MASSEY UNIVERSITY
COLLEGE OF SCIENCE
SCHOOL OF NATURAL AND COMPUTATIONAL SCIENCES

Ph.D. Thesis

FOX GLACIER 2021

MARKÉTA VLKOVÁ

MASSEY UNIVERSITY
COLLEGE OF SCIENCE
SCHOOL OF NATURAL AND COMPUTATIONAL SCIENCES

Natural variation in bacterial gene regulation

A thesis submitted in partial fulfilment of the requirements for the degree of

Ph.D.

in

Microbiology & Genetics

Submitted by

Markéta Vlková

Supervisors: Olin Silander & Tim Cooper **Fox Glacier 2021**

Abstract

It has been over 160 years since Charles Darwin set out the theory of evolution by natural selection. This theory is broadly accepted these days. However, it is still not completely understood how natural selection shapes particular cell mechanisms and behaviours. There is a limited research about selection acting on gene regulation.

To address the questions about how selection shapes gene regulation we used a collection of environmental *E. coli* isolates. We quantified the genetic variability of 605 promoters within this collection of highly diverged strains. We then selected ten promoters (*aldA*, *yhjX*, *lacZ*, *aceB*, *mtr*, *cdd*, *dctA*, *ptsG*, *purA*, and *tpiA*) which highly differ in their genetic variability to analyse their phenotypic variability. We used fluorescent reporter assays with flow cytometry to measure changes in gene expression with high-throughput and at single cell resolution. In order to discern natural selection acting on gene regulation we compared phenotypes from segregating promoter variants, which have been subject to natural selection and random promoter variants that have never been subject to natural selection. We generated the random variants using PCR random mutagenesis. Beside focusing on the changes in the overall expression (i.e., transcription and translation), we examine selection acting on transcription only. This we achieved be implementing self-cleaving ribozyme insulation.

In this thesis we showed that natural selection towards high plasticity and low noise is common among regulated *E. coli* promoters. We also verify that the self-cleaving ribozyme RiboJ activity is highly effective and that this genetic tool can be used to detect changes in transcription alone. Utilizing the RiboJ we were then able to detect both directional and diversifying selection acting on *lacZ* promoter.

This thesis thus broadens the knowledge about natural selection acting on gene regulation and provides a new insights into how promoters are shaped in nature by selection, including some most well-characterized bacterial promoters. This work also demonstrates a new application of RiboJ ribozyme that has not been published before.

Aknowledgement

First, I would like to thank my main supervisor Dr. Olin Silander for all the help, time and encouragement while at the same time for providing me a lot of freedom not only to explore new approaches. I also very much appreciate your support when I raised my wish to move from the big Auckland city to remote Fox Glacier village for the final stage of my PhD. I am also grateful for all your comments and discouragement from going deeper into the rabbit holes when there is an easier path.

I further thank my co-supervisor Prof. Tim Cooper for his valuable comments, suggestions and for always finding alternatives when I hit a wall. My big thanks go also to Dr. Nikki Freed for all her support in the lab from the beginning of my PhD and her invaluable comments when it comes to presentations.

I also thank Andrea Sajuthi for being an awesome friend with whom I can talk about science whole day. I hope we will stay in the same country a bit longer this time. I value a lot the time Stella Pearless and Bhargava Morampalli invested when helping me with a few experiments. And I thank all the members of the Silander lab for accepting me among them.

To my partner Jan I am incredibly grateful for always cheering me up, distracting me from work and making me enjoy the beauties of the country into which I moved and last, but by far not least for surviving Auckland.

Table of Contents

List of Abbreviations

Here I present an alphabetical list of used abbreviations for easier orientation in the following text.

API	average pairwise identity
ATP	adenosine triphosphate
cAMP	cyclic adenosine monophosphate
CRP	cAMP receptor protein
DNA	deoxyribonucleic acid
EPEC	enteropathogenic *Escherichia coli*
FFL	feed-forward loop
gDNA	genomic DNA
GFP	green fluorescent protein
HGT	horizontal gene transfer
HK	histidine kinase
IGR	intergenic region
IPTG	isopropyl-β-D-thiogalactoside
LB	lysogeny broth
mRNA	messenger RNA
NAP	nucleoid-associated protein
OD	optical density
ORF	open reading frame
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
PSS	proportion of segregating sites
RBS	ribosome binding site
RNA	ribonucleic acid
RR	response regulator
RT-qPCR	reverse transcription quantitative PCR
SNP	single nucleotide polymorphism
STEC	Shiga-toxin producing *Escherichia coli*
TF	transcription factor

TMG	thiomethyl-β-D-galactoside
TSS	transcription start site
UTR	untranslated region
UV	ultraviolet radiation

List of Figures

List of Tables

Chapter 1

Introduction

Bacteria exhibit genotypic and phenotypic variability even at the level of species and strain. This variability is continuously shaped by natural selection, as the cells that are better adapted to the current conditions are more likely to survive and proliferate. In the case of unicellular organisms, adaptation to environmental changes is even more critical as they have limited ability to escape. Cells must survive for example osmotic stress, nutrient-poor environments, UV radiation. They have evolved multiple ways to overcome such conditions, many of which rely on responding to such environmental changes with physiological adaptations. These are all mediated by sensing external signals and translating them into internal responses. However, the responses to the same signal might differ among bacteria. These differences could be driven by other selection forces acting simultaneously, or could be a result of previous evolution in response to a similar environment. In any case, cells which do not manage to adapt to the new conditions are less likely to leave offspring.

It is thus essential for a living cell to sense what is happening in the environment and react accordingly and in a timely manner. If the timescale of an appropriate reaction should be on the order of seconds, there is no time for changes in transcription or translation. In that case, the cell fate depends on the proteins or other small molecules it has already synthesised. One example of a fast response to a signal is chemotaxis. When a motile bacterium is in a gradient of an attractant (e.g., nutrients such as sugars or amino acids) or repellent (e.g., toxic metabolites), it can swim towards or away from the higher concentration of the attractant or repellent, respectively. In *E. coli*, the change from a random bacterial walk to active chemotaxis is mediated by the protein CheY. Its phosphorylation depends on sensing of external signals and determines the probability of flagella rotating clockwise (i.e., cell rotation) and switching into counter-clockwise motion (i.e., straight run) (Shimizu et al. 2010, Mears et al. 2014).

On the other hand, if it is only necessary to react in minutes or hours, responses can be mediated through translational or transcriptional changes. Sometimes rapid responses can be detrimental, especially in the case of extensive physiological reprogramming, which may consume a lot of energy. In these cases, it might be more profitable for the cell to not respond at once. Cells might even benefit from not responding at all for some time, for example, if the signal triggering costly response is likely to be transient. To manage and orchestrate many responses to multiple signals cells have evolved a network of extra- and intracellular signalling pathways.

As I focus here on transcriptional responses, the following text covers mostly this process and describes the effects of different mechanisms on the speed of the cell response, phenotypic plasticity, transcriptional noise, and sensitivity. Response times depend on the topology of the circuit, the signals that bacteria have observed previously (memory), and any predictive behaviour that the bacteria are capable of. Phenotypic plasticity relates to the magnitude of changes in expression in response to changes in the environmental signals. Memory refers to heritable changes in gene expression without simultaneous changes in DNA primary sequence. The transcriptional noise is understood as variability in gene expression among individual cells with the same genetic background. Below I explore each of these properties in more detail to establish the current state of the field. I focus on *E. coli* as it is a well-studied model organism, however little is known about non-laboratory isolates, which may be adapted to survive and proliferate outside their hosts (Byappanahalli and Fujioka 2004, Ishii et al. 2006, Somorin et al. 2016).

1.1 Signal recognition

The first step of cellular response is the sensing of internal and external cues. There are several means by which bacteria can detect such signals, and these are described below. However, there is little known about how these different signal sensing mechanisms affect the cell response dynamics and whether they differ in terms of speed, sensitivity, phenotypic plasticty or noise.

1.1.1 Direct signalling pathways

Direct signalling occurs when a signal molecule is imported through a cell membrane into the cytosol, or is produced by the cell and subsequently interacts with regulatory proteins. This is often the case for genes involved in carbohydrate catabolism and amino acid biosynthesis (Charlier et al. 1992, Weickert and Adhya 1992, Pittard 1996, Wheatley et al. 2013). In these cases, the nutrient import into a cell is utilized also for signalling besides consumption purposes. Physical signals such as UV or heat are also considered direct, although no active import takes place in the signalling.

1.1.2 Indirect signalling pathways

It is not always beneficial for a cell to import signal molecules into the cytosol as some of these might be toxic or too big (e.g., host-pathogen cell-cell interactions). In order to sense environmental cues without the necessity to import them, bacteria possess transmembrane sensory proteins. Upon binding of a signal molecule by extracellular domain, the protein changes its conformation transferring the received signal into the cell interior.

A classic example of indirect signalling is the two-component system. This consists of a transmembrane sensory kinase which is autophosphorylated at its histidine residue after receiving a signal from the environment (**Figure 1.1a**). To be able to phosphorylate itself the kinase requires ATP or other phosphate donors. Next, the phosphorylated kinase transfers the phosphate group to its partner regulatory protein enabling its activity, most often DNA-binding and transcriptional regulation (Lynch and Sonnenburg 2012, Gao and

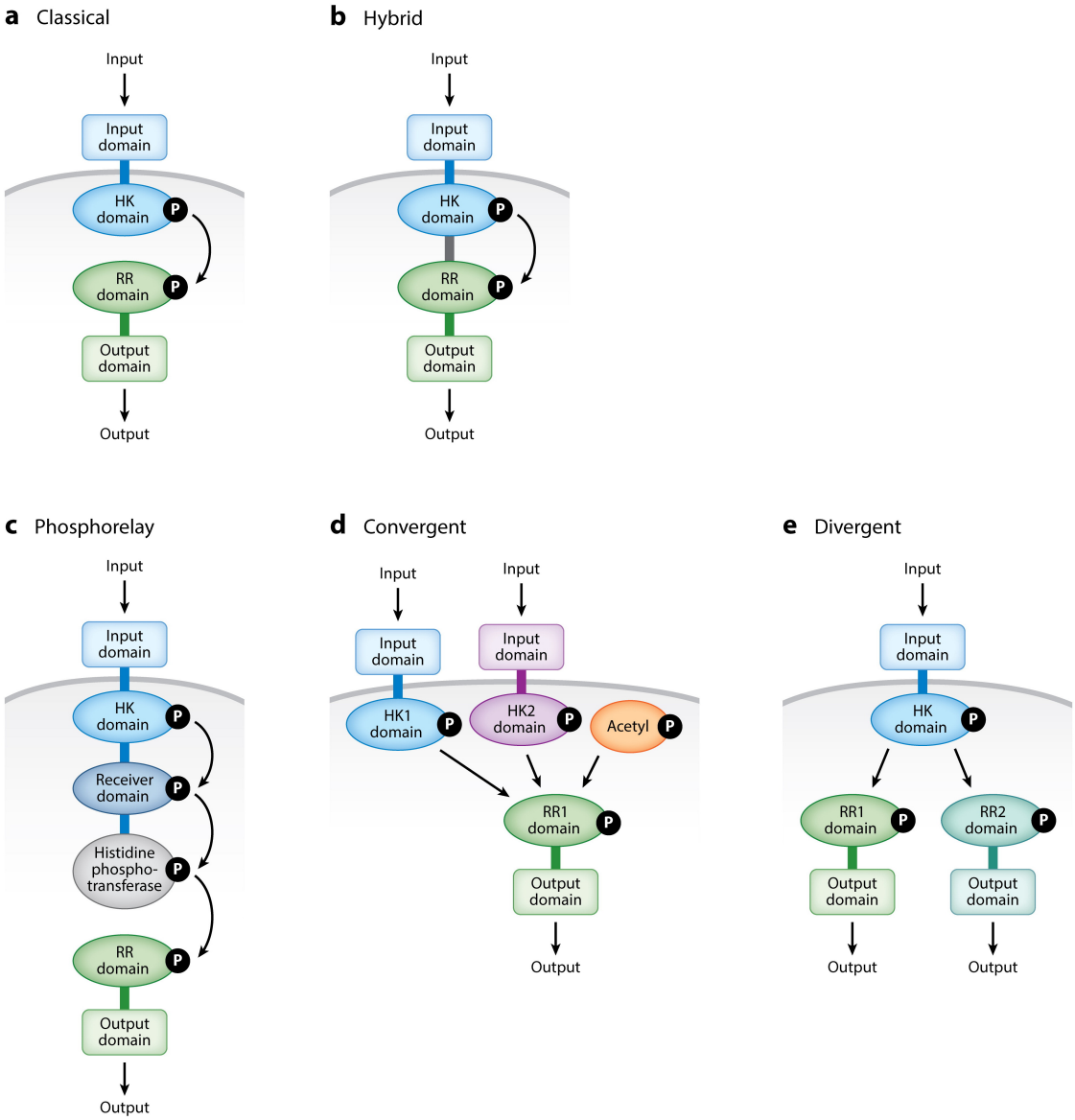

Figure 1.1: **Two-component signalling pathways. a)** histidine kinase (HK) autophosphorylates in response to a signal and subsequently transfers the phosphate to a response regulator (RR) which generates an output; **b)** a hybrid system comprises both components (HK and RR) into a single protein; **c)** in phosphorelay the phosphate is transferred multiple times before reaching its final RR; **d)** some RR can be activated by several HK including metabolites such as acetyl phosphate; **e)** one HK might phosphorylate multiple RR generating various outputs. (reproduced from (Groisman 2016))

Stock 2015, Cui et al. 2018). This protein receives the phosphate to its aspartate residue and might act as both gene repressor and/or activator. Alternatively, the regulatory protein can modify biochemical activity of target proteins or RNAs instead or on the top of the change in gene expression (Shu and Zhulin 2002, Chambonnier et al. 2016). Hybrid two-component systems also exist, and are usually located in the cell membrane with their sensory domains in periplasmic space (**Figure 1.1b**) (Lynch and Sonnenburg 2012, Hirano

et al. 2013).

Another version of two-component system is the phosphorelay (**Figure 1.1c**). In this case, multiple steps of phosphate transfers occur before the final phosphorylation of target regulator. These multiple steps allow for easier regulation by other signals as the phosphorelay can be silenced at any of the intermediate steps during the transmission (Perego and Brannigan 2001, Groisman 2016).

In some cases, multiple different signals are detected by the same sensory kinase or converge into the same regulatory protein (**Figure 1.1d**). This results in similar outputs in response to various signals (Kaczmarczyk et al. 2014, Chambonnier et al. 2016). In contrast, a divergent signal transmission results in multiple responses to the same signal (**Figure 1.1e**) as one sensory kinase is able to phosphorylate aspartate domains of different acceptor proteins (Mika and Hengge 2005, Groisman 2016).

1.2 Signal processing

After a signal is received by the cell, it triggers changes in the activity of proteins and/or gene expression. Control of protein activity helps to respond to rapid environmental changes. Translational regulation is a useful tool for controlling the differential production of proteins coded within the same operon and thus transcribed on the same mRNA (Dar and Sorek 2018). Transcriptional regulation might be considered as the most economical response as it inhibits the synthesis of products which are not needed at the moment. This saves resources and energy for synthesis of desired proteins. As the latter type of regulation is the focus of this work, I outline the important aspects of this process in more detail below.

1.2.1 Regulation of transcription initiation

The initiation of transcription is the first step of gene expression and is highly regulated. The initial step is RNA polymerase binding to the promoter sequence, which contains all motifs: the -35 element, the extended -10 element, the -10 element, the discriminator region, the UP element and the core recognition element. These elements mediate the interaction of the promoter and RNA polymerase, and their proximity to consensus sequences influences the strength of the promoter. The rate of transcription initiation is proportional to the total amount of produced mRNA as long as an early transcription termination does not occur (Kennell and Riezman 1977, Iyer and Struhl 1996). This step is also regulated by transcription factors. The activity of these factors corresponds to signals the cell acquires from the internal or external environment. Although some factors control only one promoter, the majority of at least *E. coli* promoters is regulated by more than one transcription factor (Karp et al. 2018).

Repression

Repression of promoters is often based on the placement of an obstacle at or near the -10 and -35 elements, which blocks RNA polymerase binding to the promoter. This is achieved through various mechanisms. The simplest one is repressor binding to the operator which overlaps one or both of those promoter elements (Brent and Ptashne 1981)

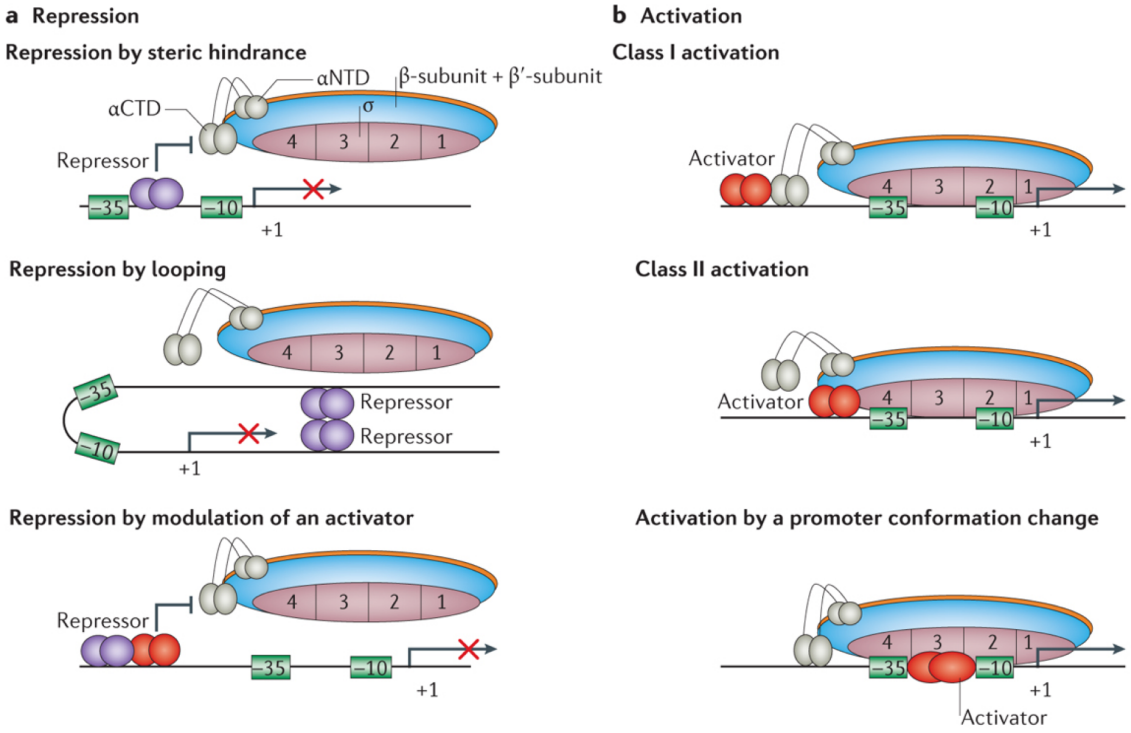

Figure 1.2: **Regulation of transcription initiation by repressors and activators.** (reproduced from (Browning and Busby 2016))

(**Figure 1.2a**). Other promoters have distant operators outside -10 and -35 elements, but bound repressors cluster creating a DNA loop which also constitutes a blockage (Semsey et al. 2004) (**Figure 1.2a**). A more complex and indirect way of repression is by inhibiting activators (described below) - i.e., repressors acting as anti-activators (Søgaard-Andersen and Valentin-Hansen 1993). Because some promoters require activation for transcription initiation, anti-activation might be sufficient for complete repression (**Figure 1.2a**).

The activity of specific promoters can be controlled gradually. Such promoter sequences have arrays of operators, and the amount of bound repressors affects the strength of the repression. Alternatively, some promoters can be recognized by multiple repressors which usually do not affect each other (El Qaidi et al. 2009).

For instance, arginine acts as a co-repressor of its own biosynthesis (**Figure 1.3a**). When a bacterium does not have enough of this amino acid available for protein production the transcription of arginine genes is active (Charlier and Glansdorff 2004, Caldara et al. 2006). On the other hand, when arginine is abundant in the culture medium, and thus in the cell, there is no need to make more of it. The cell has ArgR repressor present in the cytosol, but ArgR itself cannot bind to DNA and inhibit transcription in the absence of arginine (Clark 2005, Caldara et al. 2006). However when there is enough arginine in the cell, it binds to ArgR and as a co-repressor inhibits expression of arginine structural genes by binding to an appropriate promoter (Charlier et al. 1992, Charlier and Glansdorff 2004, Clark 2005).

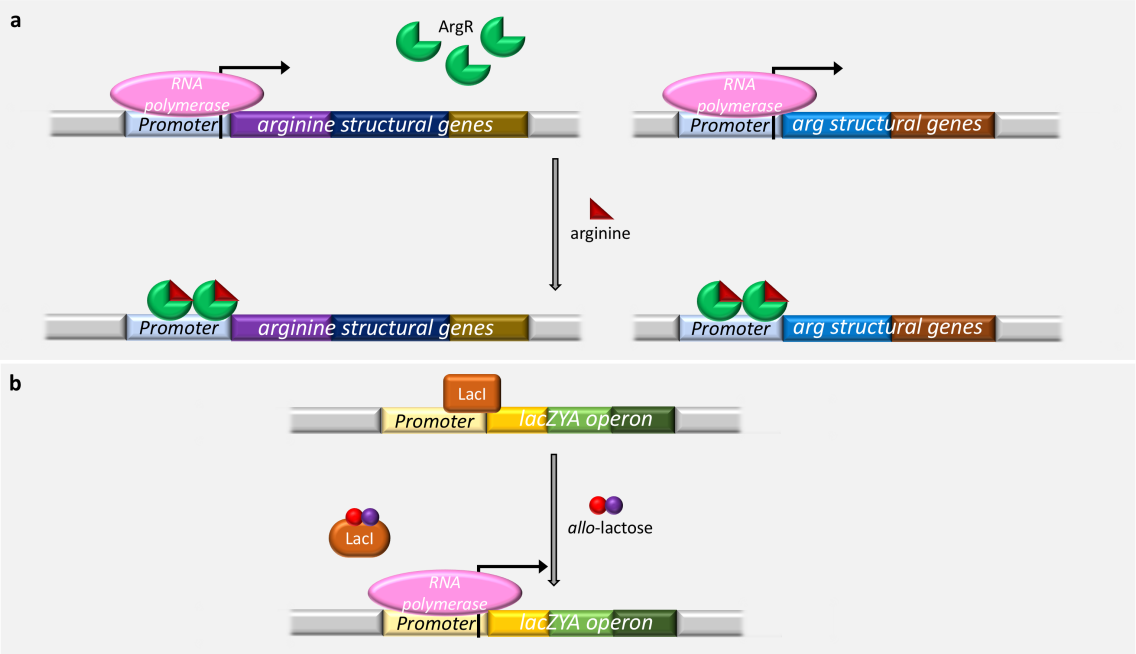

Figure 1.3: **Scheme of repression and activation. a)** expression of arginine structural genes is co-repressed by arginine itself, because without it ArgR repressor cannot bind to the DNA; **b)** if *lac* operon anti-repressor (allolactose) is present it binds to LacI repressor and induces *lac* operon transcription.

Activation

Activation enables particular gene transcription or increases the strength of RNA polymerase binding to the promoter. Similar to repression, some activators interact directly with RNA polymerase but others act indirectly as anti-repressors most often by releasing bound repressors and preventing their further binding to the operator (Frederix et al. 2011). Besides affecting sequence specific repressors, some indirect activators also affect binding of nucleoid-associated proteins (NAPs; described more in the next section), altering access for RNA polymerase to the promoter sequence (Santana et al. 2001). Direct activators mostly facilitate the interaction of RNA polymerase and promoter. Three mechanisms of direct activation are usually described: class I activation, class II activation, and activation by conformational change.

Class I activators bind upstream of -35 element and interact with the α subunit of RNA polymerase (Ushida and Aiba 1990) (**Figure 1.2b**). Operators of class II activators in turn overlap with -35 promoter element and the activators interact with a given σ factor of RNA polymerase (Igarashi et al. 1991) (**Figure 1.2b**). Moreover, bound class II activators prevent binding of the RNA polymerase α subunit to the preferred position right upstream of -35 element. Class I and class II activation thus can co-occur on the same promoter when two different activators are required to trigger the transcription (Lloyd et al. 2002).

The last classical activation mechanism causes a conformational change of the sequence between -10 and -35 elements. RNA polymerase binding to these promoters without an activator is weakened or impaired due to an inappropriate distance between -10 and -35 elements. This is also where the operators of such activators are located (**Figure 1.2b**). A

bound activator then brings the elements into the optimal position for RNA polymerase recruitment (Heldwein and Brennan 2001).

Gene regulation has been best studied in the lac operon of the model organism *E. coli* MG1655 K12. To avoid wasteful production of the enzyme hydrolysing lactose (LacZ), if no lactose is available, *E. coli* expresses lac repressor (LacI) (Hudson and Fried 1990). However, *lac* operon transcription is desired if lactose is present in the media, and this is ensured by lac operon inducers. Allolactose - *lac* operon inducer - is imported by lactose permease into the cell where it binds directly to the repressor LacI. Lactose has to be modified to allolactose by β-galactosidase (LacZ) to act as a *lac* operon inducer (Jobe and Bourgeois 1972, Wheatley et al. 2013). Allolactose then induces lac operon gene indirectly by acting as an anti-repressor. LacI binding to the promoter DNA is then released triggering *lacZYA* expression (**Figure 1.3b**) together with binding of general regulatory protein CRP (Hudson and Fried 1990, Clark 2005). CRP levels rise when the availability of a preferred carbon source (e.g., glucose) drops. Low levels of CRP thus prevent lactose utilization favouring glucose consumption even when lactose is available.

1.2.2 Regulation of mRNA elongation and transcription termination

After a successful transcription initiation, the elongation of mRNA follows until termination occurs. Even though the main transcription regulation occurs at the level of transcription initiation, elongation and termination can also be modulated, ranging from the repression of elongation (Monsalve et al. 1996), through elongation pausing and backtracking (Mustaev et al. 2017) to transcription attenuation and anti-termination (Naville and Gautheret 2009).

1.2.3 Transcription network topology

For a better understanding of the whole transcription network regulation, it is important to look at how individual transcriptional motifs affect the speed of regulatory circuits, phenotypic plasticity, and transcriptional noise (cell-to-cell variability). The speed in transcriptional response is usually characterised by a rise-time (or response time), i.e., a time when the concentration of a gene product reaches its half maximal concentration after gene induction. While the effects of different signal sensing mechanisms (described above) on the transcriptional response dynamics (e.g., speed, plasticity, noise, sensitivity) are not well studied, the ways in which network motifs affect these aspects are better understood.

Promoter strength and protein lifetime

The stronger a promoter the more easily and often RNA polymerase binds to it and the more protein is produced per time unit. The strength of the promoter is determined by its nucleotide sequence and binding of transcription factors increases or decreases RNA polymerase affinity to it. On the other side of this equilibrium is the degradation of the product, as its stability in certain conditions determines its lifetime. A functional protein is also diluted during cell division, which further reduces its concentration in a cell.

Feedback loops

A feedback loop is a transcriptional network motif in which the products affect their own production (**Figure 1.4a** and **Figure 1.4b**). The influence of the output on its own transcription might be direct or indirect if multiple effectors are involved. Feedback loops are quite common in bacterial gene regulation and affect the speed, variation, and sensitivity of cellular responses and can also lead to epigenetic switches.

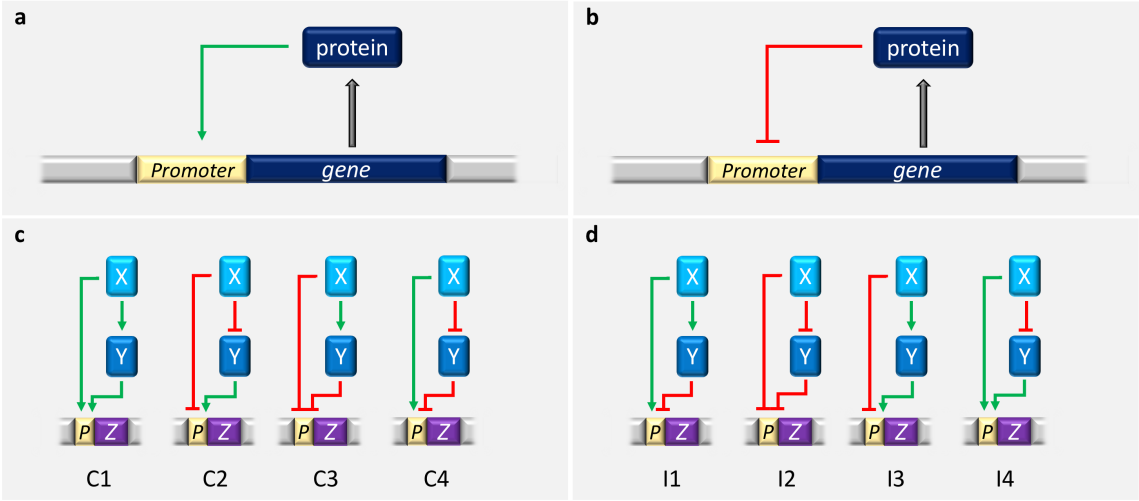

Figure 1.4: **Schemes of transcriptional network motifs. a)** positive feedback loop - gene product induces its own transcription; **b)** negative feedback loop - gene product represses transcription of its own gene; **c)** coherent feed-forward loops; **d)** incoherent feed-forward loops.

A positive feedback loop is a motif in which the product further induces the expression of the gene coding it (**Figure 1.4a**). Transcriptional activation is slower and noisier in case of positive feedback loops compared to no feedback systems (Maeda and Sano 2006, Sayut et al. 2007). This is most likely due to low initial amounts of the auto-inducer which might either be diluted before the expression is fully auto-induced or are bound to other functions within the cell. This stochasticity in auto-inducer binding to its promoter is likely to cause higher transcription noise upon the induction among individuals and delay the effect of the strong induction by positive feedback. Although slower and noisier, this motif enables cells to react to much lower concentrations of a signal once the positive feedback loop establishes. Both graded and hysteretic expression (i.e., ON or OFF expression state depending on recent events) can occur in genes regulated this way (Maeda and Sano 2006).

A negative feedback loop, i.e., a circuit where the product inhibits its own expression, exhibits the opposite traits of a positive feedback, as it decreases the response time (Rosenfeld et al. 2002) (**Figure 1.4b**). It also reduces the variability in product concentration decreasing cell-to-cell and temporal variability (Becskei and Serrano 2000). This enables the cell to use stronger promoters for proteins needed in a short time but at low concentration without overshoots and risks of protein toxicity. Hysteresis in terms of simple negative feedback loop has not been reported but could be readily achieved by double-negative feedback (Toman et al. 1985). An example of this is detailed below.

Positive feedback loops and double-negative feedback loops are known to play a role in memory of bacteria. A case of positive feedback loop was demonstrated in 1957 in which

an *E. coli* sub-population primed by high concentration of lactose analogue TMG is able to maintain *lac* operon fully induced even in low non-inducible TMG concentrations (Novick and Weiner 1957). While if the same bacteria are exposed to the low non-inducible TMG concentration without priming the *lac* operon remains repressed by LacI. The mechanism behind this resides in different levels of β-galactoside permease (LacY) present in TMG induced and naive cells. The induced population has a high level of the permease in the cell membrane, because *lacY* gene expression has been activated by previous high TMG concentrations. This state is preserved in a sub-population of induced cells after transfer into low TMG concentration as the abundant permease is able to maintain the intracellular TMG level high enough to avoid LacI repression. Thus the high level of permease leads to high intracellular TMG concentration which in turn acts as permease transcription activator. On the other side, naive cells have minimal amounts of LacY if any, so they are not able to obtain enough TMG from the solution to activate *lac* operon expression (Smits et al. 2006, Casadesús and Low 2013).

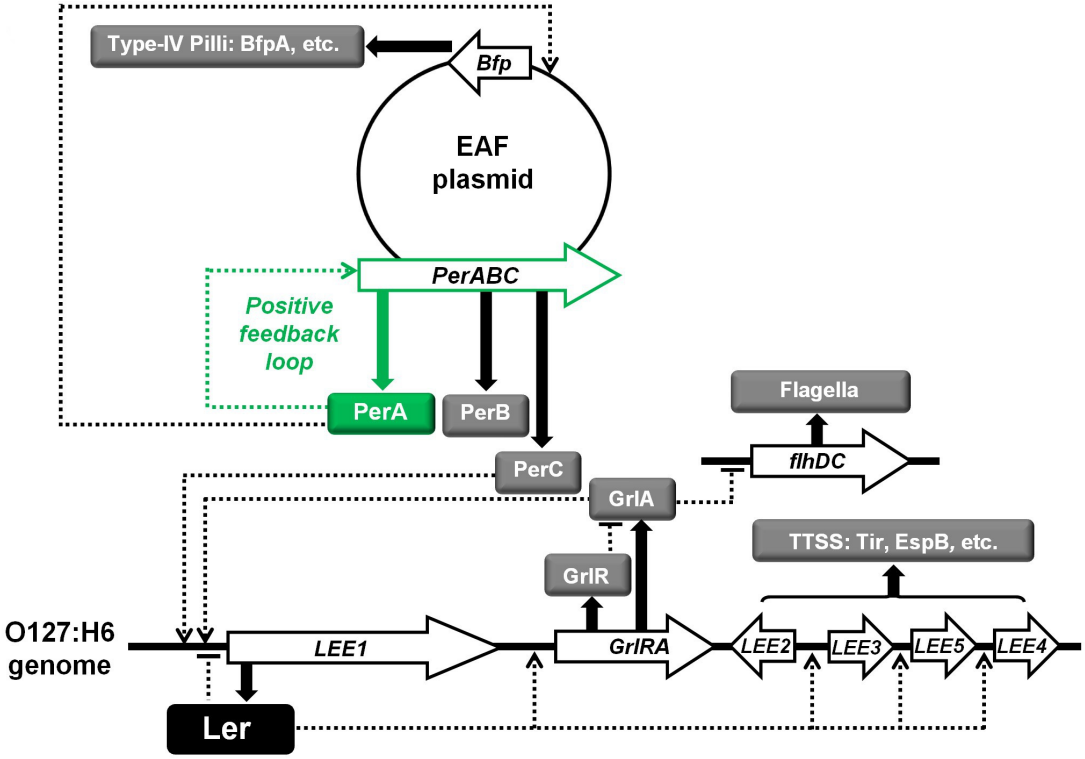

Figure 1.5: **Scheme of EPEC virulence regulation.** PerA positivelly autoregulates *perABC* operon and triggers expression of type IV pili. PerC activates transcription of Ler which is the main regulator of T3SS secretion system machinery. (reproduced from (Ronin et al. 2017))

Another example of a long-term epigenetic switch mediated by a positive feedback loop was published recently. Enteropathogenic *E. coli* (EPEC) has been revealed to co-exists in two, non-virulent and hyper-virulent, sub-populations, each having lag phases of different length (Ronin et al. 2017). At the transcriptional level it is a change in *per* operon expression (located on EAF plasmid) and *per* regulated genes. EPEC cultivation in virulence-activating conditions gives rise to *per*-ON hyper-virulent aggregative sub-population (SMALL phenotype) reaching nearly 100% (Ronin et al. 2017). Interestingly,

this high ratio of *per*-ON vs. *per*-OFF cells remained for many generations even after transferring the culture back into non-activating conditions, although naive EPEC culture contains only a minority of *per*-ON cells. Transition back from *per*-ON to *per*-OFF phase is achieved when cells are grown up to stationary phase. The long-term stability of *per*-ON state relies on a positive feedback of PerA which acts as an activator of its own gene beside *perB* and *perC* in the *per* operon (**Figure 1.5**) (Ibarra et al. 2003, Ronin et al. 2017).

A double negative feedback loop is known to play a role in epigenetic switching between a lytic and lysogenic cycle of *E. coli* phage λ (Smits et al. 2006, Casadesús and Low 2013). After inserting its DNA, the virus might undergo two different cycles - i.e., either replicate producing new phages and heading for bacterial lysis or integrate its DNA into the bacterial chromosome and persist there within lysogeny. The fate of the phage lies in two repressors, CI and Cro, each repressing transcription of the other (Eisen et al. 1970, Neubauer and Calef 1970). Both CI and Cro are produced at the beginning of the infection. If the level of CI reaches a certain threshold and outcompetes Cro, whose activity is suppressed, the phage enters a lysogenic cycle staying dormant in the cell thanks to a CI-*cI* positive feedback loop. Otherwise, Cro levels rise further inhibiting CI production and establishing phage proliferation with subsequent cell lysis (Svenningsen et al. 2005). It should be noted that although it is not predictable whether the phage enters the lytic or lysogenic cycle, the ratio of lysed/lysogenized cells depends on the bacterial physiologic situation and other environmental factors as well. Interestingly, Toman et al. used the CI-Cro system for epigenetic regulation of *gal* operon in *E. coli* (Toman et al. 1985).

Feed forward loops

Feed forward loops (FFLs) are more complex, but also common regulatory motifs. A FFL is a regulatory circuit when a transcription factor X regulates another transcription factor, Y, and both (X and Y) regulate a gene or operon Z. It has been predicted that coherent FFLs (C1-4; **Figure 1.4c**) delay transcriptional responses, while incoherent FFLs (I1-4; **Figure 1.4d**) should act the opposite as transcription accelelators and exhibit transient pulses of expression (Mangan and Alon 2003). This was later experimentally confirmed, but only in FFL C1 and I1, as these are the most abundant forms shown in **Figure 1.4c** and **Figure 1.4d** (Mangan et al. 2003, Kalir et al. 2005, Mangan et al. 2006). The distinction between coherent and incoherent FFL depends on whether the sign (positive or negative) of Z regulation caused by factor X *through* Y is the same as regulation of Z by X itself (Shen-Orr et al. 2002, Mangan and Alon 2003). Compared to feedback loops there is little known whether FFLs are involved in any epigenetic mechanism, following text thus covers only their effect on transcriptional dynamics.

Coherent FFL type 1 (**Figure 1.4c**, C1), in which both X and Y must bind to trigger transcription (**Table 1.1**; AND-gate logic) delays activation compared to non-FFL systems (Mangan et al. 2003). No delay occurs when transcriptional induction is turned off. On the other hand, if only factor X or Y is sufficient to induce Z (**Table 1.1**; OR-gate logic in C1 FFL) gene activation is not delayed, but its deactivation is delayed considerably (Kalir et al. 2005). The AND-FFL prevents Z expression when the concentration of X fluctuates close to a threshold, while the OR-FFL enables continuous production of Z even if X concentration drops below a threshold for a while.

Table 1.1: Relationship among expression activity (ON vs. OFF) in FFL C1, gate logic (AND vs. OR) in regulation and absence or presence (0 vs. 1) of trascription inducers

inducer X	inducer Y	AND-gate logic	OR-gate logic
0	0	OFF	OFF
0	1	OFF	ON
1	0	OFF	ON
1	1	ON	ON

From incoherent FFL motifs, type 1 is predominant in *E. coli* transcriptional network. Factor X here positively regulates both Y and Z; however factor Y acts as an inhibitor of the Z gene (**Figure 1.4d**, I1). This incoherent FFL speeds up transcriptional responses and was used to create effective transcriptional pulses (or transcriptional bursting), although weak pulse-generating was expected (Mangan and Alon 2003, Basu et al. 2004, Mangan et al. 2006). The acceleration trait was confirmed despite the fact that other motifs (e.g., feedback loop) contributed to the used regulatory system. This suggests that the motifs' fundamental functions are preserved even when combined with other motifs. Overall this type of regulation similarly, to negative feedback loop shortens the rise-time, but unlike the negative feedback does not avoid overshoots rather the opposite. The overshoot is seen as a pulse of high gene Z expression which is subsequently inhibited by factor Y, but not necessarily back to the basal level.

1.3 Global control of transcription

Differential physical access to DNA within a chromosome also affects promoter activity based on their position in the chromosome (Bryant et al. 2014). Nucleoid-associated proteins (NAPs) and supercoiling are general mechanisms affecting this accessibility. NAPs' role in several cellular processes such as replication, horizontal gene transfer or transcription regulation was shown (Dixon and Kornberg 1984, Kayoko et al. 1992, Aznar et al. 2013). NAPs, in general, have a dual effect on transcription, i.e., can act as both enhancers or silencers of genes.

H-NS is one of the most abundant nucleoid-associated proteins of the *E. coli* chromosome which occurs during all growth phases (Azam et al. 1999). Its expression is negatively autoregulated, but another NAP Fis acts as an activator of *hns* gene (Ueguchi et al. 1993, Falconi et al. 1996). H-NS binds AT-rich DNA regions and forms polymers bridging distant DNA sequences (Navarre et al. 2006, Arold et al. 2010). This leads to promoter silencing especially in cases when the α subunit of RNA polymerase uses AT-rich sites to stabilize binding of the whole complex to the promoter (Singh and Grainger 2013). However, H-NS can compete with RNA polymerase and other transcriptional factors binding to AT-rich sites. RNA polymerase might get stuck in a DNA loop due to H-NS polymerization and become unable to elongate nascent mRNA (Dame et al. 2002). The silencing of affected promoters is not strict though. RNA polymerase sometimes bypasses

this by association with an alternative σ factors instead of conventional σ^{70} (Grainger et al. 2008). Although H-NS is generally considered to be a global gene silencer, H-NS binding to the promoter sequence of a *ehxCABD* operon in Shiga toxin-producing *E. coli* (STEC) is needed for the *ehxCABD* operon expression (Singh and Grainger 2013).

Fis protein, similarly to H-NS, prefers binding at AT-rich sites and is negatively auto-regulated (Ball et al. 1992, Stella et al. 2010). The ability of Fis to affect gene expression in both positive and negative ways is better known (Choi et al. 2005, Karambelkar et al. 2012). Moreover, Fis antagonizes H-NS silencing of some promoters (Falconi et al. 2001). Even though Fis shares its binding preferences for AT-rich DNA with H-NS, it does not polymerase but bends the DNA sequence at the binding site (Hübner et al. 1989). This bending is essential for, e.g., transcription initiation of ribosomal gene *rrnB* (Gosink et al. 1993).

These two NAPs described here belong among the most well-understood but are not an exhaustive outline of all bacterial NAPs, see (Dillon and Dorman 2010, Aznar et al. 2013).

Supercoiling is another general mechanism that affects the access to the genome (Brahms et al. 1985). Overwinding (positive supercoiling) and underwinding (negative supercoiling) is generated during transcription when a transcription bubble forms (Wu et al. 1988). This happens due to the helical structure of DNA. Although some NAPs such as Fis and H-NS might affect the supercoiling levels (Ouafa et al. 2012) topoisomerases take care of this process. *E. coli* possess two major topoisomerases - gyrase (topoisomerase II) and topoisomerase I. The former releases positive the latter negative supercoiling (Wang 1971, Gellert et al. 1976). If high levels of supercoiling are not released the access to the DNA is reduced, and even already initiated transcription is slowed down or terminated (Chong et al. 2014).

Clearly, transcriptional regulation is a complex and multipartite process. Even when both global and specific regulation of transcription initiation permit production of a transcript the gene expression is still regulated further downstream, e.g., during mRNA elongation or translation. Moreover, one transcription factor can regulate multiple genes as well as being co-regulated by several factors including itself. Regulation of gene expression thus exhibits a huge network of mutually controlled outputs based on the information acquired by a cell, and each of these components can be influenced by selection.

1.4 Natural selection on responses

The current knowledge about gene expression in prokaryotes shows that transcription is a fairly regulated process. However, how this regulation has evolved under selection in natural conditions is not yet fully understood yet.

1.4.1 Selection on expression level and plasticity

Phenotypic plasticity is a crucial process by which cells adapt to changing environmental conditions. As mentioned above, the expression of the vast majority of *E. coli* genes are known to be under some sort of regulation. This implies that the expression of many genes is selected upon to be plastic to some extent. There is a considerable amount of knowledge about regulatory promoter sequences in *E. coli* that affect expression level in

specified conditions, such as activity and binding of various sigma factors and transcription factors (Ireland et al. 2020). In addition, the specific biochemical reactions affecting gene expression have been investigated (Brewster et al. 2012, 2014, Karp et al. 2018) and there is an experimental evidence that expression levels of many genes are close to optimal (Keren et al. 2016, Hawkins et al. 2020). This suggests that plasticity should be close to optimal as well. However, there is little information about what type of selection (stabilizing, directional, or diversifying) has shaped the expression levels and phenotypic plasticity in nature, especially in bacteria. Even in eukaryotes only one promoter has been studied in detail (Metzger et al. 2015, Duveau et al. 2017).

1.4.2 Selection on noise

In the case of phenotypic variability within an isogenic population (noise), prior works have clearly shown that it is an evolvable trait (Richard and Yvert 2014). Generally, essential, constitutive genes and genes with high effect on overall expression (i.e., some transcription factors) tend to have low transcriptional noise (Silander et al. 2012, Metzger et al. 2015). In contrast, there is also some evidence that there is higher noise in native promoters in *E. coli* compared to synthetic promoters (Wolf et al. 2015). These results suggest that for different genes and/or at different niches there is a different optimum of transcriptional noise. Indeed, a very recent study of the effect of expression noise on the fitness of *Saccharomyces cerevisiae* reveals that high noise is beneficial when the expression level is far from optimum and vice versa (Duveau et al. 2018), supporting the results of (Wolf et al. 2015). These results also indicate that after the introduction of an organism into a new environment, selection might act to increase noise in the expression of some genes before appropriate expression plasticity evolves. However, note that mutations in promoter regions usually change both expression level and noise at the same time (Metzger et al. 2015). Some work in this area on other model organisms such as *Drosophila melanogaster* (Schor et al. 2017) and mouse (Barroso et al. 2017) has been done as well. But, it remains to validate that it is actually selected for certain levels of noise in the natural environment.

In the next three chapters I describe a set of experiments aimed at exploring the selective forces acting upon gene regulation and transcriptional control in *E. coli*. I quantify regulation in promoter variants from environmental isolates of *E. coli* and compare these regulatory phenotypes to those observed in the laboratory strain MG1655 K12 and to those having random mutations never subject to selection.

Chapter 2

Gene regulation is commonly selected for high plasticity and low noise

Markéta Vlková, Olin K. Silander

Article under peer-review

(Nature Ecology & Evolution)

Author contributions:
Markéta Vlková: Conceptualization (equal); Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Project administration (supporting); Visualization (lead); Writing-original draft (lead); Writing-review & editing (equal).
Olin K. Silander: Conceptualization (equal); Formal analysis (supporting); Funding acquisition (lead); Methodology (supporting); Project administration (lead); Resources (lead); Supervision (lead); Writing-original draft (supporting); Writing-review & editing (equal).

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Marketa Vlkova
Name/title of Primary Supervisor:	Dr. Olin Silander
In which chapter is the manuscript /published work:	2

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

◉ The manuscript is currently under review for publication – please indicate:

- The name of the journal:

Nature Ecology & Evolution

- The percentage of the manuscript/published work that was contributed by the candidate: 90.00

- Describe the contribution that the candidate has made to the manuscript/published work:

She conceived the project and designed the experiments and analyses in collaboration with her supervisor Olin Silander. She performed all the experiments (master's student Stella Pearless and PhD student Bhargava Morampalli sequenced several plasmids she constructed). She carried out the bioinformatic analysis and data curation with inputs from her supervisors. She wrote the manuscript with contribution from her supervisor Olin Silander.

○ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Marketa Vlkova Digitally signed by Marketa Vlkova Date: 2021.09.23 14:06:40 +12'00'
Date:	23-Sep-2021
Primary Supervisor's Signature:	Olin Silander Digitally signed by Olin Silander Date: 2021.09.23 14:19:33 +12'00'
Date:	23-Sep-2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

2.1 Preface

Most of the existing knowledge about the selection forces acting on gene expression so far is based on several individual eukaryotic promoters (*S. cerevisiae*) and their cell-to-cell variability or expression noise. This is in part due to the high importance of hysteresis in the development of complex eukaryotic organisms, because high expression noise in an isogenic population can and often does lead to hysteresis. Most studies also focus on selection acting on an expression phenotype in a single experimental envrionment. However, in order to fully understand how gene expression and regulation is shaped in nature there is a need to study other aspects of gene expression besides noise through the lens of natural selection. These other phenotypes should also be investigated in multiple environments in the cases on non-constitutive promoters and in other biological systems, such as prokaryotes.

We tried to fill these gaps in this chapter by studying ten individual promoters that differ in their genetic variability within the population of environmental *E. coli* isolates. Using fluorescent reporter assays coupled with flow cytometry we quantify expression level, plasticity, and noise. To cover most of the regulatory dynamic range of each promoter we use three different environments specific to each promoter. In order to infer selection acting on the measured expression phenotypes we compare the expression level, plasticity, and noise between segregating variants which were subject to natural selection and random variants which were never subject to natural selection.

Olin Silander and I conceived the project and designed the experiments and analyses. I performed all the experiments covered in this chapter, including library preparation and flow cytometry assays. Stella Pearless and Bhargava Morampalli sequenced several plasmids that I constructed. I also carried out the bioinformatic analysis and data curation with inputs from Olin Silander and Tim Cooper. I wrote this chapter with contribution from Olin Silander and comments from Tim Cooper, Andrea Sajuthi and Nikki Freed. We published the first draft of this chapter as a preprint on bioRxiv (DOI: 10.1101/2021.07.18.452581) and it is currently under revision for *Natural Ecology & Evolution*. In this chapter I have since implemented several of the changes suggested by the reviewers and also adjusted the structure to match the formatting style of the other chapters in this thesis.

2.2 Abstract

Bacteria often respond to dynamically changing environments by regulating gene expression. Despite this regulation being critically important for growth and survival, little is known about how selection shapes gene regulation in natural populations. To better understand the role natural selection plays in shaping bacterial gene regulation, here we compare differences in the regulatory behaviour of naturally segregating promoter variants from *Escherichia coli* (which have been subject to natural selection) to randomly mutated promoter variants (which have never been exposed to natural selection). We quantify gene expression phenotypes (expression level, plasticity, and noise) for hundreds of promoter variants across multiple environments, and show that segregating promoter variants are enriched for mutations with minimal effects on expression level. In many promoters, we infer that there is strong selection to maintain high levels of plasticity, and direct selection to decrease or increase cell-to-cell variability in expression. Finally, taking an integrated view, we show that across all phenotypes combined, segregating promoter variants are far more phenotypically similar than would be expected given their genetic divergence. This is the consequence of both stabilizing and directional selection acting on individual phenotypes to minimize differences among segregating variants. Taken together, these results expand our knowledge of how gene regulation is affected by natural selection and highlight the power of comparing naturally segregating polymorphisms to *de novo* random mutations to quantify the action of selection.

2.3 Introduction

Gene regulation plays a critical role in determining the physiology and behaviour of unicellular organisms. As such, regulation of gene expression is subject to natural selection. There are at least three aspects of gene regulation that are under selection and which affect transcription and downstream phenotypes. The first of these is expression level (the amount of protein produced from a gene). There is experimental evidence that expression levels of some genes are close to optimal (Hawkins et al. 2020, Keren et al. 2016), and furthermore, can rapidly be tuned to optimise fitness within an environment (Dekel and Alon 2005). However, only a few studies have systematically tested the range of expression levels that have no effect on fitness, and whether this range differs between genes (Keren et al. 2016).

In contrast to the understudied role of natural selection on gene expression level, a considerable amount is known about the molecular mechanisms controlling expression level, especially in model organisms. For example, for the vast majority of genes and operons in the bacterium *E. coli*, the specific sigma factors, transcription factors, and their binding sites are well documented (de Boer et al. 2020, Ireland et al. 2020). In many cases, the specific biochemical interactions that affect expression level have been thoroughly investigated (Brewster et al. 2012, 2014, Gertz et al. 2009).

A secondary aspect of expression level is plasticity - changes in gene expression that occur when cells encounter different environmental signals. There are several aspects of phenotypic plasticity that are subject to selection; among these are the level of the transcriptional response, the speed of the response, and the sensitivity of the response

(Maeda and Sano 2006, Mangan et al. 2006, Rosenfeld et al. 2002). Although these plastic responses are well characterised, as with gene expression levels, little is known about whether stabilizing, directional, or diversifying selection is most influential in shaping phenotypic plasticity. One exception to this is the work that has been done to characterise the action of selection on the plasticity of the TDH3 promoter in yeast (Duveau et al. 2017). However, there is limited knowledge about phenotypic plasticity in bacteria and even in eukaryotes only a limited number of promoters have been studied (Duveau et al. 2017, Hill et al. 2021, López-Maury et al. 2008).

As with expression level, an abundance of research has been done on the molecular mechanisms underlying phenotypic plasticity. As noted above, many of the transcription factors that cause changes in transcription are well characterised. In addition, the specific topologies of regulatory networks that promote specific types of responses have been extensively studied (Basu et al. 2004, Becskei and Serrano 2000, Eisen et al. 1970, Kalir et al. 2005, Mangan and Alon 2003, Novick and Weiner 1957, Shen-Orr et al. 2002, Smits et al. 2006), including how different topologies influence the phenotypic effects of mutations (Madan Babu et al. 2006, Mayo et al. 2006, Metzger and Wittkopp 2019, Schaerli et al. 2018).

In contrast to expression level or plasticity, a large number of studies have quantified how selection acts on expression noise - the amount by which individual isogenic cells differ in expression level. For example, several studies have compared the noise exhibited by different promoters within the same organism to understand whether specific promoters have been selected such that they confer high or low levels of noise (Bar-Even et al. 2006, Elowitz et al. 2002, Rossi et al. 2019, Silander et al. 2012, Süel et al. 2007, Taniguchi et al. 2010, Wolf et al. 2015). There is evidence that promoters of essential genes exhibit low levels of noise, and this has been observed for both bacteria and eukaryotes (Metzger and Wittkopp 2019, Silander et al. 2012, Urchueguía et al. 2019). In addition, modeling has shown that there can be a selective advantage for genes with high plasticity to exhibit high noise levels if the regulation of gene expression is not precise (Wolf et al. 2015). High noise levels from some promoters can lead to phenotypically distinct populations of isogenic cells coexisting in the same environment (Govers et al. 2017, Kotte et al. 2014, Ronin et al. 2017). This can be selectively advantageous as a bet-hedging phenomenon, which is sometimes beneficial in highly variable or unpredictable environments (Acar et al. 2008, Veening et al. 2008). One of the most studied examples of bet-hedging is bacterial persistence, which is known to result in a small subpopulation of cells which tolerate antibiotic exposure without acquisition of antibiotic resistance genes (Lewis 2007).

The specific molecular mechanisms that can affect noise levels are also well-studied. For example, promoters in yeast with TATA boxes are generally more noisy than those lacking these sequences. Random mutations in TATA boxes have been found to decrease both expression noise and phenotypic plasticity (Hornung et al. 2012, Richard and Yvert 2014). In contrast, yeast promoters lacking TATA boxes have been found to be mutationally robust (Hornung et al. 2012). In *E. coli*, specific transcription factors are associated with promoters conferring high or low levels of noise (Silander et al. 2012). Finally, there are general features of gene regulation that promote or dampen noise. For example, genes that have low rates of transcription and high rates of translation will generally have higher noise levels (Jones et al. 2014).

Here, we investigate how selection acts on different gene expression phenotypes by selecting ten *E. coli* promoters which exhibit different levels of segregating variation. We quantify expression phenotypes (expression level, plasticity, and noise) from these segregating promoter variants in three environments specific for each promoter. Finally, to elucidate whether stabilizing, directional, or diversifying selection has shaped gene regulation, we compare the expression from the segregating promoter variants, which have been exposed to natural selection, to a set of variants generated through random mutation, and which have never been exposed to selection.

2.4 Materials and Methods

2.4.1 Promoter selection

We downloaded a database of transcription start sites (TSSs) driven by the σ^{70} factor from RegulonDB (Santos-Zavaleta et al. 2019). From this database, we filtered out TSSs denoted as "weak" in RegulonDB, or those outside of intergenic regions (IGRs). We also removed duplicated IGRs if multiple TSSs were present in the same IGR with the same orientation. This filtering resulted in 605 IGRs, each having one or more TSS, that were used for further analysis. Based on the TSS annotations, the IGRs together with 100 bp of flanking upstream and downstream sequences were extracted from a reference MG1655 genome using GenBank file annotation (Blattner et al. 1997). Throughout, we refer to IGRs together with the flanking sequences as "promoters". The genomes of 135 environmental *E. coli* isolates (Breckell and Silander 2020, Ishii et al. 2006, Sakoparnig et al. 2021) were then blasted against the promoter sequences extracted from MG1655 with an e-value cut-off set to 10^{-10}; Blast version 2.2.31+ (Altschul et al. 1990). Blast hits not more than 100 bp shorter than the MG1655 reference sequence were extracted as segregating promoter variants from each of the 135 environmental *E. coli* isolates. All segregating promoter variants of each promoter were then aligned via t-coffee version 11.0.8 with parameter mode procoffee (Notredame et al. 2000). We then calculated the average pairwise identity (API; t-coffee) and proportion of segregating sites (PSS) for each promoter and IGR (excluding the 100 bp flanking sequences).

We also extracted open reading frames (ORFs) directly downstream of each of the 605 filtered promoters from the MG1655 reference genome (GenBank annotation). The same workflow for extracting segregating promoter variants was followed when extracting segregating ORF variants from the 135 isolates of *E. coli* (changing the t-coffee parameter setting -type=dna). No extra flanking regions were included in the ORFs. Out of the 605 promoters originally identified in the MG1655 genome, 429 were found in at least 130 of the 135 environmental *E. coli* isolates together with their downstream ORFs. For all further analyses and tests we used only these 429 promoters, IGRs, and ORFs.

To check whether any functional classes of ORFs were enriched for promoters with IGRs that have high or low sequence variation, we grouped the IGRs into the major categories defined by MultiFun (Serres and Riley 2000) according to their downstream ORF. IGR variation (PSS and API) from each group was then compared to the rest of the groups together using a two-sided Wilcoxon rank-sum test. We also tested the correlation in sequence variation (PSS and API) between IGRs only and promoters (IGRs with 100 bp

of upstream and downstream sequences flanking each IGR) using Spearman's correlation test. All the scripts used with the workflow described above can be accessed through https://doi.org/10.5281/zenodo.5515765.

From the resulting promoter dataset we selected ten promoters exhibiting various levels of sequence variation. We aimed at selecting promoters with obvious and easy differential regulation, i.e, either with known environments in which differential expression occurs or those whose annotated transcription factors suggested possible environments for differential expression. The selected promoters - *aldA*, *yhjX*, *lacZ*, *aceB*, *mtr*, *cdd*, *dctA*, *ptsG*, *purA*, and *tpiA* - differ more than ten-fold in their levels of genetic variation, while they all control expression of genes that are involved in bacterial metabolism or transport of material involved in the metabolism.

2.4.2 Promoter library construction

For each promoter (*aldA*, *yhjX*, *lacZ*, *aceB*, *mtr*, *cdd*, *dctA*, *ptsG*, *purA*, and *tpiA*) we created two libraries, one with segregating promoter variants and a second with randomly mutated promoter variants. The two promoter libraries were aliquoted into separate microplates, except for the *lacZ* promoter, which was aliquoted into a single microplate. Each microplate also contained the MG1655 variant of the promoter, a positive control consisting of the highly active murein lipoprotein (*lpp*) promoter driving GFP expression, and a negative expression control consisting of a promoter-less pUA66 plasmid (Zaslaver et al. 2006). We describe the construction of each library type separately below. All cloned plasmid constructs are available from the corresponding author upon request.

Variants segregating in *E. coli* population

Both the vector pUA66 backbone (a low-copy number plasmid with a SC101 ori, a strong RBS, and GFPmut2) and segregating promoter variants were PCR amplified using Phusion High-Fidelity DNA polymerase with HF buffer (New England Biolabs). For promoter PCR amplification, 5μl of pooled DNA from isolates with different variants of each promoter was used as a DNA template. The primers for promoter amplification contained 17 nucleotide overhangs which were homologous to the ends of the vector backbone for subsequent DNA assembly. All primers used in this study are listed in **Table S2.1**.

For vector PCR amplification, 0.5 ng of plasmid DNA with pUA66 backbone served as a template. After confirming a successful PCR amplification of the products on 1% agarose gel, the template DNA was digested by DpnI from the remaining reaction volume (Li et al. 2011). The reactions were then column-purified and we assembled the PCR amplified vector and promoters using Gibson assembly (Gibson et al. 2009) with NEBuilder® HiFi DNA Assembly Master Mix (New England Biolabs). The assembly mix was then electroporated directly into the electrocompetent MG1655 strain. Transformed colonies which grew on LB agar plates with 50μg/ml Kanamycin were picked for Sanger sequencing across the insert in the vector backbone and stored as glycerol stocks. Clones with confirmed promoter variants matching the segregating ones were then grown in liquid LB with Kanamycin and used to create 96 well microplate glycerol stock libraries.

Random variants from PCR mutagenesis

We amplified the pUA66 backbone, DpnI treated, and column-purified it the same way as described for segregating promoter variants above. We produced the promoter inserts by performing error-prone PCR using the GeneMorph II Random Mutagenesis Kit (Agilent Technologies). We used the plasmid constructs with MG1655 promoter variants cloned into them as template DNA for the error-prone PCR aiming to achieve approximately 1.5 SNPs per promoter sequence. We used the same primers as for the segregating promoter variants (**Table S2.1**). Each reaction with randomly mutated promoter variants was column-purified before Gibson assembly with the NEBuilder® HiFi DNA Assembly Master Mix (New England Biolabs). Each assembly mix was then electroporated into the MG1655 strain and colonies that grew on LB with Kanamycin were picked for Sanger sequencing and stored as glycerol stocks. Clones which had none or more than three SNPs in the cloned promoter insert were excluded as well as those with SNPs detected in the vector backbone (rare occasion). The rest of the clones were then re-grown in liquid LB with Kanamycin in 96 deep-well microplates overnight and 96 well microplate glycerol stock libraries were prepared from them.

2.4.3 Bacterial clones and environments

We performed all the promoter activity assays in the MG1655 genetic background of *E. coli*. All the clones had a pUA66 vector (Zaslaver et al. 2006) differing by promoter variants cloned upstream of the GFPmut2 gene (for details see **Promoter library construction**). We also included positive and negative expression controls which consisted of the highly active plpp::GFPmut2 strain from the Alon Zaslaver library collection and promoter-less pUA66 vector, respectively (Zaslaver et al. 2006).

The assays were performed in 96 well microplates in M9 minimal media (Sigma-Aldrich) supplemented with $MgSO_4$, $CaCl_2$, a carbon source, and any additional reagents needed to induce or repress expression from a particular promoter (**Table 2.1**). We grew all the clones in the presence of 50μg/ml Kanamycin to prevent plasmid loss.

Before introducing the clones into the assay media we first inoculated them into 0.5ml of M9 minimal media with 0.4% glucose and Kanamycin (M9 glucose) from a 96 well microplate glycerol stock using a 96-well pin replicator (Enzyscreen B.V.). We then incubated these microplates overnight at 37°C with shaking to revive the cells. After overnight growth in M9 glucose, we inoculated each grown clone into 0.5ml of one of the assay media in a 96 deep-well microplate using the pin replicator and incubated them for 24h at 37°C with shaking. After this second incubation we inoculated the clones into the same fresh assay media they grew in for 24h, but into three separate 96 deep-well microplates (to obtain triplicates for each clone). These cultures were incubated until they reached an exponential phase (**Table 2.1**).

Table 2.1: **Promoter-environment combinations**

Promoter	Assay media	Incubation time prior flow cytometry	Environment abbreviation
aldA	0.4% glucose	4.5h	Glu
	0.4% fucose	4.5h	Fuc
	0.4% glycerol	4.5h	Gly
yhjX	0.4% glucose	4.5h	Glu
	0.4% pyruvic acid	4.5h	Pyr
	0.2% pyruvic acid	4.5h	Pyr2
lacZ	0.4% glucose	4.5h	Glu
	0.4% galactose	4.5h	Gal
	0.4% lactose	4.5h	Lac
aceB	0.4% glucose	4.5h	Glu
	0.4% pyruvic acid	4.5h	Pyr
	0.4% L-malic acid	7.0h	Mal
mtr	0.4% glucose	4.5h	Glu
	0.4% glucose + 1mM L-tryptophan	4.5h	Try
	0.4% glucose + 1mM L-phenylalanine	4.5h	Phe
cdd	0.4% glucose	4.5h	Glu
	0.4% glycerol + 2mM Cytidine	4.5h	Cyt
	0.4% glycerol + 3mM Adenosine monophosphate	4.5h	Amp
dctA	0.4% glucose	4.5h	Glu
	0.4% pyruvic acid	4.5h	Pyr
	0.4% L-malic acid	7.0h	Mal
ptsG	0.4% glucose	4.5h	Glu
	0.4% glycerol	4.5h	Gly
	0.4% D-mannose	4.5h	Man
purA	0.4% glucose	4.5h	Glu
	0.4% glycerol	4.5h	Gly
	0.4% L-arabinose	5.5h	Ara
tpiA	0.4% glucose	4.5h	Glu
	0.4% glycerol	4.5h	Gly
	0.4% L-arabinose	5.5h	Ara

Note: All environments included M9 minimal salt media supplemented with $MgSO_4$, $CaCl_2$, and 50 μg/ml Kanamycin.

2.4.4 Flow cytometry analysis

Once the cells reached exponential growth (**Table 2.1**) in the appropriate assay media, we diluted them into 1x PBS with $\sim$2.5% formaldehyde and kept them on ice until performing flow cytometry analysis the same day. We performed the flow cytometry on

a BD FACSCanto II machine using BD FACSDiva software version 6.1.3. We obtained the GFP fluorescence data through the 488 nm laser and a 513/17 nm bandpass filter. We set the number of events to record from each well to 20,000. We exported the acquired data from FACSDiva software into Flow Cytometry Standard files, and performed all cell gating and fluorescence analysis using custom R scripts (flowCore package version 2.0.1; available through https://doi.org/10.5281/zenodo.5515765). We gated cells based on their maximal kernel density of forward and side scatter values, using the ellipsoidGate function from the flowCore package, and keeping about 1/3 of all events (**Figure S2.1**).

The modal population expression was calculated as the mean of the three maximal kernel density values from the GFP fluorescence signal of three replicates. The modal coefficient of variation (mCV) was calculated separately for each of the three biological replicates (standard deviation divided by the modal population expression), and the mean of these values was used as the mCV of the promoter variant. Replicates with fewer than 2,500 and 5,000 recorded events were excluded from the calculation of modal population expression and mCV, respectively. We found that when fewer than 5,000 events were collected, a larger number of these were likely to be outlier events (e.g. machine noise), which affected the calculations of variance.

When comparing segregating and random variants of the same promoter that came from two separate microplates, we obtained an offset for both modal population expression and mCV to minimise plate-effects. We calculated these offsets as the mean of the differences between the three controls present in each microplate, i.e., the MG1655 promoter variant, plpp::GFPmut2 (positive control), and pUA66 (promoter-less negative control). All figures contain modal expression and mCV values that are corrected using these offsets. All scripts using the workflow described here, including the original data files can be accessed through https://doi.org/10.5281/zenodo.5515765.

2.4.5 Testing the correlation between segregating genotypic and phenotypic variation

To compare sequence promoter variation with variation in expression levels among segregating variants we used the promoter sequences (IGRs with 100 bp flanking regions) of the ten selected promoters (*aldA*, *yhjX*, *lacZ*, *aceB*, *mtr*, *cdd*, *dctA*, *ptsG*, *purA*, and *tpiA*). We measured modal population expression level in triplicates from each segregating promoter variant in three different environments (**Table 2.1**), and used the mean of these triplicate measures. We then calculated the standard deviation of modal population expression levels from all segregating variants of each promoter in each environment as a measure of phenotypic variation. For each promoter we thus had three values of segregating phenotypic variation (one for each environment). The MG1655 variant of the *mtr* promoter was excluded from the calculations due to a SNP in GFP causing lower fluorescence (**Supplementary Note**). We then used Spearman's correlation test to calculate the correlation between the phenotypic variation and genotypic variation using four metrics: PSS, API, total number of existing segregating variants and number of cloned segregating variants (the code for these calculations can be accessed through https://doi.org/10.5281/zenodo.5515765).

2.4.6 Mapping the effects of single SNPs to promoter sequence

When mapping the effect size and direction of randomly introduced SNPs we used modal population expression values only from those random variants which had contained a single SNP relative to the MG1655 promoter variant. Information about the TF binding sites, -10 and -35 elements, and ORFs was taken from the EcoCyc database (Karp et al. 2018), and only the annotations associated with σ^{70} driven TSSs were used. Using the Sanger sequencing results, we identified the location of all SNPs for each random variant (the code for these calculations can be accessed through https://doi.org/10.5281/zenodo.5515765).

2.4.7 Comparison of phenotypic variation between segregating and random variants

To test for differences in the variation in expression between the segregating and random variants, we calculated the modal population expression values for each promoter variant in both groups and all environments. We then tested for significant differences in variation in modal population expression levels using the Fligner-Killeen test of homogeneity of variances. We included values from the non-mutated MG1655 promoter variants in the sets with the segregating variants, except for the *mtr* promoter due to a SNP in GFP (**Supplementary Note**; the code for these calculations can be accessed through https://doi.org/10.5281/zenodo.5515765).

2.4.8 Comparing plasticity between segregating and random variants

We calculated the phenotypic plasticity of promoter variants across all three environments by calculating the Euclidean distance of each datapoint in three dimensions to an isocline representing null plasticity. The isocline is defined by equal values in the three dimensional space, i.e., x = y = z. The three dimensions are defined by the modal population expression values in the three environments specific for each promoter. Each datapoint (promoter variant) is thus defined by its expression values in all three environments. The closer a datapoint is to the isocline, the lower the plasticity.

To calculate plasticity in pairs of environments, we used an analogous method for two environments. To test whether natural selection had acted on plasticity (across all three environments and in pairs of environments) we compared plasticity values from all segregating (including MG1655 variants, except the one from *mtr* promoter - **Supplementary Note**) and all random variants for each of the ten promoters using a two-sided Wilcoxon rank-sum test (the code for these calculations can be accessed through https://doi.org/10.5281/zenodo.5515765).

2.4.9 Comparing noise between segregating and random variants

To determine the noise within an isogenic cell population of each promoter variant, we first excluded fluorescence values from the population that were lower or higher than three standard deviations from the modal population expression level. Then we calculated the mCV from this isogenic cell population as the standard deviation of the fluorescence divided by the modal population expression level. We next fitted a cubic smoothing spline (smoothing

parameter lambda = 0.01) to the modal population expression vs. the mCV values, using all (segregating and random) promoter variants (**Figure S2.5**). We determined the noise levels as the difference in measured mCV from the mCV predicted from the fitted spline. We then compared the noise values between segregating and random variants using a two-sided Wilcoxon rank-sum test to determine whether selection had acted on noise (the code for these calculations can be accessed through https://doi.org/10.5281/zenodo.5515765). The MG1655 variant of the *mtr* promoter was excluded from the calculations due to a SNP in GFP causing lower fluorescence (**Supplementary Note**).

2.4.10 Comparing overall promoter activity between segregating and random variants

We calculated the mean and standard deviation for each promoter and phenotype (modal population expression in all three environments, plasticity across all three environments and transcriptional noise deviation for each of the three environments) using just segregating variants. We then used these mean and standard deviation values of segregating variants to calculate z-score for each individual promoter variant (both segregating and random) to determine how much it differs in a particular phenotype from an average segregating variant. To focus on the relative size of the change, rather than on directionality, we converted all z-scores into absolute values. Giving all the phenotypes an equal weighting, we summed the z-scores for each variant to measure its deviation from an average segregating variant in all the phenotypes together. Due to a strong correlation between the z-scores for expression levels in the three environments, we used only expression level z-scores from a single environment. This environment was chosen so that the median expression from the segregating variants was higher than 2.5 to remove environments causing promoter repression. If more than one environment remained, we used the one least correlated with plasticity. This was done to minimize non-independence of individual z-scores when calculating cumulative z-scores. We then checked for statistically significant differences between the cumulative z-scores of segregating (including MG1655 variants, except the one from *mtr* promoter - **Supplementary Note**) and random variants using a two-sided Wilcoxon rank-sum test (the code for these calculations can be accessed through https://doi.org/10.5281/zenodo.5515765).

2.5 Results

2.5.1 Sequence variation in segregating promoters

To characterise the relationship between genetic variation, phenotypic variation, and selection on gene regulation, we first quantified genetic variation in intergenic regions (IGRs) and open reading frames (ORFs) from 135 environmental isolates of *E. coli* (Ishii et al. 2006) that span the genetic diversity of *E. coli* (Sakoparnig et al. 2021). We assumed that a substantial portion of gene regulatory phenotypes depend on the sequence of these IGRs. We used annotations in RegulonDB to identify 605 of IGRs with known transcription start sites in *E. coli* MG1655, and which are controlled by the sigma factor σ^{70} (Santos-Zavaleta et al. 2019). We also included the corresponding downstream annotated ORFs (Blattner

et al. 1997). We then identified and aligned homologous regions from the 135 environmental isolates of *E. coli* (Breckell and Silander 2020, Sakoparnig et al. 2021). To increase the likelihood of capturing homologous IGRs, we included 100 base pairs (bp) of the upstream and downstream ORFs (see **Materials and Methods**). Using these alignments, we calculated the proportion of segregating sites (PSS) and average pairwise identity (API) for each IGR (excluding the ORF portions of the alignments), as well as for the downstream ORFs (**Figure 2.1a** and **Figure S2.2a**). We found that the PSS in IGRs varied by more than an order of magnitude, from 0.369 (one indel and 104 SNPs across 282 bp for the IGR acting as a bidirectional promoter for the transcriptional dual regulator *melR* and α-galactosidase *melA*) to 0.007 (the IGR upstream of 30S ribosomal subunit protein *rpsM*, one SNP in 146 bp). We also observed perfect conservation for the IGRs upstream of the zinc metalloprotease ftsH (99 bp in length), the octaprenyltransferase *menA* (66 bp), the PF03966 family protein *ycaR* (51 bp), and the IGR that acts as as a bidirectional promoter for the transcriptional dual regulators *soxR* and *soxS* (85 bp). The API of IGRs ranged from 88.14% (one indel and 27 SNPs across 75 bp for the IGR upstream of bacterioferritin *bfr*) to 100% (*ftsH*, *menA*, *ycaR*, *soxR*, and *soxS*).

In order to understand whether selection had acted in a general manner to affect the levels of polymorphism, we tested whether ORFs in certain functional categories had upstream IGRs with high or low sequence variation. We found that genes involved in information transfer were enriched for IGRs with lower levels of polymorphism (median PSS 0.084 compared to 0.104 for all other IGRs, p = 0.005; median API 98.87% compared to 98.51% for all other IGRs, p = 0.004, two-sided Wilcoxon rank-sum test; **Figure 2.1b** and **Figure S2.2b**). This lower level of variation may be due to stronger stabilising selection on the regulation of ORFs involved in information transfer. The correlation between PSS and inverted API values for IGRs is high (Spearman's rho = 0.946, p = 4.3e-211) making the two measures interchangeable. PSS and API data for all IGRs and ORFs present in at least 130 out of 135 environmental isolates are in **Table S2.2**.

2.5.2 A system to quantify the effects of mutations on promoter phenotypes

To investigate in more detail the selective mechanisms responsible for conferring different levels of genetic variation in IGRs, we selected ten IGRs varying widely in PSS and API (**Table 2.2**, **Figure 2.1a** and **Figure S2.2a**; red points). We selected these based solely on genetic variation and for which literature data suggested phenotypic plasticity in certain environments.

For each of these ten IGRs, we cloned the segregating variants upstream of green fluorescent protein (GFP) on a low copy-number plasmid (**Figure 2.2a**, **Figure 2.2b** and **Figure 2.2d**). In all cases, the cloned regions included the IGRs together with 40 - 110 bp of coding sequences flanking the IGRs to include regulatory sequences that might be present outside the IGRs. We PCR amplified the promoter region from a DNA pool of the segregating variants. Accordingly, across all ten promoters, we obtained 75% (134 out of 179) of the segregating variants (**Table 2.2**). We transformed all the resulting plasmids into *E. coli* MG1655, such that they were all present in a single isogenic background. We expect that if promoter phenotypes are more similar in their respective native genetic backgrounds,

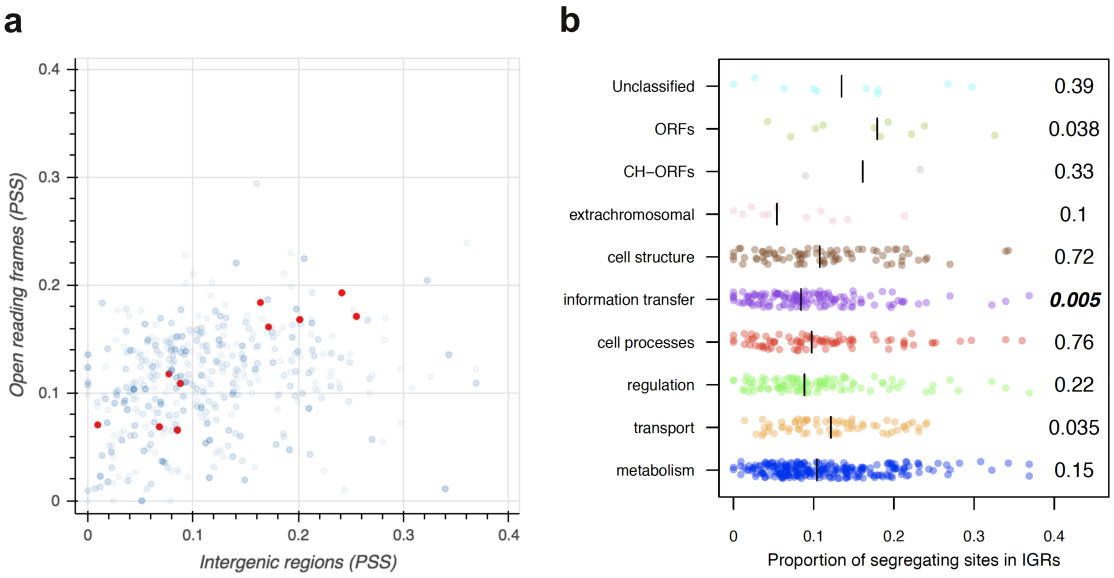

Figure 2.1: **Polymorphisms in intergenic regions (IGRs) and open reading frames (ORFs) across 135 environmental isolates of *E. coli* and MG1655. a)** The proportion of segregating sites (PSS) for IGRs and downstream ORFs varies by more than an order of magnitude. Each blue dot indicates PSS for an IGR-ORF pair when the IGR contains a transcriptional start site for the ORF. In red are IGRs that we selected for further study due to their different levels of sequence variation (*aldA*, *yhjX*, *lacZ*, *aceB*, *mtr*, *cdd*, *dctA*, *ptsG*, *purA*, and *tpiA*; see **Table 2.2**). **b)** The proportion of segregating sites for IGRs differs little among different functional groups, as classified by the downstream ORF. CH denotes conserved-hypothetical. Unclassified, ORFs, and CH-ORFs all represent groups of ORFs with very limited information on function (Serres and Riley 2000). Numbers next to each function group represent the p-value obtained by performing a two-sided Wilcoxon rank-sum test to test for differences in PSS in each group from all other PSS values for IGRs outside of the group. Bold indicates significant p-values after the Bonferroni correction for multiple comparisons among functional groups. The black lines indicate the median PSS values of each functional group.

then this change in genetic background will only increase phenotypic differences. Thus, any conclusions we make on the similarity in the behaviour of segregating variants are conservative.

We refer to these IGRs and proximal regions as promoters. We note that while differences between promoters in the expression phenotypes they confer are necessarily due to differences in the DNA sequence, it is not necessarily due to differences solely in transcription, but to any differences in the levels of transcriptional or translational regulation. Secondly, although this is a plasmid-based system, we have shown previously that expression phenotypes are well-correlated with those observed when these constructs are assayed in a chromosomal context (Silander et al. 2012).

Inclusion of the regions from the flanking ORFs had little effect on the relative genetic variabilities we observed (Spearman's correlation for the regions inclusive and exclusive of the flanking ORFs, R = 0.918 for PSS, p = 1.09e-180; R = 0.885 for API, p = 1.05e-149; **Figure S2.2c** and **Figure S2.2d**).

Table 2.2: **Characteristics of promoters selected for phenotypic assays**

Promoter	PSS	API	Number of segregating variants		Function(s) of downstream ORF	Environment abbreviation*
			cloned	total		
aldA	0.20	96.70%	25	34	metabolism	Glu Fuc Gly
yhjX	0.20	95.50%	16	24	cell structure, transport	Glu Pyr Pyr2
lacZ	0.18	96.19%	20	27	metabolism	Glu Gal Lac
aceB	0.14	97.87%	18	27	metabolism	Glu Pyr Mal
mtr	0.13	97.83%	19	26	metabolism, cell structure, transport	Glu Try Phe
cdd	0.09	98.78%	12	19	metabolism	Glu Cyt Amp
dctA	0.07	99.52%	7	8	metabolism, regulation cell structure, transport	Glu Pyr Mal
ptsG	0.06	99.16%	10	15	metabolism, regulation, cell structure, transport, information transfer	Glu Gly Man
purA	0.04	99.57%	4	7	metabolism	Glu Gly Ara
tpiA	0.02	99.76%	3	5	metabolism	Glu Gly Ara

Note: PSS - proportion of segregating sites in promoters, API - average pairwise identity in promoters. The functional groups of downstream genes were obtained using MultiFun (Serres and Riley 2000).
* For the full description of the assay environments see **Table 2.1**.

For each promoter we then identified specific environments in which it exhibited different expression levels. As it is well-established that glucose is the preferred carbon source for *E. coli* (Monod 1949), we included this as an environment for all promoters, assaying the activity during exponential growth in M9 minimal salts media with 0.4% glucose (Glu,

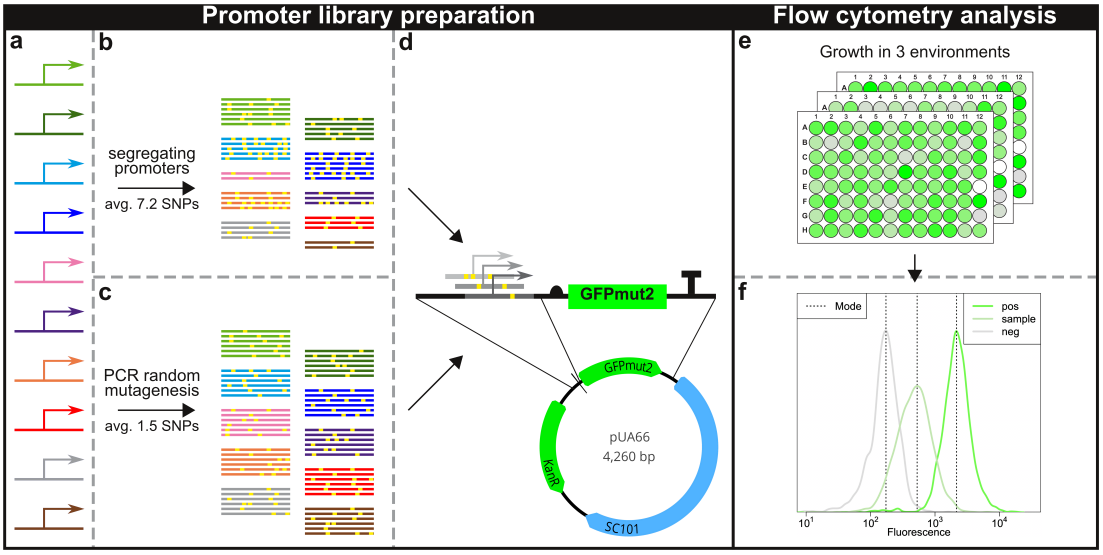

Figure 2.2: **Experimental design to assay the effects of segregating and random mutations on gene expression. a)** We isolated ten promoters (*aldA*, *yhjX*, *lacZ*, *aceB*, *mtr*, *cdd*, *dctA*, *ptsG*, *purA*, and *tpiA*) originating from MG1655. **b)** We then PCR amplified variants of these ten promoters segregating among environmental *E. coli* isolates from DNA pools. The average number of mutation across all segregating variants (as compared to MG1655) is 7.2 (ranging from 1 to 12.7 for individual promoters). **c)** We also performed PCR random mutagenesis using each of the ten MG1655 promoters with a target mutation rate of 1.5 mutations per promoter sequence. **d)** We cloned the resulting PCR amplicons (both segregating and random) into the pUA66 vector upstream of GFPmut2 (Zaslaver et al. 2006). We Sanger sequenced all the promoter variants to confirm the presence and location of mutations. From mutagenesis only the variants containing 1 to 3 SNPs were used for further phenotypic assays. **e)** We then cultured each of these individual promoter variants (1000 in total) in three different environments in triplicates, and **f)** quantified the modal population expression and modal coefficient of variation levels using flow cytometry.

Table 2.1 and **Table 2.2**). We used information from the literature on the behaviour of the MG1655 promoter variants to identify additional environments in which we expected a promoter to exhibit different expression levels. We confirmed this behaviour using flow cytometry. In this way, for each promoter, we identified two additional environments in which we observed differential expression (**Table 2.1** and **Table 2.2**, **Figure 2.2e** and **Figure 2.2f**). We found that all segregating variants behaved consistently in all the environments that we tested. In the case of *yhjX*, we were not able to obtain expression levels higher than cellular background fluorescence in any environment that we tested except pyruvic acid. Nevertheless, we found that the promoter was very responsive to the concentration of pyruvic acid. We thus used two different concentrations of pyruvic acid as a carbon source to achieve differential levels of expression.

2.5.3 Relationship between segregating genetic variation and phenotypic variation is correlated

We hypothesized that the differences we observed in the genetic variation present for different promoters might be a consequence of differences in the range of optimal expression levels. If the optimal range was very narrow, then all variants of that promoter should exhibit similar expression levels, and new mutations that affect expression should be filtered by selection, decreasing genetic diversity. In contrast, if there were a wide range of optimal expression levels within an environment, then variants can differ in expression levels, selection would act weakly against new mutations, and segregating genetic diversity would be high. Under this hypothesis, there should be a positive correlation between genetic variation and phenotypic variation. A second possibility is that segregating variants have very little effect on phenotype and are selectively neutral. In this case we would expect no correlation between genetic variation and phenotypic variation.

To test for such a correlation, for each promoter, we measured the modal population expression level from each segregating variant in each environment. In all cases, we measured the modal population expression from three full biological replicates (**Materials and Methods**). As a measure of phenotypic variation, we calculated the standard deviation in log expression level across all segregating variants of each promoter. We found that promoters with more segregating genetic variation tended to have larger phenotypic variation in expression (Spearman's rho = 0.666, p = 5.8e-05; **Figure 2.3a**, **Figure 2.3b**, **Figure S2.3a**, and **Figure S2.3b**). While this evidence does not conclusively show that the lower level of genetic variation in some promoters is due to the increased strength of stabilising selection, it shows that our experimental system is capable of detecting the effects of segregating mutations on expression phenotype.

2.5.4 Effects on expression level are environment-dependent

Segregating promoter variants have evolved under the pressure of natural selection. Despite this, some segregating variants differ by up to 27 SNPs from the MG1655 variant (the putative transporter *yhjX*), and we have found that these mutations have detectable phenotypic effects - promoters with larger levels of sequence variation tend to have a larger range in modal expression levels (**Figure 2.3a** and **Figure 2.3b**). To investigate how selection has affected the level of segregating variation, we created a set of promoter variants with mutations that had not been subject to selection. To do this we used PCR mutagenesis on the MG1655 promoter variant, as a representative of segregating variants of each of the ten promoters. We cloned these upstream of GFPmut2 gene on a plasmid and transformed them into MG1655, creating a set of promoter variants exactly analogous to the segregating variants, but containing mutations that had never been subject to the action of selection (**Figure 2.2c** and **Figure 2.2d**). Similar approaches comparing segregating variants with new mutations have been used in previous studies (Denver et al. 2005, Metzger et al. 2015).

To identify the specific mutations that each of these new promoters contained, we sequenced 192 clones for each of the ten promoters (1,920 clones in total; see **Materials and Methods**). We filtered out all duplicated variants and any containing none or more than three *de novo* random mutations. Next, for nine of the ten promoters, we selected 93 randomly mutated variants to produce a well-characterised library of promoter variants

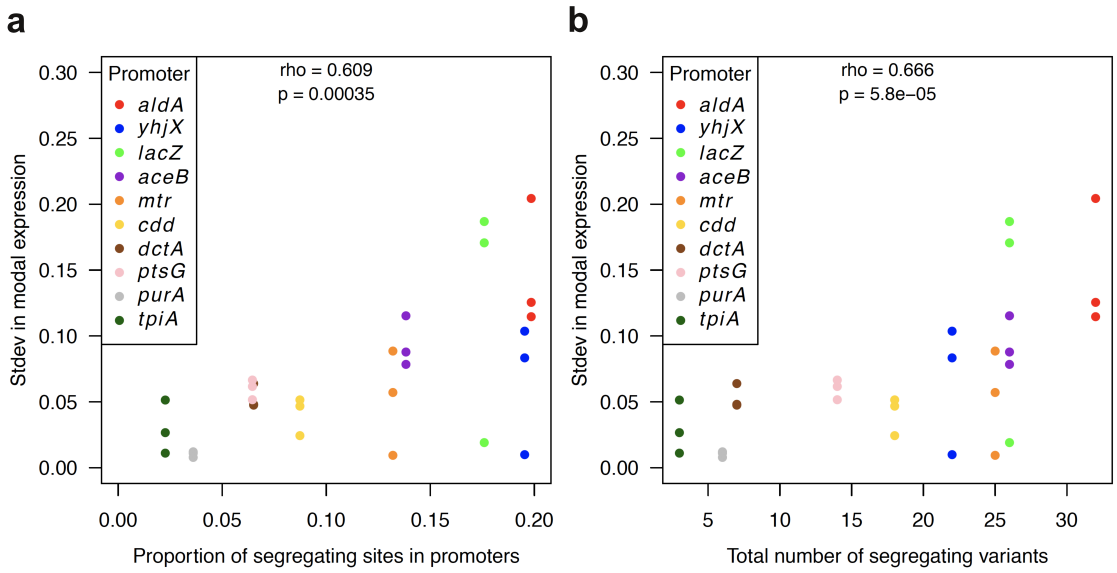

Figure 2.3: **Segregating genetic variation in promoters correlates with phenotypic variation in expression levels. a)** and **b)** Standard deviations in modal expression levels from segregating variants are correlated with the genetic variation of the promoter (IGR with 100 bp flanking regions). Panel **a** shows the correlation with PSS and panel **b** shows the correlation with the total number of segregating promoter variants. For each promoter, the standard deviation of modal population expression was measured in three environments (three dots per promoter, **Table 2.1**). The rho and p-values were calculated using Spearman's correlation test.

containing mutations that have never experienced the action of natural selection. For the tenth promoter, *lacZ*, we obtained only 29 random variants (see **Materials and Methods**). Given the small sample size, selective effects will be more difficult to detect, as statistical significance will generally require larger effect sizes.

Using these libraries, we first characterised the effects of the randomly introduced SNPs on expression level, considering only promoters with a single SNP (**Figure 2.4a**). As expected (Belliveau et al. 2018, Kinney and McCandlish 2019), we found that mutations with the largest effects on expression in all environments and promoters were present in the regions required for σ^{70} binding (-35 or -10 elements) or annotated transcription factor (TF) binding sites (p = 2.27e-09 for larger fold-changes in expression for SNPs within vs. outside of σ^{70} and TF binding sites, one-sided Wilcoxon rank-sum test). However, these data also showed that both the direction (i.e. an increase or decrease in expression level) and size of the mutational effects were dependent on the growth environment. In some promoter-environment combinations almost half of the mutations resulted in increased expression, while in a different environment these same mutations decreased expression. For example, comparing the effects on expression level in pyruvic acid vs. L-malic acid, 44.2% of all single mutations (both inside and outside TF binding sites) in *aceB* and 31.11% of all single mutations in *dctA* had opposing effects (**Figure 2.4b**). In other cases, the effect sizes of the mutations were environment-dependent (e.g. compare *aldA* in glucose vs. fucose or *yhjX* in glucose vs. pyruvic acid; **Figure 2.4a**).

We note that the MG1655 *mtr* variant is expressed at lower levels, which is likely caused by a non-synonymous mutation in GFP (**Supplementary Note**). The changes in

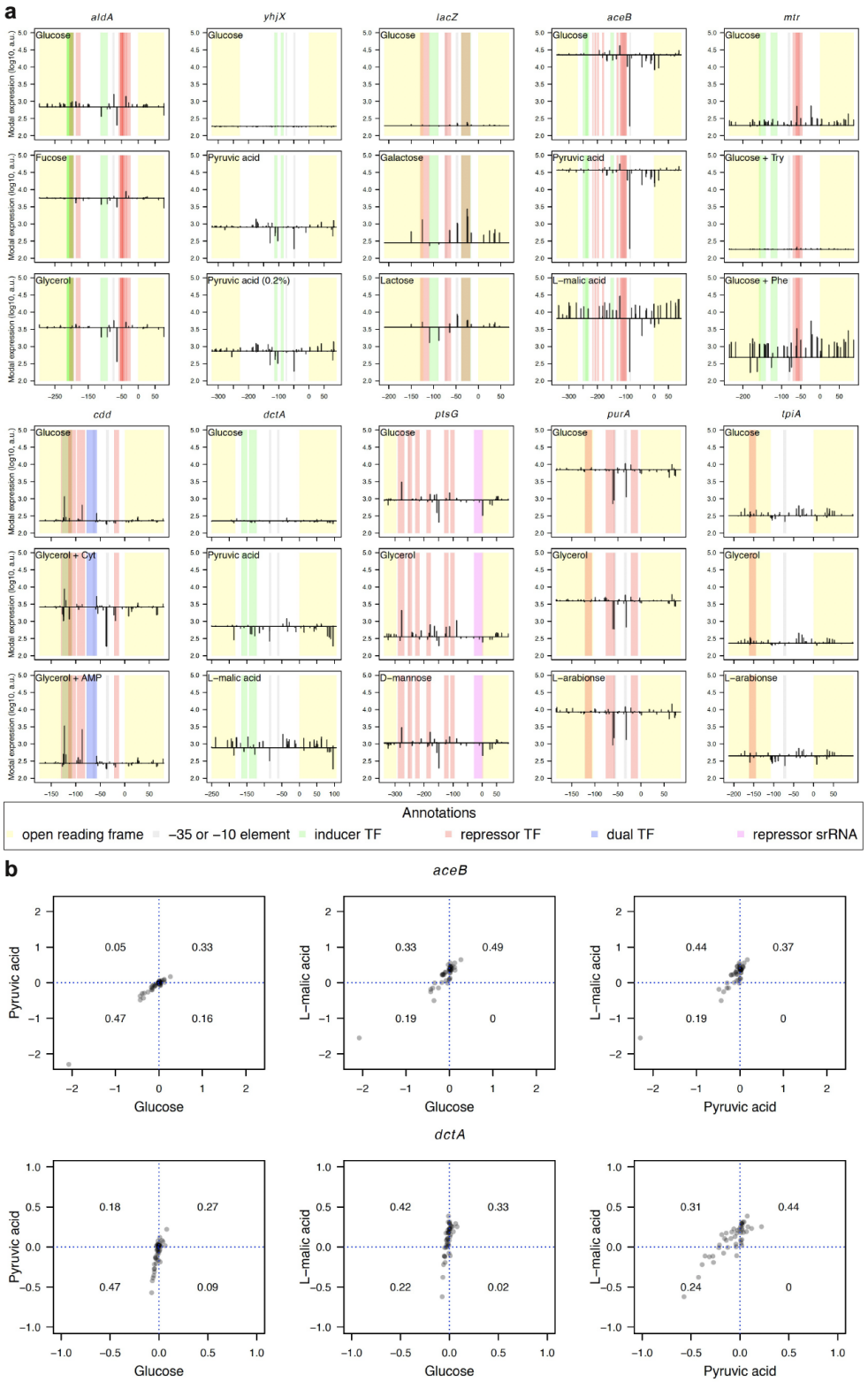

Figure 2.4: **The effects of random mutations on expression level are promoter and environment dependent.** *(continues on the next page)*

Figure 2.4: *(continues from the previous page)* **a)** Changes in expression from MG1655 promoter variants due to single random mutations for all environments. For each promoter, the x-axes indicate the position relative to the start codon of the gene downstream of each promoter. The horizontal black line indicates the expression level of the MG1655 promoter variant in that environment. The vertical black lines show the direction and size of the change in expression level when the MG1655 promoter variant is mutated at that specific position. While in certain environments some mutations cause no change in expression phenotype in other environments, the same mutations can cause up to 10-fold changes in expression. The results shown here are only from random promoter variants containing a single mutation relative to the MG1655 promoter variant. Note that for *mtr*, the MG1655 variant has a mutation in the downstream GFP, causing lower fluorescence (**Supplementary Note**), and thus the random mutations exhibit larger effects when the promoter is activated. The brown stripes result from overlapping repressor and activator TF binding sites. **b)** Comparison of differences in modal expression from random variants with single SNP relative to the MG1655 variant for *aceB* and *dctA* promoters. Differences are always compared between a pair of environments. Blue dotted lines indicate a null change in expression as compared to the MG1655 variant. These lines also divide each plot into four quadrants. Data points in the bottom left and top right quadrants show changes in expression in the same direction (decrease or increase) in both environments. Variants clustering in top left and bottom right quadrants exhibit changes in the opposite directions in the two environments (increase in one environment and decrease in the other). The numbers in each quadrant represent the proportion of variants falling into that quadrant. Each data point in **a** and **b** is a result of three full biological replicates (**Materials and Methods**).

expression in **Figure 2.4** for the *mtr* promoter, most notably for glucose with phenylalanine (Glucose + Phe) thus do not accurately represent the magnitude of the changes to native MG1655 promoter variant.

All together, these results emphasize that in order to gain general insights into the effects of mutations on gene expression phenotypes, it is necessary to measure promoter activity across different growth environments. Not only did we observe that the relative effects of random mutations on promoter activity varied across promoters, but both the size and direction of effects varied across environments (**Figure 2.4b**). However, due to the action of selection, this may not be true for naturally segregating promoter variants - for example all large-effect mutations may be filtered by selection. To quantify the action of selection, we next investigated differences between the phenotypic effects of naturally segregating and *de novo* random mutations.

2.5.5 Segregating polymorphisms are enriched for mutations with small effects on expression level

We first compared the expression levels of each of the 93 random variants per promoter (29 in the case of *lacZ*) to the expression levels of segregating promoter variants. Specifically, for each promoter and each environment, we tested whether expression levels among random mutants varied more than expression levels among segregating promoter variants. In all cases in which we observed significant differences, we found that random promoter variants differed more in expression levels than did segregating variants (**Figure 2.5**). This is despite the far larger numbers of segregating mutations (on average 7.2 mutations different compared to MG1655) compared to randomly generated mutations (on average

1.5 mutations different compared to MG1655). This shows conclusively that segregating mutations are enriched for mutations with small effects on expression levels, across a range of environments and promoters having different levels of segregating diversity (**Figure 2.5**).

In the majority of cases, the MG1655 promoter variant is representative of other segregating variants. However, we found two cases in which the MG1655 promoter variant

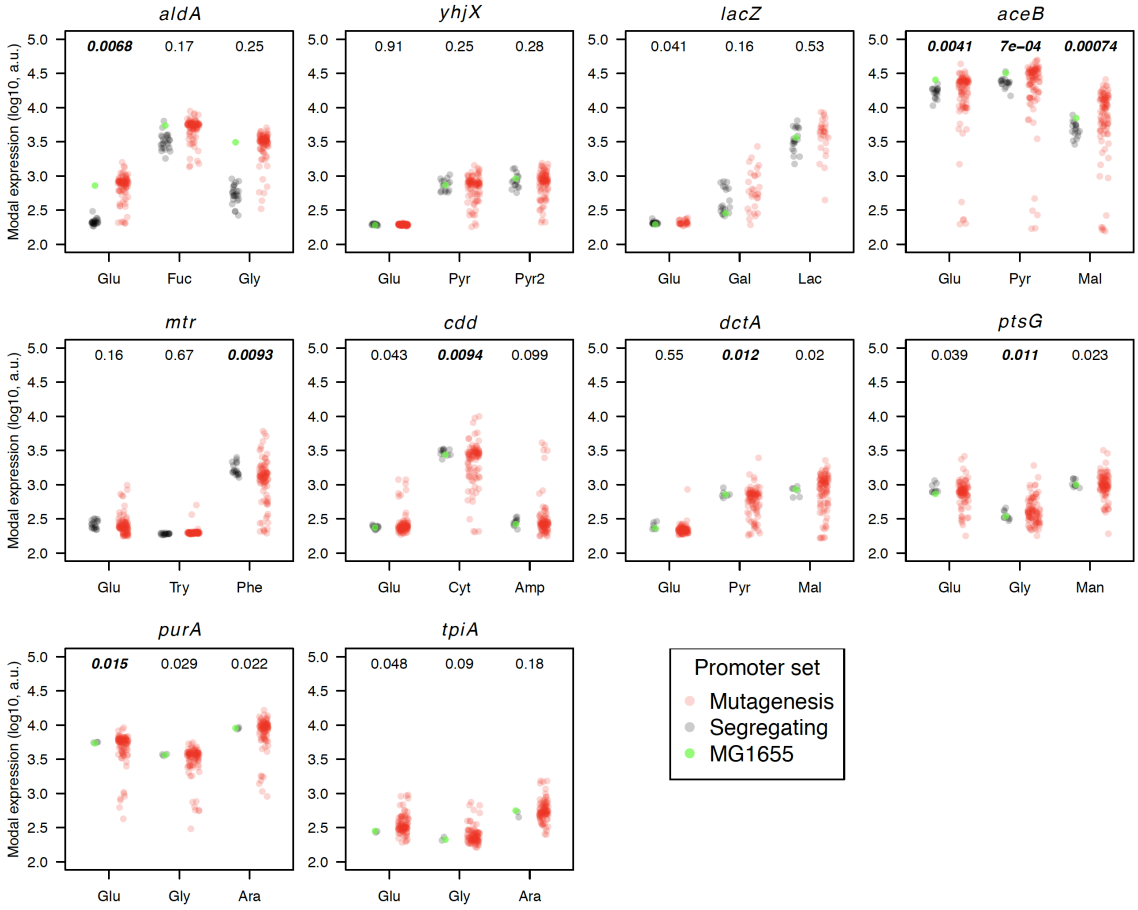

Figure 2.5: **Selection acts against mutations with large effects on expression levels.** We calculated the modal population expression level for all segregating and random promoter variants, and tested whether segregating variants differed in expression levels more or less than random variants. If segregating promoter variants exhibit similar expression levels (i.e. they have low variance), compared to random promoter variants, then we can conclude that large-effect mutations are filtered out via selection. For each promoter, the expression level of each group of variants is plotted. Segregating variants are shown in grey, random variants in red, and the MG1655 variant in green. The expression levels during growth in each of three different environments are shown in each plot and indicated on the x-axis (**Table 2.1**). Here, as in the other figures, the panels are arranged in decreasing order of segregating genetic variation (PSS). As noted in **Figure 2.4**, the MG1655 *mtr* promoter variant has a mutation in GFP, causing lower fluorescence, which is most apparent when the promoter is activated (**Supplementary Note**), we thus omitted it from the calculations. The numbers above each pair of segregating and random variants is the p-value obtained through the Fligner-Killeen test to test for differences in variances between the two groups. The numbers in bold indicate significant p-values after the Bonferroni correction for multiple comparisons within the promoter.

behaved differently than all other segregating promoter variants. In the case of *aldA*, all segregating variants exhibited lower expression levels than the MG1655 variant across all conditions, except for a single variant in fucose (**Figure 2.5**). In addition, the expression levels of promoter variants obtained via random mutagenesis of the MG1655 variant clustered around MG1655. The increased expression in the MG1655 variant is likely due to a single SNP that changes the -35 element from TGCCGT to TTCCGT, which is more similar to the canonical -35 motif (Harley and Reynolds 1987). In addition, the MG1655 *mtr* variant was expressed at lower levels likely linked to a single SNP in the GFP gene, as mentioned above (**Supplementary Note**) and was thus excluded from the analysis.

2.5.6 Segregating polymorphisms are selected to maintain high levels of phenotypic plasticity

A key role of promoters is to control expression of downstream genes in order to provide optimal levels of gene product in particular environments, i.e., phenotypic plasticity. To understand how selection acts on phenotypic plasticity, we again compared the phenotypes of the library of random variants to the segregating variants. We hypothesised that frequently, genes involved in the metabolism of specific substrates will be selected such that their expression is highly plastic (i.e. environment-dependent). We thus compared the environmental dependence of expression levels for all segregating and random promoter variants in all three environments. When considering two environments, highly plastic promoters should result in high expression levels in one environment, and low expression in the second environment. Less plastic promoters would have similar expression levels across both environments. We quantify this plasticity by calculating the distance of each promoter variant from an isocline specifying identical expression in both growth environments (**Figure 2.6a** and **Figure S2.4**). This same logic can be applied to three environments: promoters with similar levels in all three environments are defined as having low plasticity, while strong differences in expression levels in each of the three environments will be observed in highly plastic promoters. Thus to quantify plasticity, we calculated the Euclidean distance of each promoter variant in three dimensions (i.e. the three environments) from an isocline representing equal expression levels in all three environments (**Figure 2.6b**). This distance is 0 for promoter variants that are expressed at the same level in all environments, and greater than zero for promoter variants that are expressed at different levels, with a maximal value that is dependent on the maximum difference in expression levels between any two environments.

As shown above (**Figure 2.5**), many random mutations have large effects on expression, and we observed many mutations that increased as well as decreased expression levels. However, this was not true for plasticity. Only rarely did random mutations increase plasticity. For four out of the ten promoters, we found significant differences in plasticity between segregating and random variants (**Figure 2.6c**; excluding *aldA* promoter results - see below), with random promoter variants exhibiting lower plasticity. Using the median plasticity of segregating promoter variants, 92.47% (*aceB*), 75.27% (*mtr*), 76.34% (*cdd*), and 86.02% (*ptsG*) of the random variants exhibited lower plasticity. This strongly suggests that segregating promoter variants are under strong selection for high plasticity in these cases. Furthermore, this is almost certainly a conservative estimate of the strength of

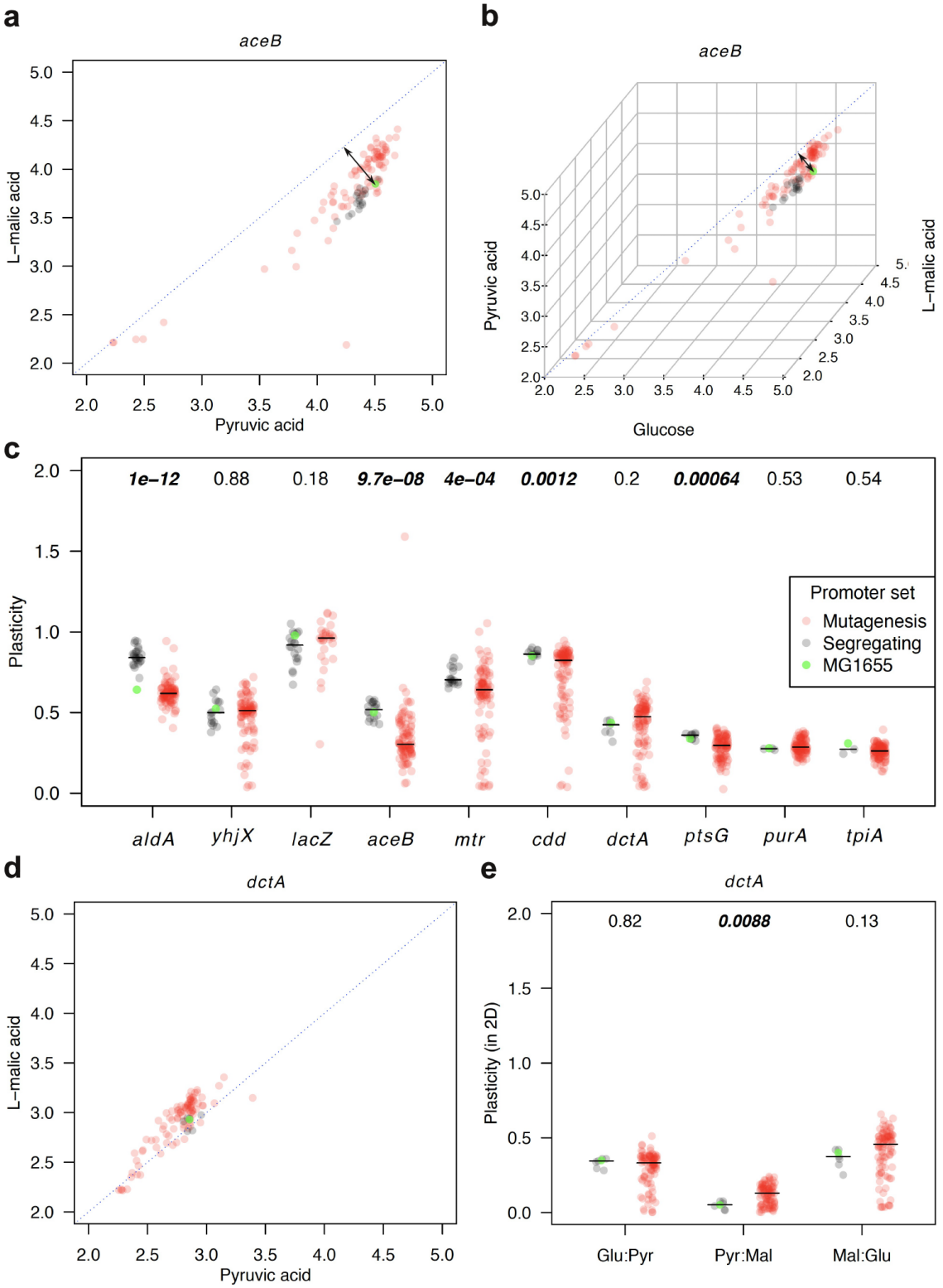

Figure 2.6: **Selection on plasticity.** Comparison of expression levels from promoters in a combination of environments. *(continues on the next page)*

Figure 2.6: *(continues from the previous page)* **a)** For the *aceB* promoter in pyruvic acid and L-malic acid, the segregating variants tend to have high expression in pyruvic acid and low expression in L-malic acid, indicating high plasticity in these two dimensions. **b)** The same projection as a but using all three environments. The blue dotted lines in **a** and **b** indicate the isocline of equal expression levels in all environments, i.e, no phenotypic plasticity. The black arrows illustrate the plasticity for the MG1655 promoter variant (measured as the minimal distance of each point from the blue isocline, i.e., x = y for **a** and x = y = z for **b**. The further from this isocline a promoter variant is, the higher its phenotypic plasticity. **c)** Plasticity calculated as in **b** for all environments and all ten promoters. Segregating promoter variants often have higher levels of plasticity across the three environments. The promoters are arranged in decreasing order of segregating genetic variation (PSS). The MG1655 variant of the *mtr* promoter was omitted from calculation due to a SNP in GFP (**Supplementary Note**). **d)** Expression of all *dctA* variants in pyruvic acid and L-malic acid. In this promoter, we detected low levels of plasticity in segregating variants as compared to random variants in these two environments. **e)** Calculating the plasticity of *dctA* using pairs of environments we found that in pyruvic acid and L-malic acid segregating promoters exhibited lower plasticity ($p = 0.009$). Numbers above each pair of segregating and random variants in **c** and **e** are the p-values obtained via two-sided Wilcoxon rank-sum test to test for differences between the two groups. The numbers in bold indicate significant p-values. The horizontal black lines in **c** and **e** indicate the median plasticity values of each group.

selection on plasticity, as we have assayed expression in only three environments.

Interestingly, when we calculated plasticity only for pairs of environments (**Figure S2.4**) we found a case when the random variants exhibited significantly higher plasticity than the segregating variants - *dctA* in pyruvic vs. L-malic acid (76.34% of random variants have higher plasticity than is the median for segregating variants; median plasticity of 0.184 for random variants as compared to a median of 0.073 for segregating variants, $p = 0.009$, two-sided Wilcoxon rank-sum test; **Figure 2.6d** and **Figure 2.6e**, Pyr:Mal). However, the difference in plasticity across all three environments was not significant ($p = 0.199$, two-sided Wilcoxon rank-sum test; **Figure 2.6c**). This was the only case in which the data implied that there was selection for low plasticity, i.e., for similar levels of expression across different environments.

Finally, in the case of *aldA*, we found that the MG1655 variant exhibited considerably lower plasticity than all other *aldA* segregating variants (**Figure 2.5**). The random variants have similar plasticity as MG1655, resulting in a strong difference in plasticity between the segregating and random variants (**Figure 2.6c** and **Figure S2.4**, *aldA*). Despite this, if we compare the plasticity of the random variants to that of MG1655, we find that 67.74% exhibit lower plasticity. This suggests that even though the MG1655 variant has lower plasticity than other segregating variants, selection has still acted to filter variants that decrease plasticity.

2.5.7 Segregating polymorphisms are enriched for mutations that both increase or decrease noise

Finally we also looked at the role of selection on noise levels among our ten promoters. Here, we use the term "noise" to refer to the differences between isogenic cells in their protein expression level in the same environment. Promoters conferring high levels of

noise result in isogenic cells having different protein expression levels, while promoters conferring low noise result in almost all cells having the same expression level. We note that noise is highly dependent on expression level due to the stochastic production of mRNA and protein. In general, promoters that confer low expression exhibit higher noise, while promoters that confer high expression exhibit lower noise. Because there is strong selection on the expression level of a protein, we first calculated a metric that allowed us to decouple noise from expression level. Specifically, we calculated the vertical deviation from a fitted smooth spline on the log of the modal expression levels versus the modal coefficients of variation (mCV; **Materials and Methods**), a measure analogous to the coefficient of variation. We then fitted the spline using all variants (segregating and random) for a particular environment (**Figure S2.5**). Thus, positive deviations from the fitted spline indicate promoters with high noise, while negative values indicate promoters conferring low noise.

We found that eight out of ten promoters had significantly different levels of noise (after Bonferroni correction for multiple tests for each promoter) in segregating variants as compared to random variants in at least one environment (**Figure 2.7**). In the majority of cases, noise was lower for segregating variants. As we calculated the noise metric separately for each promoter and each condition, this strongly suggests that selection has acted directly to decrease noise. The differences in noise are not due to differences in regulatory inputs, or to growth conditions, but to subtle changes in the sequence of the promoter and the resulting phenotypes. It is possible that some of these effects are mediated by changes in the binding strength of various transcription factors. However, this is unlikely to be the primary factor. We would expect the major effect of these changes to be expression level, and this noise metric explicitly accounts for expression level.

However, segregating promoter variants did not exhibit lower noise in all cases. Specifically, segregating *aceB* variants exhibited higher noise when grown in L-malic acid (median noise deviation 7e-03 for segregating variants compared to -2e-03 for random variants, p = 0.004, two-sided Wilcoxon rank-sum test; **Figure 2.7b**). Notably, in the two other environments that we tested, pyruvic acid and glucose, segregating *aceB* variants exhibited lower noise. In 13 cases out of 30 we did not detect a significant difference in noise between the two groups. In these cases, we expect that the differences are below our limit of detection, or that random mutations are just as likely to result in increased noise as they are in decreased noise. Finally, we note that we have measured noise in only a very limited number of environments. It is possible that in other environments, such as those with high stress, the majority of promoters would exhibit higher noise.

2.5.8 Segregating polymorphisms are enriched for mutations with small phenotypic effects across multiple expression phenotypes

So far we have focused on how natural selection affects individual expression phenotypes: expression level, phenotypic plasticity, and noise. However, in natural environments all these aspects are simultaneously under selection. While a particular random variant may exhibit behaviour comparable to the segregating variants for a single phenotype, this may not be the case across all the phenotypes. To compare the behaviours of segregating and random variants across all the expression phenotypes, we first made the assumption that

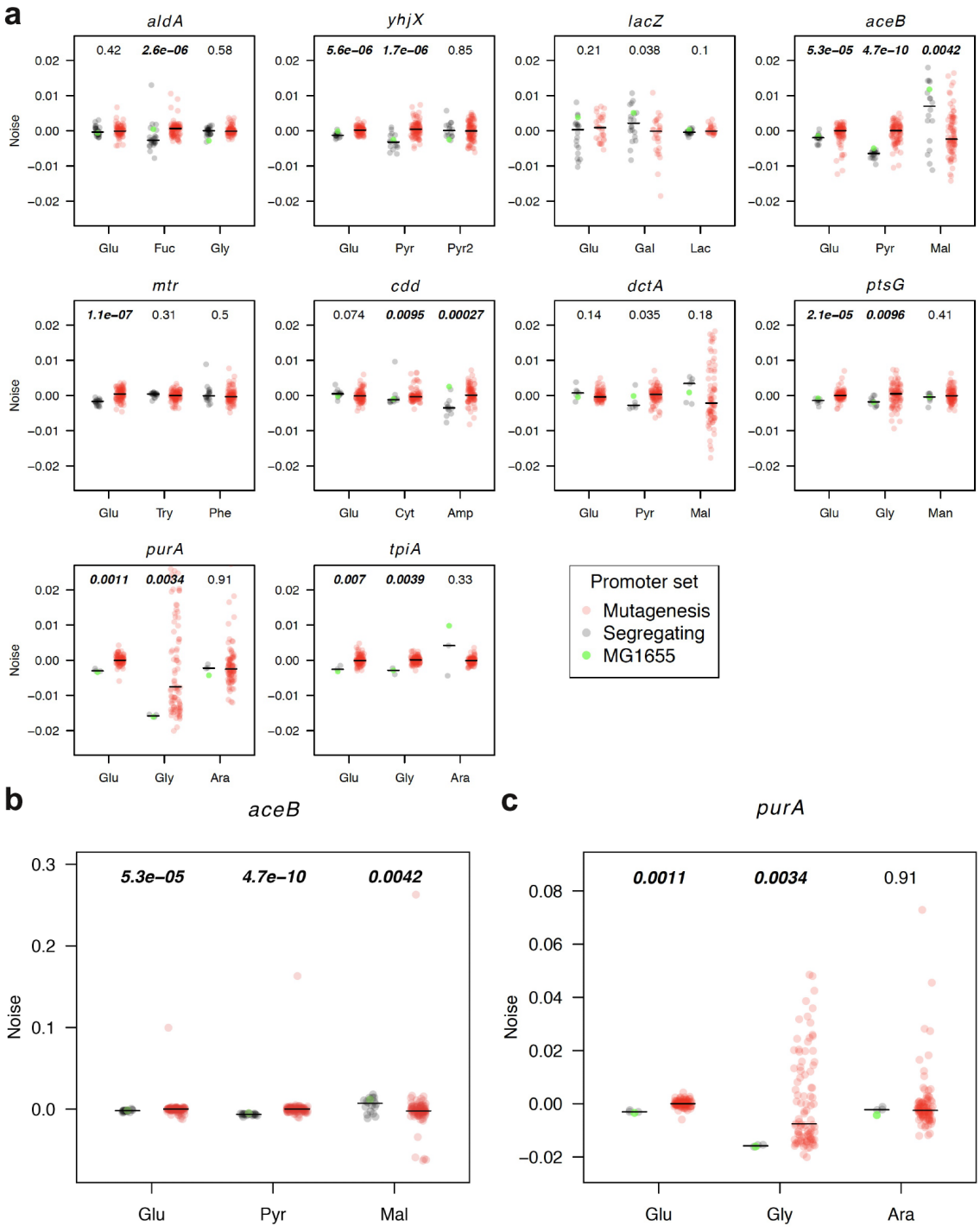

Figure 2.7: **Selection on noise.** Segregating promoter variants are selected such that they have low noise in most environments. Here, noise refers to the deviation from a spline fitted of modal population expression vs. the standard deviation in expression divided by the modal population expression (**Figure S2.5**). **a)** Differences in noise between segregating (black) and random (red) variants. The black lines indicate the median noise value for each group. We tested for differences in noise between segregating and random variants using a two-sided Wilcoxon rank-sum test; the resulting p-values are indicated above each pair. *(continues on the next page)*

Figure 2.7: *(continues from the previous page)* The numbers in bold indicate significant p-values after the Bonferroni correction for multiple comparisons within the promoter. The specific environment in which the promoters were tested is indicated on the x-axis (**Table 2.1**). The panels are arranged in decreasing order of segregating genetic variation (PSS). The MG1655 variant of the *mtr* promoter was omitted from calculation due to a SNP in GFP (**Supplementary Note**). **b)** and **c)** show noise deviations for the *aceB* and *purA* promoters using a wider scale.

regulatory phenotypes are under stabilizing selection, and thus that the mean phenotype of the segregating promoters is close to the optimum. This is clearly supported by the above results, in particular that segregating mutations are enriched for mutations of small effect. Under this assumption, regulatory phenotypes that deviate from this mean are less fit. Thus, we calculated the mean and standard deviation of each phenotype for all segregating variants. We then used these values to calculate a z-score for all variants and phenotypes. We converted these z-scores into absolute values to focus only on the relative difference of the phenotype of the variant, rather than the directionality. Finally, given that we have little information on which of these phenotypes is under stronger selection, we weighed all phenotypes equally, and calculated the sum of absolute values of the z-scores. This metric is indicative of how individual variants differ in their overall phenotype set from the phenotype of the average segregating variant. We found a correlation between the z-scores of different expression levels, as well as between expression levels and plasticity. For expression levels we thus used z-scores from just one out of the three environments. This eliminated the non-independence of individual z-scores among expression levels. However, some non-independence between the z-scores of expression level and plasticity remained (**Materials and Methods**).

Using this metric, we found that in all cases except for *lacZ*, random variants exhibited significantly higher cumulative z-scores than did segregating variants (**Figure 2.8**). Surprisingly, some random promoter variants (e.g. *purA*) exhibited cumulative z-scores over 100, indicating that their phenotypes were more than 100 standard deviations outside of the segregating variants. In the case of both *purA* and *tpiA*, all random variants had cumulative z-scores higher than the highest z-score of any segregating variant. These results highlight that although for individual phenotypes there are not necessarily strong consistent differences between all segregating and all random variants, considering all phenotypes together clearly illustrates that random mutations generally have much larger effects on regulatory phenotypes than do segregating mutations, and that such mutations are effectively filtered by selection.

2.6 Discussion

The effect of selection on gene expression phenotypes has not been well-studied, despite gene expression phenotypes having critical effects on cellular physiology and behaviour. To gain insight into the action of selection on gene expression, here we have compared the phenotypes of segregating variants for ten promoters having different levels of sequence conservation. To understand how selection has acted, we quantified three important phenotypes: modal population expression level, phenotypic plasticity, and noise.

By comparing the variation in expression levels to the level of segregating genetic

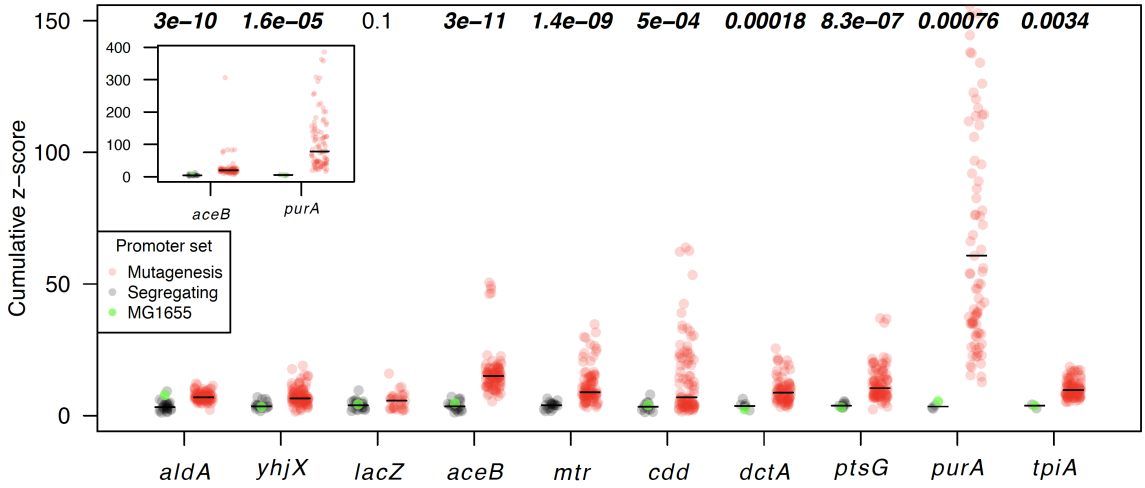

Figure 2.8: **Overall selection pressure.** Comparison of cumulative z-scores of all measured phenotypes (expression level, plasticity, and noise) between segregating and random variants. The numbers above each pair of segregating and random variants indicate the p-values for two-sided Wilcoxon rank-sum test to test for differences between the two groups. The numbers in bold indicate significant p-values. The horizontal black lines indicate the median values of cumulative z-scores of each group. The promoters are arranged in decreasing order of segregating genetic variation (PSS). The MG1655 variant of the *mtr* promoter was omitted from calculation due to a SNP in GFP (**Supplementary Note**). The inset shows the full scale of cumulative z-scores on the y-axis for comparison of high values in *aceB* and *purA* promoters.

variation, we established that promoters with higher levels of genetic variation also have higher levels of phenotypic variation (**Figure 2.3** and **Figure S2.3**). This suggests that the decreased genetic variation observed in some promoters is due to stronger stabilizing selection.

To more directly measure the action of selection, we created a large number of random variants for each of the ten promoters. To create these we used MG1655 variants, which proved to be representative of segregating variants in the majority of cases. As the random variants were produced via PCR random mutagenesis, they had never been subject to selection. We first quantified the effects of these random unselected mutations on expression level. This comparison highlighted the necessity of using multiple environments to understand mutational effects, as we frequently observed environment-specific effects on expression (**Figure 2.4**). Only a small number of studies have examined expression across multiple environments to understand promoter evolution (Duveau et al. 2017, Schaerli et al. 2018, Urchueguía et al. 2019); many more have used only a single environment (Duveau et al. 2018, Hodgins-Davis et al. 2019, Metzger et al. 2015, Schmiedel et al. 2019, Silander et al. 2012). However, bacteria often inhabit multiple different environments.

By comparing expression levels from segregating and random promoter variants, we showed that natural selection has acted such that segregating variants tend to have small-effect mutations. However, the strength of the observed differences between the two groups varies among both promoters and environments. In general, promoters are more robust to mutations affecting expression level when expression is low or completely off (**Figure 2.5**, *yhjX*, *mtr* or *dctA* in glucose). More specifically, we observed few strongly repressed

promoters exhibiting increased expression due to random mutations.

Using a similar strategy to that above, we compared the phenotypic plasticity of segregating promoter variants and random promoter variants. Overall, we found that for five out of ten promoters, random variants exhibited lower plasticity. In these cases, there may be stabilising selection on plasticity in natural populations. However, given that random variants generally exhibited lower plasticity, this leaves open the possibility that there is strong directional selection for regulation to be as plastic as possible, but that higher levels of plasticity are difficult to achieve as there is not sufficient genetic variation.

The random variants of *yhjX*, *lacZ*, *purA*, and *tpiA* did not exhibit different levels of plasticity than segregating variants. Interestingly, three out of four of these promoters are regulated solely via activation (*yhjX*) or repression (*purA* and *tpiA*). This contrasts with all the promoters for which random variants exhibited lower plasticity than segregating variants - all five are regulated by both activation and repression. It is possible that plastic responses from promoters with this simplified architecture are more robust to random mutations, while regulation from promoters having both repressors and activators is easily disrupted.

For *dctA*, we observed significantly higher plasticity for random variants compared to segregating variants in one pair of environments (pyruvic acid vs. L-malic acid, $p = 0.009$, two-sided Wilcoxon rank-sum test; **Figure 2.6e**). This result suggests that the *dctA* promoter is under selection to maintain the same expression level in both environments, and surprisingly, that this regulatory robustness is easily destroyed via very few mutations (**Figure 2.6d**) - most random variants differ from the MG1655 variant by only one or two mutations.

Finally, we have shown that cell-to-cell expression variability, or noise, is often subject to strong selection - as has been shown previously (Hornung et al. 2012, Metzger et al. 2015, Rossi et al. 2019, Schmiedel et al. 2019, Silander et al. 2012, Urchueguía et al. 2019). Importantly, the evidence we present here suggests that selection acts directly to decrease noise, rather than this being a correlated response to other selection pressures. In addition, we found that even single mutations have subtle but detectable effects on noise.

It is very possible that in cases in which we did not find significant changes in a phenotype between segregating and random variants it is possible that those promoters are not optimized for certain phenotypes in the environments considered. This might be true especially in cases when the bacteria do not encounter such environments very often.

We did not observe any strong correlations between the extent of segregating variation (**Figure 2.1**) and the phenotypic consequences of *de novo* random mutations. For example, there was no evidence that random mutations generally had smaller phenotypic effects in highly polymorphic promoters. There are two possible explanations for this pattern. First, although our assays of expression phenotypes can detect differences between random and segregating mutations, there may be more subtle differences in mutational effects that we cannot detect. For example, there may be mutations with small effects that are detected by selection but not in this experimental context. The second possibility is that *de novo* mutations do have similar effects in highly polymorphic and conserved promoters, but selection is not as strong in highly polymorphic promoters, allowing through more small-effect mutations. This is supported by the correlation we observed between genetic variation and phenotypic variation in expression levels (**Figure 2.3** and **Figure S2.3**). To

confirm this hypothesis fitness assays are required. However, previous work suggests that fitness differences can be difficult to detect even when expression levels change by more than ten-fold (Keren et al. 2016). In the majority of results here, random mutations have less than two-fold effects on expression level. Thus, the approach we take here, comparing the phenotypes of segregating and random variants, may be a more powerful method to infer the action of selection than assaying the effects of random mutations on fitness relative to a single wild type.

2.7 Conclusion

We showed that taking into account all the regulatory behaviours we have quantified here - expression level, phenotypic plasticity, and noise - that selection has acted such that segregating SNPs have only minimal phenotypic effects. This conclusion is made more compelling through the fact that the sequences of segregating promoter variants differ from each other by, on average, more than three-fold compared to the sequences of random variants. Thus, there are many single mutations that have large phenotypic effects - yet none of these are segregating.

2.8 Supplementary Information

2.8.1 Supplementary Note

We found that the MG1655 *mtr* variant is expressed at lower levels than all other *mtr* segregating variants when grown in glucose with phenylalanine. In addition, the promoter variants generated via random mutagenesis of the MG1655 variant exhibited expression levels similar to the segregating variants rather than MG1655. We thus sequenced the plasmid containing the MG1655 promoter variant, and found that there was a single non-synonymous mutation in GFP at position 521 changing the codon GGA (G) to GAA (E). A small fraction of other constructs may also contain a mutation outside of the promoter region, but we expect that the large sample sizes we use will negate the effect of this small fraction on our conclusions.

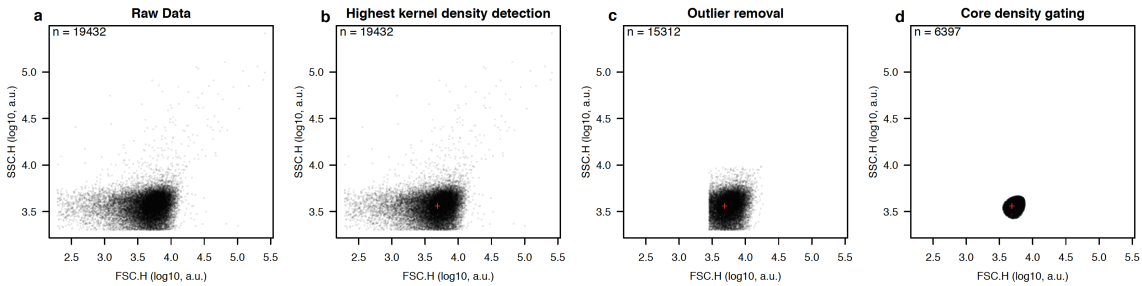

Figure S2.1: **Cell gating strategy from flow cytometry data. a)** Raw data from the flow cytometer of one of the samples. Each point is an individual event recorded by the flow cytometer, the majority of which are expected to be cells. **b)** Identification of the highest kernel density of forward and side scatter values is displayed as the red cross. **c)** Removing events that are too far from the highest kernel density point. This ensured compactness of the final gating step. **d)** Final gating step. The function ellipsoidGate from the flowCore R package was used to isolate the densest homogenous population within the sample.

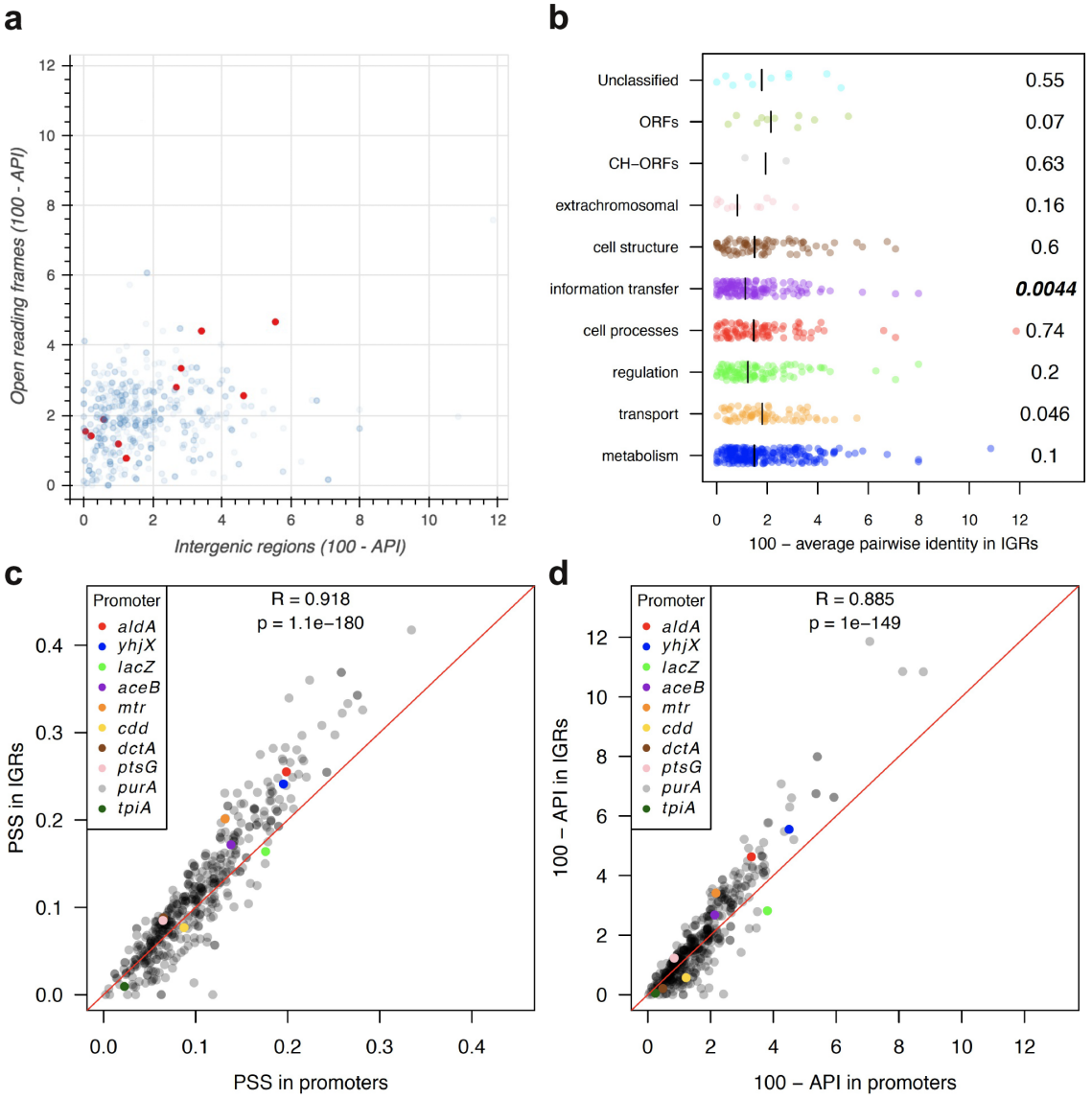

Figure S2.2: **Polymorphisms in intergenic regions (IGRs) and open reading frames (ORFs) across 135 environmental isolates of *E. coli* and MG1655. a)** The inverted average pairwise identity (100 - API) for IGRs and downstream ORFs varies by more than an order of magnitude. Each blue dot indicates an inverted API for an IGR-ORF pair when the IGR contains a transcriptional start site for the ORF. In red are IGRs that we selected for further study due to their different levels of sequence variation (*aldA*, *yhjX*, *lacZ*, *aceB*, *mtr*, *cdd*, *dctA*, *ptsG*, *purA*, and *tpiA*; see **Table 2.2**). **b)** The inverted API for IGRs differs little among different functional groups, as classified by the downstream ORF. CH denotes conserved-hypothetical. Unclassified, ORFs, and CH-ORFs all represent groups of ORFs with very limited information on function (Serres and Riley 2000). Numbers next to each function group represent the p-value obtained via two-sided Wilcoxon rank-sum test to test for differences in API in each group from all other inverted API values for IGRs outside of the group. Bold indicates significant p-values after the Bonferroni correction for multiple comparisons among functional groups. The black lines indicate the mean inverted API values of each functional group. **c)** and **d)** Correlation in sequence variation between IGRs and promoters (IGRs with 100 bp of flanking ORF regions). **c** shows the proportion of segregating sites (PSS) and **d** displays the inverted average pairwise identity (100 - API).

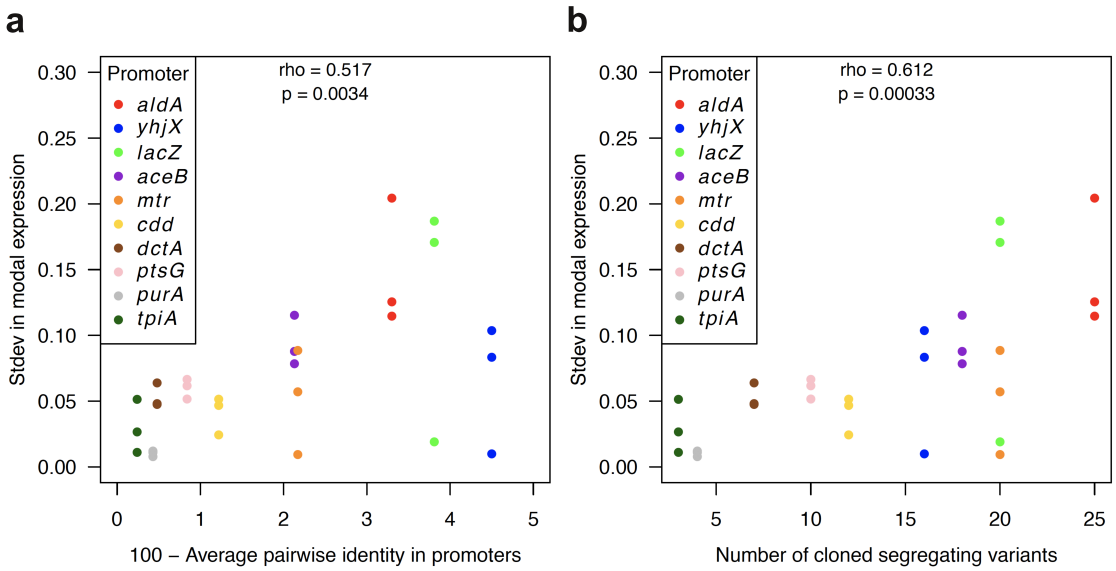

Figure S2.3: **Segregating genetic variation in promoters correlates with variation in expression levels. a)** and **b)** Standard deviations in modal expression levels from segregating variants are correlated with the genetic variation of the promoter (IGR with 100 bp flanking regions). Panel **a** shows the correlation with API and panel **b** shows the correlation with the number of segregating promoter variants cloned and used for phenotypic assay. For each promoter, the standard deviation of modal population expression was measured in three environments (three dots per promoter, **Table 2.1**). The rho and p-values were calculated using Spearman's correlation test.

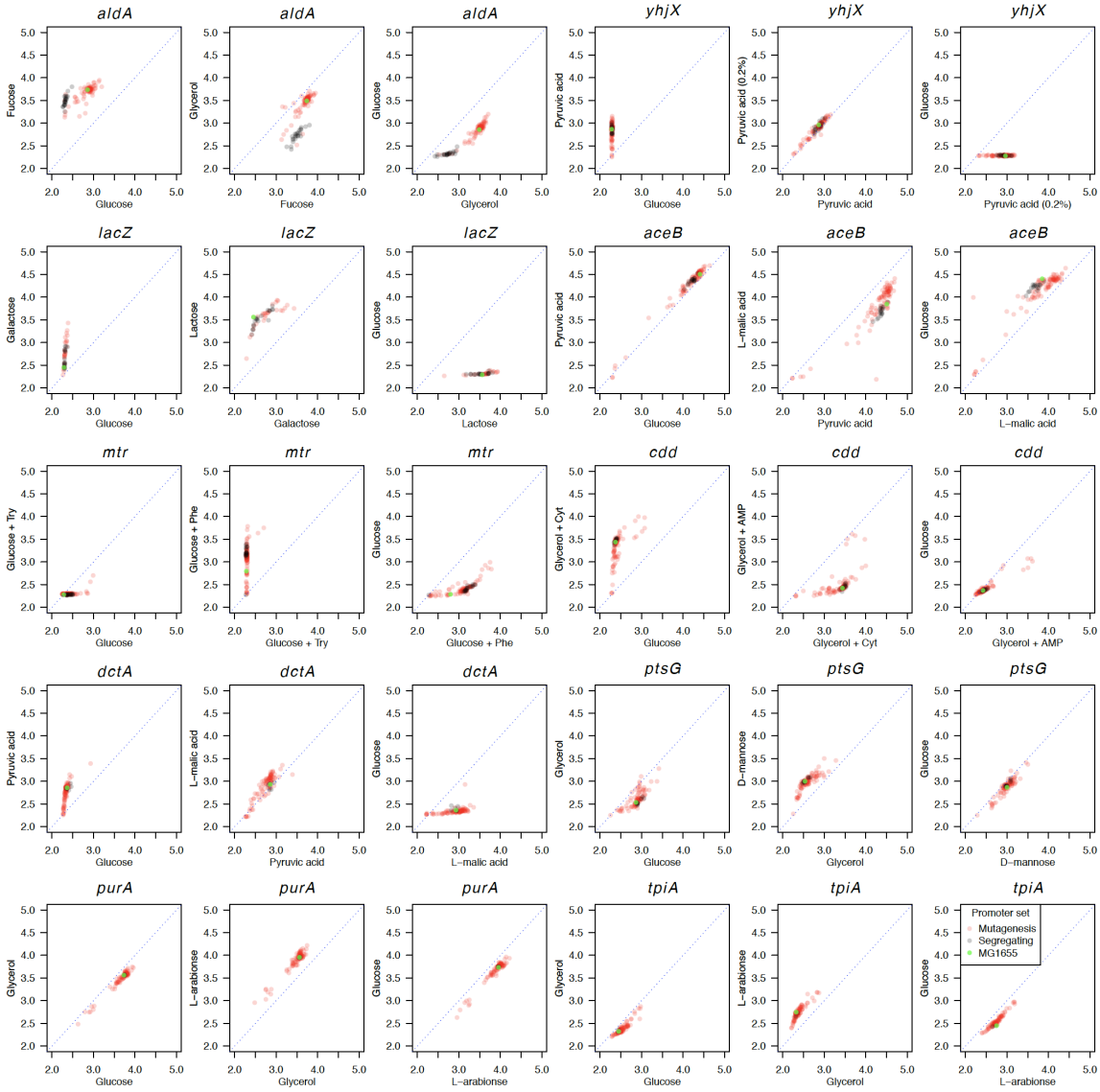

Figure S2.4: **Comparison of expression levels from promoters in pairs of environments.** All x and y axes represent the modal population expression level in the particular environment written on the axis label. The blue dotted lines indicate equal expression levels in both environments, i.e. no phenotypic plasticity. The further from the line a promoter is, the higher the absolute difference is in expression from the promoter between the two environments (i.e., the higher its phenotypic plasticity).

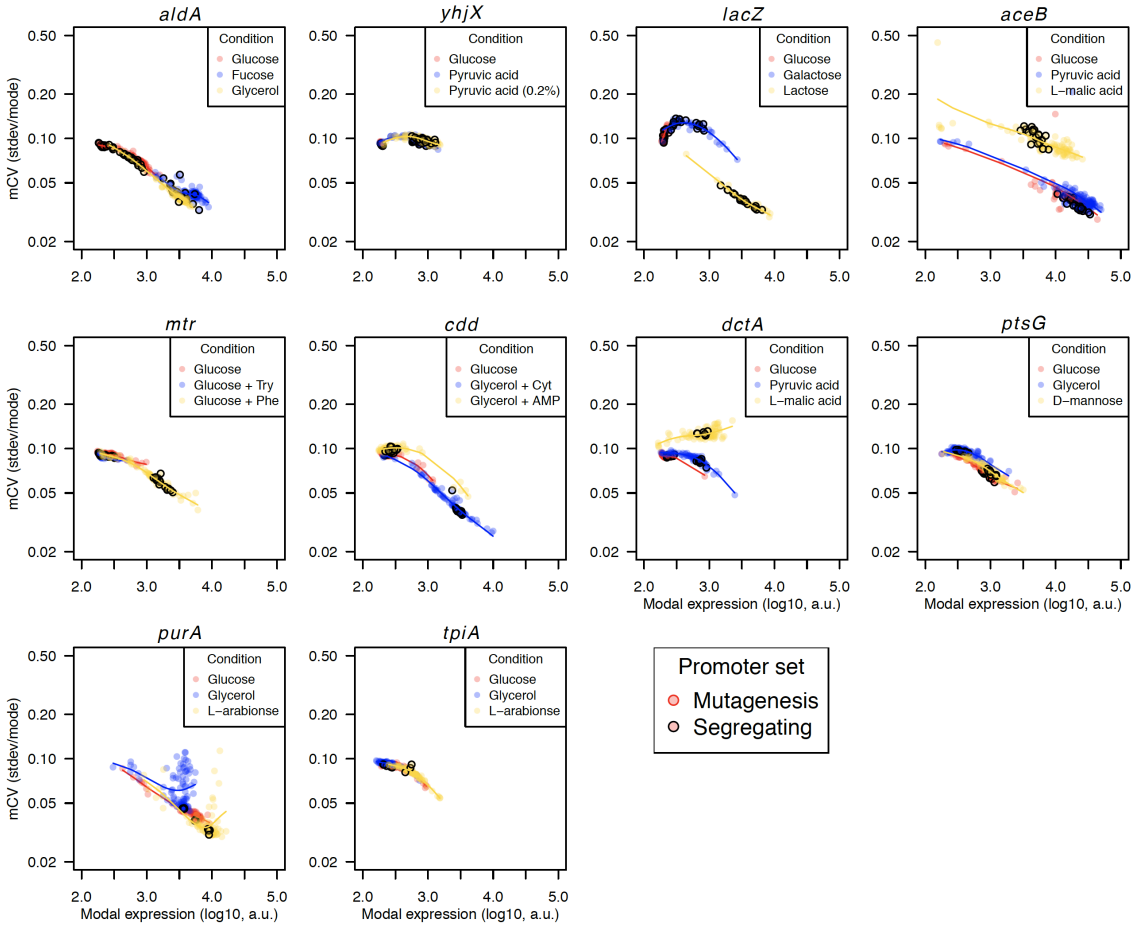

Figure S2.5: **Fits of smoothing splines to modal population expression and modal coefficient of variation (mCV).** A smoothing spline was fitted to all variants (segregating and mutagenized) in each environment. The term "noise" is used for the vertical deviation of each variant, i.e., deviation in the mCV from the fitted spline. The mCV is a measure analogous to the coefficient of variation. It was calculated as a standard deviation of log transformed expression levels (stdev) divided by modal population expression level (mode).

Table S2.1: **Primers used in this work**

Primer ID	Sequence	Purpose
pUA66_insert_F3965	5' - TTG TCT GTT GTG CCC AGT CAT AGC - 3'	PCR & Sanger sequencing
pUA66_insert_R232	5' - TCG CAA AGC ATT GAA GAC CAT ACG C - 3'	PCR & Sanger sequencing
pUA66_vector_F2	5' - GGG ATC CTC TAG ATT TAA GAA GG - 3'	PCR of pUA66 vector for DNA assembly
pUA66_vector_R	5' - TCG AGG TGA AGA CGA AAG G - 3'	PCR of pUA66 vector for DNA assembly
aceB_FastClonIN_F	5' - TTT CGT CTT CAC CTC GAA CCA ACT GGC TCA ACT ATT ACG - 3'	PCR of *aceB* promoter for DNA assembly
aceB_FastClonIN_R	5' - TAA ATC TAG AGG ATC CCA AAT TCT ACC GCT TCG GCA G - 3'	PCR of *aceB* promoter for DNA assembly
aldA_FastClonIN_F	5' - TTT CGT CTT CAC CTC GAG TTC ACG TTG ATT TTC CGG CAG - 3'	PCR of *aldA* promoter for DNA assembly
aldA_FastClonIN_R	5' - TAA ATC TAG AGG ATC CCC ACA TCA ATC CAT GCG TCT CC - 3'	PCR of *aldA* promoter for DNA assembly
cdd_FastClonIN_F	5' - TTT CGT CTT CAC CTC GAA TCG CAC CGT TCC TTT TCC - 3'	PCR of *cdd* promoter for DNA assembly
cdd_FastClonIN_R	5' - TAA ATC TAG AGG ATC CCC TTG TCT GCC AGA ATA GGT TCC - 3'	PCR of *cdd* promoter for DNA assembly
dctA_FastClonIN_F	5' - TAA ATC TAG AGG ATC CCG TTT CAT TTG CTC GCC TAT TTC AGG - 3'	PCR of *dctA* promoter for DNA assembly
dctA_FastClonIN_R	5' - TTT CGT CTT CAC CTC GAA GGG CTT CCT TTT TGC TCG - 3'	PCR of *dctA* promoter for DNA assembly
lacZ_FastClonIN_F	5' - TTT CGT CTT CAC CTC GAC AAT ACG CAA ACC GCC TCT CC - 3'	PCR of *lacZ* promoter for DNA assembly
lacZ_FCINpUA66-D9_R	5' - TAA ATC TAG AGG ATC CCG TGT AAC GCC AGG GTT TTC C - 3'	PCR of *lacZ* promoter for DNA assembly (69A SNP)
lacZ_FCINpUA66-K12_R	5' - TAA ATC TAG AGG ATC CCG GGT AAC GCC AGG GTT TTC C - 3'	PCR of *lacZ* promoter for DNA assembly (69C SNP)
mtr_FastClonIN_F	5' - TAA ATC TAG AGG ATC CCA CAT CCC TGC GCC AAT AAT GG - 3'	PCR of *mtr* promoter for DNA assembly
mtr_FastClonIN_R	5' - TTT CGT CTT CAC CTC GAT TTT AGC GGC GAA CGT CGT G - 3'	PCR of *mtr* promoter for DNA assembly
ptsG_FastClonIN_F	5' - TTT CGT CTT CAC CTC GAT CTG TTT CAC ATC GAC GCT TCC - 3'	PCR of *ptsG* promoter for DNA assembly
ptsG_FastClonIN_R	5' - TAA ATC TAG AGG ATC CCC CAG CAG AAT ACC TGC GAT AGG - 3'	PCR of *ptsG* promoter for DNA assembly

Continued on the next page

Table S2.1 Primers used in this work – continuation

Primer ID	**Sequence**	**Purpose**
purA_FastClonIN_F	5' - TTT CGT CTT CAC CTC GAA TAT TTT ACG TCG TTT TGG CGG TGG - 3'	PCR of *purA* promoter for DNA assembly
purA_FastClonIN_R	5' - TAA ATC TAG AGG ATC CCT TTA GCC CGT TCA GTC AGA AGA TCG - 3'	PCR of *purA* promoter for DNA assembly
tpiA_FastClonIN_F	5' - TAA ATC TAG AGG ATC CCC AAC ACC TGC CAG CTC TTT ACG - 3'	PCR of *tpiA* promoter for DNA assembly
tpiA_FastClonIN_R	5' - TTT CGT CTT CAC CTC GAT GCA CTG CGT TAT GTT GTC GC - 3'	PCR of *tpiA* promoter for DNA assembly
yhjX_FastClonIN_F	5' - TAA ATC TAG AGG ATC CCC TCC AGG TAT AAA CCG ACC C - 3'	PCR of *yhjX* promoter for DNA assembly
yhjX_FastClonIn-A1_R	5' - TTT CGT CTT CAC CTC GAA ATA TTG CTG CCT ATG CTG C - 3'	PCR of *yhjX* promoter for DNA assembly (615A & 624A SNPs)
yhjX_FastClonIN-K12 _R	5' - TTT CGT CTT CAC CTC GAA ATA TTG CCG CCT ATG CCG - 3'	PCR of *yhjX* promoter for DNA assembly (615G & 624G SNPs)

Note: Underscored sequences = regions homologous to PCR amplified pUA66 vector. In the case of the *lacZ* and *yhjX* promoter, two versions of reverse primer exist differing by a single SNP in downstream *lacZ* gene (C69A) and two SNPs in upstream *yhjY* gene (G615A and G624A), respectively.

Table S2.2: **Genetic variability of IGRs present in the majority of environmental isolates**

Promoter	IGR present in N isolates	N of IGR versions	Inverted API of IGRs	PSS of IGRs	IGR length	Inverted API of ORFs	PSS of ORFs
accB	135	12	0.64	0.064	407	0.81	0.045
accD	136	4	0.23	0.019	155	1.52	0.084
aceB	136	24	2.68	0.172	268	2.80	0.161
aceE	136	9	0.49	0.066	166	1.20	0.081
acnB	135	17	1.18	0.080	361	2.77	0.128
acpP	136	3	0.24	0.024	210	0.00	0.000
acrA	136	2	0.35	0.028	141	1.57	0.102
acrZ	136	6	0.80	0.086	128	0.29	0.027
acs	136	18	1.09	0.109	393	2.57	0.131
ada	136	13	2.79	0.205	73	4.48	0.224
ade	136	5	0.30	0.034	175	2.63	0.143
adhE	136	15	1.63	0.107	478	1.33	0.081
adiA	136	21	3.28	0.283	198	2.11	0.169
agaR	138	23	1.88	0.133	248	2.41	0.130
ahpC	131	10	0.90	0.129	373	0.54	0.041
aidB	136	7	0.95	0.060	83	2.46	0.151
aldA	136	28	4.63	0.255	196	2.56	0.171
alkA	136	13	2.10	0.098	133	3.35	0.221
amiA	136	9	1.21	0.080	213	2.92	0.140
ampC	136	6	2.65	0.129	62	3.36	0.162
ansB	136	9	0.82	0.069	175	2.84	0.149
apt	136	5	0.72	0.066	152	1.72	0.125
araB	136	14	1.11	0.080	338	2.68	0.150
araC	136	14	1.11	0.080	338	2.42	0.122
araE	136	25	3.35	0.200	314	2.14	0.120
araJ	136	10	1.39	0.088	125	2.68	0.138
arcA	136	25	1.52	0.115	399	0.18	0.021
arfA	136	7	3.57	0.127	55	2.20	0.128
argC	136	7	0.63	0.059	153	2.31	0.126
argD	136	12	4.01	0.153	85	3.00	0.160
argE	136	7	0.63	0.059	153	2.89	0.148
argI	136	16	1.80	0.118	161	4.61	0.202
argR	136	9	0.85	0.069	434	1.34	0.089
argS	136	18	1.58	0.149	215	2.22	0.115
argX	136	6	0.66	0.078	102	0.00	0.000
aroF	136	27	5.78	0.230	213	1.63	0.115

Continued on the next page

Table S2.2 Genetic variability of IGRs present in the majority of isolates – continuation

Promoter	IGR present in N isolates	N of IGR versions	Inverted API of IGRs	PSS of IGRs	IGR length	Inverted API of ORFs	PSS of ORFs
aroG	136	8	1.49	0.079	315	2.72	0.142
aroH	134	14	3.49	0.231	156	1.95	0.136
aroK	136	9	1.04	0.085	400	0.48	0.029
aroL	136	13	2.18	0.126	182	1.93	0.135
artP	136	13	1.39	0.120	217	1.11	0.089
asnA	136	20	2.64	0.159	151	2.30	0.119
asnC	136	20	2.64	0.159	151	2.22	0.122
aspA	136	10	1.06	0.060	336	0.70	0.052
asr	134	18	3.24	0.212	424	2.90	0.168
atpI	136	18	1.18	0.070	616	0.31	0.037
azoR	136	13	1.40	0.105	200	2.87	0.172
azuC	133	36	5.21	0.326	267	2.05	0.115
bcsE	136	14	2.85	0.180	272	2.48	0.155
bfr	131	6	11.86	0.360	75	7.58	0.239
bhsA	136	15	1.43	0.108	240	1.33	0.078
bolA	136	14	1.60	0.112	304	0.63	0.053
bssS	136	15	1.20	0.080	289	0.44	0.027
cadB	135	27	2.87	0.186	365	1.07	0.059
cadC	135	20	2.93	0.176	256	1.51	0.128
caiF	135	11	1.78	0.137	262	1.78	0.098
caiT	135	33	4.66	0.192	473	1.83	0.106
can	136	2	0.01	0.009	108	1.40	0.087
carA	136	40	5.48	0.268	456	2.18	0.113
cbl	136	14	2.25	0.198	101	2.18	0.178
cdd	136	12	0.57	0.077	130	1.88	0.118
cfa	136	9	1.04	0.076	290	1.92	0.122
cirA	136	19	1.74	0.142	295	1.50	0.121
clpP	136	6	0.39	0.033	246	1.12	0.056
cmk	136	6	0.53	0.052	172	1.54	0.080
copA	135	16	2.05	0.129	263	3.46	0.153
cpdB	136	7	2.51	0.185	189	3.05	0.160
cpxP	136	2	0.17	0.013	149	0.55	0.042
crp	135	8	0.64	0.063	301	1.02	0.073
csgB	136	31	2.91	0.210	758	1.16	0.086
csgD	136	31	2.91	0.210	758	1.70	0.118
cspA	136	8	1.18	0.089	282	1.38	0.131
cspD	135	11	0.74	0.056	322	0.32	0.027

Continued on the next page

Table S2.2 Genetic variability of IGRs present in the majority of isolates – continuation

Promoter	IGR present in N isolates	N of IGR versions	Inverted API of IGRs	PSS of IGRs	IGR length	Inverted API of ORFs	PSS of ORFs
cspE	136	17	3.38	0.178	174	6.76	0.119
csrA	136	4	0.71	0.051	234	0.00	0.000
csrB	136	14	3.05	0.186	210	0.75	0.062
cstA	136	23	3.42	0.178	180	2.97	0.141
cusC	136	13	2.22	0.109	156	2.80	0.156
cusR	136	13	2.22	0.109	156	2.94	0.143
cutA	136	10	2.04	0.097	124	1.59	0.112
cyaA	136	12	1.08	0.075	386	2.79	0.142
cydD	136	7	1.14	0.082	122	2.43	0.130
cyoA	136	39	3.65	0.179	464	0.87	0.052
cysD	136	9	0.57	0.060	252	1.99	0.123
cysJ	136	15	2.18	0.130	315	3.50	0.182
cysK	136	10	1.54	0.098	184	1.83	0.109
cytR	136	11	2.22	0.088	159	1.89	0.116
dacA	136	4	0.37	0.036	139	1.94	0.122
dadA	136	14	1.25	0.094	329	2.41	0.128
dam	136	7	1.48	0.104	106	1.21	0.087
dapA	136	3	0.60	0.034	146	0.96	0.065
dctA	136	5	0.21	0.088	182	1.41	0.109
dcuA	136	8	1.65	0.077	117	1.10	0.079
dcuB	135	26	3.00	0.215	572	1.71	0.134
dcuS	135	15	3.65	0.322	180	2.71	0.204
def	136	4	0.27	0.030	132	0.85	0.057
degQ	136	9	0.83	0.078	167	2.25	0.146
degS	136	6	1.51	0.079	89	2.81	0.140
deoB	136	4	0.63	0.039	51	1.95	0.100
deoC	136	12	0.81	0.074	257	2.25	0.100
dinI	136	8	2.00	0.149	74	1.66	0.126
dksA	136	4	0.43	0.023	177	0.39	0.031
dld	136	10	0.79	0.056	195	2.69	0.147
dmsA	134	13	1.49	0.113	238	2.67	0.139
dps	136	16	3.24	0.197	300	1.41	0.097
dsrA	136	20	4.49	0.270	200	0.60	0.057
dusB	136	19	3.19	0.159	328	0.37	0.031
edd	136	13	2.04	0.141	234	1.80	0.113
eno	136	7	0.59	0.057	87	0.93	0.048
entS	136	5	0.31	0.036	110	3.41	0.177

Continued on the next page

Table S2.2 Genetic variability of IGRs present in the majority of isolates – continuation

Promoter	IGR present in N isolates	N of IGR versions	Inverted API of IGRs	PSS of IGRs	IGR length	Inverted API of ORFs	PSS of ORFs
fabB	136	11	1.21	0.101	158	2.01	0.089
fadB	136	15	1.38	0.106	189	2.38	0.135
fadD	136	13	1.49	0.146	205	2.55	0.136
fadR	135	9	1.78	0.108	223	1.71	0.086
fdnG	136	11	1.89	0.160	231	2.25	0.294
fepA	136	18	3.12	0.213	244	2.99	0.140
fepD	136	5	0.31	0.036	110	3.21	0.157
fes	136	18	3.12	0.213	244	3.37	0.193
fhuF	136	23	2.29	0.193	140	4.41	0.229
fieF	136	5	0.68	0.037	81	1.94	0.137
fixA	135	33	4.66	0.192	473	2.25	0.119
fldB	136	6	1.24	0.099	111	1.52	0.107
fliL	136	8	1.10	0.125	104	2.27	0.144
fnr	136	12	1.43	0.113	194	1.27	0.072
focA	135	17	0.92	0.044	406	1.18	0.061
folA	136	12	2.69	0.212	193	2.42	0.121
frdA	136	12	1.35	0.083	324	1.31	0.071
frsA	136	3	1.11	0.065	92	2.44	0.140
ftnA	136	9	2.09	0.101	179	1.33	0.106
ftsH	136	1	0.00	0.000	99	1.63	0.107
fumA	136	7	0.42	0.040	198	11.06	0.302
galE	130	21	4.29	0.210	262	1.58	0.086
galR	136	16	1.49	0.124	217	2.35	0.149
galS	136	7	0.73	0.064	141	2.22	0.144
gapA	136	18	0.86	0.073	341	0.40	0.035
gcd	136	11	1.12	0.083	205	2.88	0.140
gcvA	136	8	1.05	0.055	128	1.76	0.103
gcvR	136	3	0.60	0.034	146	1.32	0.091
gcvT	136	24	4.85	0.186	290	2.26	0.116
glcC	132	18	1.50	0.108	251	1.71	0.116
glcD	132	18	1.50	0.108	251	2.39	0.126
glgB	136	9	1.13	0.066	272	2.37	0.138
glk	136	5	0.68	0.039	203	2.54	0.139
glmY	136	33	6.30	0.280	346	0.67	0.054
glnA	136	14	1.03	0.040	372	1.94	0.100
glnB	135	2	0.21	0.017	60	2.17	0.094
glpD	136	11	1.15	0.079	189	3.58	0.165

Continued on the next page

Table S2.2 Genetic variability of IGRs present in the majority of isolates – continuation

Promoter	IGR present in N isolates	N of IGR versions	Inverted API of IGRs	PSS of IGRs	IGR length	Inverted API of ORFs	PSS of ORFs
glpF	136	17	0.51	0.045	426	1.13	0.070
glrK	136	3	1.56	0.103	58	2.82	0.148
gltB	136	31	2.94	0.176	676	3.19	0.152
gltX	136	11	0.89	0.066	258	2.24	0.117
glyA	136	20	1.49	0.130	332	1.71	0.100
glyQ	136	3	0.28	0.032	95	1.02	0.063
gmk	136	17	3.18	0.140	257	1.02	0.063
gntK	136	12	1.23	0.086	139	2.11	0.106
gntP	136	29	2.19	0.156	340	2.94	0.150
gpmA	136	15	3.46	0.181	204	2.05	0.098
grcA	136	24	3.18	0.145	304	1.28	0.060
greA	135	11	0.79	0.073	248	0.92	0.067
groS	136	11	1.47	0.076	276	0.23	0.031
guaB	136	7	0.69	0.055	164	3.19	0.130
hdeA	136	13	1.81	0.122	254	1.41	0.075
hdeD	136	13	1.81	0.122	254	2.55	0.131
hemA	136	8	1.16	0.066	213	3.48	0.150
hemB	136	39	10.85	0.308	532	1.96	0.105
hemN	136	5	0.53	0.048	188	2.16	0.133
hepA	136	4	1.03	0.079	164	2.27	0.120
hfq	136	2	0.15	0.012	85	0.42	0.027
hisJ	136	12	1.46	0.132	220	2.28	0.126
hisS	136	5	0.96	0.045	110	1.56	0.085
hns	136	21	1.86	0.134	605	0.78	0.065
hpt	136	11	1.12	0.083	205	1.11	0.067
hscB	136	5	0.49	0.084	95	1.34	0.079
hupB	136	4	1.00	0.072	208	11.82	0.179
hybO	136	5	0.58	0.042	189	1.98	0.105
icdA	136	8	0.72	0.045	177	2.35	0.110
ihfB	136	4	0.25	0.019	159	0.32	0.018
ilvC	136	14	1.83	0.141	149	3.37	0.146
ilvIH	136	12	2.82	0.151	119	3.52	0.172
ilvL	136	25	4.07	0.244	353	0.22	0.010
ilvY	136	14	1.83	0.141	149	6.07	0.220
imp	136	5	0.63	0.039	254	2.56	0.171
infA	136	8	0.36	0.039	284	0.18	0.023
iraP	136	25	3.72	0.248	464	1.85	0.092

Continued on the next page

Table S2.2 Genetic variability of IGRs present in the majority of isolates – continuation

Promoter	IGR present in N isolates	N of IGR versions	Inverted API of IGRs	PSS of IGRs	IGR length	Inverted API of ORFs	PSS of ORFs
katG	136	21	4.01	0.252	341	3.29	0.177
kbaZ	138	23	1.88	0.133	248	2.76	0.151
kdpF	136	25	1.91	0.121	314	1.16	0.044
lacZ	136	17	2.82	0.164	122	3.34	0.184
lapA	136	9	1.29	0.128	148	1.93	0.120
leuL	135	29	4.10	0.199	672	1.91	0.092
leuO	135	29	4.10	0.199	672	2.54	0.138
lexA	136	3	0.34	0.018	109	0.99	0.072
lgt	136	5	0.44	0.027	150	1.47	0.100
livK	136	26	2.17	0.128	423	4.12	0.292
loiP	136	10	0.57	0.051	277	3.22	0.177
lolA	136	19	3.55	0.148	169	1.70	0.119
lon	136	3	0.42	0.037	187	1.35	0.085
lrp	136	15	0.76	0.071	546	0.67	0.036
lysP	136	12	1.57	0.122	205	1.86	0.129
lysU	135	17	1.95	0.106	236	10.89	0.271
malE	136	18	1.82	0.137	364	2.21	0.134
malI	136	9	0.51	0.069	174	2.74	0.169
malP	136	21	3.10	0.194	623	2.73	0.163
malT	136	21	3.10	0.194	623	2.66	0.152
malX	136	9	0.51	0.069	174	2.20	0.135
manA	136	7	0.42	0.040	198	2.47	0.139
manX	135	7	0.85	0.060	266	0.94	0.062
mdh	136	9	0.85	0.069	434	1.64	0.092
mdtJ	136	18	1.99	0.131	411	2.70	0.169
melA	135	28	7.99	0.369	282	2.16	0.122
melR	135	28	7.99	0.369	282	1.62	0.109
menA	136	1	0.00	0.000	66	1.96	0.110
metA	136	12	1.88	0.141	156	2.30	0.123
metB	136	18	3.38	0.138	276	2.09	0.109
metG	134	6	0.60	0.076	132	2.78	0.147
metJ	136	18	3.38	0.138	276	0.76	0.041
metN	136	16	1.38	0.091	187	1.76	0.115
mfd	136	9	0.80	0.094	127	2.83	0.145
mgrR	136	3	1.51	0.133	45	0.52	0.051
mgtA	136	11	2.33	0.147	184	0.13	0.037
mlc	136	4	0.96	0.090	134	2.65	0.152

Continued on the next page

Table S2.2 Genetic variability of IGRs present in the majority of isolates – continuation

Promoter	IGR present in N isolates	N of IGR versions	Inverted API of IGRs	PSS of IGRs	IGR length	Inverted API of ORFs	PSS of ORFs
mlrA	136	9	0.66	0.107	225	1.94	0.154
mnmG	136	28	1.25	0.114	378	1.91	0.113
mntH	136	20	1.90	0.104	338	2.71	0.158
mntP	136	31	2.06	0.128	430	1.87	0.123
modA	136	20	3.30	0.162	167	2.96	0.143
moeA	136	13	1.02	0.093	204	3.85	0.201
mpl	136	9	0.90	0.051	175	2.40	0.130
mprA	136	3	0.03	0.022	90	0.70	0.045
mraZ	136	20	4.36	0.297	602	0.94	0.072
mreB	136	4	0.64	0.046	304	1.36	0.096
mrp	134	6	0.60	0.076	132	3.28	0.243
mtn	136	3	0.54	0.048	83	1.90	0.137
mtr	136	16	3.41	0.201	154	4.41	0.168
nadE	136	10	0.90	0.080	201	2.54	0.151
nagB	136	15	0.55	0.062	340	1.38	0.085
nagE	136	15	0.55	0.062	340	2.50	0.124
nanA	136	7	0.77	0.090	123	2.89	0.138
napF	136	3	0.46	0.028	108	1.52	0.097
narK	136	18	1.13	0.092	338	1.92	0.109
nemR	136	11	2.16	0.117	103	3.16	0.190
nfo	136	3	0.19	0.027	73	1.74	0.098
nfsB	136	7	4.13	0.194	93	2.06	0.130
nhaB	135	9	1.78	0.108	223	2.67	0.160
nirB	136	23	3.65	0.218	261	1.94	0.103
nlpD	136	7	1.01	0.072	139	1.36	0.089
nrdA	136	10	2.51	0.141	128	2.31	0.121
nrdD	136	28	4.63	0.270	396	2.28	0.116
nrdR	136	8	1.10	0.080	150	0.98	0.076
nrfA	136	18	1.09	0.109	393	1.73	0.101
nuoA	130	30	3.17	0.191	632	0.70	0.047
nupC	136	20	1.90	0.104	338	1.83	0.104
nupG	136	16	2.87	0.204	201	1.82	0.099
ompA	136	15	1.02	0.084	356	2.74	0.091
ompC	136	19	1.33	0.093	335	5.73	0.228
ompF	135	31	3.98	0.224	604	4.02	0.136
ompR	136	5	0.35	0.035	227	2.25	0.093
ompX	136	7	1.61	0.124	234	0.61	0.045

Continued on the next page

Table S2.2 Genetic variability of IGRs present in the majority of isolates – continuation

Promoter	IGR present in N isolates	N of IGR versions	Inverted API of IGRs	PSS of IGRs	IGR length	Inverted API of ORFs	PSS of ORFs
opgG	136	14	1.54	0.117	394	1.62	0.108
osmB	136	26	1.86	0.146	268	2.00	0.132
osmC	136	26	6.61	0.297	364	2.43	0.123
osmE	136	10	0.90	0.080	201	0.86	0.065
osmY	136	21	1.64	0.109	402	1.78	0.116
oxyR	136	7	2.05	0.084	95	2.05	0.114
pabA	136	3	1.66	0.065	31	2.92	0.163
panD	136	20	3.95	0.179	273	2.02	0.097
pck	136	19	1.52	0.095	378	2.46	0.117
pcnB	136	2	0.21	0.017	59	2.63	0.124
pepA	136	5	0.11	0.019	266	2.92	0.141
pepD	136	6	0.50	0.050	260	2.24	0.124
pfkA	136	6	1.52	0.110	182	1.65	0.105
pflA	136	8	0.39	0.037	191	2.02	0.105
pgpB	136	5	1.50	0.112	169	2.16	0.144
pheL	136	4	0.80	0.049	103	1.09	0.104
phoE	136	18	1.02	0.080	288	3.30	0.145
pka	136	4	0.46	0.097	31	2.86	0.158
plsX	136	4	0.33	0.025	80	0.95	0.061
pncB	136	7	0.72	0.068	265	2.13	0.127
poxB	130	16	1.81	0.121	132	3.79	0.176
ppiA	136	13	1.87	0.122	270	1.77	0.129
ppiD	130	6	1.17	0.084	191	1.82	0.104
pps	135	21	1.41	0.096	332	2.11	0.117
proP	136	14	1.91	0.175	263	1.84	0.156
proS	136	8	0.53	0.045	111	1.91	0.109
proV	135	23	2.22	0.115	357	1.75	0.092
pspF	136	11	0.82	0.054	166	2.93	0.164
pssA	136	2	0.05	0.009	113	1.54	0.091
pth	136	17	1.88	0.126	277	1.43	0.099
ptrA	136	17	3.54	0.160	175	2.78	0.143
ptsG	136	10	1.23	0.085	294	0.77	0.066
ptsH	136	8	0.82	0.060	383	0.19	0.016
purA	136	4	1.00	0.068	103	1.18	0.069
purC	136	23	2.29	0.137	212	1.20	0.080
purL	136	10	1.50	0.105	258	3.22	0.154
purM	136	24	3.86	0.216	324	2.02	0.110

Continued on the next page

Table S2.2 Genetic variability of IGRs present in the majority of isolates – continuation

Promoter	IGR present in N isolates	N of IGR versions	Inverted API of IGRs	PSS of IGRs	IGR length	Inverted API of ORFs	PSS of ORFs
purR	136	5	0.61	0.044	298	1.86	0.103
putA	136	12	1.34	0.085	142	2.97	0.152
putP	136	6	0.43	0.051	158	2.90	0.135
pykA	136	6	2.50	0.197	127	2.44	0.137
pykF	136	19	2.42	0.154	558	1.61	0.109
pyrC	136	4	0.47	0.028	106	3.13	0.177
pyrF	136	20	3.69	0.202	193	2.68	0.164
pyrH	136	8	1.32	0.055	146	0.96	0.069
qseB	136	6	2.03	0.126	151	2.66	0.171
rbsD	136	14	1.94	0.162	167	1.44	0.112
rcnA	136	10	1.33	0.124	121	2.07	0.161
rcnR	136	10	1.33	0.124	121	1.38	0.128
rcsA	136	18	1.95	0.116	294	1.06	0.056
rcsD	136	25	3.35	0.240	312	2.10	0.132
recA	135	3	0.11	0.025	79	1.36	0.090
rhaB	136	18	2.63	0.171	287	2.66	0.155
rhaS	136	18	2.63	0.171	287	2.72	0.186
rhaT	136	14	1.04	0.081	285	2.56	0.134
ribA	136	5	1.50	0.112	169	1.23	0.076
rihB	133	11	1.82	0.113	168	3.59	0.169
rimP	136	8	0.21	0.029	206	0.35	0.022
rlmE	136	5	0.61	0.024	126	0.58	0.041
rnb	136	3	0.58	0.060	67	1.68	0.115
rnc	136	12	1.72	0.121	272	1.54	0.087
rne	135	7	2.26	0.120	117	0.33	0.031
rplU	136	5	0.25	0.012	258	0.14	0.013
rpoE	136	18	2.64	0.149	255	0.25	0.019
rpoH	136	9	0.86	0.057	244	1.79	0.088
rpoZ	136	2	0.46	0.037	54	0.22	0.018
rpsF	136	22	5.77	0.211	327	0.39	0.043
rpsM	136	2	0.09	0.007	146	0.06	0.011
rpsU	136	4	0.00	0.000	241	0.12	0.009
rraB	136	16	1.80	0.118	161	1.92	0.120
rsd	136	5	0.59	0.064	94	1.67	0.094
rybA	131	5	1.03	0.076	118	1.62	0.106
ryfD	136	1	0.00	0.000	7	0.00	0.000
ryhB	136	8	0.37	0.056	125	0.14	0.011

Continued on the next page

Table S2.2 Genetic variability of IGRs present in the majority of isolates – continuation

Promoter	IGR present in N isolates	N of IGR versions	Inverted API of IGRs	PSS of IGRs	IGR length	Inverted API of ORFs	PSS of ORFs
sdhC	136	13	0.77	0.053	395	0.60	0.046
sdiA	135	3	0.59	0.057	88	2.17	0.149
secG	136	19	1.64	0.141	227	1.59	0.087
serA	134	19	3.55	0.232	259	1.83	0.120
serC	136	11	3.10	0.202	198	2.73	0.133
shiA	131	29	1.84	0.127	314	2.32	0.145
sixA	136	13	0.88	0.074	202	2.64	0.148
sodA	136	14	1.04	0.081	285	1.77	0.110
sodB	136	3	0.50	0.047	127	1.16	0.082
sohB	136	10	1.54	0.132	219	2.32	0.143
soxR	136	1	0.00	0.000	85	2.71	0.135
soxS	136	1	0.00	0.000	85	2.33	0.123
spy	136	19	1.55	0.112	329	0.91	0.082
srkA	136	7	1.47	0.079	76	1.90	0.120
srlA	136	21	2.71	0.148	257	2.31	0.131
ssb	136	8	1.15	0.091	254	1.66	0.097
ssrA	135	20	7.08	0.340	215	0.16	0.011
ssrS	136	2	0.30	0.024	41	0.17	0.016
sufA	131	22	1.73	0.140	328	2.18	0.117
sulA	136	21	1.27	0.092	218	1.73	0.135
sutR	136	11	2.67	0.187	91	3.86	0.210
sxy	136	21	1.27	0.092	218	1.11	0.097
tatA	136	4	0.73	0.038	78	0.25	0.019
tig	136	4	0.26	0.020	343	0.76	0.051
tomB	136	25	3.44	0.222	567	1.26	0.096
tonB	136	19	2.04	0.148	223	2.12	0.132
topA	136	13	0.56	0.053	379	1.96	0.107
torC	136	8	1.21	0.093	129	1.86	0.110
torR	136	8	1.21	0.093	129	2.59	0.139
tpiA	136	2	0.05	0.009	107	1.54	0.070
tppB	135	27	3.59	0.239	610	1.74	0.104
tpx	136	5	0.42	0.051	118	1.66	0.112
treB	136	5	1.92	0.161	118	2.35	0.127
treR	136	11	2.33	0.147	184	2.54	0.143
trpL	136	16	1.09	0.095	137	0.29	0.022
trpR	136	10	1.51	0.138	94	1.80	0.101
trxA	136	4	0.31	0.023	130	0.80	0.052

Continued on the next page

Table S2.2 Genetic variability of IGRs present in the majority of isolates – continuation

Promoter	IGR present in N isolates	N of IGR versions	Inverted API of IGRs	PSS of IGRs	IGR length	Inverted API of ORFs	PSS of ORFs
trxC	136	11	0.99	0.063	206	1.46	0.105
tsx	136	20	1.99	0.091	298	1.88	0.084
typA	136	14	1.03	0.040	372	0.93	0.060
tyrB	135	14	4.28	0.266	252	2.75	0.168
tyrP	136	27	3.94	0.219	192	2.90	0.149
tyrR	136	7	0.90	0.088	147	3.24	0.177
udpP	136	14	2.24	0.142	261	1.89	0.105
ulaA	136	35	4.30	0.212	354	2.76	0.155
upp	136	24	3.86	0.216	324	1.66	0.104
uspA	136	8	0.48	0.044	390	1.32	0.094
uvrA	136	8	1.15	0.091	254	3.19	0.141
uvrB	136	27	3.27	0.213	578	3.36	0.165
uvrD	136	5	0.75	0.084	83	3.41	0.152
uvrY	136	14	1.95	0.108	296	1.16	0.088
uxaB	136	22	5.22	0.282	227	2.69	0.149
uxuA	136	29	2.19	0.156	340	1.86	0.100
valV	136	14	2.20	0.162	308	1.11	0.078
waaA	136	20	6.75	0.343	458	2.42	0.135
wza	132	38	2.45	0.231	674	2.23	0.137
xerD	136	6	1.24	0.099	111	2.29	0.155
yacC	136	5	0.45	0.042	165	1.65	0.109
yajG	136	14	1.60	0.112	304	1.66	0.098
ybbL	136	5	1.38	0.083	145	2.23	0.140
ybdL	136	13	1.24	0.100	120	3.97	0.210
ybeD	136	7	0.79	0.045	110	0.47	0.027
ybhL	136	13	2.15	0.165	139	2.70	0.169
ybjC	136	8	2.75	0.233	159	2.14	0.153
ycaR	136	1	0.00	0.000	51	1.52	0.098
yccA	136	9	3.43	0.231	208	1.87	0.106
yceJ	136	23	4.21	0.249	261	2.25	0.136
ycfD	136	4	1.40	0.107	75	2.84	0.156
ychF	136	8	0.76	0.069	116	1.73	0.090
ychH	136	17	1.88	0.126	277	0.47	0.047
yeaR	136	22	4.92	0.267	172	1.43	0.100
yebK	136	18	1.05	0.062	337	1.90	0.118
yegR	132	27	3.87	0.222	419	1.96	0.094
yeiQ	136	18	2.00	0.176	222	2.90	0.160

Continued on the next page

Table S2.2 Genetic variability of IGRs present in the majority of isolates – continuation

Promoter	IGR present in N isolates	N of IGR versions	Inverted API of IGRs	PSS of IGRs	IGR length	Inverted API of ORFs	PSS of ORFs
yfgG	136	11	0.78	0.071	351	0.87	0.063
ygfB	136	2	1.12	0.090	167	1.97	0.097
yhdH	136	3	0.36	0.026	151	2.72	0.149
yhfA	135	8	0.64	0.063	301	1.90	0.128
yhjR	136	14	2.85	0.180	272	1.14	0.085
yhjX	136	19	5.55	0.241	228	4.67	0.193
yiaK	133	22	3.02	0.157	204	3.33	0.155
yibN	136	13	1.78	0.102	244	1.85	0.100
yifL	136	19	1.42	0.104	192	1.35	0.093
yihI	136	2	0.19	0.015	66	1.52	0.098
yjeV	136	22	3.25	0.183	235	1.14	0.130
yjjQ	136	36	4.25	0.249	631	1.62	0.128
yjjX	136	4	0.84	0.098	51	2.50	0.138
yjjZ	136	23	2.29	0.193	140	3.49	0.198
ymiA	132	22	3.21	0.238	294	1.57	0.109
yneJ	136	11	4.15	0.150	100	3.28	0.166
yobF	136	31	3.75	0.224	673	0.34	0.028
yqhD	135	10	2.59	0.096	136	2.03	0.125
yqjA	136	11	3.12	0.142	344	1.77	0.098
yrbG	136	9	0.47	0.043	210	1.90	0.139
zntA	135	2	0.02	0.014	73	4.12	0.183
znuC	135	7	0.57	0.077	78	2.44	0.131
zwf	135	17	1.05	0.062	337	1.91	0.112

Note: IGR = intergenic region; ORF = open reading frame; API = average pairwise identity; PSS = proportion of segregating sites

Chapter 3

Efficiency of the synthetic self-splicing RiboJ ribozyme is robust to *cis*- and *trans*-changes in genetic background

Markéta Vlková, Bhargava Reddy Morampalli, Olin K. Silander

Article pulished after peer-review

MicrobiologyOpen (Volume 10, Issue 4, 2021)

Author contributions:
Markéta Vlková: Conceptualization (equal); Data curation (lead); Formal analysis (lead); Investigation (equal); Methodology (lead); Project administration (supporting); Visualization (lead); Writing-original draft (lead); Writing-review & editing (equal).
Bhargava Reddy Morampalli: Investigation (equal); Writing-review & editing (supporting).
Olin K. Silander: Conceptualization (equal); Formal analysis (supporting); Funding acquisition (lead); Methodology (supporting); Project administration (lead); Resources (lead); Supervision (lead); Writing-original draft (supporting); Writing-review & editing (equal).

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Marketa Vlkova
Name/title of Primary Supervisor:	Dr. Olin Silander
In which chapter is the manuscript /published work:	3

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Vlková, M., Morampalli, B. R., and Silander, O. K. (2021). Efficiency of the synthetic self-splicing RiboJ ribozyme is robust to cis- and trans-changes in genetic background. MicrobiologyOpen, 10(4). DOI: 10.1002/mbo3.1232

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Marketa Vlkova Digitally signed by Marketa Vlkova Date: 2021.09.23 13:55:00 +12'00'
Date:	23-Sep-2021
Primary Supervisor's Signature:	Olin Silander Digitally signed by Olin Silander Date: 2021.09.23 14:20:01 +12'00'
Date:	23-Sep-2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

3.1 Preface

When focusing on the role of natural selection on gene expression we also asked whether there is a different selection pressure acting on transcription and translation. In order to address this question we required a new system to measure changes in transcription in isolation from translational changes. One could achieve this using RNA sequencing, however this approach misses some important advantages of flow cytometry - the high throughput and single cell resolution. Instead, we aimed to incorporate a system that has been utilised previously - a self-cleaving ribozyme RiboJ. Transcribing RiboJ into the mRNA molecules results in identical mRNA molecules from identical genes being produced even when the sequence of the 5' UTR is not identical. However, we did not find any information about how efficient this ribozyme is in its cleaving activity - for example whether only 50% of mRNA was immediately cleaved, or whether 90% or more was immediately cleaved.

We thus decided quantify self-cleavage in this system and this chapter covers the approach and results. For quantification we designed an RT-qPCR assay which is able to quantify what proportion of the mRNAs that have RiboJ sequences are cleaved in bacterial population at single timepoint during an exponential growth. In this way we could measure the efficiency of RiboJ's cleaving activity, and assess if the remaining uncleaved mRNAs, if any, could impact our measurements of transcription activity using flow cytometry. In addition we tested whether *cis*- and *trans*-genetic changes have any impact on the cleaving efficiency as well.

Olin Silander and I conceived the project and designed the experiments and analyses. Bhargava Morampalli and I performed the experiments covered in this chapter. I performed the majority of the RT-qPCR assays with the help of Bhargava Morampalli and I prepared strains for RNA isolation, which was carried out by Bhargava Morampalli. Bhargava Morampalli also sequenced several of the constructed plasmids with the help of Stella Pearless. Two of the plasmids used for this study were constructed previously by technician D. Blank at the University of Basel. I constructed the other plasmids used for this work while Stella Pearless assisted me with their transformation into the different bacterial strains. I carried out the bioinformatic analysis and data curation with inputs from Olin Silander. I wrote this chapter with contribution from Olin Silander and comments from Bhargava Morampalli and Tim Cooper. This chapter has been published after peer-review in *MicrobiologyOpen* (DOI: 10.1002/mbo3.1232) and it was slightly modified to match the formatting style of the other chapters in this thesis.

3.2 Abstract

The expanding knowledge of the variety of synthetic genetic elements has enabled the construction of new and more efficient genetic circuits and yielded novel insights into molecular mechanisms. However, context dependence, in which interactions between *cis*- or *trans*-genetic elements affect the behavior of these elements, can reduce their general applicability or predictability. Genetic insulators, which mitigate unintended context-dependent *cis*-interactions, have been used to address this issue. One of the most commonly used genetic insulators is a self-splicing ribozyme called RiboJ, which can be used to decouple upstream 5' UTR in mRNA from downstream sequences (e.g., open reading frames). Despite its general use as an insulator, there has been no systematic study quantifying the efficiency of RiboJ splicing or whether this autocatalytic activity is robust to *trans*- and *cis*-genetic context. Here, we determine the robustness of RiboJ splicing in the genetic context of six widely divergent *E. coli* strains. We also check for possible *cis*-effects by assessing two SNP versions close to the catalytic site of RiboJ. We show that mRNA molecules containing RiboJ are rapidly spliced even during rapid exponential growth and high levels of gene expression, with a mean efficiency of 98%. We also show that neither the *cis*- nor *trans*-genetic context has a significant impact on RiboJ activity, suggesting this element is robust to both *cis*- and *trans*-genetic changes.

3.3 Introduction

Synthetic nucleic acid functional elements used to control protein output - such as promoters, ribosome binding sites, or terminators - are an indispensable part of engineered genetic circuits (Levskaya et al. 2005, Na et al. 2013, Neves et al. 2020) and are frequently used to study basic biological processes (Barbier et al. 2020, Bittihn et al. 2020). The value of such synthetic functional elements increases as their properties are better described and quantified - in some cases, careful quantification of the behavior of synthetic functional elements has led to fundamentally new insights into molecular mechanisms controlling protein output (Schmiedel et al. 2019, Urtecho et al. 2019).

One important type of synthetic functional elements that have been used to ensure predictable and robust protein output from mRNA are self-splicing ribozymes. These ribozymes can be used to splice mRNA at specific locations, for example, to remove the 5' untranslated region (UTR). One of the most common ribozymes used to remove 5' UTRs is RiboJ. By removing the 5' UTR of the mRNA, RiboJ enables transcripts having different promoters and thus different 5' UTRs to produce identical mRNAs. This mitigates any effects that the 5' UTR might have on mRNA folding or ribosome binding, keeping the translation initiation rate consistent (and predictable) even when promoters have different sequences (Neves et al. 2020, Urtecho et al. 2019, Yu et al. 2018). The utility of the RiboJ element was first demonstrated when it was used to ensure predictable expression in a synthetic NOT gate circuit, irrespective of the sequence of the promoter used to control the expression of a CI repressor in the system (Lou et al. 2012).

However, since the first use of RiboJ as a means of ensuring predictable expression, additional research has suggested there can also be unexpected effects of its use. (Clifton

et al. 2018) demonstrated that RiboJ insertion into the mRNA sequence led to an increase in protein expression and that the relative increase in expression depended on the strength of the promoter used. This effect was attributed to hairpin formation at the 5' end of mRNA whose 5' UTR had been removed by RiboJ, leading to higher stability and increased translation (Carrier and Keasling 1997, Clifton et al. 2018, Neves et al. 2020). Another unexpected effect was observed when (Bartoli et al. 2020) designed a tunable system to control translation initiation via binding of small regulatory RNA (srRNA). The complex secondary structure of mRNA molecules with RiboJ at the 5' end appeared to interfere with the srRNA binding, decreasing the performance of the system. These results emphasize that unknown properties and behaviors of synthetic functional elements - here, RiboJ in particular - can lead to unexpected obstacles when creating new synthetic circuits.

We hypothesized that a complicating factor in the use of the RiboJ system would be the varying efficiency of RiboJ autocatalytic splicing activity between different bacterial strains, or due to polymorphisms near the RiboJ element. To our knowledge, there has been no systematic study quantifying the efficiency of the autocatalytic RiboJ splicing, or whether the efficiency of this autocatalytic activity depends on the genetic background of the organism in which it is used. To address these questions we first developed an assay to quantify RiboJ self-splicing efficiency. We then tested the robustness of the self-splicing activity to *cis*-genetic changes by assaying efficiency in two genetic contexts that differ by a single nucleotide polymorphism (SNP) close to the autocatalytic site of RiboJ. Finally, we tested the robustness of RiboJ behavior to *trans*-genetic changes by quantifying efficiency in six widely divergent strains of *E. coli*.

3.4 Materials and Methods

3.4.1 Bacterial strains

The genetic backgrounds of *E. coli* strains used in this study are listed in **Table 3.1**. The identity of all lab strains was confirmed using whole-genome sequencing. The whole genomes of strains SC312 and SC392 have been also sequenced (Breckell and Silander 2020). Four different plasmids (**Table 3.2**) were transformed into each of the strains, providing 24 clones that we used to evaluate the efficiency of RiboJ splicing. The presence of the plasmids with correct inserts in all clones was confirmed by Sanger sequencing (Macrogen, South Korea).

3.4.2 Plasmid construction

All plasmids used to measure the autocatalytic activity of RiboJ are listed in **Table 3.2**. Plasmids p69A.RJ- and p69C.RJ- were generously gifted by D. Blank, University of Basel. Plasmids p69A.RJ+ and p69C.RJ+ were constructed using plasmid pMV001 (which was created beforehand), as follows: RiboJ was ordered as four 60nt single-stranded oligos with each 30nt of them being homologous to either another 60nt RiboJ oligo or PCR amplified pUA66 vector (**Table S3.1**). These four oligos were then assembled with PCR amplified pUA66 vector using NEBuilder HiFi DNA assembly kit (New England Biolabs). The resulting pMV001 plasmid assembly mix was then used to electroporate Top10 *E. coli*

cells (Invitrogen). The presence of the RiboJ was then confirmed by Sanger sequencing (Macrogen, South Korea) from colonies grown on selective LB agar plates with 50 μg/ ml Kanamycin.

Table 3.1: **Bacterial strains used in this study**

Strain	Relevant characteristics	Phylogroup	Source or reference
SC392	A natural isolate of *E. coli*; Soil; 7/18/05; SC15-U2out14; St. Louis Clyde; Upshore (2m) outside the box	B1	(Ishii et al. 2006)
SC312	A natural isolate of *E. coli*; Water; 6/15/05; SC14-W8; St. Louis Clyde; Surface water	B1	(Ishii et al. 2006)
MG1655	F- λ- ilvG- rfb-50 rph-1	A	(Blattner et al. 1997)
DH5α	F- ϕ80lacZΔ M15 Δ(lacZYA -argF) U169 recA1 endA1 hsdR17 (rK- mK+) phoA supE44 λ- thi-1 gyrA96 relA1	A	Invitrogen
BW25113	F- DE(araD-araB)567 lacZ4787 (del)::rrnB-3 LAM- rph-1 DE(rhaD- rhaB)568 hsdR514	A	(Datsenko and Wanner 2000)
BL21 Star (DE3)	F-ompT hsdSB (rB-, mB-) galdcmrne131 (DE3)	A	Invitrogen

Table 3.2: **Plasmids used in this study**

Plasmid	Relevant characteristics	Source or reference
p69A.RJ-	*lacZ* promoter 69A, without RiboJ	D. Blank, University of Basel
p69C.RJ-	*lacZ* promoter 69C, without RiboJ	D. Blank, University of Basel
p69A.RJ+	*lacZ* promoter 69A, with RiboJ	this study
p69C.RJ+	*lacZ* promoter 69C, with RiboJ	this study

Note: All plasmids carry Kan^R selection marker and were created using pUA66 backbone (Zaslaver et al. 2006).

To create inserts for p69A.RJ+ and p69C.RJ+ plasmids the lacZ promoter regions from p69A.RJ- and p69C.RJ- were PCR amplified. The primers used contain 17nt overhangs that are homologous to PCR amplified pMV001 vector (**Table S3.1**). We ligated the vector with the inserts through Gibson assembly (Gibson et al. 2009) using the NEBuilder

HiFi DNA assembly kit (New England Biolabs). All primers and oligos used including sequencing primers (Integrated DNA Technologies) are listed in **Table S3.1**. In all cases, the same method for insert and vector PCR amplification from existing plasmids was used as described by (Li et al. 2011).

3.4.3 Flow cytometry

Strains for flow cytometry were grown in M9 minimal media (Sigma) supplemented with MgSO4, CaCl2, 0.4% (w/v) carbon source (glucose, galactose, or lactose), and 50 μg/ml Kanamycin. They were first inoculated from a glycerol stock library into a 96 well microplate using a pin replicator (Enzyscreen B.V.) and incubated at 37°C. After overnight incubation, the cultures were diluted into the same fresh media with the pin replicator and incubated the same way until they reached the mid-exponential phase ($\sim$4 h). At that point, the cells were diluted into 1x PBS with $\sim$1% formaldehyde and kept on ice until measuring the GFP levels using the flow cytometer.

Cytometry was performed with a BD FACSCanto II and BD FACSDiva software version 6.1.3. The GFP fluorescence was measured using the 488 nm laser and a 513/17 nm bandpass filter. The data from FACSDiva were exported into Flow Cytometry Standard files, and cell gating and fluorescence analysis was performed using custom R scripts (flowCore package v2.0.1; the scripts are available through https://doi.org/10.5281/zenodo.5154246. Cells were gated based on their maximal kernel density of forward and side scatter values, keeping about 1/3 of all events. The modal fluorescence was calculated from gated cells as the maximal kernel density from the fluorescence signal.

3.4.4 RNA isolation

RNA was isolated from four clones a day, while clones with the same genetic background were processed together on the same day. We isolated RNA from MG1655 clones twice on two different days, all other clones were isolated just once. Each strain containing one of the four plasmids (**Table 3.1** and **Table 3.2**) was grown from a single colony overnight in 3 ml of LB with 50 μg/ml Kanamycin and 2 mM IPTG (Isopropyl β-D-1-thiogalactopyranoside) with shaking (250 rpm) at 37°C. Because the high IPTG concentration impaired the growth of SC312 strain with RiboJ plasmids (i.e., p69A.RJ+ and p69C.RJ+), we grew all SC312 clones for RNA isolation in LB with 0.2 mM of IPTG instead. The next day 15 ml of the same fresh media in 50 ml Falcon tubes was inoculated by 15 μl of this overnight culture. This was incubated under the same conditions. Once the cultures reached an exponential phase (between 1.75 h and 2.5 h) it was placed on an ice slurry.

Next, we added 7.5 ml of ice-cold 5% phenol in ethanol to each 15 ml of culture and kept them on ice for 15 min. The cultures were then spun at 7000G for 7 min at 4°C, the supernatant was discarded and the pellet was redispersed in 350 μl of 3 mg/ml Lysozyme solution (in TE buffer). After incubating for 3 min, an equal volume of RNA lysis buffer was added and RNA isolated using Monarch Total RNA Miniprep Kit (New England Biolabs). Each sample was treated by DNase I twice: (1) on-column during the RNA extraction and then (2) in-tube after RNA extraction. This was done to avoid any amplification from residual gDNA during RT-qPCR. After the second treatment with DNase I the samples

were column-purified and concentrated using RNA Clean & Concentrator-5 kit (Zymo Research). The quality of RNA in each sample was checked on 1% agarose gel and its concentration was measured on a Qubit 4 fluorometer (Invitrogen). The isolated RNA samples were then stored in a -80°C freezer.

3.4.5 RT-qPCR

To assess the efficiency of PCR amplification by our primers we used RNA from MG1655 strain (all four plasmids). A ten-fold serial dilution was performed on all the RNA samples up to 10^{-4}. RT-qPCR was run on all the dilutions in triplicates using two different master mixes differing by the forward primer used - F1 and F2 (**Figure 3.2** and **Table S3.1**). The total reaction volume was 20 μl with 2 μl of template RNA. We used SensiFAST Probe No-ROX One-Step Kit (Meridian Bioscience) and PikoReal Real-Time PCR System (Thermo Scientific) with following cycling conditions: Reverse transcription for 10 min at 45°C; Polymerase activation for 2 min at 95°C; 40 cycles of denaturation for 5 s at 95°C and Annealing & extension for 20 s at 55°C. The C_t values were obtained via PikoReal software version 2.2, exported into .xlsx file, and converted into .csv to be further analyzed using custom R scripts (available through https://doi.org/10.5281/zenodo.5154246).

To assess the autocatalytic efficiency of RiboJ, the RNA from all samples was first diluted from its original concentration (~2-3 μg/μl) to 20 pg/μl to obtain C_t values between 20 and 40 and to dilute out any potential residual of gDNA (to less than one molecule per reaction). We confirmed that no amplification occurred when omitting reverse transcriptase from the master mix. Each RNA sample was then run in three or more replicates using both primer sets (**Figure 3.2**) with the same conditions described above. We exported the data from the PikoReal software version 2.2 into .xlsx files, converted these into .csv, and performed all analyses using custom R scripts (available through https://doi.org/10.5281/zenodo.5154246). In brief, we determined the mean C_t value of all the replicates for the uncut and cut RiboJ transcripts and calculated the efficiency as the ratio of the cut and uncut transcripts using the Pfaffl method (Pfaffl 2001):

$$Eff = 100 - 100\frac{E^{(a-b)}}{E^{(c-d)}} \tag{3.1}$$

where E is constant mean amplification efficiency (1.95766), a and b are mean C_t values of transcripts without and with RiboJ, respectively, using F1 primer, and c and d are mean C_t values of the same transcripts using F2 primer. To obtain a measure of the error in these estimates, we bootstrapped the data 10,000 times and recalculated the ratio for each bootstrap replicate.

3.4.6 Plasmid sequencing

Plasmid DNA was isolated from overnight cultures using the StrataPrep Plasmid Miniprep Kit (Agilent), per the manufacturer's instructions. These were then prepared for Oxford Nanopore sequencing using the Oxford Nanopore rapid barcoding library prep, per the manufacturer's instructions, with a separate barcode used for each plasmid. These were

run on a single MinION flowcell for 1 h and 50 min. The reads were base-called using the guppy_basecaller v5.0.7 high accuracy model and demultiplexed using guppy_barcoder, resulting in between 28.8 Mbp and 35.1 Mbp for each plasmid. These reads were used as input for medaka, using medaka_consensus to correct the original plasmid sequence.

3.5 Results

Our first motivation for quantifying the behavior of RiboJ arose during experiments aimed at understanding the effects of promoter polymorphisms segregating in the environmental *E. coli* population on transcription and translation. Here, "promoter" is defined as the entire intergenic region upstream of an open reading frame, as well as part of the upstream and downstream open reading frames (Zaslaver et al. 2006). We include parts of the upstream and downstream open reading frames as it is well established that many open reading frames contain transcriptional regulatory elements affecting their own regulation or that of downstream genes. We assay the effects of the promoter on transcription by quantifying the fluorescence that occurs due to a GFP open reading frame that lies downstream of this "promoter".

In the case of the *lacZ*, here we define the "promoter" as the *lacI-lacZ* intergenic region, plus 88 and 71 bp of each flanking upstream (*lacI*) and downstream (*lacZ*) coding regions, respectively. We discovered a single SNP at position 69 relative to the *lacZ* gene start codon (C to A) that resulted in a change in downstream protein levels. We found that the effect of this SNP on GFP protein expression was consistent in different genetic backgrounds as well as during growth in different carbon sources (**Figure 3.1a** and **Figure S3.1a**). To check whether this C69A SNP affected transcription or translation (or both), we incorporated RiboJ as an insulator downstream of it (**Figure 3.2**, Top panels). This ensured that the mRNA being translated was identical regardless of which SNP was present. Thus, if the change in protein expression remained in the presence of RiboJ, we could infer that the change was due solely to the SNP affecting transcription. If the difference disappeared in the presence of RiboJ, we could infer that the change was due to the SNP affecting translation. However, it was also possible that the SNP itself interfered with RiboJ cutting (a *cis*-effect). If so, we could not unambiguously infer that the cause of the changes in fluorescence we observed was due to translation or transcription (or both). In addition, the genetic background of the strain itself might have affected RiboJ cutting (a *trans*-effect). To exclude the possibility of *cis*- or *trans*-effects on RiboJ cutting efficiency, we quantified efficiency in the presence of *cis*- and *trans*-genetic changes.

We designed an RT-qPCR assay to quantify the autocatalytic cutting activity of RiboJ. This assay is based on the principle that for a pair of forward and reverse primers that span the RiboJ cut site, an amplification product should only be produced for uncut mRNA molecules. In contrast, for primer pairs that do not span the cut site, an amplification product should be produced for all molecules. By examining the relative numbers of cut and uncut molecules, we can infer the efficiency of RiboJ cutting relative to the rate of production of all transcripts controlled by the same promoter (i.e., the rate of transcription). To this end, we designed two qPCR primer sets. The first set produced an amplicon from a region spanning the RiboJ cut site, while the second produced an amplicon from a region downstream of the RiboJ cut site (**Figure 3.2**). Both sets shared the same reverse primer,

differing solely by the location of the forward primer. Because one forward primer binds upstream of the RiboJ cut site, no amplification can occur if the 5' UTR sequence has been cut off (**Figure 3.2**, Middle panel). The second forward primer binds downstream of the cut site and results in an amplification product from all transcripts. To quantify differences in amplification that might result from primer binding or other unforeseen mechanisms, we calculated the relative fold change in the abundance of these two amplicons when RiboJ is absent. In the absence of RiboJ, any difference in amplification between the two primer sets should be due solely to differences in primer efficiency or related effects, as without RiboJ, both amplification products will always be produced (**Figure 3.2**, bottom panel).

We first assessed whether *trans*-genetic changes affected the self-splicing activity of RiboJ, by assaying RiboJ activity in six widely divergent strains of *E. coli* (**Table 3.1**). To test for *cis*-effects, we assayed activity in two promoter contexts, each varying by a single SNP that was 8 bp upstream of the RiboJ cut site (2 bp upstream of RiboJ sequence). We

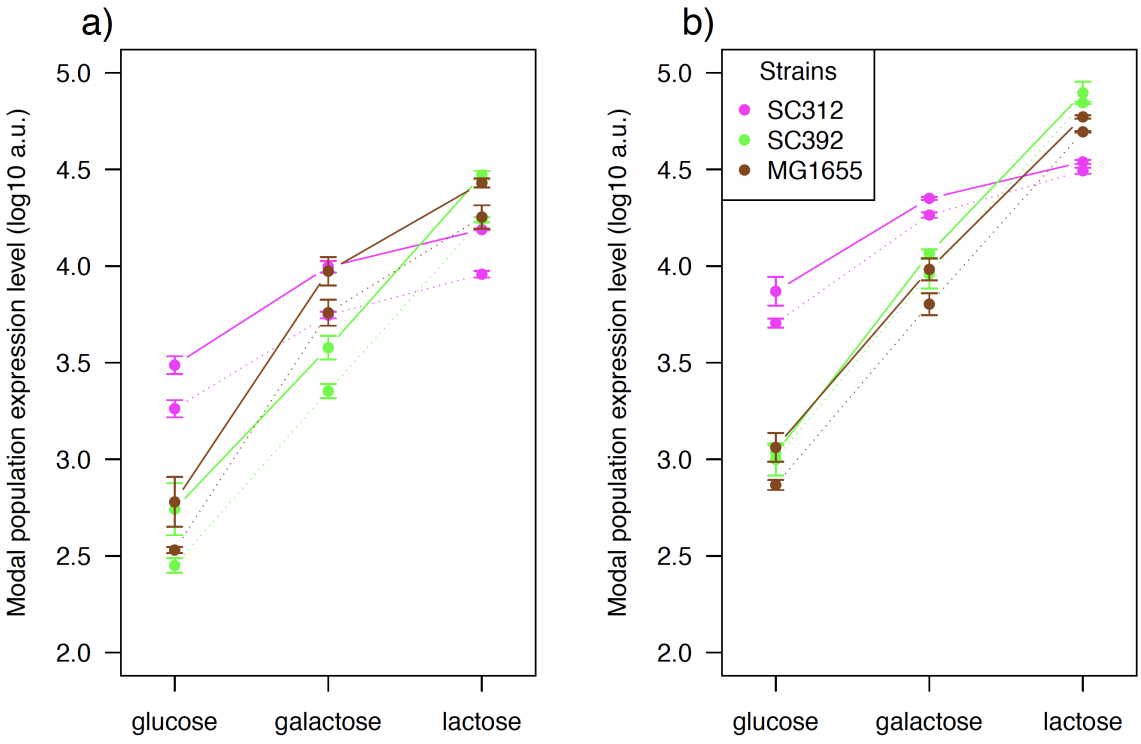

Figure 3.1: **Modal expression levels differ consistently due to a single A to C change at position +69 of the *lacZ* open reading frame both without RiboJ (a) and with RiboJ (b) and across genetic backgrounds.** Shown are the modal population fluorescence levels for GFP driven by an upstream *lacZ* promoter for three divergent strains of *E. coli*. In all genetic backgrounds tested and all growth conditions (glucose, galactose, and lactose), the fluorescence levels from a promoter with the 69C polymorphism (dotted line) were consistently lower than those from the promoter with the 69A polymorphism (solid line). On the log10 scale that is shown, a 0.1 difference is equivalent to a 25% change in expression. There are clear effects of genetic background on both the level and dynamic range of protein expression. In particular, SC312 has a narrow dynamic range, with relatively high expression in non-lactose environments compared to the other strains, but relatively low expression in lactose. Despite this, the effect of the A to C change is nearly constant. Whiskers show one standard deviation of three replicates.

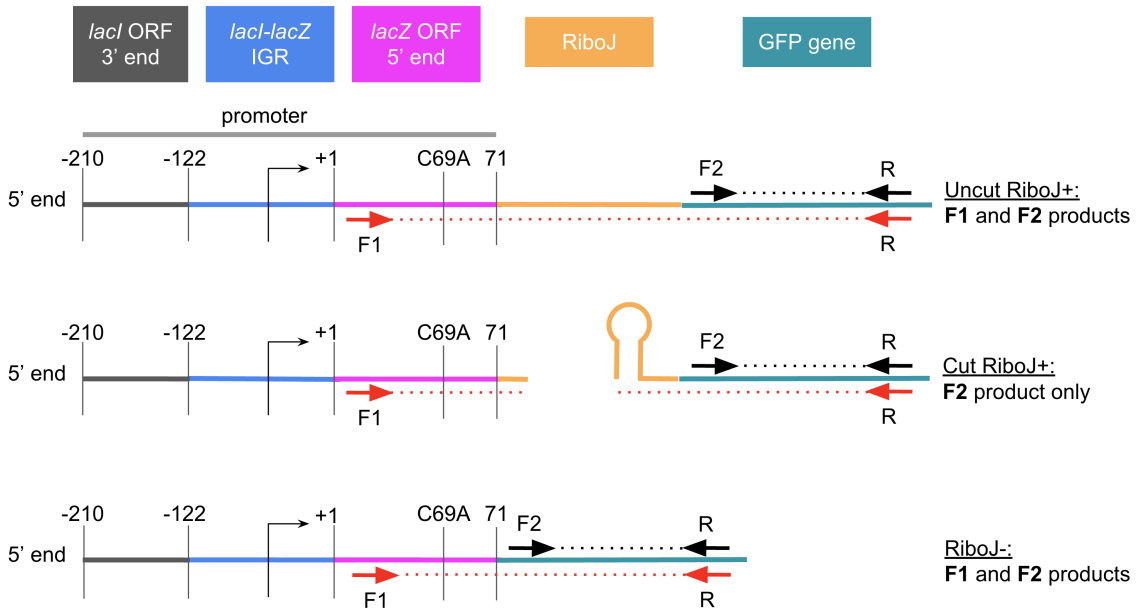

Figure 3.2: **Scheme of the RT-qPCR primer design to quantify the efficiency of RiboJ cutting.** Each primer is represented by an arrow, with pairs colored the same. The dotted lines indicate amplicons. If the dotted line between a primer pair is interrupted, the amplicon is not produced. When RiboJ cleaves off the 5' UTR (Middle panel), the amplicon from primer F1 is not produced, while the amplicon from the F2 primer is still produced. The *lacI-lacZ* intergenic region (IGR) with the first 71 bp and last 88 bp of the *lacZ* and *lacI* open reading frames, respectively, were placed upstream of GFP (and RiboJ) as a promoter. The arrow in the middle of *lacI-lacZ* IGR indicates the transcription start site. The translation is driven from a strong synthetic ribosome binding site downstream of the *lacZ* gene sequence (here as a part of the GFP gene).

thus transformed each of the six strains with each of four plasmids differing by the C69A SNP in the *lacZ* promoter and either with RiboJ or without RiboJ (**Figure 3.2**, **Table 3.2**). We isolated RNA from exponentially growing cultures for all strains and confirmed that the amplification efficiency of all primer combinations with all templates was within the range of 90–110% (**Figure S3.2**). We used the resulting mean efficiency value across all strains (95.8%) for all subsequent calculations of RiboJ autocatalytic activity. We assayed the efficiency of RiboJ autocatalytic activity using at least triplicates for each strain and promoter combination (**Materials and Methods**). RiboJ cutting efficiency was high in all cases. Overall we found that 98% of all mRNA molecules containing RiboJ were cleaved. This was extremely robust for almost all strain and promoter combinations, with the lowest median value being 97% (**Figure 3.3**). We also found that RiboJ activity was robust to *cis*-changes, with no consistent differences between the 69A and 69C versions of the promoter.

However, we observed one exception to this robust behavior. In strain SC392, the 69A version of the promoter construct exhibited a 10-fold higher cutting efficiency (**Figure 3.3**, inset). To obtain an estimate of error for our measurements and test whether sampling alone could account for this or other differences, we bootstrapped the data 10,000 times and recalculated the efficiencies (**Figure 3.3**, **Materials and Methods**). We found that sampling alone was unlikely to account for the higher efficiency of 69A that we observed.

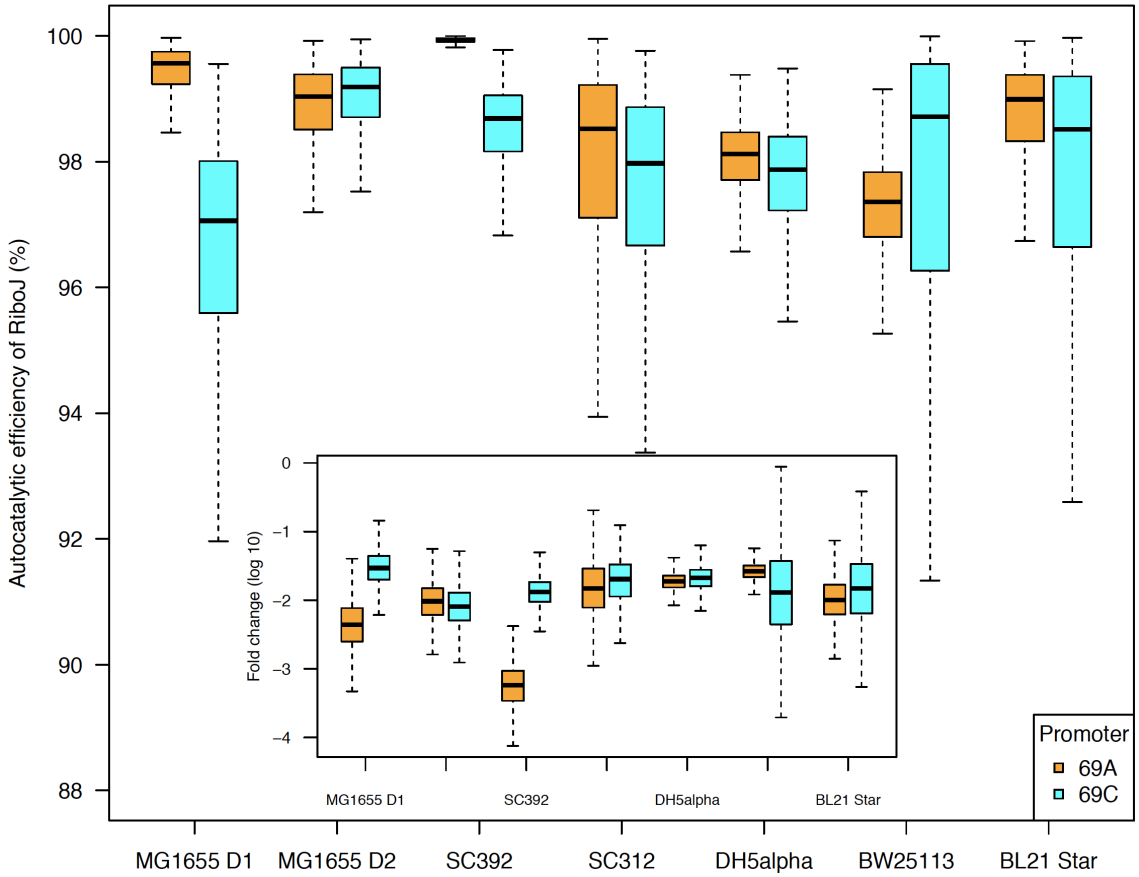

Figure 3.3: **Autocatalytic efficiency of RiboJ.** The boxplots in the figure show the minimum and maximum value (whiskers), the first and third quartile (boxes), and the median. These values were obtained through bootstrapping the RT-qPCR data (see **Materials and Methods**). D1 and D2 in MG1655 strain labels indicate that this data is from different biological replicates for which the RNA was extracted on different days (D1 and D2 denoting day 1 and day 2, respectively). The inset shows fold changes in the abundance of uncut transcripts with RiboJ relative to all transcripts. Note that the smaller range in cutting efficiency of RiboJ in SC392 strain for promoter 69A is simply a consequence of converting the C_t fold change of the two different amplicons into catalytic efficiency in percentages. The inset shows that the range and error in fold changes for SC392 with promoter 69A is comparable to the other samples.

We thus sequenced all plasmids from the SC392 strain to check for possible SNPs in the vector backbone that might lead to what we see as an increased RiboJ autocatalytic activity in this strain. We discovered single SNPs in each of the two plasmids with RiboJ. They were not identical, but each was located close to the origin of replication and thus could have affected the copy number of these plasmids. It is thus possible that the shift in RiboJ activity we observed for 69A in SC392 is due to a SNP in plasmid p69A.RJ+.

3.6 Discussion

Given the efficient activity of RiboJ (resulting in 98% of all mRNA molecules being cut), the differences in splicing measured between the two *lacZ* promoters and among the strains cannot explain the changes in expression we observed (**Figure 3.1** and **Figure S3.1**). Rather than being due to the C69A SNP affecting the activity of RiboJ, the changes in expression are a consequence of this SNP affecting the regulation of both transcription and translation. The C69 SNP lowers both transcription and translation by approximately 25% (**Figure 3.1** and **Figure S3.1**).

Considering the depth at which the activity of the *llacZ* promoter has been studied, we do not expect this SNP to be part of some unknown transcription factor binding site. The SNP may create a binding site causing a polymerase or transcription factor to pause during its linear search on the DNA strand for functional binding sites, thus inhibiting transcription. Processes such as transcriptional pausing and attenuation have been previously described in bacteria (Bailey et al. 2021, Blainey et al. 2006, Mustaev et al. 2017, Naville and Gautheret 2009). These provide a more plausible explanation for the effect we see.

At the level of translational regulation, there is a possibility that the C69A SNP causes a difference in the secondary structure of the mRNA. This could then lead to differential accessibility of the mRNA to ribosomes. It is also possible that it inhibits proper translation by causing spurious ribosomal binding (Whitaker et al. 2015). However, it is beyond the scope of this paper to uncover the very mechanism of the regulation the SNP is involved in.

3.7 Conclusion

In this study, we have confirmed that the autocatalytic activity of the ribozyme RiboJ is robust in *cis*- and *trans*-genetic contexts. This robust behavior of RiboJ suggests that the differences in expression that we observed in **Figure 3.1** and **Figure S3.1** are a result of changes in both transcription and translation due to the single C69A SNP, and not to changes in RiboJ autocatalytic efficiency. We note that there have been no previous reports that this region is involved in *lacZ* gene regulation. We proposed possible ways by which this SNP can be affecting both transcription and translation, however, more in-depth research is needed to determine the actual mechanism.

3.8 Supplementary Information

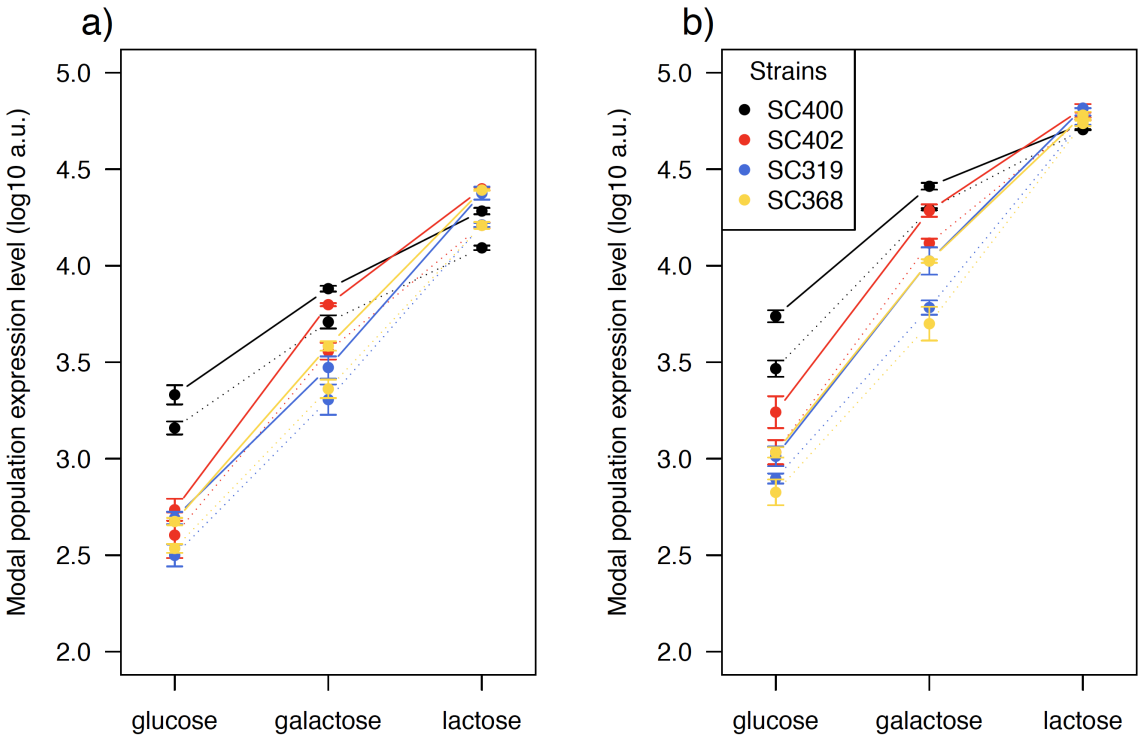

Figure S3.1: **Modal expression levels differ consistently due to a single A to C change at position +69 of the *lacZ* open reading frame both without RiboJ (a) and with RiboJ (b) in additional strains.** Shown are the modal population fluorescence levels for GFP driven by an upstream *lacZ* promoter for four additional *E. coli* strains in **Figure 3.1**. In all genetic backgrounds tested and all growth conditions (glucose, galactose, and lactose), the fluorescence levels from a promoter with the 69C polymorphism (dotted line) were consistently lower than those from the promoter with the 69A polymorphism (solid line). On the log10 scale that is shown, a 0.1 difference is equivalent to a 25% change in expression.

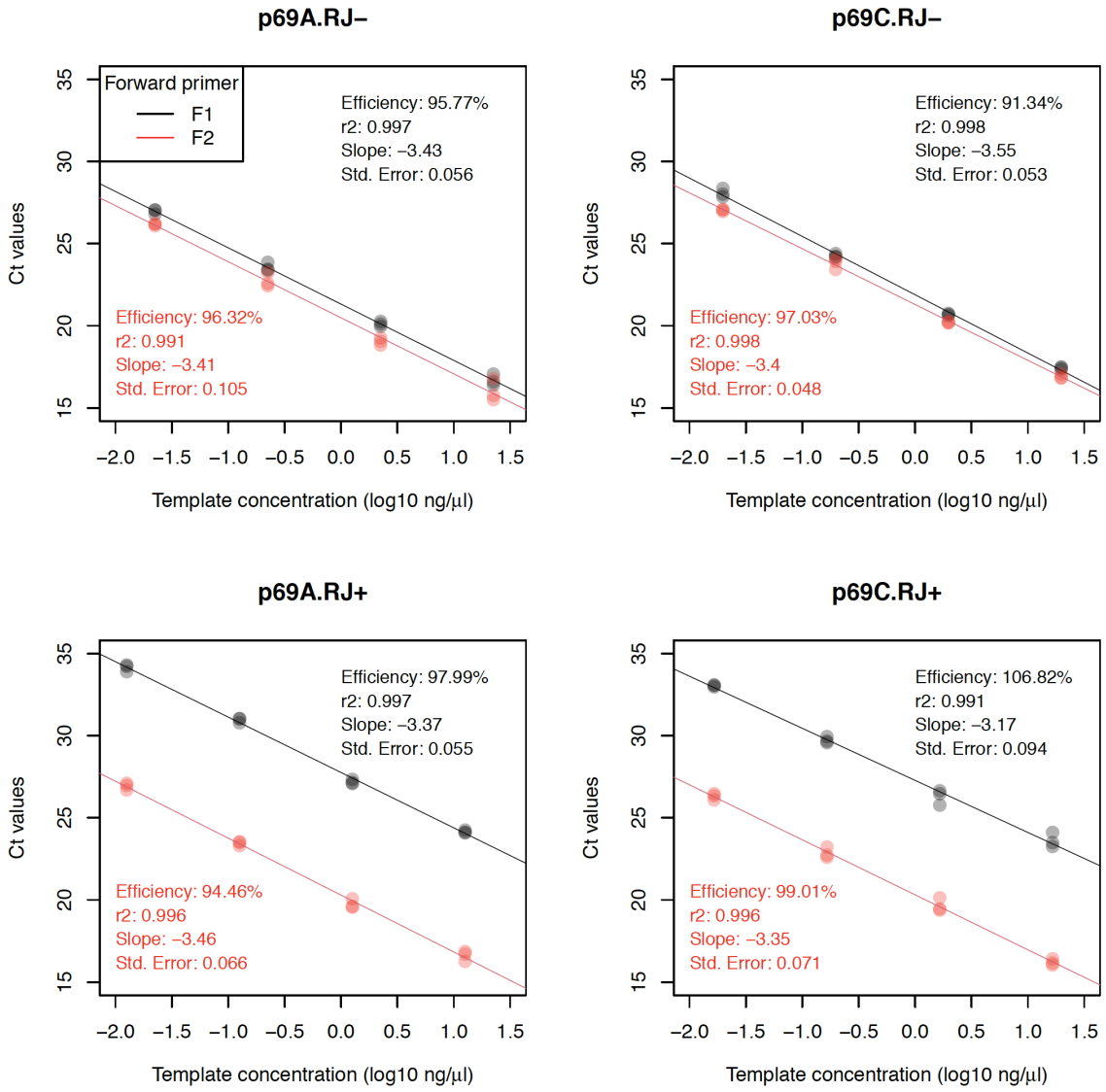

Figure S3.2: **Amplification efficiency of all primer-template combinations.** Each point indicates the C_t value for one technical replicate at different template concentrations, with each panel indicating one template and each color (red or black) indicating one primer combination. The lines show linear regressions, calculated using all data points for a given primer-template combination. For the templates without RiboJ (top panels), both primer pairs result in nearly equal C_t values. For the templates with RiboJ (bottom panels), the F1 primer pair has consistently larger C_t values, as expected. To calculate the catalytic activity of RiboJ, we used the mean amplification efficiency across all primer-template combinations, 95.8%.

Table S3.1: **Primers and oligos used in this work**

Primer and oligo ID	Sequence	Purpose
pUA66_insert_F3965	5' - TTG TCT GTT GTG CCC AGT CAT AGC - 3'	PCR & Sanger sequencing
pUA66_insert_R232	5' - TCG CAA AGC ATT GAA GAC CAT ACG C - 3'	PCR & Sanger sequencing
RiboJ_oligo1_Rev	5' - GAA AGC ACA TCC GGT GAC AGC TGG ATC CCC TCG AGG TGA AGA CGA AAG GGC CTC GTG ATA - 3'	DNA assembly of pMV001
RiboJ_oligo2_For	5' - GGG GAT CCA GCT GTC ACC GGA TGT GCT TTC CGG TCT GAT GAG TCC GTG AGG ACG AAA CAG - 3'	DNA assembly of pMV001
RiboJ_oligo3_Rev	5' - TCT TAG TTT AAA CAA AAT TAT TTG TAG AGG CTG TTT CGT CCT CAC GGA CTC ATC AGA CCG - 3'	DNA assembly of pMV001
RiboJ_oligo4_For	5' - CCT CTA CAA ATA ATT TTG TTT AAA CTA AGA AGG AGA TAT ACA TAT GAG TAA AGG AGA - 3'	DNA assembly of pMV001
pUA66_vector_F	5' - GAA GGA GAT ATA CAT ATG AGT AAA GG - 3'	PCR of pUA66 for DNA assembly of pMV001 & RT-qPCR assay (**primer F2**)
pUA66_vector_R	5' - TCG AGG TGA AGA CGA AAG G - 3'	PCR of pUA66 for DNA assembly of pMV001
pMV001_FastClonV_F	5' - AGC TGT CAC CGG ATG TGC - 3'	PCR of pMV001 for DNA assembly of p69A.RJ+ and p69C.RJ+
pMV001_FastClonV_R	5' - TCG AGG TGA AGA CGA AAG GGC - 3'	PCR of pMV001 for DNA assembly of p69A.RJ+ and p69C.RJ+
lacZ_FastClonIN_F	5' - TTT CGT CTT CAC CTC GAC AAT ACG CAA ACC GCC TCT CC - 3'	PCR of p69A.RJ- and p69C.RJ- for DNA assembly of p69A.RJ+ and p69C.RJ+
lacZ_FastClonIN-D9_R	5' - CAC ATC CGG TGA CAG CTG TGT AAC GCC AGG GTT TTC C - 3'	PCR of p69A.RJ- for DNA assembly of p69A.RJ+
lacZ_FastClonIN-K12_R	5' - CAC ATC CGG TGA CAG CTG GGT AAC GCC AGG GTT TTC C - 3'	PCR of p69C.RJ- for DNA assembly of p69C.RJ+
lacZ_qPCR_F	5' - ATG ATT ACG GAT TCA CTG GC - 3'	RT-qPCR assay (**primer F1**)

Continued on the next page

Table S3.1 Primers and oligos used in this work – continuation

Primer and oligo ID	**Sequence**	**Purpose**
GFP_qPCR_R	5' - GAA AAT TTG TGC CCA TTA ACA TCA CC - 3'	RT-qPCR assay (**primer R**)
GFP_Probe	5' - /56-FAM/ TTC AAC AAG/ZEN/AAT TGG GAC AAC TCC AGT GAA AAG TT/3IABkFQ/ - 3'	RT-qPCR assay (**probe**)

Note: Underscored sequences = regions homologous to PCR amplified pMV001 vector. Primers and the probe used for RT-qPCR assay are highlighted in bold with their simplified names that were used in the main text and **Figure 3.2** (Purpose column).

Chapter 4

Transcriptional control of the *lacZ* promoter is under directional and diversifying selection in nature

Markéta Vlková, Olin K. Silander

Article in preparation

(Molecular Biology and Evolution)

Author contributions:
Markéta Vlková: Conceptualization (equal); Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Project administration (supporting); Visualization (lead); Writing-original draft (equal); Writing-review & editing (equal).
Olin K. Silander: Conceptualization (equal); Formal analysis (supporting); Funding acquisition (lead); Methodology (supporting); Project administration (lead); Resources (lead); Supervision (lead); Visualization (supporting); Writing-original draft (equal); Writing-review & editing (equal).

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Marketa Vlkova
Name/title of Primary Supervisor:	Dr. Olin Silander
In which chapter is the manuscript /published work:	4

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

○ The manuscript is currently under review for publication – please indicate:

- The name of the journal:
- The percentage of the manuscript/published work that was contributed by the candidate:
- Describe the contribution that the candidate has made to the manuscript/published work:

◉ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Marketa Vlkova Digitally signed by Marketa Vlkova Date: 2021.09.28 12:55:09 +13'00'
Date:	28-Sep-2021
Primary Supervisor's Signature:	Olin Silander Digitally signed by Olin Silander Date: 2021.09.28 12:59:07 +13'00'
Date:	28-Sep-2021

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.

4.1 Preface

Given the transcription is the first step crucial for gene expression to initiate, the natural selection acting on it might be more stringent or of a different type from selection acting on translation. One can also easily imagine that limiting resources and energy such that transcription is kept at minimal rate while increasing translation to achieve high protein expression could be beneficial. However, there may also be a trade-off in transcriptional versus translational regulation as some regulatory networks might not be sustained by the low transcription-high translation arrangement. There are also some indications from a recent experimental evolution study that changes in expression are far more frequent than changes in translation. However, little is known how selection shapes transcriptional activity alone in nature and whether it differs from selection acting on translation.

We addressed this gap by using a similar methodology as described in the chapter studying selection acting on ten individual promoters (Chapter 2.), using flow cytometry and fluorescent reporter assay to discern selection when comparing segregating (under selection) and random (selection neutral) promoter variants. However, to measure changes in transcription only, we implemented a system that allows us to decouple transcription from translation, which has been described and verified in the previous chapter (Chapter 3.). In this chapter we focus on a single promoter, *lacZ*, but we quantify regulatory phenotypes, which are equivalents to the expression phenotypes used earlier, i.e., transcriptional activity, plasticity, and noise again. In addition, we look at the association with specific changes in promoter genotype and phenotype, and show that there is evidence of diversifying selection on noise phenotypes. In this case we also use three environments (glucose, galactose, and lactose) to examine the highest possible dynamic range of *lacZ* transcriptional activity.

Olin Silander and I conceived the project and designed the experiments and analyses. I performed all the experiments covered in this chapter and I carried out the bioinformatic analysis and data curation with inputs from Olin Silander. This manuscript was written in collaboration with Olin Silander, with comments from Tim Cooper and Andrea Sajuthi. This chapter is unpublished, but it is in the process of being submitted for peer-review to journal *Molecular Biology and Evolution* in coming days.

4.2 Abstract

Bacterial cells often respond to changes in the environment by modifying protein expression. This can be achieved through changes in transcriptional or translational activity, or both. Recent research has shed some light on how natural selection shapes overall protein expression. Still, little is known about how selection acts on transcription or translation individually, and whether one is under more stringent selection than the other. To address part of this question, we implemented an experimental system which allows us to measure how genetic changes affect transcription only, excluding the effects on translation. We used this system to quantify changes in three regulatory phenotypes of the *lacZ* promoter: transcriptional activity, plasticity, and cell-to-cell variability. We compared these phenotypes from segregating variants that have been subject to natural selection, and random variants that have never been subjected to natural selection. We detected both directional and diversifying selection acting on different phenotypes of *lacZ* regulation. Our results thus provide new insight into how one of the most well-characterized bacterial promoters is shaped in nature by selection.

4.3 Introduction

Bacteria respond to dynamically changing environmental conditions by altering their physiology and metabolism. Natural selection acts on aspects of these responses, such as the level, speed, sensitivity, and variability (Hawkins et al. 2020, Hodgins-Davis et al. 2019, Keren et al. 2016, Schmiedel et al. 2019, Silander et al. 2012, Vlková and Silander 2021). Frequently these aspects are controlled by mediating protein levels, which in turn are mediated by changes in transcriptional and translational activity.

However, there is little data on whether over evolutionary time, genetic variation in transcriptional or translational mechanisms are most often responsible for mediating changes in protein levels and the downstream cellular responses. From an energetic perspective, it seems straightforward that in order to produce a specific level of protein, it is optimal to produce as little mRNA as possible while still producing sufficient protein - minimising transcription level and maximising translation (Frumkin et al. 2017). Nevertheless, there are trade-offs that arise. Low transcription rates can slow down the response time to a signal, and increase variability between cells in their response (Maeda and Sano 2006, Sayut et al. 2007, Taniguchi et al. 2010). If cells generally experience selection to minimise response times, it is not clear whether there will still be directional selection to minimise transcription while maximising translation, or whether selection for rapid response will result in stabilising selection on transcription. In addition, some regulatory changes and networks are likely to be easier to evolve than others (Schaerli et al. 2018). In some cases, it may not be possible to mediate proper responses simply by maximising translation, and in extreme cases selection might act to maximise both transcription and translation, resulting in directional selection on transcriptional activity. Thus, it is unclear how regulatory changes are generally mediated over evolutionary time (mostly via transcription, translation, or a combination of both), and what evolutionary forces are acting on regulatory phenotypes (e.g. directional, stabilising, or diversifying).

Yet, it is readily apparent that changes in regulation can evolve rapidly (Gresham et al. 2008, Khademi et al. 2019, Tenaillon et al. 2016, Yona et al. 2018) via very few mutations (Duveau et al. 2017, Metzger et al. 2015, Vlková and Silander 2021). There are a number of mechanisms by which this can occur. The frequency of transcription can be changed by increasing or decreasing the strength of transcription factor binding sites, either activators or repressors. Alternatively, the strength of the sigma factor binding site can be adjusted (Browning and Busby 2016, Iyer and Struhl 1996, Kennell and Riezman 1977). Changes to translational regulation can also affect downstream protein levels, and these can occur through adjustments to the binding energy of the ribosomal binding site, including specific binding of antisense RNAs, the stability and folding energy of the 5' end of the mRNA, or by changing codon content, which may affect ribosomal speed or ribosomal pausing (Bailey et al. 2021, Desnoyers et al. 2013, Kudla et al. 2009, Mustaev et al. 2017, Naville and Gautheret 2009, Whitaker et al. 2015).

Here we focus on the evolution of regulatory responses for a canonical example of physiological adaptation, the *E. coli lac* operon. This operon is involved in the metabolism of lactose, and is regulated by the presence of lactose in the environment and cAMP in the cell (Hudson and Fried 1990). If glucose levels are high, the levels of free cAMP drop, subsiding activation of the *lac* operon; if glucose levels are low and lactose levels are high, then the operon is activated (Jobe and Bourgeois 1972, Kuo et al. 2003, Wanner et al. 1978, Wheatley et al. 2013). The expression is largely dependent on the relative ratios, and absolute levels, of intracellular cAMP and lactose (Kuhlman et al. 2007, Ozbudak et al. 2004).

However, most of what is known about the *lacZ* promoter activity has been obtained using classical laboratory strains only. Previously, we showed that there is a substantial genetic and phenotypic variability in *lacZ* promoter among environmental isolates of *E. coli* (Vlková and Silander 2021). Here, we implement an experimental system that allows measurement of transcriptional effects alone. This allows us to test whether many of the phenotypic differences between isolates are dependent solely on regulatory changes in translation, or whether they involve changes to transcriptional regulation.

In addition, compared to previous work (Vlková and Silander 2021), the experimental system we use has substantially increased sensitivity for detecting changes in transcriptional regulation. This increased sensitivity allows us to detect a wider range of evolutionary forces acting on transcriptional regulation, including both directional and diversifying selection. Finally, we also test how genetic background affects the behaviour of the operon by quantifying transcriptional regulation in natural isolates of *E. coli*. This work yields new insight into how the activity of one of the most well-characterized bacterial promoters is shaped in nature.

4.4 Materials and Methods

4.4.1 Construction of *lacZ* variant libraries

We created four types of *lacZ* promoter variant libraries: (1) segregating *lacZ* promoter variants placed into strain MG1655 on a plasmid; (2) random *lacZ* promoter variants placed into strain MG1655 on a plasmid; (3) segregating *lacZ* promoter variants placed into

their native *E. coli* isolate on a plasmid; (4) segregating *lacZ* promoter variants placed into chromosome of strain MG1655 replacing the MG1655 promoter variant. The promoter libraries were aliquoted into 96 well microplates. Each microplate also contained a positive control consisting of the highly active murein lipoprotein (*lpp*) promoter driving GFP expression (Zaslaver et al. 2006), and a negative expression control consisting of a promoter-less plasmid pMV001 (Vlková et al. 2021). Chromosome-based library had an additional negative expression control consisting of wild-type MG1655 strain. We describe the construction of each library type separately below.

Segregating *lacZ* variants in MG1655 genetic background

Low-copy number plasmid pMV001 with a SC101 ori, a strong RBS, self-cleaving ribozyme RiboJ and GFPmut2 gene was used for the vector backbone (Vlková et al. 2021). Both the vector backbone and segregating *lacZ* promoter variants were PCR amplified using Phusion High-Fidelity DNA polymerase with HF buffer (New England Biolabs). For promoter PCR amplification, 5 μl of pooled DNA from isolates with various variants of *lacZ* promoter was used as a DNA template (Vlková and Silander 2021). The primers for promoter amplification contained 17 nucleotide overhangs which were homologous to the ends of the vector backbone for subsequent DNA assembly. All primers used in this study are listed in **Table S4.3**.

For vector PCR amplification, 0.5 ng of pMV001 plasmid DNA served as a template. After confirming a successful PCR amplification of the products on 1% agarose gel, the template DNA was digested by DpnI from the remaining reaction volume (Li et al. 2011). Both the insert and vector PCR reactions were then column-purified and we assembled the vector and promoter variants using Gibson assembly (Gibson et al. 2009) with NEBuilder® HiFi DNA Assembly Master Mix (New England Biolabs). The assembly mix was then electroporated into the electrocompetent MG1655 strain. Transformed colonies which grew on LB agar plates with 50 μg/ml Kanamycin were picked for Sanger sequencing across the insert in the vector backbone and stored as glycerol stocks. Clones with confirmed segregating promoter variants were then grown in liquid LB with Kanamycin and used to create 96 well microplate glycerol stock libraries.

Random *lacZ* variants in MG1655 genetic background

We PCR amplified the pMV001 backbone, DpnI treated, and column-purified it the same way as described for segregating promoter variants above. We produced the promoter inserts by performing error-prone PCR using the GeneMorph II Random Mutagenesis Kit (Agilent Technologies). We used 25 ng of the plasmid construct with the MG1655 *lacZ* promoter variant cloned into it as a template DNA for the error-prone PCR to achieve approximately 1.5 SNPs per variant sequence. We used the same primers as for the segregating promoter variants **Table S4.3**. The reaction with randomly mutated promoter variants was DpnI-treated and column-purified before Gibson assembly with the NEBuilder® HiFi DNA Assembly Master Mix (New England Biolabs). The assembly mix was then electroporated into the MG1655 strain and colonies that grew on LB with Kanamycin were picked for Sanger sequencing and stored as glycerol stocks. Clones which had none or more than three SNPs in the cloned promoter insert were excluded as well as those with SNPs detected in

the vector backbone (rare occasion). The rest of the clones were then re-grown in liquid LB with Kanamycin in 96 deep-well microplates overnight and 96 well microplate glycerol stock libraries were prepared from them.

Segregating *lacZ* variants in native genetic background

We selected three environmental *E. coli* isolates: SC312, SC400, and SC418 (Ishii et al. 2006) for comparing the expression from identical segregating promoter variants inside and outside of their native genetic background (i.e., in MG1655 strain). These three isolates were selected for the different expression phenotypes produced from their *lacZ* promoter variants in the MG1655 genetic background.

We isolated the plasmids containing the three segregating *lacZ* promoter variants from the MG1655 strain using StrataPrep Plasmid Miniprep Kit (Agilent). We also isolated the promoter-less vector pMV001 the same way. We then electroporated the plasmids containing the segregating *lacZ* promoter variants into the *E. coli* isolates of the variant origin. All three isolates were also transformed with the promoter-less vector pMV001. We confirmed the presence of all desired plasmids by Sanger sequencing across the promoter insert region from clones that grew on LB with Kanamycin and aliquoted them into a microplate library.

We prepared the environmental *E. coli* isolates for electroporation as follows: we inoculated 100 ml of liquid LB in 1 l flask with 1 ml of overnight culture of environmental *E. coli* isolate (grown in liquid LB at 37°C with shaking). We then incubated the flask at 37°C with shaking (250 rpm) until the isolate reached mid-exponential phase. Once the optical density (OD_{600}) of the media reached values between 0.5 and 0.6 we cooled down the culture by swirling the flask in ice slurry for at least 5 min and then placed it into a 4°C fridge inside the ice slurry for 1 h. Next, we spinned the cold culture at 4,200 G for 5 min at 4°C, discarded the supernatant and washed the cell pellet twice in cold 10% glycerol with centrifugation at 4,500 G for 5 min at 4°C. At the end we aliquoted 70 μl of cells in the small amount of remaining 10% glycerol into cold 1.5 ml tubes. We stored these aliquots at -80°C until use.

Segregating *lacZ* variants in MG1655 chromosome

To modify the *lacZ* promoter of MG1655 strain at the native locus in the chromosome and obtain translational ligation of *lacZ* and GFP genes, we implemented a landing pad assay (Tas et al. 2015). In detail, we first replaced the whole *lacZ* promoter and gene, including a small part of *lacI* gene by the landing pad (TetR gene). However, using the original pTKRED helper plasmid interfered with the homologous recombination during this first step due to *lacI* gene presence in the pTKRED sequence. Because this step relies only on the λ Red system we replaced pTKRED with pKD46 helper plasmid (Datsenko and Wanner 2000). We confirmed successful landing pad integration by Sanger sequencing across the TetR insert site and heat-cured the resulting strain of the pKD46 helper plasmid using its temperature sensitive origin of replication.

We constructed the first donor plasmid pMV002 so that it contained the MG1655 *lacZ* promoter variant and MG1655 *lacZ* gene translationally ligated to GFPmut2. For this purpose we first digested the synthesized insert (**Supplementary Note**) from the plasmid

on which it was delivered (Twist Biosciences) and the vector pTKDP-*neo* (Tas et al. 2015) using I-SceI restriction enzyme followed by rSAP treatment to avoid self-ligation (New England Biolabs). We then ran the digested products on 1% agarose gel and extracted the desired fragments using StrataPrep DNA Gel Extraction Kit (Agilent). We then assembled these two fragments using the NEBuilder® HiFi DNA Assembly Master Mix (New England Biolabs) with the help of four single strand oligo bridges (**Table S4.3**). The other donor plasmids were constructed using the approach described above for plasmid-based libraries, just replacing the pMV001 vector with pMV002 and using a different set of primers for vector and insert PCR amplification as listed in **Table S4.3**.

The donor plasmids were then transformed into the MG1655 strain having the TetR landing pad in *lac* operon and pTKRED helper plasmid. To confirm successful integration of the target sequence into the MG1655 chromosome we sequenced the whole genome of colonies that appeared blue on LB agar plates with 2 mM IPTG and 20 μg/ml X-gal after heat-curing the pTKRED helper plasmid. We then constructed a microplate library from the obtained clones.

4.4.2 Flow cytometry assays

The assays and data analysis were performed the same way as previously described for the *lacZ* promoter (Vlková and Silander 2021). In short, the bacterial clones in the libraries were first inoculated into 0.5 ml of M9 minimal media with 0.4% glucose (and Kanamycin for plasmid-based libraries). After overnight growth in M9 glucose, we re-inoculated the libraries into 0.5 ml of one of the three assay media: 0.4% glucose, 0.4% galactose or 0.4% lactose. After this second overnight growth, we inoculated the libraries into the same fresh assay media into three separate microplates to obtain triplicates for each clone. The library containing segregating *lacZ* variants in the MG1655 chromosome had a different layout. This means that each triplicate was inoculated into separate wells since the first overnight in M9 with 0.4% glucose. Once the cells reached exponential growth, we diluted them into 1x PBS with ∼2.5% formaldehyde.

We performed the flow cytometry on a BD FACSCanto II machine using BD FACSDiva software version 6.1.3. We used a 488 nm laser and a 513/17 nm bandpass filter to obtain the GFP fluorescence data. We set the number of events to record from each well to 20,000. We exported the acquired data from FACSDiva software into Flow Cytometry Standard files, and performed all cell gating and fluorescence analysis using custom R scripts (flowCore package version 2.0.1; available through https://doi.org/110.5281/zenodo.5525368). We gated cells based on their maximal kernel density of forward and side scatter values, using the ellipsoidGate function from the flowCore package, and keeping about 1/3 of all events.

The modal population fluorescence was calculated as the mean of the three maximal kernel density values from the GFP fluorescence signal of three replicates. The modal coefficient of variation (mCV) was calculated separately for each of the three biological replicates (standard deviation divided by the modal population fluorescence), and the mean of these values was used as the mCV of the promoter. Replicates with fewer than 2,500 and 5,000 recorded events were excluded from the calculation of modal population fluorescence and mCV, respectively. We set the limit to 5,000 events for mCV to avoid outlier events such as machine noise to affect the mCV calculations (Vlková and Silander 2021).

When comparing *lacZ* variants that came from two separate microplate layouts, we obtained an offset for both modal population fluorescence and mCV to minimise plate-effects. We calculated these offsets as the mean of the differences between the two or three controls present in each microplate, i.e., the MG1655 promoter variant (in plasmid-based libraries only), plpp::GFPmut2 (positive control), and negative control (pMV001 or MG1655 wild-type strain). All figures contain modal fluorescence and mCV values that are corrected using these offsets. All scripts using the workflow described here, including the original data files can be accessed through https://doi.org/110.5281/zenodo.5525368.

4.4.3 Quantifying transcriptional activity

To test for differences in the variation in transcriptional activity between the segregating and random variants, we calculated the modal population fluorescence values for each promoter variant in all libraries and environments as a proxy for transcriptional activity. We then tested for significant differences in variation in modal population fluorescence levels between groups using the Fligner-Killeen test of homogeneity of variances. We further tested whether increases in modal fluorescence in random variants are equally probable as decreases using two-sided binomial tests. These tests were performed against modal population fluorescence values from the MG1655 variant in each environment. The code for these calculations can be accessed through https://doi.org/110.5281/zenodo.5525368.

4.4.4 Quantifying phenotypic plasticity

We calculated the phenotypic plasticity of promoter variants across all three environments as described earlier (Vlková and Silander 2021). In short, we calculated the Euclidean distance of each datapoint in three dimensions to an isocline representing null plasticity. The isocline is defined by equal values in the two or three dimensional space, i.e., x = y or x = y = z. The dimensions are defined by the modal population fluorescence values in the two or three environments compared. Each datapoint (promoter variant) is thus defined by its fluorescence values in compared environments. The closer a datapoint is to the isocline, the lower the plasticity. To test whether natural selection has acted on plasticity, we compared plasticity values from segregating and random promoter variants using a two-sided Wilcoxon rank-sum test. We also tested whether increases in plasticity in random variants are as frequent as decreases in plasticity using a two-sided binomial test. We used plasticity values from the MG1655 variant as reference for these tests in each environment combination (the code for these calculations can be accessed through https://doi.org/110.5281/zenodo.5525368).

4.4.5 Quantifying transcriptional noise

We used the same metric of noise as described previously (Vlková and Silander 2021). In detail, to determine the noise within an isogenic cell population of each promoter variant, we first excluded fluorescence values from the population that were lower or higher than three standard deviations from the modal population fluorescence level. Then we calculated the mCV from the isogenic cell population as the standard deviation of the fluorescence divided by the modal population fluorescence level. We next fitted a cubic smoothing

spline (smoothing parameter lambda = 0.01) to the modal population fluorescence vs. the mCV values, using all (segregating and random) promoter variants. We determined the noise levels as the difference in measured mCV from the mCV predicted from the fitted spline. We then compared the noise values between segregating and random variants using a two-sided Wilcoxon rank-sum test to determine whether selection has acted on noise. We also tested whether increase in noise due to random mutations is as likely as decrease in noise using a two-sided binomial test. Noise levels from the MG1655 variant served as a reference for the binomial test (the code for these calculations can be accessed through https://doi.org/110.5281/zenodo.5525368).

4.4.6 Comparison of fluorescence values from variants with and without RiboJ

The pMV001 vector we used in this study was constructed by introducing self-cleaving ribozyme RiboJ into pUA66 vector upstream of GFPmut2 gene (Lou et al. 2012, Vlková et al. 2021, Zaslaver et al. 2006). In order to assess the changes in expression caused by RiboJ presence we used part of the dataset published earlier that contains segregating and random *lacZ* promoter variants that are identical to those used in this study, but which are cloned into pUA66 vector and thus without RiboJ (Vlková and Silander 2021).

When mapping the effect size and direction of randomly introduced SNPs we used modal population fluorescence values only from those random variants which contained a single SNP relative to the MG1655 *lacZ* promoter variant. Information about the TF binding sites, -10 and -35 elements, and ORFs was taken from the EcoCyc database (Karp et al. 2018), and only the annotations associated with σ^{70} driven TSS were used. Using the Sanger sequencing results, we identified the location of all SNPs for each random variant. We also compared the modal population fluorescence values from all variants shared between pUA66 and pMV001-based systems using Spearman's correlation test (the code for all calculations can be accessed through https://doi.org/110.5281/zenodo.5525368).

4.5 Results

4.5.1 A plasmid based approach to infer the action of selection on transcriptional control

In order to investigate the selection pressure acting on the transcriptional regulation of the *lacZ* promoter in *E. coli*, we first quantified genetic diversity in the regulatory region of the *lac* operon in a collection of 135 environmental *E. coli* isolates (Ishii et al. 2006). Here, we have designated the *lacZ* promoter as the entire *lacI-lacZ* intergenic region (IGR) as well as 88 bp of the 3' end of the upstream *lacI* ORF and 69 bp of the 5' end of the downstream *lacZ* ORF (**Figure 4.1a**). We include these regions, as ORFs that flank IGRs are known to frequently affect regulation of gene expression (Vlková et al. 2021, Zaslaver et al. 2006). We identified 26 *lacZ* promoter variants segregating in this set of environmental *E. coli* isolates (**Figure S4.1** and **Table S4.1**).

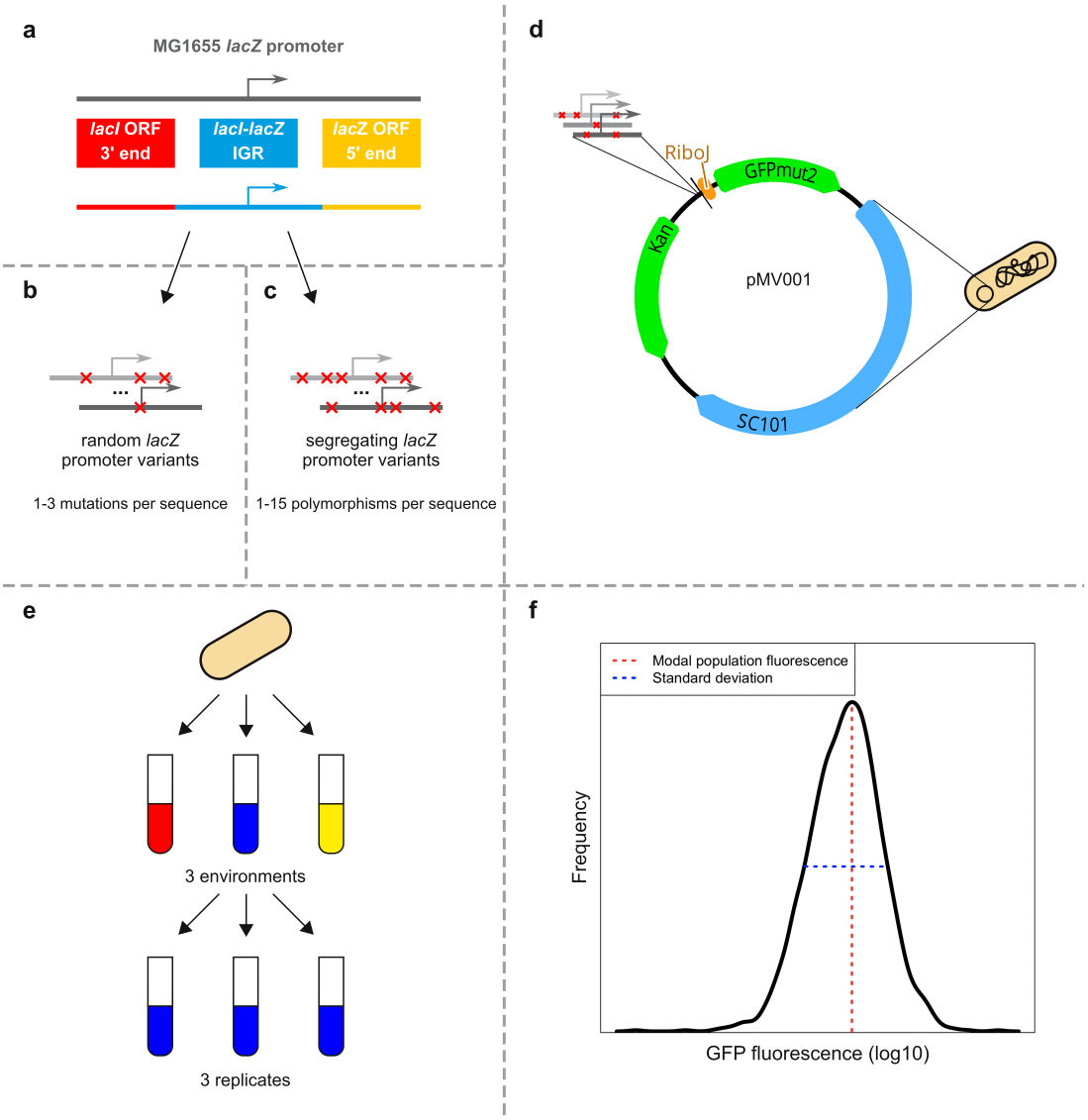

Figure 4.1: **Experimental design to assay the effects of segregating polymorphisms and random mutations on transcriptional control from *lacZ* promoter. a)** The sequences we refer to as *lacZ* promoter variants consist of the *lacI-lacZ* intergenic region (IGR), 88 bp of the 3' end of the upstream *lacI* open reading frame (ORF), and 69 bp of the 5' end of the downstream lacZ ORF. **b)** We performed random PCR mutagenesis using the MG1655 *lacZ* promoter variant as a template with a target mutation rate of 1.5 mutations per promoter variant. **c)** We also PCR-amplified variants of *lacZ* promoter segregating among environmental *E. coli* isolates from a DNA pool. The average number of polymorphisms across obtained segregating variants (as compared to MG1655) is almost 10-times higher than the average number of mutations for random variants. **d)** We cloned the resulting PCR amplicons (both segregating and random) upstream of GFPmut2 (Vlková et al. 2021). We Sanger sequenced all the promoter variants to confirm the presence and location of mutations. From mutagenesis, only the variants containing 1 to 3 mutations were used for further phenotypic assays. **e)** Each bacterial clone was cultured in media containing glucose, galactose, or lactose as the primary carbon source in triplicate. **f)** Using flow cytometry we quantified the modal population fluorescence and the standard deviation of population fluorescence.

In nature, all newly generated variants (i.e. those arising through mutation or recombination) are filtered by natural selection based on their phenotypic effects. Our goal here is to understand the effects of selection on transcriptional regulation. However, many of the polymorphisms present in segregating promoter variants are downstream of the *lacZ* transcription start site, and thus present in the transcribed mRNA. These changes in the *lacZ* mRNA sequence may result in phenotypic changes by altering translation, for example by changing mRNA folding or ribosome initiation. Thus, we sought a method to disentangle phenotypic differences that are mediated by changes in transcription from phenotypic differences mediated by changes in translation.

To do this, we designed a system in which a GFP open reading frame is placed downstream of a promoter variant on a low-copy number plasmid (Zaslaver et al. 2006). We placed a self-cleaving ribozyme, RiboJ, upstream of the GFP open reading frame (**Figure 4.1d**) (Lou et al. 2012, Vlková et al. 2021). When this mRNA is transcribed, RiboJ immediately cleaves and removes the 5' end of the mRNA, including any polymorphisms or mutations that are present (**Figure 4.2**). This cleavage is extremely rapid, and at any particular time, more than 95% of all mRNAs in a cell are cleaved. Thus, any observed effects of polymorphisms on GFP expression and fluorescence are necessarily mediated solely by changes in transcription (Vlková et al. 2021). This contrasts with mRNA molecules lacking RiboJ, as sequence changes present in these mRNAs may also affect GFP expression and fluorescence by affecting translation.

To understand the filtering process of natural selection on transcriptional regulation, we used PCR mutagenesis to generate a set of promoter variants with random mutations that have not been filtered by selection (**Materials and Methods**). By comparing the regulatory phenotypes of segregating and random promoter variants, we can infer how natural selection has acted. Note that we designate genetic differences in segregating promoter variants as "polymorphisms," and genetic differences in the random promoter variants as "mutations" to differentiate genetic differences that have or have not been filtered by selection.

To generate this set of promoter variants unfiltered by selection, we took the MG1655 variant of the *lacZ* promoter (**Figure 4.1a**) and performed PCR mutagenesis, generating approximately 1.5 mutations per promoter variant (**Figure 4.1b** and **Figure S4.2**). We cloned these promoters upstream of the GFP open reading frame on the plasmid-based system with RiboJ described above. At the same time, we cloned the 18 of the 26 segregating promoter variants into the same plasmid (**Figure 4.1c**, **Figure 4.1d**, **Table S4.1**, and **Table S4.2**).

We sequenced all the random promoter variants, and selected 92 that contained between one and three mutations. We note that this random promoter library is considerably less diverse than the segregating variants: the average pairwise identity for all segregating variants is 96.25%, compared to 98.97% for the random variants.

We next quantified the regulatory phenotypes of each of the promoter variants in each of the libraries. We transformed all the promoter constructs into *E. coli* K12 MG1655 (**Table S4.1** and **Table S4.2**), and grew each clone in the presence of three different carbon sources (0.4% glucose, 0.4% galactose, and 0.4% lactose) until exponential growth was reached (**Figure 4.1e**). We then measured the amount of fluorescence from each cell using flow cytometry. We used the changes in fluorescence as a proxy for changes in transcriptional activity (Vlková et al. 2021). For each sample population we calculated

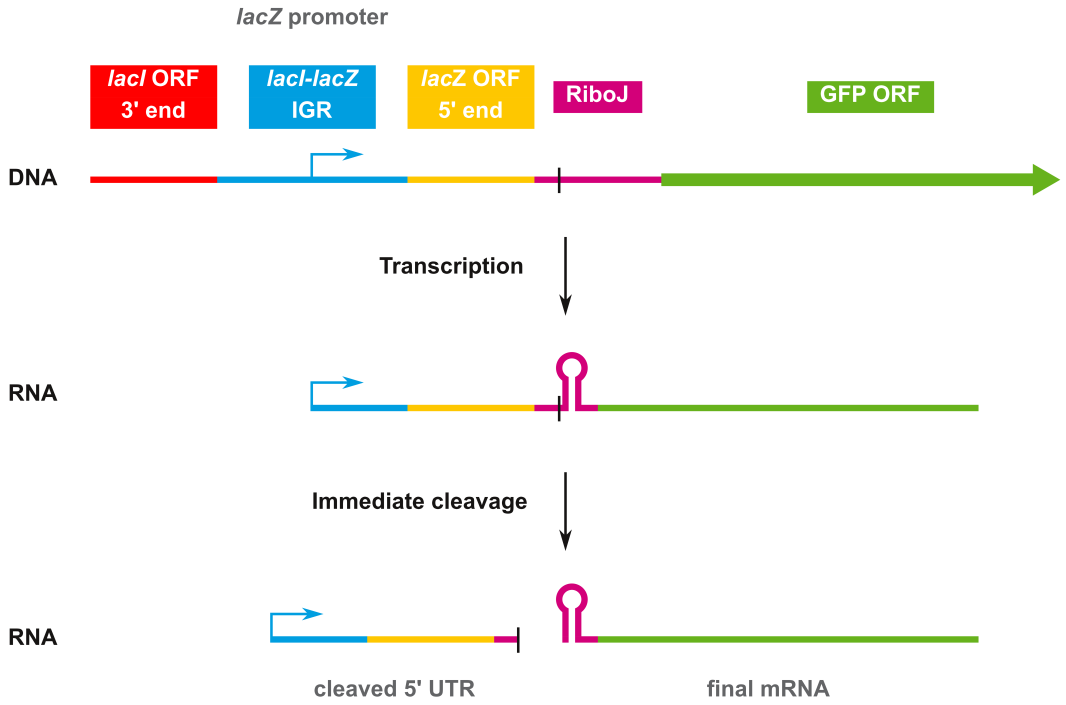

Figure 4.2: **Removal of 5' sequence variation via RiboJ ribozyme autocatalytic cleavage activity.** The RiboJ sequence is placed between the *lacZ* promoter and GFP open reading frame (ORF). The arrow in the middle of the *lacI-lacZ* intergenic region (IGR) represents the transcription start site. When transcription is induced from the *lacZ* promoter, the RiboJ RNA sequence produces a strong secondary RNA structure resulting in fast and efficient cleavage at a specific site, cutting off most of the 5' untranslated region (UTR). The final mRNAs that undergo translation from different *lacZ* promoter variants are thus identical, even when there are polymorphisms downstream of the transcription start site. The differences in GFP fluorescence measured from the promoter variants are the result of differences solely in transcription.

two metrics of regulatory phenotypes, the modal population fluorescence and the standard deviation (**Figure 4.1f**). All measures were taken from three full biological replicates (**Materials and Methods**).

We first compared the fluorescence levels we observed from promoter variants with RiboJ to the fluorescence levels of identical promoter variants lacking RiboJ that we had previously generated (Vlková and Silander 2021). We found that the fluorescence levels from the promoter variants in these two systems were highly correlated, with correlations of 0.76, 0.83, and 0.87 in the glucose, galactose, and lactose environments, respectively (Spearman's rho; **Figure S4.3**). We also observed an approximately ten-fold systematic increase in fluorescence levels when RiboJ was present. This increase was expected, and has been ascribed as resulting from the specific mRNA folding after the RiboJ self-cleaving (Carrier and Keasling 1997, Clifton et al. 2018, Neves et al. 2020). Thus, the addition of RiboJ brings two advantages. First, it allows us to determine the extent to which selection has acted on regulatory phenotypes mediated solely via transcriptional mechanisms. Second, the systematic increase in fluorescence in the presence of RiboJ increases the sensitivity

of our measurements.

4.5.2 Directional selection acts on *lacZ* promoter activity in glucose and lactose

To infer the filtering effects of natural selection we compared the fluorescence levels of the random promoter variants to the segregating promoter variants and to the progenitor MG1655 variant. We expected that if there were systematic differences in fluorescence between these sets of variants, that this is caused by selection tending to remove some new variants. For example, if most random variants exhibit systematically higher fluorescence than segregating variants, this suggests that selection has removed newly generated variants with higher fluorescence levels.

We found that the segregating variants differed by approximately three-fold in transcriptional activity in each of the environments (**Figure 4.3a**). While the majority of random variants also ranged approximately three-fold in transcriptional activity, a small number had more than ten-fold higher or lower transcriptional activity than the MG1655 variant from which they were derived. However, segregating variants are approximately four-fold more genetically diverse. We stratified the random variants according to the number of random mutations they had compared to the progenitor MG1655 promoter variant. We found that as the number of random mutations in a promoter increased, their transcriptional activity diverged more and more from the transcriptional activity of the segregating variants (**Figure 4.3b**).

Furthermore, the divergence in transcriptional activity was not symmetric across environments. Specifically, we found that in the glucose environment, in which the *lac* operon is highly repressed, the vast majority of random variants increased transcriptional activity. We calculated the probability of the observed distribution of effects to a distribution symmetric around the activity of the MG1655 variant, and found that the fraction of random variants that increased activity was much greater than expected (68 out of 92, p = 4.9e-06, two-sided binomial test). In contrast, in the lactose environment, in which the *lac* operon is highly active, the majority of random variants decreased their transcription (71 out of 92, p = 1.6e-07, two-sided binomial test). We found no difference in the galactose environment (p = 0.75). However, we note that these results are highly dependent on the precision of the fluorescence levels that we have measured for the MG1655 variant, although these are likely to be accurate as we performed all reported measurements in triplicate.

In addition, in all environments, we found that segregating variants exhibited lower phenotypic variation compared to random variants (**Figure 4.3b**) although these results were not significant (p = 0.07 for glucose; p = 0.10 for galactose; p = 0.80 for lactose, Fligner-Killeen test of homogeneity of variances). This is despite segregating variants harboring much higher levels of genetic variation. These data thus suggest that selection has acted to remove mutations that have large effects on transcriptional activity. Moreover, the results in the glucose environment suggest that in MG1655, selection has acted to minimize transcriptional activity, as the majority of random variants increase activity. Conversely, in MG1655, selection has acted to maximize transcriptional activity in the lactose environment.

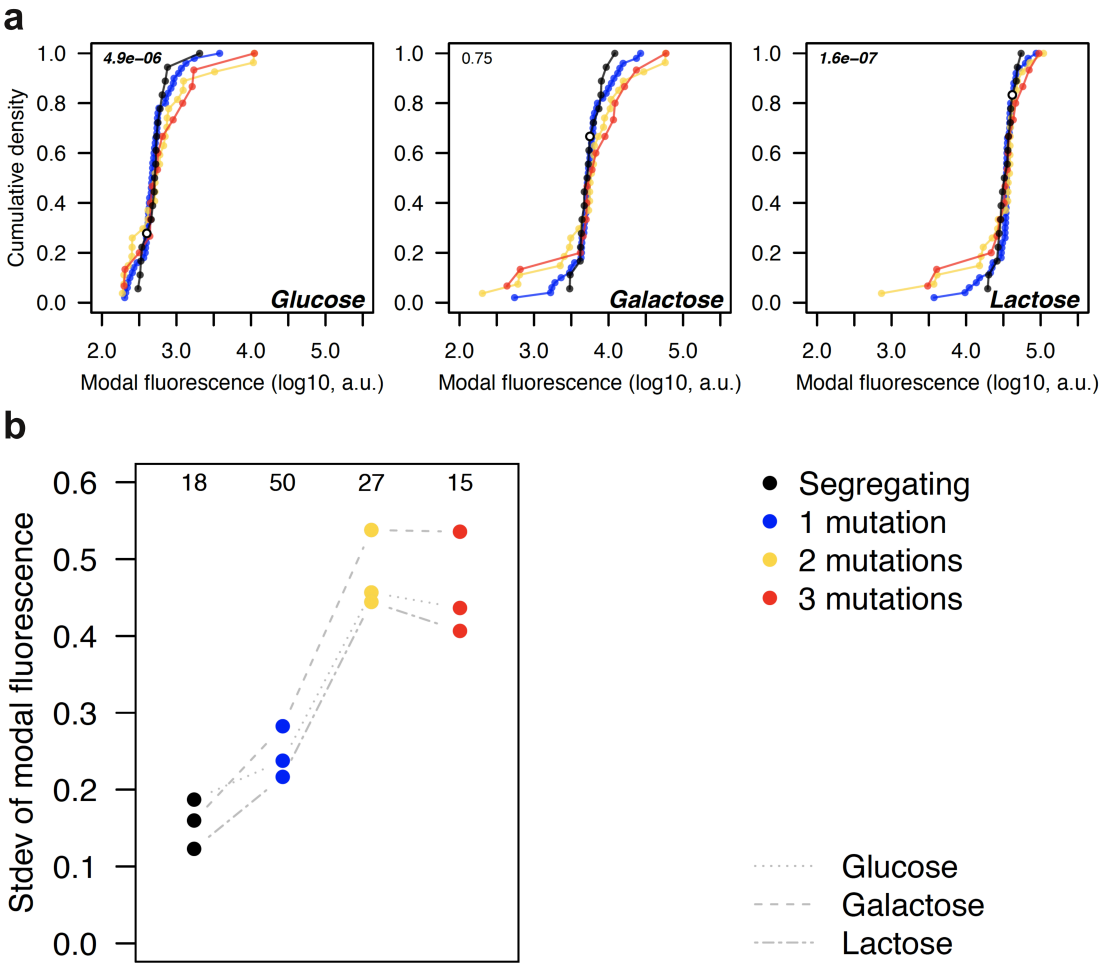

Figure 4.3: **Selection acts to maintain similar transcription levels in segregating variants. a)** Cumulative density of modal population transcriptional levels of random and segregating *lacZ* promoter variants in the MG1655 genetic background in all environments (glucose, galactose, and lactose). Random variants are stratified by the number of mutations they differ from the MG1655 variant from which they were derived. Segregating variants include the MG1655 variant which is displayed as a white circle in each plot. The numbers in the top-left corner represent p-values from the two-sided binomial test indicating whether random variants are as likely to increase as decrease transcriptional level when the MG1655 variant is randomly mutated. **b)** Standard deviation (stdev) of modal population fluorescence values for each promoter variant group in all environments. The standard deviation generally rises with an increasing number of mutations among random variants. The segregating variants have among the lowest standard deviations while having up to 15 polymorphisms relative to the MG1655 variant. The numbers above each group of points indicate the sample sizes.

4.5.3 Selection on transcriptional plasticity

Regulating transcription such that it is maximised in one environment and minimised in a second can be viewed as maximising transcriptional plasticity - the amount by which levels of transcriptional activity are dependent on environmental conditions. To visualize plasticity in two environments, we can plot the fluorescence level in one environment versus

the fluorescence level in the second environment, with the x = y isocline indicating equal fluorescence and thus equal transcriptional activity in both environments. We quantified plasticity as the shortest distance from this isocline (**Figure 4.4a**). Thus, promoter variants that lie directly on this line have equal transcriptional activity in both environments and no plasticity; the further they are from this line, the higher their plasticity. When considering three environments, we quantified plasticity using an analogous measure: the shortest distance to the isocline from a point defined by the fluorescence levels in all three environments considered here: glucose, galactose, and lactose (**Figure 4.4c**). Note that plasticity calculated from the combination of all three environments (**Figure 4.4d**) strongly mirrors the pattern of plasticity from the combination of glucose and lactose only (**Figure 4.4b**, Glu:Lac). This is due a strong dependence of the plasticity from three environments on the largest difference in fluorescence levels between any two environments, which in the case of *lacZ* promoter variants are glucose and lactose, as expected.

We found that almost all random variants had reduced plasticity compared to MG1655 (81 out of 92, p = 2.5e-14 for glucose vs. galactose; 90 out of 92, p = 1.7e-24 for glucose versus lactose; 83 out of 92, p = 4e-16 for galactose vs. lactose, two-sided binomial test, **Figure 4.4b**). Together with the results above, this suggests that for MG1655, selection has acted to maximise plasticity by minimising transcription in glucose and maximising it in lactose.

Relative to MG1655, we found that segregating variants generally exhibited lower plasticity, and in fact did not differ in plasticity compared to the random variants (p = 0.75 for glucose vs. galactose; p = 0.48 for glucose vs. lactose; p = 0.81 for galactose vs. lactose; p = 0.47 for all three environments, two-sided Wilcoxon rank-sum test). However, random promoters with more than two mutations exhibited a long tail of lower plasticity (**Figure 4.4b**).

The absence of strong differences between segregating and random variants in plasticity indicates that selection on transcriptional activity and plasticity in the *lacZ* promoter is quite relaxed, and new mutations are not strongly filtered by natural selection. Previously, we showed that relative to other promoters in *E. coli*, segregating variants of the *lacZ* promoter have the largest differences in expression level, supporting this hypothesis of relaxed selection (Vlková and Silander 2021). Alternatively, plasticity may be highly dependent on the genetic background; thus, the plasticity that we have measured in MG1655 for each of the segregating variants may be lower than would be observed in the native background of the variants.

4.5.4 Both diversifying and directional selection act on transcriptional noise

Finally, we tested whether selection has acted to filter mutations according to their effects on transcriptional noise. Here, "transcriptional noise" refers to the cell-to-cell variability in transcriptional activity within an isogenic population. As the variation in fluorescence is highly dependent on the modal fluorescence level, we quantified the level of transcriptional noise by calculating the vertical deviation from a spline fitted to modal fluorescence versus the standard deviation divided by the modal fluorescence, a measure analogous to the coefficient of variation and which we refer to as the modal CV, or mCV (Vlková and Silander

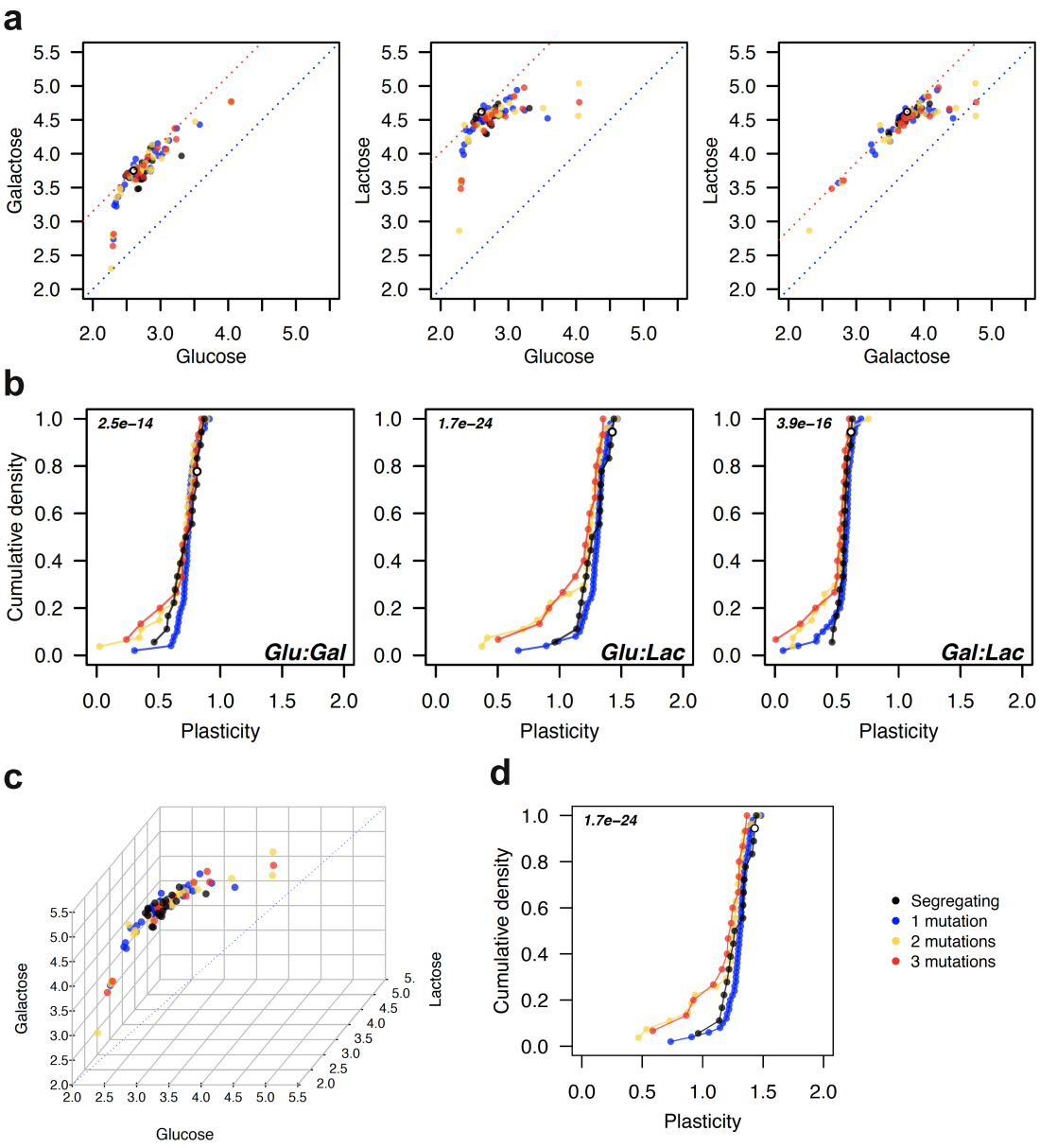

Figure 4.4: **Directional selection acts to maintain high transcriptional plasticity in segregating variants. a)** Comparison of modal fluorescence levels from random and segregating variants in pairs of environments as indicated by the x and y axis labels. The plasticity from a pair of environments is calculated as the distance of a datapoint (promoter variant) from an isocline of equal fluorescence levels (x = y, blue dotted line). Promoter variants lying on the red dotted line have plasticity levels equal to the MG1655 variant through which the line runs. **b)** Cumulative plasticity levels from random and segregating variants in pairs of environments. **c)** Comparison of modal fluorescence levels from random and segregating variants in all three environments. The plasticity from three environments is calculated as the distance of a datapoint (promoter variant) from an isocline of equal fluorescence levels (x = y = z, blue dotted line), analogous to the approach in **a**. *(continues on the next page)*

Figure 4.4: *(continues from the previous page)* **d)** Cumulative transcriptional plasticity levels from random and segregating variants in all three environments. Across all panels the random variants are stratified by the number of mutations they differ from the MG1655 variant from which they were derived. Segregating variants include the MG1655 variant which is displayed as a white circle. The numbers in the top-left corner in panels **b** and **d** represent p-values from the two-sided binomial test indicating whether random variants are as likely to increase as decrease plasticity when the MG1655 variant is randomly mutated.

2021). This measure of noise is independent of fluorescence level. In addition, because the variation in population fluorescence is highly dependent on the growth environment, we fit a spline for all promoter variants in each environment. This allows us to infer how selection is acting on transcriptional noise within a specific growth environment, rather than between environments, for which other selective factors (e.g. selection on transcriptional activity) may overwhelm selection on noise. With this noise metric, variants with a higher mCV than expected given their modal fluorescence value have "high" noise, while variants with lower mCV than expected have "low" noise. As noted above, this metric also results in a different value of noise being calculated for each promoter variant in each environment (**Figure 4.5a**).

We first considered differences in transcriptional noise between segregating variants and random variants in lactose, where we expect full transcriptional activation. We found that all except for one of the 92 random variants (derived from the MG1655 promoter variant) exhibited higher levels of noise than the MG1655 variant (**Figure 4.5b**). The MG1655 variant exhibited similar noise levels as the other segregating variants, and considering all segregating variants together, we found that they had significantly lower levels of transcriptional noise compared to random variants (median of -0.0021 $\pm$ 0.0003 for segregating variants vs 0.0005 $\pm$ 0.0013 for random variants, p = 9.4e-11, two-sided Wilcox rank-sum test). This suggests that in natural populations almost all mutations that detectably increase transcriptional noise when lactose is the primary carbon source are filtered out, indicative of selection for low transcriptional noise in lactose.

In strong contrast, in glucose, where we expect almost full repression of the *lac* operon, the majority of random variants had lower levels of noise than their MG1655 progenitor (61 out of 92, p = 2.3e-03, two-sided binomial test; **Figure 4.5b**). In addition, we found that transcriptional noise varied considerably between segregating variants, with nine segregating variants, including MG1655, exhibiting high noise (between 0 and 0.01), and nine exhibiting low noise (all less than -0.015, **Figure 4.5b**). Surprisingly, a small number of the random variants had noise levels that approached those of the low-noise segregating variants. This suggests that low levels of noise are easy to achieve from a mutational standpoint, but are not always selected for. This might be due to most low noise phenotypes having deleterious effects on other regulatory phenotypes (transcriptional activity, plasticity, or both). It is also possible that for some isolates, high noise is advantageous when glucose is the primary carbon source.

Finally, in galactose, we found no strong differences in noise levels between the random and segregating variants, although the random variants followed a similar pattern to that observed for other phenotypes, as variants with greater genetic divergence from MG1655 also exhibited great phenotypic divergence. This is apparent in the tails of the distributions, which are dominated by the random mutants with two and three mutations relative to the

progenitor MG1655 variant (**Figure 4.5b**).

We next sought to gain some insight into the observation that nine segregating variants exhibited high noise in glucose, while nine others exhibited low noise. We first tested whether the high- and low-noise phenotypes were a general phenomenon for each promoter variant, or whether the noise each variant exhibited was specific to regulation in glucose. Although we observed a strong positive correlation between noise in glucose and noise in galactose for the segregating variants (Spearman's rho = 0.78, p = 2.4e-04), we found no

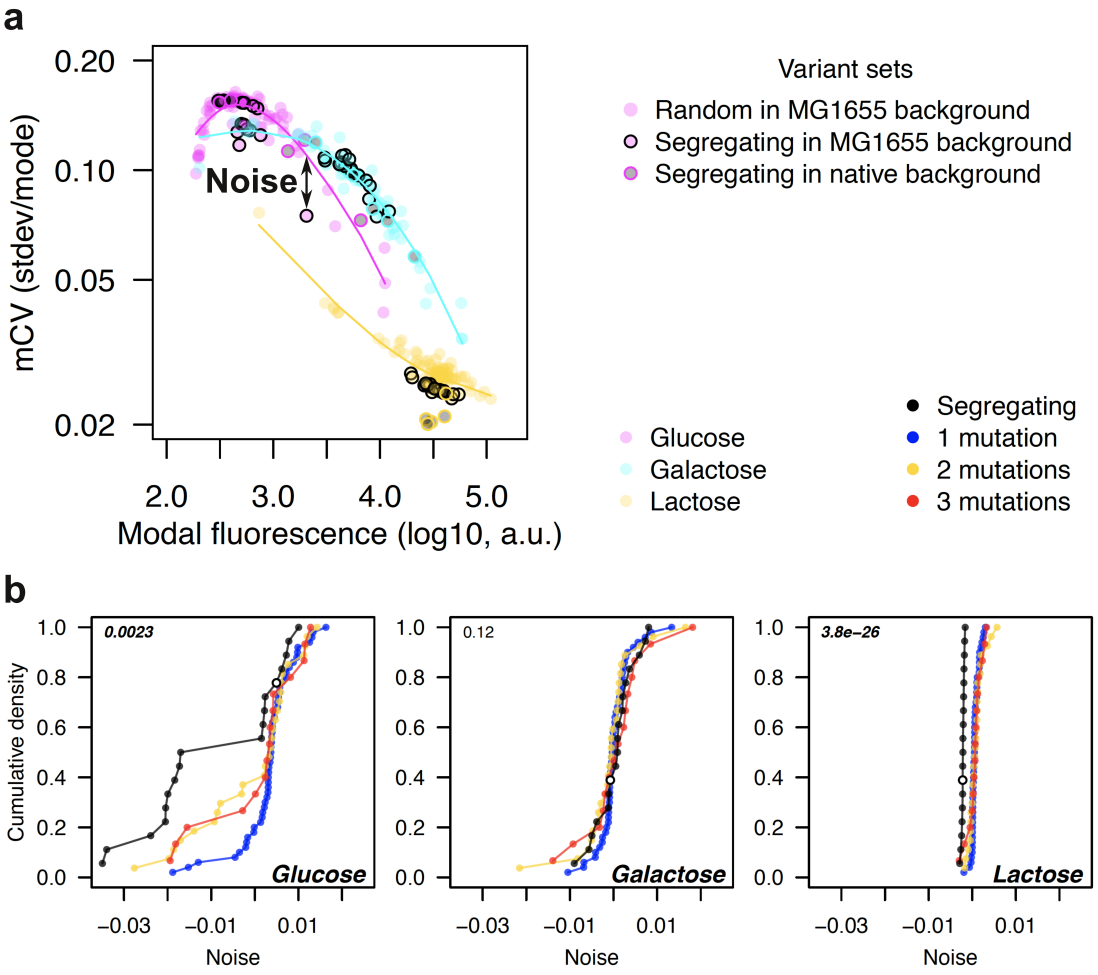

Figure 4.5: **Selection on transcriptional noise levels in segregating *lacZ* promoter variants. a)** We fit smoothing splines to modal population fluorescence vs. the standard deviation of population fluorescence normalised by modal fluorescence (mCV) for all variants (segregating and random) in all genetic backgrounds. We fit a separate spline for each environment. The term "noise" is used for the vertical deviation of each variant, i.e., deviation in the mCV from the fitted spline as indicated by the arrow. **b)** Cumulative distributions of transcriptional noise from random and segregating variants. Random variants are stratified by the number of mutations they differ from the MG1655 variant from which they were derived. Segregating variants include the MG1655 variant, which is displayed as a white circle. The numbers in the top-left corner represent p-values from a two-sided binomial test indicating whether random variants are as likely to increase as decrease transcriptional noise.

correlation between noise levels in glucose and lactose (Spearman's rho = 0.22, p = 0.39; **Figure 4.6b**). This suggests that there is specific selection for high or low transcriptional noise in glucose (diversifying selection) while in lactose, directional selection has led to the filtering of mutations that increase transcriptional noise.

We looked in more detail at the genotypes of the segregating *lacZ* promoter variants with high and low transcriptional noise phenotypes in glucose (**Figure 4.5b**). We found that two - and only two - polymorphisms segregated perfectly between these two phenotypes. These two polymorphisms were located 7 bp upstream (-7 bp) and 6 bp downstream (+6 bp) of the *lacZ* gene start codon (**Figure 4.6a**). The low noise variants had T at both positions ("TT genotypes"), while high noise variants had A and C at the respective positions ("AC genotypes"). We compared the sequences at the *lac* locus across the entire collection of 135 environmental *E. coli* isolates, and found that the frequency of TT and AC genotypes among the isolates was close to equal (62 TT vs. 63 AC). Ten of the 135 environmental *E. coli* isolates had a TC genotype. We used a core genome phylogeny to determine the distribution of these polymorphisms, and found that they segregated almost perfectly, with all isolates in the B2, D, E, and F clades having the TT genotype associated with low noise in glucose, and the majority of isolates in the A and B1 clades having the glucose high-noise AC genotype, with six exceptions (**Figure 4.6c**). These six exceptions appeared to be horizontal gene transfers from the B2, D, E, or F clades. Almost all isolates with the TC genotype were in a single highly diverged cluster of *E. marmotae*, suggesting that this genotype is ancestral, or that a rare recombination event or mutational reversion has occurred at this locus.

Notably, we found a single random variant that had changed from a high-noise AC genotype to a TC genotype, but no other random variants had changes at either the -7 bp or +6 bp positions. The -7 bp T polymorphism is associated with low-noise phenotypes, and we found that this variant exhibited a low-noise phenotype (-0.014), in contrast to the phenotype of its progenitor MG1655 variant (0.008). However, this random variant also had one additional mutation located in the LacI O3 binding site, which may also have an effect on its noise phenotype.

Nevertheless, these results provide circumstantial evidence that there are single mutations that may have strong effects on noise. Furthermore there has been long-term maintenance of genetic variants associated with either high-noise or low-noise phenotypes. This is consistent with there being balancing selection on the transcriptional noise levels conferred by *lacZ* promoter variants in a glucose environment. However, there is no strong evidence that these genotypic associations with noise phenotypes are causal, and there are a range of other reasons that could lead to the long-term maintenance of these polymorphisms, for example correlated selection on other characters, or simply chance.

4.5.5 Effects of genetic background on transcriptional phenotypes

Above we have quantified the *cis*-effects of segregating polymorphisms and random mutations, and have shown that in general there is directional selection acting to minimise transcriptional activity in glucose and maximise activity in lactose. We found that the vast majority of random mutations decrease plasticity relative to the MG1655 promoter variant, but there is no clear evidence that segregating promoters have been selected to maximise

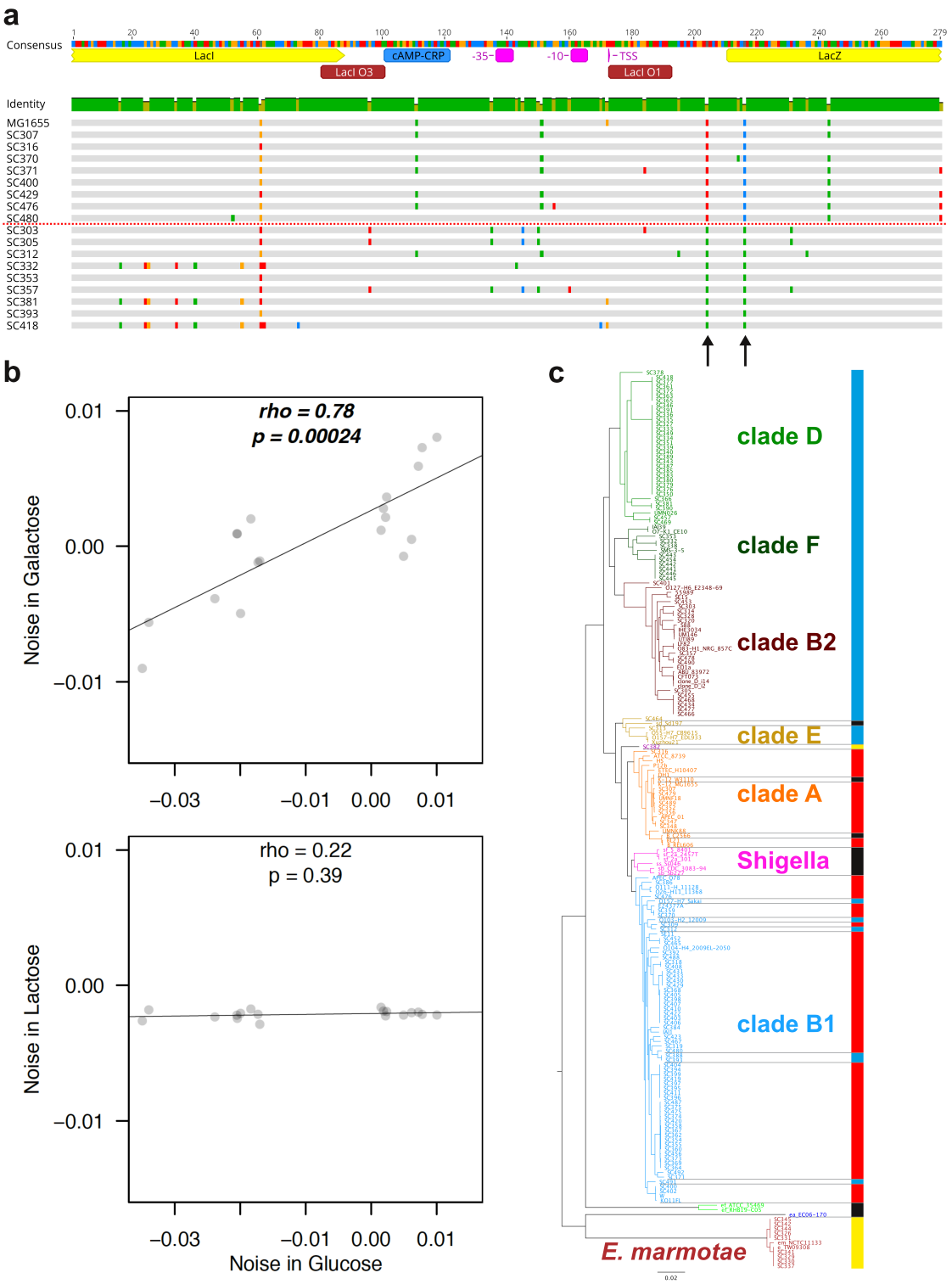

Figure 4.6: **Segregating genotypes associated with high and low transcriptional noise in glucose. a)** Alignment of segregating promoter variants. The horizontal red dotted line divides variants with high noise in glucose (above the line) and variants with low noise in glucose (below the line). *(continues on the next page)*

Figure 4.6: *(continues from the previous page)* The two arrows highlight the position 7 bp upstream and 6 bp downstream of the *lacZ* gene start codon. Polymorphisms in these positions are associated with high (A and C) or low noise (T at both positions). Nucleotide colors: red = A, green = T, yellow = G, blue = C, grey = matches the consensus (top sequence with annotations). **b)** Correlation between transcriptional noise in glucose and transcriptional noise in the other environments for segregating promoter variants. The black lines show linear regression of the data points. The rho and p-values were calculated using Spearman's correlation test. **c)** Maximum likelihood phylogenetic tree constructed from the core genomes of all 135 environmental *E. coli* isolates and 56 known laboratory, type or clinical isolates of *Escherichia* species using REALPHY (Bertels et al. 2014). The colors in strain names indicate standard phylogenetic clades of the genus *Escherichia*, as indicated by the adjacent names. The bar next to the tree represents the mutational combination the strains possess at the positions highlighted by arrows in **a**: blue = TT genotype; yellow = TC genotype; red = AC genotype; black = no *lacZ* promoter found (all of these are *Shigella*, *E. fergusonii*, *E. albertii* species or laboratory *E. coli* strains with deletion of the *lac* operon). The horizontal grey lines touching the bar serve as visual aids for identifying the precise taxa.

plasticity, perhaps due to genetic background having strong effects on plasticity. Finally, we have found evidence of diversifying selection on transcriptional noise in glucose, but directional selection in lactose. However, these *cis*-regulatory phenotypic effects can be strongly dependent on genetic background (*trans*-regulatory effects). To gain insight into the effects of genetic background on the regulatory behaviour of the *lac* operon, we transformed three of the segregating *lacZ* promoter variants into the isolates that each had originated from, i.e. their native genetic background (**Table S4.1**). We also include the MG1655 promoter variant in its native MG1655 background in the analyses below.

In both glucose and galactose, we found that all variants in their native genetic background maintained the same rank order of fluorescence levels as in the non-native MG1655 background, except for a single case in galactose (**Figure 4.7a**, SC418). However, the range of fluorescence levels increased considerably, with promoter variants varying more than 20-fold in fluorescence levels in glucose. This was primarily due to an almost three-fold increase in transcription for the SC312 and SC418 isolates. In strong contrast, all promoter variants converged in fluorescence levels in lactose. In the non-native genetic background, promoter variants differed by approximately three-fold; in their native genetic background, this decreased to less than two-fold, with three segregating promoter variants converging to almost exactly the same level (**Figure 4.7a**). This indicates that *trans*-effects on transcriptional activity can be divergent, increasing activity in some environments (SC312 in glucose) but decreasing it in others (SC312 in lactose. These changes will necessarily affect plasticity as well.

When we considered transcriptional plasticity, we also found that in all combinations of environments the majority of promoter variants maintained the same rank order in the native genetic background as in MG1655 strain, except for SC400 in galactose and lactose (**Figure 4.7b**, Gal:Lac). The maintenance of the rank order implies that the transcriptional plasticity of promoter variants follows the same pattern in both genetic backgrounds, and that from a relative standpoint, the conclusions we have derived from segregating promoter variants in the MG1655 genetic background are indicative of how the variants behave in nature. However, we also observed substantial decreases in transcriptional plasticity for SC312 and SC418 promoter variants in their native genetic backgrounds (**Figure 4.7b**).

This is due almost entirely to the considerable increase in transcription that occurred in glucose for these two isolates. As previously (**Figure 4.4b** and **Figure 4.4d**) the plasticity in all three environments (**Figure 4.7b**, All) primarily reflected the patterns of plasticity in glucose and lactose.

Finally, the segregating variants in their native genetic backgrounds exhibited much lower transcriptional noise in lactose as compared to random variants. This is emphasized by the fact that while there appears to be selection for low levels of noise in this environment (**Figure 4.5b**, Lactose), the MG1655 variant, even when in its native background, exhibited

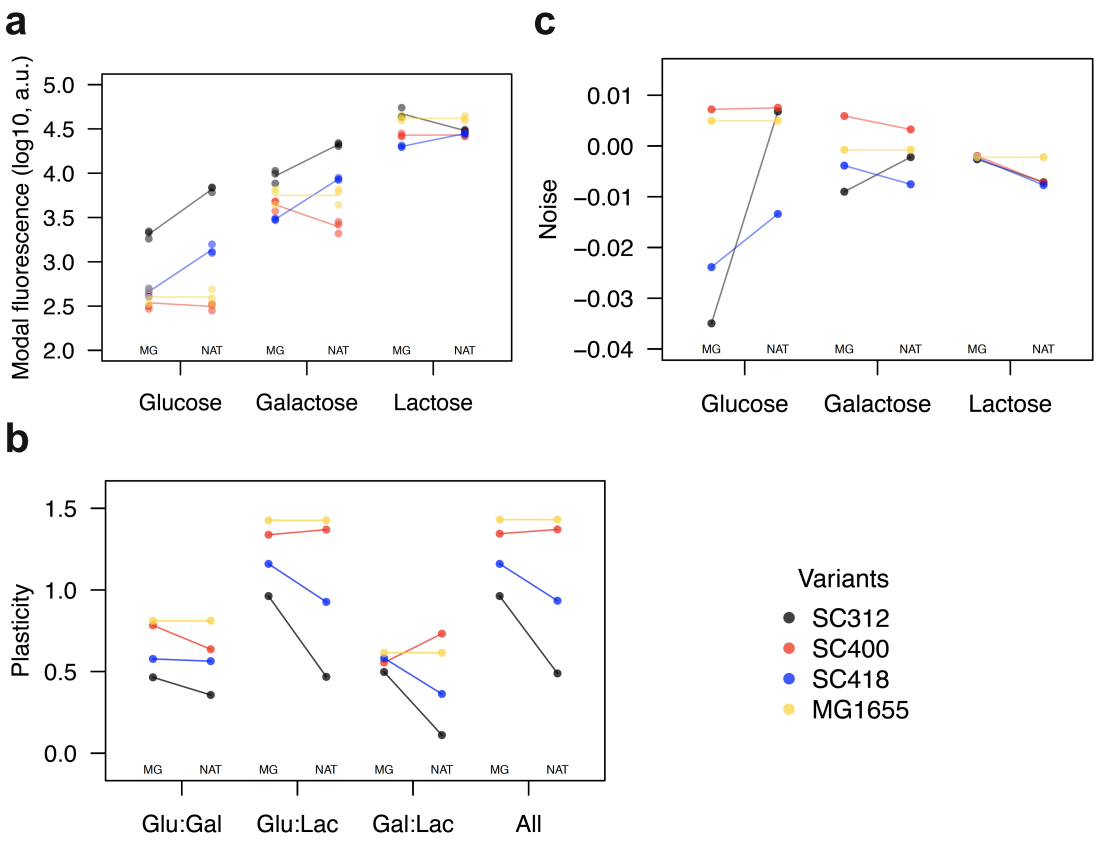

Figure 4.7: **Comparison of regulatory phenotypes between segregating promoter variants in MG1655 and native genetic backgrounds. a)** Modal population fluorescence in all three environments as indicated by x-axis labels. Each datapoint represents modal population fluorescence from one out of a total of three full biological replicates. The lines connect mean modal population fluorescence values calculated from each of the three replicates. **b)** Plasticity in all pairs of environments (Glu:Gal, Glu:Lac, and Gal:Lac) and in combination of all three environments (All). **c)** Transcriptional noise in all three environments as indicated by x-axis labels. Identical promoter variants are compared when being placed either to MG1655 genetic background (MG) or the genetic background from which each promoter variant originates (NAT). The colors distinguish promoter variants and they are connected with the same color lines within each environment or combination of. The MG1655 variant is always present in the MG1655 genetic background that serves as non-native genetic background for other variants. All the values compared between MG and NAT are thus identical.

higher levels of noise than other segregating variants. Surprisingly, in glucose, we found a very large increase in transcriptional noise of the SC312 promoter variant, which had exhibited the least noise in the MG1655 background, and which has a low-noise TT genotype (**Figure 4.6a**, **Figure 4.6c**, and **Figure 4.7c**). Interestingly, SC312 is one of six isolates in the A and B1 clades that have an apparently horizontally transferred *lac* locus (**Figure 4.6c**). These results suggest that *trans*-effects (i.e. genetic background) can have strong and opposing effects to the *cis*-effects of the promoter itself.

4.6 Discussion

Here we have investigated the actions of selection on transcriptional regulation of the canonical *lac* operon. We utilised an experimental system with a self-cleaving ribozyme RiboJ (**Figure 4.1** and **Figure 4.2**) that allows quantification of transcriptional effects on phenotypes while excluding translational effects (Lou et al. 2012, Vlková et al. 2021). This system also has increased sensitivity for measuring changes in transcriptional regulation (**Figure S4.3**). To ensure that this system also reflects the patterns of regulatory behavior in the native *lac* locus, we replaced the native *lacZ* promoter in MG1655 with promoter variants from the natural isolates. In this experimental system, the promoter sequence can have both transcriptional and translational effects on protein expression level. Nevertheless, these results confirmed that in almost all cases, the relative transcriptional activity of each promoter in the plasmid-based system reflected the relative expression levels in the chromosome, with the rank order of fluorescence level and plasticity remaining the same in all environments except for one case, in which the promoter from isolate SC312 exhibited decreased relative fluorescence in lactose (**Figure S4.4**). This may be due to low levels of LacI being bound more frequently to the single copy of the *lacZ* promoter on the chromosome, and repressing transcription, whereas on the plasmid, the LacI molecules are titrated out due to the slightly higher copy number of the promoter. However, since in the chromosomal system we also measure changes in translation it might be possible that the SC312 variant has significantly decreased translation rate as compared to the other segregating variants. We were not able to quantify noise levels for the chromosomal copies of the promoter variants, primarily because a large sample size is required to accurately calculate the noise metric. In addition, promoter variants in the chromosome exhibit fluorescence levels that are close to the limit of detection, decreasing sensitivity.

In previous work, we showed that mutations with large *cis*-effects on expression phenotypes (expression level, plasticity, and noise) are often filtered out from the *E. coli* population by both directional and stabilizing selection (Vlková and Silander 2021). However, for some regulatory phenotypes, we previously did not find evidence of selection, suggesting that changes in these phenotypes had only weak effects on fitness. While the work here is consistent with our previous results, the increased sensitivity of the assays on transcriptional phenotypes, together with additional analyses, have allowed us to discern other selective forces.

First, we have shown that there has been selection to minimise transcriptional activity in glucose and maximise activity in lactose, with random variants generally exhibiting higher transcription in glucose or lower transcription in lactose (**Figure 4.3a**). Related to this, the majority of random promoter variants exhibited decreased transcriptional plasticity

compared to their MG1655 progenitor (**Figure 4.4**). The decreased plasticity of random variants suggests directional selection for high transcriptional plasticity in MG1655. Even so, we found no clear differences between the plasticity of segregating and random *lacZ* promoter variants. However, the segregating variants are far more diverged from MG1655 than the random variants are, still supporting the hypothesis that new mutations that decrease plasticity are filtered out by selection.

Second, using similar comparisons of segregating and random promoter variants, we have shown that selection has acted to decrease transcriptional noise in the segregating *lacZ* promoter variants in lactose. This contrasts with previous results in which we found no significant evidence of selection on noise for the *lacZ* promoter, although there was a trend for segregating variants to exhibit lower noise (Vlková and Silander 2021). The previous lack of clear evidence of selection was most likely due to a lower sample size of random variants and a lack of sensitivity in glucose.

More surprisingly, we found evidence that in glucose, segregating variants constituted two phenotypic clusters, one having high noise and the other low noise (**Figure 4.5b**, Glucose). We also identified two specific polymorphisms near the start codon of the *lacZ* ORF that were associated with the two noise phenotypes (**Figure 4.6a**). In addition, these two polymorphisms are ancient, having been segregating in *E. coli* since the divergence of phylogroups A and B1, in contrast to every other polymorphism segregating at the *lac* locus (**Figure 4.6a** and **Figure 4.6c**). However, we could not establish that these two polymorphisms were causal for the noise phenotypes, nor that their long-term maintenance is due specifically to selection on noise. We did find a single random mutant that shared one of these segregating polymorphisms, and the noise phenotype of this random mutant changed considerably from the progenitor MG1655 in the direction expected if this polymorphism is causal. It has previously been shown that only a small number of genetic changes can affect noise phenotypes (Hornung et al. 2012, Metzger et al. 2015, Schmiedel et al. 2019, Urchueguía et al. 2019, Vlková and Silander 2021, Wolf et al. 2015).

Finally, we confirmed that differences in transcriptional noise measured in MG1655 strain are qualitatively similar to those observed in the native genetic backgrounds of the isolates from which the segregating variants originate, with the relative levels of noise remaining the same in most cases (**Figure 4.7c**). Nevertheless, we found one clear exception, the segregating *lacZ* promoter variant from the SC312 isolate. Despite this promoter variant exhibiting the lowest noise in the MG1655 genetic background, and having the low-noise TT genotype (at positions -7 bp and +6 bp relative to *lacZ* gene start codon) in glucose, it manifested with a high-noise phenotype when present in its original genetic background (**Figure 4.7c**, SC312). Interestingly, the phylogenetic analysis suggested horizontal gene transfer (HGT) of the *lac* locus into SC312 from an isolate in the B2, D, F, or E clades. Thus, one reason for the change in the noise level of the *lac* locus in the SC312 isolate may be that there are *trans*-effects that mediate noise as well as *cis*-effects. Because of the HGT of this locus, there might be suboptimal transcriptional control of the *lac* operon in this isolate. We also observed unusually high transcriptional activity in glucose and low plasticity for the SC312 variant compared to other segregating variants (**Figure 4.7a** and **Figure 4.7b**). We thus propose that a relatively recent HGT event has resulted in suboptimal regulation of the *lacZ* promoter. However, this suboptimal regulation may be mitigated by this isolate having evolved increased noise at the *lac* locus. Theoretical models have predicted that high

noise can be beneficial when the precise expression control and plasticity is not optimal (Schmiedel et al. 2019, Schmutzer and Wagner 2020, Wolf et al. 2015). This has also been experimentally tested in *S. cerevisiae*, in which high-noise variants of the TDH3 promoter resulted in on average higher fitness as compared to low noise variants with suboptimal expression levels (Duveau et al. 2018). As we observed unusually high transcriptional activity in glucose (**Figure 4.7a**) with low plasticity (**Figure 4.7b**), the high transcriptional noise might be a way of mitigating the possible misregulation in glucose. Interestingly, this high noise mitigation is mediated by *trans*- rather than *cis*-genetic changes in the SC312 isolate.

4.7 Conclusion

The data here show that for the *lacZ* promoter, natural selection acts on expression phenotypes at least in part solely through adjusting transcriptional control. We found evidence of diversifying selection acting on transcriptional noise in glucose, and two single nucleotide polymorphisms that are associated with this phenotype, but not necessarily causal. Furthermore, we found that one segregating promoter variant exhibited atypical regulatory responses both inside and outside of the genetic background of its origin. This may be a result of recent horizontal gene transfer, resulting in suboptimal regulatory phenotype. In consensus with theoretical predictions and recent experimental evidence showing that high noise can be advantageous when expression levels are suboptimal, we observed an increase in transcriptional noise in this variant. Interestingly, this increase in noise seems to be environment-specific and results from *trans*- rather than *cis*-genetic changes.

4.8 Supplementary Information

4.8.1 Supplementary Note

Here I provide the synthetized sequence of *lacZ* promoter and gene, which is translationally ligated to GFPmut2 (Twist Biosciences). The sequence is flanked by I-SceI restriction sites as well.

TTCCCCGAAAAGTGCCACCTTAGGGATAACAGGGTAATAATCAGCTGTTGCCCGTCTCACTGGTGAAAAGAAAAACCACC
CTGGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCCGACT
GGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTT
CCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATGATTACGGATTCAC
TGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCTTTC
GCCAGCTGGCGTAATAGCGAAGAGGCCCGCACCGATCGCCCTTCCCAACAGTTGCGCAGCCTGAATGGCGAATGGCGCTT
TGCCTGGTTTCCGGCACCAGAAGCGGTGCCGGAAAGCTGGCTGGAGTGCGATCTTCCTGAGGCCGATACTGTCGTCGTCC
CCTCAAACTGGCAGATGCACGGTTACGATGCGCCCATCTACACCAACGTGACCTATCCCATTACGGTCAATCCGCCGTTT
GTTCCCACGGAGAATCCGACGGGTTGTTACTCGCTCACATTTAATGTTGATGAAAGCTGGCTACAGGAAGGCCAGACGCG
AATTATTTTTGATGGCGTTAACTCGGCGTTTCATCTGTGGTGCAACGGGCGCTGGGTCGGTTACGGCCAGGACAGTCGTT
TGCCGTCTGAATTTGACCTGAGCGCATTTTTACGCGCCGGAGAAAACCGCCTCGCGGTGATGGTGCTGCGCTGGAGTGAC
GGCAGTTATCTGGAAGATCAGGATATGTGGCGGATGAGCGGCATTTTCCGTGACGTCTCGTTGCTGCATAAACCGACTAC
ACAAATCAGCGATTTCCATGTTGCCACTCGCTTTAATGATGATTTCAGCCGCGCTGTACTGGAGGCTGAAGTTCAGATGT
GCGGCGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTATGGCAGGGTGAAACGCAGGTCGCCAGCGGCACCGCGCCT
TTCGGCGGTGAAATTATCGATGAGCGTGGTGGTTATGCCGATCGCGTCACACTACGTCTGAACGTCGAAAACCCGAAACT
GTGGAGCGCCGAAATCCCGAATCTCTATCGTGCGGTGGTTGAACTGCACACCGCCGACGGCACGCTGATTGAAGCAGAAG
CCTGCGATGTCGGTTTCCGCGAGGTGCGGATTGAAAATGGTCTGCTGCTGCTGAACGGCAAGCCGTTGCTGATTCGAGGC
GTTAACCGTCACGAGCATCATCCTCTGCATGGTCAGGTCATGGATGAGCAGACGATGGTGCAGGATATCCTGCTGATGAA
GCAGAACAACTTTAACGCCGTGCGCTGTTCGCATTATCCGAACCATCCGCTGTGGTACACGCTGTGCGACCGCTACGGCC
TGTATGTGGTGGATGAAGCCAATATTGAAACCCACGGCATGGTGCCAATGAATCGTCTGACCGATGATCCGCGCTGGCTA
CCGGCGATGAGCGAACGCGTAACGCGAATGGTGCAGCGCGATCGTAATCACCCGAGTGTGATCATCTGGTCGCTGGGGAA
TGAATCAGGCCACGGCGCTAATCACGACGCGCTGTATCGCTGGATCAAATCTGTCGATCCTTCCCGCCCGGTGCAGTATG
AAGGCGGCGGAGCCGACACCACGGCCACCGATATTATTTGCCCGATGTACGCGCGCGTGGATGAAGACCAGCCCTTCCCG
GCTGTGCCGAAATGGTCCATCAAAAAATGGCTTTCGCTACCTGGAGAGACGCGCCCGCTGATCCTTTGCGAATACGCCCA
CGCGATGGGTAACAGTCTTGGCGGTTTCGCTAAATACTGGCAGGCGTTTCGTCAGTATCCCCGTTTACAGGGCGGCTTCG
TCTGGGACTGGGTGGATCAGTCGCTGATTAAATATGATGAAAACGGCAACCCGTGGTCGGCTTACGGCGGTGATTTTGGC
GATACGCCGAACGATCGCCAGTTCTGTATGAACGGTCTGGTCTTTGCCGACCGCACGCCGCATCCAGCGCTGACGGAAGC
AAAACACCAGCAGCAGTTTTTCCAGTTCCGTTTATCCGGGCAAACCATCGAAGTGACCAGCGAATACCTGTTCCGTCATA
GCGATAACGAGCTCCTGCACTGGATGGTGGCGCTGGATGGTAAGCCGCTGGCAAGCGGTGAAGTGCCTCTGGATGTCGCT
CCACAAGGTAAACAGTTGATTGAACTGCCTGAACTACCGCAGCCGGAGAGCGCCGGGCAACTCTGGCTCACAGTACGCGT
AGTGCAACCGAACGCGACCGCATGGTCAGAAGCCGGGCACATCAGCGCCTGGCAGCAGTGGCGTCTGGCGGAAAACCTCA
GTGTGACGCTCCCCGCCGCGTCCCACGCCATCCCGCATCTGACCACCAGCGAAATGGATTTTTGCATCGAGCTGGGTAAT
AAGCGTTGGCAATTTAACCGCCAGTCAGGCTTTCTTTCACAGATGTGGATTGGCGATAAAAAACAACTGCTGACGCCGCT
GCGCGATCAGTTCACCCGTGCACCGCTGGATAACGACATTGGCGTAAGTGAAGCGACCCGCATTGACCCTAACGCCTGGG
TCGAACGCTGGAAGGCGGCGGGCCATTACCAGGCCGAAGCAGCGTTGTTGCAGTGCACGGCAGATACACTTGCTGATGCG
GTGCTGATTACGACCGCTCACGCGTGGCAGCATCAGGGGAAAACCTTATTTATCAGCCGGAAAACCTACCGGATTGATGG
TAGTGGTCAAATGGCGATTACCGTTGATGTTGAAGTGGCGAGCGATACACCGCATCCGGCGCGGATTGGCCTGAACTGCC
AGCTGGCGCAGGTAGCAGAGCGGGTAAACTGGCTCGGATTAGGGCCGCAAGAAAACTATCCCGACCGCCTTACTGCCGCC
TGTTTTGACCGCTGGGATCTGCCATTGTCAGACATGTATACCCCGTACGTCTTCCCGAGCGAAAACGGTCTGCGCTGCGG
GACGCGCGAATTGAATTATGGCCCACACCAGTGGCGCGGCGACTTCCAGTTCAACATCAGCCGCTACAGTCAACAGCAAC
TGATGGAAACCAGCCATCGCCATCTGCTGCACGCGGAAGAAGGCACATGGCTGAATATCGACGGTTTCCATATGGGGATT
GGTGGCGACGACTCCTGGAGCCCGTCAGTATCGGCGGAATTCCAGCTGAGCGCCGGTCGCTACCATTACCAGTTGGTCTG
GTGTCAAAAAGGATCCGGCGGAGGCGGAATGAGTAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAAT
TAGATGGTGATGTTAATGGGCACAAATTTTCTGTCAGTGGAGAGGGTGAAGGTGATGCAACATACGGAAAACTTACCCTT
AAATTTATTTGCACTACTGGAAAACTACCTGTTCCATGGCCAACACTTGTCACTACTTTCGCGTATGGTCTTCAATGCTT
TGCGAGATACCCAGATCATATGAAACAGCATGACTTTTTCAAGAGTGCCATGCCCGAAGGTTATGTACAGGAAAGAACTA

TATTTTTCAAAGATGACGGGAACTACAAGACACGTGCTGAAGTCAAGTTTGAAGGTGATACCCTTGTTAATAGAATCGAG
TTAAAAGGTATTGATTTTAAAGAAGATGGAAACATTCTTGGACACAAATTGGAATACAACTATAACTCACACAATGTATA
CATCATGGCAGACAAACAAAAGAATGGAATCAAAGTTAACTTCAAAATTAGACACAACATTGAAGATGGAAGCGTTCAAC
TAGCAGACCATTATCAACAAAATACTCCAATTGGCGATGGCCCTGTCCTTTTACCAGACAACCATTACCTGTCCACACAA
TCTGCCCTTTCGAAAGATCCCAACGAAAAGAGAGACCACATGGTCCTTCTTGAGTTTGTAACAGCTGCTGGGATTACCCA
TGGTATGGATGAATTGTACAAATAATAATAACCGGGCAGGCCATGTCTGCCCGTATTTCGCGTAAGGAAATCCATTAGGG
ATAACAGGGTAATGCAGGATGCTGCTGGCTACCCT

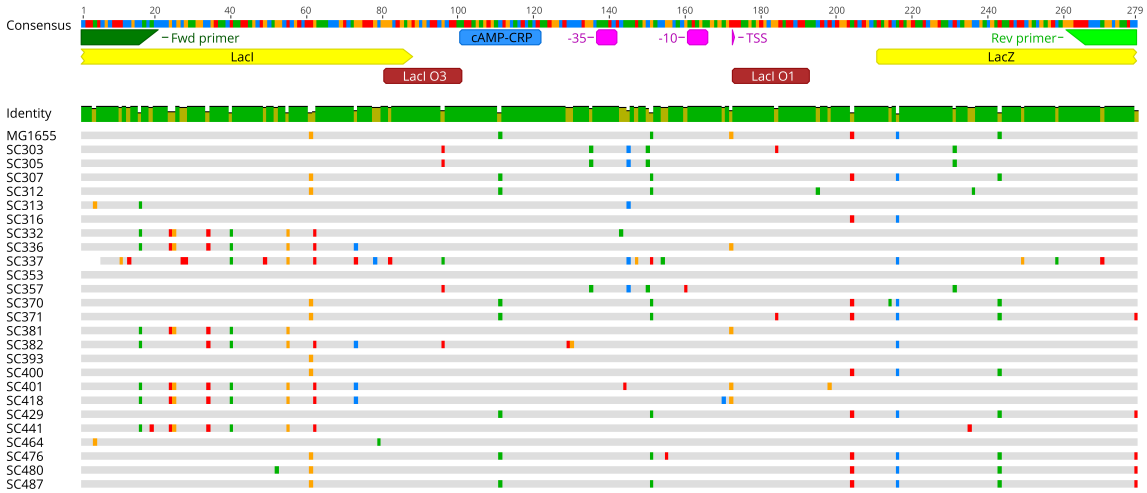

Figure S4.1: **Sequence alignment of all segregating promoter variants found among 135 environmental *E. coli* isolates and MG1655 laboratory strain.** Nucleotide colors: red = A, green = T, yellow = G, blue = C, grey = matches the consensus (top sequence with annotations). The green primer annotations indicate the primers used for PCR amplification and assembly of the variants with the pMV001 vector (**Figure 4.1d**).

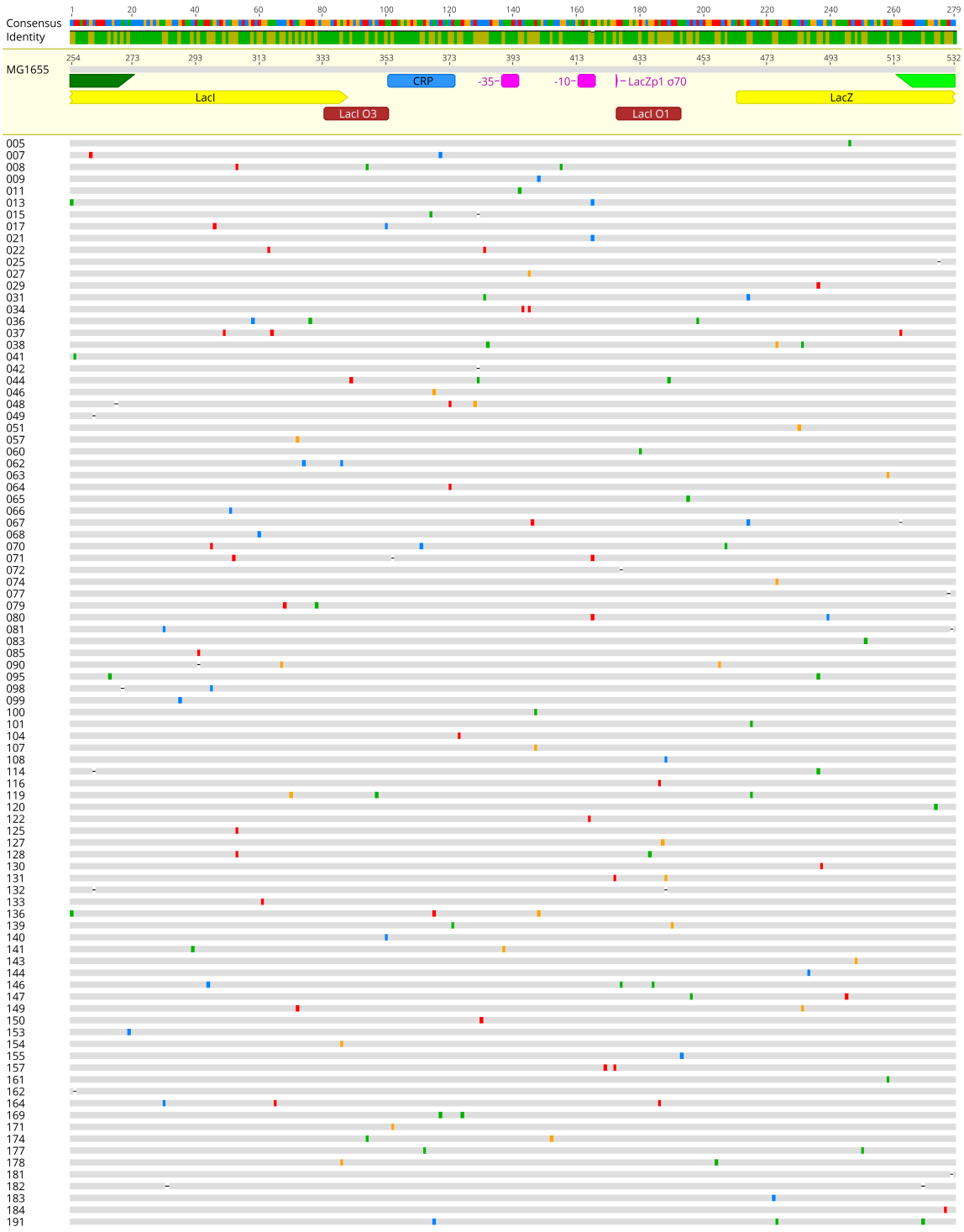

Figure S4.2: **Sequence alignment of random promoter variants.** Nucleotide colors: red = A, green = T, yellow = G, blue = C, grey = same as the MG1655 reference variant from which the random variants are derived (the sequence with annotations). The dashes in the sequences indicate indels.

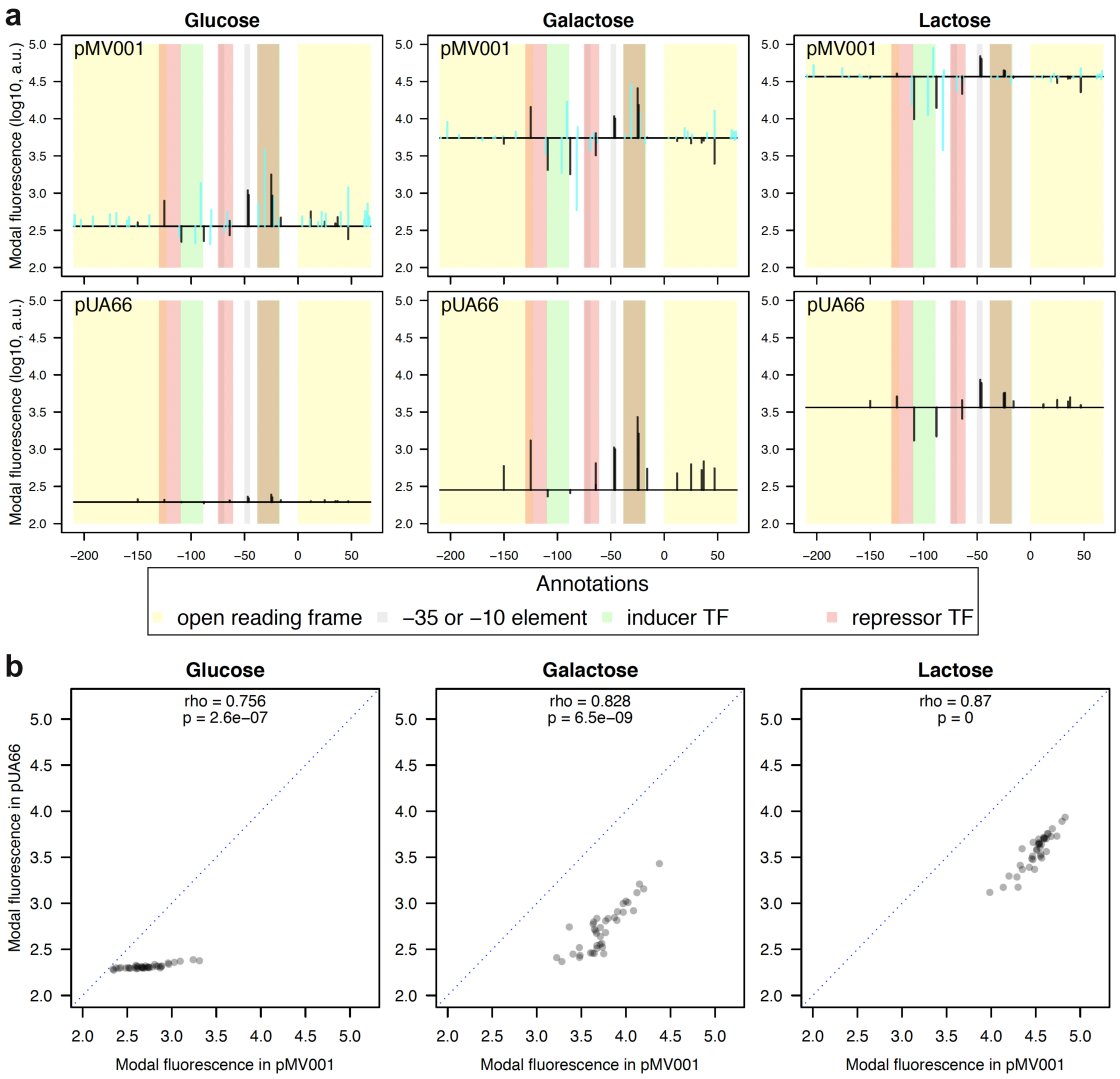

Figure S4.3: **The changes in fluorescence due to random mutations are correlated between the two plasmid systems. a)** Changes in fluorescence from the MG1655 promoter variant due to single random mutations for all environments and both plasmid systems (pUA66 and pMV001). The x-axes indicate the position relative to the *lacZ* gene start codon. The horizontal black line indicates the fluorescence level of the MG1655 *lacZ* promoter variant in that environment and with that plasmid system. The vertical lines show the direction and size of the change in fluorescence level when the MG1655 promoter variant is mutated at that specific position. The black vertical lines show mutation changes that are present in both plasmid systems, while the cyan vertical lines are mutations specific just for the dataset with RiboJ. The brown stripes result from overlapping repressor and activator TF binding sites. **b)** Comparison of modal fluorescence from all variants which are shared by both plasmid systems in MG1655 genetic background. The modal fluorescence levels are strongly correlated in all environments. Each data point in **a** and **b** is a result of three full biological replicates (**Materials and Methods**). The blue dotted lines indicate identical modal fluorescence in both plasmid systems (x = y). The rho and p-values were calculated using Spearman's correlation test.

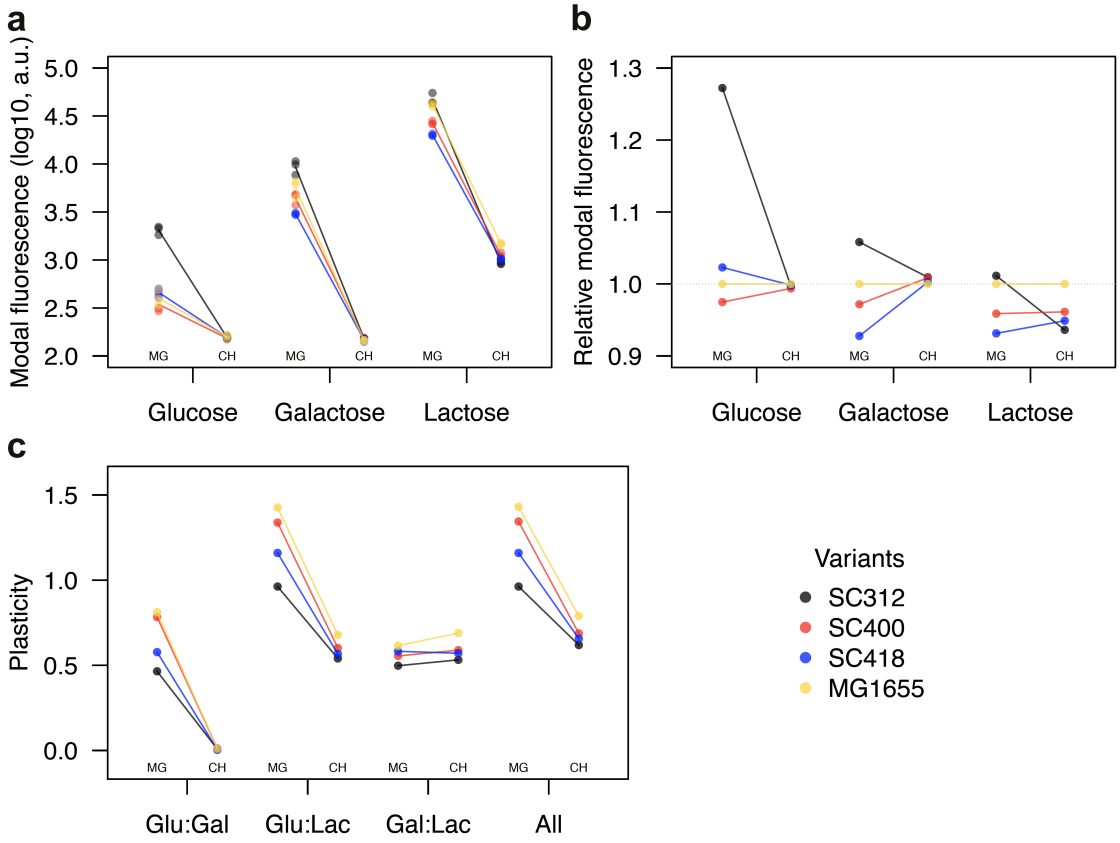

Figure S4.4: **Comparison of regulatory phenotypes between segregating promoter variants in MG1655 on a plasmid and in the chromosome. a)** Modal population fluorescence in all three environments as indicated by x-axis labels. Each datapoint represents modal population fluorescence from one out of a total of three full biological replicates. The lines connect mean modal population fluorescence values calculated from each of the three replicates. **b)** Modal population fluorescence relative to the MG1655 variant in all three environments as indicated by x-axis labels. **c)** Plasticity in all pairs of environments (Glu:Gal, Glu:Lac, and Gal:Lac) and in combination of all three environments (All). Identical promoter variants are compared when being placed either to MG1655 genetic background on a plasmid (MG) or into the lac operon on chromosome (CH). The colors distinguish promoter variants and they are connected with the same color lines within each environment or combination of.

Table S4.1: **List of segregating *lacZ* promoter variants**

Variant	Polymorphisms relative to MG1655 variant	Present in N of isolates out of 135
MG1655*	0	0
SC307	1	17
SC370	2	2
SC371	3	1
SC400*	3	7
SC429	3	4
SC476	3	1
SC316	5	1
SC480	5	1
SC312*	6	2
SC393	6	2
SC353	7	1
SC305	12	12
SC381	12	2
SC303	13	1
SC357	13	1
SC332	15	2
SC418*	15	8
SC487†	2	29
SC464†	9	1
SC313†	10	1
SC336†	14	22
SC382†	15	1
SC401†	16	1
SC441†	16	6
SC337†	19	9

Note: The variant names correspond to representative isolates in which the variants were identified. The same variant naming is used throughout the manuscript. Bolded variants were cloned into both pMV001 vector and the pUA66 vector (lacking RiboJ), and used for comparison of fluorescence in the **Figure S4.3**.

* These variants were also cloned into the MG1655 chromosome and were assayed both in MG1655 and in the isolate of their origin.

† Segregating variants that were not cloned.

Table S4.2: **List of random *lacZ* promoter variants**

Variant	Mutations relative to the MG1655 variant
005	1
009	1
011	1
021	1
025	1
027	1
029	1
041	1
042	1
046	1
049	1
051	1
057	1
060	1
063	1
064	1
065	1
066	1
068	1
072	1
074	1
077	1
083	1
085	1
099	1
100	1
101	1
104	1
107	1
108	1
116	1
120	1
122	1
125	1
127	1
130	1
133	1
140	1
143	1

Continued on the next page

Table S4.2 List of random *lacZ* promoter variants – continuation

Variant	Mutations relative to the MG1655 variant
144	1
150	1
153	1
154	1
155	1
161	1
162	1
171	1
181	1
183	1
184	1
007	2
013	2
015	2
017	2
022	2
031	2
034	2
062	2
079	2
080	2
081	2
095	2
098	2
114	2
128	2
131	2
132	2
139	2
141	2
147	2
149	2
157	2
169	2
174	2
177	2
178	2
182	2
008	3
036	3

Continued on the next page

Table S4.2 List of random *lacZ* promoter variants – continuation

Variant	Mutations relative to the MG1655 variant
037	3
038	3
044	3
048	3
067	3
070	3
071	3
090	3
119	3
136	3
146	3
164	3
191	3

Note: The same variant naming is used throughout the manuscript. The variants in bold were also cloned into the pUA66 vector and used for comparison of fluorescence in the **Figure S4.3**.

Table S4.3: **Primers and oligos used in this work**

Primer and oligo ID	Sequence	Purpose
pUA66_insert_F3965	5' - TTG TCT GTT GTG CCC AGT CAT AGC - 3'	PCR & Sanger sequencing
pUA66_insert_R232	5' - TCG CAA AGC ATT GAA GAC CAT ACG C - 3'	PCR & Sanger sequencing
pMV001_FastClonV_F	5' - AGC TGT CAC CGG ATG TGC - 3'	PCR of pMV001 vector for DNA assembly
pMV001_FastClonV_R	5' - TCG AGG TGA AGA CGA AAG GGC - 3'	PCR of pMV001 vector for DNA assembly
lacZ_FastClonIN_F	5' - TTT CGT CTT CAC CTC GAC AAT ACG CAA ACC GCC TCT CC - 3'	PCR of *lacZ* promoter for DNA assembly into pMV001 vector
lacZ_FastClonIN-D9_R	5' - CAC ATC CGG TGA CAG CTG TGT AAC GCC AGG GTT TTC C - 3'	PCR of *lacZ* promoter for DNA assembly into pMV001 vector (69A SNP)
lacZ_FastClonIN-K12_R	5' - CAC ATC CGG TGA CAG CTG GGT AAC GCC AGG GTT TTC C - 3'	PCR of *lacZ* promoter for DNA assembly into pMV001 vector (69C SNP)
pMV002_FastClonV_F	5' - CGT CGT GAC TGG GAA AAC C - 3'	PCR of pMV002 vector for DNA assembly
pMV002_FastClonV_R	5' - TTT TCA CCA GTG AGA CGG GCA ACA GCT GAT TAT TAC CCT GTT ATC CCT AAG GTG GC - 3'	PCR of pMV002 vector for DNA assembly & oligo bridge for pMV002 DNA assembly
placZ_DP_F	5' - GGT TTT CCC AGT CAC GAC G - 3'	PCR of *lacZ* promoter for DNA assembly into pMV002 vector
placZ_DP-K12_R	5' - CGT CTC ACT GGT GAA AAG AAA AAC CAC CCT GGC GCC CAA TAC GCA AAC CGC CTC TC - 3'	PCR of *lacZ* promoter for DNA assembly into pMV002 vector (1011C SNP)
placZ_DP-B6_R	5' - CGT CTC ACT GGT GAA AAG AAA AAC CAC CCT GGC GCC CAA TAC GCA AAC CGC TTC TCC - 3'	PCR of *lacZ* promoter for DNA assembly into pMV002 vector (1011T SNP)
lacZ_LP_F	5' - AAT CAG CTG TTG CCC GTC TCA CTG GTG AAA AGA AAA ACC ACC CTG GCG CCT ACG GCC CCA AGG TCC AAA	PCR of Landing pad for electroporation

Continued on the next page

Table S4.3 Primers and oligos used in this work – continuation

Primer and oligo ID	Sequence	Purpose
	CGG TGA - 3'	
lacZ_LP_R	5' - ATG GAT TTC CTT ACG CGA AAT ACG GGC AGA CAT GGC CTG CCC GGT TAT TAT TGG CTT CAG GGA TGA GGC GCC ATC - 3'	PCR of Landing pad for electroporation
pTKDP_vector_F	5' - CAT GTC TGC CCG TAT TTC GCG TAA GGA AAT CCA TTA GGG ATA ACA GGG TAA TGC AGG ATG -3'	oligo bridge for pMV002 DNA assembly
pTKDP_backbone_F	5' - CAC ATT TCC CCG AAA AGT GCC ACC TTA GGG ATA ACA GGG TAA TAA TCA -3'	oligo bridge for pMV002 DNA assembly
pTKDP_backbone_R	5' - GGG TAG CCA GCA GCA TCC TGC ATT ACC CTG TTA TCC CTA ATG G - 3'	oligo bridge for pMV002 DNA assembly

Note: There are two versions of the reverse primer sets due to a SNP presence in the priming area of the *lacZ* gene (C69A) and *lacI* gene (C1011T).

Chapter 5

Concluding remarks

In this thesis I have investigated how natural selection shapes gene regulation in bacteria, a topic which has been under-studied until this point. While multiple studies have focused on the role of natural selection on gene regulation before, many of them used only eukaryotic systems which have a different promoter structure than prokaryotes. At the same time majority of the studies published so far measured mainly expression noise and used a single experimental condition when investigating selection acting on gene regulation (Barroso et al. 2017, Duveau et al. 2018, Hodgins-Davis et al. 2019, Hornung et al. 2012, Metzger et al. 2015, Schmiedel et al. 2019, Schor et al. 2017, Silander et al. 2012, Wolf et al. 2015). This prevents detecting effects of natural selection on plasticity and determining whether selection on noise differs among environments. One exception is work that has been done on quantifying plasticity of the TDH3 promoter in yeast (Duveau et al. 2017). In this thesis we aimed to expand the knowledge about selection forces shaping gene regulation in bacteria using a collection of environmental *E. coli* isolates (Ishii et al. 2006).

In the **Chapter 2** we first examined the genetic variability of promoters in the *E. coli* population. We showed that there is more than an order of magnitude difference in the genetic diversity of promoters present in the *E. coli* population. Based on these results and known expression phenotypes or known function of downstream genes we selected ten promoters for further study. In each of these ten promoters we focused on understanding how selection affects three expression phenotypes: expression level, plasticity, and noise, using three environments specific for each promoter. In order to discern selection acting on those phenotypes we compared expression from promoter variants segregating in the *E. coli* population which have been subject to selection, and random promoter variants, which have never been subject to selection (Denver et al. 2005, Metzger et al. 2015). We found a strong correlation between genetic variation of a promoter and phenotypic variation in expression levels driven from its segregating variants. We also found that the effect size and direction of changes in expression level caused by random mutations can depend on the environment in which it is assayed. This highlights the importance of the use of multiple experimental environments in phenotypic assays to fully understand selection acting on gene regulation. The natural selection in general has filtered mutations that cause large changes in expression level. We further reported frequent selection towards high plasticity and low noise. Nevertheless, selection did not appear consistent across all promoters. Each promoter had its own selection pattern and not in all cases we detected significant difference

between random and segregating promoters. However, we generally observed that despite segregating variants having had greater genetic variability compared to random variants, they were phenotypically more similar. In some cases the sample size of segregating variants was very low, reducing the statistical power and making it difficult to quantify differences in phenotypic variability. By further considering the selection acting on all three phenotypes together we clearly showed that segregating variants of the promoters are enriched for polymorphisms that have minimal effect on the overall expression phenotype. The **Chapter 2** thus expands the knowledge about common selection patterns in bacterial gene regulation.

Since, transcription is the first crucial step for gene expression to begin, we wanted to examine selection acting on transcriptional regulation alone. Transcriptomics is most commonly used to measure transcriptional activity, however even with the recent advances in single cell RNA sequencing it still lacks the high-throughput of flow cytometry (Creecy and Conway 2015, Güell et al. 2011, Luecken and Theis 2019). We thus decided to test whether the self-cleaving ribozyme RiboJ (Lou et al. 2012) can be utilized to decouple transcription from translation in a way that would allow detection of changes in transcription only. In the **Chapter 3** we showed that the efficiency of self-cleaving RiboJ activity is over 95% in bacterial cells growing exponentially in rich LB media. We also confirmed that *cis*- and *trans*-genetic context have a minimal impact on RiboJ activity and thus changes in fluorescence measured between promoter variants are a consequence of changes in transcriptional activity alone. The **Chapter 3** thus demonstrates a new application for RiboJ in gene expression studies and it also increases sensitivity to changes in transcription due to increased translation.

The **Chapter 4** combines the methodology from **Chapter 2** with results from **Chapter 3** in order to examine selection acting on transcriptional control alone. For this purpose we picked the *lacZ* promoter due to its well understood mechanism of transcriptional regulation. We also applied additional analysis, such as stratifying the effects of random variants by number of mutations acquired through random mutagenesis. The **Chapter 4** describes directional selection acting to minimize transcriptional activity of MG1655 *lacZ* promoter variant in glucose, while maximizing in lactose. This is coupled with the selection towards high transcriptional plasticity. Stratifying random variants by number of mutations allowed us to detect selection acting to filter out mutations that cause big changes in the transcriptional activity from the segregating variants even when we detected no significant differences in direct comparison of transcriptional activity variation between segregating and random variants. This is thus consistent with the results of overall expression (transcription and translation) presented in **Chapter 2** for *lacZ* promoter. The detection of selection towards low transcriptional noise in lactose by simple comparison of segregating and random variants is however in contrast with the results from **Chapter 2**. It is likely that this discrepancy is caused by the lower sample size used for *lacZ* random promoter variant dataset in **Chapter 2**.

Interestingly we also detect diversifying selection acting on transcriptional noise in glucose together with two genotypes associated with it. Similar to the results in lactose, the sample size might have played a role in not detecting this selection on noise in glucose when testing overall expression as described in **Chapter 2**, although the lack of sensitivity in the expression system compared to the system with RiboJ is more likely the cause.

We also found two genotypes associated with either high or low transcriptional noise. These genotypes differed two SNPs near the *lacZ* gene start codon and they segregated almost perfectly in the *E. coli* population. The exceptions to this are probably due to horizontal gene transfer (HGT) within the genomic region of the *lac* operon, namely the SC312 segregating variant. Next to observing unusually high transcriptional activity in glucose for SC312 as compared to other segregating variants, we also detected a great increase in transcriptional noise in its native genetic background as compared to MG1655 strain. We thus propose that a recent HGT event within the *lac* operon of SC312 isolate has caused suboptimal regulation of transcriptional activity when lactose is absent in the environment. The negative impact of suboptimal regulation might have been mitigated by the observed increase in transcriptional noise. This is consistent with theoretical and experimental studies showing that high noise can be beneficial when precise regulation is suboptimal (Duveau et al. 2018, Schmiedel et al. 2019, Schmutzer and Wagner 2020, Wolf et al. 2015). Surprisingly, the high noise in SC312 isolate comes from changes in genetic background rather than from changes in *lacZ* promoter sequence.

In conclusion, the results in this thesis constitute of new insights into how natural selection shapes gene regulation in bacterium *E. coli*. Particularly, we showed that selection on high expression plasticity and low expression noise is relatively common among regulated promoters. We have further identified directional and diversifying selection acting on transcriptional control alone in *lacZ* promoter. Together with this we also present a new application of previously described genetic tool, the ribozyme RiboJ, which we used to detect changes in transcription alone. This work thus increases the understanding of how selection shapes bacterial gene regulation in nature.

References

Acar, M., Mettetal, J. T., and van Oudenaarden, A. (2008). Stochastic switching as a survival strategy in fluctuating environments. *Nat Genet*, 40(4):471–475.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–410.

Arold, S. T., Leonard, P. G., Parkinson, G. N., and Ladbury, J. E. (2010). H-NS forms a superhelical protein scaffold for DNA condensation. *Proceedings of the National Academy of Sciences*, 107(36):15728–15732.

Azam, T. A., Iwata, A., Nishimura, A., Ueda, S., and Ishihama, A. (1999). Growth phase-dependent variation in protein composition of the *Escherichia coli* nucleoid. *Journal of Bacteriology*, 181(20):6361–6370.

Aznar, S., Paytubi, S., and Juárez, A. (2013). The hha protein facilitates incorporation of horizontally acquired dna in enteric bacteria. *Microbiology*, 159(3):545–554.

Bailey, S. F., Alonso Morales, L. A., and Kassen, R. (2021). Effects of synonymous mutations beyond codon bias: The evidence for adaptive 4 synonymous substitutions from microbial evolution experiments. *Genome Biology and Evolution*, page evab141.

Ball, C. A., Osuna, R., Ferguson, K., and Johnson, R. (1992). Dramatic changes in Fis levels upon nutrient upshift in *Escherichia coli*. *Journal of Bacteriology*, 174(24):8043–8056.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., and Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. *Nat Genet*, 38(6):636–643.

Barbier, I., Perez-Carrasco, R., and Schaerli, Y. (2020). Controlling spatiotemporal pattern formation in a concentration gradient with a synthetic toggle switch. *Mol Syst Biol*, 16(6):e9361.

Barroso, G. V., Puzovic, N., and Dutheil, J. Y. (2017). The evolution of gene-specific transcriptional noise is driven by selection at the pathway level. *Genetics*, pages genetics–300467.

Bartoli, V., Meaker, G. A., di Bernardo, M., and Gorochowski, T. E. (2020). Tunable genetic devices through simultaneous control of transcription and translation. *Nat Commun*, 11(1):2095.

Basu, S., Mehreja, R., Thiberge, S., Chen, M.-T., and Weiss, R. (2004). Spatiotemporal control of gene expression with pulse-generating networks. *Proceedings of the National Academy of Sciences*, 101(17):6355–6360.

Becskei, A. and Serrano, L. (2000). Engineering stability in gene networks by autoregulation. *Nature*, 405(6786):590.

Belliveau, N. M., Barnes, S. L., Ireland, W. T., Jones, D. L., Sweredoski, M. J., Moradian, A., Hess, S., Kinney, J. B., and Phillips, R. (2018). Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proc Natl Acad Sci USA*, 115(21):E4796–E4805.

Bertels, F., Silander, O. K., Pachkov, M., Rainey, P. B., and van Nimwegen, E. (2014). Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads. *Molecular Biology and Evolution*, 31(5):1077–1088.

Bittihn, P., Didovyk, A., Tsimring, L. S., and Hasty, J. (2020). Genetically engineered control of phenotypic structure in microbial colonies. *Nat Microbiol*, 5(5):697–705.

Blainey, P. C., van Oijen, A. M., Banerjee, A., Verdine, G. L., and Xie, X. S. (2006). A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proceedings of the National Academy of Sciences*, 103(15):5752–5757.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., et al. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, 277(5331):1453–1462.

Brahms, J., Dargouge, O., Brahms, S., Ohara, Y., and Vagner, V. (1985). Activation and inhibition of transcription by supercoiling. *Journal of Molecular Biology*, 181(4):455–465.

Breckell, G. and Silander, O. K. (2020). Complete Genome Sequences of 47 Environmental Isolates of *Escherichia coli*. *Microbiol Resour Announc*, 9(38):e00222–20.

Brent, R. and Ptashne, M. (1981). Mechanism of action of the *lexA* gene product. *Proceedings of the National Academy of Sciences*, 78(7):4204–4208.

Brewster, R. C., Jones, D. L., and Phillips, R. (2012). Tuning Promoter Strength through RNA Polymerase Binding Site Design in *Escherichia coli*. *PLoS Computational Biology*, 8(12):e1002811.

Brewster, R. C., Weinert, F. M., Garcia, H. G., Song, D., Rydenfelt, M., and Phillips, R. (2014). The Transcription Factor Titration Effect Dictates Level of Gene Expression. *Cell*, 156(6):1312–1323.

Browning, D. F. and Busby, S. J. (2016). Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*, 14(10):638–650.

Bryant, J. A., Sellars, L. E., Busby, S. J., and Lee, D. J. (2014). Chromosome position effects on gene expression in *Escherichia coli* K-12. *Nucleic Acids Research*, 42(18):11383–11392.

Byappanahalli, M. and Fujioka, R. (2004). Indigenous soil bacteria and low moisture may limit but allow faecal bacteria to multiply and become a minor population in tropical soils. *Water Science and Technology*, 50(1):27–32.

Caldara, M., Charlier, D., and Cunin, R. (2006). The arginine regulon of *Escherichia coli*: whole-system transcriptome analysis discovers new genes and provides an integrated view of arginine regulation. *Microbiology*, 152(11):3343–3354.

Carrier, T. A. and Keasling, J. D. (1997). Engineering mRNA Stability in *E. coli* by the Addition of Synthetic Hairpins Using a 5' Cassette System. *BIOTECHNOLOGY AND BIOENGINEERING*, 55(3):4.

Casadesús, J. and Low, D. A. (2013). Programmed heterogeneity: epigenetic mechanisms in bacteria. *Journal of Biological Chemistry*, 288(20):13929–13935.

Chambonnier, G., Roux, L., Redelberger, D., Fadel, F., Filloux, A., Sivaneson, M., De Bentzmann, S., and Bordi, C. (2016). The hybrid histidine kinase LadS forms a multicomponent signal transduction system with the GacS/GacA two-component system in *Pseudomonas aeruginosa*. *PLoS Genetics*, 12(5):e1006032.

Charlier, D. and Glansdorff, N. (2004). Biosynthesis of arginine and polyamines. *EcoSal Plus*, 1(1).

Charlier, D., Roovers, M., Van Vliet, F., Boyen, A., Cunin, R., Nakamura, Y., Glansdorff, N., and Piérard, A. (1992). Arginine regulon of *Escherichia coli* k-12: a study of repressor-operator interactions and of i*n vitro* binding affinities *versus in vivo* repression. *Journal of Molecular Biology*, 226(2):367–386.

Choi, H.-S., Kim, K., Park, J. W., Jung, Y. H., and Lee, Y. (2005). Effects of FIS protein on *rnpB* transcription in *Escherichia coli*. *Molecules and Cells*, 19(2):239–245.

Chong, S., Chen, C., Ge, H., and Xie, X. S. (2014). Mechanism of transcriptional bursting in bacteria. *Cell*, 158(2):314–326.

Clark, D. P. (2005). *Molecular Biology*. Elsevier.

Clifton, K. P., Jones, E. M., Paudel, S., Marken, J. P., Monette, C. E., Halleran, A. D., Epp, L., and Saha, M. S. (2018). The genetic insulator RiboJ increases expression of insulated genes. *J Biol Eng*, 12(1):23.

Creecy, J. P. and Conway, T. (2015). Quantitative bacterial transcriptomics with RNA-seq. *Current opinion in microbiology*, 23:133–140.

Cui, C., Yang, C., Song, S., Fu, S., Sun, X., Yang, L., He, F., Zhang, L.-H., Zhang, Y., and Deng, Y. (2018). A novel two-component system modulates quorum sensing and pathogenicity in *Burkholderia cenocepacia*. *Molecular Microbiology*, 108(1):32–44.

Dame, R. T., Wyman, C., Wurm, R., Wagner, R., and Goosen, N. (2002). Structural basis for H-NS-mediated trapping of RNA polymerase in the open initiation complex at the *rrnB* P1. *Journal of Biological Chemistry*, 277(3):2146–2150.

Dar, D. and Sorek, R. (2018). Extensive reshaping of bacterial operons by programmed mRNA decay. *PLoS Genetics*, 14(4):e1007354.

Datsenko, K. A. and Wanner, B. L. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proceedings of the National Academy of Sciences*, 97(12):6640–6645.

de Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N., and Regev, A. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol*, 38(1):56–65.

Dekel, E. and Alon, U. (2005). Optimality and evolutionary tuning of the expression level of a protein. *Nature*, 436(7050):588–592.

Denver, D. R., Morris, K., Streelman, J. T., Kim, S. K., Lynch, M., and Thomas, W. K. (2005). The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nat Genet*, 37(5):544–548.

Desnoyers, G., Bouchard, M.-P., and Massé, E. (2013). New insights into small RNA-dependent translational regulation in prokaryotes. *Trends in Genetics*, 29(2):92–98.

Dillon, S. C. and Dorman, C. J. (2010). Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nature Reviews Microbiology*, 8(3):185.

Dixon, N. E. and Kornberg, A. (1984). Protein HU in the enzymatic replication of the chromosomal origin of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 81(2):424–428.

Duveau, F., Hodgins-Davis, A., Metzger, B. P., Yang, B., Tryban, S., Walker, E. A., Lybrook, T., and Wittkopp, P. J. (2018). Fitness effects of altering gene expression noise in *Saccharomyces cerevisiae*. *eLife*, 7:e37272.

Duveau, F., Yuan, D. C., Metzger, B. P., Hodgins-Davis, A., and Wittkopp, P. J. (2017). Effects of mutation and selection on plasticity of a promoter activity in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, 114(52):E11218–E11227.

Eisen, H., Brachet, P., Da Silva, L. P., and Jacob, F. (1970). Regulation of repressor expression in λ. *Proceedings of the National Academy of Sciences*, 66(3):855–862.

El Qaidi, S., Allemand, F., Oberto, J., and Plumbridge, J. (2009). Repression of *galP*, the galactose transporter in *Escherichia coli*, requires the specific regulator of N-acetylglucosamine metabolism. *Molecular Microbiology*, 71(1):146–157.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic Gene Expression in a Single Cell. *Science*, 297(5584):1183–1186.

Falconi, M., Brandi, A., La Teana, A., Gualerzi, C. O., and Pon, C. L. (1996). Antagonistic involvement of FIS and H-NS proteins in the transcriptional control of *hns* expression. *Molecular Microbiology*, 19(5):965–975.

Falconi, M., Prosseda, G., Giangrossi, M., Beghetto, E., and Colonna, B. (2001). Involvement of FIS in the H-NS-mediated regulation of *virF* gene of *Shigella* and enteroinvasive *Escherichia coli*. *Molecular Microbiology*, 42(2):439–452.

Frederix, M., Edwards, A., McAnulla, C., and Downie, J. A. (2011). Co-ordination of quorum-sensing regulation in *Rhizobium leguminosarum* by induction of an anti-repressor. *Molecular Microbiology*, 81(4):994–1007.

Frumkin, I., Schirman, D., Rotman, A., Li, F., Zahavi, L., Mordret, E., Asraf, O., Wu, S., Levy, S. F., and Pilpel, Y. (2017). Gene Architectures that Minimize Cost of Gene Expression. *Molecular Cell*, 65(1):142–153.

Gao, R. and Stock, A. M. (2015). Temporal hierarchy of gene expression mediated by transcription factor binding affinity and activation dynamics. *MBio*, 6(3):e00686–15.

Gellert, M., Mizuuchi, K., O'Dea, M. H., and Nash, H. A. (1976). DNA gyrase: an enzyme that introduces superhelical turns into DNA. *Proceedings of the National Academy of Sciences*, 73(11):3872–3876.

Gertz, J., Siggia, E. D., and Cohen, B. A. (2009). Analysis of combinatorial *cis*-regulation in synthetic and genomic promoters. *Nature*, 457(7226):215–218.

Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., and Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, 6(5):343–345.

Gosink, K., Ross, W., Leirmo, S., Osuna, R., Finkel, S., Johnson, R., and Gourse, R. (1993). DNA binding and bending are necessary but not sufficient for Fis-dependent activation of *rrnB* P1. *Journal of Bacteriology*, 175(6):1580–1589.

Govers, S. K., Adam, A., Blockeel, H., and Aertsen, A. (2017). Rapid phenotypic individualization of bacterial sister cells. *Scientific Reports*, 7(1):8473.

Grainger, D. C., Goldberg, M. D., Lee, D. J., and Busby, S. J. (2008). Selective repression by Fis and H-NS at the *Escherichia coli dps* promoter. *Molecular Microbiology*, 68(6):1366–1377.

Gresham, D., Desai, M. M., Tucker, C. M., Jenq, H. T., Pai, D. A., Ward, A., DeSevo, C. G., Botstein, D., and Dunham, M. J. (2008). The Repertoire and Dynamics of Evolutionary Adaptations to Controlled Nutrient-Limited Environments in Yeast. *PLoS Genet*, 4(12):e1000303.

Groisman, E. A. (2016). Feedback control of two-component regulatory systems. *Annual Review of Microbiology*, 70:103–124.

Güell, M., Yus, E., Lluch-Senar, M., and Serrano, L. (2011). Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nature Reviews Microbiology*, 9(9):658–669.

Harley, C. B. and Reynolds, R. P. (1987). Analysis of *E. coli* Promoter Sequences. *Nucleic Acids Research*, 15(5):2343–2361.

Hawkins, J. S., Silvis, M. R., Koo, B.-M., Peters, J. M., Osadnik, H., Jost, M., Hearne, C. C., Weissman, J. S., Todor, H., and Gross, C. A. (2020). Mismatch-CRISPRi Reveals the Co-varying Expression-Fitness Relationships of Essential Genes in *Escherichia coli* and *Bacillus subtilis*. *Cell Systems*, 11(5):523–535.

Heldwein, E. E. Z. and Brennan, R. G. (2001). Crystal structure of the transcription activator BmrR bound to DNA and a drug. *Nature*, 409(6818):378.

Hill, M. S., Vande Zande, P., and Wittkopp, P. J. (2021). Molecular and evolutionary processes generating variation in gene expression. *Nat Rev Genet*, 22(4):203–215.

Hirano, T., Beck, D. A., Wright, C. J., Demuth, D. R., Hackett, M., and Lamont, R. J. (2013). Regulon controlled by the GppX hybrid two component system in *Porphyromonas gingivalis*. *Molecular Oral Microbiology*, 28(1):70–81.

Hodgins-Davis, A., Duveau, F., Walker, E. A., and Wittkopp, P. J. (2019). Empirical measures of mutational effects define neutral models of regulatory evolution in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA*, 116(42):21085–21093.

Hornung, G., Bar-Ziv, R., Rosin, D., Tokuriki, N., Tawfik, D. S., Oren, M., and Barkai, N. (2012). Noise-mean relationship in mutated promoters. *Genome Research*, 22(12):2409–2417.

Hübner, P., Haffter, P., Iida, S., and Arber, W. (1989). Bent DNA is needed for recombinational enhancer activity in the site-specific recombination system cin of bacteriophage P1 the role of FIS protein. *Journal of Molecular Biology*, 205(3):493–500.

Hudson, J. M. and Fried, M. G. (1990). Co-operative interactions between the catabolite gene activator protein and the *lac* repressor at the lactose promoter. *Journal of Molecular Biology*, 214(2):381–396.

Ibarra, J. A., Villalba, M. I., and Puente, J. L. (2003). Identification of the DNA binding sites of PerA, the transcriptional activator of the *bfp* and *per* operons in enteropathogenic *Escherichia coli*. *Journal of Bacteriology*, 185(9):2835–2847.

Igarashi, K., Hanamura, A., Makino, K., Aiba, H., Mizuno, T., Nakata, A., and Ishihama, A. (1991). Functional map of the alpha subunit of *Escherichia coli* RNA polymerase: two modes of transcription activation by positive factors. *Proceedings of the National Academy of Sciences*, 88(20):8958–8962.

Ireland, W. T., Beeler, S. M., Flores-Bautista, E., McCarty, N. S., Röschinger, T., Belliveau, N. M., Sweredoski, M. J., Moradian, A., Kinney, J. B., and Phillips, R. (2020). Deciphering the regulatory genome of *Escherichia coli*, one hundred promoters at a time. *eLife*, 9:e55308.

Ishii, S., Ksoll, W. B., Hicks, R. E., and Sadowsky, M. J. (2006). Presence and growth of naturalized *Escherichia coli* in temperate soils from lake superior watersheds. *Applied and Environmental Microbiology*, 72(1):612–621.

Iyer, V. and Struhl, K. (1996). Absolute mRNA levels and transcriptional initiation rates in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, 93(11):5208–5212.

Jobe, A. and Bourgeois, S. (1972). *lac* repressor-operator interaction: Vi. the natural inducer of the *lac* operon. *Journal of Molecular Biology*, 69(3):397–408.

Jones, D. L., Brewster, R. C., and Phillips, R. (2014). Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346(6216):1533–1536.

Kaczmarczyk, A., Hochstrasser, R., Vorholt, J. A., and Francez-Charlot, A. (2014). Complex two-component signaling regulates the general stress response in *Alphaproteobacteria*. *Proceedings of the National Academy of Sciences*, 111(48):E5196–E5204.

Kalir, S., Mangan, S., and Alon, U. (2005). A coherent feed-forward loop with a SUM input function prolongs flagella expression in *Escherichia coli*. *Molecular Systems Biology*, 1(1).

Karambelkar, S., Swapna, G., and Nagaraja, V. (2012). Silencing of toxic gene expression by Fis. *Nucleic Acids Research*, 40(10):4358–4367.

Karp, P. D., Ong, W. K., Paley, S., Billington, R., Caspi, R., Fulcher, C., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P. E., Subhraveti, P., Gama-Castro, S., Muñiz-Rascado, L., Bonavides-Martinez, C., Santos-Zavaleta, A., Mackie, A., Collado-Vides, J., Keseler, I. M., and Paulsen, I. (2018). The EcoCyc Database. *EcoSal Plus*.

Kayoko, Y., Naotaka, H., Naoki, G., Kyoko, K., Fumio, I., and Yasunobu, K. (1992). Histone-like proteins are required for cell growth and constraint of supercoils in DNA. *Gene*, 122(1):9–15.

Kennell, D. and Riezman, H. (1977). Transcription and translation initiation frequencies of the *Escherichia coli lac* operon. *Journal of Molecular Biology*, 114(1):1–21.

Keren, L., Hausser, J., Lotan-Pompan, M., Vainberg Slutskin, I., Alisar, H., Kaminski, S., Weinberger, A., Alon, U., Milo, R., and Segal, E. (2016). Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. *Cell*, 166(5):1282–1294.e18.

Khademi, S. M. H., Sazinas, P., and Jelsbak, L. (2019). Within-Host Adaptation Mediated by Intergenic Evolution in *Pseudomonas aeruginosa*. *Genome Biology and Evolution*, 11(5):1385–1397.

Kinney, J. B. and McCandlish, D. M. (2019). Massively Parallel Assays and Quantitative Sequence–Function Relationships. *Annu. Rev. Genom. Hum. Genet.*, 20(1):99–127.

Kotte, O., Volkmer, B., Radzikowski, J. L., and Heinemann, M. (2014). Phenotypic bistability in *Escherichia coli*'s central carbon metabolism. *Molecular Systems Biology*, 10(7):736.

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

Kuhlman, T., Zhang, Z., Saier, M. H., and Hwa, T. (2007). Combinatorial transcriptional control of the lactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 104(14):6043–6048.

Kuo, J.-T., Chang, Y.-J., and Tseng, C.-P. (2003). Growth rate regulation of *lac* operon expression in *Escherichia coli* is cyclic AMP dependent. *FEBS Letters*, 553(3):397–402.

Levskaya, A., Chevalier, A. A., Tabor, J. J., Simpson, Z. B., Lavery, L. A., Levy, M., Davidson, E. A., Scouras, A., Ellington, A. D., Marcotte, E. M., and Voigt, C. A. (2005). Engineering *Escherichia coli* to see light. *Nature*, 438(7067):441–442.

Lewis, K. (2007). Persister cells, dormancy and infectious disease. *Nat Rev Microbiol*, 5(1):48–56.

Li, C., Wen, A., Shen, B., Lu, J., Huang, Y., and Chang, Y. (2011). FastCloning: a highly simplified, purification-free, sequence- and ligation-independent PCR cloning method. *BMC Biotechnology*, 11(1):92.

Lloyd, G. S., Niu, W., Tebbutt, J., Ebright, R. H., and Busby, S. J. (2002). Requirement for two copies of RNA polymerase α subunit C-terminal domain for synergistic transcription activation at complex bacterial promoters. *Genes & Development*, 16(19):2557–2565.

López-Maury, L., Marguerat, S., and Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat Rev Genet*, 9(8):583–593.

Lou, C., Stanton, B., Chen, Y.-J., Munsky, B., and Voigt, C. A. (2012). Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nature Biotechnology*, 30(11):1137–1142.

Luecken, M. D. and Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746.

Lynch, J. B. and Sonnenburg, J. L. (2012). Prioritization of a plant polysaccharide over a mucus carbohydrate is enforced by a Bacteroides hybrid two-component system. *Molecular Microbiology*, 85(3):478–491.

Madan Babu, M., Teichmann, S. A., and Aravind, L. (2006). Evolutionary Dynamics of Prokaryotic Transcriptional Regulatory Networks. *Journal of Molecular Biology*, 358(2):614–633.

Maeda, Y. T. and Sano, M. (2006). Regulatory dynamics of synthetic gene networks with positive feedback. *Journal of Molecular Biology*, 359(4):1107–1124.

Mangan, S. and Alon, U. (2003). Structure and function of the feed-forward loop network motif. *Proceedings of the National Academy of Sciences*, 100(21):11980–11985.

Mangan, S., Itzkovitz, S., Zaslaver, A., and Alon, U. (2006). The incoherent feed-forward loop accelerates the response-time of the *gal* system of *Escherichia coli*. *Journal of Molecular Biology*, 356(5):1073–1081.

Mangan, S., Zaslaver, A., and Alon, U. (2003). The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *Journal of Molecular Biology*, 334(2):197–204.

Mayo, A. E., Setty, Y., Shavit, S., Zaslaver, A., and Alon, U. (2006). Plasticity of the *cis*-Regulatory Input Function of a Gene. *PLoS Biol*, 4(4):e45.

Mears, P. J., Koirala, S., Rao, C. V., Golding, I., and Chemla, Y. R. (2014). *Escherichia coli* swimming is robust against variations in flagellar number. *eLife*, 3:e01916.

Metzger, B. P., Yuan, D. C., Gruber, J. D., Duveau, F., and Wittkopp, P. J. (2015). Selection on noise constrains variation in a eukaryotic promoter. *Nature*, 521(7552):344.

Metzger, B. P. H. and Wittkopp, P. J. (2019). Compensatory *trans*-regulatory alleles minimizing variation in TDH3 expression are common within *Saccharomyces cerevisiae*. *Evolution Letters*.

Mika, F. and Hengge, R. (2005). A two-component phosphotransfer network involving ArcB, ArcA, and RssB coordinates synthesis and proteolysis of σ^S (RpoS) in *E. coli*. *Genes & Development*, 19(22):2770–2781.

Monod, J. (1949). THE GROWTH OF BACTERIAL CULTURES. *Annu. Rev. Microbiol.*, 3(1):371–394.

Monsalve, M., Mencía, M., Salas, M., and Rojo, F. (1996). Protein p4 represses phage ϕ29 A2c promoter by interacting with the alpha subunit of *Bacillus subtilis* RNA polymerase. *Proceedings of the National Academy of Sciences*, 93(17):8913–8918.

Mustaev, A., Roberts, J., and Gottesman, M. (2017). Transcription elongation. *Transcription*, 8(3):150–161.

Na, D., Yoo, S. M., Chung, H., Park, H., Park, J. H., and Lee, S. Y. (2013). Metabolic engineering of *Escherichia coli* using synthetic small regulatory RNAs. *Nat Biotechnol*, 31(2):170–174.

Navarre, W. W., Porwollik, S., Wang, Y., McClelland, M., Rosen, H., Libby, S. J., and Fang, F. C. (2006). Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*. *Science*, 313(5784):236–238.

Naville, M. and Gautheret, D. (2009). Transcription attenuation in bacteria: theme and variations. *Briefings in Functional Genomics and Proteomics*, 8(6):482–492.

Neubauer, Z. and Calef, E. (1970). Immunity phase-shift in defective lysogens: Non-mutational hereditary change of early regulation of λ prophage. *Journal of Molecular Biology*, 51(1):1–13.

Neves, D., Vos, S., Blank, L. M., and Ebert, B. E. (2020). *Pseudomonas* mRNA 2.0: Boosting Gene Expression Through Enhanced mRNA Stability and Translational Efficiency. *Front. Bioeng. Biotechnol.*, 7:458.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217.

Novick, A. and Weiner, M. (1957). Enzyme induction as an all-or-none phenomenon. *Proceedings of the National Academy of Sciences*, 43(7):553–566.

Ouafa, Z.-A., Reverchon, S., Lautier, T., Muskhelishvili, G., and Nasser, W. (2012). The nucleoid-associated proteins H-NS and FIS modulate the DNA supercoiling response of the *pel* genes, the major virulence factors in the plant pathogen bacterium *Dickeya dadantii*. *Nucleic Acids Research*, 40(10):4306–4319.

Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I., and van Oudenaarden, A. (2004). Multistability in the lactose utilization network of *Escherichia coli*. *Nature*, 427(6979):4.

Perego, M. and Brannigan, J. A. (2001). Pentapeptide regulation of aspartyl-phosphate phosphatases. *Peptides*, 22(10):1541–1547.

Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research*, 29(9):45e–45.

Pittard, J. (1996). The various strategies within the tyrr regulon of *Escherichia coli* to modulate gene expression. *Genes to Cells*, 1(8):717–725.

Richard, M. and Yvert, G. (2014). How does evolution tune biological noise? *Frontiers in Genetics*, 5:374.

Ronin, I., Katsowich, N., Rosenshine, I., and Balaban, N. Q. (2017). A long-term epigenetic memory switch controls bacterial virulence bimodality. *eLife*, 6.

Rosenfeld, N., Elowitz, M. B., and Alon, U. (2002). Negative autoregulation speeds the response times of transcription networks. *Journal of Molecular Biology*, 323(5):785–793.

Rossi, N. A., El Meouche, I., and Dunlop, M. J. (2019). Forecasting cell fate during antibiotic exposure using stochastic gene expression. *Commun Biol*, 2(1):259.

Sakoparnig, T., Field, C., and van Nimwegen, E. (2021). Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *eLife*, 10:e65366.

Santana, F. J., Calva, E., Puente, J. L., et al. (2001). Transcriptional regulation of type III secretion genes in enteropathogenic *Escherichia coli*: Ler antagonizes H-NS-dependent repression. *Molecular Microbiology*, 39(3):664–678.

Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., García-Sotelo, J. S., Alquicira-Hernández, K., Muñiz-Rascado, L. J., Peña-Loredo, P., Ishida-Gutiérrez, C., Velázquez-Ramírez, D. A., Del Moral-Chávez, V., Bonavides-Martínez, C., Méndez-Cruz, C.-F., Galagan, J., and Collado-Vides, J. (2019). RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic Acids Research*, 47(D1):D212–D220.

Sayut, D. J., Kambam, P. K. R., and Sun, L. (2007). Noise and kinetics of LuxR positive feedback loops. *Biochemical and biophysical research communications*, 363(3):667–673.

Schaerli, Y., Jiménez, A., Duarte, J. M., Mihajlovic, L., Renggli, J., Isalan, M., Sharpe, J., and Wagner, A. (2018). Synthetic circuits reveal how mechanisms of gene regulatory networks constrain evolution. *Molecular Systems Biology*, 14(9):e8102.

Schmiedel, J. M., Carey, L. B., and Lehner, B. (2019). Empirical mean-noise fitness landscapes reveal the fitness impact of gene expression noise. *Nat Commun*, 10(1):3180.

Schmutzer, M. and Wagner, A. (2020). Gene expression noise can promote the fixation of beneficial mutations in fluctuating environments. *PLoS Comput Biol*, 16(10):e1007727.

Schor, I. E., Degner, J. F., Harnett, D., Cannavò, E., Casale, F. P., Shim, H., Garfield, D. A., Birney, E., Stephens, M., Stegle, O., et al. (2017). Promoter shape varies across populations and affects promoter evolution and expression noise. *Nature Genetics*, 49(4):550.

Semsey, S., Tolstorukov, M. Y., Virnik, K., Zhurkin, V. B., and Adhya, S. (2004). DNA trajectory in the Gal repressosome. *Genes & Development*, 18(15):1898–1907.

Serres, M. H. and Riley, M. (2000). MultiFun, a Multifunctional Classification Scheme for *Escherichia coli* K-12 Gene Products. *Microbial & Comparative Genomics*, 5(4):205–222.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, 31(1):64.

Shimizu, T. S., Tu, Y., and Berg, H. C. (2010). A modular gradient-sensing network for chemotaxis in *Escherichia coli* revealed by responses to time-varying stimuli. *Molecular Systems Biology*, 6(1):382.

Shu, C. J. and Zhulin, I. B. (2002). ANTAR: an RNA-binding domain in transcription antitermination regulatory proteins. *Trends in Biochemical Sciences*, 27(1):3–5.

Silander, O. K., Nikolic, N., Zaslaver, A., Bren, A., Kikoin, I., Alon, U., and Ackermann, M. (2012). A genome-wide analysis of promoter-mediated phenotypic noise in *Escherichia coli*. *PLoS Genetics*, 8(1):e1002443.

Singh, S. S. and Grainger, D. C. (2013). H-NS can facilitate specific DNA-binding by RNA polymerase in AT-rich gene regulatory regions. *PLoS Genetics*, 9(6):e1003589.

Smits, W. K., Kuipers, O. P., and Veening, J.-W. (2006). Phenotypic variation in bacteria: the role of feedback regulation. *Nature Reviews Microbiology*, 4(4):259.

Søgaard-Andersen, L. and Valentin-Hansen, P. (1993). Protein-protein interactions in gene regulation: the cAMP-CRP complex sets the specificity of a second DNA-binding protein, the CytR repressor. *Cell*, 75(3):557–566.

Somorin, Y., Abram, F., Brennan, F., and O'Byrne, C. (2016). The general stress response is conserved in long-term soil-persistent strains of *Escherichia coli*. *Applied and Environmental Microbiology*, 82(15):4628–4640.

Stella, S., Cascio, D., and Johnson, R. C. (2010). The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. *Genes & Development*, 24(8):814–826.

Süel, G. M., Kulkarni, R. P., Dworkin, J., Garcia-Ojalvo, J., and Elowitz, M. B. (2007). Tunability and Noise Dependence in Differentiation Dynamics. *Science*, 315(5819):1716–1719.

Svenningsen, S. L., Costantino, N., Adhya, S., et al. (2005). On the role of Cro in λ prophage induction. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4465–4469.

Taniguchi, Y., Choi, P. J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X. S. (2010). Quantifying *E. coli* Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science*, 329(5991):533–538.

Tas, H., Nguyen, C. T., Patel, R., Kim, N. H., and Kuhlman, T. E. (2015). An Integrated System for Precise Genome Modification in *Escherichia coli*. *PLoS ONE*, 10(9):e0136963.

Tenaillon, O., Barrick, J. E., Ribeck, N., Deatherage, D. E., Blanchard, J. L., Dasgupta, A., Wu, G. C., Wielgoss, S., Cruveiller, S., Médigue, C., Schneider, D., and Lenski, R. E. (2016). Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature*, 536(7615):165–170.

Toman, Z., Dambly-Chaudiere, C., Tenenbaum, L., and Radman, M. (1985). A system for detection of genetic and epigenetic alterations in *Escherichia coli* induced by DNA-damaging agents. *Journal of Molecular Biology*, 186(1):97–105.

Ueguchi, C., Kakeda, M., and Mizuno, T. (1993). Autoregulatory expression of the *Escherichia coli hns* gene encoding a nucleoid protein: H-NS functions as a repressor of its own transcription. *Molecular and General Genetics*, 236(2-3):171–178.

Urchueguía, A., Galbusera, L., Bellement, G., Julou, T., and Nimwegen, E. v. (2019). Noise propagation shapes condition-dependent gene expression noise in *Escherichia coli*. preprint, Systems Biology.

Urtecho, G., Tripp, A. D., Insigne, K. D., Kim, H., and Kosuri, S. (2019). Systematic Dissection of Sequence Elements Controlling Σ70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in *Escherichia coli*. *Biochemistry*, 58(11):1539–1551.

Ushida, C. and Aiba, H. (1990). Helical phase dependent action of CRP: effect of the distance between the CRP site and the- 35 region on promoter activity. *Nucleic Acids Research*, 18(21):6325–6330.

Veening, J.-W., Smits, W. K., and Kuipers, O. P. (2008). Bistability, epigenetics, and bet-hedging in bacteria. *Annu. Rev. Microbiol.*, 62:193–210.

Vlková, M., Morampalli, B. R., and Silander, O. K. (2021). Efficiency of the synthetic self-splicing RiboJ ribozyme is robust to *cis*- and *trans*-changes in genetic background. *MicrobiologyOpen*, 10(4).

Vlková, M. and Silander, O. K. (2021). Gene regulation is commonly selected for high plasticity and low noise. preprint, Evolutionary Biology.

Wang, J. C. (1971). Interaction between DNA and an *Escherichia coli* protein ω. *Journal of Molecular Biology*, 55(3):523–IN16.

Wanner, B. L., Kodaira, R., and Neidhardt, F. C. (1978). Regulation of *lac* Operon Expression: Reappraisal of the Theory of Catabolite Repression. *J Bacteriol*, 136(3):947–954.

Weickert, M. J. and Adhya, S. (1992). Isorepressor of the *gal* regulon in *Escherichia coli*. *Journal of Molecular Biology*, 226(1):69–83.

Wheatley, R. W., Lo, S., Jancewicz, L. J., Dugdale, M. L., and Huber, R. E. (2013). Structural explanation for allolactose (*lac* operon inducer) synthesis by *lacZ* β-galactosidase and the evolutionary relationship between allolactose synthesis and the *lac* repressor. *Journal of Biological Chemistry*, pages jbc–M113.

Whitaker, W. R., Lee, H., Arkin, A. P., and Dueber, J. E. (2015). Avoidance of Truncated Proteins from Unintended Ribosome Binding Sites within Heterologous Protein Coding Sequences. *ACS Synth. Biol.*, 4(3):249–257.

Wolf, L., Silander, O. K., and van Nimwegen, E. (2015). Expression noise facilitates the evolution of gene regulation. *eLife*, 4:e05856.

Wu, H.-Y., Shyy, S., Wang, J. C., and Liu, L. F. (1988). Transcription generates positively and negatively supercoiled domains in the template. *Cell*, 53(3):433–440.

Yona, A. H., Alm, E. J., and Gore, J. (2018). Random sequences rapidly evolve into de novo promoters. *Nat Commun*, 9(1):1530.

Yu, H., Wang, Z., Xu, H., Guo, J., Ma, Q., Mu, X., and Luo, Y. (2018). A method for Absolute Protein Expression Quantity Measurement Employing Insulator RiboJ. *Engineering*, 4(6):881–887.

Zaslaver, A., Bren, A., Ronen, M., Itzkovitz, S., Kikoin, I., Shavit, S., Liebermeister, W., Surette, M. G., and Alon, U. (2006). A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nature Methods*, 3(8):623–628.

Appendix

Received: 25 March 2021 | Revised: 4 August 2021 | Accepted: 4 August 2021

DOI: 10.1002/mbo3.1232

ORIGINAL ARTICLE

MicrobiologyOpen Open Access WILEY

Efficiency of the synthetic self-splicing RiboJ ribozyme is robust to *cis*- and *trans*-changes in genetic background

Markéta Vlková | Bhargava Reddy Morampalli | Olin K. Silander

School of Natural and Computational Sciences, Massey University, Auckland, New Zealand

Correspondence
Markéta Vlková and Olin K. Silander, School of Natural and Computational Sciences, Massey University, Auckland, New Zealand.
Emails: marketa.m.vlkova@gmail.com (M.V.); olinsilander@gmail.com (O.K.S.)

Funding information
Royal Society Te Apārangi, Marsden Fund, Grant/Award Number: MAU1703

Abstract

The expanding knowledge of the variety of synthetic genetic elements has enabled the construction of new and more efficient genetic circuits and yielded novel insights into molecular mechanisms. However, context dependence, in which interactions between *cis*- or *trans*-genetic elements affect the behavior of these elements, can reduce their general applicability or predictability. Genetic insulators, which mitigate unintended context-dependent *cis*-interactions, have been used to address this issue. One of the most commonly used genetic insulators is a self-splicing ribozyme called RiboJ, which can be used to decouple upstream 5′ UTR in mRNA from downstream sequences (e.g., open reading frames). Despite its general use as an insulator, there has been no systematic study quantifying the efficiency of RiboJ splicing or whether this autocatalytic activity is robust to *trans*- and *cis*-genetic context. Here, we determine the robustness of RiboJ splicing in the genetic context of six widely divergent *E. coli* strains. We also check for possible *cis*-effects by assessing two SNP versions close to the catalytic site of RiboJ. We show that mRNA molecules containing RiboJ are rapidly spliced even during rapid exponential growth and high levels of gene expression, with a mean efficiency of 98%. We also show that neither the *cis*- nor *trans*-genetic context has a significant impact on RiboJ activity, suggesting this element is robust to both *cis*- and *trans*-genetic changes.

KEYWORDS
insulation, RiboJ, ribozyme, RT-qPCR, splicing

1 | INTRODUCTION

Synthetic nucleic acid functional elements used to control protein output—such as promoters, ribosome binding sites, or terminators—are an indispensable part of engineered genetic circuits (Levskaya et al., 2005; Na et al., 2013; Neves et al., 2020) and are frequently used to study basic biological processes (Barbier et al., 2020; Bittihn et al., 2020). The value of such synthetic functional elements increases as their properties are better described and quantified—in some cases, careful quantification of the behavior of synthetic functional elements has led to fundamentally new insights into molecular mechanisms controlling protein output (Schmiedel et al., 2019; Urtecho et al., 2019).

One important type of synthetic functional elements that have been used to ensure predictable and robust protein output from mRNA are self-splicing ribozymes. These ribozymes can be used to splice mRNA at specific locations, for example, to remove the 5′ untranslated region (UTR). One of the most common ribozymes used to remove 5′ UTRs is RiboJ. By removing the 5′ UTR of the mRNA, RiboJ enables transcripts having different promoters and

thus different 5′ UTRs to produce identical mRNAs. This mitigates any effects that the 5′ UTR might have on mRNA folding or ribosome binding, keeping the translation initiation rate consistent (and predictable) even when promoters have different sequences (Neves et al., 2020; Urtecho et al., 2019; Yu et al., 2018). The utility of the RiboJ element was first demonstrated when it was used to ensure predictable expression in a synthetic NOT gate circuit, irrespective of the sequence of the promoter used to control the expression of a CI repressor in the system (Lou et al., 2012).

However, since the first use of RiboJ as a means of ensuring predictable expression, additional research has suggested there can also be unexpected effects of its use. Clifton et al. (2018) demonstrated that RiboJ insertion into the mRNA sequence led to an increase in protein expression and that the relative increase in expression depended on the strength of the promoter used. This effect was attributed to hairpin formation at the 5′ end of mRNA whose 5′ UTR had been removed by RiboJ, leading to higher stability and increased translation (Carrier & Keasling, 1997; Clifton et al., 2018; Neves et al., 2020). Another unexpected effect was observed when Bartoli et al. (2020) designed a tunable system to control translation initiation via binding of small regulatory RNA (sRNA). The complex secondary structure of mRNA molecules with RiboJ at the 5′ end appeared to interfere with the sRNA binding, decreasing the performance of the system. These results emphasize that unknown properties and behaviors of synthetic functional elements—here, RiboJ in particular—can lead to unexpected obstacles when creating new synthetic circuits.

We hypothesized that a complicating factor in the use of the RiboJ system would be the varying efficiency of RiboJ autocatalytic splicing activity between different bacterial strains, or due to polymorphisms near the RiboJ element. To our knowledge, there has been no systematic study quantifying the efficiency of the autocatalytic RiboJ splicing, or whether the efficiency of this autocatalytic activity depends on the genetic background of the organism in which it is used. To address these questions we first developed an assay to quantify RiboJ self-splicing efficiency. We then tested the robustness of the self-splicing activity to *cis*-genetic changes by assaying efficiency in two genetic contexts that differ by a single nucleotide polymorphism (SNP) close to the autocatalytic site of RiboJ. Finally, we tested the robustness of RiboJ behavior to *trans*-genetic changes by quantifying efficiency in six widely divergent strains of *E. coli*.

2 | EXPERIMENTAL PROCEDURES

2.1 | Bacterial strains

The genetic backgrounds of *E. coli* strains used in this study are listed in Table 1. The identity of all lab strains was confirmed using whole-genome sequencing. The whole genomes of strains SC312 and SC392 have been also sequenced (Breckell & Silander, 2020). Four different plasmids (Table 2) were transformed into each of the strains, providing 24 clones that we used to evaluate the efficiency of RiboJ splicing. The presence of the plasmids with correct inserts in all clones was confirmed by Sanger sequencing (Macrogen, South Korea).

2.2 | Plasmid construction

All plasmids used to measure the autocatalytic activity of RiboJ are listed in Table 2. Plasmids p69A.RJ− and p69C.RJ− were generously gifted by D. Blank, University of Basel. Plasmids p69A.RJ+ and p69C.RJ+ were constructed using plasmid pMV001 (which was created beforehand), as follows: RiboJ was ordered as four 60nt single-stranded oligos with each 30nt of them being homologous to either another 60nt RiboJ oligo or PCR amplified pUA66 vector (Table A1). These four oligos were then assembled with PCR amplified pUA66 vector using NEBuilder HiFi DNA assembly kit (New England Biolabs). The resulting pMV001 plasmid assembly mix was then used to electroporate Top10 *E. coli* cells (Invitrogen). The presence of the RiboJ was then confirmed by Sanger sequencing (Macrogen, South Korea) from colonies grown on selective LB agar plates with 50 μg/ml Kanamycin.

To create inserts for p69A.RJ+ and p69C.RJ+ plasmids the *lacZ* promoter regions from p69A.RJ− and p69C.RJ− were PCR amplified.

TABLE 1 Bacterial strains used in this study

Bacterial strains			
Strain	**Relevant characteristics**	**Phylogroup**	**Source or reference**
SC392	*A natural isolate of E. coli; Soil; 7/18/05; SC15-U2out14; St. Louis Clyde; Upshore (2m) outside the box*	B1	(Ishii et al., 2006)
SC312	*A natural isolate of E. coli; Water; 6/15/05; SC14-W8; St. Louis Clyde; Surface water*	B1	(Ishii et al., 2006)
MG1655	*F− λ− ilvG− rfb−50 rph−1*	A	(Blattner et al., 1997)
DH5α	*F− φ80lacZΔ M15 Δ (lacZYA-argF) U169 recA1 endA1 hsdR17 (rK− mK+) phoA supE44 λ− thi−1 gyrA96 relA1*	A	Invitrogen
BW25113	*F− DE(araD−araB)567 lacZ4787(del)::rrnB−3 LAM− rph−1 DE(rhaD-rhaB)568 hsdR514*	A	(Datsenko & Wanner, 2000)
BL21 Star (DE3)	*F-ompT hsdSB (rB−, mB−) galdcmrne131 (DE3)*	A	Invitrogen

TABLE 2 Plasmids used in this study

Plasmids		
Plasmid	**Relevant characteristics**	**Source**
p69A.RJ−	*lacZ* promoter 69A, without RiboJ	D. Blank, University of Basel
p69C.RJ−	*lacZ* promoter 69C, without RiboJ	D. Blank, University of Basel
p69A.RJ+	*lacZ* promoter 69A, with RiboJ	This study
p69C.RJ+	*lacZ* promoter 69C, with RiboJ	This study

Note: All plasmids carry Kan^R selection marker and were created using pUA66 backbone (Zaslaver et al., 2006).

The primers used contain 17nt overhangs that are homologous to PCR amplified pMV001 vector (Table A1). We ligated the vector with the inserts through Gibson assembly (Gibson et al., 2009) using the NEBuilder HiFi DNA assembly kit (New England Biolabs). All primers and oligos used including sequencing primers (Integrated DNA Technologies) are listed in Table A1. In all cases, the same method for insert and vector PCR amplification from existing plasmids was used as described by (Li et al., 2011).

2.3 | Flow cytometry

Strains for flow cytometry were grown in M9 minimal media (Sigma) supplemented with $MgSO_4$, $CaCl_2$, 0.4% (w/v) carbon source (glucose, galactose, or lactose), and 50 μg/ml Kanamycin. They were first inoculated from a glycerol stock library into a 96 well microplate using a pin replicator (Enzyscreen B.V.) and incubated at 37°C. After overnight incubation, the cultures were diluted into the same fresh media with the pin replicator and incubated the same way until they reached the mid-exponential phase (~4 h). At that point, the cells were diluted into 1× PBS with ~1% formaldehyde and kept on ice until measuring the GFP levels using the flow cytometer.

Cytometry was performed with a BD FACSCanto II and BD FACSDiva software version 6.1.3. The GFP fluorescence was measured using the 488 nm laser and a 513/17 nm bandpass filter. The data from FACSDiva were exported into Flow Cytometry Standard files, and cell gating and fluorescence analysis was performed using custom R scripts (flowCore package version 2.0.1; the scripts are available through https://doi.org/10.5281/zenodo.5154246). Cells were gated based on their maximal kernel density of forward and side scatter values, keeping about ⅓ of all events. The modal fluorescence was calculated from gated cells as the maximal kernel density from the fluorescence signal.

2.4 | RNA isolation

RNA was isolated from four clones a day, while clones with the same genetic background were processed together on the same day. We isolated RNA from MG1655 clones twice on two different days, all other clones were isolated just once. Each strain containing one of the four plasmids (Table 1 and Table 2) was grown from a single colony overnight in 3 ml of LB with 50 μg/ml Kanamycin and 2 mM IPTG (Isopropyl β-D-1-thiogalactopyranoside) with shaking (250 rpm) at 37°C. Because the high IPTG concentration impaired the growth of SC312 strain with RiboJ plasmids (i.e., p69A.RJ+ and p69C.RJ+), we grew all SC312 clones for RNA isolation in LB with 0.2 mM of IPTG instead. The next day 15 ml of the same fresh media in 50 ml Falcon tubes was inoculated by 15 μl of this overnight culture. This was incubated under the same conditions. Once the cultures reached an exponential phase (between 1.75 h and 2.5 h) it was placed on an ice slurry.

Next, we added 7.5 ml of ice-cold 5% phenol in ethanol to each 15 ml of culture and kept them on ice for 15 min. The cultures were then spun at 7000G for 7 min at 4°C, the supernatant was discarded and the pellet was redispersed in 350 μl of 3 mg/ml Lysozyme solution (in TE buffer). After incubating for 3 min, an equal volume of RNA lysis buffer was added and RNA isolated using Monarch Total RNA Miniprep Kit (New England Biolabs). Each sample was treated by DNase I twice: (1) on-column during the RNA extraction and then (2) in-tube after RNA extraction. This was done to avoid any amplification from residual gDNA during RT-qPCR. After the second treatment with DNase I the samples were column-purified and concentrated using RNA Clean & Concentrator-5 kit (Zymo Research). The quality of RNA in each sample was checked on 1% agarose gel and its concentration was measured on a Qubit 4 fluorometer (Invitrogen). The isolated RNA samples were then stored in a −80°C freezer.

2.5 | RT-qPCR

To assess the efficiency of PCR amplification by our primers we used RNA from MG1655 strain (all four plasmids). A ten-fold serial dilution was performed on all the RNA samples up to 10^{-4}. RT-qPCR was run on all the dilutions in triplicates using two different master mixes differing by the forward primer used—F1 and F2 (Figure 2 and Table A1). The total reaction volume was 20 μl with 2 μl of template RNA. We used SensiFAST Probe No-ROX One-Step Kit (Meridian Bioscience) and PikoReal Real-Time PCR System (Thermo Scientific) with following cycling conditions: Reverse transcription for 10 min at 45°C; Polymerase activation for 2 min at 95°C; 40 cycles of denaturation for 5 s at 95°C and Annealing & extension for 20 s at 55°C. The C_t values were obtained via PikoReal software version

2.2, exported into .xlsx file, and converted into .csv to be further analyzed using custom R scripts (available through https://doi.org/10.5281/zenodo.5154246).

To assess the autocatalytic efficiency of RiboJ, the RNA from all samples was first diluted from its original concentration (~2–3 μg/μl) to 20 pg/μl to obtain C_t values between 20 and 40 and to dilute out any potential residual of gDNA (to less than one molecule per reaction). We confirmed that no amplification occurred when omitting reverse transcriptase from the master mix. Each RNA sample was then run in three or more replicates using both primer sets (Figure 2) with the same conditions described above. We exported the data from the PikoReal software version 2.2 into .xlsx files, converted these into .csv, and performed all analyses using custom R scripts (available through https://doi.org/10.5281/zenodo.5154246). In brief, we determined the mean C_t value of all the replicates for the uncut and cut RiboJ transcripts and calculated the efficiency as the ratio of the cut and uncut transcripts using the Pfaffl method (Pfaffl, 2001):

$$\text{Eff} = 100 - 100 * \frac{E^{(a-b)}}{E^{(c-d)}}$$

where E is constant mean amplification efficiency (1.95766), a and b are mean C_t values of transcripts without and with RiboJ, respectively, using F1 primer, and c and d are mean C_t values of the same transcripts using F2 primer. To obtain a measure of the error in these estimates, we bootstrapped the data 10,000 times and recalculated the ratio for each bootstrap replicate.

2.6 | Plasmid sequencing

Plasmid DNA was isolated from overnight cultures using the StrataPrep Plasmid Miniprep Kit (Agilent), per the manufacturer's instructions. These were then prepared for Oxford Nanopore sequencing using the Oxford Nanopore rapid barcoding library prep, per the manufacturer's instructions, with a separate barcode used for each plasmid. These were run on a single MinION flowcell for 1 h and 50 min. The reads were base-called using the guppy_basecaller v5.0.7 high accuracy model and demultiplexed using guppy_barcoder, resulting in between 28.8 Mbp and 35.1 Mbp for each plasmid. These reads were used as input for medaka, using medaka_consensus to correct the original plasmid sequence.

3 | RESULTS

Our first motivation for quantifying the behavior of RiboJ arose during experiments aimed at understanding the effects of promoter polymorphisms segregating in the environmental *E. coli* population on transcription and translation. Here, "promoter" is defined as the entire intergenic region upstream of an open reading frame, as well as part of the upstream and downstream open reading frames

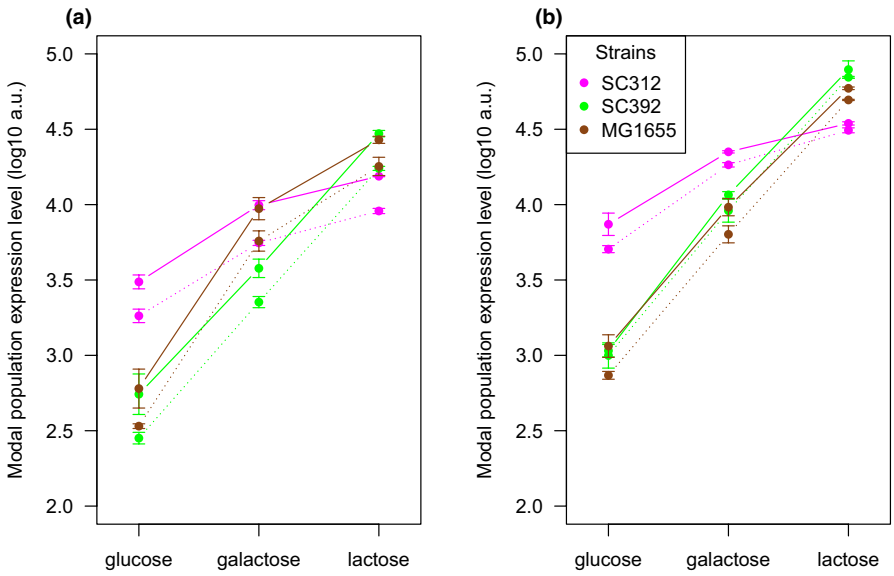

FIGURE 1 Modal expression levels differ consistently due to a single A to C change at position +69 of the *lacZ* open reading frame both without RiboJ (a) and with RiboJ (b) and across genetic backgrounds. Shown are the modal population fluorescence levels for GFP driven by an upstream *lacZ* promoter for three divergent strains of *E. coli*. In all genetic backgrounds tested and all growth conditions (glucose, galactose, and lactose), the fluorescence levels from a promoter with the 69C polymorphism (dotted line) were consistently lower than those from the promoter with the 69A polymorphism (solid line). On the log10 scale that is shown, a 0.1 difference is equivalent to a 25% change in expression. There are clear effects of genetic background on both the level and dynamic range of protein expression. In particular, SC312 has a narrow dynamic range, with relatively high expression in non-lactose environments compared to the other strains, but relatively low expression in lactose. Despite this, the effect of the A to C change is nearly constant. Whiskers show one standard deviation of three replicates

(Zaslaver et al., 2006). We include parts of the upstream and downstream open reading frames as it is well established that many open reading frames contain transcriptional regulatory elements affecting their own regulation or that of downstream genes. We assay the effects of the promoter on transcription by quantifying the fluorescence that occurs due to a GFP open reading frame that lies downstream of this "promoter".

In the case of the *lacZ*, here we define the "promoter" as the *lacI-lacZ* intergenic region, plus 88 and 71 bp of each flanking upstream (*lacI*) and downstream (*lacZ*) coding regions, respectively. We discovered a single SNP at position 69 relative to the *lacZ* gene start codon (C to A) that resulted in a change in downstream protein levels. We found that the effect of this SNP on GFP protein expression was consistent in different genetic backgrounds as well as during growth in different carbon sources (Figure 1a and Figure A1a). To check whether this C69A SNP affected transcription or translation (or both), we incorporated RiboJ as an insulator downstream of it (Figure 2, Top panels). This ensured that the mRNA being translated was identical regardless of which SNP was present. Thus, if the change in protein expression remained in the presence of RiboJ, we could infer that the change was due solely to the SNP affecting transcription. If the difference disappeared in the presence of RiboJ, we could infer that the change was due to the SNP affecting translation. However, it was also possible that the SNP itself interfered with RiboJ cutting (a *cis*-effect). If so, we could not unambiguously infer that the cause of the changes in fluorescence we observed was due to translation or transcription (or both). In addition, the genetic background of the strain itself might have affected RiboJ cutting (a *trans*-effect). To exclude the possibility of *cis*- or *trans*-effects on RiboJ cutting efficiency, we quantified efficiency in the presence of *cis*- and *trans*-genetic changes.

We designed an RT-qPCR assay to quantify the autocatalytic cutting activity of RiboJ. This assay is based on the principle that for a pair of forward and reverse primers that span the RiboJ cut site, an amplification product should only be produced for uncut mRNA molecules. In contrast, for primer pairs that do not span the cut site, an amplification product should be produced for all molecules. By examining the relative numbers of cut and uncut molecules, we can infer the efficiency of RiboJ cutting relative to the rate of production of all transcripts controlled by the same promoter (i.e., the rate of transcription). To this end, we designed two qPCR primer sets. The first set produced an amplicon from a region spanning the RiboJ cut site, while the second produced an amplicon from a region downstream of the RiboJ cut site (Figure 2). Both sets shared the same reverse primer, differing solely by the location of the forward primer. Because one forward primer binds upstream of the RiboJ cut site, no amplification can occur if the 5′ UTR sequence has been cut off (Figure 2, Middle panel). The second forward primer binds downstream of the cut site

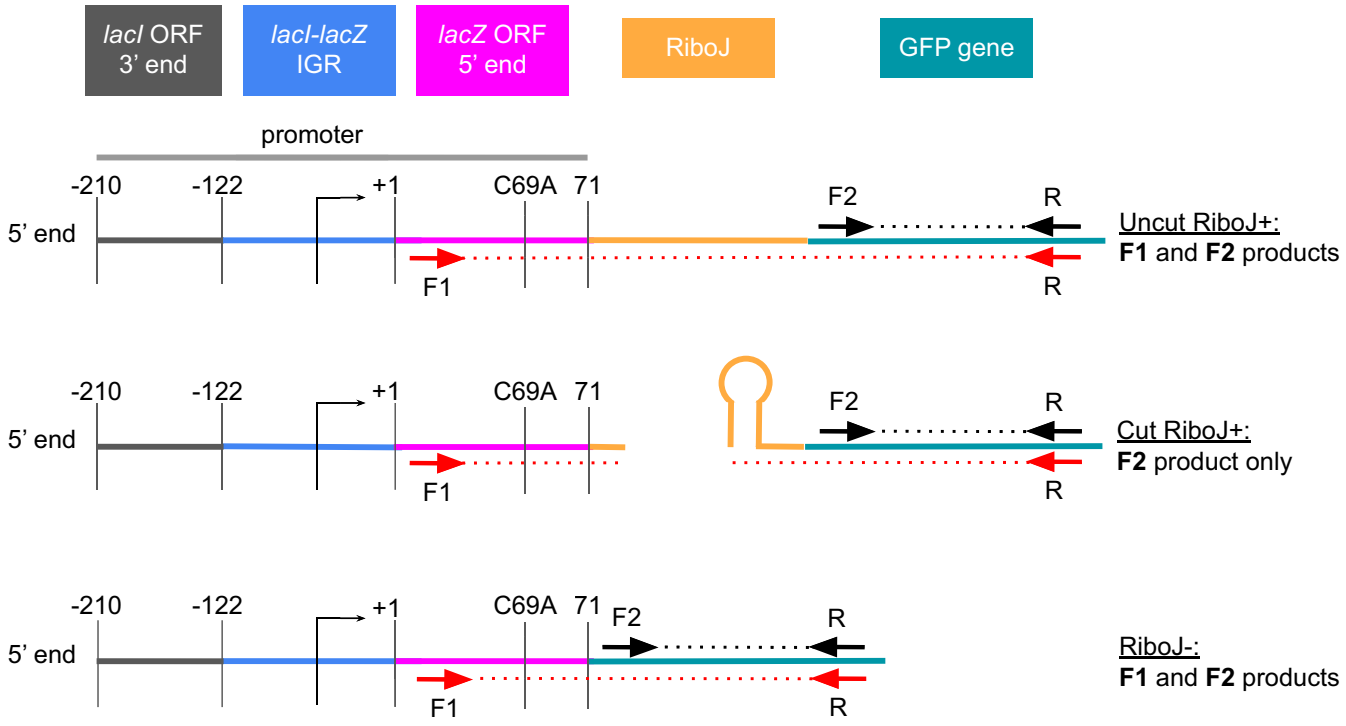

FIGURE 2 Scheme of the RT-qPCR primer design to quantify the efficiency of RiboJ cutting. Each primer is represented by an arrow, with pairs colored the same. The dotted lines indicate amplicons. If the dotted line between a primer pair is interrupted, the amplicon is not produced. When RiboJ cleaves off the 5′ UTR (Middle panel), the amplicon from primer F1 is not produced, while the amplicon from the F2 primer is still produced. The *lacI-lacZ* intergenic region (IGR) with the first 71 bp and last 88 bp of the *lacZ* and *lacI* open reading frames, respectively, were placed upstream of GFP (and RiboJ) as a promoter. The arrow in the middle of *lacI-lacZ* IGR indicates the transcription start site. The translation is driven from a strong synthetic ribosome binding site downstream of the *lacZ* gene sequence (here as a part of the GFP gene)

and results in an amplification product from all transcripts. To quantify differences in amplification that might result from primer binding or other unforeseen mechanisms, we calculated the relative fold change in the abundance of these two amplicons when RiboJ is absent. In the absence of RiboJ, any difference in amplification between the two primer sets should be due solely to differences in primer efficiency or related effects, as without RiboJ, both amplification products will always be produced (Figure 2, bottom panel).

We first assessed whether *trans*-genetic changes affected the self-splicing activity of RiboJ, by assaying RiboJ activity in six widely divergent strains of *E. coli* (Table 1). To test for *cis*-effects, we assayed activity in two promoter contexts, each varying by a single SNP that was 8 bp upstream of the RiboJ cut site (2 bp upstream of RiboJ sequence). We thus transformed each of the six strains with each of four plasmids differing by the C69A SNP in the *lacZ* promoter and either with RiboJ or without RiboJ (Figure 2, Table 2). We isolated RNA from exponentially growing cultures for all strains and confirmed that the amplification efficiency of all primer combinations with all templates was within the range of 90–110% (Figure A2). We used the resulting mean efficiency value across all strains (95.8%) for all subsequent calculations of RiboJ autocatalytic activity. We assayed the efficiency of RiboJ autocatalytic activity using at least triplicates for each strain and promoter combination (Experimental procedures). RiboJ cutting efficiency was high in all cases. Overall we found that 98% of all mRNA molecules containing RiboJ were cleaved. This was extremely robust for almost all strain and promoter combinations, with the lowest median value being 97% (Figure 3). We also found that RiboJ activity was robust

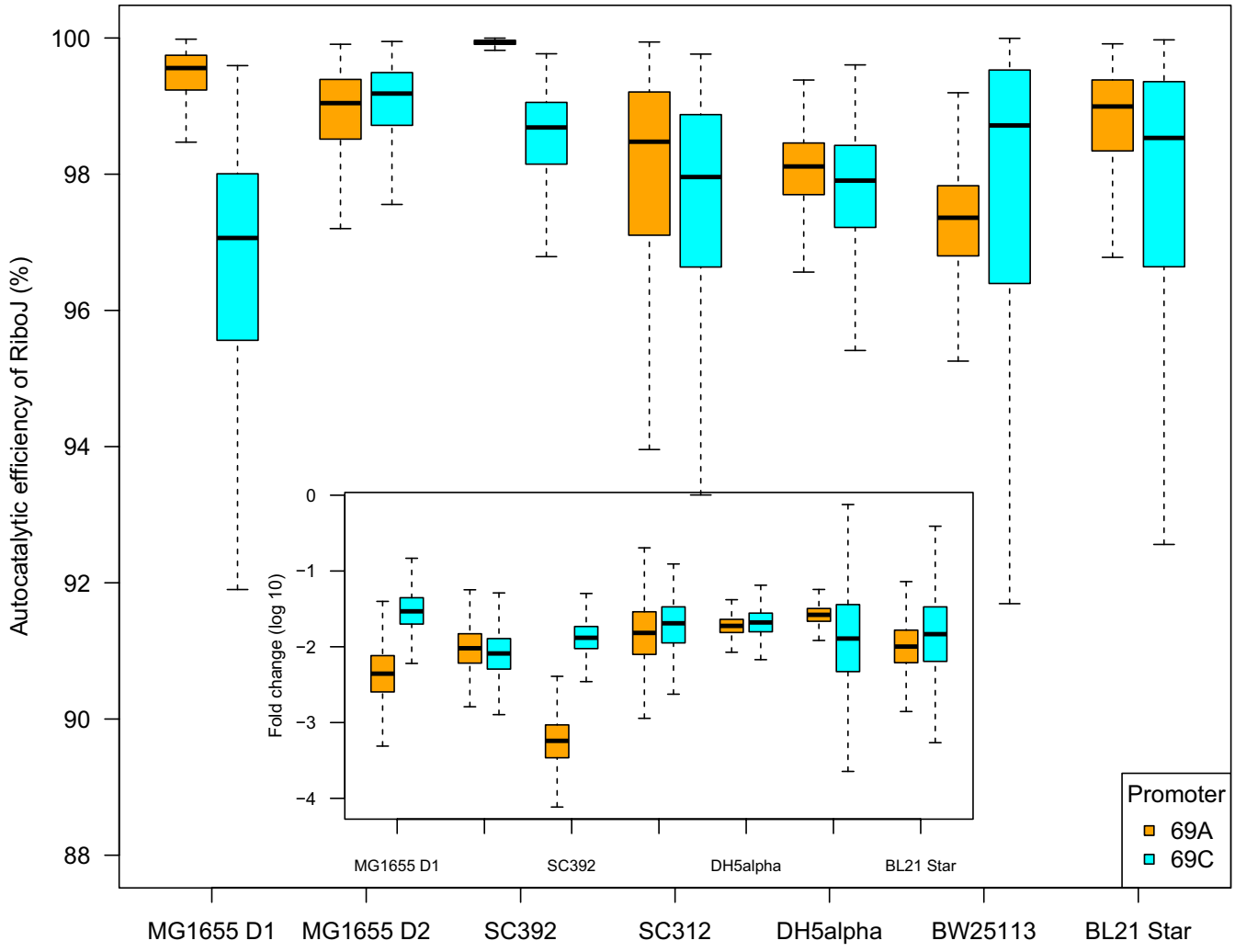

FIGURE 3 Autocatalytic efficiency of RiboJ. The boxplots in the figure show the minimum and maximum value (whiskers), the first and third quartile (boxes), and the median. These values were obtained through bootstrapping the RT-qPCR data (see Experimental procedures). D1 and D2 in MG1655 strain labels indicate that this data is from different biological replicates for which the RNA was extracted on different days (D1 and D2 denoting day 1 and day 2, respectively). The inset shows fold changes in the abundance of uncut transcripts with RiboJ relative to all transcripts. Note that the smaller range in cutting efficiency of RiboJ in SC392 strain for promoter 69A is simply a consequence of converting the C_t fold change of the two different amplicons into catalytic efficiency in percentages. The inset shows that the range and error in fold changes for SC392 with promoter 69A is comparable to the other samples

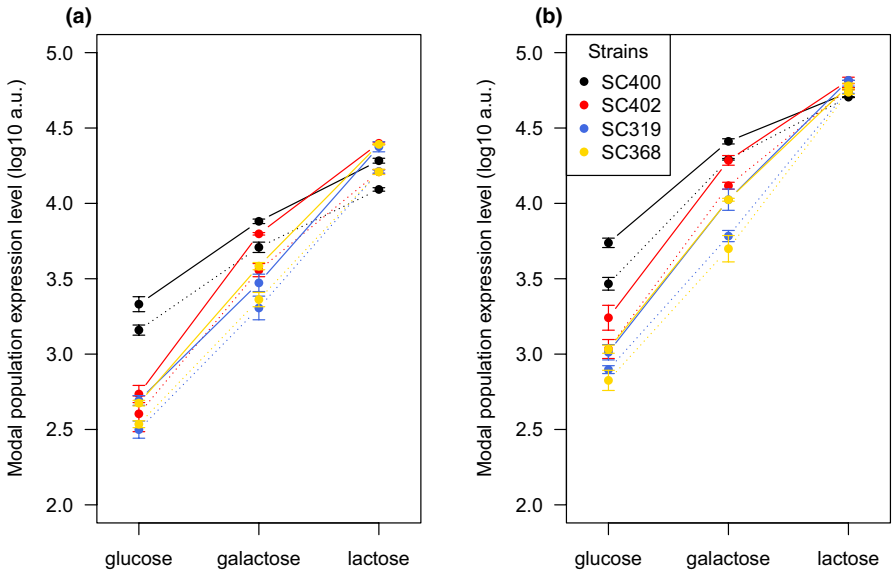

Figure A1 Modal expression levels differ consistently due to a single A to C change at position +69 of the *lacZ* open reading frame both without RiboJ (a) and with RiboJ (b) in additional strains. Shown are the modal population fluorescence levels for GFP driven by an upstream *lacZ* promoter for four additional *E. coli* strains in Figure 1. In all genetic backgrounds tested and all growth conditions (glucose, galactose, and lactose), the fluorescence levels from a promoter with the 69C polymorphism (dotted line) were consistently lower than those from the promoter with the 69A polymorphism (solid line). On the log10 scale that is shown, a 0.1 difference is equivalent to a 25% change in expression

to *cis*-changes, with no consistent differences between the 69A and 69C versions of the promoter.

However, we observed one exception to this robust behavior. In strain SC392, the 69A version of the promoter construct exhibited a 10-fold higher cutting efficiency (Figure 3, inset). To obtain an estimate of error for our measurements and test whether sampling alone could account for this or other differences, we bootstrapped the data 10,000 times and recalculated the efficiencies (Figure 3, Experimental procedures). We found that sampling alone was unlikely to account for the higher efficiency of 69A that we observed. We thus sequenced all plasmids from the SC392 strain to check for possible SNPs in the vector backbone that might lead to what we see as an increased RiboJ autocatalytic activity in this strain. We discovered single SNPs in each of the two plasmids with RiboJ. They were not identical, but each was located close to the origin of replication and thus could have affected the copy number of these plasmids. It is thus possible that the shift in RiboJ activity we observed for 69A in SC392 is due to a SNP in plasmid p69A.RJ+.

4 | DISCUSSION

Given the efficient activity of RiboJ (resulting in 98% of all mRNA molecules being cut), the differences in splicing measured between the two *lacZ* promoters and among the strains cannot explain the changes in expression we observed (Figure 1 and Figure A1). Rather than being due to the C69A SNP affecting the activity of RiboJ, the changes in expression are a consequence of this SNP affecting the regulation of both transcription and translation. The C69 SNP lowers both transcription and translation by approximately 25% (Figure 1 and Figure A1).

Considering the depth at which the activity of the *lacZ* promoter has been studied, we do not expect this SNP to be part of some unknown transcription factor binding site. The SNP may create a binding site causing a polymerase or transcription factor to pause during its linear search on the DNA strand for functional binding sites, thus inhibiting transcription. Processes such as transcriptional pausing and attenuation have been previously described in bacteria (Bailey et al., 2021; Blainey et al., 2006; Mustaev et al., 2017; Naville & Gautheret, 2009). These provide a more plausible explanation for the effect we see.

At the level of translational regulation, there is a possibility that the C69A SNP causes a difference in the secondary structure of the mRNA. This could then lead to differential accessibility of the mRNA to ribosomes. It is also possible that it inhibits proper translation by causing spurious ribosomal binding (Whitaker et al., 2015). However, it is beyond the scope of this paper to uncover the very mechanism of the regulation the SNP is involved in.

5 | CONCLUSION

In this study, we have confirmed that the autocatalytic activity of the ribozyme RiboJ is robust in *cis*- and *trans*-genetic contexts. This robust behavior of RiboJ suggests that the differences in expression that we observed in Figure 1 and Figure A1 are a result of changes in both transcription and translation due to the single C69A SNP, and not to changes in RiboJ autocatalytic efficiency. We note that

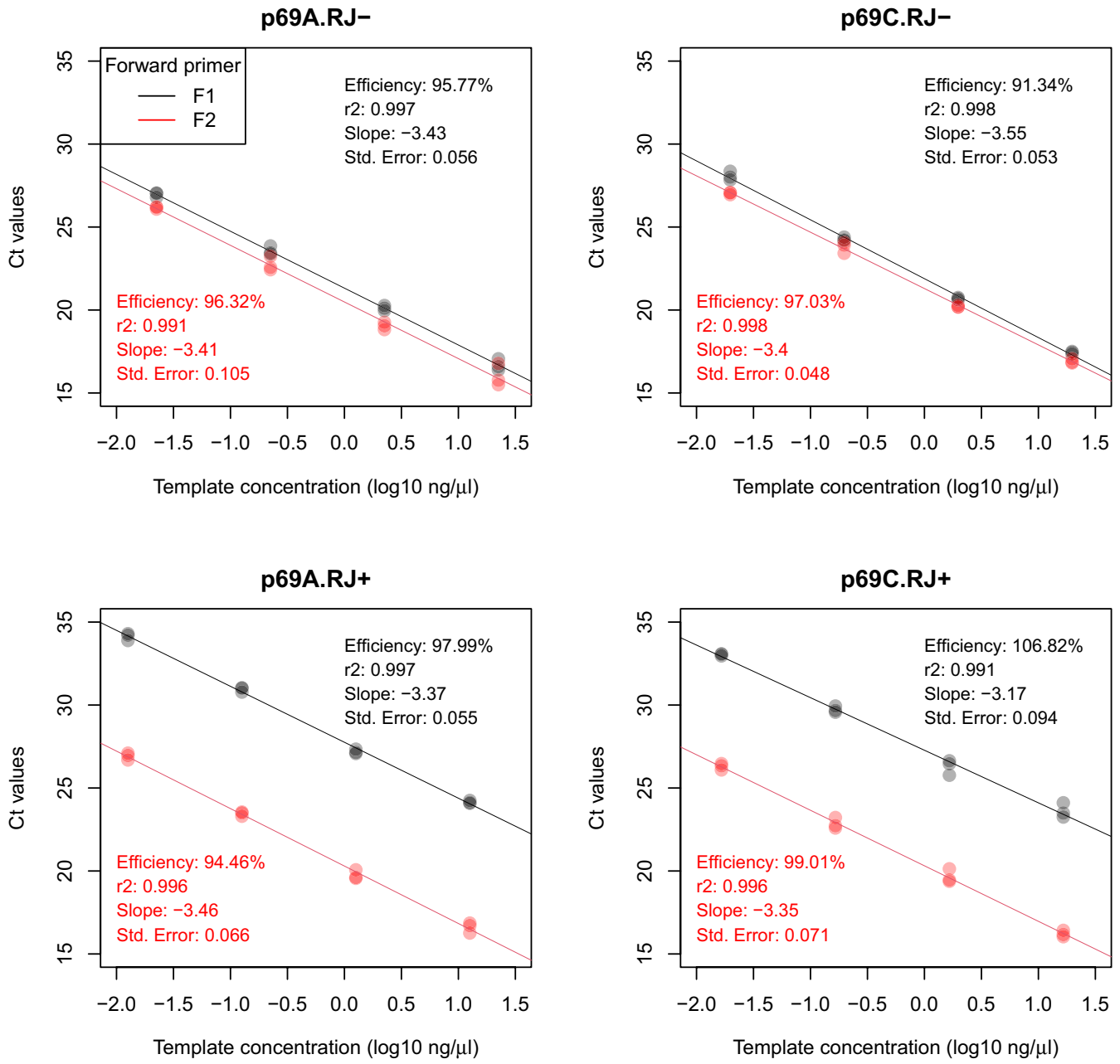

Figure A2 Amplification efficiency of all primer-template combinations. Each point indicates the C_t value for one technical replicate at different template concentrations, with each panel indicating one template and each color (red or black) indicating one primer combination. The lines show linear regressions, calculated using all data points for a given primer-template combination. For the templates without RiboJ (top panels), both primer pairs result in nearly equal C_t values. For the templates with RiboJ (bottom panels), the F1 primer pair has consistently larger C_t values, as expected. To calculate the catalytic activity of RiboJ, we used the mean amplification efficiency across all primer-template combinations, 95.8%.

there have been no previous reports that this region is involved in *lacZ* gene regulation. We proposed possible ways by which this SNP can be affecting both transcription and translation, however, more in-depth research is needed to determine the actual mechanism.

ACKNOWLEDGEMENTS

We would like to thank Tim Cooper for constructive input, D. Blank for providing p69A.RJ− and p69C.RJ− plasmids and Stella Pearless for plasmid preparation for Nanopore sequencing. This work was supported by the Royal Society Te Apārangi, Marsden Grant (MAU1703) awarded to Olin K. Silander. The funder had no role in study design, data collection, and interpretation, or the decision to submit the work for publication.

CONFLICT OF INTEREST

None declared.

AUTHOR CONTRIBUTIONS

Markéta Vlková: Conceptualization (equal); Data curation (lead); Formal analysis (lead); Investigation (equal); Methodology (lead); Project administration (supporting); Visualization (lead); Writing-original draft (lead); Writing-review & editing (equal). **Bhargava Reddy Morampalli:** Investigation (equal); Writing-review & editing (supporting). **Olin K. Silander:** Conceptualization (equal); Formal analysis (supporting); Funding acquisition (lead); Methodology (supporting); Project administration (lead); Resources (lead); Supervision (lead); Writing-original draft (supporting); Writing-review & editing (equal).

ETHICS STATEMENT

None required.

DATA AVAILABILITY STATEMENT

The original data files and scripts with data analysis that support the findings of this study are available in GitHub and accessible through the Zenodo repository at https://doi.org/10.5281/zenodo.5154246

ORCID

Markéta Vlková https://orcid.org/0000-0002-0272-5115
Bhargava Reddy Morampalli https://orcid.org/0000-0003-2741-6826
Olin K. Silander https://orcid.org/0000-0003-4105-8316

REFERENCES

Bailey, S. F., Alonso Morales, L. A., & Kassen, R. (2021). Effects of synonymous mutations beyond codon bias: The evidence for adaptive 4 synonymous substitutions from microbial evolution experiments. *Genome Biology and Evolution*, evab141. https://doi.org/10.1093/gbe/evab141

Barbier, I., Perez-Carrasco, R., & Schaerli, Y. (2020). Controlling spatiotemporal pattern formation in a concentration gradient with a synthetic toggle switch. *Molecular Systems Biology*, *16*(6), e9361. https://doi.org/10.15252/msb.20199361

Bartoli, V., Meaker, G. A., di Bernardo, M., & Gorochowski, T. E. (2020). Tunable genetic devices through simultaneous control of transcription and translation. *Nature Communications*, *11*(1), 2095. https://doi.org/10.1038/s41467-020-15653-7

Bittihn, P., Didovyk, A., Tsimring, L. S., & Hasty, J. (2020). Genetically engineered control of phenotypic structure in microbial colonies. *Nature Microbiology*, *5*(5), 697–705. https://doi.org/10.1038/s41564-020-0686-0

Blainey, P. C., van Oijen, A. M., Banerjee, A., Verdine, G. L., & Xie, X. S. (2006). A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(15), 5752–5757. https://doi.org/10.1073/pnas.0509723103

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., & Shao, Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, *277*(5331), 1453–1462. https://doi.org/10.1126/science.277.5331.1453

Breckell, G., & Silander, O. K. (2020). Complete genome sequences of 47 environmental isolates of *Escherichia coli*. *Microbiology Resource Announcements*, *9*(38), e00222-20. https://doi.org/10.1128/MRA.00222-20

Carrier, T. A., & Keasling, J. D. (1997). Engineering mRNA stability in *E. coli* by the addition of synthetic hairpins using a 5′ cassette system. *Biotechnology and Bioengineering*, *55*(3), 4. https://doi.org/10.1002/(SICI)1097-0290(19970805)55:3<577:AID-BIT16>3.0.CO;2-D

Clifton, K. P., Jones, E. M., Paudel, S., Marken, J. P., Monette, C. E., Halleran, A. D., Epp, L., & Saha, M. S. (2018). The genetic insulator RiboJ increases expression of insulated genes. *Journal of Biological Engineering*, *12*(1), 23. https://doi.org/10.1186/s13036-018-0115-6

Datsenko, K. A., & Wanner, B. L. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proceedings of the National Academy of Sciences*, *97*(12), 6640–6645. https://doi.org/10.1073/pnas.120163297

Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, *6*(5), 343–345. https://doi.org/10.1038/nmeth.1318

Ishii, S., Ksoll, W. B., Hicks, R. E., & Sadowsky, M. J. (2006). Presence and growth of naturalized *Escherichia coli* in temperate soils from lake superior watersheds. *Applied and Environmental Microbiology*, *72*(1), 612–621. https://doi.org/10.1128/AEM.72.1.612-621.2006

Levskaya, A., Chevalier, A. A., Tabor, J. J., Simpson, Z. B., Lavery, L. A., Levy, M., Davidson, E. A., Scouras, A., Ellington, A. D., Marcotte, E. M., & Voigt, C. A. (2005). Engineering *Escherichia coli* to see light. *Nature*, *438*(7067), 441–442. https://doi.org/10.1038/nature04405

Li, C., Wen, A., Shen, B., Lu, J., Huang, Y., & Chang, Y. (2011). FastCloning: A highly simplified, purification-free, sequence- and ligation-independent PCR cloning method. *BMC Biotechnology*, *11*(1), 92. https://doi.org/10.1186/1472-6750-11-92

Lou, C., Stanton, B., Chen, Y.-J., Munsky, B., & Voigt, C. A. (2012). Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nature Biotechnology*, *30*(11), 1137–1142. https://doi.org/10.1038/nbt.2401

Mustaev, A., Roberts, J., & Gottesman, M. (2017). Transcription elongation. *Transcription*, *8*(3), 150–161. https://doi.org/10.1080/21541264.2017.1289294

Na, D., Yoo, S. M., Chung, H., Park, H., Park, J. H., & Lee, S. Y. (2013). Metabolic engineering of Escherichia coli using synthetic small regulatory RNAs. *Nature Biotechnology*, *31*(2), 170–174. https://doi.org/10.1038/nbt.2461

Naville, M., & Gautheret, D. (2009). Transcription attenuation in bacteria: Theme and variations. *Briefings in Functional Genomics and Proteomics*, *8*(6), 482–492. https://doi.org/10.1093/bfgp/elp025

Neves, D., Vos, S., Blank, L. M., & Ebert, B. E. (2020). Pseudomonas mRNA 2.0: Boosting gene expression through enhanced mRNA stability and translational efficiency. *Frontiers in Bioengineering and Biotechnology*, *7*, 458. https://doi.org/10.3389/fbioe.2019.00458

Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Research*, *29*(9), 45e–445. https://doi.org/10.1093/nar/29.9.e45

Schmiedel, J. M., Carey, L. B., & Lehner, B. (2019). Empirical mean-noise fitness landscapes reveal the fitness impact of gene expression noise. *Nature Communications*, *10*(1), 3180. https://doi.org/10.1038/s41467-019-11116-w

Urtecho, G., Tripp, A. D., Insigne, K. D., Kim, H., & Kosuri, S. (2019). Systematic dissection of sequence elements controlling σ70 promoters using a genomically encoded multiplexed reporter assay in *Escherichia coli*. *Biochemistry*, *58*(11), 1539–1551. https://doi.org/10.1021/acs.biochem.7b01069

Whitaker, W. R., Lee, H., Arkin, A. P., & Dueber, J. E. (2015). Avoidance of truncated proteins from unintended ribosome binding sites within heterologous protein coding sequences. *ACS Synthetic Biology*, *4*(3), 249–257. https://doi.org/10.1021/sb500003x

Yu, H., Wang, Z., Xu, H., Guo, J., Ma, Q., Mu, X., & Luo, Y. (2018). A method for absolute protein expression quantity measurement employing insulator RiboJ. *Engineering*, *4*(6), 881–887. https://doi.org/10.1016/j.eng.2018.09.012

Zaslaver, A., Bren, A., Ronen, M., Itzkovitz, S., Kikoin, I., Shavit, S., Liebermeister, W., Surette, M. G., & Alon, U. (2006). A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nature Methods*, *3*(8), 623–628. https://doi.org/10.1038/nmeth895

How to cite this article: Vlková, M., Morampalli, B. R., & Silander, O. K. (2021). Efficiency of the synthetic self-splicing RiboJ ribozyme is robust to *cis*- and *trans*-changes in genetic background. *MicrobiologyOpen*, 10, e1232. https://doi.org/10.1002/mbo3.1232

APPENDIX A

TABLE A1 Primers and oligos used in this work

Primer or oligo ID	Sequence	Purpose
pUA66_insert_F3965	5′ - TTG TCT GTT GTG CCC AGT CAT AGC -3′	PCR & Sanger sequencing
pUA66_insert_R232	5′- TCG CAA AGC ATT GAA GAC CAT ACG C -3′	PCR & Sanger sequencing
RiboJ_oligo1_Rev	5′ - GAA AGC ACA TCC GGT GAC AGC TGG ATC CCC TCG AGG TGA AGA CGA AAG GGC CTC GTG ATA - 3′	DNA assembly of pMV001
RiboJ_oligo2_For	5′ - GGG GAT CCA GCT GTC ACC GGA TGT GCT TTC CGG TCT GAT GAG TCC GTG AGG ACG AAA CAG - 3′	DNA assembly of pMV001
RiboJ_oligo3_Rev	5′ - TCT TAG TTT AAA CAA AAT TAT TTG TAG AGG CTG TTT CGT CCT CAC GGA CTC ATC AGA CCG - 3′	DNA assembly of pMV001
RiboJ_oligo4_For	5′ - CCT CTA CAA ATA ATT TTG TTT AAA CTA AGA AGG AGA TAT ACA TAT GAG TAA AGG AGA - 3′	DNA assembly of pMV001
pUA66_vector_F	5′ - GAA GGA GAT ATA CAT ATG AGT AAA GG - 3′	PCR of pUA66 for DNA assembly of pMV001 & RT-qPCR assay (**primer F2**)
pUA66_vector_R	5′ - TCG AGG TGA AGA CGA AAG G - 3′	PCR of pUA66 for DNA assembly of pMV001
pMV001_FastClonV_F	5′ - AGC TGT CAC CGG ATG TGC - 3′	PCR of pMV001 for DNA assembly of p69A.RJ+ and p69C.RJ+
pMV01_FastClonV_R	5′ - TCG AGG TGA AGA CGA AAG GGC - 3′	PCR of pMV001 for DNA assembly of p69A.RJ+ and p69C.RJ+
lacZ_FastClonIN_F	5′ - **TTT CGT CTT CAC CTC GA**C AAT ACG CAA ACC GCC TCT CC - 3′	PCR of p69A.RJ− and p69C.RJ− for DNA assembly of p69A.RJ+ and p69C.RJ+
lacZ_FastClonIN-D9_R	5′ - **CAC ATC CGG TGA CAG CT**G TGT AAC GCC AGG GTT TTC C - 3′	PCR of p69A.RJ− for DNA assembly of p69A.RJ+
lacZ_FastClonIN-K12_R	5′ - **CAC ATC CGG TGA CAG CT**G GGT AAC GCC AGG GTT TTC C - 3′	PCR of p69C.RJ− for DNA assembly of p69C.RJ+
lacZ_qPCR_F	5′ - ATG ATT ACG GAT TCA CTG GC - 3′	RT-qPCR assay (**primer F1**)
GFP_qPCR_R	5′ - GAA AAT TTG TGC CCA TTA ACA TCA CC - 3′	RT-qPCR assay (**primer R**)
GFP_Probe	5′ - /56-FAM/TTC AAC AAG/ZEN/AAT TGG GAC AAC TCC AGT GAA AAG TT/3IABkFQ/ - 3′	RT-qPCR assay (**probe**)

Note: Bold sequences = regions homologous to PCR amplified pMV001 vector. Primers and the probe used for RT-qPCR assay are highlighted in bold with their simplified names that were used in the main text and Figure 2 (Purpose column).