observed to evolve *de novo*, and other *ntrB* mutations accounted for only 22.2% of motility rescuing mutations. However, mutations were observed in known motility-granting mutation targets *glnK*, *glnA*, and PFLU1131 at frequencies of 55.5%, 7.4% and 3.7% respectively (Fig. 3.2B). In addition to these, 11.1% of mutations were unidentified with no sequence changes in any of the genes Sanger-sequenced in this study. This may indicate the existence of a previously unseen mutational route.

AR2 miniTn7[*ntrBC*-Lead] also demonstrated a far greater mutational diversity within the *ntrB* ORF, with no two mutations being the same. Only the *ntrB* SNP A683G has some similarity to previously seen mutations, resulting in the amino acid change D228G that mirrors the effect of the D228N and D228A NtrB mutations discussed above. Whilst no A289C mutations were observed for AR2 miniTn7[*ntrBC*-Lead], one deletion in this background (Δ286-297) resulted in loss of nucleotide A289 altogether and therefore the loss of T97 from the NtrB protein. Particularly striking was the *ntrB* tandem duplication of bases 412-444, which acted to duplicate the amino acid sequencing RGLAHEIKNPL including the highly conserved phospho-acceptor H-box motif HEIKNPL (Kim and Forst 2001) that is essential for NtrB kinase functionality. AR2 miniTn7[*ntrBC*-Lead] also displayed a significant diversity of *glnK* mutations. Whilst nucleotide-level parallelism is lost, gene-level parallelism persists but in this line favours *glnK* mutations. The most frequent *glnK* mutations were a 15bp deletion Δ258-272 and the SNP A5C (both 14.8% of observed mutations) which has been observed previously (Fig. 2.4). Much of the diversity of *glnK* mutations was constituted by indels. Deletions Δ191-192 and Δ194-252 resulted in frame shifts, and one mutant additionally possessed a large insertion of a predicted IS481 family transposase, the sequence of which being present in the SBW25 chromosome prior to the evolution assay as four separate ORF's (PFLU2158/ PFLU4347/ PFLU4873/ PFLU5832). Most *glnK* mutations therefore achieve a clear loss-of-function effect on the GlnK repressor. The loss of *ntrB* hotspot bias also resulted in emergence of *glnA* and PFLU1131 mutations, as well as an alternative unidentified mutational target. The glutamine synthetase encoding *glnA* presents a metabolic regulatory route to NtrB overactivation (Taylor et al. 2015). PFLU1131 is a putative HAMP-domain containing histidine-kinase, situated in an operon with a putative sigma-54 enhancer binding protein and FleQ-homolog PFLU1132. The spectrum of observed adaptations when DNA strandedness of the operon is swapped highlights the large mutational target size available for evolving motility, most of which are simply not observed when the hotspot is active.

To check that the absence of *ntrB* A289C mutation in the AR2 miniTn7[*ntrBC*-Lead] condition was not due to the DNA-strandedness change rendering the mutation non-viable, we engineered the strain AR2 miniTn7[*ntrB'C*-Lead]. In this strain *ntrB* with A289C is present (denoted as *ntrB'*). As a comparison, miniTn7[*ntrB'C*-Lag] was also included, to demonstrate that relocation of the *ntrBC* genes did not impact the motility phenotype they normally grant. Both strains remained motile and were not significantly different from one another (*P* = 0.2667, Dunn test) in an LB motility phenotype assay (Supplementary Fig. 3.1A). Both AR2 miniTn7[ntrB'C-Lag] and AR2 miniTn7[ntrB'C-Lead] were significantly faster than AR2 *ntrB* A289C in the motility phenotype assay (*P* = 0.0034 and *P* = 0.0185 respectively, Dunn test). The magnitude of this difference was small however with both strains moving 1.5mm further in 24 h than AR2 *ntrB* A289C. For growth in shaking LB broth as a measure of metabolic fitness, an inverse pattern is present, with miniTn7[ntrB'C] strains showing reduced fitness in LB broth compared to AR2 *ntrB* A289C (Supplementary Fig. 3.1B), although only the difference between AR2 *ntrB* A289C and AR2 miniTn7[*ntrB'C*-Lag] was significant (*P* = 0.0199, Dunn test). Overall these results show that the common mutational route remains a viable adaptive option when the locus is placed in a new genetic context.

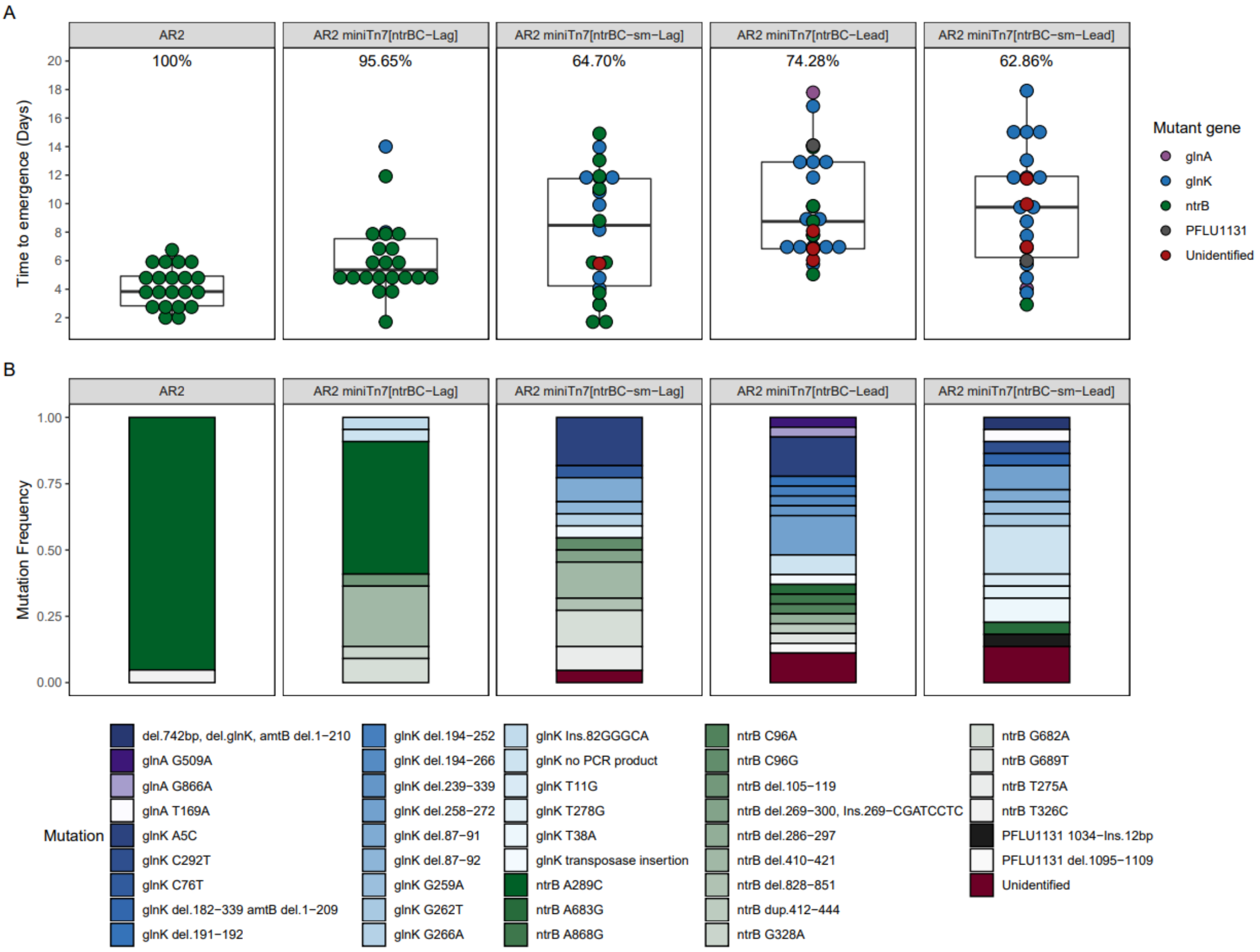


Figure 3.2. Impact of *ntrBC* translocation, DNA strandedness and local synonymous sequence variation on the observed mutational spectrum for rescuing flagellar motility in AR2-based strains. (*N* for each condition (evolved/total): AR2 – 21/21, miniTn7[*ntrBC*-Lag] – 22/23, miniTn7[*ntrBC*-sm-Lag] – 22/34, miniTn7[*ntrBC*-Lead] – 26/35, miniTn7[*ntrBC*-sm-Lead] – 22/35) **A**) Time to emergence of motility for each *ntrBC* AR2 strain background. Time to emergence given on the Y-axis in days. Boxplots display mean and quartile values. Individual replicate datapoints are shown and coloured by the mutant gene identified in that motile isolate. The percentage of replicates evolving motility within 6 weeks is given above each boxplot. **B)** Frequency of *de novo* mutations identified in motile isolates. Mutation frequency is shown on the Y axis. Each mutation is shown with its own colour, and mutations in the same gene grouped with shades of the same colour (*ntrB* = greens, *glnK* = blues, *glnA* = purples, PFLU1131 = greys, Unidentified = red).

3.6.3. Mutation hotspot-abolishing synonymous mutations reduce the frequency of *ntrB* mutations in both strand orientations

AR2 miniTn7[*ntrBC*] variants with mutation hotspot-abolishing synonymous mutations detailed in Fig. 2.4 and Fig 2.5 were also tested in this study to understand the impact of the local nucleotide context in driving *ntrB* parallel evolution in these novel genetic backgrounds. The local genetic context (which likely facilitates DNA hairpin formation; Supplementary Fig. 2.2) leads to extreme mutational parallelism for AR2 in our assay, with 95% of motility rescuing mutations being the *ntrB* SNP A289C and 100% of replicates evolving within 6 weeks (Fig. 3.2B). The synonymous mutant variants (sm) of the engineered transposon constructs, consisting of AR2 miniTn7[*ntrBC*-sm-Lag] and AR2 miniTn7[*ntrBC*-sm-Lead], did not show changed time to emergence when compared to their non-sm counterparts (*P* = 0.1083 and *P* = 0.3010 respectively, Dunn test). There was however a reduction in the percentage of replicates evolving within 6 weeks, from 95.65% to 65.70% for miniTn7[*ntrBC*-sm-Lag] and from 74.28% to 62.86% for miniTn7[*ntrBC*-sm-Lead].

In both cases, the addition of the mutation hotspot-abolishing synonymous mutations resulted in a reduction in *ntrB* mutation frequency and parallelism (Fig. 3.2B). Firstly we compared AR2 miniTn7[*ntrBC*-Lag] to AR2 miniTn7[*ntrBC*-sm-Lag], which are both translocated variants that are transcribed on the lagging strand, and differ from one another only by 6 synonymous nucleotide changes. The frequency of the *ntrB* A289C mutation reduced from 50% to 0% in the sm variant, and the frequency of all *ntrB* mutations fell from 90% to 50% respectively. This matches the previously reported effect of these synonymous mutations on the *ntrB* hotspot (Fig 2.4). For both strains the most frequent non-A289C *ntrB* mutations were the Δ410-421 12bp deletion and the SNP G682A (13.63% each of total mutations for AR2 miniTn7[*ntrBC*-sm-Lag]). Non-*ntrB* mutations included 45.45% in *glnK,* 18.18% of which being A5C. The remaining 4.55% were in an unidentified locus.

The synonymous mutations also altered the frequency of *ntrB* mutations when the gene was positioned on the leading strand. AR2 miniTn7[*ntrBC*-Lead] already lacked any emergence of *ntrB* A289C mutations, with the DNA-strandedness switch seemingly abolishing the effect of the *ntrB* mutational hotspot. As synonymous changes were similarly found to effect repeatability at this hotspot (Fig. 2.4 and Fig 2.5), introducing silent mutations was therefore not expected to significantly change the observed mutational spectrum. However, when comparing strains AR2 miniTn7[*ntrBC*-Lead] and AR2 miniTn7[*ntrBC*-sm-Lead], which are both translocated variants that are transcribed on the leading strand, and differ by only 6 synonymous nucleotide changes, *ntrB* mutations frequency dropped from 22% to 4.5%. The only *ntrB* mutation seen in the AR2 miniTn7[*ntrBC*-sm-Lead] background was A683G, which was also seen in the AR2 miniTn7[*ntrBC*-Lead] background. The rest of the mutations seen in this background were found in *glnK* (77.27%), PFLU1131 (4.5%) and the remaining 13.63% unidentified. Indels feature prominently amongst the *glnK* mutations and included a significant number of large (>400bp) deletions. Several of these deletions either affect or abolish the neighbouring gene PFLU5952 encoding an ammonium transporter *amtB* (see materials and methods). This indicates that the mutational hotspot may have a mutational bias effect on the wider *ntrB* locus, not just on the occurrence of the A289C SNP, perhaps owed to effects of the inferred secondary hairpin structure (see Supplementary Fig. 2.2).

3.6.4. Suppression of transition mutations from mismatch-repair proteins ensures repeatable evolution

Mutations occur at elevated rates and with biased spectrums, but they are only immortalised once they have circumvented or manipulated DNA mismatch repair machinery. We established above that *ntrB* A289C is able to evolve repeatedly as a consequence of multiple genetic features that greatly bias the realised mutational spectrum to favour this genetic change. But while mutation bias enriches this mutation, corrections of mutations occurring elsewhere by repair machinery may synergistically lower the immortalisation likelihood of alternative adaptive routes. Mismatch repair machinery are ubiquitous across the tree of life (Ganai and Johansson 2016) including bacteria, where they have been noted to bias observed mutational spectra by unevenly suppressing certain mutation types (Long et al. 2018). For example, transitions have been noted to become more common in their absence in model organisms *E. coli* (Schaaper and Dunn 1987) and *P. aeruginosa* (Wong et al. 2012). As such it may be that the A289C transversion mutation occurs repeatedly both because its own mutation rate is higher, and alternative adaptive routes stemming from other mutation types are suppressed by mismatch repair. To test this hypothesis, we constructed and evolved lines of a mismatch defective mutant of AR2 (AR2 Δ*mutS*), which lacks a key part of the mismatch repair complex (MutS is responsible for binding DNA; Schofield et al. 2001) and so cannot initiate repair.

We observed that, predictably, the elevated rate of mutation (see materials and methods) exhibited by Δ*mutS* strains led to more rapid emergence of motility (Fig 3.3). Whereas ancestral AR2 lines achieved motility over a range of 2-6 days, all 20 independent lines of the mutator strains achieved the motility phenotype ≤3 days (Fig 3.3). In addition, the degree of mutational parallelism across strain backgrounds differed. We analysed the genotype of each independent replicate via Sanger sequencing of key loci in the nitrogen regulatory pathway: *ntrB, glnK* and *glnA*, which were affiliated with rapid evolution of motility (Fig 3.2). Following sequencing, the adaptive mutations fixed within 25% (5/20) replicates remained unidentified, 35% (7/20) reported *ntrB* A289C, and 40% (8/20) reported a diverse set of single nucleotide polymorphisms in either *ntrB* or *glnK* (*ntrB* T323C, T407C, A608G (observed twice), A683G; *glnK* T11C, A263G, A131G). As such we could determine that mutational repeatability at *ntrB* A289C mutational hotspot had fallen from >95% to 35%. However, this was not owed to a reduction in mutation bias operating at the hotspot, but rather an elevation in the realisation of alternative adaptive mutations. This result is highlighted by the number of *ntrB* A289C mutations realised by day 2 of the mutator strains (7/20) relative to the ancestral line (2/21) which only accumulated over 7 A289C mutations by day 4 (Fig. 3.3). But while *ntrB* A289C mutations were elevated at earlier times, so too were alternative mutations (those that were identified are plotted in Fig 3.3), with the remaining 13/20 non-A289C mutations all being realised within 3 days, relative to only 1/23 non A289C mutant being realised over 6 days in AR2. Furthermore, all 8 identified alternative mutations were transition mutations. If we expect transitions to represent 33% of all mutations and assume an equal likelihood of fixation regardless of mutation type, then there is no significant enrichment for either mutation type (transition or transversion) in the mutator lines (Bootstrap test, n = 1000000, $P > 0.33$). In contrast, there is a significant omission of transitions in an AR2 background where the hotspot transversion remains in effect (Bootstrap test, n = 1000000, $P < 0.0023$). As such these results show that the mutator strains unlock alternative transition mutations that are suppressed in lines with intact mismatch repair machinery. As the A289C transversion similarly appears more frequently in mutator lines, mismatch repair complexes likely also correct transversion mutations at the hotspot site. Therefore in this model organism, mutations at hotspot sites are realised more commonly because rare mutations of alternate mutation types are suppressed, while mutations at the hotspot happen sufficiently often to overcome mismatch repair.

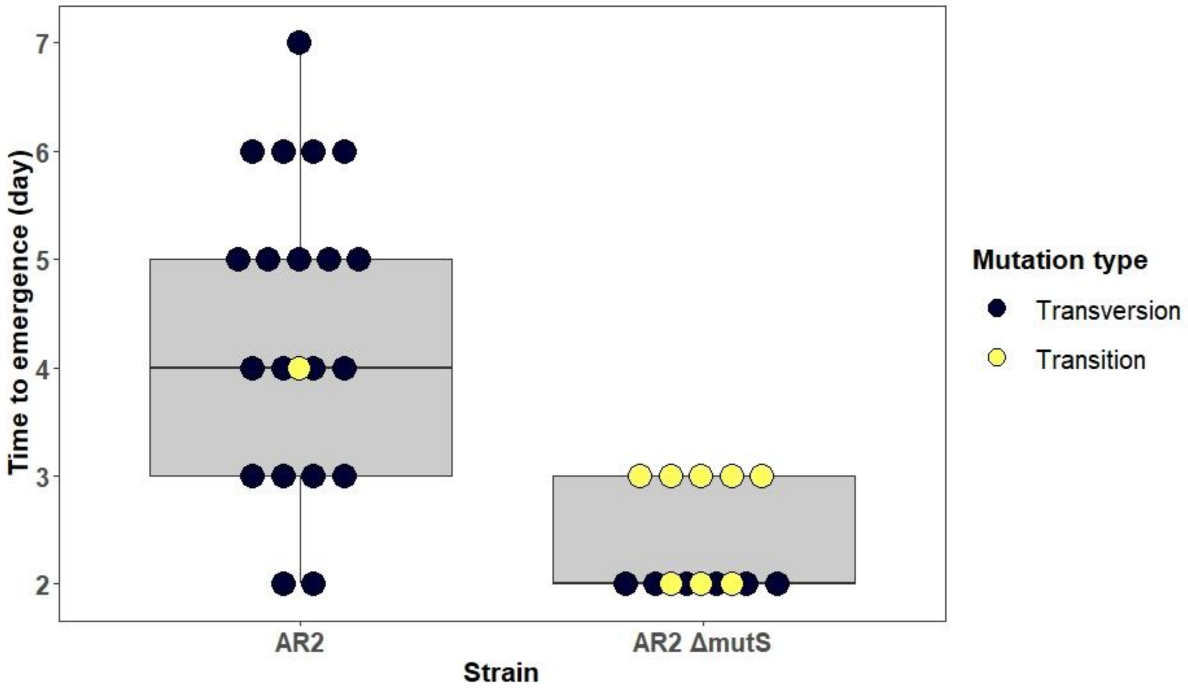


Figure 3.3. Removal of a mismatch repair complex uncovers rapidly realised adaptive transition mutations. Constructs of a mutator variant of AR2, which lacks functionality of the mismatch repair protein MutS (AR2 Δ*mutS*), exhibit different evolvability properties to the ancestral line. Independent replicates of mutator variants realised the motility phenotype in significantly less time (Kruskal-Wallis chi-squared = 16.906, df = 1, *P* < 0.001). In addition, the relative frequencies of identified mutation types differed across mutator backgrounds. While transversion mutations were realised sooner in AR2 Δ*mutS* (6/20 by 48 h, 2/23 by 48 h in AR2), more transition mutations were likewise realised relative to the AR2 ancestor (8/15 identified mutations in AR2 Δ*mutS*, 1/23 identified mutations in AR2; Pearson's Chi-squared test with Yates' continuity correction: $\chi^2 = 6.2032$, df = 1, *P* = 0.01275).

3.7. Discussion

In this work we demonstrated the interconnected influence of genomic location, strand orientation, synonymous sequence and mismatch repair on achieving repeatable evolutionary outcomes. We found that only by satisfying certain criteria for each examined variable was near-deterministic evolution possible (i.e. the realisation of an identical polymorphism in >95% of independent lines). However, by capturing the impact of these features individually and in combination, we can appreciate how the underlying mechanistic drivers interact to cause the observed changes in genetic repeatability. This allows us to form a cohesive conceptual framework that showcases the interplay of these features, and highlights how together they can ensure remarkably repeatable outcomes. This framework is depicted in Fig. 3.4, and described in detail below.

We observed that lines adapted more slowly, and parallel evolution lessened, when the hotspot locus was moved to a genomic position closer to the origin of replication (*oriC*). This may be owed to increased replication fidelity around the origin site, whereby mutation rate is measurably lower than in other regions of the genome. This effect has been observed in multiple studies including one focused on a mutator strain derived from *P. fluorescens* SBW25, the model organism used in this work (Hudson et al. 2002; Long et al. 2014). Previous research has suggested that the relationship between genomic position and mutation rate is driven by replication timing and the associated levels of deoxyribonucleotides (dNTPs) (Dillon et al. 2018). A lack of dNTPs has been suggested to increase fork stalling (Gon et al. 2006), whereas high levels of dNTPs can increase the rate of introduced mismatches (Dillon et al. 2018). It may be that in our cell lines dNTP levels are optimal during the initiation of replication but rapidly change as replication progresses, either by depletion through use, or a spike as a result of synthesis (Dillon et al. 2018). The mutational mechanism invoked would therefore vary depending on dNTP levels, but either process would increase mutagenicity at the mutational hotspot's native location relative to its new location much closer to the *oriC*.

Although parallel evolution to nucleotide resolution decreases closer to the *oriC* – and the mutant lines take longer to evolve, inferring slower mutagenesis – lines that do not fix A289C still predominantly adapt through mutations within *ntrB*. In addition, the bias toward *ntrB* mutagenesis remains the case in lines with synonymous variation around *ntrB* site 289, which breaks the local hotspot. Therefore both mutant lines suggest that the *ntrB* locus remains mutable relative to other adaptive genes (namely *glnK*) even without mutation at the 289 hotspot, suggesting the input of a wider mechanism. The argument for a wider mechanism that impacts the entire locus is supported by the lessened parallel evolution closer to the origin, as this location change appears to disproportionally effect mutation A289C, with other *ntrB* mutations becoming more common.

We can begin to unearth the wider mechanisms enabling hotspot formation through analysing the swapped strand orientation of the mutable locus. We previously theorised that the mutational hotspot, parallel to nucleotide resolution in approximately 95% of evolved lines, was the product of DNA secondary structure (Supplementary Fig. 2.2). We identified local runs of repeat nucleotides which could facilitate the formation of a stem-loop that forms on the lagging strand during genome replication. Such DNA hairpins have been noted in other studies to facilitate repeatable mutation events at the same nucleotide positions (Dutra and Lovett 2006; Klaric et al. 2020). In our case the frequently mutated base, *ntrB* A289, was predicted to lie at the base of the stem, an area particularly sensitive to mutation (Wright et al. 2003). In this study we have shown that highly repeatable evolution of mutation A289C requires both the local genetic sequence to remain unchanged (allowing the inverse repeats to stay intact) and for the gene to remain in its native orientation. This finding supports our mechanistic hypothesis, as transcription activity remains unchanged regardless of gene orientation (Supplementary Table 3.2). As such the stark differences in mutation spectrums show that transcription alone is not responsible for the hotspot. Genomic mutational biases often occur during either replication or transcription or both (see section 1.2), and so we can therefore infer that replication is involved in hotspot mutagenesis. This synergises both with the finding that translocation toward the replication origin alters repeatability, and with a mutagenic mechanism that involves hairpin formation, as these features can invoke mutation by causing stalling of the replication fork (Wang and Vasquez 2017).

An explanation that can marry the observed impact of hairpin structure (local sequence), genomic location and operon orientation is that mutability across the locus is facilitated by head-on collisions between the RNA polymerase complex and the replication fork (Pomerantz and O'Donnell 2010). Such collisions have been observed to cause mutagenesis at impact sites, and collisions are much more common when transcription and the replication fork progress in opposing directions (Merrikh 2018). Our data show a clear pattern where all lines that are transcribed toward the replication fork preferentially mutate within *ntrB* (where head-on collisions can occur), and the lines that are co-directional with the replication fork (where head-on collisions cannot occur) instead fix mutations in *glnK.*

If head-on collisions are behind the *ntrB* locus's heightened mutability, then it is unclear if the hairpin-facilitated replication stalling at site 289 is independent or similarly reliant on this mechanism. However, a previous study using *E. coli* has established that a quasipalindrome-based mutational hotspot only exerts an effect when transcribed facing the replication fork, and that the mutation occurs more readily when transcription rate increases, which suggests that hairpins and collisions are connected (Yoshiyama and Maki 2003). Stalling at hairpins naturally causes the replication fork to slow or pause (Labib and Hodgson 2007), and so this may bias the location of collisions between replication and transcription apparatus to the hotspot site (Wang and Vasquez 2017). This may also explain the difference in parallel evolution at different genomic locations, as improved fidelity close to the *oriC* may reduce replication stalling but head-on collisions would continue at the same rate. Therefore mutations at site 289 would decrease but mutations across the locus would remain high. An alternative possibility is that the mutagenic hairpin forms after the head-on collision in the resultant R-loop, a nucleic acid structure where DNA and RNA become entangled (Lang et al. 2017). However, the literature more readily supports the former hypothesis.

One consideration with this mechanistic assertion – that head-on collisions between polymerases facilitate heightened mutagenicity, and that these are biased to occur at hairpin sites because they invoke stalling – is the observation that hairpin formation still plays a role when the operon is transcribed on the leading strand. Hairpins have been noted to have more opportunity to form on the lagging strand, as this strand of DNA remains single-stranded for longer periods during replication (Bikard et al. 2010). Additionally, as head-on collisions no longer occur in this strand orientation, we may expect to not observe any mutagenic impact from a fork-stalling DNA motif. However, the predicted secondary structure formed at the 289 hotspot involves a short local tract of nucleotides (Supplementary Fig. 2.2) and so these may be able to form on the leading strand during replication. Additionally, although collisions are more common when DNA and RNA polymerases are antagonistically orientated, collisions between the two can also occur co-directionally as the replication fork migrates faster than transcription machinery (Merrikh 2018). Therefore the effect of the hairpin may be weakened when the operon is transcribed on the leading strand, and so A289C is not observed, but it still engenders some stalling and so mutations at the locus become more likely when the DNA secondary structure is present.

In this work, we have established a hierarchy of genetic features that facilitate the construction of a mutational hotspot, which we have demonstrated by removing these building blocks piece by piece. We moved a hotspot-harbouring locus to a new genomic location and observed that it lowers the hotspot's repeatability. We then introduced synonymous variation and lowered the hotspot's resolution. Finally we flipped the strand orientation of the locus and eradicated the hotspot entirely. In tandem, we demonstrated that alternative adaptive transition mutational events are disproportionally suppressed by mismatch repair proteins. Our work therefore illuminates the key genetic variables that allow mutational hotspots to occupy the "goldilocks zone" of mutability, which enables them to generate near-deterministic evolutionary outcomes from the chance event that is mutation.

Considerable inroads have already been made into forecasting evolution, but predicting the emergence of an exact genotype remains especially difficult, especially when forecasts are made prior to mutation. Adaptive mutational hotspots, however, offer a means by which we can achieve this. In this work, we have showcased the key factors one must consider when searching for these hotspots across the bacterial genome. This is a key step forward in our goal to form accurate evolutionary forecasts, as it will enable us to identify other such sites and uncover their impact on driving adaptive evolution.

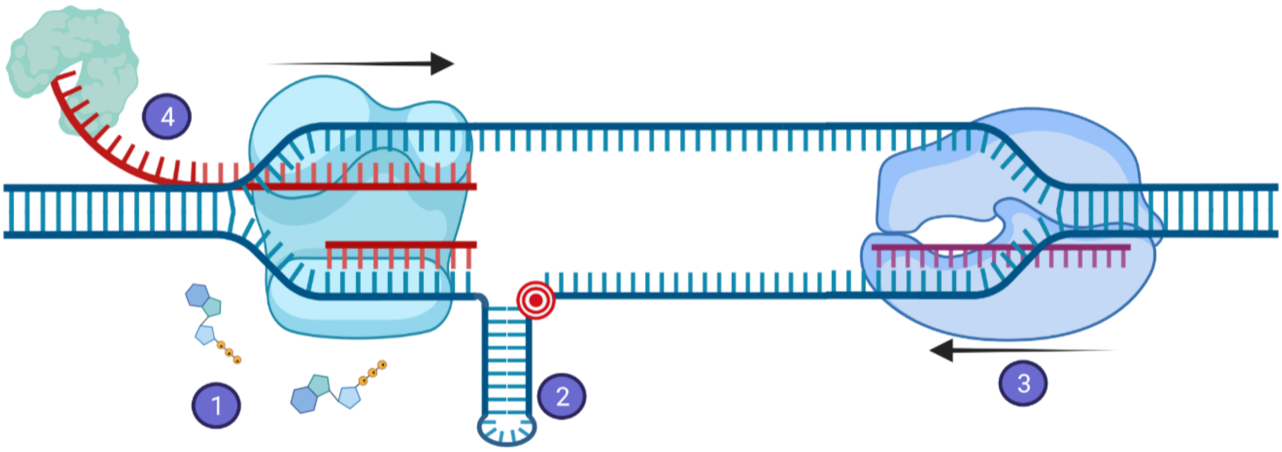


Figure 3.4. A framework for repeatable genetic evolution. In this work we have demonstrated that multiple factors must be satisfied to achieve repeatable genetic evolution of high fidelity. **1)** Replication timing. DNA polymerase complexes operate with fluctuating fidelity depending on the elapsed time since the initiation of replication, which has been proposed to be owed to the level of dNTPs available to the polymerase complex. Bacteria that demonstrate a mutational 'bulge' in mutation accumulation experiments showcase areas where complexes are more prone to stalling. In many bacteria replication timing is highly correlated with distance from the replication origin. **2)** Stem-loop DNA secondary structures. Neighbouring repeat regions of nucleotides form complementary pairs when existing as single-stranded DNA during replication or transcription. Complementary pairs form the stem, and the unpaired nucleotides that separate them form the loop. DNA on the lagging strand, which spends protracted periods as single strands, is particularly susceptible to intra-strand bonding. DNA polymerases are prone to stalling or pausing at stable secondary structure sites, which is more likely to occur depending on replication timing. **3)** Head-on collisions. Genes transcribed antagonistically to the direction of the replication fork engender head-on collisions between DNA and RNA polymerases when replication and transcription occur simultaneously. These collisions are enriched to occur at positions folded into secondary structures, as DNA polymerases stall at these sites. Sites near the base of stem-loop structures have been documented as uniquely vulnerable to mutation (**bullseye**), and unpaired bases within stems can mutate predictably owing to template-switching, wherein the complementary nucleotides in the stem are replicated in place of their complement, which converts imperfect repeats into perfect repeats. **4**) Suppression of alternate mutation types. DNA mismatch repair complexes can preferentially fix certain mutation types over others (e.g. transitions over transversions). Adaptive mutations at mutable sites have a higher likelihood of being realised if repair machinery in the genome can correct adaptive mutations of different types occurring elsewhere, and the mutations are themselves not of that type.

3.8. References

1. Alsohim AS, Taylor TB, Barrett GA, Gallie J, Zhang X, Altamirano-Junqueira AE, Johnson LJ, Rainey PB, Jackson RW. 2014. The biosurfactant viscosin produced by Pseudomonas fluorescens SBW25 aids spreading motility and plant growth promotion. Environ. Microbiol. 16:2267–2281.
2. Bikard D, Loot C, Baharoglu Z, Mazel D. 2010. Folded DNA in Action: Hairpin Formation and Biological Functions in Prokaryotes. Microbiol. Mol. Biol. Rev. 74:570–588.
3. De Boer JG, Ripley LS. 1984. Demonstration of the production of frameshift and base-substitution mutations by quasipalindromic DNA sequences.
4. Bryksin A V, Matsumura I. 2010. Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. Biotechniques 48:463–465.
5. Choi KH, Schweizer HP. 2006. mini-Tn7 insertion in bacteria with single attTn7 sites: Example Pseudomonas aeruginosa. Nat. Protoc. 1:153–161.
6. Dettman JR, Sztepanacz JL, Kassen R. 2016. The properties of spontaneous mutations in the opportunistic pathogen Pseudomonas aeruginosa. BMC Genomics [Internet] 17:1–14. Available from: http://dx.doi.org/10.1186/s12864-015-2244-3
7. Dillon MM, Sung W, Lynch M. 2018. Periodic Variation of Mutation Rates in Bacterial Genomes. MBio 9:1–15.
8. Dutra BE, Lovett ST. 2006. Cis and trans-acting effects on a mutational hotspot involving a replication template switch. J. Mol. Biol. 356:300–311.
9. Ganai RA, Johansson E. 2016. DNA Replication—A Matter of Fidelity. Mol. Cell [Internet] 62:745–755. Available from: http://dx.doi.org/10.1016/j.molcel.2016.05.003
10. Gon S, Camara JE, Klungsøyr HK, Crooke E, Skarstad K, Beckwith J. 2006. A novel regulatory mechanism couples deoxyribonucleotide synthesis and DNA replication in Escherichia coli. EMBO 25:1137–1147.
11. Hervas AB, Canosa I, Little R, Dixon R, Santero E. 2009. NtrC-Dependent Regulatory Network for Nitrogen Assimilation in Pseudomonas putida. J. Bacteriol. 191:6123–6135.
12. Hmelo LR, Borlee BR, Almblad H, Love ME, Randall TE, Tseng BS, Lin CY, Irie Y, Storek KM, Yang JJ, et al. 2015. Precision-engineering the Pseudomonas aeruginosa genome with two-step allelic exchange. Nat. Protoc. 10:1820–1841.
13. Hudson RE, Bergthorsson U, Roth JR, Ochman H. 2002. Effect of Chromosome Location on Bacterial Mutation Rates. 19:85–92.
14. Juurik T, Ilves H, Teras R, Ilmjarv T, Tavita K, Ukkivi K, Teppo A, Mikkel K, Kivisaar M. 2012. Mutation Frequency and Spectrum of Mutations Vary at Different Chromosomal Positions of Pseudomonas putida. PLoS One 7.

15. Kim DJ, Forst S. 2001. Genomic analysis of the histidine kinase family in bacteria and archaea. Microbiology 147:1197–1212.

16. Klaric JA, Perr EL, Lovett ST. 2020. Identifying Small Molecules That Promote Mutations in Escherichia coli. G3 10:1809–1815.

17. Krašovec R, Richards H, Gomez G, Gifford DR, Mazoyer A, Knight CG. 2019. Measuring microbial mutation rates with the fluctuation assay. J. Vis. Exp. 2019:1–9.

18. Labib K, Hodgson B. 2007. Replication fork barriers: Pausing for a break or stalling for time? EMBO Rep. 8:346–353.

19. Lang KS, Hall AN, Merrikh CN, Woodward JJ, Dreifus JE, Lang KS, Hall AN, Merrikh CN, Ragheb M, Tabakh H, et al. 2017. Replication-Transcription Conflicts Generate R-Loops that Orchestrate Bacterial Stress Survival Article Replication-Transcription Conflicts Generate R-Loops that Orchestrate Bacterial Stress Survival and Pathogenesis. Cell [Internet] 170:787-790.e18. Available from: http://dx.doi.org/10.1016/j.cell.2017.07.044

20. Li W, Lu CD. 2007. Regulation of carbon and nitrogen utilization by CbrAB and NtrBC two-component systems in Pseudomonas aeruginosa. J. Bacteriol. 189:5413–5420.

21. Liu Y, Rainey PB, Zhang XX. 2014. Mini-Tn7 vectors for studying post-transcriptional gene expression in Pseudomonas. J. Microbiol. Methods 107:182.

22. Long H, Miller SF, Williams E, Lynch M. 2018. Specificity of the DNA Mismatch Repair System (MMR) and Mutagenesis Bias in Bacteria. Mol. Biol. Evol. 35:2414–2421.

23. Long H, Sung W, Miller SF, Ackerman MS, Doak TG, Lynch M. 2014. Mutation rate, spectrum, topology, and context-dependency in the DNA mismatch repair-deficient Pseudomonas fluorescens ATCC948. Genome Biol. Evol. 7:262–271.

24. Meisner J, Goldberg JB. 2016. The Escherichia coli rhaSR-PrhaBAD inducible promoter system allows tightly controlled gene expression over a wide range in Pseudomonas aeruginosa. Appl. Environ. Microbiol. 82:6715–6727.

25. Merrikh H. 2018. Spatial and temporal control of evolution through replication- transcription conflicts. Trends Microbiol. 25:515–521.

26. Moxon R, Bayliss C, Hood D. 2006. Bacterial Contingency Loci : The Role of Simple Sequence DNA Repeats in Bacterial Adaptation. Annu. Rev. Genet. 40:307–335.

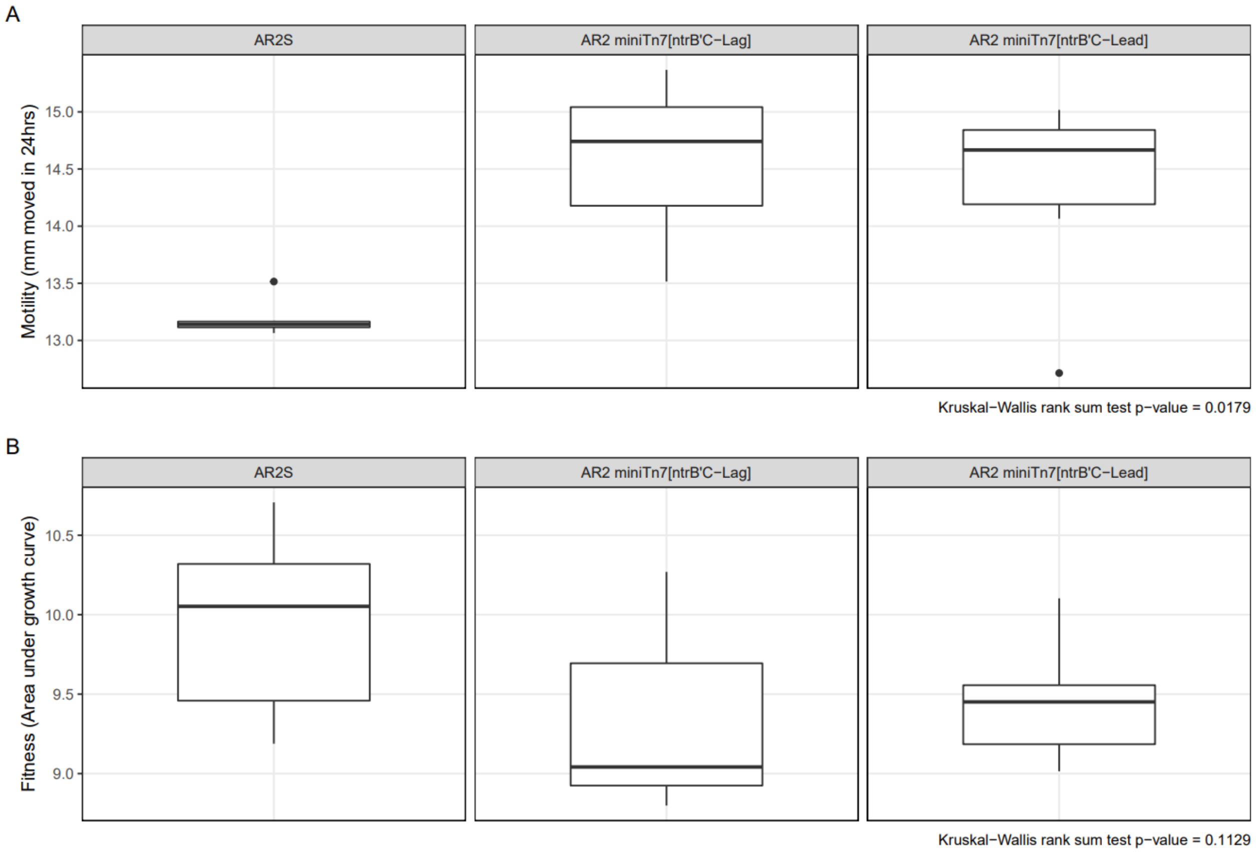
27. Paul S, Million-Weaver S, Chattopadhyay S, Sokurenko E, Merrikh H. 2013. Accelerated gene evolution via replication-transcription conflicts. Nature 495:1–13.

28. Pomerantz RT, O’Donnell M. 2010. What happens when replication and transcription complexes collide? Cell Cycle 9:2537–2543.

29. Schaaper M, Dunn RL. 1987. Spectra of spontaneous mutations in Escherichia coli strains defective in mismatch correction : The nature of in vivo DNA replication errors. PNAS 84:6220–6224.

30. Schofield MJ, Brownewell FE, Nayak S, Du C, Kool ET, Hsieh P. 2001. The Phe-X-Glu DNA Binding Motif of MutS. J. Biol. Chem. [Internet] 276:45505–45508. Available from: http://dx.doi.org/10.1074/jbc.C100449200
31. Silby MW, Cerdeño-tárraga AM, Vernikos GS, Giddens SR, Jackson RW, Preston GM, Zhang X, Moon CD, Gehrig SM, Godfrey SAC, et al. 2009. Open Access Genomic and genetic analyses of diversity and plant interactions of Pseudomonas fluorescens. Genome Biol. 10:R51.
32. Sprouffske K, Wagner A. 2016. Growthcurver: An R package for obtaining interpretable metrics from microbial growth curves. BMC Bioinformatics 17:17–20.
33. Taylor TB, Mulley G, Dills AH, Alsohim AS, McGuffin LJ, Studholme DJ, Silby MW, Brockhurst MA, Johnson LJ, Jackson RW. 2015. Evolutionary resurrection of flagellar motility via rewiring of the nitrogen regulation system. Science (80-.). 347:1014–1017.
34. Vogwill T, Kojadinovic M, Furió V, Maclean RC. 2014. Testing the role of genetic background in parallel evolution using the comparative experimental evolution of antibiotic resistance. Mol. Biol. Evol. 31:3314–3323.
35. Wang G, Vasquez KM. 2017. Effects of replication and transcription on DNA Structure-Related genetic instability. Genes (Basel). 8.
36. Wong A, Rodrigue N, Kassen R. 2012. Genomics of Adaptation during Experimental Evolution of the Opportunistic Pathogen Pseudomonas aeruginosa. PLoS Genet. 8.
37. Wright BE, Reschke DK, Schmidt KH, Reimers JM, Knight W. 2003. Predicting mutation frequencies in stem-loop structures of derepressed genes: Implications for evolution. Mol. Microbiol. 48:429–441.
38. Yoshiyama K, Maki H. 2003. Spontaneous Hotspot Mutations Resistant to Mismatch Correction in Escherichia coli: Transcription-dependent Mutagenesis Involving Template-switching Mechanisms. J. Mol. Biol. 327:7–18.

3.9. Supplementary materials



Supplementary Figure 3.1. Impact of translocation and DNA strandedness changes on motility and shaking LB growth fitness of the *ntrB* A289C mutant. **A)** Motility of A289C *ntrB* mutation in each genomic topology condition as measured by distance moved after 24 h in LB 0.25% agar (mm). AR2 NTRB A289C contains AR2 *ntrB'* (*ntrB*-A289C) in its native genomic position and strand orientation. No significant difference was found between AR2 miniTn7[*ntrB'C*-Lag] and AR2 miniTn7[*ntrB'C*-Lead] (*P* = 0.2667, Dunn test), however both moved significantly further than AR2 NTRB A289C (*P* = 0.0034 and *P* = 0.0185 respectively, Dunn test). **B)** Fitness of A289C *ntrB* mutation for each genomic topology condition as measured by area under the growth curve for 24 h growth in shaking LB broth. No significant difference was found between any strain tested (*P* = 0.1129, Kruskal-Wallis rank sum test).

Supplementary Table 3.1. Bacterial strains and culture conditions used in this study.

Strain	Media supplement
P. fluorescens AR2 and derivatives	Kanamycin sulphate 50µg/mL
P. fluorescens AR2 derivatives containing miniTn7 transposons	Kanamycin sulphate 50µg/mL, Gentamicin sulphate 5µg/mL
E. coli DH5α containing pJM220-derived plasmids	Ampicillin sodium salt 100µg/mL, Gentamicin sulphate 5µg/mL
E. coli SM10 λpir pTNS2	Ampicillin sodium salt 100µg/mL
E. coli SP50 pRK2073	Streptomycin sulphate 100µg/mL
E. coli ST18 derived strains	5-aminolevulinic acid hydrochloride (5-ALA) 50µg/mL
E. coli containing pTS1-derived plasmids, and *P. fluorescens* allelic exchange merodiploids	Tetracycline hydrochloride 10µg/mL

Supplementary Table 3.2. No significant changes in expression of the *ntrB* or *ntrC* genes was observed for each genomic topology condition. Expression was assayed for biological triplicates each in technical triplicates, using the the comparative Ct (ΔΔCt) method with *gyrB* as an endogenous reference. P-values produced using Dunn tests.

Strain	*ntrB*	
	Fold change relative to AR2	p-value
AR2 miniTn7[*ntrBC* -Lag]	0.774836401	0.4543
AR2 miniTn7[*ntrBC* -Lead]	1.314879894	0.2036
	ntrC	
	Fold change relative to AR2	p-value
AR2 miniTn7[*ntrBC* -Lag]	1.095604638	0.4226
AR2 miniTn7[*ntrBC* -Lead]	1.784101725	0.2189

Supplementary Table 3.3. Details of oligonucleotide primers used in this work.

Primer name	Sequence 5’-3’	Purpose
ntrC-down-R	GAAATTAATAGGTTGTATTGATGTTGTACCAGGGCTCCCAAAAC	Amplify downstream homologous region of ntrBC for knockout allelic exchange. Introduces overlap for incorporation into pTS1 vector by SOE-cloning.
ntrBC-KO-down F	GACCATCAGCGATGCACTGGGCGATGAAGGCTGAATC	
ntrBC-KO-up-R	GATTCAGCCTTCATCGCCCAGTGCATCGCTGATGGTC	Amplify upstream homologous region of ntrBC for knockout allelic exchange. Introduces overlap for incorporation into pTS1 vector by SOE-cloning.
ntrBC-KO-up-F	GCCGTTTCTGTAATGAAGGAGAAAACCCGAGATGGTAGGCATTGAAC	
mutS_out_NF	GCATGGGCGACTTCTACGAG	Amplify upstream homologous region of mutS for knockout allelic exchange. Contains complementary overhangs with downstream fragment.
mutS_ins_NR	GTGCATATAACATTTCGAGCGCAGGGCGGTGCGCTGGGTTT	
mutS_ins_CF	AAACCCAGCGCACCGCCCTGCGCTCGAAATGTTATATGCAC	Amplify downstream homologous region of mutS for knockout allelic exchange. Contains complementary overhangs with upstream fragment.
mutS_out_CR	AATTTAAGCTTCGCTATCAGCGTTCGAGGTC	
mutS_nest_NF	AATTTGGATCCGTTGCTGGACATCACCCTG	Amplifies across annealed upstream and downstream mutS fragments for knockout allelic exchange. Introduces restriction enzyme recognitions sites BamH1 and HindIII to termini of fragment for incorporation into pTS1 vector by restriction-ligation.
mutS_nest_CR	AATTTAAGCTTACTTCGTCCTCGGAGAAAAT	
ntrBC-SacI-F	AATTTGAGCTCCACTGTCCGAACAACACTGATC	Amplify ntrBC and their promoter+terminator regions, introducing a SacI site upstream, and a HindIII site downstream.
ntrBC-HindIII-R	AATTTAAGCTTCGGTTCATGGTGCATTGAAGC	
ntrBC-HindIII-F	AATTTAAGCTTCACTGTCCGAACAACACTGATC	Amplify ntrBC and their promoter+terminator regions, introducing a HindIII site upstream, and a SacI site downstream.
ntrBC-SacI-R	AATTTGAGCTCCGGTTCATGGTGCATTGAAGC	
SBW25-glmS	CACCAAAGCTTTCACCACCCAA	SBW25 glmS primer sequence from Liu et al., 2014.
ntrC-Down-F	GACATGAGCCGTAGTGAAACCGGGCGATGAAGGCTGAATC	Pair checks insertion of ntrBC for lagging strand orientation.
ntrB-1119-F	GAGGTCCCAATGACCATCAG	Amplification of ntrB for Sanger sequencing.
ntrB-1119-R	GACGATCCAGACGGTTTCAC	
glnK-F	GTGGGCAAAGGACTGATTTC	Amplification of glnK for Sanger sequencing.
glnK-R	GATGATGGCGAAGGTCATCT	
PFLU1131-F	CGATAAGCGAAACCTGATG	Amplification of PFLU1131 for Sanger sequencing.
PFLU1131-R	CGACTACCAGAATGTTATGCG	
glnA-F	CGGAAATCGCTCAAGGTTTA	Amplification of glnA for Sanger sequencing.
glnA-R	CTGATAATCCCCAGGCAAAA	
glnK-long-F	CTCCAGGTTCTCCAGGCG	Amplify the glnK and amtB locus for Sanger sequencing of larger deletion mutants.
glnK-long-R	GCCCATCGGCGCGCATTC	
ntrB-F-qPCR	CTTGCGCCTTGAGTACATGA	RT-qPCR primer pair for ntrB – from Taylor et al., 2015.
ntrB-R-qPCR	GTTGCTCAGGATAGGGGTC	
ntrC-F-qPCR	GCCGTAGTGAAACCGTCTG	RT-qPCR primer pair for ntrC– from Taylor et al., 2015.
ntrC-R-qPCR	CATGCGGATGTCGGAGATG	
gyrB-F-qPCR	CGTCACACCATCCAGCGAT	RT-qPCR primer pair for gyrB– from Taylor et al., 2015.
gyrB-R-qPCR	AAGTCACGACGAGGCTCGA	

Chapter IV

4.1. Investigating the evolutionary origins of a mutational hotspot

4.1.1. Abstract

Evolution can sometimes find remarkably repeatable genetic solutions due to the power of mutational hotspots, which bias mutation at adaptive sites. Yet despite their power to define evolutionary trajectories, the origins of mutational hotspots are rarely investigated. These mutagenic sites may be enforced by selection at locations where they are situationally adaptive, allowing them to operate as "contingency loci" that provide a mutational means to cope with environmental change. Alternatively, highly mutagenic sites may transiently appear through neutral evolution throughout the genome. The former scenario would showcase the importance of examining a genome's adaptive past to better predict its future. The latter scenario would exemplify the power of genetic drift in facilitating parallel evolutionary outcomes. In this work, I investigate the evolutionary origins of a mutational hotspot and reveal evidence that the mutational hotspot is not preserved by selection. In previous work I established that a mutational hotspot, which can be built and broken by just 6 synonymous nucleotide changes, was responsible for driving highly repeatable genetic evolution in immotile variants of *Pseudomonas fluorescens*. The frequently realised mutation (*ntrB* A289C) is highly adaptive in immotile variants, however in this work I show that it is maladaptive in the wild-type motile ancestral genetic background (strain SBW25). In addition, multiple sequence alignments of *P. fluorescens* strains reveal an increased rate of synonymous divergence around the hotspot site relative to the rest of the locus, providing a signature of selection in the strain's evolutionary history. I experimentally explored a role for positive selection by assessing the mutational hotspot's ability to operate as a contingency locus under fluctuating environments. These tests revealed a plethora of highly fit compensatory mutations that lead to rapid degradation of the hotspot. Together these results find no evidence for positive selection for the maintenance of a mutational hotspot, and thus instead infer that such hotspots arise through drift and may be acted upon by purifying selection. This work therefore shows that powerful mutational hotspots, rather than being preserved only in loci in which there are adaptive, may readily appear and play a role in determining evolutionary outcomes throughout the genome.

4.2. Introduction

Mutational hotspots have been revealed as key agents for defining a microbe's evolutionary trajectory. They can constrain explored genetic space, driving evolving populations to persistently realise the same adaptive solutions. Hotspots can be established by short tracts of genetic sequence, such as a run of homopolymeric nucleotides (Orsi et al. 2010), or tandem repeats (Zhou et al. 2014), or a local complimentary run of repeat nucleotides (Dutra and Lovett 2006). Due to the low number of nucleotide changes required for their formation, these important evolutionary features may readily evolve via neutral evolution. If these hotspots appear in loci wherein genetic changes will incur significant phenotypic change, they may well be suppressed by purifying selection. If they are not overtly deleterious, then mutations at hotspot sites may become fixed in the population and subsequently lose their heightened mutagenicity (Lavi et al. 2018). However, the mutagenicity at hotspot sites may likewise be maintained by selection if populations are subjected to fluctuating environments, wherein mutations within these genes are often situationally advantageous (Zhou et al. 2014). As we search for mutagenic features across the genome, a selectively enforced evolutionary origin would instruct us to consider an organism's evolutionary history. In contrast, a neutral origin points toward hotspots appearing throughout the genome, and highlights the power of drift to 'accidentally' facilitate subsequent parallel evolutionary outcomes. Whether highly deterministic mutational hotspots are products of selection or drift however remains an open, and under investigated, question.

In 1994 Moxon and colleagues proposed the notion of a contingency locus as a means of coping with environmental change (Moxon et al. 1994). Bacteria are equipped with often robust and expansive gene regulatory networks (Prud'homme et al. 2007), which are comprised of proteins that respond to molecular components in their surroundings by either supressing or enabling the expression of downstream effector genes. Gene regulatory networks therefore offer a means for bacteria to cope with environmental change by amending gene expression according to their needs in a particular environment. A contingency locus offers an alternative means of coping with environmental change by taking advantage of mutation (Zhou et al. 2014). As certain sites across the genome mutate at higher rates than elsewhere, such as homopolymeric tracts and tandem repeats which were the focus of Moxon's work (Moxon et al. 2006), in large populations standing genetic variation should appear at these sites, generating multiple combinations of alleles. The generated mutational diversity may encode multiple phenotypes, some of which may have different consequences on fitness depending on the environment. As such contingency loci offer a form of evolutionary "bet hedging" (Beaumont et al. 2009), wherein standing variation ensures that at least a portion of the population survive following environmental change. If the enriched lines maintain higher mutation rates at the adaptive positions, some will rapidly undergo genetic reversion and restore standing genetic variation in the population. This is an integral element of hotspot maintenance, as reversion mutations are constantly challenged by compensatory mutations elsewhere in the genome. While genetic reversion restores the original phenotype and retains the genetic information of the wild-type, compensatory mutation removes genetic information by committing the genotype further along the mutational path away from the wild-type genotype. Thus, the efficacy of mutations at the hotspot will change owing to the altered regulatory or coding region, and the hotspot will eventually degrade. However, an environmental shift that enriches a subset of the population possessing a persistently mutable hotspot can introduce population bottlenecks that select for these hotspots, ensuring their maintenance in a population (Moxon et al. 2006).

In previous work, we have identified a mutational hotspot that operates in an immotile variant of the soil bacterium *Pseudomonas fluorescens* SBW25. The hotspot is predicated on a small repeat region of nucleotides and therefore is sensitive to local genetic sequence, with just 6 synonymous differences between two non-motile strains of *P. fluorescens* defining the presence or absence of the hotspot. As the hotspot lies within the histidine kinase of a two-component system responsible for regulating nitrogen, mutations at the hotspot can engender significant phenotypic change. This is especially prevalent in the case of our model system, because the response-regulator of the nitrogen pathway (*ntrC*) has retained sufficient homology to the master regulator of flagellar-motility (*fleQ*), that when expressed at high enough levels it can initiate the expression of flagella genes (Taylor et al. 2015). Therefore mutation at the hotspot enables cross-talk between two core regulatory networks, providing an opportunity for situational selective advantage. Conversely, as the hotspot can be destroyed and created by very few silent mutations, its presence may instead be the consequence of genetic drift. Determining whether the hotspot has been enforced by selection or has appeared through drift therefore has powerful implications. Either positive selection has generated a hotspot that facilitates situational entanglement of two gene regulatory networks; or genetic drift has assembled a near-deterministic hotspot with powerful evolutionary consequences, which infers that such hotspots could be readily constructed during neutral evolution elsewhere in the genome.

In this work, I investigate the evolutionary origins of this mutational hotspot by addressing the role played by selection under fluctuating environments on hotspot maintenance. For the hotspot to persist under selection over multiple bottlenecks, the hotspot itself must not be completely degraded under fluctuating selective regimes. I first reveal that while *ntrB* A289C offers no discernible advantage to fitness in wild-type SBW25 (which has its native flagellar-regulator intact), in the immotile lines possession of a mutational hotspot provides a strong evolutionary advantage under selection for motility. I subsequently therefore treat our engineered lines as a toy model to investigate the maintenance of hotspots predicated on local stretches of repeat nucleotides.

I next demonstrate using a theoretical model that for the hotspot to remain stable under fluctuating environments, either: *(i)* the mutation bias that operates in one direction – i.e. the mutation from the wild type genotype to the frequently realised mutant genotype – must also operate in the reverse. Or *(ii)* if the mutation bias only operates in one direction, genetic reversion must offer superior fitness to compensatory mutations. I find that by exploiting the antagonistic pleiotropy engendered by *ntrB* A289C's disruption of nitrogen regulation, simply removing selection for motility is enough to reverse directional selection away from the *ntrB* A289C genotype. Within the fluctuated environment, all evolved replicates underwent rapid phenotypic reversion, however all independent lines achieved this through compensatory evolution rather than genetic reversion. Subsequent phenotyping of compensatory mutants revealed equivalent fitness to a genetic revertant (wild type) in agitated liquid culture. The finding of compensatory mutation preference held when selecting for the most rapidly realised phenotypic revertant and when performing mixed-culture deep sequencing on phenotypic revertant lines. Subsequent re-evolution of motility took significantly longer in the batch of compensatory strains than wild-type non-motile SBW25 lines, and overall the compensatory mutants became less evolvable with regards to acquisition of the motility phenotype.

These results provide evidence that mutational hotspots predicated on short tracts of repeat regions can be rapidly degraded in fluctuating environments, and therefore suggest that they are not maintained by selection or operate as contingency loci. As a consequence, the results indirectly implicate genetic drift as the architect of the evolution-defining hotspot, and as such infer that similarly powerful hotspots may readily appear across bacterial genomes.

4.3. Materials and Methods

4.3.1. Strains

This study employed engineered non-flagellate strains of the soil microbe *Pseudomonas fluorescens*. Non-flagellate lines were constructed from two *P. fluorescens* strain backgrounds, SBW25 (Rainey and Bailey 1996) and Pf0-1 (Compeau et al. 1988), through functional deletion of the master regulator of flagella motility, *fleQ* (Alsohim et al. 2014; Taylor et al. 2015), and the biosurfactant-production regulator *viscB* in SBW25 (Alsohim et al. 2014). Immotile SBW25 (hereafter AR2) and Pf0-1 (hereafter Pf0-2x) were further engineered by two-step allelic exchange (methodology outlined in Hmelo et al. 2015; amendments and strain construction outlined in section 2.5), to create mutant lines with 6 silent mutations within the *ntrB* locus around site 289 (hereafter AR2-sm and Pf0-2x-sm; (Fig. 2.4 and Fig. 2.5). The mutation *ntrB* A289C was introduced to the ancestral SBW25 genetic background via two-step allelic exchange using a plasmid containing this mutation constructed in previous work (section 2.5 and Supplementary Table 2.1). The plasmid utilised for this process was the same construct used to assemble strain AR2-sm (section 2.5 and Fig. 2.4), and as such SBW25 *ntrB* A289C additionally contains six synonymous mutations around the hotspot site. These are not expected to have any fitness consequences in the wild-type background following A289C mutation, which is supported by phenotyping assays performed in an AR2 background in Supplementary Fig. 2.3. Clonal cell lines were stored at -80°C in 20% glycerol. Biological replicates were prepared for assays from single clonal colonies grown overnight in 10 ml LB broth with cell densities corrected to 1 OD unit/ml prior to start of the experiment, via cell pelleting and re-suspension in phosphate buffer saline. The nutrient conditions used throughout the work were lysogeny broth (LB) and M9 minimal media containing glucose and 7.5 mM NH4, unless otherwise stated that 8mM lysine was used as the sole nitrogen source. All cells were grown at 27°C.

4.3.2. Phenotyping of SBW25 *ntrB* A289C

The impact of mutation *ntrB* A289C on the motility phenotype and growth yield within the ancestral SBW25 genetic background were explored by analysing race assays and growth in shaking liquid culture. Race assays were performed as described in previous work (section 2.5) across two nutritional environments, LB and M9 minimal media. 6 biological replicates of SBW25 and SBW25 *ntrB* A289C were used to inoculate one independent replicate of soft agar per nutrient condition. Growth yields were determined using the same biological replicates and nutrient broths as prepared for the race assay. Growth in liquid culture was achieved by adding 1 µl OD 1 cells/ml cell culture from 6 biological replicates across both strains and both nutrient conditions to a 96-well plate (Corning®) holding 99 µl of nutrient media per well. Growth was monitored autonomously by a Spark® Multimode Microplate Reader (Tecan) with readings recorded at 10-minute intervals for 24 h. Cells were kept at an agitation of 180 rpm throughout the assay. Relative growth yields were determined using area under the curve as calculated by the trapezoid rule (see Huang and Pang 2012).

4.3.3. Competitive evolution experiments

Six biological replicates of AR2, AR2-sm, Pf0-2x, and Pf0-2x-sm were prepared as outlined above and mixed at equal cell densities (determined by OD) in genotype pairs prior to the beginning the assay. 1 µl of each mixed culture was used to inoculate two soft agar replicates by piercing the top of the agar surface with a pipette tip and ejecting the volume into the cavity as the pipette was withdrawn. Mixed populations were then left to evolve and monitored daily for signs of flagella-mediated motility, which appears as a 'blebbing' that stems from the immotile inoculated cell mass. Cells were sampled from the leading edge ≤ 24 h motility emergence using an inoculating loop and streaked onto selective agar. In rare instances where two distinct motile zones appeared within the 24 h windows between monitoring, both frontier zones were sampled. Two approaches were used to identify which of the two competing genotypes had become established on the frontier. In instances where the competing pair included one AR2-derived strain and one Pf0-2x-derived strain, the sample was streaked onto two types of selective agar using the same inoculating loop. Pf0-2x strains are inhibited by 100 µg/ml kanamycin sulphate, while AR2 strains are inhibited by 250 µg/ml streptomycin sulphate. Thus the genetic background which had established itself on the frontier could be simply distinguished following 48 h of incubation, as growth was evident on one selective agar treatment and entirely absent on the other (Supplementary Fig. 4.2).

A similar approach was used when hotspot variants within the same genetic background were competed with one another. In these cases a single colony from the streak plate was analysed by PCR to identify if the hotspot allele was present on the frontier. This was achieved by using two primer sets. Common to both was a generic reverse primer (relative to operon orientation) with the sequence: 5'-CACTACGGCTCATGTCGATG-3', which contains sequence shared by AR2 and Pf0-2x and so can be used across both genetic backgrounds. This primer was separately combined with two forward primers that bind to the hotspot site, the first of which will only recognise the hotspot allele (sequence: 5'-CGACTACGCCGTGACCCCT-3'), and the second of which will only recognise a non-hotspot allele (sequence: 5'-GGATTACGCGGTGACGCCG-3'; the differences between the two sequences are underlined). Similar to identification on selective agar, following PCR amplification and visualisation on an agarose gel, one band will be clearly visible and the other absent, revealing whether a hotspot or non-hotspot allele is established at the frontier. The cycling conditions for this amplification were as follows: 95°C pre-cycle boil for 3 minutes; 30 cycles consisting of 95°C denaturation for 30 seconds, 59°C annealing for 30 seconds, 72°C extension for 1 minute; final extension for 2 minutes. PCR amplifications were performed using 2x GoTaq® Green Master Mix (Promega).

All six pair-wise competitive evolution experiments were initiated concurrently, and subsequent to this AR2 and Pf0-2x-sm mixed cultures were serially diluted and plated to determine the starting population ratios across genetic backgrounds. It was observed that 1 OD unit of cells did not provide an equal starting population, as AR2 was found to be the predominant strain (mean ratio across 6 biological replicates: AR2/Pf0-2x-sm, 1.52:1). Therefore each pair-wise competition was completed two additional times to yield experimental triplicates. When mixed cultures shared a *P. fluorescens* strain background, cells were consistently mixed at equal cell densities. However the volumes were adjusted when mixtures included those across genetic backgrounds. Initiating the assay with mixture including a cell density of 2 OD units of Pf0-2x-sm to 1 OD unit of AR2 roughly flipped the ratio observed in the original experiment (mean ratio 1:1.7), and so the third experiment used a volume ratio of OD 1.5 units of Pf0-2x-sm to 1 OD unit AR2 to achieve approximate parity. The data set formed from the triplicate experiment was pooled for subsequent statistical analysis, providing a sample range of 23-38 independent replicates per pair-wise comparison.

4.3.4. Reversion assay

Clonal colonies of AR2 *ntrB* A289C were placed under directional selection for phenotypic reversion through growth in M9 minimal medium broth agitated at 180 rpm. 20 independent replicates were prepared by aliquoting 10 ml M9 broth into 30 ml moulded glass vials with fitted screw caps, which were left slightly loosened to allow sufficient aeration. A single colony was used to inoculate each replicate and populations were allowed to incubate for 48 h before 100 µl (0.1% total volume) was transferred to 9.9 ml fresh M9 broth. This transfer was repeat into fresh media at 96 h. However, every 24 h a portion of the culture was removed from the broth, serially diluted and plated onto LB agar for identification of phenotypic revertants. Colonies that had restored wild-type fitness could be readily identified by their increased growth rate, leading to significantly larger colonies following 48 h of incubation (Supplementary Fig. 4.3). It was observed that 1 OD unit of cell density diluted to 10^{-5} and plated yielded over 100 separated colonies (Supplementary Fig. 4.3), allowing for a revertant to be identified once it represented approximately 1% of the population pool. One colony per independent replicate was re-streaked to achieve a clonal phenotypic revertant population and analysed through and directed evolution experiments on soft agar (section 2.5), growth yield in agitated broth (see above), and Sanger sequencing of the *ntrB* locus using primers: 5’-GAGGTCCCAATGACCATCAG-3’ and 5’-GACGATCCAGACGGTTTCAC-3’.

Mixed-population deep sequencing was performed on 5 independent populations evolved under the same conditions as those outlined above with the following amendments: (*i*) cells were grown in lysogeny broth (LB) and (*ii*) all lines were evolved for at total period of 96 h until phenotypic revertants had dominated the population pool (identified by daily plating as outlined above). At the end of the directed evolution experiment, 1 ml (10% of final culture volume) was withdrawn and the mixed genomic DNA within each culture was extracted using a Genomic DNA Purification Kit (ThermoFisher Scientific). Whole-genome sequencing on the five independent mixed-populations was performed by MicrobesNG, providing mean read depths ranging from 105.6 – 232.9. Lines were analysed via the Cloud Infrastructure for Microbial Bioinformatics (CLIMB). Variants were searched for using Snippy (Seemann 2015) with default parameters aside from the "--minfrac" argument which was adjusted to lower the proportion of reads required to call a variant to 20%, and Tablet (Milne et al. 2013) for manual analysis of the *ntrB* locus. The SBW25 genome was used a reference for both pieces of software (NCBI Assembly: ASM922v1, GenBank sequence: AM181176.4).

4.3.5. Mutational hotspot model

Hotspot stability under fluctuating selection was modelled using MATLAB®. A network consisting of 16 nodes was interconnected via edges that connected neighbouring nodes to form a symmetrical pattern with Nodes 1 and 2 at the centre (shown in Fig. 4.3A). For the purposes of this model, each node is representative of a distinct genotype, with connected neighbouring nodes representing adaptive genotypes that are separated by one mutational event. Node 1 was chosen to represent the wild-type genotype and Node 2 the genotype that follows mutation at the hotspot site (in this case, simulating *ntrB* A289C).

The network was formed by writing novel arrays and connecting them using a digraph. The weights of the nodes and edges were assigned manually. The weighting of the edges represents mutation rate (μ) that drives transition from one genotype state to its immediate neighbour in mutational space. To model the role of mutation rate at the hotspot site relative to mutation rates elsewhere, the rate of mutation from Node 1-2 (μ_0) and Node 2-1 (μ_1) can be manipulated as unique modifiable rates, and all other rates were standardised (μ_2). The weighting of each node represents a fitness value (λ) uniquely assigned to each genotype, allowing for the simulation of relative fitness values between adaptive mutational routes on hotspot stability. The value of λ is the sole variable to change between fluctuating selective regimes, simulating environment-dependent pleiotropy of each genotype. Finally, the proportion of the population pool which is distributed at each node was recorded every iteration (χ_j). At the start of the simulation, the entire population is assigned to Node 1 (χ_1). Population distribution is modelled in this pipeline solely as relative genotype abundance, which means that each χ will end an iteration of the simulation with a value between 0 and 1, and all χ values sum to 1 between iterations.

The model runs iteratively, re-distributing the relative population pool (χ_j) between genotypes according to network connectivity, mutation rate, and fitness values, for *r* iterations. In this work *r* is set to 10. After these iterations are complete, the fitness values (λ) of each node transitions from being derived from array Set 1 to array Set 2 for the next round of *r* iterations (see Supplementary table 1). Typically, the fitness value assigned to a node in array Set 2 is = 1- λ of array Set 1, simulating a linear relationship between the fitness boon in one environment and the severity of antagonistic pleiotropy in the alternate environment. 10 rounds of transition were used for this work, overall representing hotspot stability over 100 iterations. The novel pipeline written to enable this does so via the following:

The unique mutation rates ($\mu_{0,}\mu_{1,}\mu_2$) that weight the edges are first assigned a value between 0 and 1, with the rate determining the proportion of the population pool at a given genotype (χ) that transitions to a neighbouring genotype at each round. This simulates that mutation precedes selection, as is standard practice in theoretical evolutionary models (Chevin et al. 2010). As a μ approaching 1 would simulate almost the entire population with a given genotype mutating each iteration, mutation rates were kept low for the results presented in this work (although considerably higher than would be observed in nature). The lower boundary was set to 0.0001 and the upper boundary to 0.0095. The population is re-distributed by:

$$\chi_j == \chi_j + \sum_{m=q}^{n-1} \chi_q \bullet \mu_i$$

Where χ_j is the population assigned (χ) at node *j; n* is the amount of edges connected to node *j*; *q* is the number of the neighbouring nodes connected to node *j*; and μ_i is the mutation rate μ at an edge with allocated weight of *i* (where *i* can be 0, 1, or 2). Terms *n, q, j* and *i* are each determined by the digraph array (see Supplementary table 1). Only a small fraction of the population assigned at each node will be re-distributed during each iteration, with the rest being recycled to the same genotype at rate α. For all nodes aside from Nodes 1 and 2, α_2 is determined by:

$$\alpha_2 = 1 - \sum_1^{n-1} u_2$$

As Nodes 1 and 2 possess unique mutation rates, their α values are determined by the following, using Node 1 as an example (Node 2 will include α_1 and μ_1 in place of α_0 and μ_0):

$$\alpha_0 = 1- \left(\sum_1^{n-2} u_2 \right) - \mu_0$$

This ensures that all rates of transition sum to 1, with the non-evolving portion of the population being simulated by transitioning to the same genotype at rate α. Following mutation re-distribution, selection is simulated by multiplying the proportion of the population pool at each node (χ_j) by the node's assigned fitness (λ_j), and normalising the population by dividing this new node value by the summation of all adjusted node values:

$$\chi_j == \frac{(\chi_j \bullet \lambda_j)}{\sum_{j=1}^{16} \chi_j \bullet \lambda_j}$$

So that:

$$1 = \sum_{j=1}^{16} \chi_j$$

The values of χ_j at each iteration were recorded and plotted within Fig. 4.4. Hotspot stability was reflected in whether χ_1 and $\chi_2 \rightarrow 1$ reciprocally during the rounds in which their fitness values were high.

To measure the impact on hotspot stability of: *(i)* relative mutation rates between genotypes, $\mu_{0,}$ and μ_1 were adjusted; and to measure *(ii)* the differential fitness between genetic revertants and compensatory mutants, λ_j values were adjusted.

Given the incomplete knowledge of the mutational spectrum of the mutational hotspot of interest, the network topology encoded for this model is of a small, arbitrary design. From the starting genotype, it represents an adaptive landscape that includes 5 possible adaptive mutational routes. Each of these adaptive nodes are themselves separated by a single mutational event from 2-4 adaptive mutations (see Fig. 4.4A), but these are only adaptive in the alternate environment. This therefore simulates that local fitness optima are always reached by a single adaptive mutation, which is not truly reflective of the model system of interest (Taylor et al. 2015). The network is limited in size by having the intermediate nodes (which are immediate neighbours of either Node 1 or Node 2; Fig. 4.4A) connect to shared nodes at the extremity of the network. This simulates a case of reciprocal-sign epistasis (Poelwijk et al. 2011), which may or may not be present in the model system of interest. In addition, measuring the population using a relative genotype pool provides no opportunity to capture growth phases or population genetic factors such as frequency-dependent selection (Brisson 2018), which could be only modelled if absolute genotype numbers were included. As such network topology and the mathematical framework of the model are neither wholly reflective of the mutational spectrum surrounding the hotspot of interest, nor biologically complex with regards to microbial growth dynamics or the stochasticity of mutation.

4.3.6. *P. fluorescens ntrB* alignments

Homologs of *ntrB* (SBW25 PFLU_0344) within the *P. fluorescens* species complex were collected through The Pseudomonas Database (https://www.pseudomonas.com/). Homologous genes were found using the nucleotide Basic Local Alignment Search Tool (nBLAST), which identified 247 unique positive hits from complete or partial genomic sequences on the database. These sequences were subsequently compared through a Multiple Sequence Alignment using Clustal Omega (Park et al. 2019) and visualised using Jalview (Waterhouse et al. 2009) to establish the consensus at each nucleotide position ±15 base pairs around the mutational hotspot. Locus-wide consensus analysis was completed using a novel pipeline in R (R Core Team 2014) that utilised the R package "seqinr" to download Clustal alignment data. The number of possible synonymous changes at each nucleotide position were also determined using a novel pipeline written in R.

4.3.7. Statistics

All statistical tests were performed in R (R Core Team 2014). A combination of parametric and non-parametric tests were used for Fig. 4.1. A data point was identified and removed from the *ntrB* A289C in M9 area under the curve data set using a Dixon’s Test (R package: outliers), wherein the lowest value was established as an outlier: Q = 0.9033, $P < 0.001$. After outlier removal all area under the curve data sets reported normal distributions following Shapiro-Wilk normality tests: SBW25 in LB, W = 0.89491, *P* = 0.3447, *ntrB* A289C in LB, W = 0.85092, *P* = 0.1601; SBW25 in M9, W = 0.9602, *P* = 0.8213; *ntrB* A289C in M9 (before removing outlier), W = 0.57269, *P* < 0.001; *ntrB* A289C in M9 (after removing outlier), W = 0.95207, *P* = 0.7519. For the race assay comparisons *ntrB* A289C in M9 was found to follow a non-normal distribution (W = 0.74013, *P* < 0.02) and so a non-parametric test was used.

4.4. Results

4.4.1. Mutation *ntrB* A289C is deleterious for wild-type SBW25 in lab conditions

The mutational hotspot operating within locus *ntrB* was first identified in an engineered non-motile variant of *P. fluorescens* SBW25, AR2 (Fig. 2.1). The mutation facilitated by this hotspot was revealed to be highly adaptive in this genetic background, as it enabled binding of the transcription factor NtrC to the enhancer-binding sites of FleQ, a flagellar regulator which had been functionally deleted in this strain (Taylor et al. 2015). This allowed the mutation engendered by the hotspot (*ntrB* A289C) to restore the flagellar motility phenotype in evolved lines. However, if this mutational hotspot's origin was owed to selection, then selection for the hotspot should have been present further back in the strain's evolutionary history, to the wild-type strain SBW25. To investigate if *ntrB* A289C could prove adaptive when *fleQ* (and *viscB*, see Alsohim et al. 2014) were operational in the genome, this mutation was first engineered into wild-type SBW25.

To assess the consequences of this mutation in the wild type (WT) strain, I performed race assays on soft agar plates and compared growth yields in shaking liquid culture. This respectively allowed for an assay that placed emphasis on the motility phenotype, and for any antagonistic pleiotropy from the mutation to be adjudged independently. Furthermore, these assays were performed under high (lysogeny broth: LB) and low (M9 minimal medium) nutrient conditions, to account for the role of environment on expected pleiotropy (see Supplementary Fig. 2.1). Relative to SBW25-WT, SBW25-*ntrB* A289C mutants were observed to cover significantly smaller surface areas over the course of the race assay (Fig. 4.1A-B). This reduced surface area coverage may have been owed to regulatory competition between the native flagellar regulator and an inferior homolog for enhancer binding sites, or it may have been owed to antagonistic pleiotropy. When grown in shaking liquid culture – where motility offers no benefit but pleiotropic effects on growth yield can be measured – SBW25 *ntrB* A289C mutants reached significantly lower yields than their WT counterpart (Fig. 4.1C-D). *P. fluorescens* has been noted to occupy a variety of ecological niches in natural settings (Scales et al. 2014), and thus there may be a fitness benefit following mutation at the hotspot that was not captured by these assays. However, within a laboratory setting, these assays reveal that *ntrB* A289C offers no advantage to motility, but does cause a significant fitness cost with regards to growth yield in the SBW25 wildtype background.

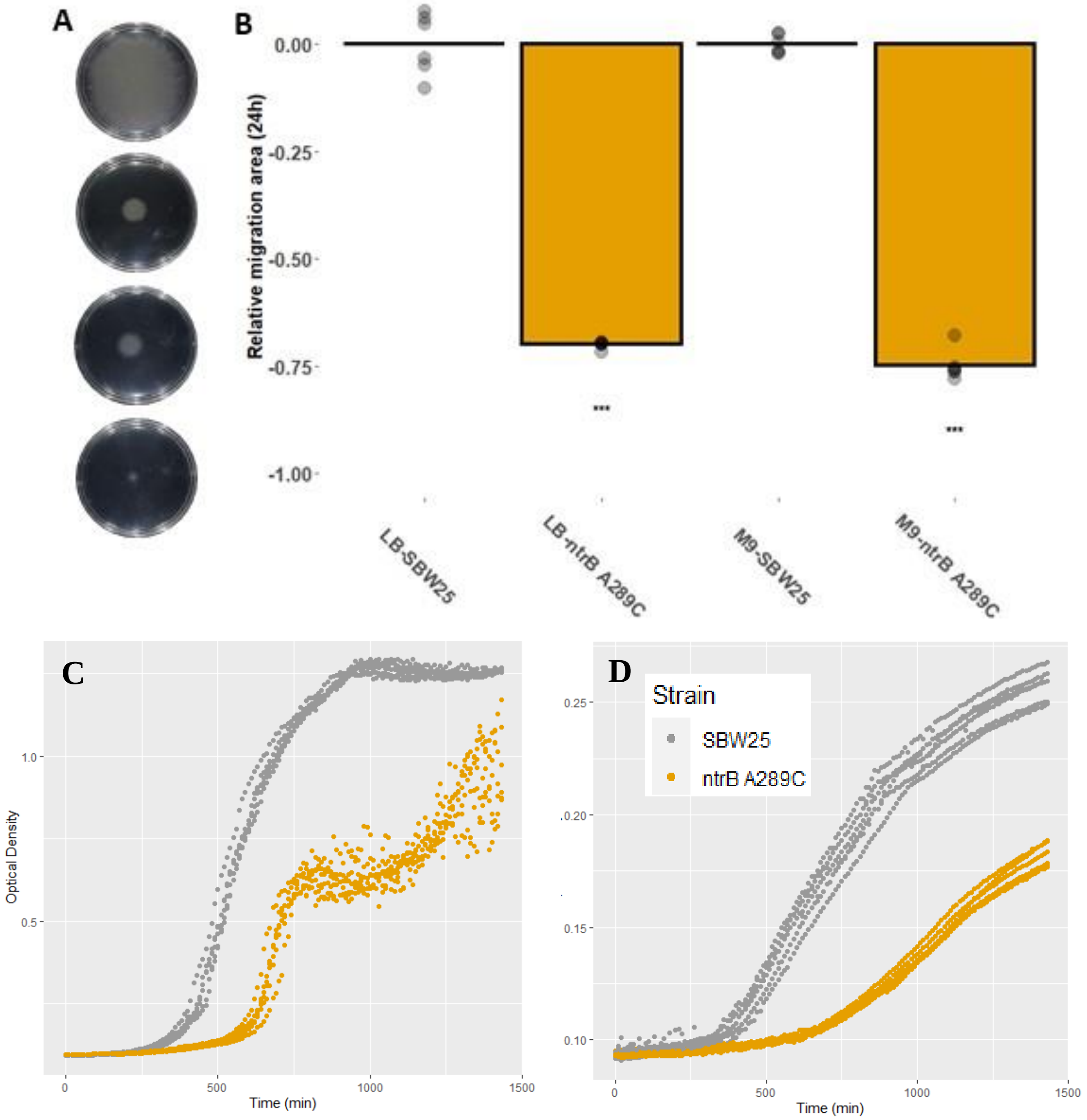


Figure 4.1. *ntrB* A289C is maladaptive in wild-type SBW25 lines under lab conditions. The effects of the *ntrB* A289C mutation on SBW25 fitness was determined by measuring the motility phenotype on soft agar (A-B) and growth yields in agitated liquid culture (C-D), under nutrient rich (lysogeny broth, LB) and nutrient poor (M9 minimal medium) growth conditions. (**A**) On both LB (top two plates) and M9 (lower two plates) 0.25% soft agar, wild-type SBW25 (top, second bottom) covers more surface area than an *ntrB* A289C mutant (second top, bottom) over 24 h. (**B**) This pattern was significant across 6 biological repeats per condition (Wilcoxon rank sum test: M9, W = 0, *P* < 0.005; LB, W = 0, *P* < 0.003). Antagonistic pleiotropy was readily identifiable in agitated liquid culture in both LB (**C**) and M9 (**D**), with the growth yields of SBW25 *ntrB* A289C being significantly lower than in the wild-type

SBW25 strain (Unpaired Two-Samples T-test: LB, t = -49.498, df = 9.2502, *P* < 0.001; M9, t = -22.982, df = 6.3021, *P* < 0.001). Note that the increase in noise observed within SBW25 *ntrB* A289C wells grown in agitated LB culture is owed to the formation of biofilm-like structures that obstruct accurate optical density readings (Supplementary Fig. 4.1).

4.4.2. Nucleotide consensus across the *P. fluorescens* species complex is significantly lower at ‘wobble’ positions around the position 289 hotspot site than elsewhere in the locus

If a mutational hotspot is situationally adaptive, or alternatively, maladaptive, then there should be signatures of selection in its evolutionary history. I searched for this by analysing the rate of evolution of synonymous substitutions surrounding *ntrB* position 289 relative to elsewhere in the *ntrB* locus. To do this I performed a multiple sequence alignment of all the *ntrB* loci, containing 247 *P. fluorescens* strains currently banked on the Pseudomonas database (https://www.pseudomonas.com/) and measured maximum consensus at each nucleotide position, with a lower consensus treated as a proxy for increased divergence across strains. Our null hypothesis is that synonymous substitutions – assuming an equal distribution of neutral or non-neutral fitness effects from these mutation types – should diverge at the same rates across the locus. However if selection has acted differentially upon sites surrounding the 289 hotspot, then we may observe markedly different levels of consensus at these sites.

The first notable observation was that the adenine nucleotide at position 289 (nucleotide position aligned to SBW25 genome) was highly conserved, with 99.6% of all aligned reads possessing an A at 289 (Fig. 4.2A). The sole exception to this was a partial read that registered as a deletion of most of the locus (derived from isolate PS865 (contig41)). Owing to synonymous variation, not all the *P. fluorescens* strains will possess heightened mutability at this nucleotide position and therefore a conserved adenine is not unexpected. However, the absence of any noted mutations at this position synergises with the earlier finding that this mutation is maladaptive in wild type lines with functional FleQ (Fig. 4.1).

I next contrasted the evolutionary divergence of nucleotide substitutions around the hotspot (±15 bp) site versus the rest of the locus. I wrote a simple pipeline to determine the amount of ‘wobble’ for each position i.e. how many substitutions at the nucleotide position would result in a synonymous change (which can be 0 to 3 as an A can change to a T, C, or G etc.) Typically sites with complete wobble were found at positions of the third base in a codon. If selection was not acting to enforce a fixation bias toward any synonymous change, and any nucleotide substitution at a given position was synonymous, then we may expect each of the four nucleotides A, T, C, G to be represented by 25% of the aligned sequences at this position. However, when performing an alignment of a homologous phylogeny of extant species, factors such as the founder effect (Bonneau et al. 2001) and evolutionary time will impact the relative frequencies of bases. So too will genomic mutational biases such as a G:C-bias observed in *P. fluorescens* (these strains are GC-rich, with SBW25 and Pf0-1 possessing genomes of 60.5% and 60.62% GC respectively; Silby et al. 2009), which means that A:T → G:C mutations may be over-represented across the genome. Likewise selection can also suppress synonymous changes regardless of their impact on mutation bias e.g. to maintain optimal mRNA stability (Kudla et al. 2009) and match desirable codon-anticodon ratios (Frumkin et al. 2018). Therefore we do not expect to see a perfect linear relationship between nucleotide consensus and the number of possible synonymous changes at each position. However we should still expect that any relationship observed between nucleotide consensus and wobble will not differ significantly across the locus, if all sites have been able to evolve neutrally.

There were several nucleotide positions around the hotspot site with notably lower consensus than at other positions (consensus mean = 89.1%, mode = 99.6%, range = 48.8 – 99.6%), with the lowest consensus nucleotides (range 48.8 – 75%) able to mutate to any other nucleotide synonymously (Fig. 4.2A). Overall, a clear negative correlation was observed around the hotspot site between the possible number of synonymous changes at each nucleotide position and nucleotide consensus (correlation coefficient = -0.8806612; Kruskal-Wallis chi-squared = 30, df = 15, $P < 0.02$; Fig. 4.2B). This reveals a degree of evolutionary freedom enjoyed by nucleotide positions where mutational events are synonymous versus positions where changes are more likely to be non-synonymous. A similar relationship was observed locus-wide (Kruskal-Wallis chi-squared = 485.17, df = 127, $P < 0.001$; Fig. 4.2B). However, while the inverse relationship between consensus and wobble was comparable between the hotspot site and the locus for nucleotide positions able to undergo 0, 1 or 2 nucleotide changes (Wilcoxon rank sum test with continuity correction, all $P > 0.37$), nucleotide positions with complete wobble (3/3) exhibited significantly lower consensus than other such sites across the locus (Wilcoxon rank sum test with continuity correction, $P < 0.006$; Fig. 4.2B). In addition, 5 of the 6 nucleotide positions associated with hotspot formation are amongst the 7/31 nucleotides with the lowest consensus, and enjoy full wobble (Fig. 4.2A). Overall, these results show a significant reduction in nucleotide consensus when nucleotides are entirely free to mutate synonymously around the 289 hotspot, and the majority of the impacted nucleotides have been incriminated in hotspot formation (Fig. 2.4 and Fig. 2.5). Therefore these results suggest that synonymous evolution at the hotspot has not occurred at the same rates as the rest of the locus, providing a signature of selection at the *ntrB* hotspot in the evolutionary history of this species. However, it is unclear whether the role played by selection was of a positive or purifying nature.

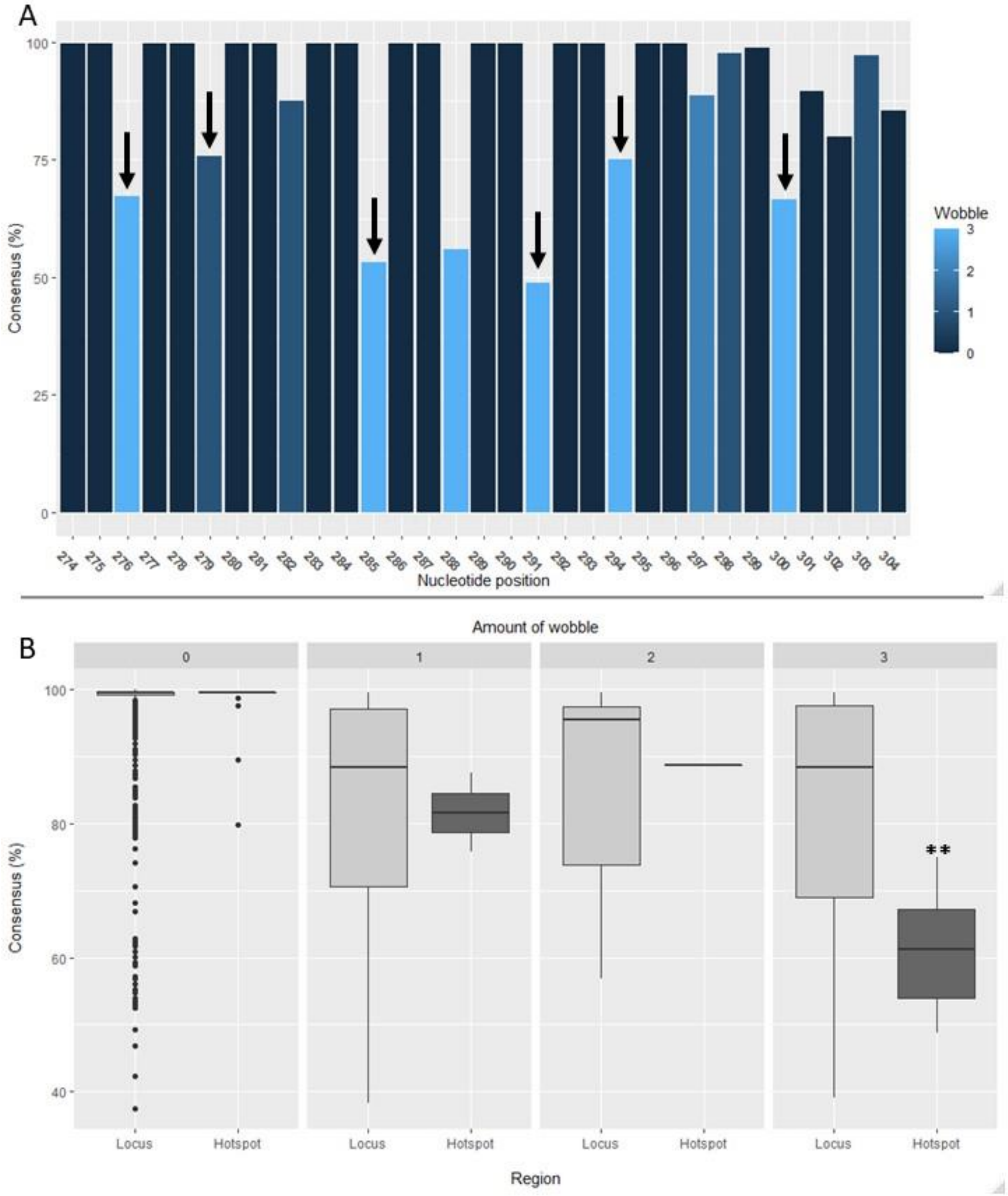


Figure 4.2. Nucleotide positions responsible for hotspot formation exhibit lower nucleotide consensus than elsewhere in the locus. The local sequence around *ntrB* position 289 (±15 bp) was contrasted against the rest of the *ntrB* locus across all 247 *P. fluorescens* strains via a multiple sequence alignment. (**A**) The *ntrB* hotspot region (with mutable nucleotide position 289 at its centre), which is highly mutable in strain SBW25, enjoys a mutation bias that is contingent on the genetic sequence of up to six nucleotide positions: 276, 279, 285, 291, 294 and 300, highlighted by **bold arrows**. Nucleotide

positions, labelled relative to the SBW25 reference genome, are shown on the x-axis. The fill colour denotes how many alternative substitution mutations at a given nucleotide position (0-3) would result in synonymous change (denoted as ‘wobble’), with lighter shades highlighting sites that have potential for increased synonymous diversity. Relative to most other nucleotides in the local region, the six sites responsible for hotspot formation exhibit lower consensus across *P. fluorescens* strains (y-axis). The proportion of strains harbouring each nucleotide variant at the hotspot-defining wobble positions (**bold arrows**) is provided in Supplementary Table 4.1. (**B**) The relationship between nucleotide consensus (y-axis) and wobble (bins) was compared across nucleotide positions around the hotspot site and elsewhere in the *ntrB* locus. The two categories significantly differ when nucleotides enjoy full wobble, with the hotspot nucleotides harbouring much lower consensus across the species-complex (Wilcoxon rank sum test with continuity correction, W = 198, *P* < 0.006).

4.4.3. Harbouring a mutational hotspot offers an evolutionary advantage across genetic backgrounds in non-motile lines

Mutation *ntrB* A289C was found to be maladaptive in the WT SBW25 background, suggesting that directional selection for motility would not drive the fixation of this mutation. Furthermore, sequence analysis revealed that the adenine at position 289 is highly conserved across the *P. fluorescens* species complex, and a signature of selection in the evolution of nucleotides around the hotspot site was found. This suggested that either positive selection or purifying selection may have acted upon the heightened mutagenicity of this genetic change in the organism's evolutionary history, with the latter argument aligning with the highly conserved adenine observed at this position. Therefore the previous results revealed both direct and possibly indirect evidence for suppression of the mutational hotspot in wild type lines. However in the immotile variant AR2, *ntrB* A289C is highly adaptive when under selection for motility, and the hotspot facilitates more rapid realisation of the motility phenotype (Fig. 2.4). Therefore the presence of the mutational hotspot may provide a competitive evolutionary advantage between immotile lines, allowing those with the hotspot to realise motility sooner and thus engage in clonal interference and reach fixation over non-hotspot genotypes in the population pool. To test this hypothesis, I performed competitive evolution experiments between the hotspot-harbouring native AR2 lines and their hotspot-lacking counterparts (AR2-sm), which had been stripped of the mutational hotspot through silent genetic changes (Fig. 2.4). To explore this effect across genetic backgrounds, these competition experiments were likewise performed between a hotspot and non-hotspot immotile variant of *P. fluorescens* Pf0-1, Pf0-2x-sm (Fig. 2.5) and Pf0-2x (Taylor et al. 2015) respectively. Finally, competition experiments were performed between strain backgrounds and hotspot variants so that the evolutionary advantage of a mutational hotspot versus broad genetic background could be investigated.

To perform the assay, immotile strain pairs were mixed in all pair-wise combinations (Fig. 4.3A) and used to seed ≥20 independent replicates of soft agar per pair. Immotile growths were then left for 11 days or until emergent motile zones appeared, at which point the genotype established at the frontier was identified (see materials and methods). The observed evolutionary advantage between strain pairs is shown in Fig. 4.3A. Both hotspot-harbouring mutants (AR2 and Pf0-2x-sm, denoted with $^{HP+}$ hereafter) were observed on the motile frontier in significantly more cases than their hotspot-lacking (AR2-sm and Pf0-2x, denotes with $^{HP-}$ hereafter) competitors (AR2^{HP+}/AR2-sm^{HP-}, 29:5, Bootstrap test: n = 1000000, $P < 0.001$; Pf0-2x-sm^{HP+}/Pf0-2x^{HP-}, 20:3, Bootstrap test: n = 1000000, $P < 0.001$). This reveals a prominent enrichment for hotspot alleles when immotile genotypes are mixed and placed under strong directional selection for motility (Fig. 4.3A). In contrast, no significant advantage was observed across genetic backgrounds when the hotspot sequence was shared (AR2^{HP+}/Pf0-2x-sm^{HP+}, 24:14, Bootstrap test: n = 1000000, $P > 0.07$; AR2-sm^{HP-}/Pf0-2x^{HP-}, 5:4, Bootstrap test: n = 1000000, $P \gg 0.05$).

This finding suggests that in mixed genotype pools, the hotspot allele is more strongly enriched than broad genetic background. However, in pair-wise combinations where both the hotspot allele and genetic background differ (AR2^{HP+}/Pf0-2x^{HP-}, Pf0-2x-sm^{HP+}/AR2-sm^{HP-}/), there is only a significant enrichment of AR2 when competed with Pf0-2x (AR2^{HP+}/Pf0-2x^{HP-}, 30:2, Bootstrap test: n = 1000000, P = < 0.001; Pf0-2x-sm^{HP+}/AR2-sm^{HP-}, 14:6, Bootstrap test: n = 1000000, $P > 0.05$). If the genetic background had no bearing on evolutionary outcomes, then we should not expect to see a difference between these two combinations. However, though there is no significant prevalence of a genotype once evolution has occurred, combinations that include AR2-sm are significantly less evolvable than those that do not include this genotype, as considerably more replicates remain immotile by the end of the experiment (Chi-squared test for given probabilities: χ^2= 20.571, df = 1, $P < 0.001$; Fig. 4.3A). AR2-sm lines have been observed to evolve less rapidly and frequently than the other genotypes assessed in this assay when allowed to evolve independently (Fig. 2.4). However, it is less clear why in mixed immotile populations the competitor genotype is also prevented from evolving.

One answer may be found in dividing AR2-sm cells forming a physical barrier that obstructs emergent motile mutants and stifles their emergence. To test if growth with an 'un-evolvable' immotile line was preventing the emergence of motility in competitor genotypes, Pf0-2x-sm lines were evolved alongside strain AR2 Δ*ntrC*. This genotype has a functional deletion of the *fleQ* homolog *ntrC*, and so cannot evolve motility via mutation within the ntr pathway. As a result approximately 90% of independent replicates of AR2 Δ*ntrC* do not evolve motility following several weeks of selection on soft agar (Shepherd et al., unpublished data not shown). When Pf0-2x-sm was evolved independently, all lines acquired motility within 5 days, whereas only two lines of AR2 Δ*ntrC* evolved within the 11-day experiment (Fig. 4.3B). However, when the two lines were mixed, evolution occurred much less readily (4/9 samples observed within 11 days) than independent Pf0-2x-sm lines (11/11 motile samples observed within 11 days; Wilcoxon rank sum tests with continuity correction: W = 10.5, $P < 0.003$).

This evolvability handicap shows that the notable reduction in competitor evolvability is not a unique trait of the AR2-sm genotype. Rather mixed cultures including lines with reduced evolvability can stifle the evolutionary advantage of possessing a hotspot allele, but only under certain genetic backgrounds. Overall in our experimental setting, when controlling for genetic background, immotile genotypes with hotspot alleles possess an evolutionary advantage that allows them to become enriched over other genotypes. As such I will treat this system as a toy model throughout the rest of this work to assess an evolving genome's ability to maintain a mutational hotspot predicated on repeat regions (Supplementary Fig. 2.2) and head-on collisions between polymerase complexes (Fig. 3.2 and Fig. 3.4).

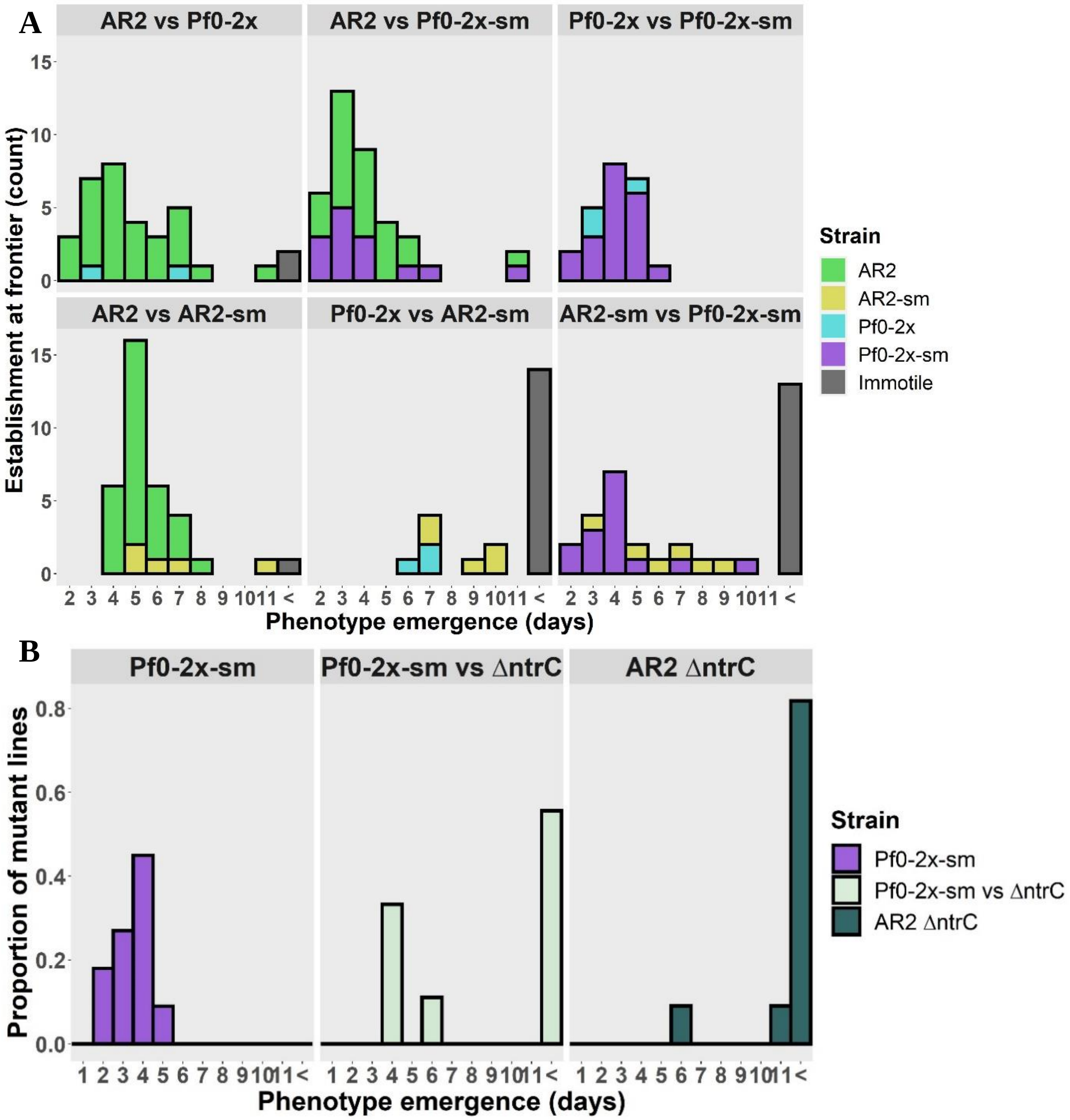


Figure 4.3. A mutational hotspot leading to the biased realisation of *ntrB* A289C provides a competitive evolutionary advantage in immotile mixed populations. (**A**) Competitive pair-wise comparisons between hotspot allele variants and genetic backgrounds. Hotspot allele variants with shared genetic backgrounds (lower left and upper right) show a clear distinction in genotype dominance favouring lines with hotspot alleles (AR2 and Pf0-2x-sm). In contrast, no dominance of evolved motile mutants was observed when hotspot alleles were shared and genetic background differed (centre, top and bottom). When both hotspot allele and genetic background differed, a significant difference in establishment at the frontier was observed between AR2 and Pf0-2x (upper left) whereas no difference was observed between AR2-sm and Pf0-2x-sm (lower right). See main text for statistical analyses. This lack of difference in motile zones was owed to a significant decrease in evolvability which led to more

immotile lines at the end of the experiment (>11 days), which was observed in mixtures that included AR2-sm (bottom row). (**B**) Comparison of Pf0-2x-sm evolvability when evolved independently (left) and when in a mixed culture with AR2 Δ*ntrC* (centre), a strain that largely fails to evolve over the course of the experiment (right).

4.4.4. Maintenance of a mutational hotspot through mutation bias and antagonistic pleiotropy in fluctuating environments

Inherent to the idea of a contingency locus is the idea of contingency – namely that a mutational route is advantageous under certain circumstances. If a mutation were to be universally beneficial to fitness, we would expect the mutation to reach fixation and as such there would be little incentive to maintain heightened mutability, leading to hotspot degradation through purifying selection (Zhang et al. 2018). Therefore for a hotspot to persevere, the two states (the wild type genotype and the frequently realised mutant genotype) must each offer higher fitness than the other under different contexts. This would mean that in hypothetical environment A, directional selection would increase the frequency of one genotype, whereas in environment B the reverse would be true. For example, a natural bacterial life cycle which may encompass such a transition is a pathogenic bacteria colonising a novel host (Moxon et al. 2006). One means of achieving fluctuating genotype fitness in fluctuating environments is by exploiting antagonistic pleiotropy, which is a common by-product of adaptive mutations under strong selection. Antagonistic pleiotropy is readily observable in the model system employed in this study (see Fig 4.5), and yet pleiotropy is overawed by the fitness boon when motility is under strong selection. Remove this directional selection however and the phenotype is no longer advantageous, yet the pleiotropy remains. As such manipulating the environment to reverse directional selection was readily achievable. However, the genetic state-switching between the two genotypes is by no means assured, as compensatory mutations – operating with their own mutational biases and rates – could commit evolving lines further away from the wild-type genotype that facilitates hotspot evolution. I wrote a theoretical model to conceptualise this idea, representing the interplay between mutation bias and antagonistic pleiotropy to maintain or degrade a mutational hotspot.

The general model represents a selection of genotypes arranged in an arbitrary, but symmetrical structure (Fig. 4.4A). The genotypes, represented as nodes, are connected in a mutational network that is visualised so that only genotypes that are adaptive in at least one environmental condition are included, allowing the network to be modest in size and of low dimensionality. Each genotype is connected via edges to their immediate neighbours in the network, representing genotypes that are separated by one mutational event (e.g., a specific single-nucleotide polymorphism). The weights of these edges are determined by an assigned mutational bias (μ) operating in a given mutational direction, either away or toward the genotype. The nodes additionally have an assigned fitness value (λ), which fluctuates through each round of mutation and population re-distribution. Nodes labelled in grey exhibit higher fitness in environment A, and those labelled in blue exhibit higher fitness in environment B (Fig. 4.4A). The structure has therefore been chosen to drive the enrichment of neighbouring nodes when the environment transitions, simulating populations finding compensatory mutations or undergoing genetic reversion. If the population allocated to a new node is re-distributed to a novel node, this is representative of compensatory evolution. In contrast, if a population at a node returns to a previous node, this is representative of genetic reversion. For simplicity, the model broadly assumes that the stronger the fitness value in environment A, the higher the pleiotropy in environment B (see Supplementary Table 4.2 for full list of parameter inputs). Furthermore, there are a number of key assumptions in this model that do not represent biological complexity. These include: *(i)* No role for neutral mutations and as such only epistasis involving adaptive mutations can be captured. *(ii)* A static population carrying capacity in which the population pool is normalised between rounds of mutation. And *(iii)* a simplified deterministic matrix used to adjust relative genotype abundance based on defined fitness and mutation rate values, which enjoy no stochasticity in this framework. Therefore the only adjustable parameters of interest for this model are the fitness values in fluctuating environments, and the relative mutation rates toward or away from each genotype in the network.

These two parameters were adjusted to produce the output displayed in Fig. 4.4B-G. When mutation bias at the hotspot position is high in both mutational directions (95x higher), we observe a stable transition between the two genotypes throughout the ten rounds of mutation and population re-distribution, instigated by shifts in environment. This produces a symmetrical wave-like pattern where the two genotypes (Nodes 1 and 2, representing the wild-type genotype and frequently realised mutation) reciprocally decline to extremely low frequencies and climb to fixation as the environment alternates (Fig. 4.4B). When mutation rate is high only in one direction (from the wild type to the frequently realised mutation) and the other mutation rates are standardised, however, we instead observe degradation of the hotspot. This is visualised by a collapse in the wave-like pattern as the genotypes wane from population dominance to extremely low frequencies as iterations continue (Fig. 4.4C). As expected therefore, removing the mutation bias between Nodes 1 and 2 entirely, results in more rapid hotspot degradation (Fig. 4.4D). In the above simulations, compensatory mutations enjoy equivalent fitness to genetic revertants. However, it is not uncommon in nature and experimental settings for compensatory mutations to offer inferior fitness to the wild-type genotype. When the hotspot nodes simultaneously enjoy shared mutational bias and possess higher fitness than other nodes, we predictably see highly stable hotspot maintenance (Fig. 4.4E). A similar pattern of hotspot maintenance can also be achieved when the mutation bias only operates in one direction, permitted the relative fitness of mutations at the hotspot site are higher than alternative adaptive routes (Fig 4.4F). However when mutation rates are entirely standardised, we still expect hotspot degradation even when compensatory are less fit than genetic revertants (Fig. 4.4G). Therefore the model captures the basic premise explored in this work: that mutation bias, antagonistic pleiotropy, and relative fitness values of adaptive mutations all interact to determine hotspot stability.

Although the network structure is arranged in an arbitrary fashion due to limited knowledge of the potential mutation spectrums from each starting genotype in our model system, we can form basic predictions based on what has been previously observed experimentally. My previous work has shown that the mutation bias from Node 1 to 2 (representing wild-type genotype → *ntrB* A289C) is considerably stronger than other adaptive mutations (representing approximately 95% of all realised mutations; Fig. 2.1). Furthermore, the relative fitness values of Nodes 2-5 (representing *ntrB* A289C and alternate *ntrB* mutations) exhibit similar fitness (Fig. 2.3). Node 6, chosen to represent an alternative lower fitness *glnK* mutant, was in contrast observed to have a lower fitness (Fig. 2.3). However what remained unknown at the outset of this work was the mutational bias operating from Node 2 to Node 1, and the relative fitness values of compensatory mutations. If, as depicted by the model, just 4 compensatory mutations are possible and offer comparable fitness to a genetic revertant, reversion to the original genotype would have to occur at a considerably higher rate to maintain hotspot stability (Fig. 4.4B). Alternatively, the hotspot could be maintained without higher mutation bias in the reverse direction, permitted compensatory mutants offered poorer fitness (Fig. 4.4F). However, if ≥4 compensatory mutations are possible and offer comparable fitness to the wild type genotype in one of the two selective regimes, and there is no mutational bias operating in the reverse direction to the wild-type genotype, hotspot degradation will soon occur (Fig. 4.4G).

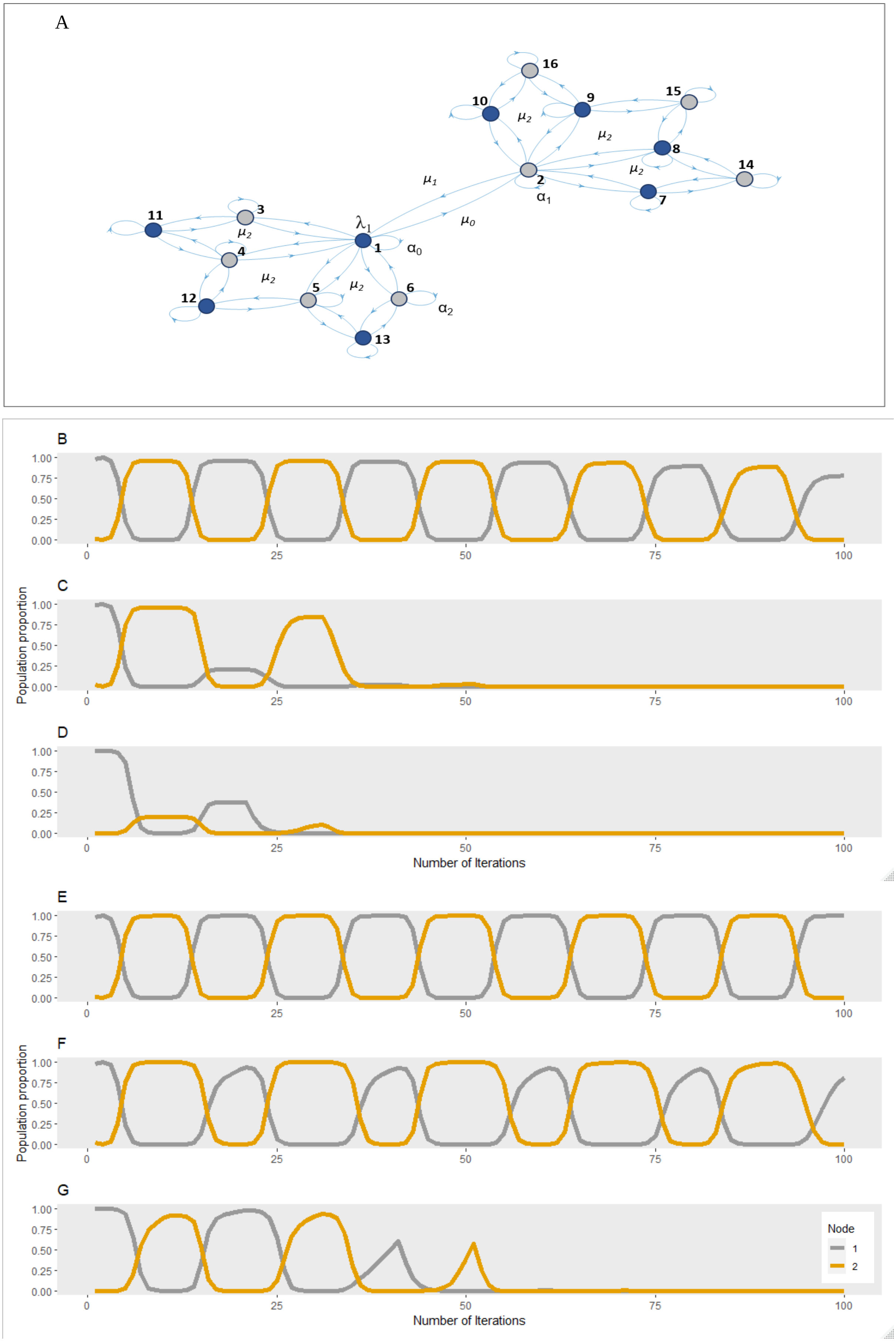

A
16
10
9
15
μ_2
μ_2
8
14
μ_1
2
μ_2
α_1
7
μ_0
λ_1
11
3
μ_2
4
1
α_0
μ_2
12
5
μ_2
6
α_2
13
B
C
D
E
F
G
1.00
0.75
0.50
0.25
0.00
0
25
50
75
100
Population proportion
Number of Iterations
Node
1
2

Figure 4.4. Mutation bias and relative fitness values of compensatory mutants determine hotspot stability in fluctuating environments. (**A**) A theoretical digraph representing a mutational network of genotypes (nodes, 1-16) connected in mutational space via single mutational events (edges). Each mutational event occurs every iteration an assigned mutation rate (*μ_0* for node 1, *μ_1* for node 2, *μ_2* for all other nodes), in which the population assigned at each node is redistributed to neighbouring nodes, or recycled to the same node at proportion α. Following mutation, selection is simulated using as assigned fitness value (λ) to adjust the population proportion represented by a given node (y-axis on B-G). Regular fluctuations in environment were simulated every 10 iterations (x-axis on B-G), by switching all assigned fitness values. Nodes coloured in grey are assigned higher fitness than blue nodes in environment 1 (iterations 0-10, 20-30… 80-90), and blue nodes higher fitness than grey in environment 2 (iterations 10-20, 30-40… 90-100). Only the proportion of the population pool represented at Nodes 1 and 2 are displayed in (B-G), with the entire population beginning the simulation at Node 1. (**B-D**) When higher fitness values are standardised across all nodes ($\lambda_1 = \lambda_2 = \dots \lambda_{16}$): (**B**) Hotspot stability is achieved over 100 iterations (as Nodes 1 and 2 alternate in reaching fixation) when mutation bias exists between Nodes 1 and 2 (*μ_0* and *μ_1* = 0.0095) when rates are higher than rates at other edges (*μ_2* = 0.0001). (**C**) Hotspot degradation occurs when mutation bias occurs at Node 1 (*μ_0* = 0.0095) only, and all other rates are standardised (*μ_2* and *μ_2* = 0.0001). (**D**) More rapid hotspot degradation occurs when all mutation rates are standardised (*μ_0*, *$\mu 1$*, and *μ_1* = 0.0001). (**E-G**) When fitness values at the hotspot nodes (1 and 2) are higher than elsewhere (high fitness of λ_1 and λ_2 = 0.9, high fitness of remaining nodes = 0.6): (**E**) As in (B) but with higher hotspot fitness, stability is achieved. (**F**) As in (C) but with higher hotspot fitness, stability is now achieved. (**G**) As in (D) but with higher hotspot fitness, degradation continues to occur. See Supplementary table 1 for full details on adjusted parameter values.

4.4.5. AR2 *ntrB* A289C mutants undergo compensatory mutation, not genetic reversion, in conditions in which the mutation is deleterious

In order to test for a mutation bias operating in the reverse mutation direction at the hotspot site (i.e. *ntrB* C289A), I evolved AR2 lines harbouring the *ntrB* A289C mutation in shaking M9 culture. In this condition, the antagonistic pleiotropy of *ntrB* A289C renders the mutation maladaptive relative to the ancestral strain (as demonstrated in wild-type SBW25 lines in Fig. 4.1B), and so lines evolve under directional selection for genetic reversion or compensatory mutation to restore wild-type fitness. As the intention of this assay was to capture the first compensatory mutant or genetic revertant to appear in each independent line, evolving cultures were serially diluted daily to approximately 100 colony forming units and plated. Colonies that had undergone mutation to restore wild type fitness were readily identifiable through superior colony growth (Supplementary Fig. 4.3). As approximately 100 colonies were plated during each daily dilution (see materials and methods), the assay was designed to identify the first evolved genotype that had reached at approximately 1% frequency in the population. All 20 independent cultures, seeded with a single clone of *ntrB* A289C, evolved within 144 h (Fig. 4.5A). Additionally, all evolved lines achieved phenotypic reversion, as each line restored at least equivalent fitness to the ancestral AR2 line in shaking culture (Fig. 4.5B-C).

I subsequently performed Sanger sequencing on the *ntrB* locus to search for genetic reversion at position 289. All 20 replicates reported a cytosine at this position, revealing that genetic reversion had not been the first adaptive solution to be realised in any line. However, sequencing of the *ntrB* locus also revealed a plethora of compensatory mutations, which has occurred elsewhere within the locus (Fig. 4.5D). Among these were ΔA67 found in the PAS domain, T485C in the phospho-acceptor domain, and G958A found in the C-terminal domain, representing a substantial compensatory mutational target size (Fig. 4.5D). Overall, compensatory mutants were found within the *ntrB* locus in 50% of independent replicates (10/20; Fig 4.5D). In contrast to over-expressing the nitrogen response regulator *ntrC*, removal of *ntrC* activity engenders no cost to fitness in M9 minimal medium supplemented with ammonia (Supplementary Fig. 4.4). As removing functionality of the nitrogen regulatory pathway is not deleterious under the experimental conditions used, the observed compensatory mutations within *ntrB* may operate by simply disrupting kinase activity. If so, the functional consequence of removing *ntrC* over-activity matches the observed phenotypes of compensatory mutants, all of which were rendered immotile following evolution. In addition, when the environment was fluctuated once more and lines were placed under selection for motility, overall the mixed batch of compensatory genotypes evolved motility significantly more slowly than AR2 lines, with several lines not achieving motility over the 11-day experiment (Fig. 4.5A). These assays therefore show that introducing an artificial bottleneck that selects for the first adaptive genotypes to appear leads to rapid degradation of the mutational hotspot.

These assays revealed that the likelihood of observing genetic reversion is reduced owing to the array of compensatory mutations that can restore wild type fitness. However, a contingency locus does not need to reach fixation in order to be maintained in the population, and instead may persist at low frequencies in the genotype pool. As such I next performed an additional assay which involved mixed-population deep-sequencing of the evolved genomes following evolution in agitated LB culture. This allowed me to see if revertant lines could appear and persist in the population at low frequency.

Using the same colony size-based identification of phenotypic revertants (see materials and methods) I observed that compensatory mutants or genetic revertants had dominated the population pool by 96 h in all five populations. This rapid evolution of compensatory mutants is likely owed to the increased growth rate of *P. fluorescens* lines in LB over M9 minimal media (see Fig. 4.1C-D). All 5 independent mixed populations were analysed by whole-genome sequencing, with mean coverage of the genotype pools ranging from 105.6 to 232.9 (Supplementary Table 1). I first searched for evidence of genetic reversion, and observed that a cytosine was found at ntrB position 289 in 100% of reads in 4/5 populations. The remaining population reported a single read that did not possess a cytosine at site 289 (1/103), however the resolution of the assay was not high enough to separate this call from read noise, which typically contained an error of 1-2 reads per genomic position (data not shown). For any mutation to be called with a degree of confidence it was determined that it must be present in 5% of reads (assuming an average of 2 errors per 100 reads, the chance of observing 5 errors: Bootstrap test: n = 1000000, $P < 0.026$). However the pipeline analysis utilised could not identify any compensatory mutations that occurred in under 20% of reads for each respective pool (see materials and methods) and no compensatory mutations were called, inferring that multiple adaptive genotypes were realised throughout the assay as none approached fixation or population dominance. Subsequent manual analysis of reads within the *ntrB* locus revealed two compensatory mutations occurring in the same population (*ntrB* ΔG375903-T375917 within ≥9/132 reads (6.8%), and *ntrB* Δ376578 within 16/132 reads (12.1%). Overall these results reveal that clonal interference between competing adaptive mutants can prevent hard sweeps to fixation by a single genotype over the short course of the experiment. However we observed no evidence for genetic reversion mutants competing within these mixed adaptive pools.

Compensatory mutations within *ntrB* had occurred in both the single-cell genotypes and mixed genotype-pool sequences. As all compensatory lines restored wild-type fitness levels (Fig 4.5B-C), lost the motility phenotype, and made subsequent evolution of motility occur less readily (Fig 4.5A), it is likely that in many instances the compensatory mutations broke functionality of the ntr pathway. However, the viable number of compensatory mutations have been observed to alter by environment (Basra et al. 2018). In this case of this experiment, compensatory mutations that debilitate nitrogen regulatory network function carry no consequences to fitness in either LB, or M9 minimal media supplemented with ammonia (Supplementary Fig. 4.4). However the nutrient condition can be manipulated so that breaking the ntr system carries severe pleiotropy, by providing lysine as the sole nitrogen source (Supplementary Fig. 4.4; Zhang and Rainey 2008). Therefore I was able to manipulate the number of viable adaptive routes, and negate functionally-debilitating mutations within *ntrB* as compensatory mutations, by substituting ammonia with lysine in the minimal media. The hypothesis of this assay was to restrict the mutational target size of compensatory mutations and so increase the likelihood of observing genetic reversion. In an echo of the original reversion assay, single adaptive colony forming units were evolved in shaking liquid culture of M9 minimal media with lysine, and their *ntrB* loci assessed by Sanger sequencing. Although 0/8 lines reported compensatory mutations within the *ntrB* locus, 8/8 reported a cytosine at site 289. As such even by manipulating the environmental conditions to prevent common compensatory mutations being realised, sufficient compensatory mutations remain viable that genetic reversion remains rare.

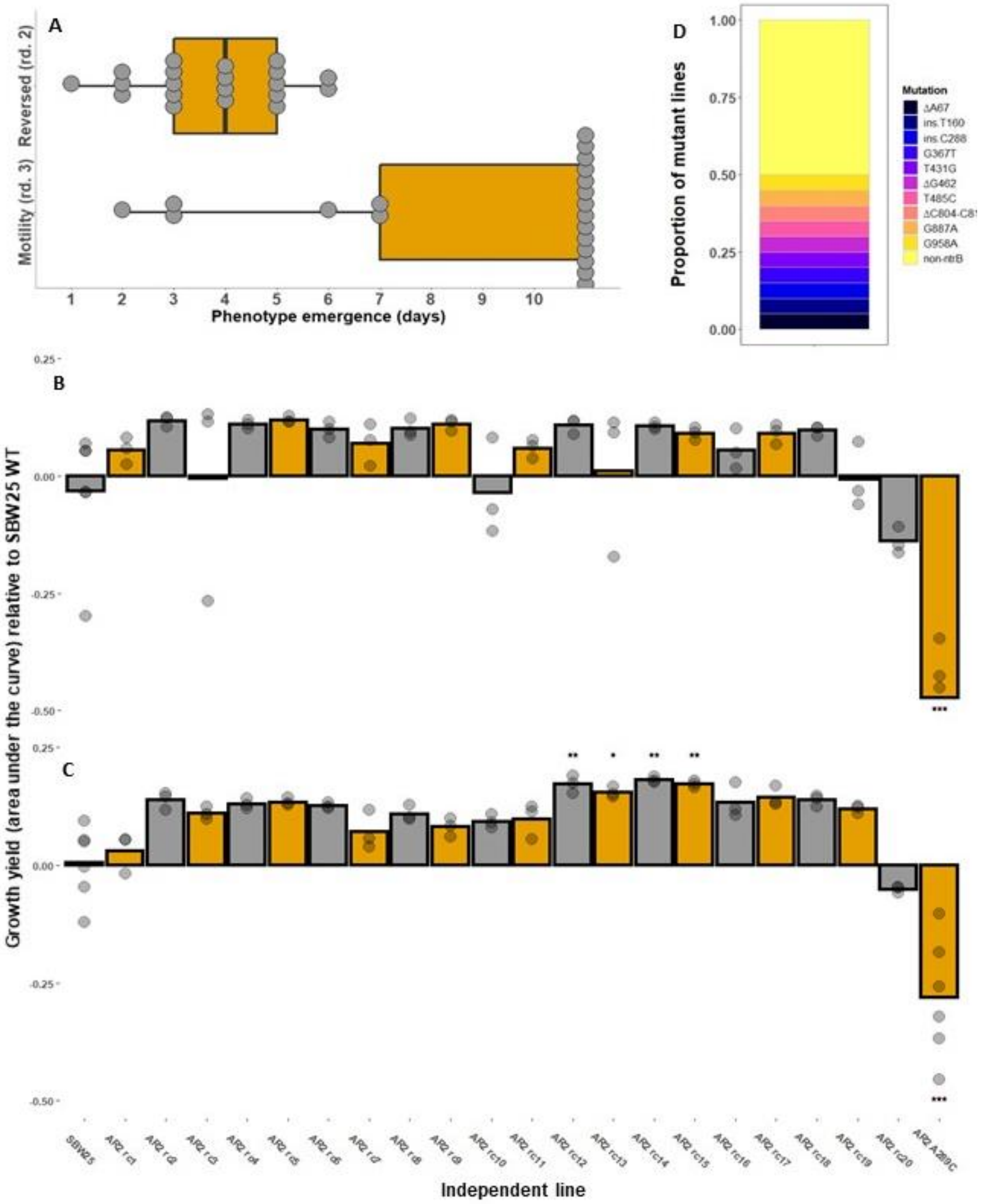


Figure 4.5. AR2 *ntrB* A289C lines evolved under directional selection to mitigate pleiotropy undergo rapid compensatory mutation and restore wild-type level fitness. (**A**) 20 independent replicates of clonal AR2 *ntrB* A289C evolved in agitated M9 minimal medium broth recovered the wild-type growth phenotype within 1-6 days (Reversed rd. 2)). When the mixed batch of recovered lines were subsequently placed under directional selection for motility, however, only 6/20 samples evolved motility over the course of the 10-day experiment (Motility rd. 3)). The recovered lines restored wild-type level fitness in both M9 (**B**) and LB (**C**) agitated liquid culture as determined by growth yield (see materials and methods), with several lines exhibiting improved yields over the wild-type in LB (one-

way ANOVA post-hoc Tukey HSD test relative to SBW25: Significance values: * = $P < 0.05$, ** = $P < 0.01$). The high range in recorded data points observed in AR2 *ntrB* A289C is discussed in Supplementary Fig. 4.1. (**D**) The *ntrB* loci of all independent replicates that had recovered wild-type fitness ($n = 20$) were sequenced through Sanger sequencing, identifying 10 compensatory mutations each unique to an independent line. All 20 lines reported the mutation A289C, and 10 replicates reported no other mutations with the *ntrB* locus.

4.5. Discussion

In this work I investigated whether the mutational hotspot at ntrB A289C could function as a contingency locus as a means to determine if the hotspot's presence derived from a selectively-enforced origin. I found no evidence to support that *ntrB* A289C was adaptive in wild-type lines, but engineered immotile variants that possessed hotspot alleles did enjoy an evolutionary advantage over non-hotspot genotypes in mixed populations. As such in artificially constructed lines and under lab conditions, I could demonstrate the selective advantage for possessing the mutational hotspot thanks to the mutation bias conferring rapid access to adaptive mutations. Furthermore, using comparative sequence analysis of the *ntrB* locus belonging to each member of *P. fluorescens* species complex, I found signatures of selection which may have been indicative of a contingency locus or purifying selection of the mutational hotspot in the evolutionary history of the species. I next designed a simple deterministic mathematical model that showed for this hotspot to be maintained under fluctuating environments, genetic reversion at site 289 must be able to compete with compensatory mutations. This would require either a heightened mutation bias operating at site 289 which converted the genotype to the wild-type state (C289A), or for the mutational target size of compensatory mutations to be low and for these compensatory mutations to offer inferior fitness. However when evolved under fluctuated selective regimes all assays showcased an array of highly fit compensatory mutations and revealed no evidence for genetic reversion. This remained true when isolating the first compensatory mutants to appear, when performing mixed-population deep sequencing to identify adaptive mutations persisting at lower frequencies, and when augmenting the environment to amend pleiotropy of certain adaptive routes. Therefore experimental work revealed no evidence for a selectively-enforced evolutionary origin. As the mutational hotspot within the locus can be assembled and dismantled by very few synonymous changes, together these results suggest that rather than positive selection maintaining a mutational hotspot, these regions may appear through drift and subsequently be worked upon by purifying selection.

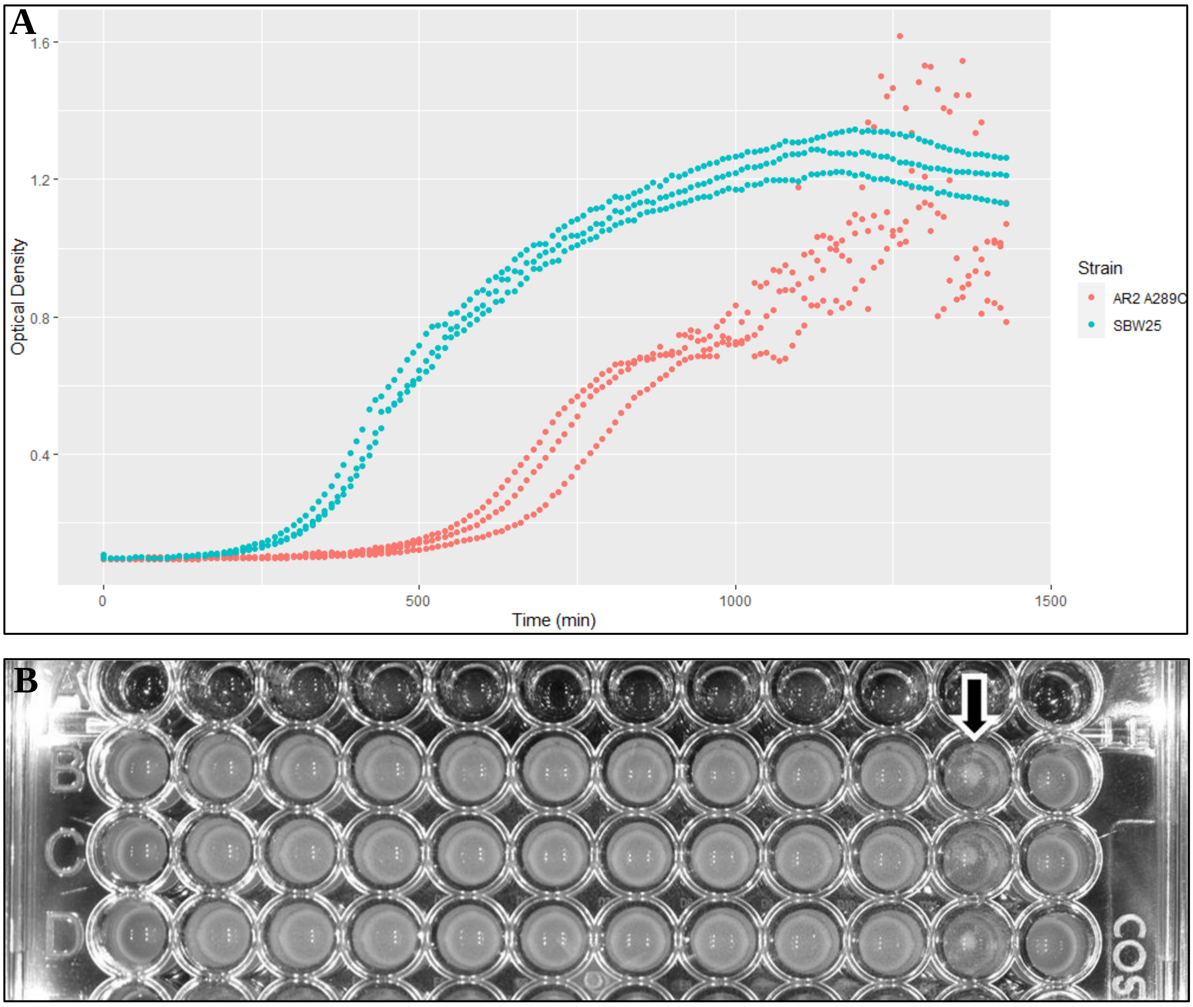
The finding of frequent compensatory mutations occurring within the *ntrB* locus matches the hotspot mechanism identified previously (Fig. 3.4). Although the hotspot site is believed to be contingent on repeat regions forming stem-loop structures (Supplementary Fig. 2.2), mutation at the site and elsewhere in the locus is synergistically heightened by head-on collisions between RNA polymerase and the replication fork, which is maintained even in hotspot-lacking strains (Fig. 3.2). As such the mechanism that ensures highly parallel evolution at site 289 when populations are under directional selection for motility is simultaneously responsible for biasing the mutational spectrum to favour compensatory mutations within the *ntrB* locus. This is because mutation bias remains high across the locus but the stem-loop structure, which has likely been converted to become more stable (De Boer and Ripley 1984), no longer offers bias unique to the hotspot site. Contingency loci predicated on strand-slippage offer heightened mutation rates in both directions (Koch 2004), but in the case of this hotspot the mutational bias enjoyed by genetic revertants is shared by rival compensatory mutations. This means that the mutational bias at site 289 is negated, and thus the outcome is degradation of the hotspot through both the loss of a highly mutable site, and further functionally disrupting mutations within the locus.

The theoretical model employed in this work was able to showcase key variables necessary for hotspot stability. However, it is unknown whether such a simple deterministic network model can form any accurate predictions and aid the researcher in elucidating the influence of unknown variables (e.g. mutation bias). Future work involving such a model should seek to expand upon the included parameter set. For example, the model employed in this work simulated regularly alternating selective regimes, yet contingency loci have been suggested to also evolve under infrequent fluctuating environments (Moxon et al. 2006). Replacing unknown features within the network structure, such as the number of nodes and their connectivity within the genetic network, with experimentally derived data from this work and future research will help to establish the current limitations of this model.

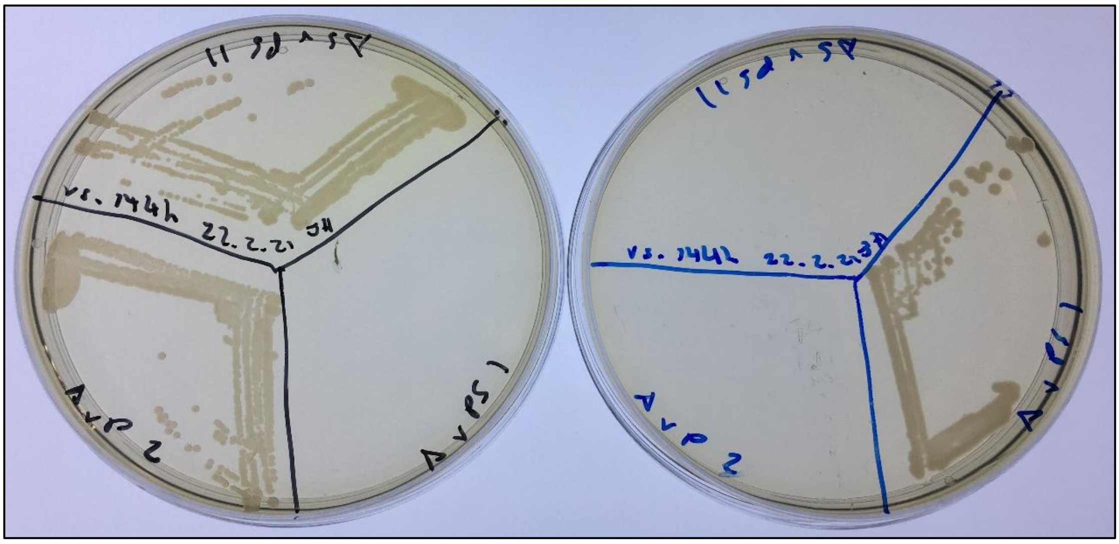
One caveat of the conclusions drawn is the assumption that mutation bias in the reverse direction i.e. from *ntrB* A289C → *ntrB* C289A must exist to a degree in order for genetic revertants to compete with the numerous highly fit observed compensatory mutants. During the reversion assay, populations were placed under very strong directional selection away from the motile genotype, which drove the rapid realisation of emergent compensatory mutations and thus provided limited generational opportunity for genetic reversion. In very large populations and in complex environmental niches, genetic revertants may have more opportunity to exist at low frequency than in the experimental context explored here. However, as the motile genotype inherently encompasses severe pleiotropy, directional selection away from the genotype in these experiments is likely not unfairly extreme. In addition, when nutrient conditions were manipulated to negate compensatory mutations occurring within *ntrB* and disrupting its function, compensatory mutation continued. The harshness of the employed assays versus more natural settings comes in the form of increased opportunity for clonal interference, which may have suppressed rare genetic revertants from persisting in the population pool. To alleviate this effect, an alternative method may have been employed that involved daily passage of single colony forming units (CFU) onto fresh solid media. By separating single CFU's across the agar surface through streaking, the opportunity for clonal interference would have been reduced, and the chance of observing genetic revertants may have increased. However, as our assays revealed that the mutational target size for compensatory mutations is considerable, the stability of the *ntrB* mutational hotspot without mutation bias in both directions is unlikely.

Mutational hotspots are powerful evolutionary agents, able to repeatedly drive the same genetic solutions across independent cell lines. Yet to fully appreciate the role they play in evolution we must understand their evolutionary origins. Generally, there seems to be evidence that structures which are likely to cause mutation bias via imperfect inverse repeat regions are reduced across genomes as following mutation they are immortalised as stable perfect palindromes (Lavi et al. 2018). However, in some instances hotspots are preserved by selection by operating as ‘contingency loci’, offering a mutational means for populations to cope with environmental change (Moxon et al. 1994). Other times hotspots may occur through happenstance, appearing through neutral evolution at sites under relaxed selection and rising in frequency through genetic drift, or existing only transiently as their deleterious mutability means they will be suppressed by purifying selection. Understanding the evolutionary origins of hotspots is important as we look toward finding them in bacterial genomes as a means to make evolutionary forecasts. A selectively enforced origin guides us to identify hotspots using an organism’s known evolutionary history; a neutral origin encourages us to scour the breadth of the genome for innocuous genetic variation that may lead to drastically different evolutionary destinies. This work suggests that a mutational hotspot predicated on repeat DNA regions (Supplementary Fig. 2.2), genome location and strand orientation (Fig. 3.2) and mismatch repair complex efficacy (Fig. 3.3); which engenders considerable phenotypic change (Fig. 2.1 and Supplementary Fig. 2.1); and is only present in a subset of *P. fluorescens* lines due to synonymous variation (Fig. 2.4 and Fig. 2.5), has likely not been maintained by positive selection. Therefore this work highlights that we should not only expect hotspots to play an active role in evolution at sites that are under selection in an organism’s evolutionary history, but that they may play a role in defining genetic trajectories across the breadth of the genome.

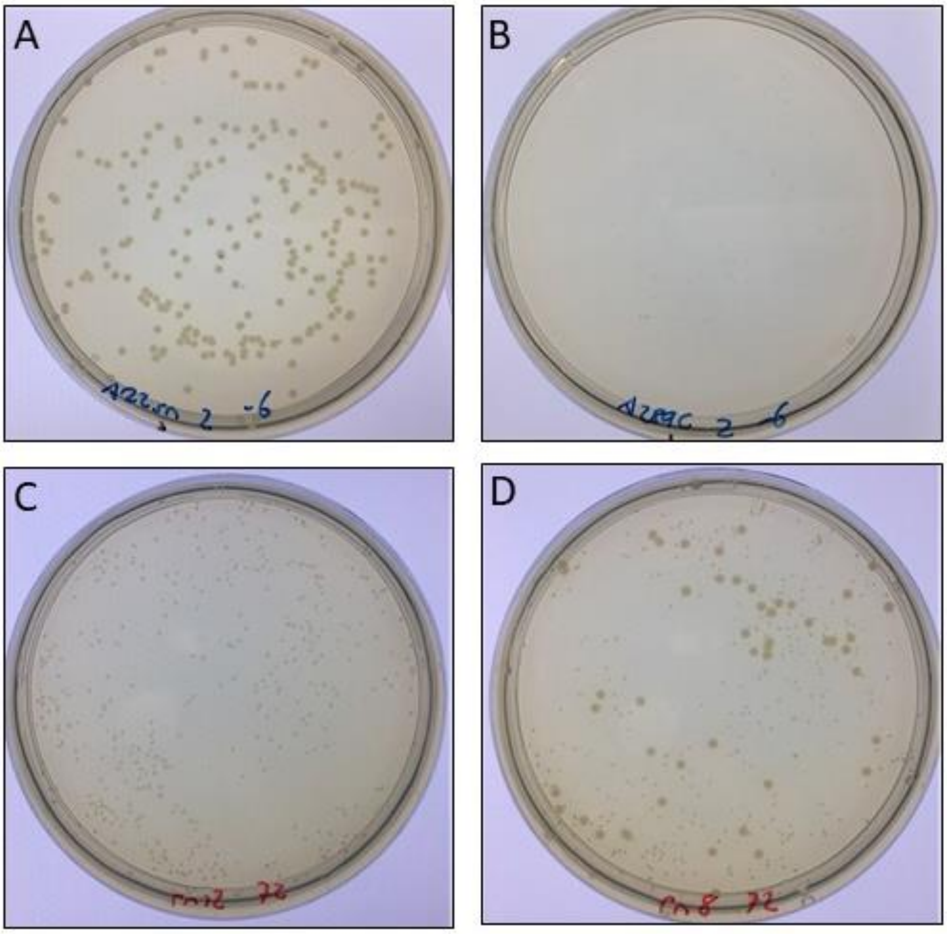
4.6. Supplementary materials



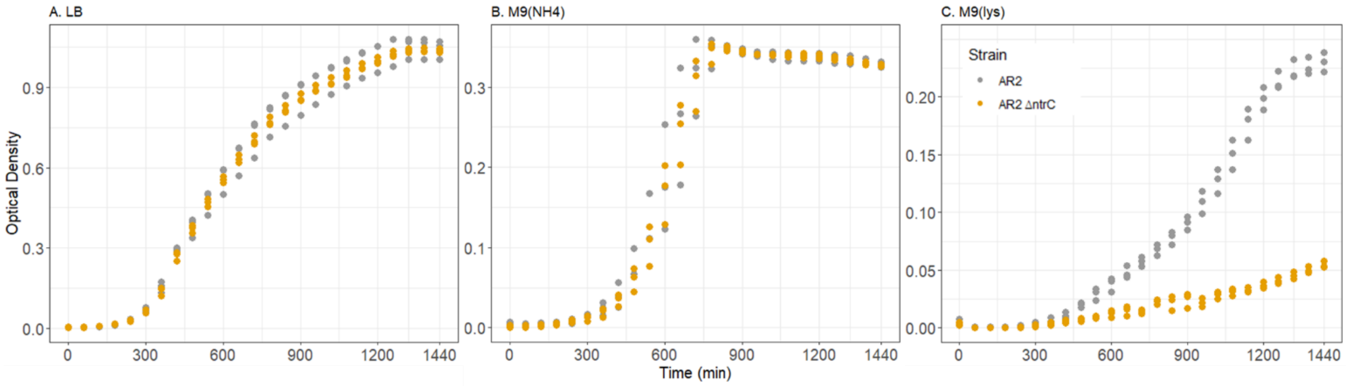
Supplementary Figure 4.1. Line growth curves of wild-type AR2 and AR2 *ntrB* A289C in agitated LB culture reveal increased noise in optical density recordings owed to biofilm-like structure formation. (**A**) Optical density readings over time (used as input data for growth yields for Fig. 4.5) show increased noise in AR2 *ntrB* A289C lines after approximately 1000 minutes, inflating optical density readings and increasing data point range of independent replicates. (**B**) This is owed to biofilm-like structures that form in the well during the assay (biological triplicates highlighted by an arrow), which are absent in the wild-type ancestral and compensatory mutant lines.



Supplementary Figure 4.2. Motile zones emerging from mixed populations derived from strains AR2 and Pf0-2x can be readily distinguished using antibiotic selection. A plate containing 100 µg/ml kanamycin sulphate (left) inhibits growth of Pf0-2x-derived strains, while a plate containing 250 µg/ml streptomycin sulphate (right) inhibits growth of AR2-derived strains. The assay is only considered reliable if clear growth is observed on one plate and the complete absence of growth is observed on the other, as shown here.



Supplementary Figure 4.3. Phenotypic revertants can be readily identified through colony size. (**A**) Serially-diluted and plated strains with wild-type fitness derived from liquid culture (AR2-sm used for the example above) grow to large colonies after 48 h incubation at 27°C. (**B**) Owing to the pleiotropic effect of the mutation, *ntrB* A289C mutants grown over the same incubation period achieve much smaller colony sizes. (**C-D**) Dilutions of independent cultures grown over a period of 72 h under directional selection for reversion. (**C**) A replicate where phenotypic reversion has not yet appeared and risen to approximately 1% population frequency (equivalent to 1/100 colony forming units, of which a higher number has been plated). (**D**) A replicate where phenotypic reversion has been realised and revertants have risen to a high frequency in the population (larger colonies represent approximately 30% of the plated population).



Supplementary Figure 4.4. Removal of nitrogen regulatory pathway function through deletion of the nitrogen response regulator *ntrC* does not impact growth rate in (**A**) LB, or (**B**) M9 minimal medium supplemented with ammonia (NH_4), but does so in (**C**) M9 supplemented with lysine (lys) as the sole nitrogen source. Independent biological triplicates of each condition were grown in agitated liquid culture (180 rpm) with optical density readings performed every 60 minutes for a total period of 24 h. The pleiotropy observed in M9 supplemented with lysine matches observations found in a previous study (Zhang and Rainey 2008).

Supplementary Table 4.1. Nucleotide distributions at wobble sites surrounding *ntrB* position 289.

Nucleotide:	276	279	285	291	294	300
A	0	0	3	23	42	1
T	4	59	1	9	8	2
C	167	188	111	94	11	79
G	76	0	132	121	186	165

Supplementary Table 4.2. Adjusted model parameters and their effect on hotspot stability.

Hotspot parameters	Mutation rates (μ)			Fitness (λ) Array set 1					Fitness (λ) Array set 2					Hotspot stability
	μ_0	μ_1	μ_2	λ_1	λ_2	λ_{3-6}	λ_{7-13}	λ_{14-16}	λ_1	λ_2	λ_{3-6}	λ_{7-13}	λ_{14-16}	
High mutation rate (μ_0 and μ_1)	0.0095	0.0095	0.0001	0.1	0.9	0.9	0.1	0.9	0.9	0.1	0.1	0.9	0.3	Near stable fixation
High mutation rate (μ_0 only)	0.0095	0.0001	0.0001	0.1	0.9	0.9	0.1	0.9	0.9	0.1	0.1	0.9	0.3	Degredation
Even mutation rate ($\mu_0 = \mu_1 = \mu_2$)	0.0001	0.0001	0.0001	0.1	0.9	0.9	0.1	0.9	0.9	0.1	0.1	0.9	0.3	Degredation
High μ_{0-1} and high fitness ($\lambda_{1-2} = 0.9$, $\lambda_{3-16} = 0.6$)	0.0095	0.0095	0.0001	0.1	0.9	0.6	0.1	0.6	0.9	0.1	0.1	0.6	0.3	Stable fixation
High μ_0 and high fitness	0.0095	0.0001	0.0001	0.1	0.9	0.6	0.1	0.6	0.9	0.1	0.1	0.6	0.3	Stable fixation
Even mutation rate and high fitness	0.0001	0.0001	0.0001	0.1	0.9	0.6	0.1	0.6	0.9	0.1	0.1	0.6	0.3	Degredation

Chapter V

5.1. Predicting evolution: Feasible, foolhardy, or futile?

"[Chaos theory:] When the present determines the future, but the approximate present does not approximately determine the future." – Edward Lorenz

Many domains of science wrestle with forces that are deterministic and others that are indeterminate – those that are predicated on chance. Isaac Newton's equations describing gravity were written as infallible, able to predict a gravitational relationship with pin-point accuracy. Like his work on thermodynamics, there was a deterministic flow to Newton's view of the universe, one of cause and effect. With a complete knowledge of the present therefore, a deterministic system provides an immensely powerful opportunity: the ability to make accurate predictions of the future. However, it is not a trivial task to understand all forces acting on a dynamic system in a given moment. As such a predictive system often falls prey to chaos – where uncertainty in the starting dynamics make accurate predictions of the future impossible. Therefore the issue in precise forecasts is often not inherent randomness, but in our own uncertainty of the starting conditions which leads to error. This is as true for the orbits of three celestial bodies (Krishnaswami and Senapati 2019) as it is true for the mutation and evolution of an organism's genome.

Biologists looking to predict genetic evolution must also deal with fundamental forces which, in many contexts, are simply indeterminate. Such random events are mutations, the basis of genetic variation, and the fate of such a novel mutation in the population. In 1968 Motoo Kimura authored a landmark study that illuminated the neutral theory as a prominent force in evolution (Kimura 1968). Kimura revealed that most mutations that persist in a population are owed to genetic drift – ultimately stating that they're fixed in a population by chance. The result is that evolution often operates at the behest of chance events that generate variation, and chance events that ensure their persistence. Under such circumstances predicting evolution is, in all likelihood, impossible.

But not all instances of evolution are so reliant on randomness. Many populations evolving in nature begin as clones and evolve under very strong directional selection. This is true of the first cell in a tissue that becomes immortal and forms the nexus of a tumour (Greaves and Maley 2012), a foundling group of pathogenic bacteria or viruses finding a new home in a patient's lung or gut (Koornhof et al. 2001), and a microbe that has endured a period of stress that killed other variants of its species (Couce et al. 2016). These extreme population bottlenecks are immensely useful for an evolutionary biologist looking to forecast adaptation, as they simplify the starting genetics to something wholly knowable.

Many of the clonal populations introduced above subsequently evolve under harsh directional selection. This is another powerful asset to those looking to forecast evolution. Harsh directional selection purges deleterious mutations out of a population's collective gene pool – a process known as purifying selection (Cvijovic et al. 2018) – and drives beneficial mutations to fixation. Therefore in stark contrast to a neutrally evolving population, harsh directional selection limits the number of mutations that have the potential to persist to a much smaller figure. Under certain selective regimes there may be only very few mutations that will benefit the evolving population (Weinreich et al. 2006).

To refine things yet further, there are instances where the adaptive potential of the population is very large i.e. when the population starts at very low fitness. In these cases, a single mutational event can create a huge disparity in fitness between the adapted mutant and others in the gene pool. This advantage allows an emergent mutant to make an adaptive "leap" as it very rapidly becomes distinct from its ancestor. This process is known as saltation and can allow such a mutant to swiftly dominate a gene pool (Theißen 2009). Therefore, in these cases we can begin a round of adaptation with a clonal population and end the round with another clonal population (when adhering to the strong selection, weak mutation model, see (Gillespie 1984)). The first clonal population is the progeny of a single ancestor, and the resulting population is the progeny of a sole or very small number of its adapted descendants. Members of the population that had fixed random neutral mutations, creating genetic diversity in the gene pool, will have been mostly purged by those who have acquired these adaptive mutations, although some may hitchhike to fixation (Denver et al. 2010).

Beginning with a clonal population and utilising strong directional selection can help cleave the number of observable mutations in adapting lines from many to very few indeed. Already this offers a great deal of hope for those aiming to forecast evolutionary outcomes. But which of these adaptive mutations will be the first to appear? As mutation is reliant on chance errors or recombination events, it is immensely difficult to guarantee the appearance of a particular genetic change. However, we can make stronger assertions about how likely a given change is to occur relative to other genetic changes in the genome.

The ability to calculate relative mutation risk for different sites in the genome is owed in part to various pieces of molecular machinery that introduce and repair errors at different rates depending on the local genetic context. This phenomenon is known as mutation rate heterogeneity, a phenomenon that can drive certain sites in the genome to be considerably more mutable than elsewhere. In extremely rare instances, such a highly mutable site can be found at a genetic position that, once mutated, offers a significant boon in fitness. These occurrences create what I refer to throughout this work as mutational hotspot: a hub for adaptation that results in a highly repeatable evolutionary event across independent replicates. Mutational hotspots can operate with such magnitude that evolution ceases to become unpredictable, but instead becomes deterministic with a high degree of confidence. Such a hotspot has formed the nexus of this thesis.

The model system investigated throughout this work, with the hotspot at its centre, has helped us acquire comprehensive knowledge of our starting conditions that allow for accurate forecasts of evolution. I have demonstrated the power of mutation bias over other key evolutionary variables in enforcing repeatable evolution. I've then shown in genetic detail the interplay of factors that create this bias.

Additionally, I’ve shown that such hotspots are not slaves to selectively-driven historical contingency, as the collected evidence suggests that these hotspots transiently appear before being suppressed by purifying selection. Armed with this knowledge, this work helps us shift toward certainty and away from chaos, by revealing that predicting evolution with high confidence is possible even before mutation has occurred. As such predicting evolution is certainly feasible, albeit perhaps foolhardy, but certainly not futile.

5.2. Future work: Toward a predictable model of evolution

"It's a subtle and powerful thing, prescience. The future becomes now." – Frank Herbert, Children of Dune

The findings of chapters 2 and 3 showcase both that evolution can be highly repeatable to nucleotide resolution, as well as showing the genetic arsenal needed to make it happen. As such these findings will move us closer to forming accurate evolutionary forecasts from genomic sequence data. By factoring in and filtering the combinatorial impact of the genetic features highlighted throughout this work, there is ample potential to write a pipeline that identifies mutational hotspots prior to any directed evolution experiments. One exciting test of this would be to generate theoretical forecasts using a pipeline that derives predictions based on an ancestral genome, and testing its efficacy in predicting adaptive mutations that have reached fixation in the descendant's evolved genome. This would be analogous to a popular machine learning approach involving algorithm training followed by testing using a similar data set. The primary advantage of this latter approach is that the required prior knowledge for generating forecasts from genetic features would be minimal. However, it can be difficult to prise apart selection, fixation, and other evolutionary forces such as the founder effect from the role played by mutation bias in enforcing adaptive outcomes from descendent sequence data. As such, taking a holistic approach to making evolutionary forecasts would be susceptible to capturing significant levels of noise. The alternative approach, guided by carefully curated experimental data, allows for a refined pipeline that focusses on mutation bias and presents the opportunity for highly accurate evolutionary forecasts.

Several of the identified features are already a trivial act to identify. For many bacterial genomes, Nanopore-based sequence breadth coupled with the Illumina-based sequencing depth allows for the annotation of resolved genomic data with high fidelity. Such annotations capture genomic position, operon orientation with respect to the replication fork, and identify mismatch repair proteins through homolog identification. Repeat regions are similarly highlighted, however some additional research is needed to fully quantify the intricacies of this feature for forecasts. It is likely that simply identifying tracts of repeats as a predictive feature will be overly reductive, as it is similarly important to identify the stability of secondary structures (Wright et al. 2003) and to identify which nucleotides will become mutable (De Boer and Ripley 1984; Dutra and Lovett 2006). With regards to the results gathered throughout this work, we have identified 6 nucleotide positions that in one combination form a mutational hotspot, and in another combination of 6 do not. Furthermore, *in silico* modelling using a pipeline developed by Wright and colleagues has demonstrated that these combinations have a clear impact on predicted DNA secondary structure (Wright et al. 2003). However, to ensure the accuracy of evolutionary forecasts, the formation of hairpin structures should first be investigated and elucidated in more detail.

This can be achieved experimentally via two methods, the first of which is designed to test the accuracy of the Wright model with regards to the model system used in this work. An experimental design that would allow this is to use the pipeline itself to generate predictions of hotspot formation. We can amend the input sequence data to create a predicted stem for a synthetically designed locus, and use features such as Mutational Index and the nucleotide of interest's distance from a stable stem (as highlighted in chapter 2) to predict if the edited locus would facilitate repeatable evolution. Mutant constructs with nucleotide sequence matching the modelled stems would then be engineered in the laboratory and the hotspot's presence would be identified through the presence or absence of parallel evolution following directed evolution, as described in chapter 2.

Preliminary work is already underway toward this goal. I have performed *in silico* modelling using the Wright pipeline for variants of the hotspot mutant Pf0-2x-sm (Fig. 5.1). This strain background was selected because the most stable secondary structures predicted by the pipeline for the ancestral Pf0-2x sequence and Pf0-2x-sm6 strain (which contains 6 synonymous changes; Fig. 5.1.A) involve nearly identical tracts of nucleotides. As such this allows us to visualise the predicted impact on secondary structures if a subset of the nucleotides are augmented (Fig. 5.1.B). From these, we can form predictions as to whether these hairpins should facilitate parallel evolution. Pf0-2x-sm4, for example, contains 4 synonymous changes which are all involved in predicted stem structure formation. (Fig. 5.1). However, unlike Pf0-2x-sm6 (which evolves highly repeatably; Fig. 2.5), this variant has two less synonymous changes, as these remaining two are predicted to exist as part of an unpaired loop in both the ancestral and 6 hairpin variant (Fig. 5.1.A). As such they should not have a bearing on secondary structure, which is reflected in the identical predicted structure between Pf0-2x-sm6 and Pf0-2x-sm4 (Fig. 5.1). Therefore our initial assumption will be that this variant will evolve as repeatably as the full 6 hairpin variant.

As an additional test of the model's efficacy, if this initial test is satisfied, is the generation of two additional mutants that contain only 2 synonymous changes involved in destabilising either the 'lower' or 'upper' stem in the ancestral Pf0-2x strain (Fig. 5.1). The hairpin variant Pf0-2x-sm2A is of particular interest, as with only two synonymous changes from the ancestral Pf0-2x, this variant's predicted most stable secondary structure involves a novel tract of nucleotides, yet produces similar characteristics (in terms of the mutable nucleotide's Mutational Index and distance from a single, stable stem; Fig. 5.1.B). In addition, this predicted secondary structures matches that of AR2, which evolves in a highly repeatable fashion (Sup Fig. 2.2). Therefore our prediction is that this hairpin should likewise evolve in parallel. Pf0-2x-sm2B (Fig. 5.1.B), offers a final test of the 'wiggle room' of these modelling predictions. In this variant the nucleotide is one base closer to the stable stem, and possesses a similar Mutational Index (-9.4) to the other variants, but it is unclear how specific the model's constraints – if found accurate in the other two variants – must be satisfied to determine the presence of a hotspot. As of this writing, all primer sets for engineering the plasmids required to generate these variants for allelic exchange have been designed and ordered (Supplementary Table 5.1), and two plasmid constructs (Pf0-2x-sm2A and Pf0-2x-sm2B) have been assembled.

An alternative approach to testing the efficacy of individual pipelines on predicting mutable secondary structures, is to generate a mutant library and develop or find a predictive tool that matches the observed experimental data. This would involve the generation of strains containing different combinations of the 6 synonymous base changes (of which there are $2^6 = 64$ in total), followed by their evolution to establish the strain's capability to evolve in a repeatable fashion. A reasonably high throughput means by which this could be achieved would involve the use of degenerate oligonucleotides with restricted assigned degeneracy at variant nucleotide positions (Supplementary Table 5.1). Using these oligos allows for the generation of a random and diverse library of plasmid constructs for allelic exchange from a single process of vector assembly (see chapter 2, Materials and Methods, for an outline of vector construction). Following mutant construction, the library could be screened initially using time to emergence data as a proxy for parallel evolution (as there is a clear positive correlation between these two characteristics: Fig. 2.4 and Fig. 2.5). This would be followed by confirmatory Sanger sequencing of the *ntrB* locus to both confirm parallel evolution (or the lack thereof) and identify the synonymous background.

Following these rounds of experiments, we would be equipped with a battery of synonymous sequence combinations and understand their resultant impact on facilitating parallel evolution. The library would therefore allow us to determine the critical nucleotides and the key combinations of synonymous sequence that enable repeatable evolution. In addition, it would reveal whether these hotspots were binary in their ability to evolve rapidly and in parallel (as in the case of the hotspot and non-hotspot variant constructs in chapter 2) or whether there exists an intermediate mutational hotspot between these two extremes. The oligonucleotides for library construction, as of the time of writing, have been designed and ordered (Supplementary Table 5.1).

The work described here focuses on the detailed nuances of mutation biases owed to secondary structure. In combination with results from chapters 2 and 3, I can develop the results from these experiments into a framework with which I can identify sites in genomes that are likely to be subjected to biased mutational outcomes, allowing for this work to be broadened beyond the model system in which it was developed.

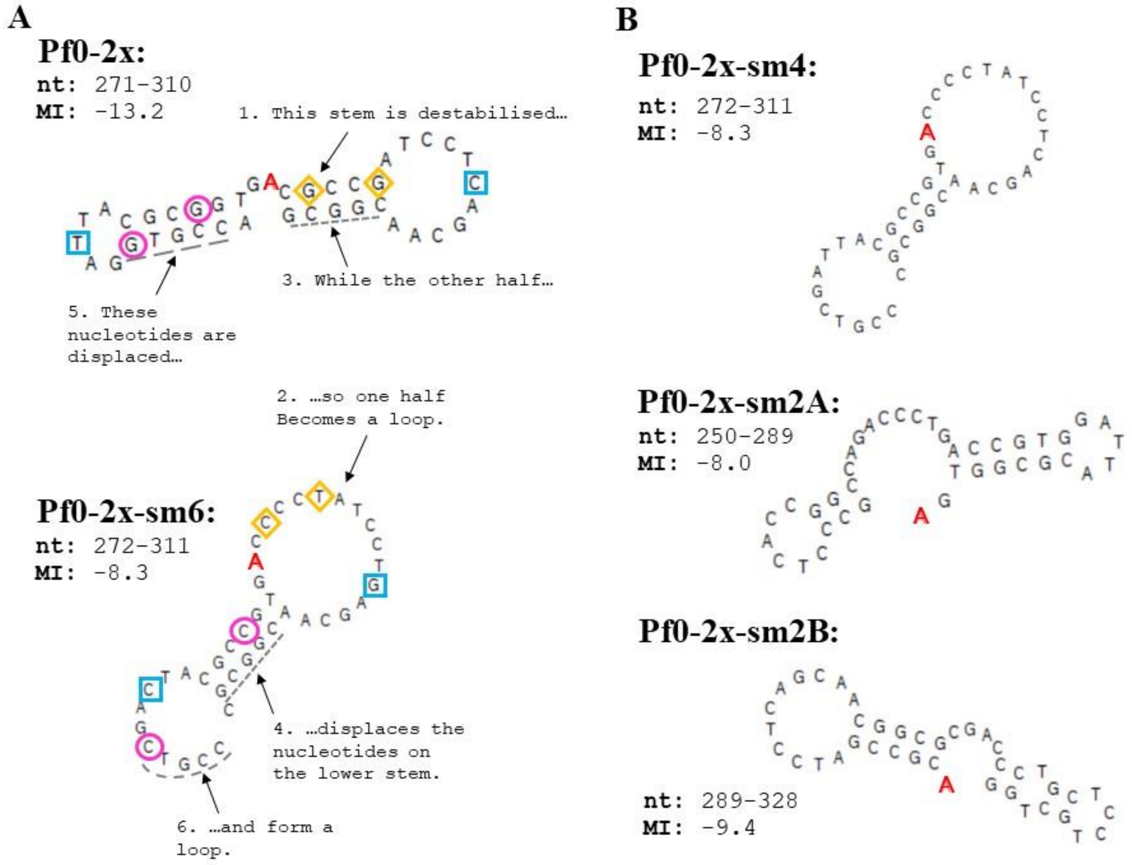


Figure 5.1. Pf0-2x hairpin variants with fewer synonymous changes possess similar or identical predicted secondary structures to the full hairpin variant. (**A**) The most stable secondary structures predicted for the ancestral Pf0-2x and the full hairpin variant, Pf0-2x-sm6, which differs from the ancestor at six nucleotide positions, as predicted by a pipeline authored by Wright et al. 2003. The six nucleotide changes are highlighted and divided into three categories based on their predicted impact on structural change. Two bases remain unpaired as part of loop structures in both variants (blue squares), and as such have no anticipated impact on secondary structure. A second set of two nucleotides are paired in the 'lower' stem of ancestral Pf0-2x (pink circles) and a third set paired in the 'upper' stem (orange diamonds). The impact of augmenting these latter two sets on predicted secondary structure is described by points 1-6 in the figure. (**B**) Alternative hairpin variants as predicted by Wright et al. 2003. Pf0-2x-sm4 has retained consensus with the ancestral Pf0-2x strain at two nucleotide positions found within loops (blue squares) but has consensus with Pf0-2x-sm6 at the other variable nucleotide positions. Pf0-2x-smA and B are entirely homologous to ancestral Pf0-2x aside from two synonymous changes involving either 'lower' or 'upper' stem (pink circles and orange diamonds, respectively). Depending on the synonymous context, the involved tract of nucleotides involved in forming the most stable secondary structure (**nt**), the Mutational Index of the mutable base (**MI**), and the distance from a single stable stem (image) either match or differ across hairpin variants. The mutable adenine, which undergoes a transversion mutation to a cytosine in instances of repeatable evolution, is highlighted as an 'A' in bold red.

5.3. In conclusion

In our forecasts, however, we have to keep in mind, as someone has said, that history makes us wise, but not necessarily clever." - Norberto J. Palleroni, The *Pseudomonas* story

Throughout this work I have demonstrated that the stochastic act of mutation can be coerced to generate the same adaptive solutions time and again. In chapter 2, I showed that highly repeatable genetic evolution of motility in previously immotile variants of *Pseudomonas fluorescens*, rather than being owed to more readily attributed explanatory variables such as mutational accessibility and clonal interference, was in fact a product of mutation bias that generated a mutational hotspot. Furthermore, this hotspot could be readily built and broken by a handful of silent mutations, revealing a key role for a typically overlooked genetic feature in ensuring predictable evolutionary outcomes. In chapter 3, I demonstrated that this mutational hotspot behaves via a 'house-of-cards' ruleset, as highly repeatable evolution is only achieved through a synergistic combination of multiple interacting genetic features. In chapter 4, I present evidence for the evolutionary suppression of these mutational hotspots, revealing that rather than being enriched at certain loci they may well appear throughout the bacterial genome. Taken together, this work demonstrates how powerful mutation bias can be in enforcing predictable evolution. It also details the genetic constrains that must be satisfied in order for a bacterium to enjoy such an effect. It then shows that these hotspots may well be pervasive across coding regions, and therefore they may be similarly deterministic in other observed adaptive events. Overall this work therefore acts as a key stepping stone toward future accurate evolutionary forecasts, by showcasing the power and prominence of these hotspots and describing in detail the features needed to identify them.

There are currently two broad approaches to forming evolutionary forecasts. The first is a 'top-down' approach which capitalises on the wealth of data and the advanced statistical tools now at our disposal. This approach employs a macro view whereby a large data set is analysed and predictive features are drawn from the data via statistical measures. An example of this is an autoregressive moving average model, which takes time-series data of an organism's evolution over many past generations to predict its future evolutionary trajectory (Nosil et al. 2020). Inherently such a broad approach carries inaccuracies and uncertainties and therefore can succumb to chaos, although some efforts have been made to mitigate against this (Rego-costa et al. 2017). Over short time windows and with an expectant degree of error, however, such approaches have already proven themselves worthy as predictive tools (Łuksza and Lassig 2014).

The alternative approach is a 'bottom-up' strategy, which aims to understand an evolving genome in forensic detail to enable highly accurate forecasts of genetic change. The primary merit of this framework is that it allows the researcher to mitigate and negate the role of chaos, as we have an intimate knowledge of our starting conditions. When working with mutational hotspots, we can go even further and shift toward deterministic principles from the stochastic process of mutation, which is typically assumed by macro approaches to be an unpredictable element of forecasting (Nosil et al. 2020). In this work I've established the foundations of the predicting microbial evolution pyramid, and future work can incrementally add layers of complexity to this model system and quantify their impact on repeatable genetic outcomes. This has been made possible thanks to the elucidation and characterisation of a powerful mutational hotspot.

It is likely that future forecasts of high fidelity will use the principles of a 'bottom-up' strategy to inform 'top-down' macro modelling efforts. As such the field investigates separately but moves together as one, driving us closer to predicting the future of evolution. These efforts are timely and of high importance. The rapid evolution of microbes (Wheatley et al. 2021) and cancers (Alves et al. 2019) presents a huge challenge to humankind. Due to their large population sizes and swift turnover in generations, microbes have eluded our efforts to combat infection with antibiotics time and again (Davies and Davies 2010). More recently, we have been struck with a rapidly evolving coronavirus (Liu et al. 2020) that has caused a global pandemic and impacted nearly every human alive today. This virus remains a huge threat, in large part because of its ability to evolve to increase infectivity (Korber et al. 2020) and thus circumvent vaccination efforts. Cancers, too, engage in an evolutionary arms race against their host's cells and therapeutic drugs (Casás-Selves and Degregori 2011). But what if we could anticipate these evolutionary events and pre-emptively counteract them before they occur? This is the ultimate ambition for those studying the predictability of evolution.

We have come a long way since Darwin, have made a giant leap since Dallinger. Yet despite our progress, accurately predicting evolutionary outcomes is most certainly an ambitious goal. But we have already made progress. This thesis has been dedicated to understanding a mutational hotspot, as a means to understand how evolutionary fates can be sealed before mutation has even occurred. As the hotspot is also found at the core of a key regulatory network, embedded within a histidine kinase of a two-component system, it also provides us with insight into the evolution of these ubiquitous bacterial tools (Tiwari et al. 2017) which mediate antibiotic resistance in many bacteria (Bhagirath et al. 2019). As such, while we continue to probe our way through the evolutionary fog in the present, the future of forecasting evolution looks bright.

5.4. Supplementary materials

Supplementary Table 5.1. List of primers for hotspot variant constructs. Sites with assigned degeneracy are underlined and labelled in bold. The controlled degeneracy corresponds to the following combinations: **S** – C or G; **Y** – C or T; **K** – G or T.

Oligo name	Oligo sequence 5’-3’	Notes
pf02x-sm4-F	CTGACCGTCGATTACGCCGTGACCCCTATCCTCAGCAACG	Used in conjunction with Pf0-2x synonymous substitution primers.
pf02x-sm4-RC	CGTTGCTGAGGATAGGGGTCACGGCGTAATCGACGGTCAG	
pf02x-sm2A-F	CTGACCGTGGATTACGCGGTGACCCCTATCCTCAGCAACG	
pf02x-sm2A-RC	CGTTGCTGAGGATAGGGGTCACCGCGTAATCCACGGTCAG	
pf02x-sm2B-F	CTGACCGTCGATTACGCCGTGACGCCGATCCTCAGCAACG	
pf02x-sm2B-RC	CGTTGCTGAGGATCGGCGTCACGGCGTAATCGACGGTCAG	
AR2-sm-degen-F	CTGACGGT**S**GA**Y**TACGC**S**GTGAC**S**CC**K**ATCCT**S**AGCAACG	Used in conjunction with AR2 synonymous substitution primers.
AR2-sm-degen-R	CGTTGCT**S**AGGAT**K**GG**S**GTCAC**S**GCGTA**Y**TC**S**ACCGTCAG	

Bibliography

1. Abbott S, Fairbanks DJ. 2016. Experiments on Plant Hybrids by Gregor Mendel. Genetics 204:407–422.
2. Alhama J, Ruiz-Laguna J, Rodriguez-Ariza A, Toribio F, López-Barea J, Pueyo C. 1998. Formation of 8-oxoguanine in cellular DNA of Escherichia coli strains defective in different antioxidant defences. Mutagenesis 13:589–594.
3. Alsohim AS, Taylor TB, Barrett GA, Gallie J, Zhang X, Altamirano-Junqueira AE, Johnson LJ, Rainey PB, Jackson RW. 2014. The biosurfactant viscosin produced by Pseudomonas fluorescens SBW25 aids spreading motility and plant growth promotion. Environ. Microbiol. 16:2267–2281.
4. Alves JM, Prado-López S, Cameselle-Teijeiro JM, Posada D. 2019. Rapid evolution and biogeographic spread in a colorectal cancer. Nat. Commun. [Internet] 10:4–10. Available from: http://dx.doi.org/10.1038/s41467-019-12926-8
5. Angus Buckling Michael A. Brockhurst & Nick Colegrave RCM. 2009. The Beagle in a bottle. Nature 457:824–829.
6. Avery OT, MacLeod CM, McCarty M. 1944. Studies on the chemical nature of the substance inducing transformation of Pneumococcal types. J. Exp. Med. 79:137–159.
7. Basener WF, Sanford JC. 2018. The fundamental theorem of natural selection with mutations. J. Math. Biol. [Internet] 76:1589–1622. Available from: https://doi.org/10.1007/s00285-017-1190-x
8. Basra P, Alsaadi A, Bernal-Astrain G, O'Sullivan ML, Hazlett B, Clarke LM, Schoenrock A, Pitre S, Wong A. 2018. Fitness Tradeoffs of Antibiotic Resistance in Extraintestinal Pathogenic Escherichia coli. Genome Biol. Evol. 10:667–679.
9. Beaumont HJE, Gallie J, Kost C, Ferguson GC, Rainey PB. 2009. Experimental evolution of bet hedging. Nature 462:90-U97.
10. Bhagirath AY, Li Y, Patidar R, Yerex K, Ma X, Kumar A, Duan K. 2019. Two component regulatory systems and antibiotic resistance in gram-negative pathogens. Int. J. Mol. Sci. 20.
11. Bhagirath AY, Li Y, Somayajula D, Dadashi M, Badr S, Duan K. 2016. Cystic fibrosis lung environment and Pseudomonas aeruginosa infection. BMC Pulm. Med. [Internet]:1–22. Available from: http://dx.doi.org/10.1186/s12890-016-0339-5
12. Bikard D, Loot C, Baharoglu Z, Mazel D. 2010. Folded DNA in Action: Hairpin Formation and Biological Functions in Prokaryotes. Microbiol. Mol. Biol. Rev. 74:570–588.
13. Blanco-romero E, Redondo-nieto M, Martínez-granero F, Garrido-sanz D, Ramos-gonzález MI, Martín M, Rivilla R. 2018. Genome-wide analysis of the FleQ direct regulon in Pseudomonas fluorescens F113 and Pseudomonas putida KT2440. Sci. Rep.:1–13.
14. Blount ZD, Lenski RE, Losos JB. 2018. Contingency and determinism in evolution: Replaying life's tape. Science (80-.). 362:50.

15. De Boer JG, Ripley LS. 1984. Demonstration of the production of frameshift and base-substitution mutations by quasipalindromic DNA sequences.
16. Bonneau KR, Mullens BA, MacLachlan NJ. 2001. Occurrence of Genetic Drift and Founder Effect during Quasispecies Evolution of the VP2 and NS3/NS3A Genes of Bluetongue Virus upon Passage between Sheep, Cattle, and Culicoides sonorensis . J. Virol. 75:8298–8305.
17. Brazda V, Fojta M, Bowater RP. 2020. Structures and stability of simple DNA repeats from bacteria. Biochem. J. 477:325–339.
18. Brisson D. 2018. Negative Frequency-Dependent Selection Is Frequently Confounding. Front. Ecol. Evol. 6:1–9.
19. Buisson R, Langenbucher A, Bowen D, Kwan EE, Benes CH, Zou L, Lawrence MS. 2019. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. Science (80-.). 2872.
20. Bush M, Dixon R. 2012. The Role of Bacterial Enhancer Binding Proteins as Specialized Activators of sigma54 -Dependent Transcriptions. Microbiol. Mol. Biol. Rev. 76:497–529.
21. Carrasco-Hernandez R, Jácome R, Vidal YL, de León SP. 2017. Are RNA Viruses Candidate Agents for the Next Global Pandemic? A Review. ILAR J. 58:343–358.
22. Carroll AC, Wong A. 2018. Plasmid persistence: costs, benefits, and the plasmid paradox. Can. J. Microbiol. 64:293–304.
23. Casás-Selves M, Degregori J. 2011. How Cancer Shapes Evolution and How Evolution Shapes Cancer. Evol. (N Y) 4:624–634.
24. Castillo-lizardo M, Henneke G, Viguera E. 2014. Replication slippage of the thermophilic DNA polymerases B and D from the Euryarchaeota Pyrococcus abyssi. Front. Microbiol. 5:1–10.
25. Chabes A, Georgieva B, Domkin V, Zhao X, Rothstein R, Thelander L, Street W, York N, York N. 2003. Survival of DNA Damage in Yeast Directly Depends on Increased dNTP Levels Allowed by Relaxed Feedback Inhibition of Ribonucleotide Reductase. Cell 112:391–401.
26. Chan K, Sterling JF, Roberts SA, Bhagwat AS, Resnick MA, Gordenin DA. 2012. Base Damage within Single-Strand DNA Underlies In Vivo Hypermutability Induced by a Ubiquitous Environmental Agent. PLOS Genet. 8:e1003149.
27. Chen F, Yang T, Zhong J, Zhang J, Li C, Yu X, Xiao J. 2018. Pan-Genomic Study of Mycobacterium tuberculosis Reflecting the Primary/Secondary Genes , Generality/Individuality , and the Interconversion Through Copy Number Variations. Front. Microbiol. 9:1–12.
28. Cheng X, Cordovez V, Etalo DW, Voort M Van Der. 2016. Role of the GacS Sensor Kinase in the Regulation of Volatile Production by Plant Growth-Promoting Pseudomonas fluorescens SBW25. Front. Plant Sci. 7:1–10.

29. Chevin L, Martin G, Lenormand T. 2010. FISHER ’ S MODEL AND THE GENOMICS OF ADAPTATION : RESTRICTED PLEIOTROPY , HETEROGENOUS MUTATION , AND PARALLEL EVOLUTION. :3213–3231.
30. Cobb M. 2014. Oswald Avery, DNA, and the transformation of biology. Curr. Biol. [Internet] 24:R55–R60. Available from: http://dx.doi.org/10.1016/j.cub.2013.11.060
31. Compeau G, Al-achi BJ, Platsouka E, Levy SB. 1988. Survival of Rifampin-Resistant Mutants of Pseudomonas fluorescens and Pseudomonas putida in Soil Systems. Appl. Environ. Microbiol. 54:2432–2438.
32. Couce A, Rodríguez-Rojas A, Blázquez J. 2016. Determinants of Genetic Diversity of Spontaneous Drug Resistance in Bacteria. Genetics 203:1369–1380.
33. Crozat E, Tardin C, Salhi M, Rousseau P, Lablaine A, Bertoni T, Holcman D, Sclavi B, Cicuta P, Cornet F. 2020. Post-replicative pairing of sister ter regions in Escherichia coli involves multiple activities of MatP. Nat. Commun. [Internet] 11:1–12. Available from: http://dx.doi.org/10.1038/s41467-020-17606-6
34. Cvijovic I, Good BH, Desai MM. 2018. The Effect of Strong Purifying Selection on Genetic Diversity. Genetics 209:1235–1278.
35. Darmon E, Leach DRF. 2014. Bacterial Genome Instability. MMBR 78:1–39.
36. Darwin C. 1832. Darwin Correspondence Project, “Letter no. 177.” Available from: https://www.darwinproject.ac.uk/letter/DCP-LETT-177.xml
37. Darwin C. 1859. On The Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life. London: John Murray
38. Dasgupta N, Wolfgang MC, Goodman AL, Arora SK, Jyot J, Lory S, Ramphal R. 2003. A four-tiered transcriptional regulatory circuit controls flagellar biogenesis in Pseudomonas aeruginosa. Mol. Microbiol. 50:809–824.
39. Davies J, Davies D. 2010. Origins and evolution of antibiotic resistance. Microbiol. Mol. Biol. Rev. 74:417–433.
40. Davis BD. 1989. Transcriptional bias : A non-Lamarckian mechanism for substrate-induced mutations. PNAS 86:5005–5009.
41. Denver DR, Howe DK, Wilhelm LJ, Palmer CA, Anderson JL, Stein KC, Phillips PC, Estes S. 2010. Selective sweeps and parallel mutation in the adaptive recovery from deleterious mutation in Caenorhabditis elegans. Genome Res. 20:1663–1671.
42. Dettman JR, Sztepanacz JL, Kassen R. 2016. The properties of spontaneous mutations in the opportunistic pathogen Pseudomonas aeruginosa. BMC Genomics [Internet] 17:1–14. Available from: http://dx.doi.org/10.1186/s12864-015-2244-3
43. Dillon MM, Sung W, Lynch M. 2018. Periodic Variation of Mutation Rates in Bacterial Genomes. MBio 9:1–15.

44. Dorman CJ, Dorman MJ. 2016. DNA supercoiling is a fundamental regulatory principle in the control of bacterial gene expression. Biophys. Rev. [Internet] 8. Available from: http://dx.doi.org/10.1007/s12551-016-0238-2
45. Drake JW. 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci. U. S. A.* 88:7160–7164.
46. Duret L. 2008. Neutral theory: The null hypothesis of molecular evolution. Nat. Educ. 1:218.
47. Dutra BE, Lovett ST. 2006. Cis and trans-acting effects on a mutational hotspot involving a replication template switch. J. Mol. Biol. 356:300–311.
48. Faure E, Kwong K, Nguyen D. 2018. Pseudomonas aeruginosa in Chronic Lung Infections: How to Adapt Within the Host? Front. Immunol. 9:1–10.
49. Fijalkowska IJ, Jonczyk P, Tkaczyk MM, Bialoskorska M, Schaaper RM. 1998. Unequal fidelity of leading strand and lagging strand DNA replication on the Escherichia coli chromosome. PNAS 95:10020–10025.
50. Foster PL, Niccum BA, Popodi E, Townes JP, Lee H, MohammedIsmail W, Tang H. 2018. Determinants of base-pair substitution patterns revealed by whole-genome sequencing of DNA mismatch repair defective Escherichia coli. Genetics 209:1029–1042.
51. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, Church GM, Pilpel Y. 2018. Codon usage of highly expressed genes affects proteome-wide translation efficiency. Proc. Natl. Acad. Sci. U. S. A. 115:E4940–E4949.
52. Ganai RA, Johansson E. 2016. DNA Replication—A Matter of Fidelity. Mol. Cell [Internet] 62:745–755. Available from: http://dx.doi.org/10.1016/j.molcel.2016.05.003
53. Ganeshan G, Kumar AM. 2007. Pseudomonas fluorescens , a potential bacterial antagonist to control plant diseases. J. Plant Interact. 1:123–134.
54. Gillespie JH. 1984. Molecular Evolution Over the Mutational Landscape. Evolution (N. Y). 38:1116.
55. Gon S, Camara JE, Klungsøyr HK, Crooke E, Skarstad K, Beckwith J. 2006. A novel regulatory mechanism couples deoxyribonucleotide synthesis and DNA replication in Escherichia coli. EMBO 25:1137–1147.
56. Gould SJ. 1989. Wonderful Life: The Burgess Shale and the Nature of History. First. New York: W. W. Norton & Co.
57. Greaves M, Maley CC. 2012. Clonal evolution in cancer. Nature 483:306–313.
58. Haas D, Défago G. 2005. Biological Control of Soil-Borne Pathogens By Fluorescent Pseudomonads. Nat. Rev. Microbiol.
59. Haas JW. 2000. The Reverend Dr William Henry Dallinger, F.R.S. (1839-1909). Notes Rec. R. Soc. Lond. 54:53–65.

60. Hayashi F, Smith KD, Ozinsky A, Hawn TR, Yi EC, Goodlett DR, Eng JK, Akira S, Underhill DM, Aderem A. 2001. The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. Nat. Lett. 410:1099–1103.

61. Henry R. 2012. entymologia Pseudomonas. Emerg. Infect. Dis. 18:1241.

62. Hermisson J, Pennings PS. 2005. Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335–2352.

63. Hervas AB, Canosa I, Little R, Dixon R, Santero E. 2009. NtrC-Dependent Regulatory Network for Nitrogen Assimilation in Pseudomonas putida. J. Bacteriol. 191:6123–6135.

64. Hervás AB, Canosa I, Santero E. 2008. Transcriptome analysis of Pseudomonas putida in response to nitrogen availability. J. Bacteriol. 190:416–420.

65. Hickman JW, Harwood CS. 2008. Identification of FleQ from Pseudomonas aeruginosa as a c-di-GMP-responsive transcription factor. Mol. Microbiol. 69:206–221.

66. Hmelo LR, Borlee BR, Almblad H, Love ME, Randall TE, Tseng BS, Lin CY, Irie Y, Storek KM, Yang JJ, et al. 2015. Precision-engineering the Pseudomonas aeruginosa genome with two-step allelic exchange. Nat. Protoc. 10:1820–1841.

67. Hoede C, Denamur E, Tenaillon O. 2006. Selection Acts on DNA Secondary Structures to Decrease Transcriptional Mutagenesis. PLOS Genet. 2:1697–1701.

68. Hogg M, Wallace SS, Doublie S. 2005. Bumps in the road: how replicative DNA polymerases see DNA damage. Curr. Opin. Struct. Biol. 15:86–93.

69. Holder IT, Wagner S, Xiong P, Sinn M, Frickey T, Hartig S, Meyer A. 2015. Intrastrand triplex DNA repeats in bacteria : a source of genomic instability. Nucleic Acids Res. 43:10126–10142.

70. de Hoon MJ, Eichenberger P, Vitkup D. 2010. Hierarchical evolution of the bacterial sporulation network. Curr. Biol. 20:1–7.

71. Huang S, Pang L. 2012. Comparing statistical methods for quantifying drug sensitivity based on in vitro dose-response assays. Assay Drug Dev. Technol. 10:88–96.

72. Hudson RE, Bergthorsson U, Ochman H. 2003. Transcription increases multiple spontaneous point mutations in Salmonella enterica. Nucleic Acids Res. 31:4517–4522.

73. Hudson RE, Bergthorsson U, Roth JR, Ochman H. 2002. Effect of Chromosome Location on Bacterial Mutation Rates. 19:85–92.

74. Huergo LF, Dixon R. 2015. The Emergence of 2-Oxoglutarate as a Master Regulator Metabolite. Microbiol. Mol. Biol. Rev. 79:419–435.

75. Jerome JP, Bell JA, Plovanich-jones AE, Barrick JE, Brown CT, Linda S. 2011. Standing Genetic Variation in Contingency Loci Drives the Rapid Adaptation of Campylobacter jejuni to a Novel Host. PLoS One 6:e16399.

76. Juurik T, Ilves H, Teras R, Ilmjarv T, Tavita K, Ukkivi K, Teppo A, Mikkel K, Kivisaar M. 2012. Mutation Frequency and Spectrum of Mutations Vary at Different Chromosomal Positions of Pseudomonas putida. PLoS One 7.

77. Jyot J, Dasgupta N, Ramphal R. 2002. FleQ, the Major Flagellar Gene Regulator in Pseudomonas aeruginosa, Binds to Enhancer Sites Located Either Upstream or Atypically Downstream of the RpoN Binding Site. J. Bacteriol. 184:5251–5260.

78. Kato T, Inagaki H, Tong M, Kogo H, Ohye T, Yamada K, Tsutsumi M, Emanuel BS, Kurahashi H. 2011. DNA secondary structure is influenced by genetic variation and alters susceptibility to de novo translocation. Mol. Cytogenet. 4:1–8.

79. Katsnelson MI, Wolf YI, Koonin E V. 2019. On the feasibility of saltational evolution. PNAS 116:21068–21075.

80. Kimura M. 1967. On the evolutionary adjustment of spontaneous mutation rates. Genet. Res. 9:23–34.

81. Kimura M. 1968. Evolutionary Rate at the Molecular Level. Nature [Internet] 217:624–626.

82. Kivisaar M. 2020. Mutation and Recombination Rates Vary Across Bacterial Chromosome. Microorganisms 8.

83. Klaric JA, Perr EL, Lovett ST. 2020. Identifying Small Molecules That Promote Mutations in Escherichia coli. G3 10:1809–1815.

84. Klasson L, Andersson SGE. 2006. Strong Asymmetric Mutation Bias in Endosymbiont Genomes Coincide with Loss of Genes for Replication Restart Pathways. MBE 23:1031–1039.

85. Klug A. 1968. Rosalind Franklin and the discovery of the structure of DNA. Nature 219:808–844.

86. Koch AL. 2004. Catastrophe and What To Do About It If You Are a Bacterium : The Importance of Frameshift Mutants. Crit. Rev. Microbiol. 30:1–6.

87. Koornhof HJ, Keddy K, Mcgee L. 2001. Clonal Expansion of Bacterial Pathogens across the World. J. Travel Med. 8:29–40.

88. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, et al. 2020. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell 182:812-827.e19.

89. Krishnaswami GS, Senapati H. 2019. An Introduction to the Classical Three-Body Problem.

90. Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science (80-.). [Internet] 324:255–258. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf

91. Kurokawa K, Nishida S, Emoto A, Sekimizu K, Katayama T. 1999. Replication cycle-coordinated change of the adenine nucleotide-bound forms of DnaA protein in Escherichia coli. EMBO 18:6642–6652.

92. Lang KS, Hall AN, Merrikh CN, Woodward JJ, Dreifus JE, Lang KS, Hall AN, Merrikh CN, Ragheb M, Tabakh H, et al. 2017. Replication-Transcription Conflicts Generate R-Loops that Orchestrate Bacterial Stress Survival Article Replication-Transcription Conflicts Generate R-

Loops that Orchestrate Bacterial Stress Survival and Pathogenesis. Cell [Internet] 170:787-790.e18. Available from: http://dx.doi.org/10.1016/j.cell.2017.07.044

93. Langenbucher A, Bowen D, Sakhtemani R, Bournique E, Wise JF, Zou L, Bhagwat AS, Buisson R, Lawrence MS. 2021. An extended APOBEC3A mutation signature in cancer. Nat. Commun. [Internet] 12. Available from: http://dx.doi.org/10.1038/s41467-021-21891-0

94. Lavi B, Karin EL, Pupko T, Hazkani-covo E. 2018. The Prevalence and Evolutionary Conservation of Inverted Repeats in Proteobacteria. Genome Biol. Evol. 10:918–927.

95. Leach DR. 1994. Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. Bioessays 16:893–900.

96. Lee H, Popodi E, Tang H, Foster PL. 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. 109.

97. Li W, Lu CD. 2007. Regulation of carbon and nitrogen utilization by CbrAB and NtrBC two-component systems in Pseudomonas aeruginosa. J. Bacteriol. 189:5413–5420.

98. Lind PA, Farr AD, Rainey PB. 2015. Experimental evolution reveals hidden diversity in evolutionary pathways. Elife 4.

99. Liu T, Gong D, Xiao J, Hu J, He G, Rong Z, Ma W. 2020. Cluster infections play important roles in the rapid evolution of COVID-19 transmission: A systematic review. Int. J. Infect. Dis. [Internet] 99:374–380. Available from: https://doi.org/10.1016/j.ijid.2020.07.073

100. Long H, Sung W, Miller SF, Ackerman MS, Doak TG, Lynch M. 2014. Mutation rate, spectrum, topology, and context-dependency in the DNA mismatch repair-deficient Pseudomonas fluorescens ATCC948. Genome Biol. Evol. 7:262–271.

101. Łuksza M, Lassig M. 2014. A predictive fitness model for influenza. Nature 507:57–61.

102. Lynch M, Conery J, Burger R. 1995. Mutation Accumulation and the Extinction of Small Populations. Am. Soc. Naturalists. 146:489-518.

103. Lynch M. 2010. Evolution of the mutation rate. Trends Genet. 26:345–352.

104. M. J. Merrick, Edwards RA. 1995. Nitrogen Control in Bacteria. Am. Soc. Microbiol. 59:604–622.

105. Martincorena I, Seshasayee ASN, Luscombe NM. 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nat. Lett. 485:95–98.

106. Maslowska KH, Makiela-dzbenska K, Mo J, Fijalkowska IJ, Schaaper RM. 2018. High-accuracy lagging-strand DNA replication mediated by DNA polymerase dissociation. PNAS 115:4212–4217.

107. Matthews TD, Rabsch W, Maloy S. 2011. Chromosomal Rearrangements in Salmonella enterica Serovar Typhi Strains Isolated from Asymptomatic Human Carriers. MBio 2:1–6.

108. Merrikh H. 2018. Spatial and temporal control of evolution through replication-transcription conflicts. Trends Microbiol. 25:515–521.

109. Million-weaver S, Samadpour AN, Moreno-habel DA, Nugent P, Brittnacher MJ. 2015. An underlying mechanism for the increased mutagenesis of lagging-strand genes in Bacillus subtilis. PNAS.

110. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D. 2013. Using Tablet for visual exploration of second-generation sequencing data. Brief. Bioinform. 14:193–202.

111. Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. TRENDS Genet. 17:589–596.

112. Mirkin E V., Mirkin SM. 2007. Replication Fork Stalling at Natural Impediments. Microbiol. Mol. Biol. Rev. 71:13–35.

113. Morita R, Nakane S, Shimada A, Inoue M, Iino H, Wakamatsu T, Fukui K, Nakagawa N, Masui R, Kuramitsu S. 2010. Molecular Mechanisms of the Whole DNA Repair System : A Comparison of Bacterial and Eukaryotic Systems. J. Nucleic Acids 2010.

114. Morreall J, Kim A, Liu Y, Degtyareva N, Weiss B. 2015. Evidence for Retromutagenesis as a Mechanism for Adaptive Mutation in Escherichia coli. PLOS Genet. 11:1–12.

115. Moxon ER, Rainey PB, Nowak MA, Lenski RE. 1994. Adaptive Evolution of Highly Mutable Loci in Pathogenic Bacteria. Curr. Biol. 4:24–33.

116. Moxon R, Bayliss C, Hood D. 2006. Bacterial Contingency Loci : The Role of Simple Sequence DNA Repeats in Bacterial Adaptation. Annu. Rev. Genet. 40:307–335.

117. Naito M, Pawlowska TE. 2016. The role of mobile genetic elements in evolutionary longevity of heritable endobacteria. Mob. Genet. Elements [Internet] 6:1–9. Available from: http://dx.doi.org/10.1080/2159256X.2015.1136375

118. Niccum BA, Lee H, Mohammedismail W, Tang H. 2019. crossm The Symmetrical Wave Pattern of Base-Pair Substitution Rates across the Escherichia coli Chromosome Has Multiple Causes. MBio 10.

119. Noble D. 2011. Neo-Darwinism, the Modern Synthesis and selfish genes: are they of use in physiology? J. Physiol. 589:1007–1015.

120. Nosil P, Flaxman SM, Feder JL, Gompert Z. 2020. Increasing our ability to predict contemporary evolution. Nat. Commun. [Internet] 10:1–6. Available from: http://dx.doi.org/10.1038/s41467-020-19437-x

121. O'Sullivan DJ, O'Gara F. 1992. Traits of Fluorescent Pseudomonas spp. Involved in Suppression of Plant Root Pathogens. 56:662–676.

122. Oliver A. 2010. Mutators in cystic fibrosis chronic lung infection: Prevalence, mechanisms, and consequences for antimicrobial therapy. Int. J. Med. Microbiol. [Internet] 300:563–572. Available from: http://dx.doi.org/10.1016/j.ijmm.2010.08.009

123. Orsi RH, Bowen BM, Wiedmann M. 2010. Homopolymeric tracts represent a general regulatory mechanism in prokaryotes. BMC Genomics 11.

124. Otto SP, Whitlock MC. 1997. The Probability of Fixation in Populations of Changing Size. Genetics 146:723–733.

125. Palleroni NJ. 2010. The Pseudomonas Story. Environ. Microbiol. 12:1377–1383.

126. Park C, Qian W, Zhang J. 2012. Genomic evidence for elevated mutation rates in highly expressed genes. EMBO Rep. 13:1123–1129.

127. Park Y, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn D, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019 F abio. Nucleic Acids Res. 47:636–641.

128. Paul S, Million-Weaver S, Chattopadhyay S, Sokurenko E, Merrikh H. 2013. Accelerated gene evolution via replication-transcription conflicts. Nature 495:1–13.

129. Pavlov YI, Newlon CS, Kunkel TA, Carolina N. 2002. Yeast Origins Establish a Strand Bias for Replicational Mutagenesis. Mol. Cell 10:207–213.

130. Payne JL, Menardo F, Trauner A, Borrell S, Gygli SM, Loiseau C, Gagneux S, Hall AR. 2019. Transition bias influences the evolution of antibiotic resistance in Mycobacterium tuberculosis. Plos Biol.:1–23.

131. Poelwijk FJ, Sorin T, Kiviet DJ, Tans SJ. 2011. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. J. Theor. Biol. 272:141–144.

132. Poon A, Chao L. 2005. The rate of compensatory mutation in the DNA bacteriophage φX174. Genetics 170:989–999.

133. Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. Proc. Natl. Acad. Sci. U. S. A. 104:8605–8612.

134. R Core Team. 2014. R: A language and environment for statistical computing. R Found. Stat. Comput. Vienna, Austria [Internet]. Available from: http://www.r-project.org/.

135. Raeside C, Gaffé J, Deatherage DE, Tenaillon O, Briska AM, Ptashkin RN, Cruveiller S, Médigue C, Lenski RE, Barrick JE. 2014. Large Chromosomal Rearrangements during a Long-Term Evolution Experiment with Escherichia coli. 5:1–13.

136. Rainey PB, Bailey MJ. 1996. Physical and genetic map of the Pseudomonas fluorescens SBW25 chromosome. Mol. Microbiol. 19:521–533.

137. Rego-costa A, Florence D, Chevin L. 2017. Chaos and the (un)predictability. Evolution (N. Y). 72:375–385.

138. Rocha EPC. 2004. The replication-related organization of bacterial genomes. Microbiology 150:1609–1627.

139. Rocha EPC, Danchin A, Bacte Â. 2003. Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res. 31:6570–6577.

140. Rodríguez-Beltrán J, DelaFuente J, León-Sampedro R, MacLean RC, San Millán Á. 2021. Beyond horizontal gene transfer: the role of plasmids in bacterial evolution. Nat. Rev. Microbiol.

141. Roth JR, Benson N, Galitski. T, Haack K, Lawrence JG, Miesel L. 1996. Rearrangement of the bacterial chromosome: Formation and Applications. 2nd ed. Washington D.C.: ASM Press 2256 2276

142. Roth JR, Kugelberg E, Reams AB, Kofoid E, Andersson DI. 2006. Origin of mutations under selection: The adaptive mutation controversy. Annu. Rev. Microbiol. 60:477–501.

143. Sabari Sankar T, Wastuwidyaningtyas BD, Dong Y, Lewis SA, Wang JD. 2016. The Nature of Mutations Induced by Replication-Transcription Collisions. Nature 535:171–181.

144. Scales BS, Dickson RP, LiPuma JJ, Huffnagle GB. 2014. Microbiology, Genomics, and Clinical Significance of the Pseudomonas fluorescens Species Complex, an Unappreciated Colonizer of Humans. Clin. Microbiol. Rev. 27:927–948.

145. Schroeder JW, Hirst WG, Szewczyk GA, Simmons LA. 2016. The effect of local sequence context on mutational bias of leading and lagging strand genes. Curr. Biol. 26:692–697.

146. Seaton SC, Silby MW, Levy SB. 2013. Pleiotropic effects of gaca on pseudomonas fluorescens pf0-1 in vitro and in soil. Appl. Environ. Microbiol. 79:5405–5410.

147. Seemann T. 2015. Snippy: fast bacterial variant calling from NGS reads. Available from: https://github.com/tseemann/snippy

148. Sekowska A, Wendel S, Fischer EC, Nørholm MHH. 2016. Generation of mutation hotspots in ageing bacterial colonies. Sci. Rep. 6:4–10.

149. Shis DL, Bennett MR, Igoshin OA. 2018. Dynamics of Bacterial Gene Regulatory Networks. Annu. Rev. Biophys. 47:447–467.

150. Silby MW, Cerdeño-tárraga AM, Vernikos GS, Giddens SR, Jackson RW, Preston GM, Zhang X, Moon CD, Gehrig SM, Godfrey SAC, et al. 2009. Open Access Genomic and genetic analyses of diversity and plant interactions of Pseudomonas fluorescens. Genome Biol. 10:R51.

151. Silby MW, Winstanley C, Godfrey SAC, Levy SB, Jackson RW. 2011. Pseudomonas genomes: diverse and adaptable. FEMS Microbiol. Rev. 35:652–680.

152. Srivatsan A, Tehranchi A, Macalpine DM, Wang JD. 2010. Co-Orientation of Replication and Transcription Preserves Genome Integrity. PLOS Genet. 6.

153. Stoltzfus A, McCandlish DM. 2017. Mutational biases influence parallel adaptation. Mol. Biol. Evol. 34:2163–2172.

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

155. Taddei F, Radman M, Maynard-Smith J, Toupance B, Gouyon PH, Godelle B. 1997. Role of mutator alleles in adaptive evolution. Nat. Lett. 387:700–702.

156. Taverna P, Sedgwick B. 1996. Generation of an Endogenous DNA-Methylating Agent by Nitrosation in Escherichia coli. J. Bacteriol. 178:5105–5111.

157. Taylor TB, Mulley G, Dills AH, Alsohim AS, McGuffin LJ, Studholme DJ, Silby MW, Brockhurst MA, Johnson LJ, Jackson RW. 2015. Evolutionary resurrection of flagellar motility via rewiring of the nitrogen regulation system. Science (80-.). 347:1014–1017.

158. Theißen G. 2009. Saltational evolution: Hopeful monsters are here to stay. Theory Biosci. 128:43–51.

159. Theunissen B. 1994. Knowledge is power: Hugo de Vries on science, heredity and social progress. Br. Soc. Hist. Sci. 27:291–311.

160. Tiwari S, Jamal SB, Hassan SS, Carvalho PVSD, Almeida S, Barh D, Ghosh P, Silva A, Castro TLP, Azevedo V. 2017. Two-component signal transduction systems of pathogenic bacteria as targets for antimicrobial therapy: An overview. Front. Microbiol. 8:1–7.

161. Tomkova M, Tomek J, Kriaucionis S, Schuster-böckler B. 2018. Mutational signature distribution varies with DNA replication timing and strand asymmetry. Genome Biol.:1–12.

162. Trinh TQ, Sinden RR. 1991. Preferential DNA secondary structure mutagenesis in the lagging strand of replication in E. coli. Nat. Lett. 352:544–547.

163. Turnbull GA, Morgan JAW, Whipps JM, Saunders JR. 2001. The role of bacterial motility in the survival and spread of Pseudomonas Fluorescens in soil and in the attachment and colonisation of wheat roots. FEMS Microbiol. Ecol. 36:21–31.

164. Voineagu I, Narayanan V, Lobachev KS, Mirkin SM. 2008. Replication stalling at unstable inverted repeats : Interplay between DNA hairpins and fork stabilizing proteins. PNAS 105:9936–9941.

165. Wang G, Vasquez KM. 2017. Effects of replication and transcription on DNA Structure-Related genetic instability. Genes (Basel). 8.

166. Warnecke T, Supek F, Lehner B. 2012. Nucleoid-Associated Proteins Affect Mutation Dynamics in E. coli in a Growth Phase-Specific Manner. PLoS Comput. Biol. 8.

167. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2 — a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191.

168. Watson J, Crick F. 1953. Molecular structure of nucleic acids. Nature 171:737–738.

169. Watt DL, Buckland RJ, Lujan SA, Kunkel TA, Chabes A. 2016. Genome-wide analysis of the specificity and mechanisms of replication infidelity driven by imbalanced dNTP pools. Nucleic Acids Res. 44:1669–1680.

170. Weigand MR, Peng Y, Loparev V, Batra D, Bowden KE, Burroughs M, Cassiday PK, Davis JK, Johnson T, Juieng P, et al. 2017. The History of Bordetella pertussis Genome Evolution Includes Structural Rearrangement. J. Bacteriol. 199:1–15.

171. Weinreich DM, Delaney NF, De Pristo MA, Hartl DL. 2006. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. Science (80-.). 312.

172. Wernegreen JJ, Kauppinen SN, Degnan PH. 2010. Slip into Something More Functional : Selection Maintains Ancient Frameshifts in Homopolymeric Sequences Research article. Mol. Biol. Evol. 27:833–839.

173. Wheatley R, Diaz Caballero J, Kapel N, de Winter FHR, Jangir P, Quinn A, del Barrio-Tofiño E, López-Causapé C, Hedge J, Torrens G, et al. 2021. Rapid evolution and host immunity drive the rise and fall of carbapenem resistance during an acute Pseudomonas aeruginosa infection. Nat. Commun. 12:1–12.

174. White JH, Bauer WR. 1987. Superhelical DNA with Local Substructures: A Generalization of the Topological Constraint in Terms of the Intersection Number and the Ladder-like Correspondence Surface. J. Mol. Biol. 195:205–213.

175. Wright BE, Reschke DK, Schmidt KH, Reimers JM, Knight W. 2003. Predicting mutation frequencies in stem-loop structures of derepressed genes: Implications for evolution. Mol. Microbiol. 48:429–441.

176. Wu Z, Liu J, Yang H, Xiang H. 2014. DNA replication origins in archaea. Front. Microbiol. 5:1–7.

177. Zhang Xiaolong, Zhang Xuehong, Zhang Xia, Liao Y, Song L, Zhang Q, Li P, Tian J, Shao Y, Li Y, et al. 2018. Spatial Vulnerabilities of the Escherichia coli Genome. Genetics 210:547–558.

178. Zhang XX, Rainey PB. 2008. Dual involvement of CbrAB and NtrBC in the regulation of histidine utilization in Pseudomonas fluorescens SBW25. Genetics 178:185–195.

179. Zheng L, Kostrewa D, Berne S, Winkler FK, Li X. 2004. The mechanism of ammonia transport based on the crystal structure of AmtB of Escherichia coli. PNAS 101.

180. Zhou K, Aertsen A, Michiels CW. 2014. The role of variable DNA tandem repeats in bacterial adaptation. Fems Microbiol. Rev. 38:119–141.

181. Zou Y. 2014. Charles Darwin’s Theory of Pangenesis. Embryo Proj. Encycl.:1–5.