Transcriptional Effects of Adaptive Synonymous Mutations in *Pseudomonas fluorescens*

Nicholas McCloskey

Thesis submitted to

The Faculty of Graduate and Postdoctoral Studies

In partial fulfillment of the requirements for

the M.Sc. degree in the

Department of Biology

University of Ottawa

ABSTRACT

Synonymous mutations have traditionally been thought to have no significant effect on fitness. However, a growing body of recent research has shown that this is not always the case. In an experimentally evolved population of *Pseudomonas fluorescens* grown in minimal glucose media, synonymous mutations arose in a glucose transport gene that resulted in beneficial fitness effects comparable to those of non-synonymous mutations. We found that the increase in fitness was a direct result of increased gene expression; however, the precise mechanism was unclear. Synonymous mutations have been shown to affect gene expression on transcriptional and translational levels through changes in mRNA secondary structure and codon usage.

Our study investigates the underlying mechanisms in which these evolved synonymous mutations lead to increased gene expression. In addition to the evolved mutations, we have a library of 42 strains with single synonymous mutations within the glucose transport gene and found a positive correlation between fitness and gene expression. To determine whether these mutations affect transcript levels, translational efficiency or a combination of both, we systematically incorporated transcriptional and translational fusions of a yellow fluorescent protein within the glucose transport operon. We found that the evolved mutations predominantly act on the level of transcription and have strong polar downstream effects. Additionally, through manipulation of the local genetic sequence, we investigated the specific molecular requirements necessary for the increased expression. We found that for one of our evolved synonymous mutants, mRNA secondary structure does not play an essential role, but we speculate that the mutation may strengthen a weak internal promoter sequence to confer its increased expression. Our study provides evidence of the adaptive mechanisms of beneficial synonymous mutations in an experimentally evolved setting.

RÉSUMÉ

Les mutations synonymes ont traditionnellement été considérées comme n'ayant aucun effet significatif sur la valeur sélective. Cependant, un nombre croissant de recherches récentes a montré que ce n'est pas toujours le cas. Dans une population expérimentalement développée de *Pseudomonas fluorescens* cultivée dans des milieux de glucose minimaux, des mutations synonymes sont apparues dans un gène de transport du glucose qui a entraîné des effets bénéfiques sur la valeur sélective comparables à ceux de mutations non synonymes. Nous avons constaté que l'augmentation de la valeur sélective était le résultat direct de l'augmentation de l'expression génique. Cependant, le mécanisme précis n'était pas clair. Notre étude examine les mécanismes sous-jacents dans lesquels ces mutations synonymes évoluées conduisent à une augmentation de l'expression des gènes. En plus des mutations évoluées, nous avons un ensemble de 42 souches avec des mutations simples synonymes dans le gène de transport du glucose et trouvé une corrélation positive entre la valeur sélective et l'expression des gènes. Pour déterminer si ces mutations affectent les niveaux de transcription, l'efficacité traductionnelle ou une combinaison des deux, nous avons systématiquement incorporé des fusions transcriptionnelles et traductionnelle d'une protéine fluorescente jaune dans l'opéron de transport du glucose. Nous avons constaté que les mutations évoluées agissent principalement sur le niveau de transcription et ont de forts effets polaires en aval. De plus, grâce à la manipulation de la séquence génétique locale, nous avons étudié les exigences moléculaires spécifiques nécessaires pour l'expression accrue. Nous avons trouvé que pour l'un de nos mutants synonymiques évolués, la mutation peut renforcer une séquence de promoteur interne faible pour conférer son expression accrue. Notre étude fournit des preuves des mécanismes adaptatifs des mutations bénéfiques synonymes dans un cadre expérimentalement évolué.

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor Dr. Rees Kassen. Your guidance over the past three years has been invaluable and your continuous support in helping me reach my academic goals is truly appreciated. Your passion for research is inspiring and made my transition into evolutionary biology an enjoyable experience.

I'd like to thank Aaron Hinz for your contributions to my experimental design and for creating the tools necessary to allow my my project to be possible. Thank you for answering all of my questions, keeping my outlandish ideas in check, and always asking me to go for coffee despite knowing the answer.

Thank you to Eleonore Lebeuf-Taylor, for your dedicated work and for making the office a little less desolate with all your beautiful plants. Thank you Patrick, for the endless supply of Oreos and discounted Halloween candy during those cold winter months. Thank you Felipe for all your encouraging words and making me not feel guilty for eating all of Patrick's candy. Thank you Mackenzie for all the delicious homemade treats and bringing me along to win OSSC volleyball gold. Thank you Sonal for providing me company for all the late nights in the lab. To Mohammed, for all the gels you've poured (correctly), I am grateful. And to Bridget, thank you for keeping Mohammed in line. To all past and present lab members, including Jonathan Dench, Alana Schick, Alanna Leale, Jeremy Dettman and Anita Melnyk, thank you for your insight, advice and company over the past three years.

I would like to also thank my committee members, Dr. Alex Wong, Dr. Alex Poulain and Dr. Morgan Fullerton, for their valuable advice and encouragement over the course of my degree.

A personal thank you goes out to my family and close friends, who despite at times became second priority to my bacterial cultures, stuck with me the entire way.

During my research, I have received financial support through OGS and University of Ottawa scholarships.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction to the Effects of Synonymous Mutations

Synonymous mutations have traditionally been thought to have no significant effect on fitness. However, a growing body of recent research has shown that this is not always the case. In light of the discovery of two beneficial adaptive synonymous mutations, our study aims to uncover the underlying mechanism in which these synonymous mutations confer their fitness effect. We will explore transcriptional and translational effects of synonymous mutations to help strengthen case for the underappreciated role of synonymous mutations in adaptive evolution.

I. Classical View of Synonymous mutations

The redundancy of the genetic code refers to the fact that 18 of the 20 amino acids can be encoded by multiple codons. Genetic variability therefore exists in two main classes of mutations: synonymous and non-synonymous. Non-synonymous mutations, also referred to as missense mutations, are single nucleotide substitutions that result in a codon that encodes a different amino acid. The effects of non-synonymous mutations have been thoroughly studied since amino acid substitutions can have significant effects on both the structure and function of proteins. Synonymous mutations, on the other hand, also referred to as silent mutations, are single nucleotide substitutions that do not alter the amino acid sequence of a protein. For this reason, they were thought to have no significant effect on fitness, and therefore have traditionally been considered neutral in the context of natural selection and evolution.

The neutrality of mutations in general made their way into early theories of evolution. Specifically, in the 1960s Kimura introduced the neutral theory of molecular evolution which was in sharp contrast to selectionist theory, as it proposed that the majority of genetic variation seen

within finite populations was not due to natural selection, but rather genetic drift and fixation of neutral mutations. The theory does not deny the existence of selection, but postulates only a small fraction of mutations are adaptive and that most segregating variation is effectively neutral with respect to fitness (Kimura 1977). Moreover, the theory predicts that slightly beneficial or deleterious mutations behave as selectively neutral when their selection coefficients (s) are less than the reciprocal of the effective population size ($s < 1/N_e$) (Kimura 1968).

The proposal that the majority of observed mutations were effectively neutral carried over to applications that quantify and predict evolutionary outcomes. Specifically, the dN/dS calculation was introduced as a useful tool to detect and estimate direction and magnitude of selection by measuring the rates of non-synonymous (dN) and synonymous (dS) substitutions at a given site (Yang and Bielawski 2000). However, the calculation makes the assumption that all synonymous substitutions are strictly neutral and so they are taken to reflect the neutral background substitution rate. A dN/dS ratio greater than 1 implies selection is promoting protein changes, while a dN/dS ratio less than 1 implies selection is acting against protein changes. The tool is used to understand DNA sequence evolution, and by comparing dN/dS ratios between lineages, can help detect specific sites that may be under positive selection. However, with advancements in both genome-wide analyses and large scale comparative and experimental studies, new evidence has emerged revealing that significant portion of synonymous sites are under selection (Lawrie et al. 2013) and can have impacts on both gene expression and fitness (Lind et al. 2010; Agashe et al. 2012). These results have updated the traditional assumption that synonymous mutations are strictly neutral and have challenged the accuracy of the tools used to study evolutionary processes.

II. Contemporary View of Synonymous Mutations

The first major development for the case of synonymous mutations under selection came in the 1980s as a result of the development of DNA sequencing. A comparative study of both unicellular prokaryotes and eukaryotes had observed that the frequency of certain synonymous codons was higher than expected by chance, which is referred to as codon usage bias. A strong codon usage bias was discovered in a set of highly expressed genes, while genes of low expression had a much weaker bias. The differences in codon bias were speculated to be mediated by the strength of translational selection across the genes (Sharp and Li 1986). Alongside technological advances in whole-genome genotyping, evidence emerged showing that selection of synonymous sites was not exclusive to unicellular organisms, but present in all domains of life, from *C. elegans* (Duret 2000), to *D. melanogaster* (Lawrie et al. 2013), and even *H. sapiens* (Sharp et al. 1988).

However, to make claims that synonymous sites are under selection, it is important to be able to create a link between genotype and phenotypic effects. With the introduction of microarray analyses and RNA sequencing, it became feasible to determine expression profiles of large sets of genes simultaneously across a genome. In *E. coli*, codon usage bias was shown to have a strong positive correlation with translational efficiency (protein abundance per mRNA level) across its genome (Tuller et al. 2010a). In *D. melanogaster*, 22% of synonymous sites are under strong selective constraints and are most frequently located in highly expressed genes (Lawrie et al. 2013). Additionally, manual introduction of synonymous mutations in targeted genes in *S. typhimurium* and *M. extorquens* were shown to have significant deleterious effects on both protein levels and overall growth (Lind et al. 2010; Agashe et al. 2012).

Although the phenotypic evidence of selection on synonymous mutations is compelling, the most direct evidence for adaptive evolution is through their effects on fitness. Fitness can be

defined as an individual's genetic contribution to the gene pool of the next generation, and for asexual organisms, fitness effects are directly assigned to their genotype. Fitness for microbes is often measured in one of two ways. The first method measures and compares exponential growth rates in the selected environment by means of optical density. The second is a competitive fitness assay, in which a desired strain is co-cultured at equal density with an easily identifiable reference strain (usually a marked ancestor) and fitness is measured as a change in frequency of the two strains over time (Bataillon and Bailey 2014). When studying fitness, more inferences can be made when there are enough genotypes to create a distribution of fitness effects. To accomplish this, studies often employ site directed mutagenesis, in which targeted genes are bombarded with either single mutations or combinations of mutations. This allows for a quick and efficient way to create a library of mutants (synonymous or non-synonymous) for which a range of fitness effects can be assessed.

That being said, there have been several cases in which synonymous mutations demonstrate considerable effects on fitness. In two separate mutagenesis studies of RNA viruses, synonymous mutations were identified to have fitness effects ranging from strongly deleterious (lethal) to wild-type-like (Sanjuán et al. 2004; Canale et al. 2018). A mutagenesis of ribosomal proteins in *S. typhimurium* showcased a distributed range of fitness effects for 38 single synonymous mutations with selection coefficients ranging from –0.0279 to 0 (Lind et al. 2010). A follow-up study had taken four of the deleterious single synonymous mutations and demonstrated that through experimental evolution, compensatory synonymous mutations were able to efficiently restore fitness (Knöppel et al. 2016). Another study performed mutagenesis on a key enzyme required in the formaldehyde oxidation pathway in *M. extorquens* but differed from Lind et al. as they assessed fitness of alleles with 46 to 150 synonymous mutations (Agashe et al. 2012). Although the mutated

strains all had deleterious fitness effects, compensatory synonymous mutations arose during evolution experiments to restore fitness in a similar manner to the *S. typhimurium* study seen before (Agashe et al. 2016). Lastly, a study that had removed a key enzyme for the production of arginine (Δ*argC*) in *E. coli*, showed the emergence of a synonymous mutation that restored fitness through increased expression in a second maladapted, yet bifunctional enzyme, *proA*, that could catalyze reactions in both the arginine and proline pathways. Interestingly, the single synonymous mutation displayed an impressive 5.1-fold increase in growth rate for the Δ*argC* strain, albeit still lower than the growth rate of the wild-type strain (Kershner et al. 2016).

In summary, the distribution of fitness effects of synonymous mutations introduced via site directed mutagenesis ranged from deleterious to neutral. Furthermore, only through experimental evolution of mutated, maladapted strains were synonymous mutations able to demonstrate their adaptive potential through compensatory mutations.

However, a new direct line of evidence for the adaptive potential of synonymous mutations emerged as work by Bailey et al. showed that in an experimentally evolved population of *Pseudomonas fluorescens*, grown in a minimal glucose media, two independently evolved synonymous mutations were recovered that were shown to have comparable beneficial fitness effects to that of a non-synonymous mutation found within the same experimental population (Bailey et al. 2014). Both synonymous mutations were located within the same gene, *gtsB*, and were on trajectories to fix within the population after only 200 generations, consistent with natural selection of strongly beneficial mutations. This was compelling evidence that beneficial synonymous mutations may not be as rare as previously thought.

To build upon that idea, a fellow Masters student, Eleonore Lebeuf-Taylor, tested how common beneficial synonymous mutations were in the *gtsB* gene (unpublished). Specifically, she

assessed the distribution of fitness effects of single synonymous mutations throughout the *gtsB* gene introduced through site directed mutagenesis. A library of over 100 single point mutants was created, including over 40 single synonymous mutants, and their fitness was assessed through head-to-head competitions against the ancestral strain. Interestingly, synonymous and non-synonymous single mutants displayed a range of both deleterious and beneficial fitness effects. Additionally, for the subset of beneficial mutations, there was no significant difference in the distribution of fitness effects between synonymous and non-synonymous mutations. The results of these two studies made a strong case for the underappreciated role of synonymous mutations in the context of adaptive evolution.

However, what remained elusive, and became the focal point of my research, was the molecular mechanism behind the fitness effect of these synonymous mutations. In the above-mentioned cases, adaptive synonymous mutations predominantly confer their effect on fitness through changes in gene expression, but the exact underlying mechanisms remain elusive (Agashe et al. 2016; Knöppel et al. 2016). Over recent years, genome-wide analyses and large scale mutagenic experiments have generated an impressive list of mechanisms to explain why synonymous sites are under selection. Validating that these conserved mechanisms can be applied to individual cases of large effect synonymous mutations has proven difficult as measurements of mRNA levels and protein abundance alone are not sufficient to elucidate an exact mechanism (Agashe et al. 2016). As it stands, there is little evidence and no strong consensus to explain how beneficial single synonymous mutations are able to confer their fitness effects.

III. Translational and Transcriptional Mechanisms of Synonymous Mutations

To better understand how synonymous mutations can impact gene expression, I will review the literature on currently known mechanisms.

i. Codon Usage Bias

As mentioned before, the first evidence of synonymous sites under selection came in the 1980s, when a bias in synonymous codon usage in *E. coli* was correlated to high levels of gene expression (Sharp and Li 1986). Codon usage bias is thought to stem from a selective pressure to utilize the most abundant tRNAs in order to maximize the translational speed and accuracy of the protein. A codon adaptation index (CAI) was introduced shortly after as a means to quantify the codon usage bias of a given gene. A reference set of highly expressed genes from a species, assumed to be under the strongest selection to optimize translational efficiency, is used to assess codon frequencies and create a relative score, from 0 to 1, for each codon. CAI of a given gene can be calculated by assessing the frequency of its codons, and be used as a predictor of gene expression as well as a predictor of the effectiveness selection has had on shaping its codon usage (Sharp and Li 1987). Some drawbacks of using CAI is that it can be difficult to create a meaningful set of reference genes if data is limited and it does not address wobble base pairing between codons and tRNA molecules.

In search of a more accurate predictor of translational efficiency, in 2004, dos Reis introduced the tRNA adaptation index (tAI). Instead of using a set of reference genes, the tAI uses relative abundance of each tRNA within the tRNA pool, or when data is not available, uses tRNA copy numbers within the genome as a predictor of relative tRNA abundance. Unlike CAI, tAI also takes into consideration wobble interactions between codon and anti-codons which accounts for the fact that a given codon can accept different tRNAs species with varying affinities.

dos Reis was able to show that strength of translational selection varied across species but is present in both prokaryotes and eukaryotes alike. 60% of the strength of selection was explained by genome size and the number of tRNA genes within the genome. Their model proposed that organisms adjust their codon usage bias to match the skew in the tRNA pool to increase translational efficiency. However, selection is only effective when a) there is high enough tRNA copy numbers to create a skewed tRNA pool and b) genome size is not too small as to limit the number of tRNA copy numbers, but also not too large. Since large genome sizes are correlated to longer cell cycles (dos Reis et al. 2004), they speculated translational optimization of large proteins is of little value to humans compared to organisms with short generation times such as *E. coli*. From this model they identified optimal genome sizes, where translational selection was the strongest, as well as upper and lower limits to genome sizes where translational selection was not seen. Their model was able to explain why organisms with small and large (*Helicobacter pylori*, *Homo sapiens*) genome sizes have no selection on codon usage compared to other organisms such as *S. cerevisiae* and *E. coli*, which fall into a selective "hotspot" for genome size (dos Reis et al. 2004).

Moreover, another mechanism involving codon usage was shown in *S. cerevisiae* that incorporates autocorrelated codons to increase translational efficiency (Cannarozzi, 2010). Autocorrelated codons, codons that reoccur within in proximity of each other on the transcript, demonstrated the ability to recycle tRNA molecules before they diffuse away from the ribosome, therefore increasing translational efficiency. This finding has implications for both CAI and tAI interpretations. In a scenario in which a rare codon is repeated in sequence within proximity, both CAI and tAI values would underestimate measures of translational efficiency. The rare codon

would presumably be translated faster than expected since its tRNA molecule is likely to be recycled.

ii. Pause Elements

Conversely, other selective features of translation have emerged that incorporate transient pause elements to increase translational efficiency and proper protein production. In 2010, Tuller had shown that several organisms of all three kingdoms (including the aforementioned *S. cerevisiae* and *E. coli*) favour rare, slowly translated codons within the first ~30-50 codons of their genes. Tuller proposed that slower translation creates a translational ramp, which reduces ribosomal traffic jamming, allowing for fewer aborted ribosomes and therefore increased translational efficiency downstream. In this model, the translational cost of a ribosomal collisions outweighs the cost of a slower translation initiation rate (Tuller et al. 2010b).

Translational pausing has also been shown to affect the conformational folding of proteins. A study in 2007 examined synonymous codon changes in a multidrug resistance gene (MRD1), which encodes for an ABC transporter efflux pump to rid the cell of harmful substrates. The synonymous mutations did not affect mRNA or protein levels for MRD1, yet were still associated with altered transport activity. Conformational-sensitive antibodies and trypsin digestions revealed folding differences between mutant and wild type proteins, which was attributed to a translational pause caused by the presence of a rare synonymous codon (Kimchi-Sarfaty, 2007).

Pause elements are not limited to codon-tRNA interactions. Anti-Shine Dalgarno (aSD) motifs are sequences that have high binding affinity to the 3' end of 16S ribosomal RNA and are typically found several nucleotides upstream of start codons to serve as a ribosomal binding sites to initiate translation. However, aSD-like motifs can also be found throughout a transcript and lead

to transient ribosomal pausing due to the mRNA-ribosome interaction. The correlation between aSD-like sequences and ribosomal pausing was extensively studied in *E. coli* by Li in 2012. It was discovered that there was strong selection against aSD-like sequences in protein-coding regions of *E coli*, but not in coding regions for tRNA or rRNA. The consensus aSD sequence in *E. coli* is AGGAGGU; as a result, it was shown that glycine-glycine codon pairs resembling the consensus sequence appear rarely (GGA GGU), while Gly-Gly codon pairs that have the lowest affinity to SD sequences appear the most frequently (GGC GGC). This lead to a proposal that codon usage bias is not shaped by the tRNA pool, but rather shaped by disfavoured SD-like codons, in which case tRNA expression levels followed codon adaptation (Li et al. 2012).

iii. Ribosomal protection

The efficiency of translation can also have its effects on transcript abundance as well. A transcript with high ribosomal density is protected from endoribonucleases, and therefore increases its half-life in the cell. Notably, in a set of regulated operons in *E. coli*, those that demonstrated differential mRNA decay between genes (ie. gene A has a longer half-life than gene B on the same transcript) were shown to be dependent on differences in ribosomal densities (Dar and Sorek 2018). In comparative studies, ribosomal density in *S. cerevisiae* has been shown to be increased by the presence of rare codons (translational ramp hypothesis) (Tuller et al. 2010b) as well as strong mRNA secondary structures (Tuller et al. 2010a). The property of ribosomal protection has also been utilized in this field of research to create ribosomal density profiles to assess the movement (ie. pausing elements) of ribosomes along the transcript. Specifically, ribosome-protected mRNA fragments are isolated and deep sequenced to assess average ribosomal occupancy across the transcriptome with single nucleotide resolution (Ingolia 2014).

iv. mRNA Secondary Structure

mRNA secondary structure has also been shown to have significant effects on gene expression. Notably, in 2009, Kudla engineered a library of synonymous mutation combinations into a green fluorescent protein that was expressed in *E. coli*. Using predicted secondary structure of the transcript, they discovered strong hairpin structures near the start gene (-4 to +32 nucleotides) decreased expression by preventing ribosomes from engaging with the transcript, hence decreasing the rate of translational initiation (Kudla et al. 2009).

However, within a year, Tuller published an extensive analysis of the transcriptome of both *E. coli* and *S. cerevisiae* that disagreed with the claims made by Kudla. From a genomic point of view, Tuller had found a significant association between codon usage and translational efficiency, yet no association for mRNA folding energy and translational efficiency. Folding energy was only found to modulate the strength of association between codon usage and translational efficiency. It was proposed that selection for weak mRNA folding at the beginning of genes works on a global scale; and consequently, folding energy is not a reliable predictor of individual gene expression as Kudla had proposed (Tuller et al. 2010a).

Strong hairpin structures in mRNA transcript have functional purposes as well. First, strong hairpins followed by a string of uracil ribonucleotides act as rho-independent (intrinsic) transcriptional terminators at the 3' end of genes. During transcription, the G-C rich stem loop hairpin structure (7-20 base pairs in length) interacts with a RNA polymerase bound protein, *nusA*, causing the RNA polymerase to stall. The string of uracil ribonucleotides creates a weak interaction between the DNA-RNA duplex and allows for the disassociation of the transcriptional complex (Wilson and Hippel 1995). Transcription terminators can be found at 5' end of coding

regions as transcriptional regulators, such that when bound to specific small regulatory molecules, can change their mRNA secondary structure conformation and allow for transcription to proceed (Millman et al. 2017).

Second, strong hairpin structures can act as protective caps on transcripts to provide protection from endo- and exoribonucleases (Dar and Sorek 2018). In a set of ABC transporter operons, it was shown that genes with strong secondary structure at their 5' and 3' ends have a longer half-life in the cell compared to the neighbouring genes in their operon. In almost all cases for ABC transporters, the gene with the most protection is the substrate binding protein, no matter its position in the operon. The increased half-life allows for more protein production of the substrate binding protein, which typically is required in larger quantities compared to the other proteins in the operon (Dar and Sorek 2018). Despite the obvious functional benefits of strong hairpin structures, we are unaware of any studies that has made associations between these features and synonymous sites under selection.

Today, there is a well-established consensus that synonymous mutations can be under selection. Although strong correlations can be extracted from genome wide analyses, the predominant effects of synonymous mutations can vary not only from organism to organism, but also on a gene to gene basis. There is also a complex dynamic between translational and transcriptional processes which makes elucidating the mechanisms of single synonymous mutants quite difficult. Building off the work done by Bailey and Lebeuf-Taylor, the overarching goal of my research is to help bridge the gap between adaptive fitness and the molecular mechanisms of synonymous mutations.

Chapter 2

Transcriptional effects of adaptive synonymous mutations in *P. fluorescens*

I. Introduction

As mentioned in chapter 1, the framework of my research stems from the research by Susan Bailey, published in Nature Communications in 2014. The two independently evolved synonymous mutations in *Pseudomonas fluorescens*, A15A and G38G, had relative fitness effects of +7.25% (±0.55 s.e.m.) and +8.73% (±0.79 s.e.m.) respectively. Both mutations were on trajectories to fix within the population after only 200 generations, consistent with natural selection of strongly beneficial mutations.

The two evolved synonymous mutations and the one evolved nonsynonymous mutation (A10T) are located within the first 38 codons of the *gtsB* gene, which encodes for a permease subunit of a ATP-binding cassette (ABC) glucose transporter. The transport system is encoded by a four gene *gts* (glucose transport system) operon, which includes: *gtsA*, *gtsB*, *gtsC* and *gtsD* (Fig 1A). *gtsA* is a glucose binding protein responsible for sequestering and transporting glucose in the periplasm towards the ABC permease transporters, *gtsB* and *gtsC*. *gtsD* is a ATP binding protein which hydrolyzes ATP to provide energy for the active transport of glucose through the permease transporters into the cytosol (Fig 1B) (Schneider 2001). The *gts* operon was shown to be induced by the presence of glucose, and results had shown that *gtsB* was a target of selection as the beneficial fitness effects provided by the mutations were lost when grown in alternative carbon sources such as succinate and mannitol. The increased fitness was shown to be caused by increased gene expression through the use of a luciferase reporter. Interestingly, overexpressing a second copy of *gtsB* in the ancestral host was not enough to mimic the fitness effect gained by the synonymous mutations; a beneficial fitness effect was only gained when a second copy of the

entire *gts* operon was overexpressed (Bailey et al. 2014). It was speculated that both *gtsB* and *gtsC* were needed to be expressed in stoichiometric quantities due to the functional properties and interaction of the two channel proteins.

Changes to mRNA secondary structure or CAI alone were not enough to explain the effects of the mutations. Consequently, Bailey tested two possible mechanisms that involved metabolism regulation through protein-mRNA and sRNA-mRNA interactions. The hypothesis was that the synonymous mutations were altering the affinities of these regulatory molecules and thus increasing overall expression. The first mechanism tested was a deletion of a catabolite repression control (*crc*) protein which regulates glucose metabolism in Pseudomonads by binding to target mRNA thus preventing translation. The second was a deletion of the *hfq* gene, a protein required for the activity of many sRNAs that target mRNA transcripts for degradation. Both deletions had no effects on gene expression, and thus the mechanism remained unknown (Bailey et al. 2014).

The first follow-up study aimed to determine the distribution of fitness effects of synonymous mutations in *gtsB*. Lebeuf-Taylor created a library of over 40 single synonymous mutants and found no significant difference in the distribution of beneficial fitness effects between synonymous and non-synonymous mutations (unpublished). Intriguingly, a subset of beneficial synonymous mutations emerged from the fitness assays, indicating that the two evolved synonymous mutants were not entirely unique, and perhaps shared similar mechanisms to confer beneficial fitness effects. *In silico* analyses were done to elucidate possible mechanisms of the synonymous mutations. Despite the large set and extensive range of fitness effects, no significant correlation was found between fitness and any of the following parameters: CAI, tAI, mRNA folding energy, codon position, and transition versus transversions. Once again, the mechanism remained elusive.

As a result, my research focused on a simple question: what is the underlying mechanism responsible for the fitness effects caused by the synonymous mutations? Although the question itself is fairly simple, it serves an important role in the broader context of adaptive evolution. Specifically, my study provides a unique opportunity to uncover the mechanisms driving the emergence of adaptive single synonymous mutations.

The design of my research project follows four main objectives. First, with a library of single synonymous mutants at our disposal, we want to assess the relationship between gene expression and fitness effects for a range of deleterious and beneficial mutations. This requires the incorporation of a yellow fluorescent protein (YFP) bioreporter within the native *gts* operon to serve as a proxy for *gtsB* transcript levels. Second, we want to characterize whether the synonymous mutations are predominantly affecting transcript levels, translational efficiency or a combination of the two. To accomplish this, we created both YFP transcriptional and translational protein fusions after *gtsB* as a proxy of *gtsB* transcript levels and *gtsB* protein levels, respectively. Third, we want to further explore the range of the transcriptional effects across the *gts* operon to single out possible mechanisms. We will incorporate YFP transcriptional fusions across the *gts* operon to test for relative transcript levels at varying distances from the mutations in both upstream and downstream directions. Lastly, we want to investigate mutations on a case by case basis in order to test hypotheses for a specific mechanism. To achieve this, we will systematically introduce additional synonymous mutations that were predicted to disrupt the mechanism and therefore nullify the effect of the original mutation. Overall, my project aims to shed light on the transcriptional and translational effects and mechanisms of adaptive single synonymous mutations.

II. METHODS

i. YFP Fusions in the Native Operon

YFP transcriptional fusions across the *gts* operon were accomplished by allelic replacement protocol previously described in Bailey et al., 2014.

The YFP sequence was PCR amplified from a YFP storage vector (AB87, pUCC3-YFP) and sequences upstream and downstream of the desired YFP insertion locus were amplified from SBW25 template DNA as two PCR products ranging from 400 – 900 base pairs in length. Upstream PCR products included predicted native ribosomal binding sites to ensure proper YFP expression. Forward and reverse primers for all PCR products were engineered with BsaI recognition sites to allow for 4-part seamless ligation between YFP PCR product, the upstream and downstream flanking PCR products and the allelic replacement vector (pAH392) using Golden Gate assembly (Engler et al. 2008). Since the BsaI endonuclease cuts distal from its recognition sequences, both BsaI and DNA ligase could be included with the allelic replacement vector and purified PCR products in a single cloning reaction. The cloning reactions were incubated at 37°C for two hours, 50°C for five minutes, and 80°C for five minutes, then held at 10°C and transformed into *E. coli* DH5αpir following the Inoue method for chemical transformation (Sambrook, Fritsch, & Maniatis, 1989) with selection on ampicillin. Transformations yielded libraries of *E. coli* strains that were used to introduce YFP fusions into *Pseudomonas fluorescens* SBW25. Recombination of each YFP fusion into the chromosome was selected for by two-step allelic replacement, with selection for TcR followed by selection for sucrose resistance as previously described (Bailey et al. 2014). Fusion junctions and *gtsB* mutations were confirmed through PCR gel electrophoresis followed by Sanger sequencing through McGill University and Genome Quebec Innovation Centre.

ii. YFP Fusions at Tn7 Site

Transcriptional, translational, truncated and hairpin modified constructs were accomplished through Golden Gate assembly (Engler et al. 2008) and the use of site specific mini Tn7 transposon (Choi et al. 2005; Choi and Schweizer 2006).

A 2.6 kb PCR product was amplified from template DNA of the desired allele, and included the 346 base pair promoter region of *gtsA*, the open reading frame of *gtsA* and the open reading frame of *gtsB*. This PCR product and the downstream YFP fusion PCR product were seamlessly ligated together into the Tn7 vector (mini-Tn7T-Gm) through Golden Gate assembly as mentioned previously.

The YFP transcriptional fusion after *gtsB* needed additional modifications, due to the stop codon of *gtsB* and start codon of *gtsC* having an 8 nucleotide overlapping sequence. To preserve the *gtsB* sequence and predicted *gtsC* ribosomal binding site, the start codon of the YFP transcriptional fusion was inserted in-frame after the first 4 codons (M-T-S-L) of *gtsC*. Translational YFP fusions had a 6-Glycine ($(GGC)_6$) linker (Chen et al. 2013) sequence between the 2nd last codon (302) of *gtsB* and the 2nd codon of YFP, which removed both the stop codon of *gtsB* and start codon of YFP to create a single peptide. Translational fusions had an additional synonymous mutation (N302N, AAT > AAC) due to the nature of the selected BsaI overhang sequence. The 3 hairpin mutations were introduced into the *gtsB* sequence by first splitting the 2.6 kb PCR fragment into two separate PCR fragments. The seamless junction of these two fragments was located at the G38G hairpin, in which a mutagenic reverse primer containing the 3 synonymous mutations was used. Lastly, the truncated *gtsB* construct was created in a similar manner as the full length *gtsB* fusions, except the sequence from codons 47-294 of *gtsB* were not included in the PCR products. The last 8 codons of *gtsB* were included to preserve the 8 nucleotide

overlapping reading frame as well as to preserve the predicted *gtsC* ribosomal binding site. The truncated *gtsB* construct had two additional amino acid substitutions at S46G and A295G due to the selected BsaI overhang. The 3' UTR sequence of the both transcriptional and translational YFP fusions includes an intrinsic transcriptional terminator.

Tn7 vectors were transformed into *E. coli* DH5αpir, using the previously mentioned Inoue method, then incorporated into SBW25 and verified as previously described in Bailey 2014.

iii. Glucose Induction Assay

Cultures were inoculated from individual colonies (biological replicate) and grown shaking at 28°C in 200 µL minimal (M9) media supplemented with 25.6mM of succinate as the only carbon source. The minimal media contained 48 mM Na_2HPO_4, 22 mM KH_2PO_4, 9 mM NaCl, 19 mM NH_4Cl, 2 mM $MgSO_4$ and 0.1 mM $CaCl_2$. After 24 hours of growth, 20uL of culture was transferred to 180uL of minimal (M9) media supplemented with 212 µM glucose in a transparent 96 well plate. The 96 well plate was incubated for 10 hours static at 28°C in a Tecan Infinite M200 Pro® fluorescent plate reader where optical density (OD) and fluorescence measurements were taken every 10 minutes. OD was measure by absorbance at 595 nm wavelength and fluorescence was measured with 500 nm excitation and 535 nm emission wavelengths. The maximum fluorescence was calculated as the maximum YFP signal (~7 hours), subtracted by the background fluorescence of unmarked SBW25 and then standardized by dividing by the blank corrected OD. Maximum fluorescence values were then divided by the ancestral SBW25 values to determine relative effect sizes.

iv. In silico Predictions

mRNA secondary structure conformations and folding energies were predicted using Mfold online software (RNA Folding Form (version 2.3 energies)) (Zuker 2003) using 42 nucleotide windows (Kudla et al. 2010). Promoter sequences in the *gts* operon were predicted using the Softberry BPROM online software.

v. Luciferase Glucose Reporters

Transcriptional luciferase fusions were constructed using the methods described in Bailey et al. 2014. Additionally, only the SBW25 allele was constructed and was incorporated into the genetic background of the desired alleles. Glucose uptake assays followed the same protocol as the glucose induction assays except luciferase activity was measure with luminescence. Average rate of induction was measured by taking the average slope of luciferase activity for each time point of the first 3 hours of the induction. Time for glucose depletion was calculated by the time point in which the slope of luciferase activity first began to decrease (critical point).

vi. Statistical Analyses

All statistical analyses were performed using R version 3.5.0. (www.r-project.org). Correlation analysis between fitness and YFP expression for the subset of synonymous mutations was performed using a linear model. We tested for significant differences in YFP expression between mutant and wildtype alleles using two tailed T-test assuming equal variance. Threshold for significance was $\alpha=0.05$.

III. Results

i. Creating a Link Between Fitness and Gene Expression

Our first objective is to create a link between fitness and gene expression of *gtsB* for the library of engineered single synonymous mutants. We hypothesize that there will be a strong correlation between transcript levels of *gtsB* and relative fitness based on the trend of the evolved mutants (A10T, A15A, G38G) seen by Bailey. To assess transcript levels of *gtsB*, we introduced the transcriptional YFP fusion after *gtsB* in the native site of a subset of 27 single synonymous mutants (Fig 2A). Relative fitness data was collected previously in head-to-head fitness competitions against the ancestral SBW25, and relative maximum fluorescence was collected from a glucose induction assay. There is a strong significant correlation (R^2=0.691, p-value =1.46 x 10^{-8}, n = 28) between relative fitness and relative maximum fluorescence for the subset of synonymous mutants (Fig 2B). Our data confirm that the fitness effects of the library of synonymous mutations was specific to changes in transcript levels within the *gts* operon. The YFP transcriptional fusions after *gtsB* are considered a proxy for mRNA transcript levels, and therefore the majority of the synonymous mutations appeared to have a polar effect downstream of their location. The strong correlation gave us confidence that YFP fusions could be a used a quantitative tool moving forward in our study.

ii. Comparison of transcriptional and translational fusions

Our second objective is to determine whether the evolved mutations have an additional translational effect on *gtsB*. Studies mentioned in chapter 1 have focused on the impact of synonymous mutations on translational efficiency through codon usage bias, translation ramps, transient pause elements and obstructive hairpins. Although our data thus far points towards a

predominant transcriptional effect by the mutations, we hypothesize that the synonymous mutations could also have an additional beneficial effect on the translational efficiency of *gtsB*. Both A15A and G38G mutations introduce less preferred codons (based on tAI and CAI values) (Lebeuf-Taylor, unpublished), and both mutations are found within the first 50 codons of *gtsB*. We thought it plausible that the introduction of rare codons could promote a translational ramp which would increase translational efficiency. To test this hypothesis we constructed YFP transcriptional fusions after *gtsB* and *gtsB*::YFP translational fusion for ancestral and evolved mutant alleles at the Tn7 site (Fig 4A). YFP transcriptional fusions were used as a proxy for *gtsB* transcript levels and *gtsB*::YFP translational fusions were used as a proxy for *gtsB* protein levels. For a given mutation, the effect on transcript levels should carry over equally to both transcriptional and translational fusions. An increase in relative effect size for the translational fusion can therefore be attributed to an increase in translational efficiency of *gtsB* caused by the mutation.

We determined the relative maximum expression for both transcriptional and translational YFP fusions by means of a glucose induction assay (Fig 4B). For the two evolved synonymous mutants, A15A and G38G, there is no significant ($p>0.06$) difference in effect size between their respective YFP transcriptional and translational fusions relative to SBW25 (A15A: +19.22% ±2.00 s.e.m., +17.77% ±0.72 s.e.m., respectively. G38G: +25.18%, ±1.80 s.e.m., +30.12% ±0.94 s.e.m., respectively). We conclude that these mutations are predominantly increasing transcript levels and have no significant impact on the translational efficiency of *gtsB*. There is a significant ($p = 2.2 \times 10^{-5}$) additional effect of the translational fusion for A10T (+53.47%, ±0.99 s.e.m.) compared to the effect of its transcriptional fusion (+42.72%, ±0.99 s.e.m.) relative to SBW25. However, we are unable to distinguish whether the additional effect of the translational fusion is due to increase

translational efficiency, or simply due to increase protein stability or signaling caused by the nonsynonymous amino acid change.

iii. Assessing the effect of transcription across the gts transcript

Knowing that the main effect of the mutations was on the transcriptional level, our third objective is to further quantify the polar transcriptional effects of the evolved mutations (A10T, A15A, G38G) across the *gts* operon in order to single out possible mechanisms. Since most transcriptional and translational processes are unidirectional, (ie. 5' to 3') we hypothesized that the mutations will only have a downstream effect on transcript levels. However, we are unsure if the mutations will maintain their effect for regions of the transcript after *gtsC*.

To test this, we independently inserted YFP transcriptional fusions before *gtsA*, after *gtsA*, after *gtsB* and after *gtsC* at the native site in the ancestral and three evolved mutant strains (Fig 5A). Relative maximum fluorescence was quantified during a glucose induction assay (Fig 5B). As expected, all three evolved mutants had significantly ($p<5.0x10^{-5}$) increased transcript levels after *gtsB*: A10T (+26.97% ±0.83 s.e.m.), A15A (+8.78% ±1.16 s.e.m.), G38G (+19.28% ±2.27 s.e.m.). The effect of all three mutations is significantly ($p<0.03$) maintained after *gtsC*, however, in comparison to *gtsB*, the effect sizes are significantly ($p<0.02$) smaller for all three mutations: A10T (+11.66% ±1.42 s.e.m.), A15A (+4.49% ±0.67 s.e.m.), and G38G (+8.79% increase ±1.40 s.e.m.). The synonymous mutations have no significant ($p>0.60$) effect on upstream regions of the transcript (before *gtsA* or after *gtsA*). However, the nonsynonymous mutation has a slight deleterious effect on transcript levels upstream of *gtsA* (-5.17% ±0.57 s.e.m. $p<0.0006$), and a slight increase in transcript levels after *gtsA* (+4.49% ±0.48 s.e.m. $p<0.005$).

These results show that all three evolved mutations have strong polar effects on transcript levels. For the two synonymous mutations, the effects are exclusively downstream of the mutation, and notably have the ability to increase transcript levels more than 2.8 kb away from their location. Although the non-synonymous mutation has significant upstream effects, the relative effect size is small, and therefore we are unsure if it is biologically significant. Overall, these results provide our first major insight into the possible mechanisms of the mutations, specifically, the long range downstream effects of the mutations suggest further investigation into processes involved with transcription regulation.

iv. G38G Local Hairpin Manipulation

Our final objective is to test specific mechanisms of the evolved mutations on a case by case basis. Using Mfold, an online software to predict mRNA secondary structure, we were intrigued by the local secondary structure of the G38G mutation. The G38G site is part of a stem region of a relatively strong local hairpin, which is followed by a string of 4 uracil nucleotides, a structure similar to that of a transcriptional terminator (Fig 6) (Wilson and Hippel 1995). Since the G38G mutation (GGC > GGT) weakens the hairpin by changing a stable C-G bond to a weaker T-G bond, we hypothesize that the mutation is decreasing the frequency of premature transcriptional termination events. To test this hypothesis, we systematically chose 3 synonymous mutations to introduce into the local hairpin structure to weaken it. We hypothesize if there was no longer a hairpin structure, premature transcriptional termination would be eliminated, and therefore there would be no difference in YFP expression between the SBW25 and G38G alleles. To test this hypothesis, we created YFP transcriptional fusions after *gtsB* at the Tn7 site for the ancestral and evolved synonymous mutants (A15A and G38G), each with the three additional synonymous

hairpin mutations (denoted as +3HM) to weaken the G38G hairpin (Fig 6). The A15A allele is included as a control because the A15A mutation does not participate in the local G38G hairpin structure; therefore, we predict that the relative effect size of A15A will be unaffected by the introduction of the 3 hairpin mutations.

We assessed maximum YFP fluorescence through a glucose induction assay (Fig 7). Our results showed that the +3HM constructs had an overall deleterious effect; YFP expression for the SBW25+3HM allele was lowered by 51.34% compared to the SBW25 allele. This is not expected as our model predicted a weak hairpin to confer higher levels of gene expression. Second, the relative effect size of the A15A mutation is only slightly affected by the presence of the hairpin mutations (5.97% increase, $p<0.02$). This is partially in line with our prediction, as we did not expect the hairpin mutations to have any interaction with A15A since they were separated by 75 base pairs. Third, relative to SBW25, the effect size of G38G increases 3-fold ($p<3x10^{-7}$) in the presence of the hairpin mutations (+75.16% ±3.25 s.e.m.) compared to without the hairpin mutations (+25.18% ±1.80 s.e.m.). This is also not expected as our model predicts the relative effect size of G38G to decrease with a weakened hairpin structure. Interestingly, when a truncated version of *gtsB* (first 46 codons) is used, the relative effect size of the G38G+3HM construct is significantly ($p<10^{-7}$) lowered to +11.72% (±1.39 s.e.m.). The relative effect size of the A15A+3HM construct remains unaffected by the truncated *gtsB* sequence ($p>0.09$).

Due to the deleterious effects of the 3 hairpin mutations and the fact they do not decrease the relative effect size for the G38G mutation, we conclude that premature transcriptional termination is not the mechanism at work. However, since the effect size of the G38G mutation is significantly decreased when a truncated *gtsB* sequence was used, we are lead to believe that the

proximal downstream region of G38G has an essential role in conferring its effect on transcript levels. A post hoc interpretation of these results is addressed in the discussion.

IV. Discussion

i. Synonymous mutants have polar transcriptional effects

We began the study seeking a high-throughput and reliable tool to measure gene expression for a library of single synonymous mutants in *gtsB*. Due to the number of replicates required and issues with repeatability, quantitative PCR and Western blotting were not feasible for the scope of my project. The YFP bioreporter, however, was an in-house alternative that allowed for fast data collection and accurate repeatability.

Our first objective was to create a link between *gtsB* transcript levels and fitness for the subset of synonymous mutations. We were able to insert a transcriptional YFP fusion after *gtsB* into a library of 27 synonymous mutants. Although the initial library of synonymous mutants spread across the entire length of the *gtsB* gene (302 codons), we were only able to successfully incorporate the YFP fusions into mutants within the first 92 codons. Mutations after the 92nd codon of *gtsB* were subject to homologous recombination with the ancestral *gtsB* sequence during the allelic replacement protocol, and consequently, had substantially lower success rates of retaining their mutated allele.

Due to the possible differential decay of the *gts* operon transcript, the YFP transcriptional fusion after *gtsB* is not a measure mRNA abundance of the *gtsB* region specifically, but rather a general measure of transcript level of the downstream region of *gtsB*. Despite this caveat, there was still a strong correlation between the fitness of the subset of synonymous mutants and their effect on transcript levels downstream of *gtsB* (Fig 2). This finding made logical sense as we would expect an increase in fitness to require not only an increase expression of *gtsB*, but also of its matching subunit permease, *gtsC*, located downstream of *gtsB* (Bailey et al. 2014). Importantly, our results validated that our system is not specific to beneficial mutations but can also explain

deleterious fitness effects of synonymous mutations through decreased levels of gene expression. Also, given the small amount of fitness data for synonymous mutations, we believe this is one of the first sets of data that presents a strong linear correlation between gene expression and fitness effects of synonymous mutations.

69% of the fitness variation was explained by the expression of the transcriptional YFP fusion. The fitness data was collected from competitions using unmarked synonymous mutants (without the YFP fusion); therefore, a portion of the unexplained variance may stem from a fitness cost associated with the presence of the YFP protein in the operon. Evidence against this hypothesis comes from a comparison of growth rates between ancestral SBW25 and YFP marked SBW25: we found no significant ($p>0.20$) difference in growth (OD_{595}).

The most intriguing finding was the consistent downstream polar effect displayed by the synonymous mutations. Polar effects in operons have been studied previously in the 1960s but have been associated with deleterious translational and transcriptional effects of downstream genes due to nonsense mutations and mutations within intergenic regions (Newton et al. 1965; Yanofsky and Ito 1966). To our knowledge, beneficial polar effects of single synonymous mutations has only been identified once before in a study by Kershner in 2016. The synonymous point mutation was identified at the 3' end of a *proB* gene, which was the upstream gene in a *proBA* operon in *E. coli*. The mutation was shown to create a new weak promoter for *proA*, therefore increasing transcriptional levels of the downstream *proA* gene (Kershner et al. 2016). The synonymous mutations we had assessed were all located over 600 base pairs away from start codon of *gtsC*; therefore, it seemed unlikely that our mutants would share the same promoter mechanism based on the physical differences between the two systems.

i. Tn7 site is a reliable substitution for the native site

Upon moving forward, if we wished to make further manipulations to the *gts* operon, such as truncated proteins, early stop codons, sequence deletions and translational fusions, we ran into the risk of creating a non-functional glucose transport system. Since glucose is required to induce transcription of the operon, we also risked having insufficient YFP expression. The use of a site specific mini Tn7 transposon allowed us the ability to insert a second copy of the *gts* operon (functional or not) with the desired alleles and YFP fusions. Each construct therefore had the unaltered, fully functional, ancestral *gts* operon at the native site, which removed possible unaccounted variables such as rate of glucose uptake and growth. It was important to establish that the mutations maintained their effect at the Tn7 site in comparison to the native operon (Fig 3). Relative effect sizes for YFP transcriptional fusions after *gtsB* had no significant difference at the Tn7 versus the native site for the A10T and A15A alleles ($p>0.10$), and there was only a slight 3.96% decrease in effect size for the G38G allele ($p<0.02$). These results provided assurance that the Tn7 site would be a reliable tool to mimic the effects we would expect to see at the native site, without the risk of disrupting glucose intake.

ii. Mutations predominantly act on the transcriptional level

Although our data was presenting strong effects on transcript levels, we still wanted to address whether the mutations were also increasing translational efficiency of *gtsB*. A previous study has estimated translational efficiency as a protein:mRNA ratio (Agashe et al. 2012); in other words, quantity of protein produced per mRNA molecule. Since we encountered issues with western blotting and qPCR, we used *gtsB*::YFP translational fusions as a reporter for *gtsB* protein level and our established YFP transcriptional fusions after *gtsB* as a reporter for mRNA levels.

There are two main factors that contribute to protein expression: transcript levels and translational efficiency. Since the YFP transcriptional fusion is translated as a separate protein, it is only affected by the level of transcript. The *gtsB*::YFP translational fusion, however, is translated as a single protein, therefore is affected by the level of transcript but also by the translational efficiency of *gtsB* (Fig 4a). Therefore, any difference between the relative effect sizes of transcriptional and translational fusions for a given mutation can be attributed to changes in its translational efficiency of *gtsB*.

Needless to say, both evolved synonymous mutants had no significant difference between the transcriptional and translational YFP fusions after *gtsB*. From this we inferred that the predominant effect of the synonymous mutations is on the transcriptional level, and there was no significant effect on translational efficiency (Fig 4b). This agrees with previous data, as supplementation of *gtsB* alone does not confer a fitness advantage (Bailey et al. 2014), therefore we would expect a mechanism in which both *gtsB* and *gtsC* are affected. These results also align with the notion that changes in CAI of a gene need to be extensive in order to see significant changes in translational efficiency (Kudla et al. 2009). It stands to reason that these fine-tuning translational mechanisms may not be apt to explain the large effects in gene expression that are present in the case of our adaptive single synonymous mutations.

iii. Mutations have strong polar downstream transcriptional effects

Focusing in on transcriptional effects, we wanted to further assess the polarity of the synonymous mutations by assessing transcript levels across the *gts* operon. Through the use of YFP transcriptional fusions, the three evolved mutants we tested, A10T, A15A and G38G, had shown an increase in transcript levels after *gtsB*, and remarkably, all three mutations maintained a

significant effect after *gtsC* as well (Fig 5). Although the magnitude of the effect significantly decreased after *gtsC*, the three mutations demonstrated an ability to increase transcript levels over 2.8 kb away from their location, spanning two full genes. We believe this is the first evidence of multiple beneficial single synonymous mutations displaying such a long range polar effect.

However, there was concern that these downstream transcriptional effects were solely due to a positive feedback loop. Hypothetically, the synonymous mutations could first cause an increase in glucose uptake, for which the increased amount of glucose in the cell could then increase transcriptional induction of the entire operon. If that were the case, we would expect to see a significant increase of transcript levels of across the entire operon.

Intriguingly, there is no significant effect of the evolved synonymous mutations on the transcript for upstream regions of *gtsB*. This leads us to assume that increasing levels of *gtsA* is not necessary to confer a beneficial fitness effect. Bailey had shown that supplementation of the entire *gts* operon was able to confer a fitness advantage, while supplementation of *gtsB* alone was not. However, what was not tested was whether supplementation of *gtsA* was truly necessary. It would be useful to know whether supplementation of other combinations, such as *gtsBCD*, *gtsBC*, or *gtsCD*, would be sufficient to confer the same fitness advantage. To further address this hypothesis, we turned to a recent paper that had extensively studied a maltose ABC-transporter operon, *malEFG*, in *E coli*. *malE*, the maltose binding protein is needed in higher quantities compared to the permease subunits, *malF* and *malG*, and this need is addressed by differential decay patterns throughout the transcript. The *malF* and *malG* regions of the transcript are rapidly degraded compared to *malE*, which is protected from exoribonucleases at its 3' end (intergenic region between *malE* and *malF*) by strong mRNA secondary structure. Independent of where the substrate binding protein is located in the operon, they found four other *E. coli* ABC transporter

operons in which the substrate binding protein was always the most stabilized gene. The need for increased levels of substrate binding protein may stem from the advantage of being able to sequester more sugars in the periplasm of the cell, out of reach from competitors.

That being said, the *gts* operon has similar physical properties to the *malEFG* operon. Notably, *gtsA* and *malE*, both substrate binding proteins, are the most upstream protein in the operon, and both have strong mRNA secondary structures at their 3' end in the intergenic region between the next downstream gene (*gtsB*, *malF*). This lead us to predict that, stoichiometrically, *gtsA* may already be in excess due to increased protection from degradation, and therefore follow an independent degradation pattern compared to *gtsB* and *gtsC*. This would therefore place a selective pressure on increasing expression levels of downstream genes (as seen in our data), as they may be the limiting factors in the transport system, especially in an experimental setting where glucose is the only available carbon source. Although more work would need to be done to validate this hypothesis, it provides an explanation as to why *gtsB* specifically is a strong target for selection and why increased transcript levels upstream of *gtsB* are not needed to confer a fitness advantage.

iv. Speculated Mechanism: Premature Transcriptional Termination

Two likely explanations as to how the mutations were increasing downstream transcript levels are 1) increasing transcriptional efficiency/accuracy or 2) decreasing the rate of transcript degradation. One possible way to increase transcriptional efficiency is to decrease the number of aborted transcriptional events. As stated before, intrinsic transcriptional terminators rely on sharp hairpins followed by a string of uracil ribonucleotides (Wilson and Hippel 1995). Therefore, one hypothesis was that hairpin structures within the transcript could mimic transcriptional terminators and cause a premature termination. Another hypothesis was that synonymous substitutions are

creating sites for new internal promoters (Kershner et al. 2016). Both hypotheses would agree with our data as the mechanism would only display a downstream effect, as transcription is performed in a 5' to 3' direction.

The hypothesis that the mutations are decreasing the rate of degradation of the transcript hypothesis could be explained in two ways: 1) strong secondary structures formed by the mutations offer protection from either endo- or exoribonucleases, or 2) the mutations are altering ribosome movement/occupancy, which confers a higher density of ribosomes and thus an increase in protection from endoribonucleases to areas of possible high susceptibility. Although exoribonucleases in bacteria were thought to only degrade transcripts from 3' to 5', a study in 2007 had identified an exoribonuclease in *Bacillus subtilis*, RNase J, that degrades in a 5' to 3' direction. They had also shown that aSD-like motifs within the transcript lead to significant ribosomal pausing and thus downstream protection by blocking the 5'end of the transcript from RNase J (Mathy et al. 2007). However, whether a 5' to 3' exoribonuclease exists in Pseudomonads is currently unknown, and thus the mechanism seemed unlikely. Protection via higher ribosome density also seemed unlikely, as significant changes in ribosome occupancy would have likely manifested itself through changes in translational efficiency, which we did not detect in our *gtsB*::YFP translational fusions.

v. G38G hairpin is not essential to confer the transcriptional effect

Since no significant correlation was previously found between fitness and possible mechanisms, we deduced that we would have to investigate the mutations on a case by case basis. Our hypothesis that the G38G mutation conferred its effect by weakening its local hairpin structure (Fig 6) and thus decreasing the frequency of premature transcriptional termination events decrease

was rejected. Since the three synonymous hairpin mutations had increased the relative effect size of the G38G mutation, we had to rethink the possible mechanism at play. The fact the effect size of G38G significantly decreased when a 46 codon truncated *gtsB* sequence was used lead us to believe that the effect of the G38G mutation was dependent on proximal downstream elements in the *gtsB* sequence (Fig7).

vi. Proposed Mechanism: Internal Promoters

In a post hoc analysis to attempt to interpret the results of the G38G hairpin, we used the Softberry BPROM online software to search the *gts* operon for internal sigma 70 bacterial promoter sequences. Sigma 70 proteins are the most abundant transcriptional factors in bacteria. They bind to specific DNA sequences often located -10 and -35 nucleotides upstream of transcriptional initiation sites to enable the binding of RNA polymerase (Potvin et al. 2008). The top hit out of five promoter sequences was a recognizable promoter 86 base pairs upstream of the *gtsA* start codon. Interestingly, the second hit was a promoter sequence spanning codons 36-46 of *gtsB*, for which the G38G mutation locus is two base pairs downstream of the -35 box (Fig 8A). Although BPROM states 80% accuracy and specificity for its results (Solovyev 2011), at this point we propose that the G38G synonymous mutation strengthens a weak internal promoter sequence within *gtsB*, leading to increased transcript levels downstream of the mutation.

There are three lines of evidence that corroborate this hypothesis. First, one of the three synonymous hairpin mutations is located within the predicted -10 box sequence and lowers the promoter score when inputted into BPROM. This is a possible explanation for the the overall deleterious affects of the hairpin mutations. Second, the predicted -10 box sequence extends to the exact location of the truncated *gtsB*/YFP fusion junction. The significantly decreased effect size

of G38G+3HM allele in the presence of the truncated *gtsB* sequence could be explained by interference with the promoter sequence. Third, aside from A15A and G38G, the synonymous mutant the highest fitness (+7.98%) is Y39Y (TAC > TAT). This mutation occurs one codon away from G38G (GGC > GGT), and when inputted into BPROM, returns a new hit for an even stronger promoter sequence (Fig 8B). The new promoter sequence, pushes the predicted -10 box further upstream, so that now both G38G and Y39Y loci are within the -10 box sequence. The -10 box consensus sequence is TATAAT, and since both mutations introduce thymine nucleotides, it provides an explanation how these mutations are strengthening the internal promoter. A phylogenetic analysis by Lebeuf-Taylor revealed that the Y39Y allele is present in 24/77 closely related Pseudomonads, in contrast to an analysis done by Bailey which showed the G38G allele present in only 1/23 closely related Pseudomonads (Bailey et al. 2014), despite both mutations having comparable beneficial fitness effects.

As stated earlier, this mechanism of a weak internal promoter has been identified once before by Kushner in 2016, in which a synonymous mutation within the weak promoter sequence (-10 box) increased expression of a downstream gene by 2-fold. Further investigation is required to validate this hypothesis, which would include transcription start site predictions using 5'RACE as well as manipulating the predicted promoter regions to see if they have strong beneficial or deleterious effects. Additionally, it should be tested whether the resultant transcript of an internal *gtsB* promoter would translate into a truncated yet functional *gtsB* protein. The remaining three promoter hits from the BPROM software were located twice in *gtsA* and once in *gtsC*, thus the promoter hypothesis maybe suitable for the G38G and Y39Y mutations, but at this time cannot convincingly be used to explain the effects of the A15A mutation or other beneficial synonymous mutations outside of the predicted promoter region.

Lastly, to test whether the mutations had increased glucose uptake, we designed an assay to test the length of time for the mutants to deplete all available glucose. We used a common transcriptional luciferase fusion after *gtsB* at the Tn7 site (*gtsAB*-lux) and introduced it into the genetic backgrounds of SBW25, A15A and G38G. The common luciferase fusion contained the SBW25 *gtsB* allele, therefore luciferase expression is unaffected by the mutations and only affected by the presence of glucose in the cell. It is presumed that when glucose is depleted, induction of luciferase and its activity is stopped. We therefore used the critical point of the luciferase curve as the time point in which glucose was deemed depleted. We performed a glucose induction assay for SBW25, A15A and G38G and had two main findings. The first finding is that there was no significant ($P>0.28$) difference between the rate of induction of the *gtsAB*-lux operon between SBW25 and A15A or G38G. This confirmed our prediction that the mutations would have no effect on the expression of the luciferase fusion, and that it was only affected by the presence of glucose. Our second finding had shown that the two synonymous mutations depleted glucose significantly ($P<0.05$) quicker than SBW25 (Fig 9).

Therefore, our final proposal for a mechanism to explain the fitness effects of the synonymous mutations is as follows. A weak internal promoter within *gtsB* is strengthened by both G38G and Y39Y synonymous mutations. The mutations increase transcription initiation events downstream of their location to increase transcript levels of *gtsC*, *gtsD*, and a truncated yet possibly functional *gtsB* sequence. The increased transcript levels carry over to produce more protein subunits, in turn increases the number of functional glucose channels in the inner membrane. This allows for a more efficient intake of glucose, thus provides a fitness advantage by 1) providing the cell with more energy and 2) depleting glucose in the environment which limits the growth of its competitors.

Chapter 3

Conclusions

I. Summary of Findings

The findings of our study clearly show that the fitness effects of both beneficial and deleterious synonymous mutations can be correlated to levels of gene expression. Specifically, these synonymous mutations primarily act on the transcript level and have shown to have strong polar effects. Our evolved synonymous mutations have little effect on the translational efficiency of *gtsB*, despite an overwhelming amount of evidence that synonymous mutations are under selection for their ability to fine-tune translational efficiency.

My research presents clear evidence of the ability of synonymous mutants to have long range polar effects, increasing transcript levels over 2.8 kb away from their location. Beneficial polar effects of synonymous mutations are not well documented, but our findings showcase the ability of synonymous mutants to act on a much higher level of regulation, compared to conventional mechanisms in which the effects of the mutations are localized to one gene. Lastly, we provide speculative evidence of two individual beneficial synonymous mutations that can strengthen weak internal promoter sequences to produce strong polar downstream effects. Interestingly, at this point the mechanism does not appear to be a suitable explanation for all the mutations within *gtsB*, leading us to believe that there is the possibility more than one mechanism is at play.

The next steps of this research project would be to further investigate the internal promoter hypothesis at the G38G and Y39Y site and investigate other potential likely candidate mutations. Additionally, our research could benefit from an RNA-seq analysis of the *gts* operon after glucose induction. The data would provide a high-resolution profile of transcript abundance patterns across the operon for the different mutations, which could help to single out possible mechanisms for

each one. Lastly, although we have shown there is a strong correlation between the expression of the YFP fusions and fitness, we believe that it would be important to demonstrate that introducing additional synonymous mutations that alter transcript levels also have significant and predictable effects on fitness.

II. Applications

We believe there are a few important takeaways from our research. First, many of the genome-wide and high through-put studies mentioned in chapter 1 have provided evidence to explain mechanisms in which synonymous mutations can affect transcript levels and translational efficiency. Most of these mechanisms, such as codon usage bias, weak 5' mRNA secondary structure, and ribosomal pausing, appear to be ways to fine-tune transcriptional and translational processes which have been selected for over millions of years (Boucher et al. 2016). In adaptive settings, where single synonymous mutations can have considerably large effects on gene expression and fitness, these mechanisms often fall short in fully explaining the fitness effects. We propose the notion that large effect synonymous mutations often work on much higher levels of transcriptional and translation regulation, which at this point have not been fully characterized. Although these large effect synonymous mutations may seem rare, and in cases, idiosyncratic, they serve important biological roles in the context of adaptive evolution.

Second, our study focuses on the regulation of an ABC transporter operon, which exists across all domains of life, has a wide variety of structures and functions, and has been shown to share many common regulatory properties. In archaea and bacteria, 1 to 3% of their genomes encode for subunits of ABC transporters (Wilkens 2015). Further characterization of the *gts* operon and the mechanisms in which synonymous mutations can affect its regulation could have wide spread applicability to the characterization of other ABC transporter families.

Third, although it was not addressed directly in my results or discussion, we proposed that there may be an underappreciated dual-role of nonsynonymous mutations. We often associate the effects of nonsynonymous mutations with their changes in protein function, but what if the main effect of our evolved nonsynonymous mutant, A10T, came from "synonymous-like mechanisms?" After all, the A10T mutation had shown to carry significant downstream effects along the *gts* operon in a similar manner to the two synonymous mutations, which cannot be explained by protein function alone. Second, Bailey (unpublished) had shown that a truncated (nonfunctional) A10T *gtsB* protein could still maintain its positive effects on transcript levels. Therefore, as the mechanisms behind synonymous mutations become more clear, we believe that it is important to remember that those mechanisms should be inclusive to nonsynonymous mutations as well.

Lastly, despite that most research cited so far has focused on molecular biology and evolutionary concepts, synonymous mutations have also made their way into the clinical research as well. In the clinical field, many synonymous SNPs (sSNP) have been identified in association with human disease. sSNPs have been discovered in genes associated to diseases such as Treacher–Collins syndrome, X-linked infantile spinal muscular atrophy, Seckel syndrome and cystic fibrosis. Through association studies, sSNPs have even been found in genes currently unlinked to the mechanisms of certain diseases, but have been shown to help predict disease outcomes (Sauna and Kimichi-Sarfaty 2011). That being said, a full characterization of the possible mechanisms in which synonymous mutants can affect gene expression has the possibility to help characterize human disease. In the future, this may have impacts on treatment and personalized medicine, especially in cases where sSNPs are associated with genes involved in drug delivery and metabolism.

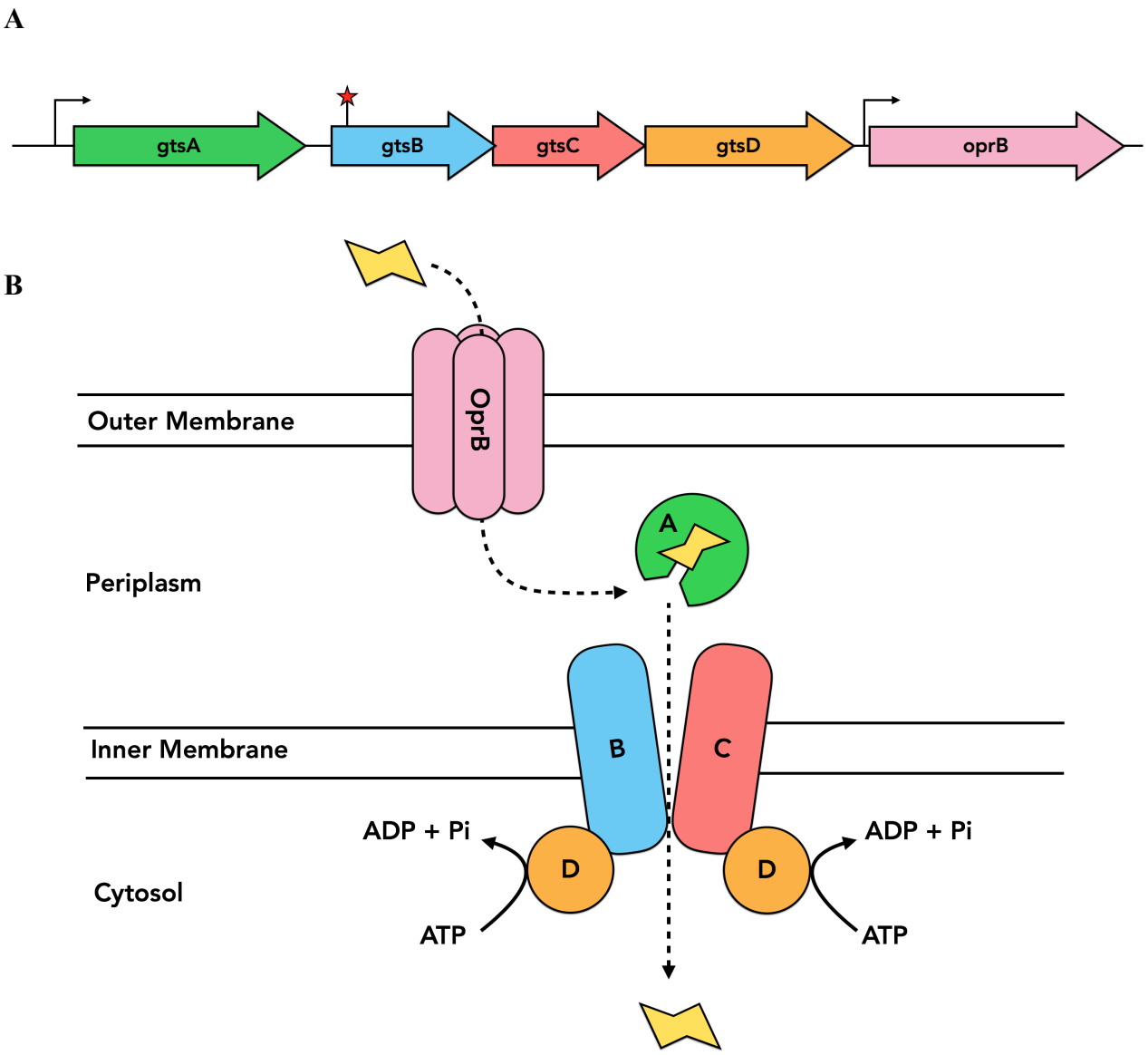

Figure 1. Schematic of gts operon and predicted transport system configuration A) Schematic of the *gts* and oprB operon with red star indicating general location of the three evolved mutants. Thin bent arrows represent promoter regions as predicted by SoftBerry BPROM (Solovyev, 2011). B) Schematic of the predicted *gts* transport system configuration (Schneider, 2001) and the movement of glucose into the cytosol. *gtsA*: glucose binding protein. *gtsB*, *gtsC*: permease subunits. *gtsD*: ATP binding protein (B).

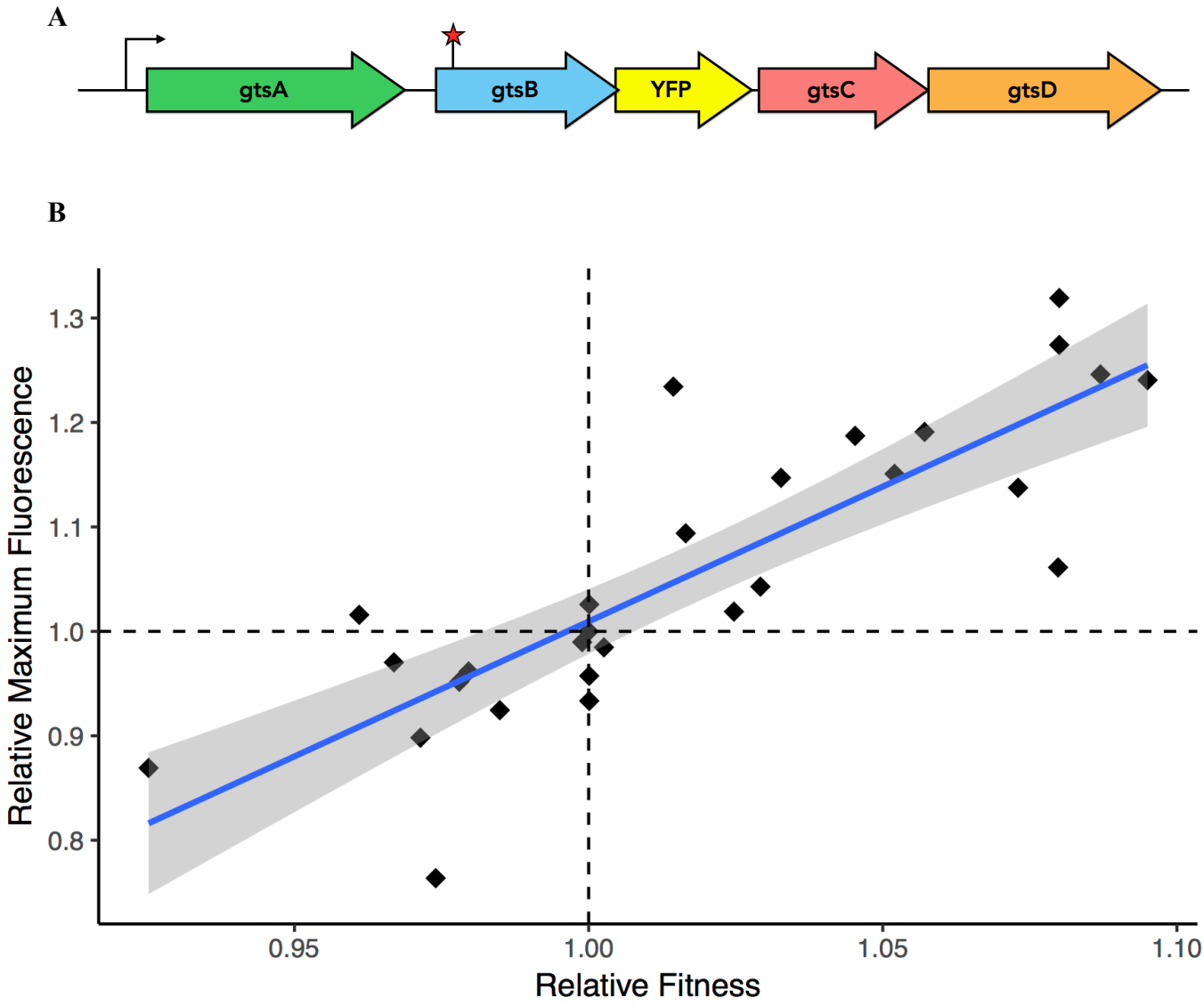

Figure 2. Correlation of relative fitness and relative maximum fluorescence between synonymous mutants. A) Schematic of YFP transcriptional fusion after *gtsB* in the native operon with red star indicating general location of evolved mutants. B) Maximum fluorescence was obtained from a glucose induction assay. Fluorescent signal is used as a proxy for transcript levels. Blue line represents linear regression with the 95% confidence interval shaded in grey. Dotted line indicates relative fitness and relative maximum fluorescence values of SBW25. Adjusted R^2= 0.69. p-value =1.46 x 10^{-8}. N=28.

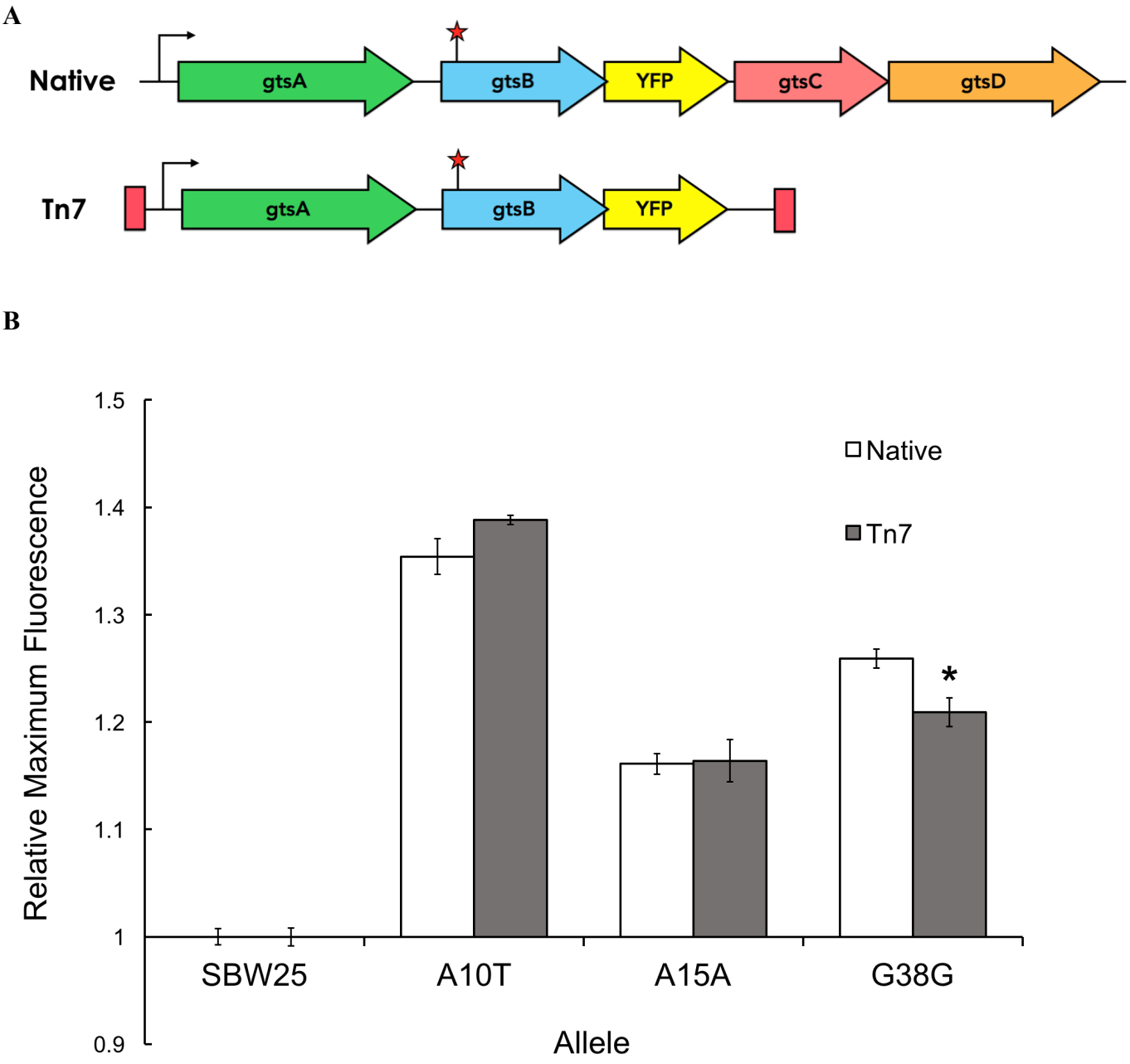

Figure 3. Comparison of YFP expression at native and Tn7 site. A) Schematic of YFP transcriptional fusions after *gtsB* at the native site an Tn7 site. Red rectangles indicate Tn7 junctions. B) Comparison of relative maximum YFP expression for transcriptional fusions after *gtsB* during glucose induction assay between the native site and Tn7 site ±S.E.M. presented. N=[5,14]. Significance with respect to native site: * $p<0.05$.

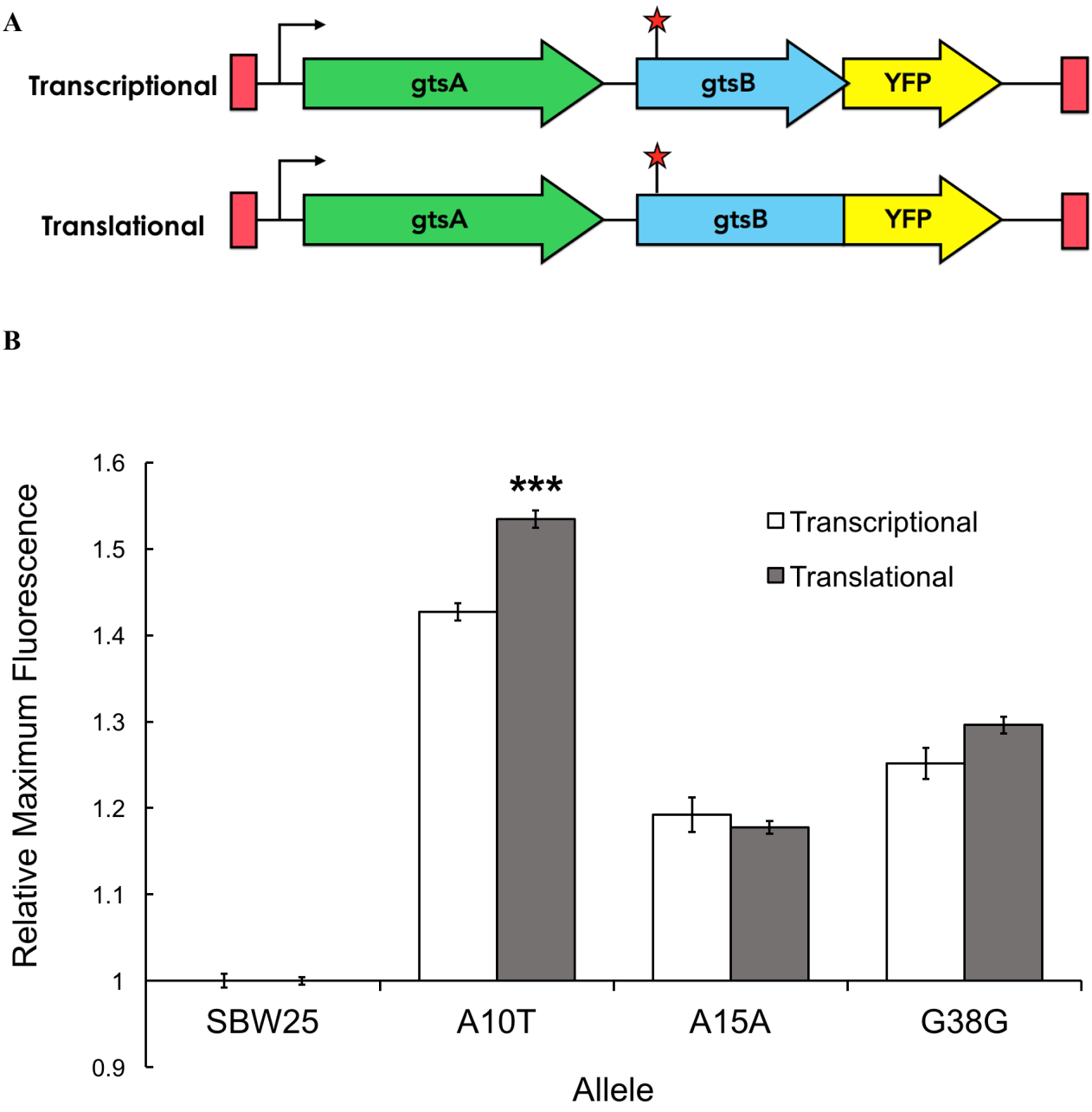

Figure 4. Comparison of transcriptional and translational effects of the evolved mutants

A) Schematic of the transcriptional and translational YFP fusions at the Tn7 site. B) Comparison of maximum YFP expression between transcriptional and translational YFP fusions for the evolved mutants at the Tn7 site during a glucose induction assay. ±S.E.M. presented. N=[5,14]. Significance with respect to transcriptional fusion: *** $p<0.001$.

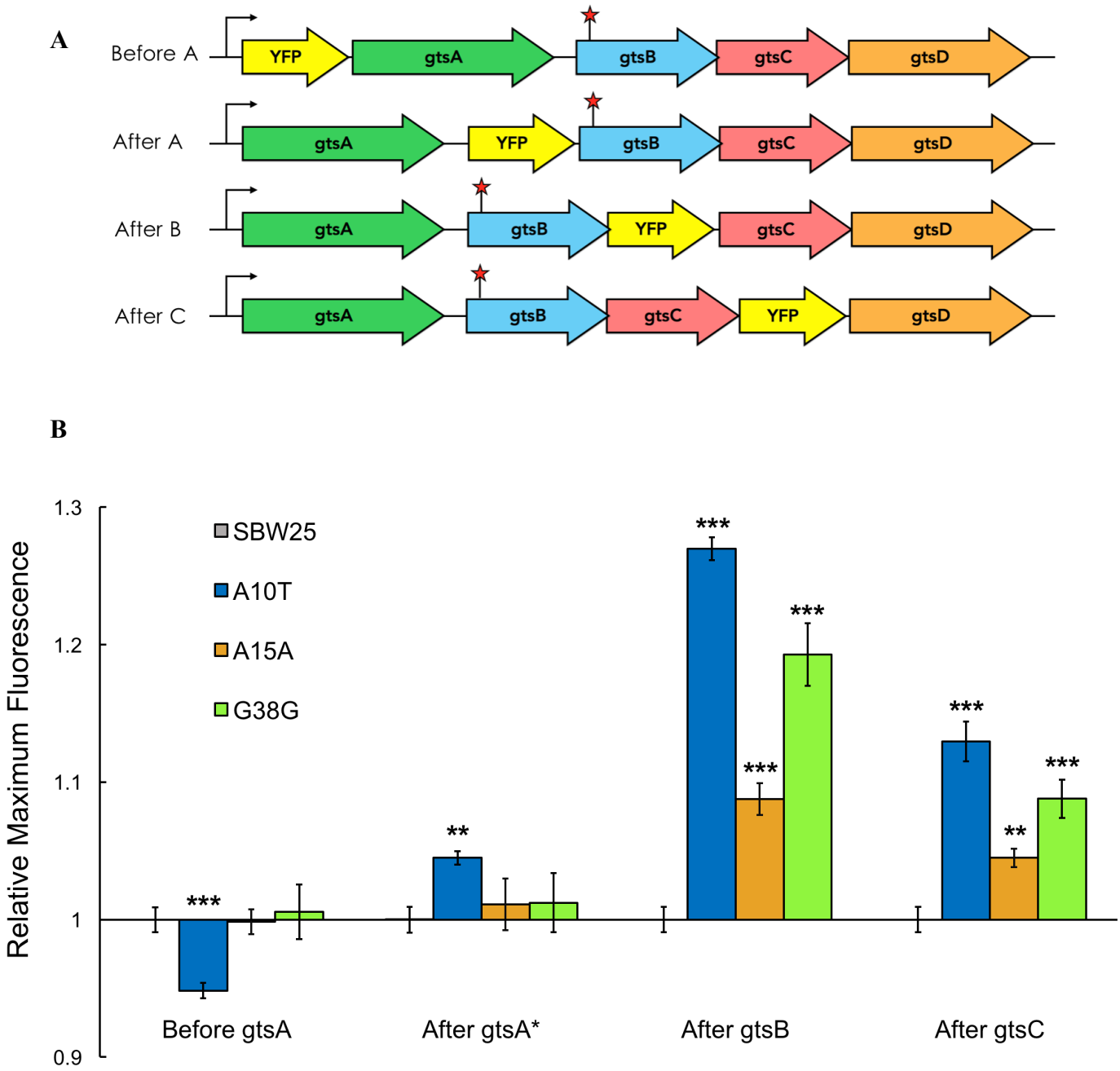

Figure 5. Analysis of transcriptional effects across the gts operon for the evolved mutants.

A) Schematic of the four different YFP transcriptional fusions within the native *gts* operon. B) Maximum fluorescence of the YFP transcriptional fusions across the *gts* operon relative to SBW25. After *gtsA** data was collected from a reconstruction at the Tn7 site. ± S.E.M presented. N=[4,12]. Significance with respect to SBW25: ** $p<0.01$, ***$p<0.001$.

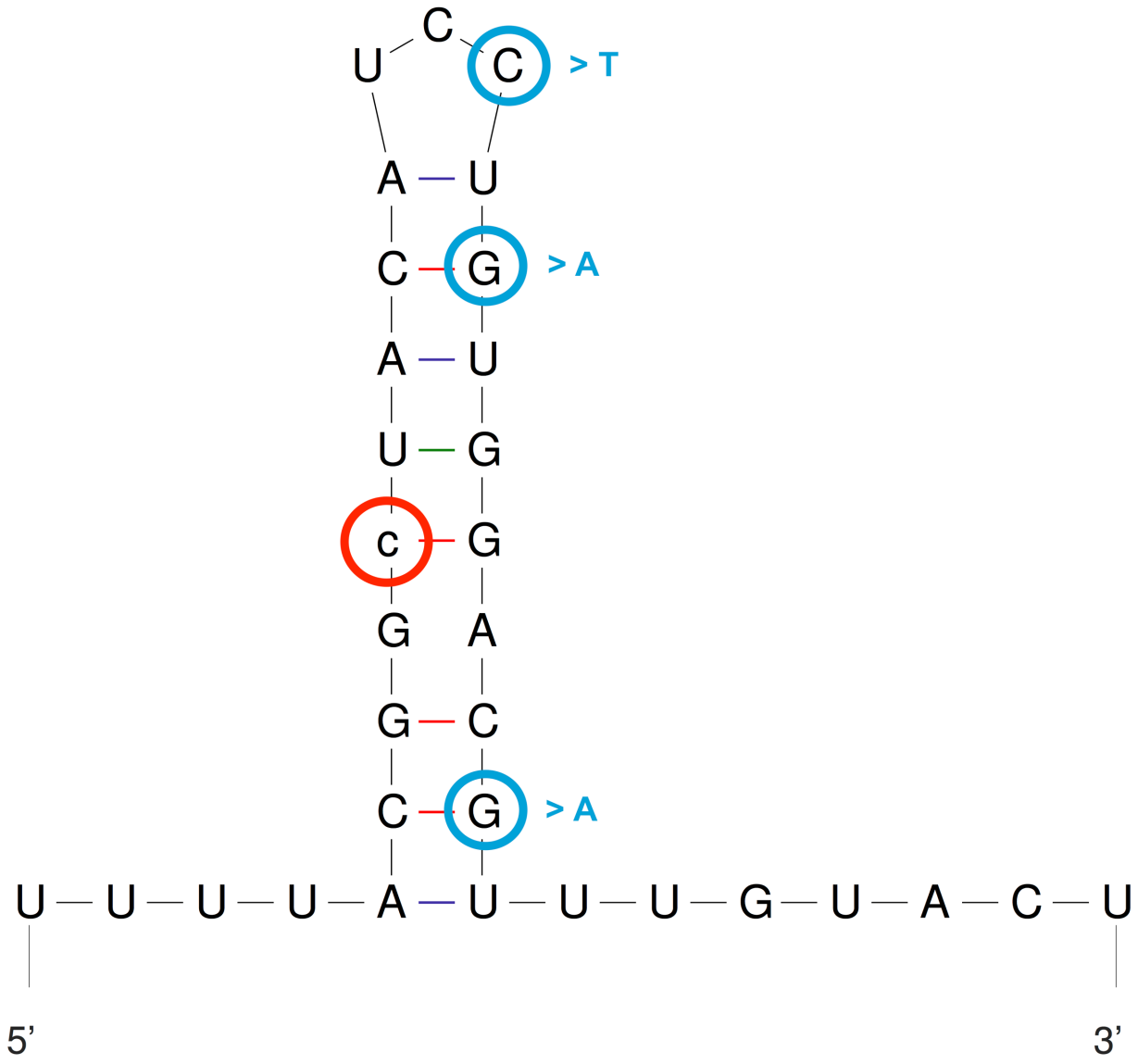

Figure 6. Schematic of the local G38G hairpin and location of hairpin mutations.

Ancestral SBW25 sequence shown as predicted by Mfold. Red circle indicates the location of the G38G mutation (GGC > GGT) and blue circles indicate the location of the three synonymous Hairpin Mutations (+3HM).

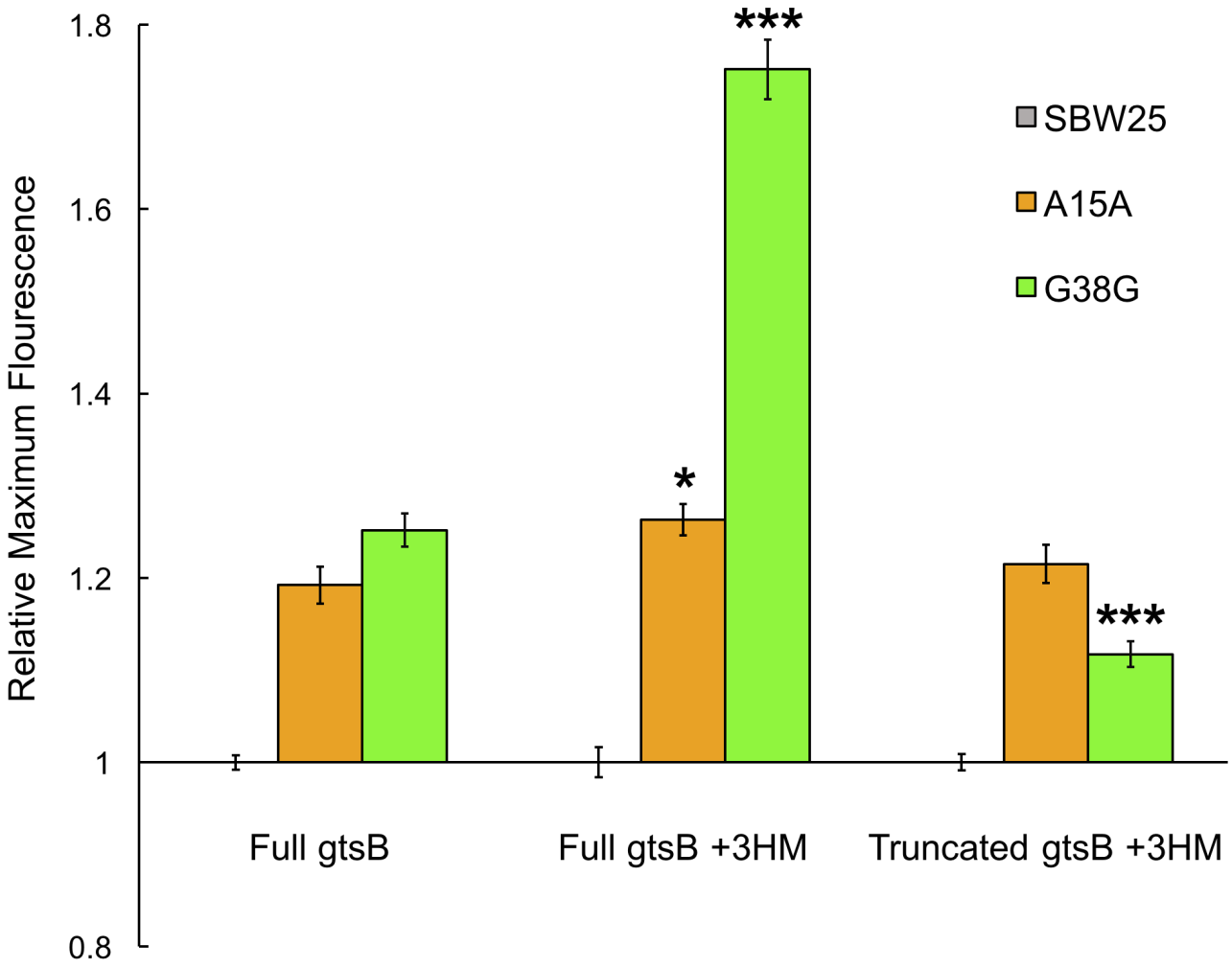

Figure 7. Effect of synonymous hairpin mutations on transcript levels for evolved synonymous mutants. Relative maximum YFP expression for SBW25, A15A and G38G mutations, in the presence and absence of the three synonymous hairpin mutations (+3HM), and in the presence and absence of the truncated *gtsB* sequence. Not shown are the absolute levels which vary significantly for each set of constructs (see results). ±S.E.M. presented. N=[6,14]. Significance with respect to Full gtsB: *p<0.05, ***p<0.001.

A

B

Figure 8. Predicted weak internal promoter sites in gtsB. A) Internal sigma 70 promoter sequence in *gtsB* as predicted by BPROM. B) New predicted sequence when Y39Y (TAC > TAT) mutation is introduced. -35 and -10 boxes sequences are underlined and highlighted yellow. G38G mutation location is highlighted red, locations of the 3 synonymous hairpin mutations are highlight blue and the location of the junction for the truncated *gtsB* YFP transcriptional fusion is highlighted orange.

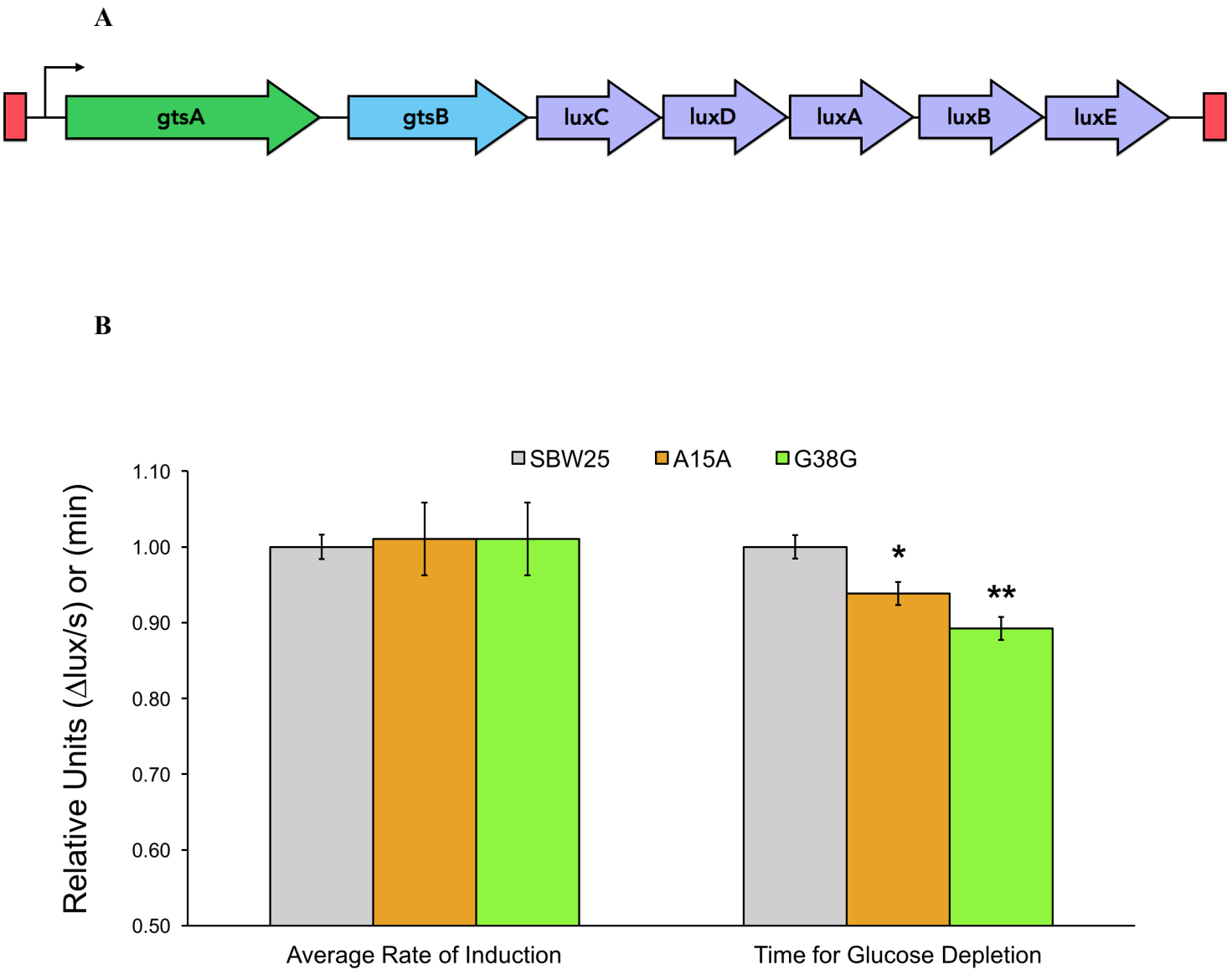

Figure 9. Comparison of glucose uptake rates for evolved synonymous mutants. A) Schematic of the luciferase reporter at the Tn7 site with SBW25 *gtsB* allele. Luciferase genes are not to scale. B) Relative rate of luciferase induction as measured by Δlux/min for the first 3 hours of the glucose induction. Relative time taken to deplete glucose as measured by luciferase activity critical point) ±S.E.M. presented. N=3. Significance with respect to SWB25: * $p<0.05$, ** $p<0.01$.

LITERATURE CITED

Agashe, D., N. C. Martinez-Gomez, D. A. Drummond, and C. J. Marx. 2013. Good codons, bad transcript: Large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Mol. Biol. Evol. 30:549–560.

Agashe, D., M. Sane, K. Phalnikar, G. D. Diwan, A. Habibullah, N. C. Martinez-Gomez, V. Sahasrabuddhe, W. Polachek, J. Wang, L. M. Chubiz, and C. J. Marx. 2016. Large-effect beneficial synonymous mutations mediate rapid and parallel adaptation in a bacterium. Mol. Biol. Evol. 33:1542–1553.

Bailey, S. F., A. Hinz, and R. Kassen. 2014. Adaptive synonymous mutations in an experimentally evolved Pseudomonas fluorescens population. Nat. Commun. 5. Nature Publishing Group. 5:4076 DOI: 10.1038

Bataillon, T., S.F. Bailey. 2014. Effects of new mutations on fitness: insights from models and data. Ann. N.Y. Acad. Sci. 1329:76-92

Boucher J.I., Bolon D, Tawfik D. 2016. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. Protein Science. 25(7):1219-1226.

Canale, A.S., et al. 2018. Synonymous Mutations at the Beginning of the Influenza A Virus Hemagglutinin Gene Impact Experimental Fitness. J Mol Biol. 430(8):1098-1115.

Cannarozzi, G., N.N. Schraudolph, M. Faty, P. von Rohr, M.T. Friberg, A.C. Roth, P. Gonnet, G. Gonnet, Y. Barral. 2010. A role for codon order in translation dynamics. Cell. 141(2):355-67.

Chamary, J. V, J. L. Parmley, and L. D. Hurst. 2006. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat. Rev. Genet. 7:98–108.

Chao, Y., L. Li, D. Girodat, K.U. Förstner, N. Said, C. Corcoran, M. Śmiga, K. Papenfort, R. Reinhardt, H.J. Wieden, B.F. Luisi, J. Vogel. 2017. In Vivo Cleavage Map Illuminates the Central Role of RNase E in Coding and Non-coding RNA Pathways. Mol Cell. 65(1):39-51.

Chen, X., J. Zaro, W. Shen. 2013. Fusion Protein Linkers: Property, Design and Functionality. Adv Drug Deliv Rev. 65(10): 1357-1369.

Choi, K. H. et al. A Tn7-based broad-range bacterial cloning and expression system. Nat. Methods 2, 443–448 (2005).

Choi, K. H. & Schweizer, H. P. mini-Tn7 insertion in bacteria with single attTn7 sites: example Pseudomonas aeruginosa. Nat. Protoc. 1, 153–161 (2006)

del Castillo, T. et al. 2007. Convergent peripheral pathways catalyze initial glucose catabolism in Pseudomonas putida: genomic and flux analysis. J. Bacteriol. 189, 5142–5152.

Dar, D., S. Rotem. 2018. Extensive reshaping of bacterial operons by programmed mRNA decay. PLoS Genet. 14(4): e1007354.

dos Reis, M., R. Savva, and L. Wernisch. 2004. Solving the riddle of codon usage preferences: A test for translational selection. Nucleic Acids Res. 32:5036–5044.

Duret, Laurent. 2000. tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. Genome Analysis. 16(7):287-289.

Engler, C., R. Kandzia, and S. Marillonnet. 2008. A one pot, one step, precision cloning method with high throughput capability. PLoS One 3.

Hunt, R. C., V. L. Simhadri, M. Iandoli, Z. E. Sauna, and C. Kimchi-Sarfaty. 2014. Exposing synonymous mutations. Trends Genet. 30:308–321. Elsevier Ltd.

Ingolia, N.T. 2014. Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet. 15(3):205-13.

Kershner, J.P., S.Y. McLoughlin, J. Kim, A. Morgenthaler, C.C. Ebmeier, W.M. Old, S.D. Copley. 2016. A Synonymous Mutation Upstream of the Gene Encoding a Weak-Link Enzyme Causes an Ultrasensitive Response in Growth RateJ. Bacteriol. 198(20): 2853-2863.

Kimchi-Sarfaty, C., J. M. Oh, I.-W. Kim, Z. Sauna, A. M. Calcagno, S. Ambudkar, and M. Gottesman. 2007. A "Silent" Polymorphism in the MDR1 Gene Changes Substrate Specificity. Science (80-.). 315:525–529.

Kimura, M. 1968. Evolutionary Rate at the Molecular Level. Nature. 217: 624–626.

Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267: 275–276.

Knöppel, A., J. Näsvall, and D. I. Andersson. 2016. Compensating the fitness costs of synonymous mutations. Mol. Biol. Evol. 33:1461–1477.

Kudla, G., A. W. Murray, D. Tollervey, and J. B. Plotkin. 2009. Coding-sequence determinants of gene expression in Escherichia coli. Science. 324(5924):255–258.

Lawrie, D. S., P. W. Messer, R. Hershberg, and D. A. Petrov. 2013. Strong purifying selection at synonymous sites in D. melanogaster. PLoS Genet. 9:33–40.

Li, G.-W., E. Oh, and J. S. Weissman. 2012. The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature 484:538–541.

Lind, P. A., O. G. Berg, and D. I. Andersson. 2010. Mutational robustness of ribosomal protein genes. Science (80-.). 330:825–827.

Mathy, N., L. Bénard, O. Pellegrini, R. Daou, T. Wen, C. Condon. 2007. 5'-to-3' exoribonuclease activity in bacteria: role of RNase J1 in rRNA maturation and 5' stability of mRNA. Cell. 129(4):681-92.

Millman, A., D. Daniel, M. Shamir, R. Sorek. 2017. Computational prediction of regulatory, premature transcription termination in bacteria. Nucleic Acids Research. 45(2):886–893.

Newton, W.A., J.R. Beckwith, D. Zipser, S. Brenner. 1965. Nonsense mutants and polarity in the lac operon of Escherichia coli. J Mol Biol. 14(1):290-6.

Potvin, E., F. Sanschagrin, R.C. Levesque. 2008. Sigma factors in Pseudomonas aeruginosa. FEMS Microbiology Reviews. 32(1): 38–55.

Sanjuán, R., A. Moya, and S. F. Elena. 2004. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc. Natl. Acad. Sci. U. S. A. 101:8396–401.

Sauna, Z. E., C. Kimichi-Sarfaty. 2011. Understanding the contribution of synonymous mutations to human disease. Nature Reviews Genetics. 12:683–691.

Schneider, E. 2001. ABC transporters catalyzing carbohydrate uptake. Res. Microbiol. 152, 303–310.

Sharp, P. M., W. Li. 1986. An Evolutionary Perspective on Synonymous Codon Usage in Unicellular Organisms. J Mol Evol. 24:28-38.

Sharp, P. M., and W. Li. 1987. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281– 1295.

Sharp, P. M., E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe, and F. Wright. 1988. Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. Nucleic Acids Res. 16:8207–8211.

Solovyev, V., A. Salamov. 2011. Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies. Nova Science Publishers. 61-78

Tuller, T., Y.Y. Waldman, M. Kupiec, E. Ruppin. 2010a. Translation efficiency is determined by both codon bias and folding energy. PNAS. 107(8):3645-3650.

Tuller**,** T., A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, and Y. Pilpel. 2010b. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141:344–354. Elsevier Ltd.

Wilkens, S. 2015. Structure and mechanism of ABC transporters. F1000 Prime Reports. 7(14): doi:10.12703

Wilson, K.S., P.H. von Hippel. 1995. Transcription termination at intrinsic terminators: The role of the RNA hairpin. Proc Natl Acad Sci USA. 91(19):8793-7.

Yanofsky, C., J. Ito. 1966. Nonsense Codons and Polarity in the Tryptophan Operon. J. Mol. Biol. 21, 313-334.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496–503.

Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31:3406–3415.