# The influence of sequence context on the evolution of bacterial gene expression

by

Magdalena Steinrück

October, 2018

*A thesis presented to the*
*Graduate School*
*of the*
*Institute of Science and Technology Austria, Klosterneuburg, Austria*
*in partial fulfillment of the requirements*
*for the degree of*
*Doctor of Philosophy*

IST AUSTRIA
Institute of Science and Technology

The dissertation of Magdalena Steinrück, titled *The influence of sequence context on the evolution of bacterial gene expression*, is approved by:

**Supervisor**: Prof. Calin C. Guet, IST Austria, Klosterneuburg, Austria

Signature: ___________________________________

**Committee Member**: Prof. Nicholas H. Barton, IST Austria, Klosterneuburg, Austria

Signature: ___________________________________

**Committee Member**: Prof. Martin Ackermann, ETH Zürich, Switzerland

Signature: ___________________________________

**Exam Chair**: Prof. Daria Siekhaus, IST Austria, Klosterneuburg, Austria

Signature: ___________________________________

I hereby declare that this dissertation is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: ________________________

Magdalena Steinrück

October 30, 2018

# Abstract

Expression of genes is a fundamental molecular phenotype that is subject to evolution by different types of mutations. Both the rate and the effect of mutations may depend on the DNA sequence context of a particular gene or a particular promoter sequence. In this thesis I investigate the nature of this dependence using simple genetic systems in *Escherichia coli*. With these systems I explore the evolution of constitutive gene expression from random starting sequences at different loci on the chromosome and at different locations in sequence space. First, I dissect chromosomal neighborhood effects that underlie locus-dependent differences in the potential of a gene under selection to become more highly expressed. Next, I find that the effects of point mutations in promoter sequences are dependent on sequence context, and that an existing energy matrix model performs poorly in predicting relative expression of unrelated sequences. Finally, I show that a substantial fraction of random sequences contain functional promoters and I present an extended thermodynamic model that predicts promoter strength in full sequence space. Taken together, these results provide new insights and guides on how to integrate information on sequence context to improve our qualitative and quantitative understanding of bacterial gene expression, with implications for rapid evolution of drug resistance, *de novo* evolution of genes, and horizontal gene transfer.

## Acknowledgments

I am very grateful to my advisor Călin Guet. From the very beginning, he gave me trust and freedom, much more than I could have asked for. He always nurtured my own trust in my capabilities and encouraged me to push my boundaries. Whenever it was important, he was available. When I became a mother, he motivated me to protect the scientist in me, in the light of my new role. To me, Calin is an extraordinary role model as freethinker, intrepid of doing things differently, and nurturing the freshness of the beginner's mind.

I thank my committee members: Martin Ackermann for his inspiring joy in doing science and for moral support; Nick Barton for the rotation in his group, and for his fast and helpful feedback on manuscripts; Jon Bollback for bringing microbial evolutionary biology to IST and for advice on statistics. I thank Daria Siekhaus for her charming way of chairing my defense, and for being an important female role model.

I thank all Guet group members for their peer support, for their constructive criticism in our lab meetings, and for many fun hours in the lab. In particular, I thank Maroš Pleška for his critical comments on many versions of the *eLife* manuscript; Remy Chait for building the macroscopes and for all his little tinkering, making me laugh so often; Kathrin Tomašek for keeping the lab in an organized state; Bella Tomanek for taking the amplification reporter idea to the next level; Tobi Bergmiller for bringing solid molecular microbiology experience to the young lab. I thank Kirti, Bor, Dora, Michael, Allen, Anna, Anna, Moritz, Karin, Marta, Marjon, Guillaume, Bella, Kathrin and many others for engaging and relaxing lunch and coffee breaks.

Special thanks go to my collaborators, to Murat Tuğrul for his persistent genuine interest in the experiments, to our data magician Srdjan Sarikas for finding those project hours while being on paternity leave, to Mato Lagator for his efforts in pulling it all together, and to Gašper Tkačik for his sharp and instructive analyses and for his positive attitude.

Many thanks to everyone dedicated to building IST Austria. It was a unique experience to be part of it as it was growing from infancy to adolescence. In particular, I thank everyone involved in making it a family-friendly place. I am grateful for the friends I made here, and for the kindergarten community.

My final thanks go to my family, to my parents for their constant and manifold support, to Filip for being my coach and for being there when I needed him, and to my kids for challenging me in refreshing ways, for their laughter and love.

## About the Author

Magdalena Steinrück completed a BSc in Food Sciences and Biotechnology at the University of Agriculture and Life Sciences in Vienna (BOKU). During her MSc in Biotechnology (BOKU) she also studied and performed research at Cornell University, Ithaca, NY, and at the Medical University of Vienna. For her diploma work and studies, she was awarded the 'Würdigungspreis' of the Austrian ministry of research. She joined IST Austria as part of the first cohort of seven PhD students in 2010. Magdalena did rotations in the research groups of Michael Sixt and Nick Barton, with whom she developed her interest in evolution, setting course for the direction of her PhD as first member in the lab of Calin Guet. Participating in the 2012 Advanced Bacterial Genetics course at the Cold Spring Harbor Laboratories became another influential experience for her research on constraints in bacterial regulatory evolution. She presented her work at two Gordon Research Conferences (Microbial Stress Response 2016 and Marine Microbes 2018) and published her results on the context-dependency of adaptation via increased gene expression in the online journal *eLife*. Magdalena enjoys addressing basic evolutionary questions using experiments with simple genetic model systems, and hopes to connect this approach in the future with her more recently developed interest in microbial life in the diverse wild and in the distant past.

## List of Publications Appearing in this Thesis

Steinrueck, M. and Guet, C.C. Complex chromosomal neighborhood effects determine the adaptive potential of a gene under selection. **eLife** 2017;6:e25100.

# Table of Contents

## List of Symbols/Abbreviations

| | |
|---|---|
| **AF95** | 95th percentile of a no-plasmid autofluorescence control culture |
| **cDNA** | complementary DNA |
| **FACS** | fluorescence associated cell sorting |
| **gDNA** | genomic DNA |
| **IS** | insertion sequence |
| **MIC** | minimum inhibitory concentration |
| **PCR** | polymerase chain reaction |
| **$P_{on}$** | probability of RNAP being bound ('ON' state of the promoter) |
| **qPCR** | quantitative PCR |
| **RFU** | relative fluorescence units |
| **RNAP** | RNA polymerase holoenzyme |
| **ROUT** | robust regression and outlier removal |
| **rpm** | rotations per minute |
| **sort-seq** | FACS sorting of cells into bins of different fluorescence, followed by bin-specific barcoding PCR and high-throughput sequencing of fluorescence determinants |
| **TF** | transcription factor |
| **TSS** | transcription start site |
| **wt** | wild type |

# Preface

In biology, the correct answer to almost any question needs to start with the same two words: '*It depends …*'

At all levels, nothing in biology exists in isolation; everything exists in and depends on a biological context that may become important in unexpected ways. In molecular genetics, *context* has a quite literal meaning. The four different nucleotide 'letters' are each others' neighbors on the linear DNA molecule. The same is true for genes, the 'words' formed by these letters. Genes physically exist next to each other on the DNA like beads on a string. They also evolve like that (and, in bacteria, where sex is rare, not like beans in a bag). In this thesis, I investigate, at the two levels of nucleotides and genes, whether and *how* evolution of gene expression in bacteria depends on sequence context.

Strictly speaking, to form a correct answer to a biological question, a full stop after the two words above is sufficient: *'It depends.'* However, what makes an answer a useful one is what follows after these first two words. I hope that this thesis will provide a few such useful answers.

# 1 Introduction

With this thesis, I hope to contribute a minuscule bit to one of the fundamental goals of biology at the intersection of molecular and evolutionary biology: to understand biological phenotypes from genotypes. To this end, I study how genes come to be expressed by way of mutations. Using simple synthetic genetic constructs, instead of dissecting natural genetic model systems, I seek to identify generally applicable factors in the sequence context of these constructs that constrain the evolution of gene expression. The hope is that, if we take these factors into account, we can get a better understanding of the function and evolution of gene expression in naturally evolving systems.

Before I give a preview of how this endeavor is pursued in the individual chapters, I briefly locate this work in the larger context of current molecular evolutionary biology. More specific introductions can be found in the respective sections of the individual chapters.

## *1.1 'The middle way' between molecular model systems and the sequencing data deluge*

The foundations of molecular biology in the 1950s and 1960s were laid by the detailed dissection of individual model systems such as the *lac* operon of *E. coli* (Jacob & Monod 1961) and the lifecycle control of phage λ (E. M. Lederberg & J. Lederberg 1953; Gottesman & Weisberg 2004). Many questions concerning the ecology and evolution of these model systems await being addressed (for examples, see the theses of my colleagues Fabienne (Jesse 2017) and Maroš (Pleška 2017)). On a particular molecular level however, after countless lessons have been learned from *lac*, λ, et cetera, our understanding of their function has saturated. What remains to be seen is how much trouble we cause ourselves by the narrow focus on model systems when trying to generalize to other genes and organisms. We will come across this problem in chapter 4, where we will see that the description of the interaction between RNA polymerase and the *lac* promoter applies poorly to the full promoter sequence space.

Starting in the mid-2000s, 'next generation sequencing' put an end to the sequencing bottleneck that before then had been limiting molecular biology (Schuster 2007), just to, as it goes with bottlenecks, create a new one. Today, sequencing data of all kinds (genomic, metagenomic, RNAseq, ChIPseq, …) is pouring in at a much faster rate than we can make sense of it. Also, systems biology, which set out to explain biology by integrating components and models (Ehrenberg et al. 2003), does not simply scale up with the amount of data one would hope to apply it on (Brenner 2009). Models of interacting components often lose their explanatory power when applied to a large scale. This is why, borrowing from a seminar title by Nick Barton, ‘systems biology may be doomed to fail’ (FrisBi seminar, Feb 27th, 2015).

What we are left with is a gap between two extremes: detailed local data on molecular model systems and poorly understood data on a global scale. The gap calls for a ‘middle way’ that allows us to upscale insights from the reductionist study of model genes to the scale of -omics data. It is in this gap where the trickiness of biology lies: Since no two systems under study are alike, it is crucial to identify what differences are the important ones. Only if we manage to identify these ‘differences that make a difference’ (Bateson), will we make progress in putting together a picture that will work more generally. In chapter 2, I will argue that one such important difference between two genes, when it comes to evolution of their expression, is their position on the chromosome. And fortunately, it appears that we can understand why.

## *1.2 Using synthetic systems and random sequences to learn about natural ones*

As said above, every evolved biological system comes with its particularities that may complicate the abstraction of general principles. For example, when trying to understand the strength of a constitutive promoter, as we will in chapter 4, the fact that most genes are regulated, constitutes a complication. Or, when trying to carve out the isolated effect of chromosome position on evolution, as we do in chapter 2, moving a native gene to different chromosome positions may yield results that are difficult to generalize, due to a history of co-adaptation between a gene and its locus.

Throughout this thesis, I circumvent this problem by using well-defined synthetic genetic systems that function independently from the host cell machinery. They include fluorescent reporters and, in chapter 2, a ‘stand-alone’ antibiotic pump gene. ‘Synthetic’ here refers solely to the orthogonality of components with respect to the host cell and should not be confused with efforts to engineer an artificial function for any purpose other than learning about natural systems.

Also, instead of starting from functional model promoters, I use random sequences to drive expression. This approach offers two advantages: First, by acquiring data from random sequences covering a larger sequence space (chapter 4), results are expected to generalize well over the full space of promoter functionality, instead of being locally fitted to some particular sequence. Second, studying the emergence of function in and from random sequences may mimic a possibly important, only recently discovered mode of evolution, that is the evolution of transcripts and genes from scratch, i.e. *de novo* (McLysaght & Guerzoni 2015).

## *1.3 Questions addressed in this thesis*

Chapter 2 describes evolution experiments, in which I subjected engineered strains of *E. coli* to selection for increased expression of an antibiotic pump gene, placed at different positions of the chromosome, to answer the following questions:

- How does the adaptive potential of a gene vary with the position of the gene on the chromosome?
- What mutation types contribute to adaptation via increased gene expression at different chromosomal positions?
- What are the determinants of chromosomal neighborhood that underlie differences in adaptive potential and what do these determinants predict about the distribution of adaptive potential on the chromosome?

After considering the full range of possible mutation types in chapter 2, chapter 3 zooms in on the effect of point mutations on promoter strength and on modeling these effects using a thermodynamic framework. Together with my collaborators Srdjan Sarikas, Murat Tugrul and Gašper Tkačik, we quantify the effect of single nucleotide mutations on three different starting sequences, one of which we had already used in chapter 2. We use three promoter-GFP libraries and sort-seq to address the following questions:

- Can we predict promoter-generating point mutations observed in evolution experiments using energy matrix models of RNA polymerase binding?
- How specific is the predictive power of energy matrix models to distinct sequence contexts?
- How does the effect of promoter mutations depend on the promoter sequence context?
- What possible reasons are there for the context-specificity of energy matrix models?

Unexpected observations in chapter 3 raise the question how the sequence context of an RNA polymerase binding site, i.e. its location in the full sequence space, influences promoter function. In chapter 4, continuing the collaboration with Srdjan Sarikas and Gašper Tkačik, we address these unexpected results using sort-seq data of another promoter-GFP library that covers a much larger area of random sequence space. We also test different extensions of a thermodynamic model of RNA polymerase binding. Thereby we address the following questions:

- What is the distribution of promoter strength in random sequence space?
- How can we improve predictions of promoter strength in the full sequence space and what does that tell us about the emergence of promoter function?

# 2 Complex chromosomal neighborhood effects determine the adaptive potential of a gene under selection

This chapter was originally published in (Steinrueck & Guet 2017). Figure supplements are in the appendix of this thesis, additional source data and video files are available on the webpage of the article (open access): *https://elifesciences.org/articles/25100*

## *2.1 Abstract*

How the organization of genes on a chromosome shapes adaptation is essential for understanding evolutionary paths. Here, we investigate how adaptation to rapidly increasing levels of antibiotic depends on the chromosomal neighborhood of a drug-resistance gene inserted at different positions of the *Escherichia coli* chromosome. Using a dual-fluorescence reporter that allows us to distinguish gene amplifications from other up-mutations, we track in real-time adaptive changes in expression of the drug-resistance gene. We find that the relative contribution of several mutation types differs systematically between loci due to properties of neighboring genes: essentiality, expression, orientation, termination, presence of duplicates. These properties determine rate and fitness effects of gene amplification, deletions, and mutations compromising transcriptional termination. Thus, the adaptive potential of a gene under selection is a system-property with a complex genetic basis that is specific for each chromosomal locus, and it can be inferred from detailed functional and genomic data.

## 2.2 *Introduction*

In the process of regulatory evolution, a finite set of genes are continuously combined to form new gene expression patterns and create a myriad of phenotypes (Carroll 2000; Wittkopp et al. 2004; Wray 2007). Acquiring mutations that increase the expression of a single gene can be sufficient to make an individual substantially fitter than its competitors. For example, increased expression of drug target or efflux genes is a common mechanism for the evolution of resistance to antibiotics (Li et al. 2015; Palmer & Kishony 2014), chemotherapeutics (Cole et al. 1992), and insecticides (Devonshire & Field 1991; Coderre & Beverley 1983). Increased expression of individual genes also provides access to new nutrient resources (Notebaart et al. 2014) and tolerance to diverse toxins (Soo et al. 2011). The fitness effect of increased expression of individual genes has mostly been determined in plasmid-based overexpression libraries (Notebaart et al. 2014; Soo et al. 2011). However, the large majority of genes reside on chromosomes, neighboring other genes, and thus mutations affecting gene expression occur in a specific chromosomal context. Unequal mutation rates along the genome (Foster et al. 2013; Anderson & Roth 1981) imply that the chromosomal location can affect the adaptive potential of a gene, i.e. the probability that adaptive mutations increasing expression of the gene will spread in a population under given selective conditions.

Adaptation by increased gene expression can result from mutations of different types (Blank et al. 2014; Lind et al. 2015): point mutations, promoter insertion by mobile elements (Mahillon & Chandler 1998; Ellison & Bachtrog 2013; Stoebel et al. 2009), promoter capture by chromosomal rearrangements (ar-Rushdi et al. 1983; Blount et al. 2012; Xiao et al. 2008), and gene duplication or amplification, which increases expression by way of gene dosage (Andersson & Hughes 2009; Elliott et al. 2013). How the rate of mutation of these individual mutation types depends on chromosomal position has in part been determined experimentally (Foster et al. 2013; Hudson et al. 2002; Mahillon & Chandler 1998; Craig 1997; Touchon et al. 2009; Anderson & Roth 1981; Seaton et al. 2011; Wahl et al. 1984). Despite considerable experimental data, we currently lack an understanding of how position biases of the different mutation types together combine across different chromosomal loci, and therefore how the chromosomal context of a gene under selection affects overall adaptation.

Here, we investigate how the complex interplay of different mutation types and mutation rate biases gives rise to an effect of chromosome position on adaptation in *Escherichia coli*. To this end, we use a single chromosomal drug resistance gene as the target of selection and a two-color fluorescence reporter readout for adaptive mutations in evolution experiments. We quantify the effect of the chromosomal position of the selected gene on adaptation and identify the mutation types underlying this effect. We find that a strong effect of chromosome position on adaptation is largely explained by rate differences of gene duplications and fitness effect differences of two types of promoter co-opting mutations (promoter capture deletions and mutations that cause read-through across upstream transcriptional terminators). Both the observed rate differences and fitness effect differences depend on simple features of the chromosomal neighborhood of the gene under selection. This suggests that the adaptive potential of a gene can be estimated by looking for respective features of chromosomal neighborhoods in genomics data. Based on these results, we propose that the chromosomal context of a gene under selection is an important factor in adaptation.

## *2.3 Results*

### 2.3.1 A dual-fluorescence reporter cassette for tracking the dynamics of adaptive mutations of different types

We devised an evolution experiment with *Escherichia coli,* in which we use a single target of selection embedded in a genetic cassette that serves as a reporter of adaptive potential and mutation types. The reporter cassette can be inserted at any chromosomal position (Figure 1A and Figure 1B), and it allows us to distinguish amplifications from other adaptive mutations in real-time using two-color fluorescence measurements. The reporter cassette contains a promoterless, translational *tetA-yfp* gene fusion followed by a transcriptional terminator and a constitutively expressed *cfp* gene. Mutations that increase expression of the tetracycline efflux pump TetA-YFP can be selected with antibiotic and monitored through YFP fluorescence (Figure 1C, left). Due to the immediate proximity of the *tetA-yfp* and *cfp* genes, the large majority of *tetA-yfp* amplifications are expected to contain the *cfp* gene as well. Thus, adaptation by reporter cassette amplification is expected to be distinguishable from other up-mutations by a fluorescence increase of both YFP and CFP (Figure 1C, right). We integrated the reporter cassette at four different intergenic loci (*A, B, C, D*) along the chromosome of an *E. coli ΔtolC* strain (Figure 1A), giving rise to four strains (strain *A*, *B*, *C,* and *D*). The four loci were chosen to lie in intergenic regions between divergently transcribed genes in order to exclude transcription from upstream genes into the *tetA-yfp* gene (Figure 1B). Loci *A* and *C* are located approximately in the middle of the right and left replichore respectively. Since we wanted to also include a locus close to the origin of replication, where no pair of divergently oriented genes is present, we chose a locus in the relatively large intergenic region between the co-oriented *rsmG* and *atpI* genes (locus *D*, Figure 1B), a locus previously used for large insertions (Kuhlman & Cox 2010). Locus *B* was chosen based on its vicinity to several insertion sequences (IS).

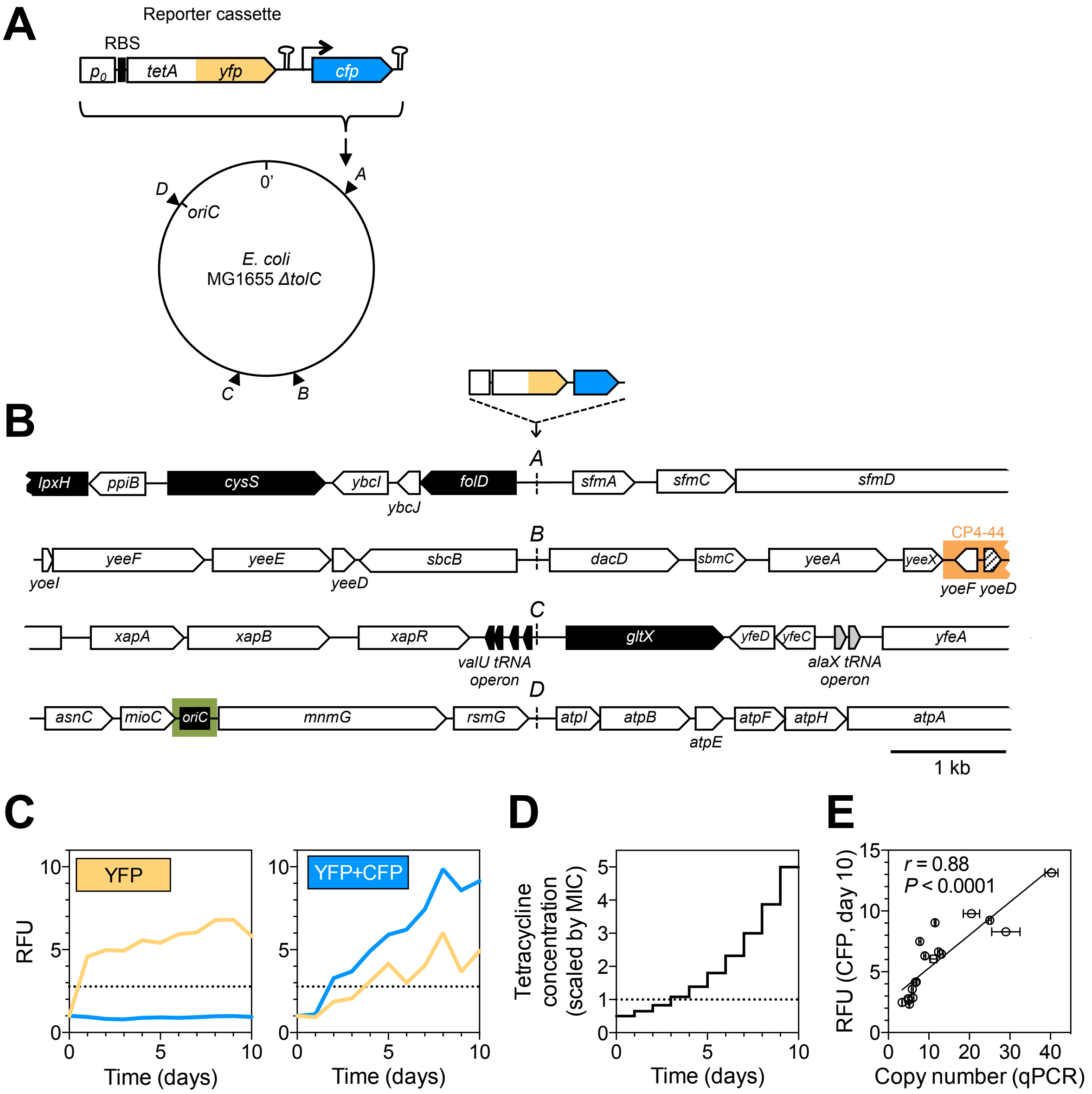


*Figure 1. A dual-fluorescence reporter cassette for real-time tracking of adaptive mutations of different types. (A) Reporter cassette construct for chromosomal insertion. $p_0$ = 188 bp random DNA sequence, RBS = ribosomal binding site, hairpins = transcriptional terminators, tetA-yfp = selected gene, cfp = constitutively expressed amplification reporter. A, B, C, D = intergenic chromosomal insertion loci, oriC = origin of replication. (B) Immediate chromosomal neighborhoods of loci A-D. Black arrows = essential genes. White arrows = non-essential genes. Grey arrows = no essentiality data available. Patterned arrow (yoeD) = pseudogene. Orange = cryptic prophage CP4-44. Green = origin of replication (oriC). Chromosomal neighborhoods of loci B, C, and D are shown reversed with respect to conventional chromosome coordinates, so that the orientation relative to the reporter cassette is shown in the same way for all four loci. Reporter cassette genes are not drawn to scale. (C) Example fluorescence trajectories of rescued populations with YFP or YFP+CFP (amplification) fluorescence phenotype. RFU = relative fluorescence units (see Methods), yellow and blue lines = YFP and CFP fluorescence, dotted lines = threshold for phenotype classification. (D) Increase of tetracycline concentration in ten-day experiment, normalized to strain-specific minimal inhibitory concentration (MIC, dotted line). (E) qPCR validation of CFP fluorescence as an indicator of extent of amplifications. x-axis: tetA-yfp copy number as determined by qPCR on genomic DNA of rescued population with a YFP+CFP fluorescence phenotype. Error bars = SD of technical qPCR triplicates. r is the Pearson correlation coefficient and P its p-value. RFU = relative fluorescence units, line = linear fit.*

We used a *ΔtolC* genetic background in order to constrain the spectrum of possible adaptive mutations to the reporter cassette locus. TolC is an outer membrane porin and an essential part of several *E. coli* multi-drug efflux pumps, which are a frequent target of selection during drug exposure (Li et al. 2015) and which cause low-level intrinsic resistance of *E. coli* to tetracyclines (Sulavik et al. 2001). By employing daily increasing levels of tetracycline (Figure 1D)* and constant daily dilution we created an experimental evolutionary rescue scenario (Carlson et al. 2014), in which populations of ancestral cells rapidly undergo extinction. Rescue from extinction requires the spread of adaptive mutations activating *tetA-yfp* expression in a race against population decline.

The probability of evolutionary rescue depends on the size and decline rate of an unadapted population, and on a combination of rate and fitness effect of adaptive mutations (Martin et al. 2013). We chose selective conditions such that the initial population size and decline rate are approximately equal for all strains. In this way, the probability of rescue (estimated by performing a large number of replicate rescue experiments) is expected to be informative about the strain-specific rate and fitness effect of adaptive mutations of all types. Specifically, we adjusted the tetracycline concentrations used in evolution experiments to strain-specific minimum inhibitory concentrations (MICs), which we measured precisely (Figure 1 – Supplement 1). Given the otherwise isogenic background of the strains, we interpret MICs as a proxy for initial expression of *tetA-yfp*. MIC measurements revealed locus-dependent differences in the initial sensitivity to tetracycline, and all strains showed an increased MIC compared to the cassette-free ancestor, which indicates low baseline expression of *tetA-yfp*. For evolution experiments, we used tetracycline concentrations starting at 50% of the strain-specific MICs (Figure 1D).

We evolved 95 populations of each strain and measured optical density ($OD_{600}$) and fluorescence daily. Populations yielding $OD_{600}$ above a fixed threshold after ten days were regarded as rescued. Rescued populations were assigned to fluorescence phenotypes (YFP or YFP+CFP) based on the increase in fluorescence at the end of the experiment compared to the ancestor (Figure 1C). We performed qPCR on genomic DNA of populations displaying

* The increase of tetracycline concentration was chosen such that not all populations are extinct after the first day and that those surviving until the end of the experiment require a substantial increase in resistance that can only be provided by increased expression of the *tetA*-gene. The choice for a geometric increase of the daily concentrations (and not for e.g. a linear increase) was arbitrarily taken.

increased CFP fluorescence and found a good correlation between CFP fluorescence and the chromosomal copy number of the *tetA-yfp* gene (Figure 1E). Thus, CFP fluorescence is a valid proxy for the extent of high level amplification of the reporter cassette.

### 2.3.2 The chromosomal location of a selected gene has large effects on adaptation

The number of rescued populations differed significantly between strains (Figure 2A), showing that the chromosomal location of the *tetA-yfp* gene is critical for its adaptive potential. No rescue was observed without the reporter cassette (Figure 2 – Supplement 1), and all rescued populations displayed increased YFP fluorescence, suggesting that rescue depended on the presence and overexpression of *tetA-yfp*. To test if increased expression of *tetA-yfp* was indeed causative for rescue, we deleted the reporter cassette genes in single clones isolated from three different rescued populations. Deletions eliminated growth on tetracycline in all three cases (Figure 2B). A minority of populations went extinct despite transiently increased YFP fluorescence (37/290 extinct populations), illustrating how our experimental selection filters for mutations that increase *tetA-yfp* expression above a minimum level. Two sets of replicate experiments yielded qualitatively similar results (Figure 2C), although the number of rescued populations fluctuated considerably between replicates, which likely reflects both technical variability (e.g. in the precise amount of transferred inoculum from day to day) as well as the inherent stochasticity of evolutionary rescue processes. Time-trajectories of $OD_{600}$ and OD-normalized YFP and CFP fluorescence of all evolved populations are available in Supplementary File 1 and fluorescence phenotype classifications in Supplementary File 2.

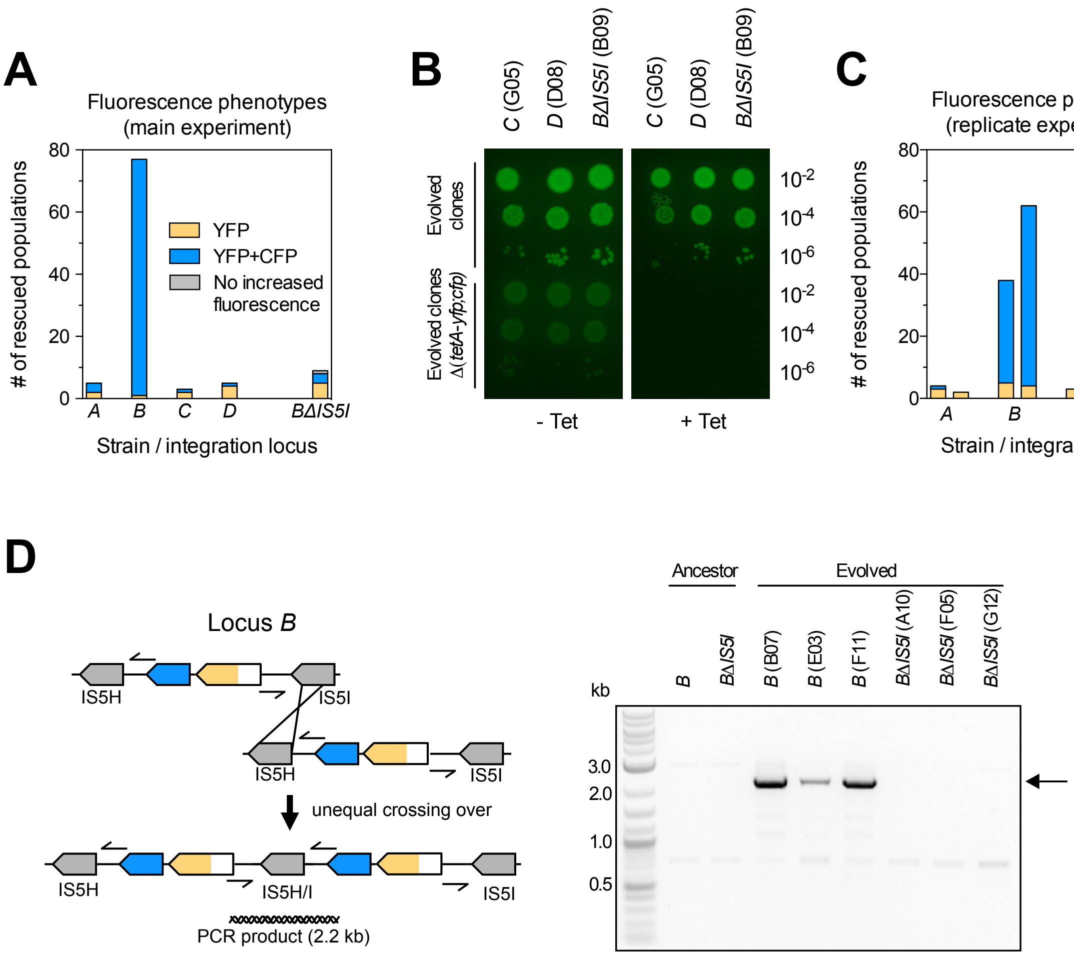


*Figure 2. Large differences in adaptation by amplification depend on flanking homology in the chromosomal neighborhood. (A) Numbers of rescued populations by fluorescence phenotype. The numbers of rescued vs. extinct populations and the distribution of fluorescence phenotypes (YFP or YFP+CFP) differ among strains A, B, C, and D ($p<10^{-16}$ and $p<10^{-7}$, Fisher's exact test). (B) The ability of evolved clones to grow on tetracycline depends on the reporter cassette. Pictures show YFP-fluorescence of cultures spotted at different dilutions on solid medium with and without tetracycline (2.25 µg/mL). Top rows: evolved clones sampled from rescued populations of three different strains. Bottom rows: respective deletion mutants lacking reporter cassette genes. In parentheses: position of the sampled populations on 96-well plates in evolution experiments. (C) Numbers of rescued populations by fluorescence phenotype in two additional replicate sets of evolution experiments. (D) IS5 copies flanking locus B promote duplication. Left: Cartoon showing the position of the reporter cassette between two copies of IS5 (distances not drawn to scale, genes in between omitted) and the putative unequal crossing over-event causing initial duplications. Right: The expected amplicon junction is present in amplifications in strain B, but not in the ancestor or in amplifications in strain BΔIS5I. Arrow: junction PCR product obtained with outward facing primers shown as pointers in the cartoon on the left.*

### 2.3.3 Amplification mediated by flanking homology is a main determinant of neighborhood-dependent adaptation

We next set out to identify which mutation types were responsible for locus-dependent differences in the number of rescued populations. Strain *B* gave the highest number of rescued populations, and 76/77 rescued populations of this strain had reporter cassette amplifications (Figure 2A). Rescue by amplification in the other three strains was rare (Figure 2AC), implying that large differences between strains were related to locus-specific amplification. According to the 'canonical' model, formation of amplifications is limited by the rate at which initial duplications are generated (Romero & Palacios 1997). Rates of spontaneous duplication are elevated between homologous sequences such as rRNA operons or duplicate copies of insertion sequences (IS) due to frequent unequal crossing-over (Anderson & Roth 1981; Andersson & Hughes 2009). We found homologous copies of IS5 at either side of locus *B* (IS5H and IS5I), but no flanking homology in the chromosomal neighborhood of the other three loci. We verified the presence of IS5 at the boundary of the amplicon in rescued *B* populations by obtaining a PCR product of the expected junction in 16/16 tested clones of evolved populations (PCR products of three populations shown in Figure 2D). The junction was undetectable in the ancestor. Deleting one of the two flanking IS (strain *BΔIS5I*) gave highly reduced numbers of rescued populations (Figure 2A) and only a minority (3/9) had increased CFP fluorescence, which was not connected to amplification between the IS5H and IS5I (Figure 2D). These results confirm flanking homology and its effect on gene amplification as a main factor of chromosomal neighborhood on adaptation by increased gene expression.

### 2.3.4 Adaptation involves a broad diversity of mutation types

Given the above result, we expected differences to disappear in the absence of IS and we repeated the evolution experiments with four strains that had the reporter cassette integrated at the same four loci as before, but that are derived from a multiple deletion strain (MDS42) free of all IS elements (Pósfai et al. 2006). MDS42 lacks around 15% of the MG1655 chromosome, including all prophages and many nonessential genes. Apart from the absence of IS-related mutations, the rates of other mutation types in MDS42 are similar to those in MG1655 (Pósfai et al. 2006). Loci *A-D* are not immediately next to genes absent in MDS42, the chromosomal neighborhood at a larger scale however is different between IS-wt and IS-free versions of the strains (Figure 3 – Supplement 1).

Despite the expected absence of frequent amplification of locus *B* in the IS-free genetic background, the fraction of rescued populations was still different among strains ($P = 3 \times 10^{-5}$, Fisher's exact test), and rescue was observed only in strains *B* and *D* (10 and 8 rescued populations, respectively). To explain these remaining differences, we identified candidate rescue mutations in strains with and without IS. Sequencing ~1 kb of DNA upstream of *tetA-yfp* revealed mutations of different types: point mutations (including small insertions and deletions), larger deletions, and insertions of mobile elements (Figure 3AB and Figure 3 – Supplement 2). The relative contribution of the different mutation types to adaptation differed between different chromosomal loci in both IS-containing and IS-free strains ($P=10^{-9}$ and $P=0.003$, Fisher's exact test). In several cases, mutations co-occurred with other mutations or amplifications (colored dots in Figure 3A), suggesting interactions between mutations, some of which we explored in more depth later (section 'Chromosomal neighborhood influences adaptation by affecting the fitness cost of amplifications').

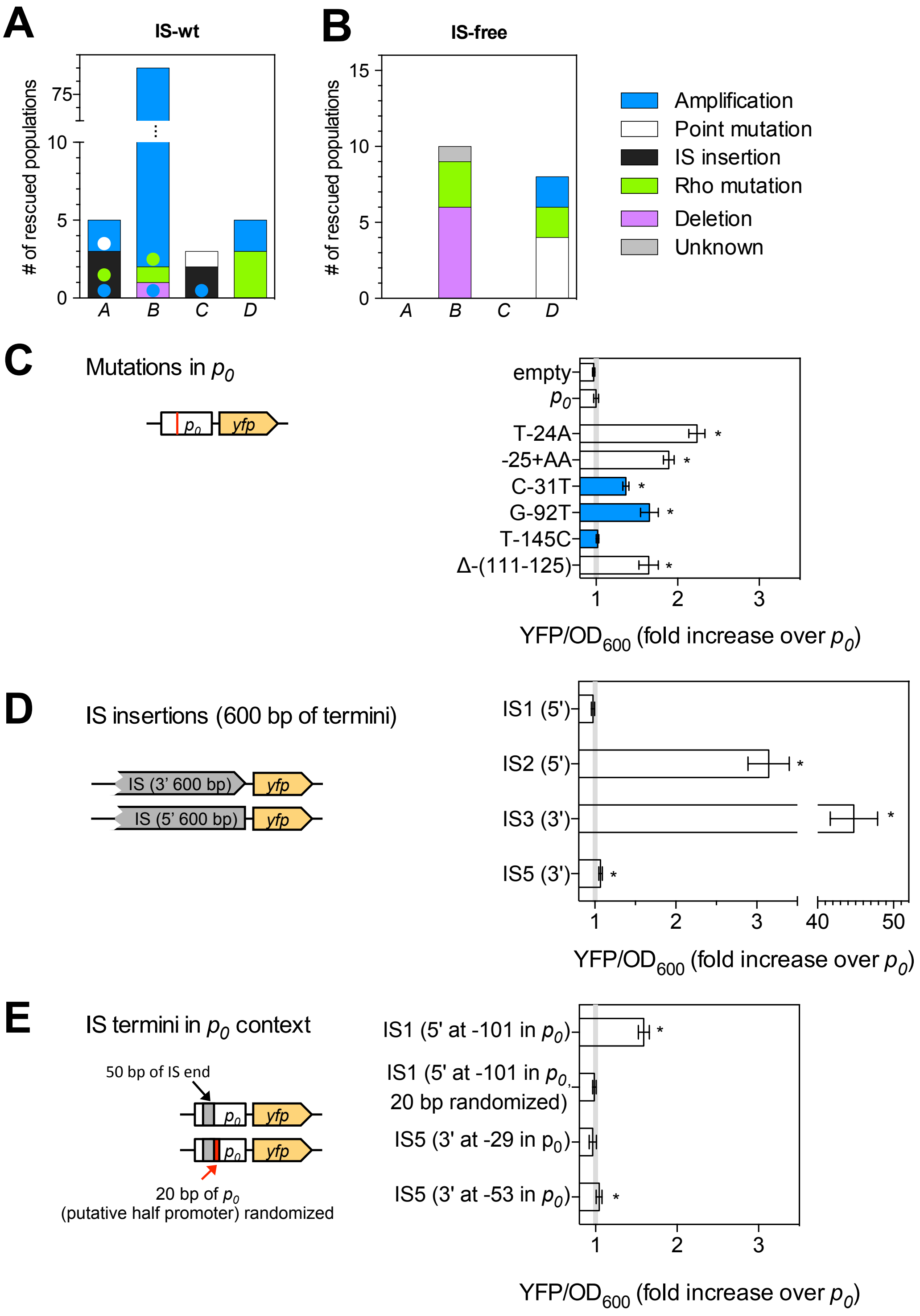


*Figure 3. Adaptation involves a broad diversity of mutation types. Mutation types in rescued populations of IS-wt (A) and IS-free (B) strains. Colored dots = later mutations occurring on top of other mutations (see Methods). Mutation types differ between loci ($p=10^{-9}$ (A) and $p=0.003$ (B), Fisher's exact test). (C–E) Effect of reconstructed point mutations and IS insertions on reporter expression on plasmids. Plasmids contain mutations reconstructed upstream of a ribosomal binding site (not shown) and a yfp reporter gene as shown in cartoons. Empty = auto-fluorescence control (plasmid backbone); $p_0$ = ancestral 188 bp random sequence. Error bars = 95% confidence intervals of six technical replicates. Grey shading: 95% confidence interval of $OD_{600}$-normalized $p_0$ fluorescence. Asterisks = $p<0.05$, two-tailed t-test on mean fluorescence difference in comparison with $p_0$. (**C**) Reporter fluorescence driven by small mutations within $p_0$ (single bp substitutions and small insertions or deletions). Mutation coordinates = distance of mutation to start codon of yfp. Blue bars =*

*mutations that co-occur with amplifications and show overlapping peaks in the sequence chromatogram of evolved clones, indicating presence of mutations only in a subset of copies in an amplification.* ***(D)*** *Reporter fluorescence driven by IS insertions. Plasmids contain the termini of IS which were truncated to 600 bp. 5' and 3' refers to the direction of the IS-contained transposase gene. IS2 and IS3 drive strong fluorescence of yfp in the plasmid context; IS1 and IS5 do not.* ***(E)*** *Reporter fluorescence driven by IS in the precise sequence context of $p_0$. IS1, but not IS5, contains a partial promoter whose activity depends on the adjacent sequence in $p_0$. Numbers in parentheses = distance between insertion point and the yfp start codon. 'rnd' = random shuffling of 20 bp of $p_0$ downstream of the IS1 insertion point.*

We then continued to identify the mutation types responsible for the remaining differences in adaptation among strains, independent of neighborhood-dependent amplifications as described above. In order to test the effect of mutations on downstream expression independent of chromosomal locus, we constructed *yfp* reporter plasmids with all mutations found within the $p_0$ region of clones from rescued populations of the first replicate set of evolution experiments (IS-wt strains, IS-free strains and strain *BΔIS5I*, Figure 3CDE). Five of six small mutations altering the sequence of $p_0$ increased *yfp* fluorescence in plasmid reconstructions (Figure 3C), presumably by increasing the affinity of RNA polymerase to $p_0$. One mutation (T-145C) did not affect fluorescence and likely did not contribute to adaptation. Instead, rescue of the respective population, which also displayed a YFP+CFP fluorescence phenotype, likely depended on amplification alone. In contrast, two other point mutations identified in conjunction with amplifications (C-31T and G-92T), did increase reporter fluorescence on plasmids, providing examples of a combined beneficial effect of amplifications and additional mutations. Two of four insertions sequences that we had found inserted into $p_0$ increased reporter fluorescence on plasmids greatly (IS2 and IS3, Figure 3D), which is consistent with the delivery of outward-facing promoters within the termini of IS (Mahillon & Chandler 1998). The two other IS (IS1 and IS5) had no or no strong effect on plasmid reporter fluorescence. Since some IS have been reported to contain partial outward-facing promoters that can drive downstream expression after insertion next to a resident complementary partial promoter site (Mahillon & Chandler 1998), we tested IS1 and IS5 in the precise sequence context of $p_0$ in which these IS were found in evolution experiments (Figure 3E). In this sequence context, IS1 indeed increased reporter fluorescence, which depended on the 20 bp downstream of the insertion point within $p_0$ (Figure 3E), consistent with the delivery of a half-promoter within the terminus of this IS. Insertion of IS5, which we repeatedly observed in evolution experiments, had very weak, but significant effects on downstream fluorescence on plasmids (Figure 1DE). To confirm the adaptive role of upstream IS5 insertions in the evolution experiments, we transduced one of the observed upstream IS5 insertions into the ancestral background, which restored growth on tetracycline as well as a marked increase in YFP fluoresence (Figure 3 – Supplement 3). Thus, in the chromosomal context, IS5 does increase expression of downstream genes, possibly due to effects on DNA bending (Zhang & Saier 2009), which may not be recapitulated on the plasmid reconstruction. These results illustrate the diverse ways in which IS can adaptively affect gene expression, both dependent (IS1, IS5) and independent (IS2, IS3) of the insertion context. Given the reporter plasmid results and the fact that the same $p_0$ sequence is part of the reporter cassette at all four chromosomal loci, point mutations and IS insertions likely were not responsible for the observed differences in the frequency of rescue between strains that are not explained by amplifications.

### 2.3.5 Properties of upstream genes determine the availability of two different types of adaptive promoter co-option mutations

Whole genome sequencing of clones from three rescued populations with neither upstream genetic changes nor amplifications (Figure 3 – Supplement 4), as well as subsequent screening of other rescued populations, revealed another candidate type of adaptive mutations, which altered the protein sequence of *rho* (Figure 4 – Supplement 2). Unlike mutations of the other types, *rho* mutations occurred in *trans* with respect to the reporter cassette. The *rho* gene of *E. coli* is an essential gene that encodes a transcriptional termination factor estimated to be required for termination at around half of all termination sites in *E. coli* (Ciampi 2006). Contrary to point mutations and IS insertions, which we found upstream of all four loci (Figure 3A and Figure 3 – Supplement 5), Rho mutations and also upstream deletions were only found in evolved clones of strains with the reporter cassette at locus *B* or *D*, with one exception of a Rho mutation co-occurring with an upstream IS insertion in strain *A*. Thus, upstream deletions and Rho mutations provide candidates for locus-dependent adaptive mutations. Comparing the upstream neighborhood of the four different loci revealed the basis of this locus-dependency (Figure 4A and Figure 4 – Supplement 1). The orientation and expression of upstream transcripts as determined in a different study (Conway et al. 2014) suggests that in strains *B* and *D*, active upstream promoters were co-opted to *tetA-yfp*, either by deletion of intervening genes, or by compromising Rho-dependent termination by partial-loss-of-function mutations in Rho that cause transcriptional read-through into *tetA-yfp*. At loci *A* and *C*, such adaptive mutations were not available because of two kinds of constraints from neighboring genes: either intervening genes were essential (constraining adaptive deletions, Figure 4A), or no upstream Rho-terminated transcripts were present (constraining adaptive Rho mutations, Figure 4A).

Since active transcripts shown in Figure 4A were experimentally determined under conditions different from our evolution experiments (Conway et al. 2014), and classification of termination sites as intrinsic or Rho-dependent was done only computationally (Kingsford et al. 2007; Conway et al. 2014), we experimentally assessed the effect of Rho mutations on transcriptional read-through across candidate upstream terminators at all four loci under experimental conditions approximating those in evolution experiments. We first confirmed the neighborhood-dependent effect of two different Rho mutations (S153F and M416I) on the phenotype of interest, i.e. tetracycline resistance, by transduction into the ancestral IS-wt strains, which are isogenic except for the position of the reporter cassette (Figure 4B). Consistent with the presence of upstream Rho-terminated transcripts as shown in Figure 4A, an increased tolerance of Rho-mutants to tetracycline was observed only in strains with the reporter cassette at loci *B* and *D*, matching our observation that Rho-mutants were only found in rescued populations of these strains. We then performed PCR on cDNA prepared from a Rho-wt strain and a Rho mutant (M416I) strain grown in sub-inhibitory tetracycline (Figure 4C). We obtained PCR products consistent with read-through across candidate terminators upstream of locus *B* (downstream of *yeeD*) and locus *D* (*mnmG*), but not upstream of locus *A* (*cysS*) and locus *C* (*xapR*). A read-through transcript at locus *D* was detectable even in the Rho-wt background, which offers an explanation for the higher initial TetA-YFP expression observed in strain *D* (Figure 1 – Supplement 1). Mutations found in rescued populations of additional replicate experiments (fluorescence phenotypes in Figure 2B) are consistent with the above constraints on promoter co-opting mutations (Figure 3 – Supplement 5). Thus, upstream deletions and *trans* mutations that compromise transcriptional termination are mutation types that depend on the chromosomal neighborhood of the gene under selection. Specifically, the orientation, expression, essentiality and termination mode of neighboring genes shape the fitness effect of these promoter co-option mutations.

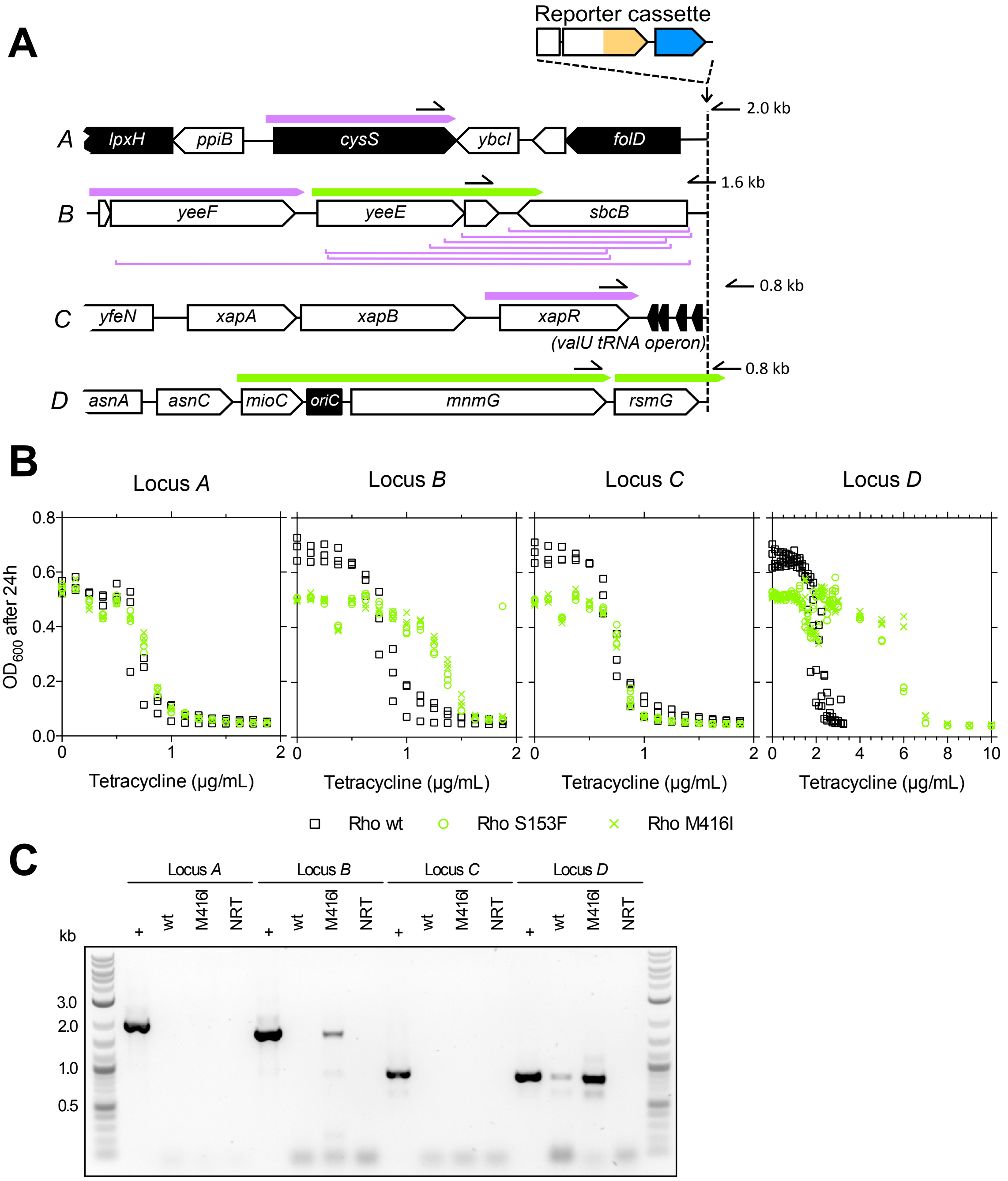


*Figure 4. The fitness effect of promoter co-opting deletions and Rho-mutations depends on properties of upstream neighboring genes. (A) Genes and transcripts upstream of loci A, B, C, and D. Promoters of intrinsically terminated transcripts (purple) can be co-opted by deletions (purple brackets) if no essential gene (black arrows) is deleted. Promoters of Rho-terminated transcripts (green) can be co-opted by deletions or by partial loss-of-function mutations in Rho. Only putatively expressed transcripts oriented toward the reporter cassette are shown (all transcripts in Figure 4—Figure supplement 1). Pointers and numbers on the right = position and size of PCR products shown in (C). (B) Tetracycline dose-response curves of strains with wt (black squares) or transduced mutant Rho (green circles = S153F, green crosses = M416I). Final $OD_{600}$ after 24 hr (platereader units) was measured in three biological replicates. Rho mutants are more tolerant to tetracycline only with the reporter cassette at loci B and D. (**C**) Read-through transcripts spanning upstream terminators in a Rho-mutant background are detectable at loci B and D, but not at loci A and C. Bands show PCR products obtained from genomic DNA (+ control) or cDNA from a Rho-wt or Rho mutant (M416I) strain grown with sub-inhibitory levels of tetracycline. Positions of used primers as indicated in (**A**). NRT = negative control (no reverse transcriptase).*

### 2.3.6 Chromosomal neighborhood influences adaptation by affecting the fitness cost of amplifications

As seen from promoter co-opting mutations, chromosomal neighborhood may affect the adaptive potential of a gene by influencing not only mutation rates (as flanking homology does for duplications that can expand into amplifications), but also mutation fitness effects. We next asked if this applies to amplifications as well. Due to the instability of amplifications and related difficulties in detecting them, quantifying the fitness effect of amplifications is laborious (Adler et al. 2014) and has so far not been done on a genome-wide scale. The benefit of amplifying a selected gene is counteracted by a cost that arises in part due to dosage imbalances in the co-amplified neighboring genes. This cost limits the ability of amplifications to effectively expand at the population level as selection increases, an ability that comes from high rates of expansion of amplifications at the level of the individual chromosome by homologous recombination. The probability of an amplification to contain a costly gene is expected to increase with the length of the amplicon.

We used two-color fluorescence data to extract information on amplification cost and its effect on adaptation. For validating this approach, we used two strains (strain *BΔIS5I*, and the newly created strain *E*, Figure 5A) that we predicted to form amplifications of higher and lower cost respectively when compared to the IS-containing strain *B*, which serves as reference. The IS5 deletion in strain *BΔIS5I*, which reduces the rate of duplications that kick-start amplifications (see above), is also expected to increase the fitness cost of amplifications, since amplicons may be larger than the 35 kb between IS5I and IS5H. In strain *E*, we placed the reporter cassette between two copies of IS1, where duplications are expected to form frequently, and amplifications, due to small amplicon size (11 kb), are expected to expand at low cost. In our experiment, if cost is negligible, amplifications are expected to expand continuously as tetracycline selection increases, resulting in rescue. In this case, YFP+CFP fluorescence increases in correlation with the level of tetracycline selection (Figure 5B, left). If amplifications are cost-limited, two outcomes are possible: (i) amplifications fail to expand beyond a certain level lower than that required for rescue, resulting in extinction – if the level of expansion before extinction is high enough, it will appear as a transient increase of CFP fluorescence in the fluorescence trajectories of extinct populations (Figure 5B, middle), (ii) amplifications allow rescue through interaction with other adaptive mutations that increase *tetA-yfp* expression – resulting in increased YFP/CFP fluorescence ratios in rescued populations (Figure 5B, right). We compared the numbers of extinct vs. rescued populations with (transiently) increased CFP fluorescence and found, as expected, that amplifications in strain *BΔIS5I* had a significantly higher extinction risk than amplifications in the reference strain *B* (Table 1), and rescued amplifications had significantly higher final YFP/CFP ratios (Figure 5C), confirming that low numbers of rescue in strain *BΔIS5I* were in part due to amplification costs. Contrariwise, populations with increased CFP fluorescence in strain *E* never went extinct (Table 1) and had consistently low final YFP/CFP ratios (Figure 5C), indicating the absence of a cost limitation.

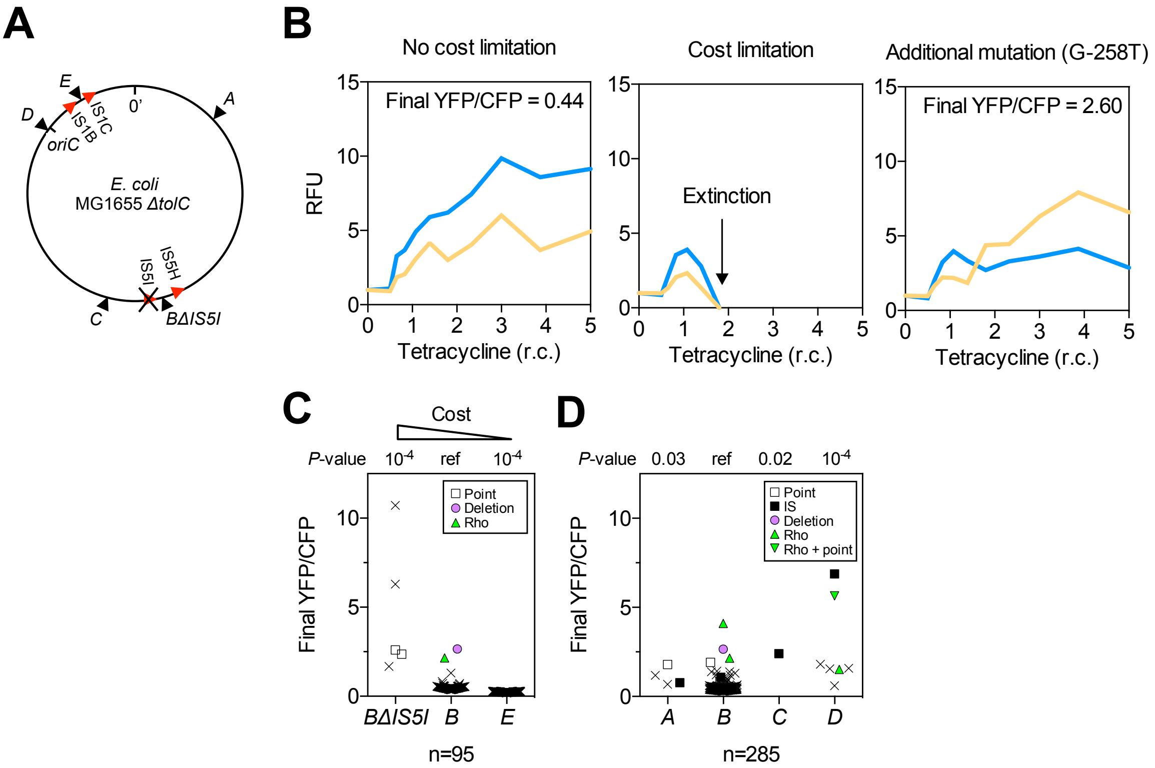


*Figure 5. Chromosomal neighborhood influences the fitness cost of amplifications. (A) Chromosomal location of reporter cassette in strains BΔIS5I and E (IS distances not drawn to scale). (B) Example fluorescence trajectories. Left: low-cost amplifications expand in correlation with the increase in tetracycline concentration over 10 days. Middle: Cost-limited amplifications fail to expand at higher tetracycline concentrations resulting in extinction. Right: Amplifications can escape extinction in combination with other mutations increasing tetA-yfp expression, resulting in higher final YFP/CFP. RFU = relative fluorescence units (see Materials and methods), r.c. = relative concentration as multiples of MIC. (C) Final YFP/CFP ratios of rescued amplifications in strains expected to have a higher (strain BΔIS5I) or lower (strain E) cost of amplifications compared to strain B. n = initial number of replicate populations used for analysis. Crosses = populations rescued by amplifications without additional mutations. Other symbols = secondary mutations (see legend). p-values: permutation tests in comparison with strain B. (D) Final YFP/CFP ratios of rescued amplifications in strains A-D. n = 285 includes replicate evolution experiments to increase statistical power. Symbols and p-values as in (C).*

*Table 1. Differences in amplification cost indicated by the extinction risk of populations with amplifications. Populations with amplifications of higher (strain BDIS5I) or lower (strain E) expected cost of amplifications have a higher or lower risk of becoming extinct, respectively. n = initial number of replicate populations used for analysis (n = 285 includes replicate evolution experiments to increase statistical power), 'Extinct' and*

*'Rescued' = numbers of extinct and rescued populations with amplifications as indicated by (transiently) increased CFP fluorescence (see Materials and methods), sample odds ratio compared to strain B, p-values: 2 x 2 Fisher's exact test.*

| n | Strain | Populations with (transiently) increased CFP fluorescence | | Sample Odds Ratio | *P* |
|---|---|---|---|---|---|
| | | Extinct | Rescued | | |
| 95 | *BΔIS5I* | 12 | 5 | 10.1 | $10^{-4}$ |
| | *B* | 18 | 76 | 1 (ref) | - |
| | *E* | 0 | 95 | 0 | $10^{-6}$ |
| 285 | *A* | 8 | 4 | 5.8 | $10^{-3}$ |
| | *B* | 58 | 168 | 1 (ref) | - |
| | *C* | 0 | 1 | 0 | n.s. |
| | *D* | 0 | 7 | 0 | n.s. |

Having validated extinction risk (Table 1) and final YFP/CFP ratios (Figure 5C) as indicators of amplification cost, we tested if neighborhood-dependent amplification costs had affected adaptation in strains *A*, *C*, and *D*. A significantly elevated extinction risk of populations with amplifications in strain *A* (Table 1), and significantly elevated final YFP/CFP ratios in connection with diverse additional mutations in strains *A*, *C*, and *D* (Figure 5D), support that the costs of amplifications in these strains were higher compared to strain *B*. Thus, amplification costs represent another neighborhood-dependent constraint on adaptation. In this perspective, the availability of neighborhood-dependent promoter co-option mutations, the most prevalent non-amplification mutation types at loci *B* and *D*, is an important determinant of adaptive potential not only in itself, but also in the interaction with amplifications.

### 2.3.7 Chromosome neighborhood effects on adaptation in a single-step plating experiment

We next investigated whether our observations of chromosomal neighborhood effects on adaptation transfer to different selective conditions. In particular, we tested the possibility that differences in rescue between strains were due to different population sizes and thus different chances for beneficial mutations to occur, rather than due to different mutation rates or fitness effects as we propose. Our experimental design corrects for population size differences between strains at the first day of selection (Figure 1 – Supplement 1, and first section of the results part), but not necessarily for population size differences at later days of the experiments. Therefore, we performed single-step plating experiments, in which approximately the same numbers of cells are plated for every strain. In these Luria-Delbrück-type experiments, we plated replicate cultures grown under non-selective conditions on solid medium with tetracycline at two-fold MIC levels. We scored the number of colonies on each plate after two days, when clearly visible colonies first appeared. These early colonies are expected to result mostly from pre-plating single-step mutations that increase *tetA-yfp* expression (point mutations, IS insertions, and promoter co-option mutations). As in evolution experiments, colony numbers in strains *B* and *D* were higher than in strains *A* and *C* (Figure 6, left), for both IS-wt and IS-free genetic backgrounds. This result is consistent with neighborhood-dependent availability of promoter co-option mutations as observed also in evolution experiments. High CFP fluorescence, indicative of amplifications, was observed only in a small fraction of early colonies (34 of 1661 across all strains and plates). During longer incubation, the number of colonies on plates of IS-wt strain *B* increased steadily (Figure 6 – Supplement 1) and almost all of these later colonies (1229/1304 on ten plates) showed high CFP fluorescence. Since tetracycline is bacteriostatic rather than bactericidal, the appearance of these late colonies can be explained by a continuous process of reporter cassette amplification expansion and increasing growth rates after plating on selective medium, starting from frequent duplications that have a slight growth advantage over single-copy cells (Andersson 1998). After five days, colony counts on plates were qualitatively similar to rescue frequencies in evolution experiments, with IS-wt strain *B* giving the highest number of colonies (Figure 6, right). In all other tested strains, late colonies appeared at much lower rates (Figure 6 – Supplement 1) and did not show high CFP fluorescence in most cases (Figure 6, right, and Figure 6 – Supplement 2), reflecting the minor role of amplification in strains that lack flanking homology in the chromosomal neighborhood of the selected gene. The consistency between liquid-culture evolutionary rescue experiments and plating experiments supports that strong effects of chromosomal neighborhood on the rate and fitness effect of adaptive mutations extend to different selective regimes.

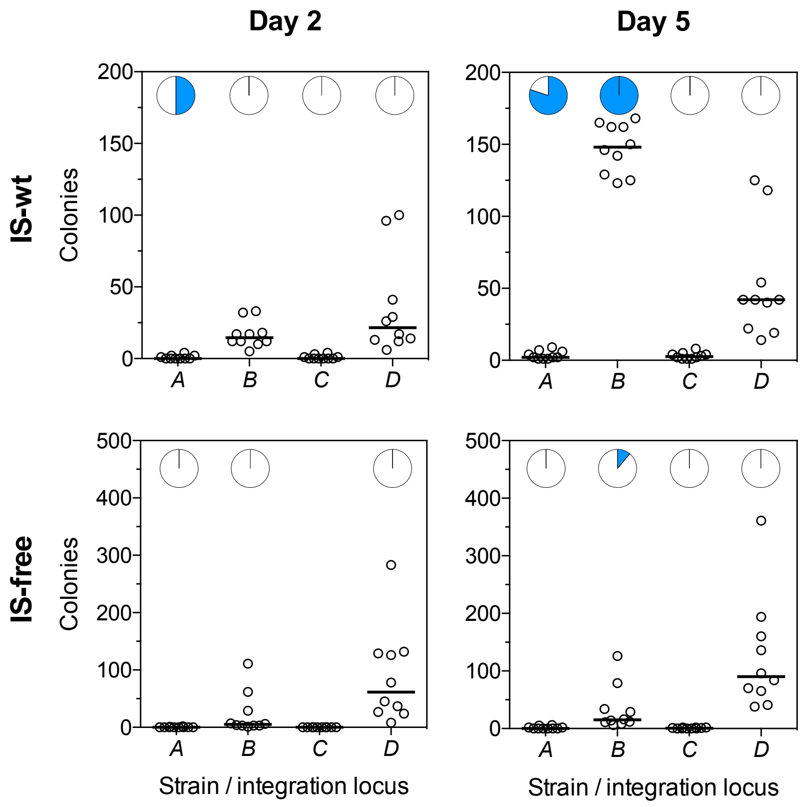
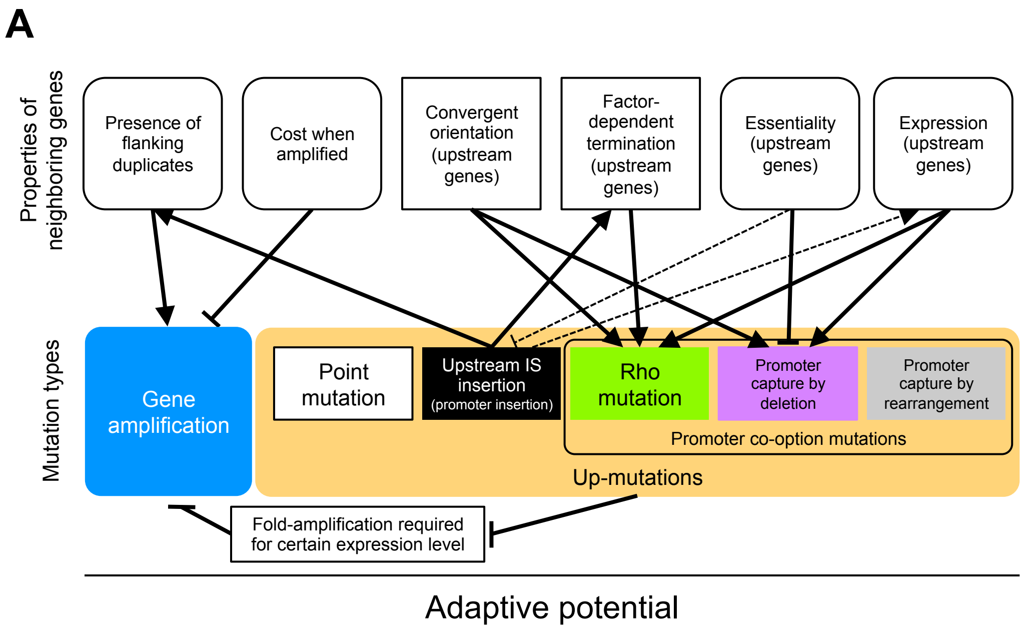


*Figure 6. Tetracycline-resistant mutants arising in a single-step plating experiment. For each strain (top panels = IS-wt, bottom panels = IS-free), 10 replicate cultures grown in the absence of tetracycline were plated on agar with tetracycline concentration two times the strain-specific MIC. Left: Colony counts after 2 days of incubation. Right: Colony counts after 5 days of incubation. Horizontal lines show the median colony number from 10 replicate plates. Pie charts = fraction of plates in which a single tested colony appearing at day 2 (left) or at days 4–5 (right) showed high CFP fluorescence indicative of amplification (Figure 6—Figure supplement 2).*

## *2.4 Discussion*

Our results reveal a complex genetic basis of strong effects of chromosomal position on the adaptive potential of a specific gene (Figure 7A). By combining time-resolved fluorescence data from the reporter cassette and end-point genetic analysis, we demonstrate how the relative contribution of previously known mutation types to adaptation (Figure 7A, bottom row) differs between chromosomal loci, how these differences arise, and how a layer of complexity is added by the interaction of mutation types. Thus, the concept of a one-dimensional mutation rate and a focus on point mutations can be misleading (Martinez & Baquero 2000), even for the simple case of adaptation by increased expression of a single gene. Instead, the adaptive potential of a given gene is a system-level property shaped by the local chromosomal genetic neighborhood. Consequently, the organization of genes on a chromosome is both cause and consequence of evolutionary change.


A
Properties of neighboring genes
Presence of flanking duplicates
Cost when amplified
Convergent orientation (upstream genes)
Factor-dependent termination (upstream genes)
Essentiality (upstream genes)
Expression (upstream genes)
Mutation types
Gene amplification
Point mutation
Upstream IS insertion (promoter insertion)
Rho mutation
Promoter capture by deletion
Promoter capture by rearrangement
Promoter co-option mutations
Up-mutations
Fold-amplification required for certain expression level
Adaptive potential

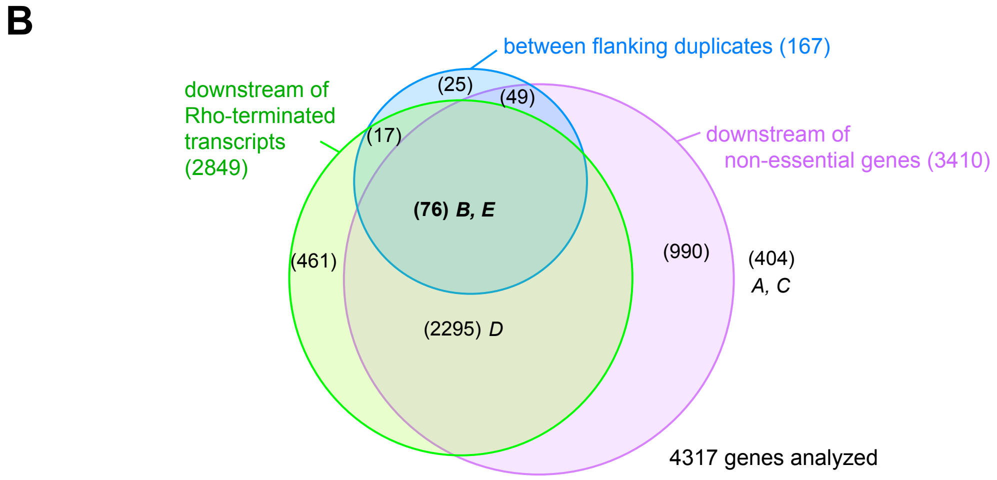


*Figure 7. The adaptive potential of a gene under selection for increased gene expression as a complex function of properties of neighboring genes that affect and are affected by mutations of diverse types. (A) Top row: Properties of neighboring genes that we identify as main determinants of the adaptive potential of a gene given its chromosomal neighborhood. Round corners indicate 'dynamic' properties that may be environment-dependent or subject to change over short evolutionary timescales. Bottom row: Different mutation types causing increased expression of a gene. Solid arrows: Effects and interactions shown or suggested by data in this study. Dashed arrows: Other effects and interactions that are likely to exist. Pointed arrowheads indicate a positive effect, T-bar ends indicate a negative effect. A sentence equivalent of each arrow is given in Figure 7—source data 1. As a sum of the above interactions, the adaptive potential of a gene emerges as a system-property. (B) Classification of chromosomal neighborhoods of E. coli genes according to adaptive potential. The chromosomal neighborhood of 4317 genes of E. coli MG1655 was assessed using published information on the position of promoters and terminators (Conway et al., 2014) and gene essentiality (see Methods for details). Numbers in parentheses = genes belonging to respective sets or intersections of sets. Genes in the intersection of all three circles (boldface) are expected to have the highest adaptive potential based on their chromosomal neighborhood. Loci A-E of this study are placed in the respective areas of the diagram.*

Importantly, the effects that we describe arise from several properties (Figure 7A, top row) of different genetic elements that are present in the vicinity of the selected gene, rather than from more global factors such as distance to the origin of replication or chromosome macro-domain organization (Bryant et al. 2014). Therefore, we propose to refer to them as 'chromosome neighborhood effects' that determine the evolution of gene expression, as opposed to 'chromosome position effects' that modulate gene expression *per se* (Bryant et al. 2014; Levis et al. 1985; Akhtar et al. 2013).

### 2.4.1 Different mutation types interact to cause neighborhood-dependent differences in adaptive potential

In our experiments, chromosomal neighborhoods facilitate or constrain adaptation mainly via amplification and promoter co-option mutations, by affecting the rate of mutations (duplication-amplification) or the fitness effects of mutations (promoter co-option mutations and amplifications). For gene amplification, a strong effect of flanking homology as provided by IS, which are often present in multiple copies, has been known for a long time (B. C. Peterson & Rownd 1985; Andersson & Hughes 2009). Our data confirm that if flanking homology is present at a given locus, amplification is the main response to selection for increased gene expression. For loci lacking nearby flanking homology, which depending on the distribution of IS elements on a chromosome may be the majority of loci (Boyd & Hartl 1997; Green et al. 1984), our data show that adaptation by amplification is limited on the level of duplication rate and fitness cost. For these loci, differences in the adaptive potential are largely due to the different availability of deletions and mutations compromising transcriptional termination, both of which co-opt upstream promoters to the selected gene. Such mutations also act in concert with amplifications and can alleviate amplification cost limitations by lowering the required fold-amplification to reach a certain level of expression of the selected gene (Figure 7A).

The multitude of mutations discovered in the termination factor Rho suggests that the function of this protein may be more 'tunable' than expected from it being an essential gene in *E. coli*. Our results may suggest that adaptation via *trans* mutations in Rho with potentially large pleiotropic effects is more likely than via local mutations that compromise upstream terminators in *cis*. Given that the sequence-dependence of Rho-dependent termination is poorly understood (Ciampi 2006), there is no clear expectation of the nature and target size of mutations that would compromise Rho-dependent termination in *cis*. This makes it difficult to compare adaptation via mutations affecting Rho-dependent termination in *cis* versus *trans*. The adaptiveness of *trans* mutations in Rho despite their pleiotropic effects is supported by a previously characterized single amino-acid substitution in Rho, which was found to have large-scale effects on the *E. coli* transcriptome and to confer higher fitness in several environments (Freddolino et al. 2012). We found substitutions at 22 different amino acid residues mapping to various regions of the Rho protein structure (Skordalakes & Berger 2003) (Figure 4 – Supplement 2 and Figure 4 – Supplement 2 – Source Data 1), which largely expands the number of Rho residues found mutated in evolution experiments (Conrad et al. 2011). This supports the idea that operons delimitated by factor-dependent terminators may be rather fluid, providing a large source of variation for adaptation to changing environments. It remains to be seen whether different Rho alleles, by revealing 'hidden' transcriptional variation, serve as capacitors of adaptation (Masel 2013) beyond laboratory evolution experiments.

### 2.4.2 Assessing properties of neighboring genes to infer the adaptive potential of a gene under selection

For both amplifications and promoter co-opting mutations, the influence of the chromosomal neighborhood arises mechanistically from several simple properties of neighboring genes – their expression, orientation, transcriptional termination, essentiality, the presence or absence of flanking gene duplicates – and from the cost of neighboring gene co-amplification (Figure 7A, top row). If these properties are known at a genomic scale, inferring a chromosome-wide 'map of adaptive potential' becomes conceivable. An understanding of adaptive potential may help assess the risk of resistance evolution via overexpression of preexisting chromosomal genes (as opposed to acquisition by horizontal transfer). Clearly, some properties of neighboring genes can be assessed on a genome-wide scale more easily (e.g. gene orientation) than others (e.g. gene essentiality or cost of genes when amplified). Once it becomes feasible to acquire data on all the main factors shaping adaptive potential, this data may improve efforts to predict specific adaptations.

As a first step towards this goal, we used published information on gene essentiality, and promoter and terminator locations (Conway et al. 2014) to assess how many of *E. coli* genes (strain MG1655) are expected to reside in a chromosomal neighborhood associated with high adaptive potential (Figure 7B). Based on the most simply assessable properties (colored circles in Figure 7B), the chromosomal neighborhood of most genes (2295/4317) is expected to have a medium adaptive potential, comparable to that of locus *D* from this study.

### 2.4.3 Adaptive potential as a dynamic property

Importantly, some properties of chromosomal neighborhoods are dynamic (rounded boxes in Figure 7A) – gene essentiality (Baba et al. 2006) and expression can be environment-dependent, and transposition causes rapid turnover of mobile element positions (Sawyer et al. 1987; Wagner 2006). Therefore, the classification of chromosomal neighborhoods of genes according to adaptive potential as in Figure 7B needs to be understood as a snapshot in time reflecting particular conditions. Also, how adaptive potential translates into the actual likelihood of adaptation depends on population parameters and the precise selection scenario.

On evolutionary timescales, the dynamics of chromosomal neighborhood properties would rapidly degrade signals that neighborhood-dependent evolution leaves in genome sequences. Nevertheless, neighborhood-dependent evolution could offer mechanistic explanations for phenomena observed in genomic data such as operon organization (Reams & Neidle 2004; Lawrence & Roth 1996), reductive genome evolution by promoter capture-deletions as suggested previously (Lind et al. 2015), or the chromosomal position of horizontally transferred genes (Touchon et al. 2009). Since horizontally transferred genes carrying selective functions are often silenced after initial integration (Navarre 2006; Cardinale et al. 2008), they depend on activating mutations to play out their benefit to the host and become stably maintained in the host chromosome. Thus, the evolutionary fate of horizontally transferred genes will be shaped by the new chromosomal neighborhood they find themselves in. For example, a drug resistance gene entering the genome at loci *B* or *D* via horizontal transfer will be more likely to enable survival of the host under drug selection, compared to insertion at loci *A* and *C*, both because of higher initial expression and the higher adaptive potential associated with these loci as described here. The common association of horizontally acquired genes with flanking mobile elements as in complex transposons and genomic islands (Dobrindt et al. 2004) may not only reflect the high transferability of such configurations, but also their high amplifiability, which may be of particular relevance for mis-expressed foreign genes.

### 2.4.4 Chromosomal neighborhood effects beyond prokaryotes

Although our results reflect many specifics of prokaryote genome organization, the importance of promoter-capture mutations (ar-Rushdi et al. 1983), modulation of transcriptional read-through (Grosso et al. 2015) and gene amplification (Cole et al. 1992; Gajduskova et al. 2007) extends to cancer evolution and cases of rapid adaptation in higher organisms (Devonshire & Field 1991). This implies that chromosomal neighborhood effects on evolution may be of wider significance and they could be investigated with similar reporter-based methods.

## *2.5 Materials and Methods*

### 2.5.1 Materials

Unless noted otherwise, we obtained chemicals from Sigma-Aldrich (St. Louis, Missouri) and enzymes from New England Biolabs (Ipswich, Massachusetts). Evolution experiments and phenotyping tests were done in in M9 medium supplemented with 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, and 0.2% glucose and 0.2% casein hydrolysate as carbon sources (M9CG medium), unless noted otherwise. A list of oligonucleotides, strains, and plasmids is available in Supplementary File 3.

### 2.5.2 Construction of the reporter cassette

The reporter cassette (*$p_0$*-RBS-*tetA-yfp-$p_R$-cfp*) was assembled on a plasmid using a combination of standard cloning techniques, ligation chain reaction (Rouillard et al. 2004), and fusion PCR. For the *$p_0$* sequence upstream of *tetA-yfp*, we generated a random 188 bp nucleotide sequence matching the average GC content of *E. coli* (CCGGAAAGACGGGCTTCAAAGCAACCTGACCACGGTTGCG CGTCCGTATCAAGATCCTCTTAATAAGCCCCCGTCACTGTTGGTTGTAGAGCCCAGGACGGGTTGGCCAGATGTG CGACTATATCGCTTAGTGGCTCTTGGGCCGCGGTGCGTTACCTTGCAGGAATTGAGGCCGTCCGTTAATTTCC). We synthesized the sequence from oligonucleotides in a ligation chain reaction. The *tetA* sequence was taken from strain TKC (Sharan et al. 2009), and the *yfp* gene from plasmid pZA21-*yfp* (Lutz & Bujard 1997). At the fusion point, we placed a 3xGS linker peptide between the C-terminus of TetA and the N-terminus of YFP. Between *$p_0$* and the start codon of *tetA-yfp* is a sequence containing a restriction site and a ribosomal binding site (GTCGACAGGAGGAATTCACC). We placed the *$p_0$-tetA-yfp* sequence on plasmid pAH81-FRT-*cfp* (Haldimann & Wanner 2001), upstream of the chloramphenicol resistance gene and the terminator-flanked *$p_R$-cfp* gene. *$p_R$* is a strong constitutive promoter originating from phage λ. We sequenced the full length of the reporter cassette on the resulting plasmid, pMS7. Replication of the pMS7 plasmid depends on the Pir protein and the plasmid was propagated in a *pir*-containing version of strain DH5α.

### 2.5.3 Strain construction

We moved the *ΔtolC::kan* allele from *E. coli* strain JW5503-1 into strain MG1655 using P1 transduction. For the IS-free genetic background, the same *ΔtolC::kan* allele was introduced into strain MDS42 (Pósfai et al. 2006) by recombineering (L. C. Thomason et al. 2014) with pKD13 (Datsenko & Wanner 2000) as PCR template. *kanR* cassettes were removed using plasmid pCP20 (Datsenko & Wanner 2000). We inserted the reporter cassette from plasmid pMS7 into the two *ΔtolC* strains by recombineering. Precise insertion points are given in Figure 1 – Source Data 1. All reporter cassette genes point towards the terminus of replication. Recombinants were selected on LB agar with chloramphenicol (10 µg/mL). The chloramphenicol marker was subsequently removed (Datsenko & Wanner 2000). We confirmed the presence of the full-length single copy insertion by PCR and verified the sequence of $p_0$-*tetA-yfp* by sequencing. The presence of functional $p_R$-*cfp* was confirmed by observing fluorescence. To obtain strain *BΔIS5I*, the *camR* cassette from pKD3 (Datsenko & Wanner 2000) was recombineered into the IS5I element of strain *B*. Recombinants were selected with choloramphenicol (10 µg/mL) and confirmed by PCR. Deletion of the reporter cassette genes in evolved clones was done by recombineering the *kanR* cassette of pKD13 into the reporter cassette such that the coding regions of both *tetA-yfp* and *cfp* were disrupted. Deletions were confirmed by absence of fluorescence and PCR with flanking primers (Figure 2 – Supplement 2). For P1 transduction of *rho* mutations, we first transduced mutations S153F and M416I from rescued clones of populations of strain *D* into MG1655. As selective marker, we used a *kanR* cassette that we had inserted upstream of *rho* by recombineering. After sequence verification, we transduced *rho* mutations into IS-wt strains *A-D*.

### 2.5.4 MIC measurements and dose-response curves

Strains were pre-grown for 16 h in M9CG medium without tetracycline and transferred to 96-well plates (200 µL/well). From there, we pin-diluted cultures with a VP408 pin replicator (V&P Scientific, San Diego, California, dilution factor ~1:820, tested with fluorescein) into fresh medium with different concentrations of tetracycline, incubated plates for 24 h at 37 °C on a Titramax plateshaker (Heidolph, Schwabach, Germany, 900 rpm), shook plates for 20 s at 1200 rpm and measured $OD_{600}$ with a H1 platereader (Biotek, Vinooski, Vermont). For obtaining fine-scale MIC measurements we tested tetracycline concentrations at intervals of 0.125 µg/mL. We defined MIC as the lowest drug concentration that yielded $OD_{600} \leq 0.075$ (plate reader units) in three replicates performed on different days.

### 2.5.5 Evolution experiments

All precultures and evolution experiments were performed in M9CG medium. We transferred an overnight culture of every strain into 95 wells of clear flat bottom 96-well plates (200 µL/well), from where we diluted cultures into medium with tetracycline using VP408. One well contained a growth medium control. As initial concentration of tetracycline, we used half of the strain-specific MIC. For ten days, we pin-diluted cultures with VP408 every 24 h into medium with geometrically increasing tetracycline concentrations such that at day 10 the concentration was ten times the initial concentration (Figure 1D). During the experiment, the maximum number of generations was set by the daily dilution factor (~1:820) and was ~97. A fresh tetracycline stock solution was prepared from powdered tetracycline-HCl every day. All incubations were done at 900 rpm on a plate shaker at 37°C in the dark and plates were wrapped in plastic bags to mitigate evaporation. Replicate evolution experiments were performed with two additional 96-well plates for each of strains *A*, *B*, *C*, and *D* (IS-wt). Each 96-well plate was started from a culture inoculated with a different colony. At the end of experiments, we froze all rescued populations.

### 2.5.6 $OD_{600}$ and fluorescence measurements

Every day during the evolution experiment, after using 24 h old cultures for inoculating fresh medium with a higher tetracycline concentrations using VP408, we shook the old plates for 20 s at 1200 rpm to resuspend cells and measured $OD_{600}$ and reporter fluorescence with a H1 Platereader (Biotek, Vinooski, Vermont; excitation/emission: YFP 515/545 nm / gain 100; CFP 433/475 nm / gain 60).

### 2.5.7 Data analysis

Populations were classified as rescued if $OD_{600}$ exceeded 0.075 (plate reader units) at the end of the experiment. Fluorescence values were normalized to $OD_{600}$ and set to zero if $OD_{600}$ fell below 0.075. As reference for calculating the fold-increase in fluorescence, we took the average OD-normalized fluorescence of 95 cultures of the respective ancestral strain, inoculated in the same way as described for the beginning of evolution experiments, and grown in 96-well plates for 24 h without tetracycline. Rescued populations were classified as YFP or YFP+CFP if the observed fold increase in respective fluorescence over the ancestor was >2.77 at the end of the experiment. This threshold corresponds to the lowest observed increase in YFP fluorescence that was sufficient for rescue in the first set of replicate experiments (IS-wt strains *A*, *B*, *C*, and *D*). To identify populations that went extinct despite elevated YFP and/or CFP fluorescence we applied more stringent criteria, requiring increased fluorescence (fold increase >2.77) for at least two days at which OD was >0.3 (platereader units). These criteria were used to exclude extinct populations that were false positive for increased fluorescence due to low $OD_{600}$ values prior to extinction. Rescued populations that met the more stringent criteria for elevated CFP fluorescence, but that did not show elevated CFP fluorescence at the end of the experiment (final fold increase <2.77), were counted as amplifications for cost analysis (Figure 5 and Table 1), but not for Figure 2. For calculating final YFP/CFP ratios of rescued amplifications, we used internal plate reader fluorescence units directly. A Matlab script used to perform the above analysis is available as a supplementary file along with the platereader raw data used as input for the script ('Source code.zip'). Plots of fluorescence trajectories of every population can be found in Supplementary File 1 and phenotype classifications in Supplementary File 2.

### 2.5.8 Quantitative PCR for reporter cassette copy number determination

We inoculated samples of all rescued populations that we had chosen for sequencing from the first set of replicate experiments and that had a YFP+CFP fluorescence phenotype. We inoculated 2 mL M9CG with 10 µL of populations that were frozen at the end of the evolution experiment. The large inoculum was used to maintain amplification-related population diversity. We added the same amount of tetracycline as on the last day of evolution experiments to maintain amplifications. From all cultures that were turbid after overnight incubation, we isolated genomic DNA (gDNA). Ancestor gDNA was isolated from cultures without tetracycline. We performed qPCR using the GoTaq qPCR mastermix (Promega, Madison, Wisconsin) and a C1000 instrument (Bio-Rad, Hercules, California). Using dilution series of one of the gDNA extracts as template, we confirmed that all primer pairs had an amplification efficiency >90%. We quantified the copy number of *tetA* in each sample with the ΔΔCq method implemented in the instrument software (Bio-Rad), taking amplification efficiency into account. As reference, we used loci equidistant from the origin of replication and compared ratios of the measured and reference locus to the ratio of the same two loci in the ancestral DNA. qPCR was done in three technical replicates.

### 2.5.9 Identification of flanking homology

We searched 400 kb around loci *A-D* for homologous sequences on either side using REPuter (Kurtz et al. 2001) with the following search criteria: forward repeats ≥200 bp, Hamming distance ≤5.

### 2.5.10 DNA sequencing

We streaked all rescued populations of strains *A*, *C*, *D* (IS-wt), of strains *B* and *D* (IS-free), and of strain *BΔIS5I* for single colonies on LB agar. For IS-wt strain *B*, we analyzed one rescued population that had a YFP-only fluorescent phenotype, two YFP+CFP populations with unusual fluorescence trajectories and 11 randomly chosen populations from the remaining 74 YFP+CFP rescued populations, which had highly similar fluorescence trajectories. Colony-PCRs were performed on a single representative clone of each streak. We amplified at least 1.5 kb of the region upstream of the *tetA* start codon. The size of PCR products was checked for insertions or deletions on an agarose gel. Sequences were obtained using primer tetA_pseq1_f. If no PCR product was obtained, we performed arbitrary PCR with primer tetA_pseq2_f and a random primer, arb1 or arb6, for upstream binding. We then did a second PCR with a nested primer tetA_arb2 and primer arb2 using the first PCR product as template, and sequenced DNA extracted from the largest distinct band on an agarose gel. The full-length sequence of the *rho* gene was amplified and sequenced with primers rho_seq_f and rho_seq_r. For additional replicate evolution experiments, we sequenced clones of all rescued populations with a YFP fluorescence phenotype and with a YFP+CFP fluorescence phenotype showing high final YFP/CFP ratios. In four cases, we identified the exact same mutation in clones isolated from two populations that had been in neighboring wells during evolution experiments. In order to ensure that a potential cross-contamination between these two wells did not influence results, we excluded one of each pair of such neighboring populations from all analyses.

### 2.5.11 Junction PCR

Colony PCR for amplification junctions was performed with primers IS5I_flank_f and IS5H_flank_r on single colonies of 16/16 evolved populations of strain *B*. For the data shown in Figure 2D, we used gDNA previously isolated from populations for qPCR to ensure a comparable amount of PCR template in all reactions.

### 2.5.12 Whole genome sequencing

We isolated gDNA from overnight cultures of single clones of four rescued D populations as well as of the ancestral D strain grown in LB. A whole genome library was prepared and sequenced by GATC biotech (Konstanz, Germany) on an Illumina sequencer (125 bp reads). Fastq files were analyzed with the breseq script (Barrick et al. 2014). We used the MG1655 genome (Genbank accession number U00096.3) as a reference for assembling the ancestral *D* genome, which then served as a reference for analyzing the genomes of the evolved clones. Fastq files are available at: http://dx.doi.org/10.15479/AT:ISTA:65

### 2.5.13 Cloning of reporter plasmids

For building the reference plasmid pAnc, which reports on expression from the ancestral $p_0$ sequence, we exchanged the pLtetO-1 promoter and RBS of pZA21-*yfp* for the $p_0$-RBS sequence upstream of *tetA-yfp* in the reporter cassette. Using a Q5 site-directed mutagenesis kit (New England Biolabs) with pAnc as template, we reconstructed small mutations (substitutions and small insertions and deletions, Figure 3C). We did the same with the terminal 50 bp of IS1 (5' terminus) and IS5 (3' terminus), which we put instead of the 50 bp of $p_0$ in the exact position where insertions were found in the experiment (Figure 3E). To confirm the IS1-$p_0$ hybrid promoter, we exchanged 20 bp of $p_0$ downstream of the IS1 insertion point in the respective reporter plasmid. The 20 bp were replaced by a randomly shuffled sequence composed of the same nucleotides. For the other IS reporter plasmids (Figure 3D), we PCR-amplified the last 600 bp of IS and cloned them into the XhoI/EcoRI sites of pZA21-*yfp*. The orientation of the truncated IS corresponds to that found in sequenced clones. As autofluorescence control, we removed the YFP fragment between EcoRI and MfeI restriction sites of pZA21-*yfp* and obtained pZA21-empty by religation of compatible ends. All changes were sequence-verified. Cloning and reporter measurements were done in strain NEB 5 alpha (New England Biolabs).

### 2.5.14 Quantifying YFP reporter fluorescence from plasmids

We grew six replicate overnight cultures of the reporter plasmid strains in LB Kanamycin (50 µg/mL) in a 96-well plate and diluted them into M9CG supplemented with Kanamycin using a VP407 pin replicator (approximate dilution factor 1:100). Diluted cultures were shaken and incubated at 37°C in the platereader and $OD_{600}$ and YFP fluorescence was monitored every 10 min (YFP gain 120). YFP readings were normalized to $OD_{600}$ and averaged for each replicate at all timepoints at which $OD_{600}$ was between 0.20 and 0.25 (platereader units, i.e. mid-exponential phase).

### 2.5.15 Tetracycline resistance phenotyping on solid medium

Clones and strains to be tested were pregrown overnight in M9CG and diluted as shown in Figure 2B and Figure 3 – Supplement 3. We spotted 2.5 µL of diluted cultures on M9CG agar plates. After 24 h incubation at 37 °C, we took YFP fluorescence images of plates using a lab-made macroscope (http://openwetware.org/wiki/Macroscope). The macroscope uses a Canon EOS 600D digital camera and a Canon EF-S 60 mm f/2.8 Macro USM lense (Canon, Tokyo, Japan). For illumination, we used a Cyan (505 nm) Rebel LED (Luxeon Star LEDs, Brantford, Canada) with a HQ500/20x excitation filter (Chroma, Bellow Falls, Vermont). As emission filter we used a camera-mounted D530/20 filter (Chroma).

### 2.5.16 Reverse transcription

Stationary cultures of MG1655 Δ*tolC* (*rho*-wt) and of the isogenic strain with the *rho* M416I mutation in LB were diluted 1:100 in M9CG supplemented with tetracycline (0.44 µg/mL, i.e. 50% of the MIC of strain MG1655 Δ*tolC* and grown overnight at 37°C with shaking. Total RNA was isolated using an Aurum Total RNA Mini kit (Bio-Rad) and DNA removed using an Ambion DNA-free kit (Life Technologies, Carlsbad, California). Isolated RNA was quantified using a Nanodrop spectrophotometer and integrity was checked on an agarose gel. cDNA was synthesized using an iScript cDNA synthesis kit (Bio-Rad) with 1 µg of total RNA as input in a 20 µL reaction. For the non-reverse-transcriptase (NRT) control reaction we used 0.5 µg of each of the two RNA samples.

### 2.5.17 Endpoint PCR on cDNA

After reverse transcription, cDNA samples and the NRT control sample were diluted by adding 150 µL of nuclease-free water. Endpoint PCR to test for the presence of transcripts resulting from possible read-through across Rho-dependent terminators were done with a OneTaq Quick-Load Mastermix (New England Biolabs), using 1 µL of diluted cDNA or NRT control as template in a 50 µL reaction. To detect rare transcripts, we used 45 amplification cycles. As a positive control template in PCR reactions, we used 1 µL of a colony of strain MG1655 Δ*tolC* resuspended in 25 µL water and heated to 95°C for 4'. For agarose gel visualization, we loaded 15 µL of cDNA and NRT control PCR reactions and 2 µL of the positive control PCR reactions.

### 2.5.18 Inferring the order of two adaptive mutations occurring in the same clone

In several cases, fluorescence analysis and sequencing revealed two potentially adaptive mutations in the same clone/population (colored dots on top of bars in Figure 3A and Figure 5 – Supplement 1). To infer which mutation came first, we proceeded as follows. For amplifications that occurred in combination with point mutations, we examined sequence chromatograms obtained from single clones. In all three cases, point mutations appeared as mixed nucleotide peaks, indicating that amplifications were initiated before the point mutations occurred. In two cases of amplifications co-occuring with upstream IS insertions, insertions occurred first. This is evident since PCR products used for sequencing appear as single bands of larger size than expected on agarose gels, whereas later insertions are expected to give two bands – a smaller one for copies without the insertion and a larger one for copies having the insertion. In one case, the insertion of IS3 upstream of locus *C* was a prerequisite for amplification initiation, as we could show by PCR that the IS3 insertion was at the amplicon junction. Cases of co-occurrence of amplifications with deletions or Rho-mutations were decided based on fluorescence trajectories. YFP/CFP ratios that remained high and relatively constant throughout the experiment indicate that amplifications expanded only after the other mutation had occurred. YFP/CFP ratios that increase at an intermediate timepoint during the experiment indicate that amplifications were initiated first. Last, we assume that a Rho mutation in strain *A* was selected only after the insertion of an upstream IS5 element, since Figure 4B indicates that Rho mutations alone would not have been adaptive in strain *A*. Rather, we assume that the Rho mutation enhanced transcriptional read-through from IS5 into the reporter cassette. Dots-on-bar color assignments in Figure 3 – Supplement 5 do not reflect the order of mutations, as we did not do such analysis for additional replicate experiments.

### 2.5.19 Assessment of gene essentiality

Essentiality data for upstream protein coding genes (Figure 4A) was taken from a published dataset (Baba et al. 2006). We did not find data on the essentiality of the *valU* tRNA operon upstream of locus *C* in the literature data. Therefore we tested experimentally, if deletions of the complete *valU* operon are tolerated. We attempted to delete the operon using recombineering with pKD13 as template plasmid for a *kanR* cassette and primers valU_ko_f and valU_ko_r. The number of colonies on the *valU* knockout selection plate was more than tenfold lower than that of a control knockout of the neighboring *xapR* gene with primers xapR_ko_f and xapR_ko_r. To exclude that the low number of recombinants was due to a hairpin structure contained in the valU_ko_r primer, we repeated recombineering with a different reverse primer, valU_ko_r2, and obtained similar results. The low recombineering efficiency was not due to a smaller amount of PCR product used in transformations. Of six tested colonies obtained on the *ΔvalU::kanR* selection plate, only one colony gave a PCR product of the expected size in a test with flanking primers, showing that 5/6 colonies are not true *valU* knockouts. This suggests that *valU* deletion mutants require rare compensatory mutations to restore growth. Therefore the *valU* operon was considered as essential.

### 2.5.20 Single-step plating experiments

We inoculated 1 mL of LB with a single colony of strains to be tested. After overnight incubation, saturated cultures were diluted 1:1000 into experimental evolution medium without tetracycline, and then split into 10 wells of a 96-well plate (220 µL / well). The 96-well plate was incubated on a plate shaker at 37°C for 24 h to obtain saturated cultures, of which 180 µL containing approximately $2 \times 10^8$ cells were plated on M9CG medium with tetracycline at a concentration two times the MIC of respective strains (cell numbers were determined by plating dilutions on non-selective medium). Plates were incubated at 37°C in the dark and colonies counted every 24 h. After 2 days, we picked one colony from every plate that had at least one colony on it and inoculated 200 µL of M9CG medium in a 96-well plate with the picked colony. After 24 h incubation at 37 °C, we used the VP407 pinner to spot approximately 2 µL on M9CG agar plates. After another 24 h incubation, we took CFP fluorescence images of plates with the macroscope (see 'Tetracycline resistance phenotyping on solid medium'). For illumination, we used a Royal Blue (447.5 nm) Rebel LED (Luxeon Star LEDs) with a D436/20x excitation filter (Chroma). As emission filter we used a camera-mounted D480/40m filter (Chroma). The mean intensity of pixels of each spot was quantified. Spots with intensity 6 times greater than the mean intensity of all ancestor spots are considered to have amplifications (Figure 6 – Supplement 2).

### 2.5.21 Statistical analysis

To test for homogeneity in the distribution of rescued vs. extinct populations, fluorescence phenotypes and mutation types, $r \times$ c Fisher's exact test for Count Data was used (fisher.test function in R (Core Team 2012)). For testing the distribution of mutation types, we used types indicated in Figure 3A by bar color, not dot color. For testing 2x2 contingency tables, Fisher's exact test was used with an alternative hypothesis of odds ratio $\neq 1$. Permutation tests were performed with the perm package (Fay & Shaw 2010) for R (permTS function, method='exact.mc', $10^4$ Monte Carlo replications, two-sided).

### 2.5.22 In-silico analysis of adaptive potential of *E. coli* gene neighborhoods (Venn Diagram)

We used the Profiling of *E. coli* Chromosome (PEC) database available at https://shigen.nig.ac.jp/ecoli/pec/genes.jsp (accession number UA00096.2) and included all 4317 genes (feature type 'gene') of *E. coli* MG1655 with essentiality information in our analysis, which excludes non-coding genes. The position and orientation of promoters was extracted from Table S2 of the same study used to identify candidate transcripts in Figure 4A (Conway et al. 2014). We only included promoters annotated as 'primary' promoters in our analysis. The 'Promoter Confidence Score' was not taken into account. The position, orientation, and termination mode (intrinsic or non-intrinsic) of all terminators was extracted from Table S3 of the same study (Conway et al. 2014). In order to identify all genes downstream of Rho-dependent terminators (green circle in Figure 7B), we identified the closest upstream co-oriented terminator of every gene and evaluated whether it was predicted to be an intrinsic terminator or not, in which case we assumed it is Rho-dependent. In order to identify all genes to which a co-oriented upstream promoter could be co-opted by deletion without disrupting an essential gene, we first identified the next essential upstream gene of every gene, and then evaluated if there is at least one co-oriented promoter and intervening co-oriented terminator between the gene of interest and the next upstream essential gene. If this was the case, the gene of interest was included in the respective set of genes (magenta circle in Figure 7B). In order to identify genes between flanking duplicates (blue circle in Figure 7B), we used the online REPuter tool (Kurtz et al. 2001) to find all forward repeats on the chromosome that satisfied the following criteria: repeat length ≥200 bp, Hamming distance ≤8, maximum distance between repeats 100 kb, minimum distance between repeats 200 bp. In this way, we identified four large regions of the MG1655 chromosome between flanking repeats: between IS1B and IS1C (containing 13 genes and locus *E*), between IS5H and IS5I (containing 42 genes and locus *B*), and between the ribosomal operons *rrnA* and *rrnC* (81 genes), and *rrnB* and *rrnE* (31 genes). We also obtained 6 genes between closely spaced repeats matching our criteria (*ybfB*, *ybfL*, *yibA*, *ldrA*, *ldrB*, *ldrC*), which we did not include in the set 'between flanking duplicates', since the behavior of such closely spaced repeats might be different than those studied in our system. The Venn diagram was drawn in Matlab using the 'ChowRodgers' method for sizes of circles and intersection areas. A list of all included genes and their assignment to the three sets is available in Figure 7B – Source Data.

# 3 Context-specific effects of promoter mutations

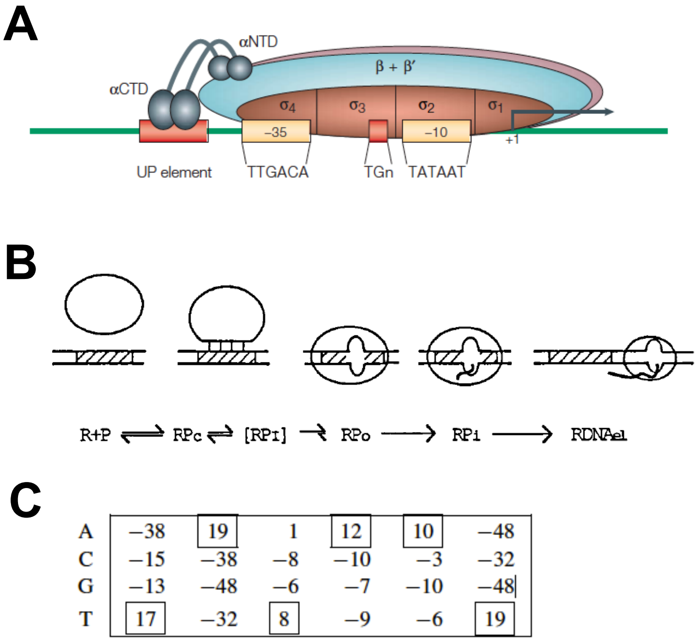
This chapter is the result of a collaboration with Murat Tugrul, Srdjan Sarikas, and Gašper Tkačik in an advisory role. Murat Tugrul did initial modeling of a preliminary dataset (see text). Srdjan Sarikas processed the sort-seq raw data.

## *3.1 Introduction*

Predicting gene expression from DNA sequence is a fundamental problem of molecular biology and is central to understanding the evolution of gene regulation in the genomic era. In the absence of a genetic code for regulatory DNA, inferring a genotype-phenotype map and effects of mutations, calls for different approaches than for protein-coding DNA. A fundamental way in which the sequence of regulatory DNA is translated into a phenotype is through sequence-specific binding of proteins, such as transcription factors and the RNA polymerase (RNAP) (Snyder & Champess 2007). Arguably the simplest regulatory molecular phenotype is the strength of a constitutive promoter as a function of the sequence-specific interaction between RNA polymerase (RNAP) and the promoter sequence.

The bacterial RNAP is a large multi-subunit protein complex, composed of subunits β, β', $\alpha_1$, $\alpha_2$, and ω (Figure 2A, (Browning & Busby 2004)). For promoter recognition, RNAP associates with another subunit, the σ factor, to form the RNAP holoenzyme. In this work, I use simply 'RNAP' to refer to the holoenzyme. Different σ factors recognize different sets of sequences associated with particular stresses or growth conditions (Ishihama 2000), with $\sigma^{70}$ being the 'housekeeping' σ factor in *E. coli*, responsible for expression of genes during exponential growth. Although the C-terminal domain of the α subunit of RNAP also contacts DNA (the 'up-element'), it is the σ factor that is the primary specificity determinant of the RNAP-DNA interaction (Figure 2A). DNA recognition by the σ factor occurs primarily at two elements of the promoter, the -35 and -10 box, located upstream of the transcription start site (TSS or '+1'), most frequently at positions between -35 to -30 and -12 to -7 respectively. The strength of a constitutive promoter, i.e. the frequency of productive transcription initiation, is therefore a function of the DNA sequence in this region.

We should keep in mind that transcription initiation is a multi-step process which is not yet fully understood (Ruff et al. 2015). Sequence-specific binding of RNAP to DNA is only the very first step (Figure 2B). Also the kinetics of later steps of the initiation process show some sequence-dependency, particularly the promoter isomerization step, in which an open 'transcription bubble' is formed (McClure et al. 1983). Other factors influencing gene expression from promoters, which I do not discuss here, include chromosome structure and position on the chromosome (Bryant et al. 2014), temperature and other environmental factors, and physiological state of the cell, including the concentration of specific metabolites (Haugen et al. 2008). Also, the concentration of proteins used to report on promoter strength depends on additional post-transcriptional factors such as mRNA structure and stability and initiation of translation (Griswold et al. 2003; Stenström & Isaksson 2002; Vind et al. 1993). Next to influencing protein binding, the DNA sequence also affects DNA shape locally, which can impact the recognition of DNA by proteins (Rohs et al. 2010) such as the RNAP.



| | | | | | | |
|---|---|---|---|---|---|---|
| A | −38 | 19 | 1 | 12 | 10 | −48 |
| C | −15 | −38 | −8 | −10 | −3 | −32 |
| G | −13 | −48 | −6 | −7 | −10 | −48 |
| T | 17 | −32 | 8 | −9 | −6 | 19 |

*Figure 8. Promoter recognition by RNAP. (A) Multiple subunits of the RNAP holoenzyme and contacts with a promoter sequence. 'TGn' denotes the extended -10 element. (B) Steps of transcription initation. R – RNAP holoenzyme, P – promoter DNA, $RP_C$ – closed complex, $RP_I$ – poorly defined intermediate step(s), $RP_O$ – open complex with DNA melting at the -10 box, $RP_i$ – initiation complex with nascent RNA, $RDNA_{el}$ – elongation complex. (C) Position weight matrix of the E. coli -10 box. (A) reproduced from (Browning & Busby 2004), (B) reproduced from (Knaus & Bujard 1990), (C) reproduced from (Stormo 2000).*

Constitutive expression is an equally fundamental and attractive molecular phenotype for studying, as it is the basis of other more complex phenotypes involving regulation by transcription factors, and it is measured easily. Early attempts to abstract a regulatory 'code' for promoter sequences in general and the sequence determinants for RNAP binding in particular were based on the detailed study of a few model promoters such as the promoter of the *E. coli lac* operon ($P_{lac}$, (Dickson et al. 1975)), the promoters of phage λ ($P_{RM}$, $P_R$, $P_L$ (Maniatis et al. 1975)) and phage T7 (Pribnow 1975). As the sequence of more promoters became known, the notion of 'consensus' sequences became central to our understanding (Pribnow 1975; Hawley & McClure 1983; Lisser & Margalit 1993), the consensus being the sequence of the most frequently found nucleotides at each position of a binding site. For the RNAP binding site, its consensus sequence is also referred to as the ‘canonical’ binding site (TTGACA and TATAAT for -35 and -10 boxes respectively). It should be noted that no single promoter in *E. coli* has the exact canonical sequence. The homology between a given sequence and the consensus RNAP binding site is often assumed to correlate with promoter strength, however this is not generally true (Knaus & Bujard 1990; Kawano 2005), and in fact ‘perfect’ consensus promoters may be dysfunctional (Graña et al. 1988; Hook-Barnard & Hinton 2007; Miroslavova & Busby 2006). The distance to consensus can be measured as number of mismatches or, a little more sophisticated, using a homology score that takes into account the frequency distribution of nucleotides found at different positions in promoter collections, as in position weight matrices (PWMs, Figure 2C). Consensus-based approaches are central in bioinformatics for the prediction of promoters (Stormo 2000), although evidence based on bioinformatic predictions, without additional experimental support, is usually classified as weak (Gama-Castro et al. 2016).

As our knowledge of transcription initiation and the interaction of RNAP with promoter sequences became more detailed, more realistic and quantitative models of gene regulation were developed. An important class of such models are thermodynamic models (Bintu et al. 2005). These models rest on the assumption that gene expression is proportional to the equilibrium probability of RNAP being bound to a promoter sequence (i.e. promoter occupancy). This assumption should be broken down into two assumptions, the thermodynamic equilibrium assumption per se, and the assumption that RNAP binding is the only sequence-dependent step in the initiation of transcription (‘single step assumption’). In thermodynamic equilibrium, the binding probability $P_{on}$ is a function of the binding energy $E$ between RNAP and the promoter DNA.

$$P_{on}\,(E) = (1 + e^{(E-\mu)/k_B T})^{-1} \qquad (1)$$

μ is the chemical potential (related to the concentration of free RNAP, generally having an unknown value), and $k_B T$ is the product of the Boltzmann constant and the temperature, a scaling factor for energy values. The binding energy $E$ is a function of the promoter sequence and is often assumed to be the sum of energy contributions of individual nucleotides (‘additivity assumption’). Thus, individual nucleotide positions are thought to contribute independently to binding, which is mathematically practical, but biologically questionable (Graña et al. 1988). The above function has a sigmoid shape with lower binding energy yielding higher expression. RNAP binding sites are expected to stand out on DNA sequences as positions of minimal binding energy.

Thermodynamic models are also widely applied to transcription factor (TF) binding and how it affects the RNAP-DNA interaction in turn. It should be noted that although both TFs and RNAP bind promoters in a sequence-specific manner, these interactions are not equivalent. Consequently, different modeling complications arise from the specifics of RNAP-DNA and TF-DNA interactions respectively. For example, one would expect the above 'single step' assumption to be less problematic for transcription factors. Additional complications for modeling the action of TFs are the wealth of possible mechanisms how TF binding can influence RNAP binding, e.g. by steric exclusion, recruitment, DNA looping (Bintu et al. 2005) and the observation that TF binding sites can affect gene expression independently of occupancy (Garcia et al. 2012).

The equivalent of a position weight matrix in the thermodynamic framework is an energy matrix, in which the energy contributions of each of the four possible nucleotides at the different positions of a promoter are given as matrix entries (Figure 9). While entries of a position weight matrix are generally inferred from homology of natural promoters, energy entries of a matrix in thermodynamic models are often derived by quantifying the effect of mutations on a given sequence. For example, the energy matrix in Figure 9 was inferred from fluorescence measurements of a mutant library of the *lac* promoter driving expression of GFP (Kinney et al. 2010).

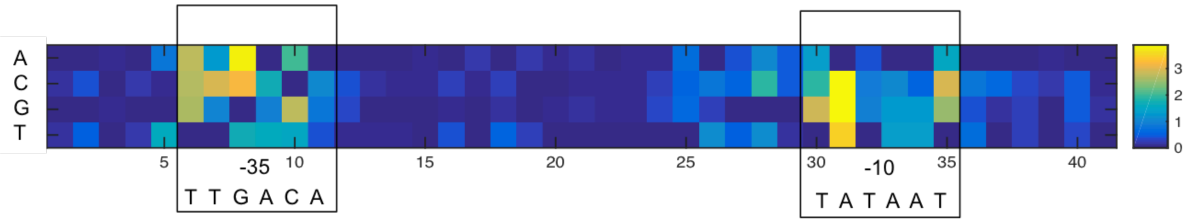


*Figure 9. Energy matrix of the interaction of RNAP with the lac promoter. Scale is in units of $k_BT$. Matrix entries are from (Kinney et al. 2010) and were provided by Murat Tugrul. Frames indicate the -35 and -10 boxes ('feet') and the consensus sequence corresponding to the minimum energy values.*

Embedded in the thermodynamics framework, energy matrices can be used to model the effect of mutations on gene expression in a biophysically more realistic and possibly more quantitative way than by using homology-based methods. In his PhD thesis, Murat Tugrul used the above energy matrix to test whether he could predict mutations in $p_0$ observed in evolution experiments preliminary to those presented in chapter 2 (Tuğrul 2016). He also tested the correspondence of the thermodynamic model with an experimental dataset of 76 single nucleotide mutants of *$p_0$*, which I generated and characterized with respect to driving expression of YFP (for details, see (Tuğrul 2016)). In both cases, correlations between experiment and the model were highly significant, which means that the model overall accurately captures an important part of the sequence dependency of promoter strength. At the same time however, correlations were not very strong, which means that the uncertainty in the prediction of a particular sequence remains large.

Overall, it is clear that promoter sequence is the most important determinant for the rate of transcription initiation. It is also clear that when trying to predict expression from sequence, currently used genotype-phenotype maps perform well in terms of significance of correlation when applied to large datasets. For individual sequences however, predictions of promoter strength are not very accurate. One possible reason for this inaccuracy is that our models of promoter function are typically inferred using naturally evolved, functional promoters as a starting point. Many of the best-studied promoters are phage promoters and very strong ($P_L$ and $P_R$ of phage λ, $P_{A1}$ of phage T7, $P_{N25}$ and $P_{H207}$ of phage T5) (Deuschle et al. 1986; Knaus & Bujard 1990). It is an open question how well insights and mathematical models derived from these model promoters apply to bacterial promoters in general and, beyond that, to the full space of sequence and function, which includes many more weak than strong promoter sequences (see section 4.2.2).

Understanding where the inaccuracy of current models of promoter function comes from, with the eventual goal of improving models of promoter function, is the main concern of this and the following chapter. We do so by applying energy matrix models to expression data from random sequence libraries. Initially, this work was motivated by the question whether we could predict point mutations seen in evolution experiments (chapter 2), and how likely *de novo* promoter evolution by point mutation is compared to evolution by other mutation types as explored in chapter 2.

In this chapter, I explore the following questions:

1. Can point mutations in *$p_0$* found in evolution experiments of chapter 2 be predicted using an energy-matrix based model? This is essentially a recapitulation of the work of Murat Tugrul’s thesis with a more comprehensive mutant dataset.
2. How specific is the power of energy matrix models to the sequence they are applied to given their inference on a particular sequence background? For this, I use three

starting sequences for single-nt promoter mutagenesis: the RNAP binding site of the *lac* promoter, a part of the random $p_0$ sequence, and a second random sequence.

3. At which positions of a promoter does the energy matrix model fail?

## *3.2 Results*

### 3.2.1 A sort-seq experiment for quantifying effects of single-nt mutations in multiple promoter sequences

To obtain genotype-phenotype data for the sequence space surrounding more than one sequence, I created three plasmid libraries with single nucleotide mutations in a 36 bp region upstream of a *gfp* reporter gene preceded by a functional RBS (Figure 10). The length of 36 bp was chosen as it is large enough to accommodate a full RNAP binding site and small enough to use the primer-based mutagenesis approach, which is a variation of classic site-directed mutagenesis (Figure 10). Also, the relatively small mutagenized region was chosen to enable the creation of libraries with near-complete coverage of all possible single-nt mutants and high sequence coverage for every single mutant. Each of the three libraries are derived from one starting sequence as shown in Table 2. Apart from the mutagenized 36 bp region, the three starting plasmids are identical.

*Table 2. Starting sequences of mutagenized region in three plasmid libraries.*

| | | |
|---|---|---|
| pMS9_1 | lacZ RNAP binding site | GGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTG |
| pMS9_2 | randomized order of nts in lacZ RNAP binding site | TTCGGCTTTCTTCGTGCATAATGCTTCGGTCTATGG |
| pMS9_3 | $p_0$ from chapter 2 | TTACCTTGCAGGAATTGAGGCCGTCCGTTAATTTCC |

Mutant expression is measured using sort-seq (Peterman & Levine 2016). In sort-seq, a library of cells with different genotypes is first sorted into bins according to reporter fluorescence using FACS. Afterwards, sorted sequences are bar-coded by bins and the identity of mutations in conjunction with bin information is obtained by Illumina sequencing, yielding a distribution of read counts for each mutant sequence (Figure 10). There are multiple experimental steps from sorting to obtaining read distributions (sorting, re-growing cultures, isolating plasmid, barcoding, sequencing, post-processing; see Figure 10, details described in Methods sections 3.4.3 to 3.4.6), each of which may introduce biases in the mapping between actual fluorescence distributions and read count distributions. We sought to minimize biases to obtain more accurate fluorescence proxies from read distributions. We did so by 'spiking' binned cultures with known numbers of cells containing a plasmid with an unrelated reference sequence instead of the mutagenized region Figure 10. The distribution of the reference sequence, added in equal amounts to each bin, is expected to become biased along the process in the same way as the mutant sequences. Therefore, dividing read counts of each mutant sequence bin-wise by the number of reference sequence reads is expected to debias fluorescence proxies calculated from read distributions. Since the number of reference sequence reads in each bin is rather high ($>10^3$, Table 5), there is no concern about introducing substantial noise due to division by small numbers.

The effect of debiasing can be seen from comparing read distributions to the known fluorescence distributions of the three starting plasmids in Figure 11A. Although the effect of debiasing on calculated fluorescence proxies is modest (Figure 11B), the close alignment of debiased read distributions with the original fluorescence distributions (Figure 11A) demonstrates the usefulness of the procedure.

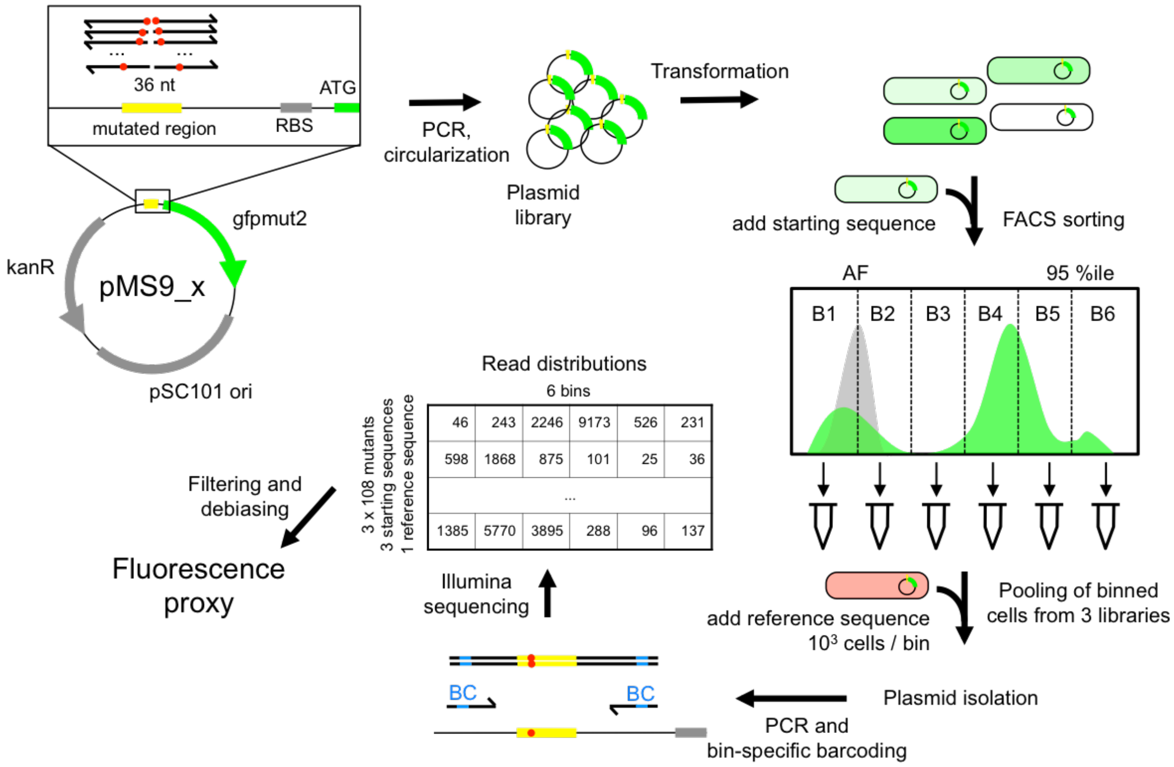
### 3.2.2 The distribution of mutational effects on three different starting sequences.

Having obtained fluorescence proxies for each mutant, we inspected the distribution of mutational effects of single nucleotide mutants (Figure 12). We notice three things.

First, and surprisingly, the two random starting sequences of pMS9_2 and pMS9_3 yield higher fluorescence than the naturally evolved RNAP binding site of the *lacZ* promoter (pMS9_1). Although the lacZ promoter is known to be a weak promoter requiring activation by CRP for full activity (Malan et al. 1984), it is unexpected that both random sequences yield higher expression and thus should be deemed ‘functional’. Still, expression from all three starting sequences is much lower (by roughly two orders of magnitude) than that of known strong promoters such as $P_L$ of phage λ.

Considering the shape of the distribution of mutation effects in the three libraries and the differences in expression between the three reference sequences, it appears that moving from a weak to a stronger promoter by single point mutations is possible but gets harder as a promoter gets stronger.

The third observation is that for pMS9_3, which contains the part of $p_0$ repeatedly found mutated in evolution experiments, the two mutations found in evolution experiments (C-31T and T-24A are among the mutations with the highest beneficial effect (arrow in Figure 12). This indicates that the effect of these mutations is similar in the chromosomal context in which they were selected, and in the context of a plasmid, as is also expected from previous reporter assays (chapter 2.3.4).



| | | | | | |
|---|---|---|---|---|---|
| 46 | 243 | 2246 | 9173 | 526 | 231 |
| 598 | 1868 | 875 | 101 | 25 | 36 |
| | | ... | | | |
| 1385 | 5770 | 3895 | 288 | 96 | 137 |

*Figure 10. Workflow for mutagenesis and measuring fluorescence of multiple mutant libraries using sort-seq. Starting from the top left: Single nucleotide mutagenesis of plasmids pMS9_x (x=1, lacZ RNAP binding site; x=2, lacZ scrambled; x=3, $p_0$). A 36 bp region (yellow) upstream of a ribosomal binding site (RBS, grey) and a gfp reporter gene (green) is mutagenized by PCR amplification with degenerate primer pools introducing exactly 1 mutation per molecule. Red dots – position of degenerate nucleotide in single primers constituting the primer pools. PCR products are circularized and transformed into E. coli. Each of the three resulting libraries, to which the ancestral starting plasmid is added, is sorted according to GFP fluorescence into six bins (B1-B6). Green FACS histogram – cartoon example of a library, grey FACS histogram, AF – autofluorescence background. After sorting, cells in respective bins 1-6 of the three libraries are pooled and a constant number of cells containing a reference sequence (red cell) is added for later debiasing. Isolated plasmid libraries are subsequently used as PCR templates with primers that incorporate bin-specific barcodes into PCR products used in Illumina sequencing. Finally, read count distributions of every sequence across bins are debiased by dividing by bin counts of the reference sequence. Debiased distributions are then transformed into a fluorescence proxy. For details see Methods.*

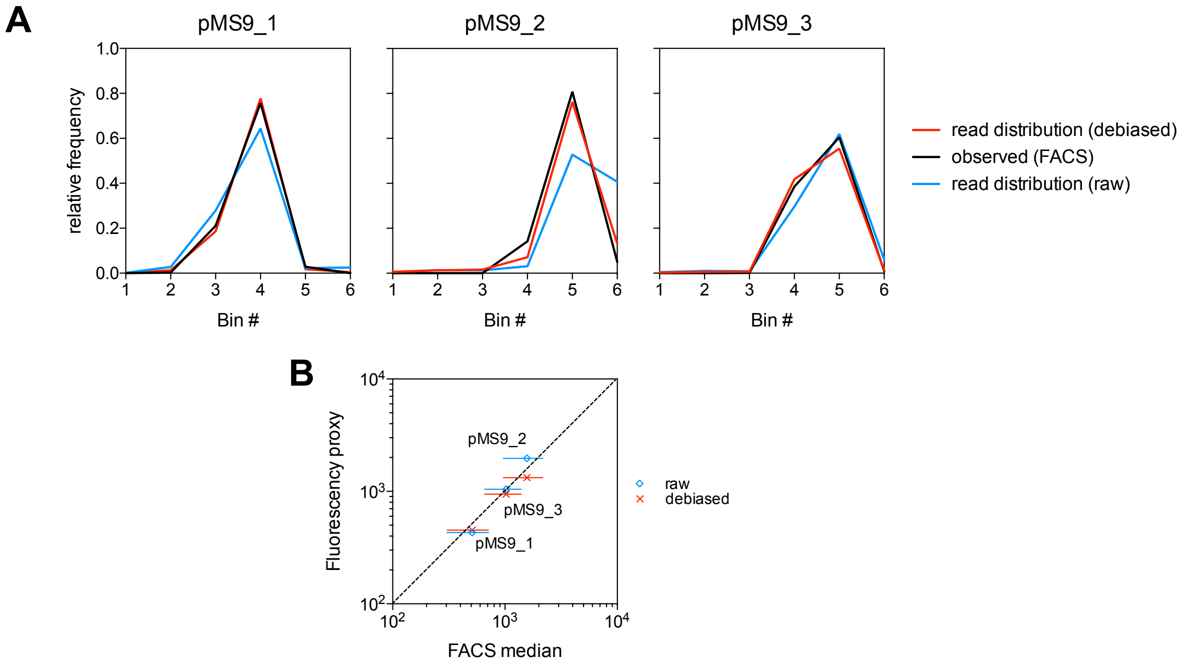


*Figure 11. Debiased read distributions approximate FACS distributions better than raw read distributions. (A) Black line – relative frequency of cells with the starting plasmid across bins as observed in FACS of a clonal culture. Blue line – raw distribution of reads of the starting sequence across bins. Red line – debiased read distribution. For debiasing, raw read counts of each bin are divided by the read counts of a reference sequence derived from cells that were added at equal numbers to each FACS bin (see Table 5 in the Methods section) . For calculating relative frequencies, the sum of the divided read counts is normalized to 1. (B) Fluorescence proxy of starting plasmids calculated as the geometric mean of raw (blue) and debiased (red) read distributions (y-axis) compared to the median of the FACS distribution of respective clonal cultures (x-axis, horizontal error bars are rSD). Dashed line is x=y.*

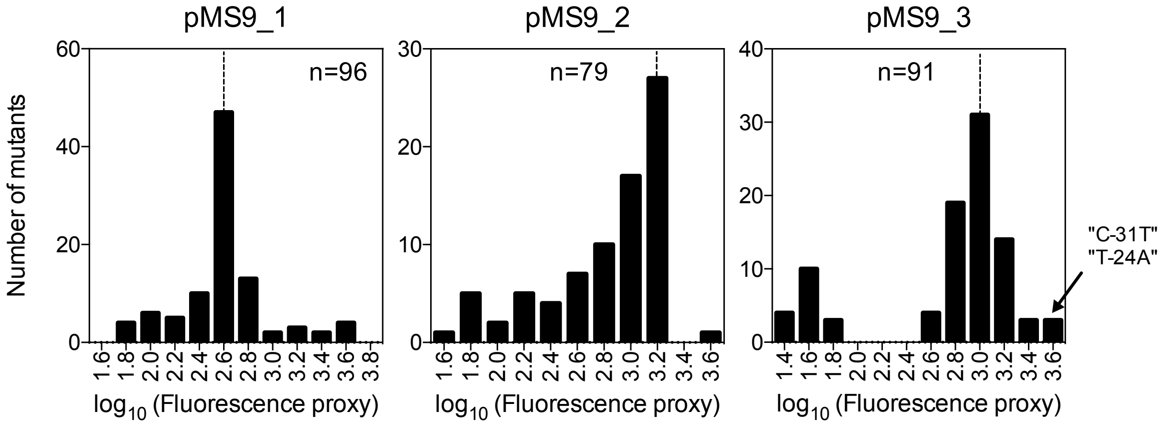


*Figure 12. Distribution of mutational effects. Dashed lines – fluorescence proxy of the starting sequence. "C-31T" and "T-24A" are the two point mutations within the tested region of $p_0$ seen in evolution experiments.*

### 3.2.3 Identifying RNAP binding sites in random sequences

As a next step, we sought to identify the specific binding sites of RNAP that apparent transcription from the three sequence libraries can be ascribed to. For pMS9_1, as this encodes the natural RNAP binding site of the *lacZ* promoter, the position of the binding site is known, but for pMS9_2 and pMS9_3, it is not. Importantly, searching for the 'core bases' of the canonical motifs (TTGnnn <spacer> TAnnnT, (Yona et al. 2018)) fails to identify a functional promoter in both pMS9_1 and pMS9_2. This illustrates the necessity for more detailed motif models such as an energy matrix model in the identification of promoters.

In our data, a functional binding site of RNAP is expected to satisfy two criteria. 1) It corresponds to the matrix position in a sequence with a minimum energy. 2) The variation in binding energy between different mutants is negatively correlated with the variation in observed expression. In Figure 13, we tested all possible binding frames of RNAP that overlap the mutagenized core of 36 nt for these two criteria. Due to flexibility in the length of the spacer between the -35 and -10 boxes, we tested frames for both spacer length 17 bp, which is the most common spacing in natural promoters (Lisser & Margalit 1993), and 18 bp as in the native *lac* promoter. For all three libraries, a single binding frame of the -10 box could be identified, although for pMS9_2 and pMS9_3 there was no clear preferred spacer length. For pMS9_1, the known RNAP binding site of *lac* promoter was correctly found (Figure 14). Going with the spacer length that gives the minimum binding energy, we continue our analysis assuming a spacer length of 18 bp for pMS9_1 and pMS9_2 and 17 bp for pMS9_3. Note that the -35 box of pMS9_2 lies outside of the mutagenized region (Figure 14) and is therefore constant for all mutants.

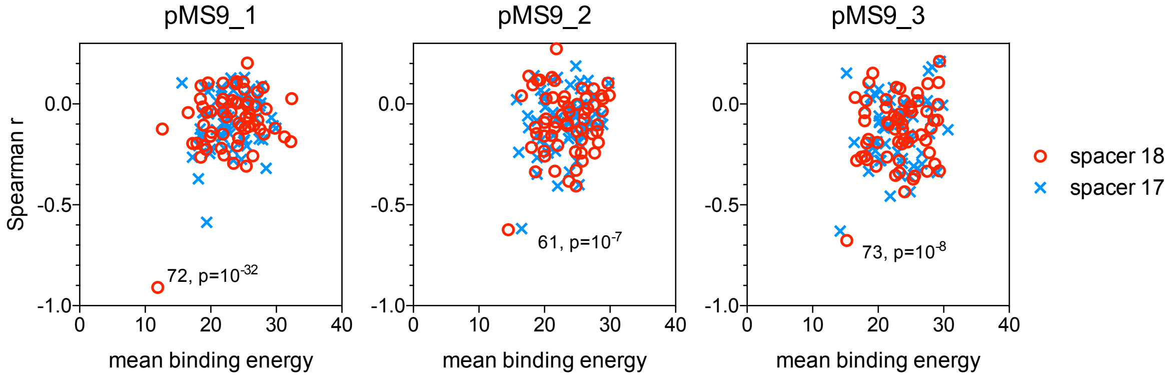


*Figure 13. Identifying a predominant frame of RNAP binding. Three panels show results for the three libraries. Each point corresponds to a single tested frame. Values are calculated with the lacZ energy matrix of spacer length 18 (red) or 17 (blue). x-axis: mean binding energy of a mutant sequence of the respective library. A frame important for binding is expected to have a low mean binding energy value. y-axis: Spearman rank*

*correlation coefficient between expression (FACS fluorescence proxy) and binding energy of all mutant sequences contained in the library. Since binding energy is inversely related to expression, a highly negative Spearman r is expected for predominant frames. Point labels indicate the position of +1 and multiplicity corrected p-values of Spearman r.*

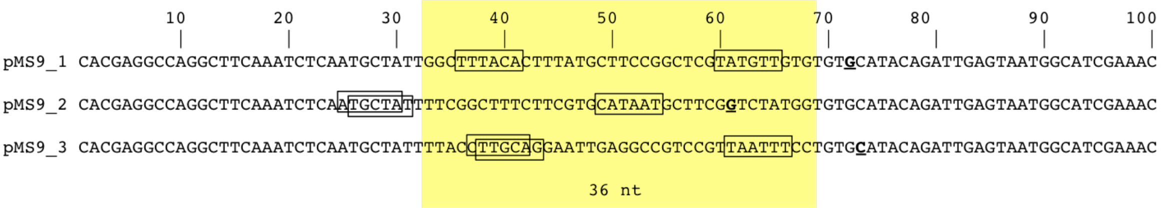

10
20
30
40
50
60
70
80
90
100
pMS9_1 CACGAGGCCAGGCTTCAAATCTCAATGCTATTGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGTGCATACAGATTGAGTAATGGCATCGAAAC
pMS9_2 CACGAGGCCAGGCTTCAAATCTCAATGCTATTTTCGGCTTTCTTCGTGCATAATGCTTCGGTCTATGGTGTGCATACAGATTGAGTAATGGCATCGAAAC
pMS9_3 CACGAGGCCAGGCTTCAAATCTCAATGCTATTTTACCTTGCAGGAATTGAGGCCGTCCGTTAATTTCCTGTGCATACAGATTGAGTAATGGCATCGAAAC
36 nt

*Figure 14. Predominant frames of RNAP binding in 3 starting sequences. The TSS (+1) is underlined in bold face, box frames show -35 and -10 regions. For pMS9_2 and pMS9_3 there are two possible positions of the -35 box.*

### 3.2.4 Binding energy matrices predict the effect of mutations locally, but not between unrelated sequences.

After having identified RNAP binding sites, we moved on to checking how well the effect of mutations is described by the *lacZ* energy matrix. A more complete way to do this would be to fit the full thermodynamic model to the data as described by equation (1) in the introduction. This involves fitting a sigmoidal function mapping binding energy to promoter occupancy and requires fitting the parameter for the chemical potential of RNAP. To keep things simple, I proceed without this step and continue using just binding energies.

Figure 15A shows correlations of binding energy with the fluorescence proxy for the three separate libraries.

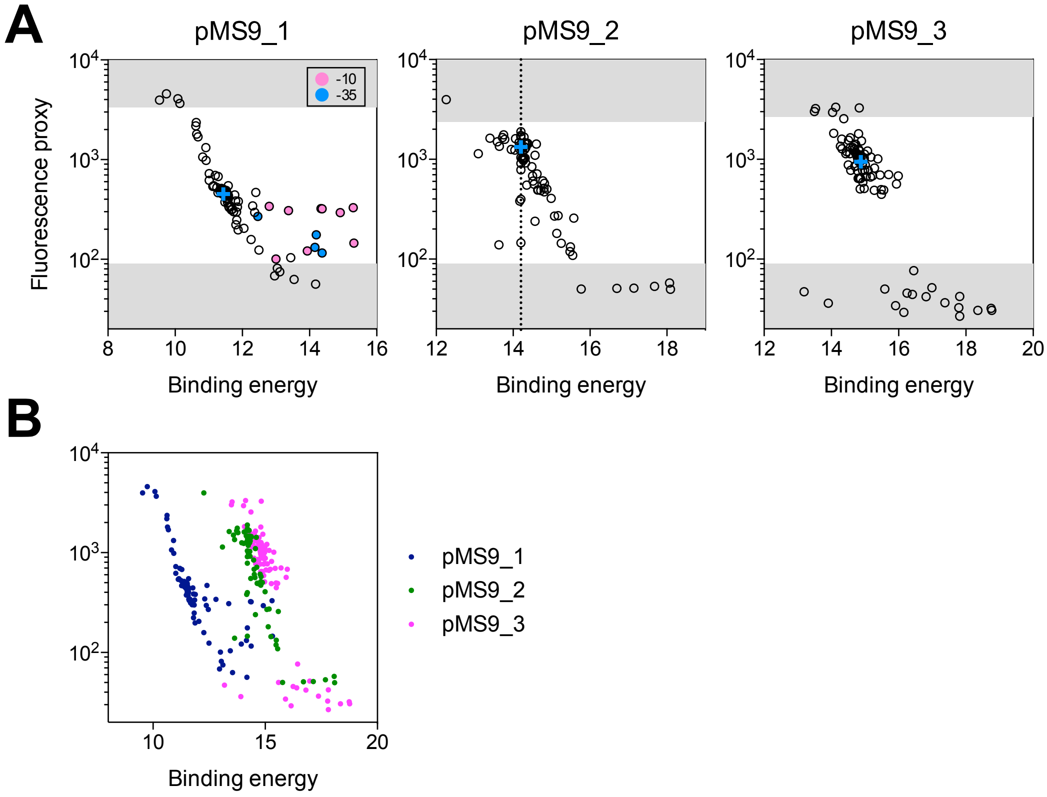


*Figure 15. Binding energy computed with the lacZ matrix and observed fluorescence. (A) Binding energy and fluorescence of three libraries shown separately. Blue cross – starting sequence. Grey areas show fluorescence ranges beyond the inner edges of the outer bins, for which fluorescence cannot be quantified reliably using read distributions. This also means that the apparent 'sigmoid' shape of a fitted curve, which could be expected by theoretical considerations, is not supported by the data (it is not excluded either). Highlighted points in the left panel indicate mutations in the -35 and -10 hexamers that have an unexpectedly small effect on fluorescence. Dashed line in middle panel – identical binding energy predictions of strong-effect mutations downstream of the binding site. (B) Data from the three panels in A overlaid.*

Overall, we find highly negative correlations between energy and fluorescence for all three libraries as expected ($P < 10^{-4}$ for Spearman correlation). Interestingly, a number of mutations expected by the binding energy model to lower expression of pMS9_1, have only a mild effect on expression in our dataset. These mutants locate to the -10 and -35 hexamers (Figure 15A, highlights in left panel).

As noted above, the RNAP binding site in pMS9_2 is only partially overlapping the mutagenized region (Figure 14). This implies that many mutants of the library are located in a region downstream of that covered by the energy matrix. Therefore, in the model, their binding energy is identical to that of the starting sequence. Interestingly, despite no difference in binding energy, fluorescence of these mutants spans an entire order of magnitude (dashed vertical line in middle panel of Figure 15A). This may indicate the importance of the downstream sequence context in which an RNAP binding site is embedded. Alternatively, the unexpected effect of mutations downstream of the hypothetical binding site could however also indicate a second RNAP binding site. In fact, the direction of mutational effects would be consistent with a second RNAP binding frame 17 bp downstream of the ‘main frame’. This second frame does not give correlations for the whole dataset (and thus does not stand out in Figure 13), but this observation raises the possibility that RNAP could bind at multiple positions in a sequence, with expression being the sum (or some other function) of occupancy at the two (or more) positions. We will explore this in more depth in chapter 4.

A particularly puzzling observation is the discordance of the binding energy not within the three libraries, but between them (Figure 15B). Based on binding energies, the pMS9_1 library is expected to have higher expression than the other two, but the opposite is the case. Possible reasons for the ‘energy offset’, i.e. the gap between the predictions for pMS9_1 and the other two libraries, are given in the discussion.

### 3.2.5 Context-dependent effects of promoter mutations

Our dataset allows us to compare the effect of single-nt mutations at corresponding positions within the RNAP binding site in the different contexts of the three plasmid libraries (Figure 16). With a few exceptions, the direction of mutation effects is independent of sequence context. The magnitude of mutation effects can be strikingly different. For example, a G-11A transition (in the -10 hexamer, TATAAT) in the context of the pMS9_2 sequence increases fluorescence by a factor of 23, in the context of pMS9_3 by a factor of 29, and in the context of pMS9_1 by a factor of only 1.4. A C-12A transversion (TATAAT) decreases expression in pMS9_2 by a factor of 4.9 but has close to no effect in either pMS9_1 or pMS9_3. Similar differences in the effect of mutations are found in the -35 hexamer, e.g. G-31C (TTGACA), increases expression in the context of pMS9_1 by a factor of 8, but has close to no effect in the context of pMS9_3.

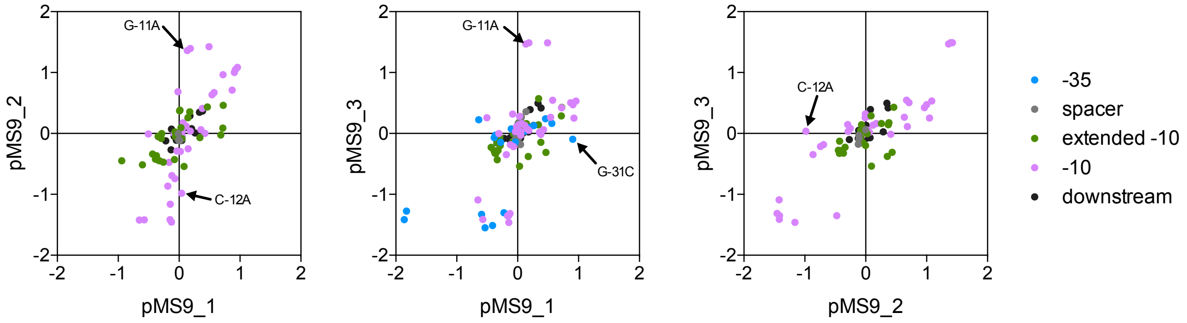


*Figure 16. Context-dependent effects of promoter mutations. Every panel compares respective mutations from two libraries (axes). Every point in the scatterplot represents a particular letter change, e.g. A→G. The values*

*on the two axes are the respective $log_{10}$-fold differences in fluorescence. Points are colored with respect to their position in the RNAP binding site. Arrows indicate mutations mentioned in the text.*

## *3.3 Discussion*

In this chapter, we have seen how single mutations on three short, unrelated sequences, one naturally evolved and two random, affect gene expression, how well mutation effects are captured by a previously published energy matrix, and how much the effect of individual mutations depends on the sequence context in which they occur. We find three surprising results.

The first surprising result is that two randomly chosen starting sequences gave higher expression than the naturally evolved *lac* promoter. In chapter 4 we follow up on this observation and, by looking at many more random sequences, quantify how unexpected this actually is.

The second surprising result is that higher expression from the two libraries with a random starting sequence is not captured by the energy matrix, which predicts the *lac* library to yield highest expression ('energy offset' in Figure 15B). This discrepancy *between* libraries is particularly interesting given the good overall match between model and data *within* the libraries. Also, up-mutations observed in evolution experiments (chapter 2) are correctly retrieved. So while the *lacZ* energy matrix 'works' locally, i.e. it produces mostly correct predictions in a small area of sequence space around a reference sequence, it appears not to work well globally, i.e. in larger sequence space. Possibly, the 'energy offset' is due to the particular fit between the pMS9_1 library and the *lacZ* energy matrix, which reflects that this matrix was inferred on a closely related sequence background. Evidence against this comes from the observation of a comparable energy offset when using two other RNAP binding energy matrices inferred on different sequences (λ $P_L$ and $P_R$, Figure 17; matrices derived by Mato Lagator and Srdjan Sarikas). At this point, it is impossible to say if this result is a peculiarity of the three sequences studied or if it holds more generally. If indeed energy matrix models work well only locally, this raises the question whether there is a universal energy matrix that works comparably well for all possible sequences. If we assume that an energy matrix is a representation of the biophysics of the RNAP, then there should be such a single universal matrix, as there is only a single RNAP. We will address this question in chapter 4.

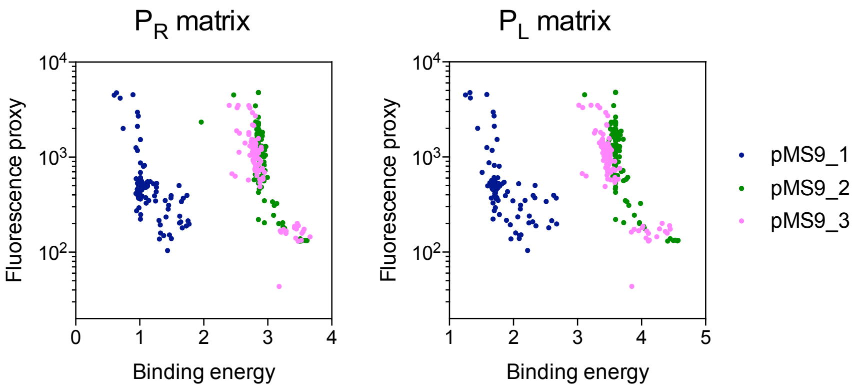


*Figure 17. The energy offset between the pMS9_1 library and the pMS9_2 and pMS9_3 libraries persists when using different energy matrices to calculate the predicted binding energy. Left and right plots: matrix inferred on a mutant library of the λ $P_R$ and $P_L$ promoter respectively (Mato Lagator and Srdjan Sarikas).*

Lastly, we find that corresponding mutations can have dramatically different effects depending on local sequence context. The results of this chapter, in particular the 'energy offset' and the context-dependency of mutations, indicate that our models of the sequence dependency of promoter strength are incomplete in important ways. We can come up with several hypotheses in which way this could be the case. In the following, we discuss four such hypotheses in more detail.

## Epistasis

RNAP binding energy may not be the sum of energy contributions of individual interacting positions. This violates the additivity assumption and calls for models that incorporate epistatic interactions between nucleotide positions, which in principle can be addressed experimentally, but quickly becomes intractable if one seeks to cover interactions between all positions.

## Alternative promoter types

The rate of transcription initiation is known to not be simply proportional to the sequence-dependent equilibrium binding probability of RNAP. Instead, already in the 80s, the existence of at least two sequence-dependent steps on the initiation pathway was discovered (McClure et al. 1983; Studnicka 1988), which led to the formulation of the 'bipartite model' of promoter function, which assumes the sequences of the -35 and -10 boxes to determine RNAP binding and promoter isomerization respectively. Recent studies have improved our understanding of the sequence determinants at these two important steps of promoter function (E. Heyduk & T. Heyduk 2014; Feklistov et al. 2006; Ruff et al. 2015; Hook-Barnard & Hinton 2007; Djordjevic & Bundschuh 2008), resulting not in a rejection of the bipartite model, but rather in its refinement. Thus the single-step assumption is certainly violated. The question is, how much this provides a complication. In the best case (from the perspective of modeling for the purpose of minimizing errors), sequence dependent effects at the promoter isomerization step simply ‘blend in’ mathematically and show up in the energy entries in our inferred matrices without introducing any distortions. In the worst case, there may actually be multiple distinct classes of promoters, possibly ‘living’ in disconnected areas of sequence space, and only a subset of them is described well by our models. ‘Alternative promoter types’ may then call for different equations or ‘matrices’, possibly containing highly epistatic interactions between different sequence positions. One suggestion has been to describe the two-step process of transcription initation using Michaelis-Menten kinetics (Ruff et al. 2015).

The ‘alternative promoter types’ hypothesis is nourished mainly by the observation of strong promoters with a relatively low homology score (exemplified by $\lambda P_L$ (Knaus & Bujard 1988)) or by the observation of individual mutations that influence transcription in the opposite direction than expected from consensus (Miroslavova & Busby 2006). We will consider this hypothesis in more detail in chapter 4.

### Multiple RNAP binding positions

We can also question our assumption that transcription is initiated only at the position of the RNAP binding energy minimum. Our finding of strong effect mutations outside of the supposed primary RNAP binding site in the pMS9_2 library may support that promoter activity emerging from random sequences is the combined result of multiple very weak RNAP binding sites, but more data is needed and will be provided in chapter 4.

### Local differences in the chemical potential of the RNAP

The offset between predicted binding energies (Figure 15B) could be due to a violation of the assumption that the same chemical potential of RNAP applies to all three libraries. There are two possible problems.

The first is a problem at the stage of applying models. At the time of inference of the energy matrix that I use (Kinney et al. 2010), the chemical potential, an additive term to binding energy (Eq. 1), was inferred as well. It might be flawed thinking that one can simply take energy values scaled in this way and apply them to a different experimental context. Possibly, a different value for the chemical potential, and analysis at the level of predicted binding ($P_{on}$) instead of predicted binding energy (E) is needed. Earlier, Murat Tugrul tried exactly this and calculated predicted binding probabilities for the three libraries using different values for the chemical potential and the same energy matrix that I use. Importantly, he could not find a value that would reconcile the offset between predictions, and correlation coefficients were low across the tested range of chemical potential values (Figure 18).

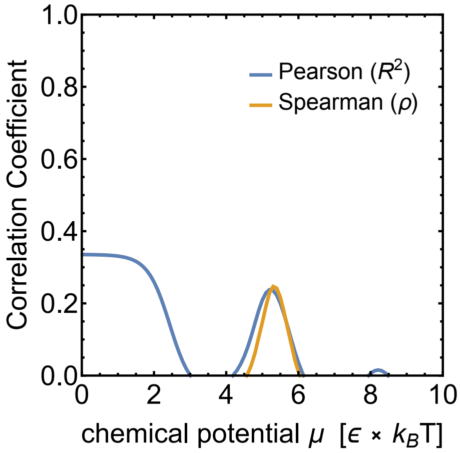


*Figure 18. Pearson and Spearman correlation coefficients between predicted and observed expression of the three single-nt libraries (pMS9_1, pMS9_2 and pMS9_3, all pooled) as a function of chemical potential. This Figure and the underlying analysis is the work of Murat Tugrul.*

The second problem concerns an actual biophysical question. There could be differences in the actual chemical potential of RNAP between the three libraries. Given however, that everything except 36 bp is identical between the three libraries and respective experiments, this explanation seems unlikely. Still, one could imagine that the three different starting sequences impose a particular sterical configuration on the DNA, making it more or less accessible to RNAP, changing its chemical potential.

Together, our results demonstrate the need for caution when using energy matrices inferred on a particular sequence background to predict expression from unrelated sequences. Obtaining RNAP binding energy matrices from a much more diverse sample of the sequence space may be essential in alleviating this problem. This is what we do in chapter 4.

## *3.4 Materials and Methods*

### 3.4.1 Plasmid cloning

We used plasmid pUA66-lacZ (Zaslaver et al. 2006) as a starting point for plasmid construction. The four 36 nt long sequences to be mutagenized in the next step were synthesized as middle part of oligonucleotides of length 100 nt. We put sequences of length 32 nt both upstream and downstream of the 36 nt core sequence. These flanking sequences serve as homology in the plasmid assembly step. Their sequence was obtained by a random shuffling of the sequences flanking the RNAP binding site in the pUA66-lacZ contained promoter fragment. The oligonucleotides (1_lacZ, 2_lacZscrambled, 3_p0) were made doublestranded using primer novo_Klenow and Klenow fragment. The pUA66-lacZ backbone was linearized using PCR amplification with primers novo_ohup and novo_ohdown. The backbone linearized in this way contains the gfpmut2 reporter gene, but leaves out the lacZ promoter fragment originally contained in pUA66-lacZ. We assembled plasmids by combining the 100 nt doublestranded fragments and the backbone fragment using an NEBuilder kit. The resulting plasmids were designated pMS9_1, pMS9_2 and pMS9_3.

The control plasmid pMS9_control was created using a Q5 site directed mutagenesis kit with primers letitshine_f and letitshine_r, and was transformed into NEB 5α cells. All newly cloned inserts were verified by sequencing.

### 3.4.2 Creation of single-nt libraries from four starting sequences of length 36 nt

Plasmids pMS9_1 to pMS9_3 were used as starting plasmids for library mutagenesis. For each plasmid, we created two pools of 18 primers each, one pool serving as forward primer (L) and one as backward primer (R). Each of the 18 primers constituting a pool was ordered such that one nucleotide of the starting sequence was replaced by an equiprobable mixture of the three alternative nucleotides (e.g. A → B = 33% C / 33% G/ 33% T). In this way, a pool of 18 primers contains all single-nt variants of one half of the 36 nt sequence to be mutagenized. At the 3' end of the primers we put a constant region homologous to sequences on the plasmid backbone. Next, we synthesized 6 plasmid pools using a Q5 site directed mutagenesis kit. For every reaction we used one starting plasmid (e.g. pMS9_1) as template, one primer pool as forward primer and a single constant reverse primer (e.g. L1_pool + R1_constant). The resulting plasmid pools were transformed into chemically competent cells (NEB 5α), incubated for 1 h at 37°C and plated on LB plates with Kanamycin (50 µg/mL) and sterile charcoal (5g / L to reduce background fluorescence). For each plasmid pool, we plated the undiluted cultures on three plates. Plates were incubated for 48 h and colonies were scraped off. After scraping, suspensions were vortexed vigorously and diluted to an $OD_{600}$ of 1, aliquoted (100 µL) and frozen after addition of glycerol (50%, 40 µL).

### 3.4.3 FACS-sorting

Prior to sorting, cells were grown in freshly filtered (0.22µm) M9 minimal medium with 0.2% CAS, 0.2% Glucose and 50 µg/mL kanamycin. Frozen aliquots of plasmid pools, starting plasmids and the control plasmid were diluted 1:10 and grown overnight. Prior to sorting, overnight cultures were diluted again 1:100 and grown for 3 h to reach exponential phase.

FACS-sorting was performed on an FACS Aria III flow cytometer (BD Biosciences, San Jose, CA) with a 70 µm nozzle for droplet formation. A 488 nm laser was used to detect forward scatter (FSC) and side scatter (SSC) with a 488/10 band-pass filter. The same laser was used for excitation of GFP (FITC channel, emission filters 502LP, 530/30). We chose the FITC channel voltage such that the median fluorescence of a plasmid-free auto-fluorescence control sample (AF) is between 0 and 100 on the FITC axes. The flow rate was set to 1.0 and samples were diluted to obtain a cell count of approximately 5000 events/second. Cells for sorting were manually gated on the densest population in an FSC/SSC scatter plot, which comprised 97-98% of all events exceeding a threshold of 1000 on the SSC axis. Six sorting gates were set on the FITC axes as follows: First we recorded autofluorescence of a culture with a plasmid lacking GFP. The median autofluorescence (42) served as upper boundary of the lowest bin (B1). Then, to obtain the three libraries, we mixed respective plasmid pools containing the left and the right mutagenized half of the 36 nt insert. For each of the resulting three libraries, we recorded fluorescence of $10^6$ cells. The lower bound of the highest bin (B6) was then taken to correspond to the 95th percentile of the fluorescence distribution. The three boundaries between intermediate bins (B2-B5) were then chosen with equidistant spacing on log scale. This procedure was done for each of the three libraries individually. Bin boundaries can be found in Table 3.

*Table 3. Upper bin boundaries for FACS.*

| *Starting plasmid* | *B1* | *B2* | *B3* | *B4* | *B5* |
|---|---|---|---|---|---|
| *pMS9_1* | 42 | 126 | 375 | 1122 | 3353 |
| *pMS9_2* | 42 | 115 | 315 | 863 | 2365 |
| *pMS9_3* | 42 | 118 | 334 | 942 | 2656 |

The number of cells to be sorted into each of the six bins B1-B6 corresponded to the number of cells previously recorded in each of the bins and can be found in Table 4. Before sorting, a culture of cells with the starting plasmid was added to the library culture at a ratio of 1:100. Cells were sorted into a 24-well plate with 500 µL sorting medium / well. The recipient plate was cooled to 4 °C to halt growth while sorting to other wells was still going on.

*Table 4. Number of cells sorted from each library.*

| *Starting plasmid* | *B1* | *B2* | *B3* | *B4* | *B5* | *B6* |
|---|---|---|---|---|---|---|
| *pMS9_1* | 34140 | 89728 | 247835 | 410198 | 171421 | 49810 |
| *pMS9_2* | 123504 | 206937 | 278145 | 167228 | 177238 | 49822 |
| *pMS9_3* | 61146 | 65106 | 56345 | 355932 | 413442 | 49804 |
| *pMS9_control* | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |

After completion of sorting, 1000 cells of the culture with the control plasmid pMS9_control were added into each of six wells. Sorted cells were spun down in a cooled centrifuge and resuspended in 1 mL medium. We then plated a dilution from each well on LB Kan to estimate viability (mean viability over 6 bins and 3 libraries was 62%, standard deviation 11%) and the frequency of mis-sorting (mean outlier frequency over 6 bins and 3 libraries was 2.2%, standard deviation 2.5%, outlier classification using ROUT with Q=1%). Finally, the cells from each bin and the different libraries (columns in Table 4) were pooled and grown overnight.

### 3.4.4 Plasmid library isolation and barcoding PCR

We isolated plasmid from the six culture pools and quantified DNA concentration using a Nanodrop spectrophotometer. Given the number of cells sorted and the plasmid pool concentrations, every sorted cell is expected to contribute 100 plasmid molecules or more to 1 ng of plasmid pool DNA, which is the amount of template we used in the subsequent PCR amplification step.

For barcoding PCR products containing the mutagenized region, we created primers mutseq_f1-6 and mutseq_r1-6. They contain a 3' constant region, a bin-specific barcode of 5 nt and a constant 5' tail of 5 nt.

PCRs were performed using Q5 high fidelity polymerase and 1 ng of the plasmid pools as template in a 50 µL reaction. We first performed five cycles with an annealing temperature calculated for the constant 3’ part of the primers, followed by 25 cycles using an annealing temperature matched to each of the full length primer pairs.

PCR products were column-purified (Zymo research, Irvine, CA) and eluted in 30 µL, of which 2 µL were run on an agarose gel for relative product quantification based on band fluorescence. PCR products were finally pooled to reach approximately equimolar concentrations of the six reaction products.

### 3.4.5 Illumina sequencing

We sent ~1 µg of pooled PCR product to sequencing by GATC biotech (Konstanz, Germany) on an Illumina sequencer (125 bp paired end).

### 3.4.6 Debiasing read distributions and calculating a fluorescence proxy

The sequencing raw data was processed by Srdjan Sarikas. For our analysis, we only used reads with matching barcodes in the forward and reverse primers and single nt mutations in the mutated core region. To account for biases in the sort-seq process, we normalized the number of reads from each bin and sequence by dividing by the number of reads of the control sequence (Table 5).

*Table 5. Distribution of control sequence reads across bins*

| *Bin* | *B1* | *B2* | *B3* | *B4* | *B5* | *B6* |
|---|---|---|---|---|---|---|
| *# of control sequence reads* | 9025 | 9223 | 6544 | 3616 | 5707 | 25394 |

In the (purely theoretical) absence of biases, the number of control reads would be the same for each bin, given the same number of cells of pMS9_control added to the sorted cultures. Then, for each single nt mutant sequence, we calculated its fluorescence value as the geometric mean from debiased (i.e. control-normalized) read distributions. For doing so, we used median bin fluorescence values of the whole library distribution recorded in FACS. We chose to use the geometric mean over the arithmetic mean to reflect that scatter in FACS appears symmetric on a log axis. We verified the effect of debiasing by applying it to the count distribution of starting sequences, the expression of which is known from FACS of clonal cultures (Figure 11).

For subsequent analysis, we only used mutant sequences with ≥50 reads.

### 3.4.7 Calculation of binding energies

For calculating binding energies, we used a previously published energy matrix inferred from a sort-seq dataset of the mutagenized *lac* promoter in *E. coli* (Kinney et al. 2010). The matrix entries were shared by Murat Tugrul and can be found in Table 6. For a matrix of spacer length 17, we deleted line 16 from the matrix.

*Table 6. LacZ energy matrix with spacer length 18. For the lacZ energy matrix with spacer length 17, line 16 was omitted.*

| | A | C | G | T |
|---|---|---|---|---|
| 1 | 0.0030200 | 0.1680000 | 0.0218000 | 0.1440000 |
| 2 | 0.0582000 | 0.4190000 | 0.0000259 | 0.5030000 |
| 3 | 0.1100000 | 0.0510000 | 0.0864000 | 0.0000129 |
| 4 | 0.0000000 | 0.1950000 | 0.0456000 | 0.1860000 |
| 5 | 0.8440000 | 0.0880000 | 0.0000000 | 1.6300000 |
| 6 | 2.7700000 | 2.7300000 | 2.6200000 | 0.0000000 |
| 7 | 1.2800000 | 2.9400000 | 1.0300000 | 0.0000000 |
| 8 | 3.8000000 | 3.2200000 | 0.0000000 | 1.6900000 |
| 9 | 0.0000000 | 1.6800000 | 0.9690000 | 1.6100000 |
| 10 | 2.0100000 | 0.0000000 | 2.7500000 | 1.5200000 |
| 11 | 0.0000000 | 1.0600000 | 0.8250000 | 0.3960000 |
| 12 | 0.0000000 | 0.4200000 | 0.3150000 | 0.0742000 |
| 13 | 0.0183000 | 0.1590000 | 0.0018200 | 0.0970000 |
| 14 | 0.0129000 | 0.1120000 | 0.0042000 | 0.0672000 |
| 15 | 0.0874000 | 0.0640000 | 0.0363000 | 0.0000639 |
| 16 | 0.0000000 | 0.2740000 | 0.0706000 | 0.3740000 |
| 17 | 0.2140000 | 0.0936000 | 0.0440000 | 0.0000122 |
| 18 | 0.0000000 | 0.2880000 | 0.1580000 | 0.5620000 |
| 19 | 0.1840000 | 0.0072100 | 0.0034600 | 0.2510000 |
| 20 | 0.0813000 | 0.0250000 | 0.0931000 | 0.0002750 |
| 21 | 0.1270000 | 0.0995000 | 0.1470000 | 0.0000000 |
| 22 | 0.0179000 | 0.0634000 | 0.1140000 | 0.0011600 |
| 23 | 0.0738000 | 0.1080000 | 0.0821000 | 0.0000000 |
| 24 | 0.1210000 | 0.2880000 | 0.3130000 | 0.0000000 |
| 25 | 0.6760000 | 0.6050000 | 0.6110000 | 0.0000000 |
| 26 | 0.0000000 | 0.8070000 | 0.2950000 | 1.1500000 |
| 27 | 0.3670000 | 0.5720000 | 0.0000000 | 0.5210000 |
| 28 | 1.0800000 | 1.9000000 | 0.0000000 | 1.1100000 |
| 29 | 0.4400000 | 0.4480000 | 0.0000000 | 0.1540000 |
| 30 | 1.3700000 | 1.9400000 | 2.9100000 | 0.0000000 |
| 31 | 0.0000000 | 3.8700000 | 3.8900000 | 3.4800000 |
| 32 | 0.3680000 | 0.8740000 | 0.9370000 | 0.0000000 |
| 33 | 0.0000000 | 1.0500000 | 1.2900000 | 1.3800000 |
| 34 | 0.0000000 | 0.6000000 | 1.2800000 | 1.3500000 |
| 35 | 1.5700000 | 2.9400000 | 2.5000000 | 0.0000000 |
| 36 | 0.0000000 | 0.8420000 | 0.4200000 | 0.7250000 |
| 37 | 0.0511000 | 0.6130000 | 0.1630000 | 0.0000001 |
| 38 | 0.0000000 | 0.3380000 | 0.2830000 | 0.2500000 |
| 39 | 0.0803000 | 0.1860000 | 0.0422000 | 0.0000527 |
| 40 | 0.0000000 | 0.4800000 | 0.4600000 | 0.3420000 |
| 41 | 0.0225000 | 0.1110000 | 0.1780000 | 0.0020500 |

To calculate the binding energy for a sequence, matrix entries corresponding to the given nucleotides at respective positions are summed up. Binding energy was calculated for each possible frame overlapping the variable 36 nt region.

To identify the predominant frame of RNAP binding, we calculated the Spearman rank correlation coefficient between the fluorescence proxy as described in section 3.4.6 and binding energies of all mutant sequences. This was done for each possible frame. P-values of spearman rank correlations were adjusted by multiplication with the number of frames tested. Mutants with a fluorescence proxy beyond the upper edge of the lowest bin and the lower edge of the highest bin were excluded for calculating correlations, because fluorescence in these limits cannot be properly quantified using our approach.

### 3.4.8 List of primers

Cloning primers

1_lacZ-RNAP binding site
CACGAGGCCAGGCTTCAAATCTCAATGCTATTGGCTTTACACTTTATGCTTCCGG
CTCGTATGTTGTGTGTGCATACAGATTGAGTAATGGCATCGAAAC

2_lacZscrambled
CACGAGGCCAGGCTTCAAATCTCAATGCTATTTCGGCTTTCTTCGTGCATAATGC
TTCGGTCTATGGTGTGCATACAGATTGAGTAATGGCATCGAAAC

3_p0
CACGAGGCCAGGCTTCAAATCTCAATGCTATTTTACCTTGCAGGAATTGAGGCCG
TCCGTTAATTTCCTGTGCATACAGATTGAGTAATGGCATCGAAAC

| | |
|---|---|
| novo_Klenow | GTTTCGATGCCATTACTCAATC |
| novo_ohup | TAGCATTGAGATTTGAAGCCTGGCCTCGTG |
| novo_ohdown | TGCATACAGATTGAGTAATGGCATCGAAAC |
| letitshine_f | TAAAGCCATATTAACGAATGTGCATACAGATTGAGTAATG |
| letitshine_r | GGTAATTTAGGTTTCCAGAATAGCATTGAGATTTGAAGC |

Barcoding primers

| | |
|---|---|
| mutseq_f1 | AAGCTATCTATCGTCTTCACCTCGAGCAC |
| mutseq_f2 | AAGCTGTACATCGTCTTCACCTCGAGCAC |
| mutseq_f3 | AAGCTAAGTGTCGTCTTCACCTCGAGCAC |
| mutseq_f4 | AAGCTCTCGTTCGTCTTCACCTCGAGCAC |
| mutseq_f5 | AAGCTATAACTCGTCTTCACCTCGAGCAC |
| mutseq_f6 | AAGCTCGTCATCGTCTTCACCTCGAGCAC |
| mutseq_r1 | CGTACATCTATTCTCCTTTACTCATATGTATATCT |

| | |
|---|---|
| mutseq_r2 | CGTACGTACATTCTCCTTTACTCATATGTATATCT |
| mutseq_r3 | CGTACAAGTGTTCTCCTTTACTCATATGTATATCT |
| mutseq_r4 | CGTACCTCGTTTCTCCTTTACTCATATGTATATCT |
| mutseq_r5 | CGTACATAACTTCTCCTTTACTCATATGTATATCT |
| mutseq_r6 | CGTACCGTCATTCTCCTTTACTCATATGTATATCT |

# 4 The distribution and prediction of promoter function in a random sample of the full sequence space

This chapter is the result of a collaboration with Srdjan Sarikas, and Gašper Tkačik in an advisory role. Srdjan Sarikas processed the raw sort-seq data, contributed to the development of data filtering criteria and applied filtering, contributed to the development of thermodynamic models and implemented their inference, and provided me with the model output.

## *4.1 Introduction*

In the final chapter of this thesis, we return to a number of questions brought up in chapter 3. We start by first addressing a new question:

- How frequently do random sequences exhibit transcriptional activity?

We approach this question by performing a similar sort-seq experiment as in chapter 3, but this time we measure expression from a plasmid library in which the variable regions of the pMS9 plasmids are replaced by a stretch of 36 random nucleotides (36N). In this way, we sample a much larger and more disperse area in sequence space.

In addition, we develop an extended thermodynamic model to predict expression from sequence. We allow an energy-penalized flexibility in the length of the spacer separating the -35 and -10 boxes, and we test the effect of summing contributions over multiple possible binding positions of RNAP within the random upstream sequence. Combining experimental data and outputs of the model, we address the following open questions from chapter 3.

- Is there an RNAP energy matrix that can make better predictions of transcription from a wide variety of sequences than matrices locally inferred on model promoters? How is such a matrix different from model promoter matrices?
- What is the distribution of spacer lengths in random promoters and how important are non-canonical spacer lengths for accurately modeling expression?
- Is there evidence that promoter activity of weakly transcribing sequences is driven by multiple RNAP binding sites?
- Is there evidence for distinct promoter types that are optimized for different steps in the transcription initiation process, in particular is there evidence of an ‘extended -10 type promoter’?

In a very recent related study, functional promoters driving expression of the *lac* operon were evolved from random sequences by selection for growth on lactose minimal medium (Yona et al. 2018). Earlier work using synthetic selection for higher expression in FACS also started from random sequences in *E. coli*, but the authors were mainly interested in the noise properties of emerging promoters, and not in the genotype-phenotype map of promoter strength (L. Wolf et al. 2015). In eukaryotes, a recent high-throughput study in yeast investigated promoter function using random libraries, with the added complication of transcription factor binding sites (de Boer et al. 2018). Our focus on functionality over a wide range from non-functional to strong binding and the associated technical challenges is shared with a recent article that quantifies eukaryote protein-DNA binding affinity in vitro (Rastogi et al. 2018). To our knowledge, no other study has investigated the genotype-phenotype map of bacterial promoter strength in a quantitative way using large random sequence libraries.

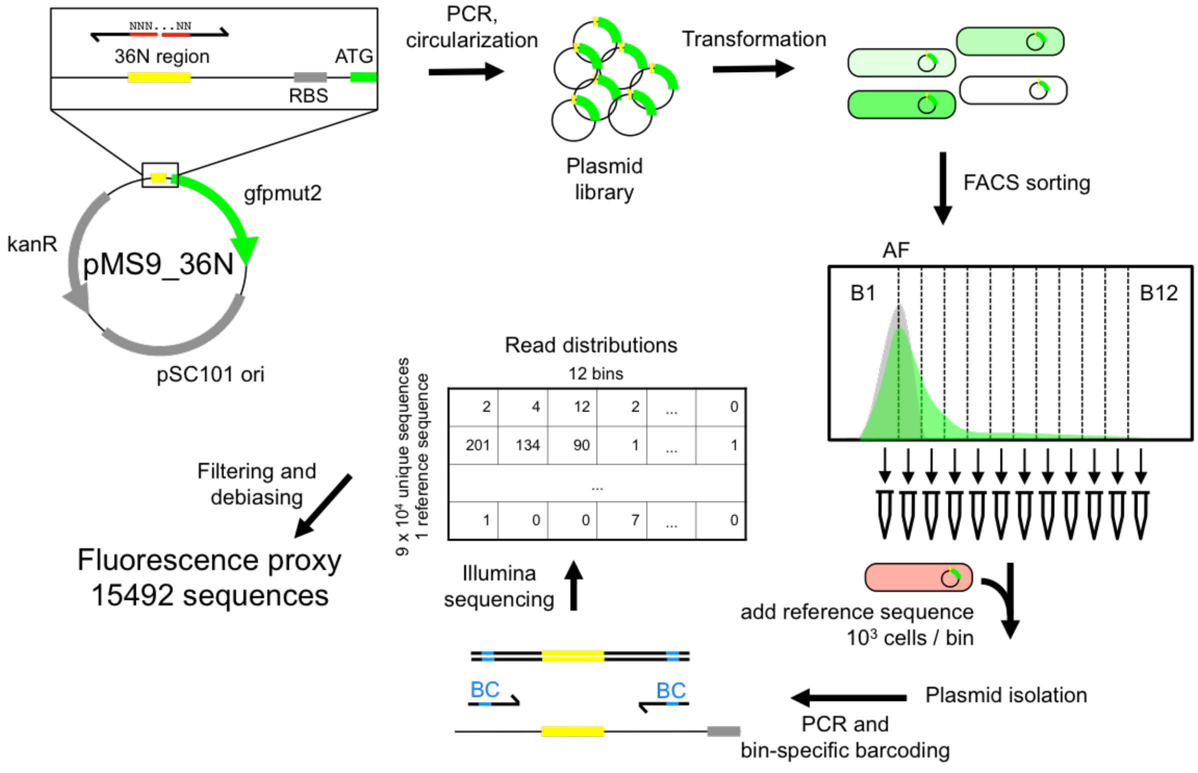
## *4.2 Results*

### 4.2.1 A sort-seq experiment for quantifying fluorescent reporter expression from a random sequence promoter library

Analogous to library creation in chapter 3, we created a plasmid library (pMS9_36N) using a variation of site-directed mutagenesis to insert a 36 nt long random sequence in front of a *gfp* reporter gene (Figure 19). Expression from the 36N inserts was measured by sorting cells transformed with the library into 12 bins according to GFP fluorescence, followed by plasmid isolation, bin specific barcoding of the variable inserts, and Illumina sequencing. In this way, we obtained a fluorescence proxy for 15492 unique clones that we use for further analyses (see Methods). As in chapter 3, we perform debiasing using read counts of a reference sequence (Table 8), added in equal cell numbers to each bin after sorting. Debiasing is particularly important given the skewed distribution of fluorescence in our library, with many more sequences in lower than in higher bins. As in chapter 3, the number of reference sequence counts in each bin is high enough ($>10^3$, Table 8), so that division by small numbers is no concern.

We compared fluorescence proxies of a dataset of 78 clones that we obtained after sorting to the fluorescence measured using a platereader. For clones with little expression, both methods are limited by autofluorescence background, which means we cannot expect a correlation between measurements. On the other end of the scale, if a clone is mostly sorted into the highest bin, the fluorescence proxy calculated from sort-seq becomes unreliable too.

This is because the highest bin has no upper bound, which means that the very high fluorescence (much higher than the lower bound of the bin) will be underestimated by the fluorescence proxy. When excluding the problematic clones on the very low and high end, we found a strong linear correlation between the two fluorescence (Figure 20). The origin of a small remaining non-linearity in the white region of Figure 20 remains unclear and probably signifies that platereader and FACS fluorescence are not perfectly comparable across the measurement range for inherent technical reasons. Since both measurements offer only indirect information on promoter activity, which is what we are eventually after, we decided not to investigate this discrepancy in more detail to find out which of the two measurements is 'right'. We conclude that, at the level that is of interest to us, platereader measurements validate the sort-seq approach across approximately three orders of magnitude of fluorescence.



| | | | | | |
|---|---|---|---|---|---|
| 2 | 4 | 12 | 2 | ... | 0 |
| 201 | 134 | 90 | 1 | ... | 1 |
| | | ... | | | |
| 1 | 0 | 0 | 7 | ... | 0 |

*Figure 19. Workflow for creating the 36N random promoter sequence library and measuring fluorescence using sort-seq. Starting from the top left: Creation of a plasmid library pMS9_36N. A stretch of 36 random nucleotides ('N') is inserted upstream of an RBS and a gfp reporter gene using PCR with 18N degenerate 5' primer ends. PCR products are circularized and transformed into E. coli. Cells are sorted according to GFP fluorescence into twelve bins (B1-B12). ). Green FACS histogram – cartoon of library fluorescence; grey FACS histogram, AF – autofluorescence background. After sorting, a constant number of cells containing a reference sequence (red cell) is added for later debiasing. Isolated plasmid libraries are subsequently used as PCR templates with primers that incorporate bin-specific barcodes into PCR products used in Illumina sequencing. Finally, read count distributions of every sequence across bins are filtered and debiased by dividing by bin counts of the reference sequence. Debiased distributions are then transformed into a fluorescence proxy. For details see Methods.*

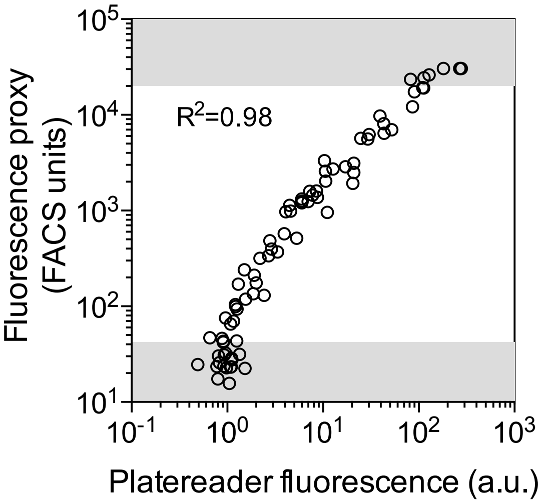


*Figure 20. Correlation between platereader fluorescence (x-axis) and the sort-seq-derived FACS fluorescence proxy (y-axis) for 78 clones. $R^2$ is the linear Pearson correlation coefficient calculated excluding points with a fluorescence proxy in the lowest or highest bin (grey shading).*

## 4.2.2 The distribution of fluorescence from random sequences

Fluorescence from random clones spanned three orders of magnitude on the FACS scale (Figure 21A). Fluorescence from the strongest expressing sequences approached that of the strong phage promoter $P_L$ (dashed line in Figure 21A). 9.1% of the library (1414 clones) showed fluorescence exceeding the 95th percentile of a no-plasmid autofluorescence control culture (AF95). For the rest of this chapter we refer to clones exhibiting fluorescence larger than AF95 as 'functional'.

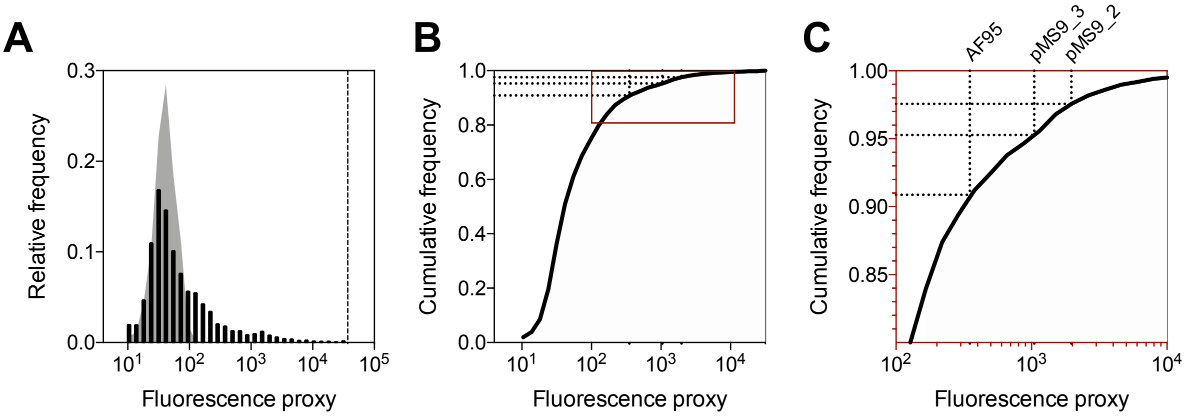


*Figure 21. Distribution of fluorescence of the pMS9_36N library. (A) Histogram of fluorescence of 15429 clones in the pMS9_36N library (black bars) and distribution of the FACS autofluorescence background from a plasmid-free culture (grey shaded area). Dashed line indicates median FACS fluorescence of the pMS9_PL reporter. (B) Cumulative frequency of library fluorescence. (C) Magnification of boxed area in (B). Dashed lines indicate the 95th percentile of the autofluorescence background (AF95), and the fluorescence proxy of the starting sequences pMS9_2 and pMS9_3 from chapter 3.*

We evaluated where fluorescence from the two random starting sequences used in chapter 3 fell with respect to the full distribution in the 36N sequence space. Fluorescence from both pMS9_2 and pMS9_3 plasmids exceeded that of 95% of the 36N distribution (Figure 21), which means that generating two sequence of such fluorescence in two attempts by chance is indeed highly unexpected (with a chance of 0.4%), but not extremely unexpected.

### 4.2.3 An extended thermodynamic model to predict expression from sequence

Due to the limitations of existing models predicting expression from sequence as described in chapter 3, we developed an extended thermodynamic model, inferred its parameters on a training subset of the 36N dataset, and did the same with simpler models for comparison. We tested the effect of introducing two extensions to the model (Figure 22). The first extension is to allow a flexible spacer length between the -35 and -10 boxes. Energy penalties of suboptimal spacer lengths are inferred from the data. The second extension is to allow multiple additive binding sites of RNAP to contribute to expression. Although this includes the possibility of overlapping RNAP binding sites, which may interfere with each other rather than adding up (M. L. Peterson & Reznikoff 1985), the rational is that interference is unlikely when binding probabilities are low overall. Specifically, we count the Boltzmann weights of all possible binding sites overlapping the 36N region, as opposed to taking the Boltzmann weight of only a single energy minimum position of RNAP binding.

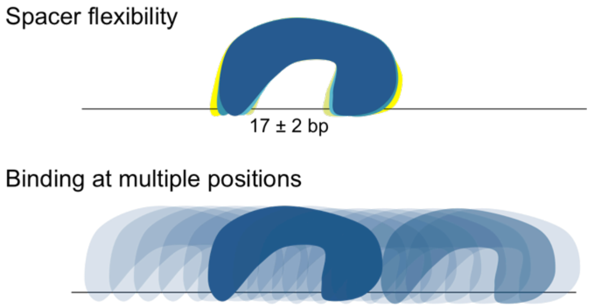


*Figure 22. Two extensions of the thermodynamic model of promoter strength.*

The fitted parameters of the models include energy matrix entries for the two ‘feet’ of RNAP (comprising the -35 box and the -10 box), chemical potential and an energy scale parameter, and, in the case of a flexible spacer, energy penalties for suboptimal spacer lengths between 15 bp and 19 bp. The model fitting procedure is outlined in the Methods section and will be published in full detail elsewhere (Srdjan Sarikas).

#### 4.2.3.1 Model selection

Figure 23 shows scatterplots for overall performance of the four tested models (with and without spacer flexibility, and with and without multiple RNAP binding positions). All parameters are inferred separately for the four models. Allowing energy-penalized spacer flexibility clearly improves correlations between predicted promoter occupancy and fluorescence. Allowing multiple RNAP binding positions has a much smaller, but positive effect on correlations.

Due to the best overall performance of the 'full' model, i.e. the model that includes both spacer flexibility and multiple binding sites (lower right scatter plot in Figure 23), we conclude that this model offers the most accurate description of the actual biophysical process generating the observed variation in the data and analyze the output of this model in more detail in the following sections.

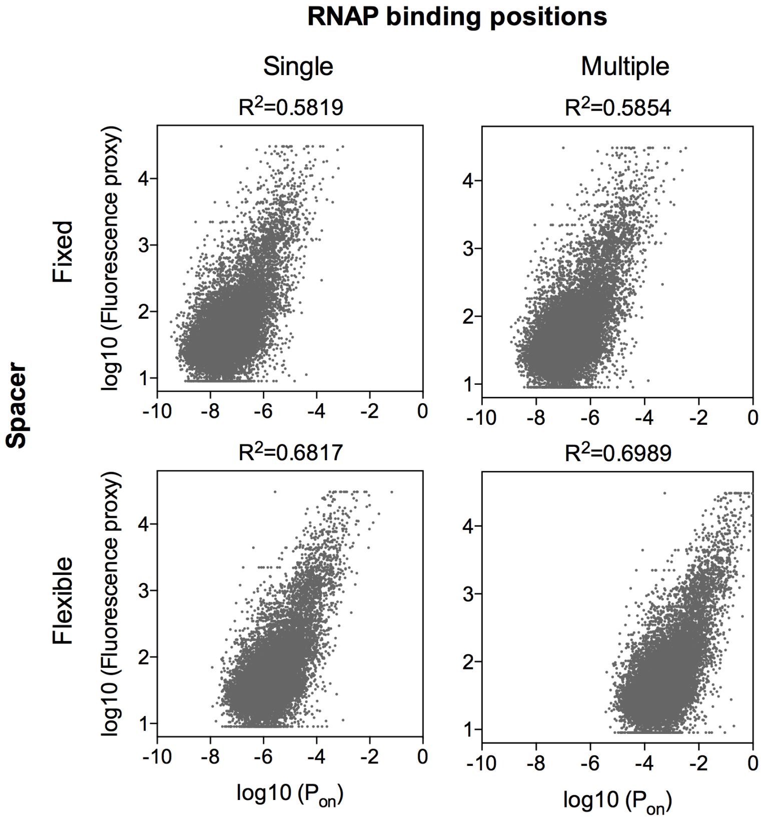


*Figure 23. Model selection. Scatter plots of promoter occupancy predictions (x-axes) and fluorescence (y-axes) for the simplest model (top left) and two model extensions (flexible spacer - bottom panels; binding at multiple sites – right panels). The slope of the fits is constrained to 1. $R^2$ is the weighted Pearson correlation coefficient. For details, see Methods.*

#### 4.2.3.2 The distribution of primary RNAP binding sites across the upstream region of the GFP reporter

The 36N library has a large enough variable region to accommodate a full RNAP binding site (29 bp), but there is nothing restricting RNAP binding outside or partially outside the variable 36N region. We can exclude that promoter function of the library is dominated by a binding site fully outside of the 36N region, as variation outside of the binding position is not expected to affect expression as much as seen in the library (Figure 21). However, a partial overlap of RNAP binding sites with the constant up- or downstream flanking region on the plasmid, could be consistent with the observed variation in fluorescence. If a particular 'half-site' in the flanking region were dominating promoter function in the 36N dataset, results would be highly specific to the flanking regions, a serious problem when trying to generalize results.

Before anything else, we therefore check if this is the case by inspecting the distribution of 'primary' RNAP binding sites identified by the model on the region upstream of the GFP reporter (Figure 24). By 'primary' RNAP binding site we mean the single binding site of every sequence that contributes maximally to the binding probability. Only 15% of all clones have a primary binding site with both of the complete -35 and -10 hexamers in the 36N region. This is lower than expected if RNAP sites were distributed evenly across the considered interval (~20%). Also, one position upstream of 36N frequently provides a -35 box (TTCAAA, first green peak in Figure 24). A second frequently predicted position of the -35 box (second high green peak in Figure 24) also overlaps the constant region, but only with an initial 'T' (TNNNNN). Despite these potential biases by overrepresentation of sequences with a -35 in the constant region, the distribution of fluorescence across possible binding site positions (boxplots in Figure 24) shows that there are no particular positions outside of the 36N region that dominate expression in the library. We therefore conclude that frequent positioning of the -35 box (and to a lesser extent, of the -10 box) in the constant flanking region may influence model fitting and introduce biases in the following analyses, but these biases do not introduce strong effects.

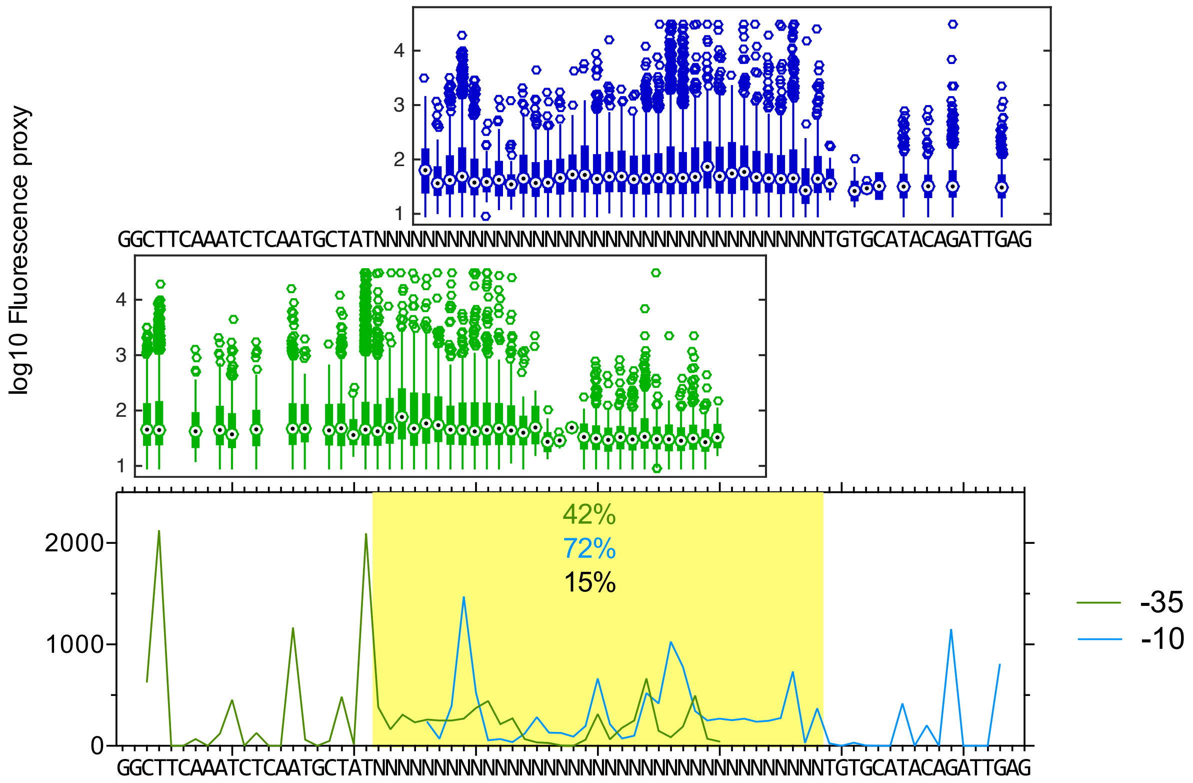


*Figure 24. Distribution of primary RNAP binding sites across the 36N variable region and left and right constant flanking regions. All panels are aligned to the sequence shown between the two boxplot panels. Bottom panel – frequency of clones with the -35 nucleotide (green line) and the -10 nucleotide (blue line) across the sequence. The 36N variable region is highlighted in yellow. Percentages are the number of clones for which the complete 6 bp -35 box (green), -10 box (blue), and both (black) are within the 36N region. Top panels show boxplots of the fluorescence proxy of all clones with the -35 (green) and -10 box (blue) at specific positions. Boxplots show median, interquartile range and whiskers extend to maximally 1.5 times the interquartile range. All panels show data from 13544 clones with an exact length of 36 bp in the variable region.*

### 4.2.3.3 A 'universal' RNAP energy matrix

The 36N dataset is best described using two energy matrices of length 12 (including the -35 box) and 15 (-10 box), although the exact choice of the two lengths is of minor importance (Srdjan Sarikas). In the following, I refer to the two matrices together as '36N matrix'. We compared the energy values of the 36N matrix to those of the lacZ matrix used in chapter 3 (Figure 9). Since there is a free energy scale parameter in the inference of both matrices, we need not worry that energy values differ by a small scaling factor (line fit in scatter plot in Figure 25). We notice that the two energy matrices are overall similar, and the energy minimum values at the -35 and -10 boxes defining the strongest binding sequence match the known RNAP consensus. There are however considerable differences in the energy penalties of non-optimal letters at critical positions. In the 36N matrix, the -35 box has an overall lower importance, while the -10 box has a higher importance as compared to the lacZ matrix.

Interestingly, the ‘TGTG’ motif at the extended -10 is part of the energy minimum sequence in the 36N matrix, but not in the lacZ matrix.

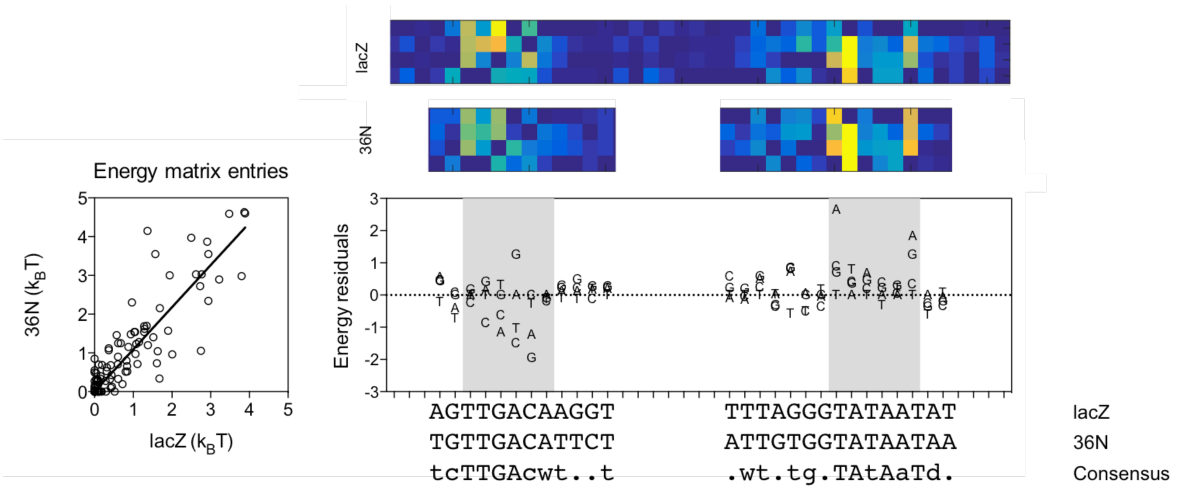


*Figure 25. Comparison of the 36N energy matrices with the lacZ energy matrix. Top: lacZ matrix (as in chapter 3) and 36N matrices for the two feet of RNAP. Values of the colorscale are in units of $k_BT$, but are not directly comparable (see text). Bottom left: Scatter plot of correlation between matrix entries. Line: linear fit. Bottom right: Energy residuals of the linear fit aligned with matrix positions. Residuals > 0 indicate larger energy penalties for non-optimal nucleotides (higher importance for promoter recognition) and residuals < 0 indicate smaller penalties (less importance). Sequences below the residual plot show energy minimum sequence of the different matrices and a previously published consensus sequence (Studnicka 1988).*

To quantify how much of the predictive power of our model is owed to the inferred energy matrix entries, we calculated binding probabilities with the full model, in which we substituted the 36N matrix entries with the lacZ energy matrix entries. This caused a large drop in the correlation coefficient $R^2$ between binding probabilities and fluorescence, from 0.69 (36N matrix) to 0.42 (lacZ matrix).

We next checked if differences between the matrices are meaningful beyond the 36N dataset and used the 36N matrix to predict expression of the single-nt mutants of the libraries from chapter 3. Figure 26 shows that the offset in calculated binding energies between the three libraries that we had observed with the lacZ matrix (and also the $P_L$ and $P_R$ matrices) became smaller, which came at the cost of lower local correlation for the pMS9_1 library (i.e. the lacZ RNAP binding site). This may indicate that the 36N matrix is indeed a more ‘universal’ energy matrix, although other explanations are possible (see discussion). The overall good fit between the binding energy obtained using the 36N matrix and the three single-nt libraries depended on allowing different spacer lengths.

On the level of predicted binding probabilities, our full model performs relatively well on the three single-nt libraries. The pMS9_1 library however is still incorrectly predicted to yield higher expression than the other two (Figure 26C).

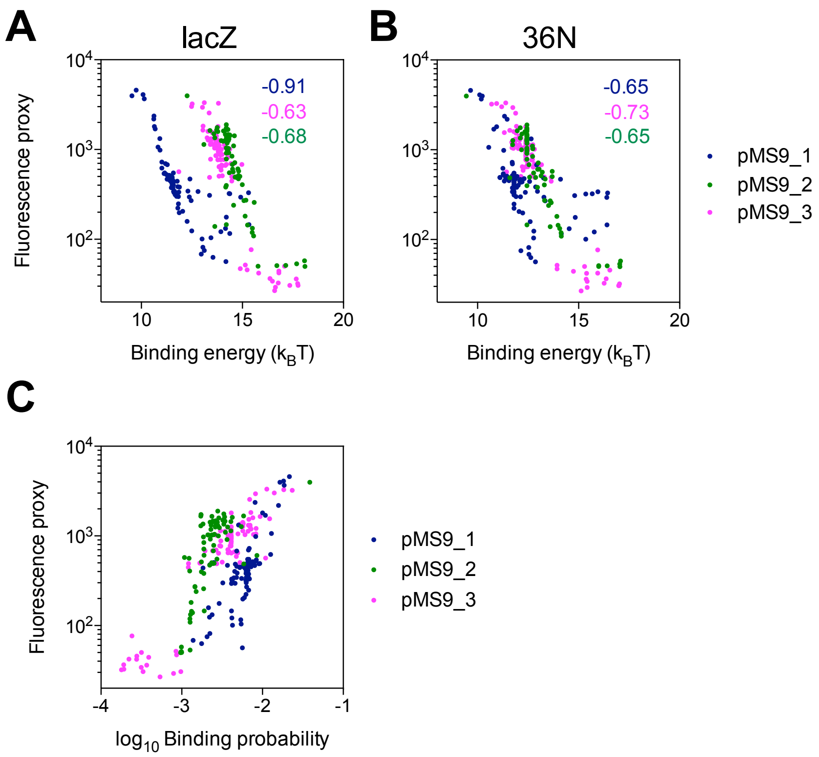


*Figure 26. Predictions of fluorescence of the three single-nt libraries from chapter 3. (A) Binding energies obtained using the lacZ energy matrix (same as Figure 15) (B) Binding energies obtained using the 36N matrix, a single frame, and a fixed optimal spacer (18 bp for pMS9_1 and pMS9_2 and 17 bp for pMS9_3). (C) Predicted binding probabilities of the full thermodynamic model.*

#### 4.2.3.4 **Effect of spacer flexibility**

Flexibility in the spacing between the -35 and -10 boxes of the RNAP binding site has been described early on (Stefano & Gralla 1982). In our model, energy values for five different spacings are inferred from the data. This allows to quantitatively compare the influence of spacer lengths on RNAP binding to the influence of nucleotide positions in the binding site (i.e. the energy matrix entries). The energy penalty of non-optimal spacer lengths increases with the distance from the canonical spacer length of 17 bp (Figure 27). Spacer penalties range from 1 to 5 $k_BT$ and are thus comparable to the energy contribution of a single non-optimal nucleotide at an important position of the energy matrix. Correspondingly, we find most primary promoters in the 36N dataset to have the canonical spacer length, as has been observed for natural promoters as well (Hawley & McClure 1983).

As already observed at the stage of model selection (Figure 23), allowing a flexible spacer length with energy penalties greatly increases the overall fit between model and data (Figure 28), which confirms that spacer flexibility is indeed important for promoter output in vivo. In particular, fluorescence of strong promoters is predicted with considerably better accuracy (Figure 28). The fixed spacer model yields a higher number of false negative predictions and a much higher number of false positive predictions compared to the full model with the flexible spacer model (quadrants in Figure 28A and B).

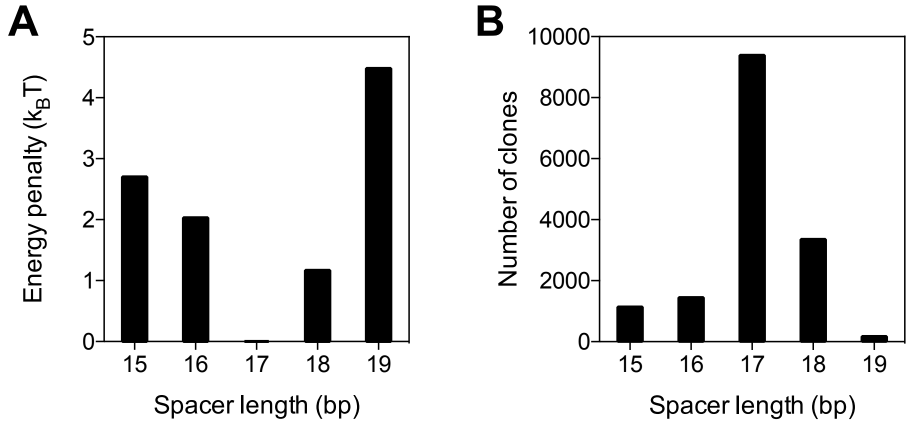


*Figure 27. Spacer flexibility in the dataset. (A) Inferred energy penalties of non-optimal spacer lengths. (B) Distribution of spacer lengths at the strongest binding position of RNAP for the full dataset.*

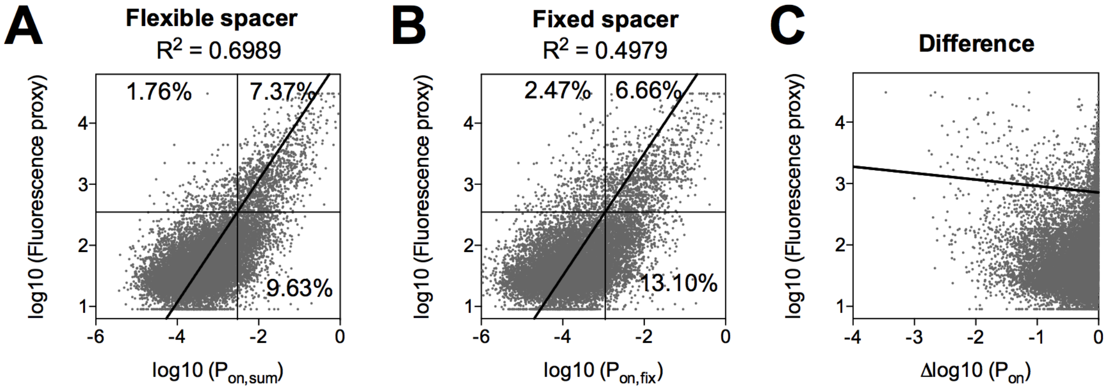


*Figure 28. Effect of allowing a flexible, energy-penalized spacer. Scatter plots of promoter occupancy predictions (x-axes) and fluorescence (y-axes). (A) Full model (B) Fixed spacer. For (B), only binding positions with spacer length 17 bp are considered as contributing towards $P_{on}$. Other parameters (energy matrix values, chemical potential and energy scale) are identical between (A) and (B). Solid lines show a weighted linear fit with a constrained slope of 1. Dashed lines indicate AF95 and respective $P_{on}$ thresholds. Percentages in quadrants are false negatives (top left), true positives (top right) and false positives (bottom right). (C) Difference between x-axis values of data shown in (A) and (B). Solid line is a weighted linear fit.*

### 4.2.3.5 Effect of additive RNAP binding at multiple sites

The textbook view of a bacterial promoter presents transcription initiation as the result of RNAP binding to a single position upstream of a gene (Snyder & Champess 2007). As is evident from Figure 21, transcription is frequently initiated at random DNA and thus we can expect that productive RNAP binding occurs relatively frequently throughout the genome. This is consistent with the observation of pervasive transcription of the bacterial genome (M. K. Thomason et al. 2014; James et al. 2017), although binding of RNAP outside of promoter regions appears to be avoided to some extent (Yona et al. 2018). Figure 29 shows that fluorescence of the 36N library is better captured by a model that allows contributions of multiple RNAP binding sites upstream of our reporter gene, although the improvement in terms of $R^2$ is modest.

Differences in predicted $P_{on}$ between the multiple sites model and the single site model are largest for weakly expressing clones and become smaller with increasing fluorescence (Figure 29C). This proportionality between the model difference between and fluorescence explains why overall correlation coefficients are almost identical.

RNAP binding at multiple sites could be pervasive in our random dataset – considering only the functional promoters, the median number of RNAP binding sites required to reach 90% of the total predicted promoter occupancy ($P_{on,90}$) is three (Figure 30A). Most promoters reach $P_{on,90}$ only with more than four binding sites. The single position model yields a higher number of false negative predictions and also a slightly higher number of false positive predictions compared to the full model with multiple binding sites (quadrants in Figure 29A and B). The total strength of a promoter is inversely correlated with the number of binding sites (Figure 30).

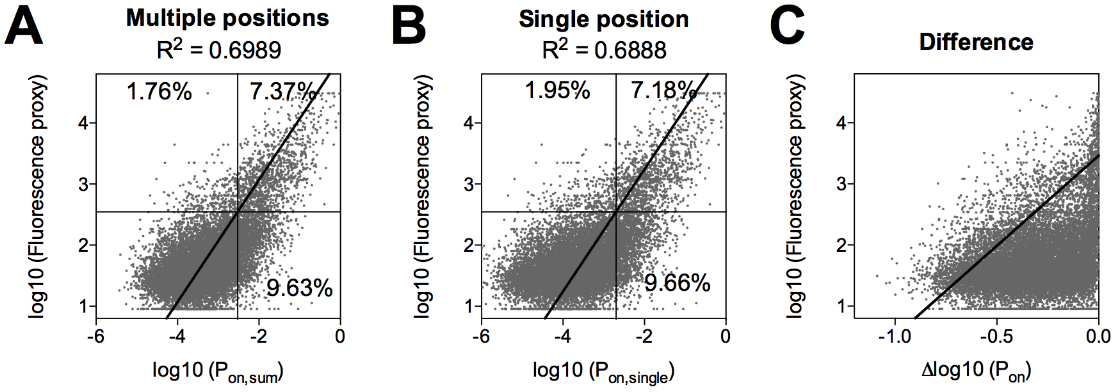


*Figure 29. Effect of allowing multiple RNAP binding positions. Scatter plots of promoter occupancy predictions (x-axes) and fluorescence (y-axes). (A) Full model, (B) single binding position. For (B), only the single energy minimum binding position and spacer length is considered as contributing towards $P_{on}$. Other parameters (energy matrix values, chemical potential, energy scale and spacer penalties) are identical between (A) and (B). Solid lines show a weighted linear fit with a constrained slope of 1. Dashed lines indicate AF95 and respective $P_{on}$ thresholds. Percentages in quadrants are false negatives (top left), true positives (top right) and false positives (bottom right). (C) Difference between x-axis values of data shown in (A) and (B). Solid line is a weighted linear fit.*

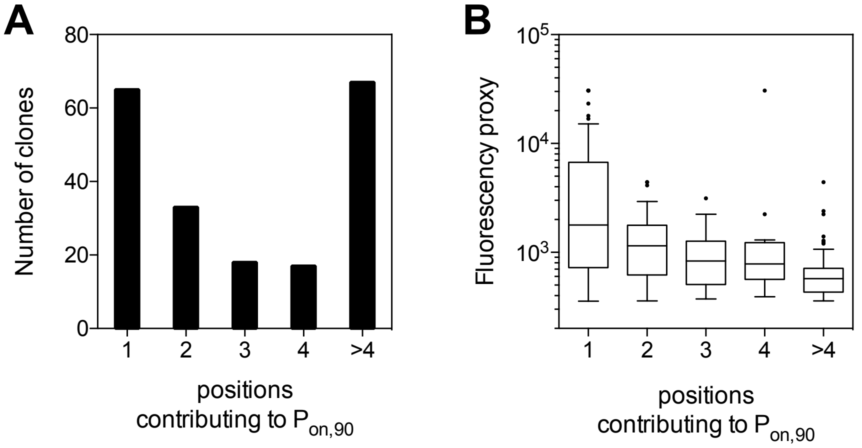


*Figure 30. Additive binding of RNAP at multiple positions is common. (A) Histogram showing the number of independent binding positions (positions of the -10 energy matrix) in a single sequence to reach at least 90% of the total binding probility of RNAP ($P_{on,90}$). (B) Weaker promoters are more likely to have multiple contributing binding sites. r=-0.4728, $P=10^{-12}$ (Spearman correlation). Both plots show data from a subset of clones with functional promoters (Fluorescence proxy > AF95) and with the four largest contributions to $P_{on}$ coming from different positions of the -10 energy matrix, n=200.*

### 4.2.3.6 Signs of 'alternative promoter types?

One possible explanation why predicting expression from sequence is hard, is that not all promoters may be described equally well by our models (see discussion in chapter 3). Due to its random nature, the 36N dataset is expected to sample promoters of all types (if such a distinction makes sense at all), and thus allows testing if specific sequences are better described by our model compared to other sequences. For this, we would however already have to have an idea of the features of 'alternative' types, or we could try to learn these features from scratch, i.e. using naïve neural networks. What is easier, is to test if sequence determinants of previously postulated promoter 'types' interact in unexpected ways, which would justify the notion of 'alternative types'. Here, we provide a small example of such an approach by testing if the 'extended -10 type' promoter exists as a recognizable class in our data.

The 'extended -10 type' of promoter was first described in *B. subtilis* (Moran et al. 1982) and later found to be conserved more weakly also in *E. coli* promoters (Mitchell et al. 2003), although it had been observed to be important earlier, for example in the context of λP$_{\mathrm{RE}}$ (Keilty & Rosenberg 1987). It has been proposed that 'extended -10' promoters, defined by a 'TG' at positions -15:-14 do not require a -35 box (Kumar et al. 1993), which could indicate that this promoter type is in fact functionally different from the bipartite -35/-10 promoter. Others have found that TG promoters have a lower requirement for homology in any of the other recognition elements, i.e. -35 or -10, and productive RNAP binding is merely the outcome of a sufficient number of contacts to the promoter, irrespective of their location at -35, -10, the extended -10, or distal elements (Mitchell et al. 2003), which is what has been referred to as the 'mix and match' model (Hook-Barnard & Hinton 2007).

We used our fluorescence data and model output to check if functional TG promoters are more likely to have a weak -35 element (Figure 31A). We found no statistical difference in the energy of the -35 box between TG and non-TG promoters. Also, we tested whether grouping functional promoters according to presence or absence of TG and binding energy of the -35 hexamer below or above the median revealed an interaction between these two elements (Figure 31B). While both a low-energy (i.e. strong) -35 box and a TG element increase fluorescence, there is no statistical support for an interaction, i.e. the positive effects of these motives on transcription appear to add up. We therefore reject the hypothesis that 'extended -10 promoters' constitute a functionally distinct promoter type. Rather, this analyses supports the additive 'mix and match' model of promoter function.

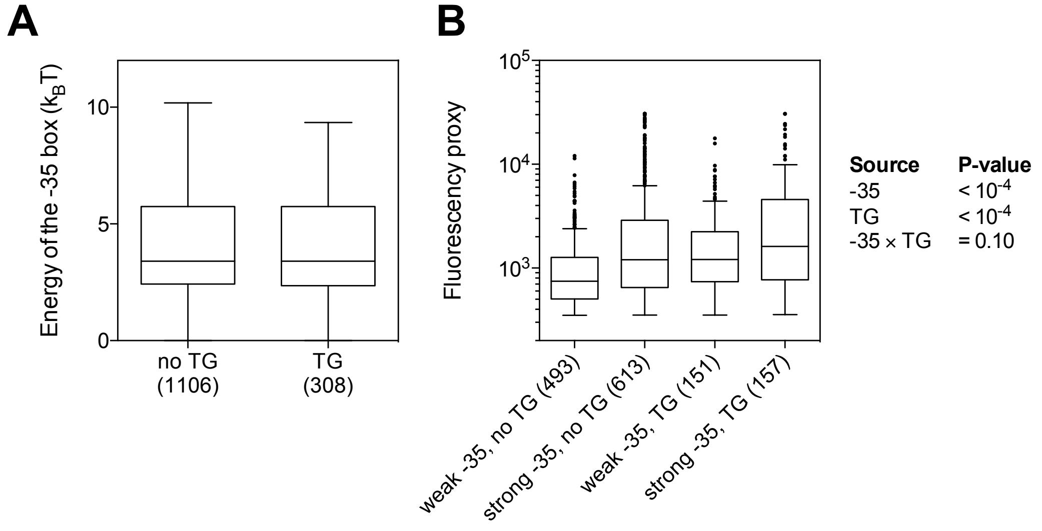


*Figure 31. No evidence for a distinct 'extended -10 type' promoter. (A) The energy of the -35 hexamer of functional promoters is the same for promoters regardless of the extended -10 'TG' motif. Numbers in parenthesis = n. (B) Presence of a strong -35 box or of a TG in the extended -10 have a positive effect on expression. The effect of the two motifs is independent. P-values are from two-way ANOVA (anovan, Matlab) and were calculated on log-transformed fluorescence values.*

This does not exclude that other functionally distinct classes of promoters may exist, in particular among very strong promoters. For example, our model fails to identify $\lambda P_L$ as a very strong promoter, with a predicted $P_{on}$ lower than the top 5.09% of the functional subset in our data, whereas its actual fluorescence is higher than we can resolve by sort-seq. For $\lambda P_L$, our model is no better than a simple homology score, which would predict it to be as strong as 4.95% of the functional promoters.

## *4.3 Discussion*

In this chapter, we have seen how promoter functionality is distributed in sequence space and we developed an extended thermodynamic model to predict promoter strength from sequence. We found that a relatively large fraction (9%) of random sequences contain functional promoters. This finding is consistent with a recently published paper that also reported 10% of random sequences to contain functional promoters (Yona et al. 2018). The authors of that study replaced the chromosomal upstream region of the *E. coli lac* operon (between the next upstream terminator and the TSS of the *lacZYA* transcript) by 40 distinct random sequences of the same length (103 bp). Of these 40 random sequences, four (10%) provided sufficient expression of the downstream *lac* operon to form colonies on plates with lactose as sole carbon source. Additional 23 of the 40 sequences (58%) evolved this capacity in evolution experiments by acquiring single point mutations. In our 36N dataset, which is substantially larger (>15000 sequences), we find a similar frequency of functional sequences. In addition, our data provides a high resolution of the shape of the distribution of promoter function in sequence space, which has a long tail that includes numerous promoters that approach the strength of RNAP binding sites of strong evolved promoters.

Our findings mean that promoters can evolve rather easily *de novo*, as weak promoters appear to be abundant in random sequence space. If one extrapolates the shape of the distribution of promoter function (Figure 21) into the region that is inaccessible to our measurements due to autofluorescence, one may be suspect that the fraction of functional promoters would even be substantially larger if we defined a lower threshold for functionality. This argues, at least in bacteria, against *de novo* promoter evolution being equivalent to a ‘bit-sum problem’ (as speculated by Gasper Tkacik inspired by the work of Murat Tugrul (Tuğrul et al. 2015)), in which selection is incapable of driving efficient adaptation due to vast areas of sequence space devoid of function. Rather, if RNAP binding is common in random sequence space, we expect selection for avoidance of RNAP binding motifs at non-promoter sites such as in coding regions, which is supported by bioinformatic evidence (Yona et al. 2018), or the existence of other mechanisms that alleviate the problem of abundant off-target binding and the resulting dilution of RNAP molecules in the cell. Experimental evidence supports that the DNA binding protein H-NS plays such a role in *E. coli* as it silences transcription from AT-rich regions in horizontally acquired genes (Lamberte et al. 2017).

Our results and those of Yona et al. on the feasibility of *de novo* promoter evolution are relevant to a broader question: Whether, in bacteria, functional genes can be ‘born’ from non-functional DNA, in addition to originating by modification of preexisting genes e.g. by duplication and divergence or rearrangements, as is widely accepted. The idea of continuous *de novo* evolution of genes is receiving increasing support in the case of eukaryotes (Schlötterer 2015; Wilson et al. 2017; Neme & Tautz 2016; McLysaght & Guerzoni 2015), but it has received little attention with respect to prokaryotes. *De novo* gene evolution offers an explanation for the existence of orphan genes, which have no recognizable homologs in species other than the one they were found in, arguing against their origin by duplication and divergence. Certainly, the specifics of prokaryote vs. eukaryote genome organization (Koonin & Y. I. Wolf 2010) makes *de novo* gene evolution, as we understand it today, less likely in prokaryotes. Since the organization of genes on prokaryote genomes is highly compact, prokaryote genomes contain little 'junk' DNA, which, free of major constraints, provides ample raw material for newly evolving gene functions in multicellular eukaryotes (Neme & Tautz 2016).

The origin of abundant prokaryote orphan genes (often termed ORFans) remains poorly understood (Yomtovian et al. 2010). More detailed hypotheses involving continuous *de novo* origination appear to be absent from the literature. The alternative hypothesis to continuous *de novo* evolution of genes, is the origin of all extant genes in a distant 'big bang', which has been loosely dated to the Archaean eon (David & Alm 2010) or even to the time before the most recent universal common ancestor (Harish et al. 2013). In this view, ORFans require other explanations such as rapid divergence that makes homologs unrecognizable.

A basic chicken-and-egg question in the *de novo* evolution of genes is whether the coding sequence evolves first or *de novo* gene expression, i.e. promoters, evolve first (Schlötterer 2015). Our results imply that, if continuous *de novo* gene evolution *does* exist in bacteria, the latter step, i.e. the evolution of functional promoters, should not provide a serious constraint in the process.

One reason for the large fraction of functional promoters can be found in the inherent flexibility of the promoter recognition machinery instantiated in RNAP. Both the flexibility of spacer length and additive binding at multiple sites, as suggested by our model, contribute to the large fraction of functional promoters in random sequence space. Spacer flexibility increases primarily the number of strong promoters; additive binding at multiple sites increases primarily the number of weak promoters.

Taken together, our two model extensions of energy-penalized spacer flexibility and additive binding of RNAP at multiple sites, combined with the 36N energy matrix, greatly improve predictions of expression using a thermodynamic model, exceeding the accuracy of a prediction based on a simple homology-score by far (Figure 32).

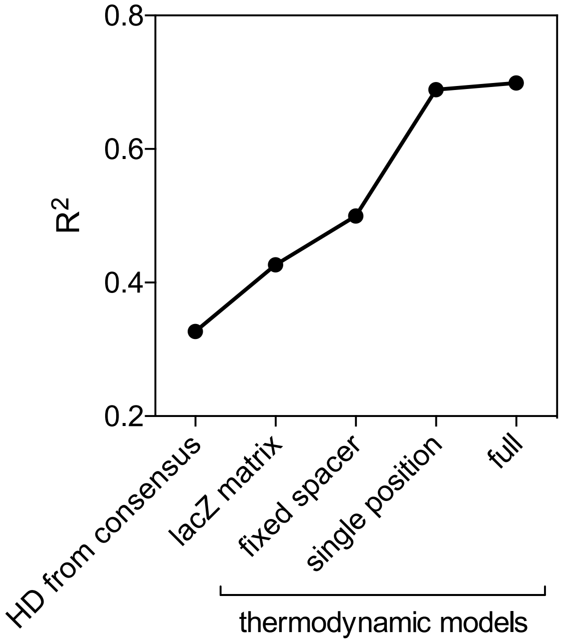


*Figure 32. Correlation coefficients between observed expression and different models of promoter strength. 'HD from consensus' is the square of the weighted correlation coefficient between expression and the hamming distance of both core hexamers (-35 and -10) to the consensus. 'full' refers to the best performing model with a flexible spacer, summing over multiple positions, and using the 36N matrix. The other three thermodynamic models are identical except for the specified modification.*

Spacer flexibility is a well known fact, but we are not aware of previously reported explicit energy penalties of suboptimal spacer lengths, and we find them to be in the same range as energy penalties of a single important position in the recognition boxes.

Additive contributions of multiple RNAP binding sites in natural promoters are less clearly supported by the literature. Although there is indirect evidence for such a case at the *dgoR* promoter as recently published (Belliveau et al. 2018), there are other reports in which multiple RNAP binding sites either at the same strand (M. L. Peterson & Reznikoff 1985) or at the opposite strand (Bendtsen et al. 2011) interfere with each other rather than add up. Also, the existence of regions with high densities of promoter like-signals on bacterial chromosomes is poorly understood (Huerta & Collado-Vides 2003). Such 'promoter islands' are typically associated with horizontally acquired genomic islands and do not initiate long transcripts (Panyukov & Ozoline 2013). In our random dataset, additive contributions of multiple RNAP binding sites, particularly in weak promoters, are mainly supported by better fitting statistics, but also by general considerations: If there are multiple sites of comparable affinity to RNAP, but only one of them contributes to expression (as in the single site model), RNAP would have to 'know' how to choose, and clearly, molecules do not 'know'.

An alternative explanation for the better fit of the multiple sites model could be that our incomplete understanding of the RNAP-promoter interaction fails to capture something important in how different potential binding sites contribute to expression. Even if it were true, that expression is driven by single RNAP binding sites, the multiple sites model may give better results. For example, a minimum-energy binding site that is further away from the start of the coding sequence may contribute little to expression, because longer untranslated transcripts could be targeted by the termination factor Rho (Ciampi 2006). If this is the case, the single site model will incorrectly assume the more distant site to determine expression levels, at the expense of missing another site further downstream. The multiple sites model would also incorrectly assume a contribution of the more distant site, but it would still incorporate sequence information from the downstream site. Therefore, the multiple sites model should be more tolerant to wrong assumptions. To clarify the question whether and how multiple possible RNAP binding sites contribute to expression, it remains to be tested experimentally how engineered mutations in predicted multiple RNAP binding sites affect fluorescence.

Possibly, contributions of multiple sites are typical in *de novo* evolution of promoters, whereas multiple sites tend to become repressed or differentially regulated (Huerta & Collado-Vides 2003) as promoters 'mature' evolutionarily, or as they switch hosts by way of horizontal transfer.

Another substantial improvement of our predictions (Figure 32) is owed to a new energy matrix that is inferred from data that covers a wide sequence space and range of expression. This is different from other energy matrices that were derived by mutagenesis in the local sequence space surrounding a functional sequence (Kinney et al. 2010; Kreamer et al. 2015). The impact of the specific reference sequence around which an energy matrix is inferred that became evident in chapter 3 was recently observed also for energy matrices of transcription factor binding (Barnes et al. 2018). Using a matrix inferred from an unbiased sample of sequence space is therefore crucial for predicting expression from any sequence. In random sequence space, and therefore possibly in natural evolution of promoters *de novo*, the -35 box appears less important than generally assumed, and the -10 element with its upstream extension appears to be more important. The closing of the 'energy gap' between the three libraries from chapter 3, by simple substitution of the matrix values from lacZ to 36N, supports that the 36N matrix is indeed a more ‘universal’ energy matrix. Alternatively, the 36N matrix could be overfitted to the sequence context on the pMS9 plasmids (and in particular the constant flanking region), which is identical for all four plasmid libraries in this thesis. The wider applicability of the 36N matrix therefore needs to be tested on unrelated datasets.

Without doubt, caution needs to be taken before applying our model to predict the strength of expression of natural chromosomally encoded promoters in vivo, where a multitude of additional factors enter the equation. Factors such as chromosome structure and physiological state of the cell, including the expression state of transcription factors etc. are constant in our experiment. Only in this way can we carve out the specific contribution of the RNAP binding site sequence to transcription. Still, applying our model to predict the strength of promoters, e.g. in intergenic regions on the *E. coli* chromosome, and comparing results to existing bioinformatics approaches should yield interesting result.

Our model builds on the assumption of expression being proportional to RNAP occupancy and additivity of energy contributions of individual nucleotides. Although we know these assumptions are wrong, we find an overall good performance of the model, indicating that ‘alternative promoter types’, if they exist, are uncommon or inconsequential, at least in our dataset. Also, we do not find evidence that the model performs worse on a particular previously proposed ‘promoter type’, the extended -10 promoter, or that this promoter type is even a meaningful class. Rather, having the features of the extended -10 element in the 36N energy matrix (Figure 25) is sufficient to improve model fits for sequences that we speculated to belong to the extended -10 promoter class in chapter 3 (Figure 26). Still, there is a possibility that there are indeed alternative promoter types that are not captured well by our model because they validate the above assumptions. So far, there is little quantitative work that addresses the sequence dependence of later steps of the transcription initiation process (E. Heyduk & T. Heyduk 2014; Djordjevic & Bundschuh 2008). As I am not aware of work that uses a multi-step model for the prediction of promoter strength from sequence, it remains to be seen how much is to be gained by a more complete model that is refined it this particular way.

One hypothesis is that alternative promoter architectures are to be found primarily at repressible promoters that, in the absence of repressor, are very strong. Promoters that bind RNAP strongly are hard to regulate by repressors (Hook-Barnard & Hinton 2007; Lanzer &

Bujard 2007). Therefore, ‘consensus-type’ promoters, whose strength is largely determined by RNAP binding, face a tradeoff between promoter strength and regulatability. This tradeoff can be circumvented in different ways to achieve expression that is both strong and responsive to regulation. The *E. coli rrn* operons, from which ribosomal RNA is transcribed, are a particularly informative example. Their transcripts constitute up to 70% of the total RNA in the cell, yet their promoters are not strong in the sense of being close to the consensus sequence and binding RNAP tightly (Haugen et al. 2006). Rather, the ribosomal RNA genes exist in multiple copies on the chromosome and their promoters are regulated by the activator Fis and additional specialized mechanisms (Haugen et al. 2006). The highly abundant protein EF-Tu is also encoded by a gene with two chromosomal copies (van der Meide et al. 1982). These examples cannot substitute a more systematic analysis, but together with the absence of the consensus sequence in the genome of *E. coli* they are consistent with the hypothesis that promoters that bind RNAP tightly are avoided in the chromosome due to the strength/regulatability tradeoff. Genes required at very high expression levels circumvent the tradeoff by increased gene copy number or by regulation via activation.

In phage genomes with strongly constrained genome sizes, increasing copy number is not an option. Thus, the $P_L$ promoter of phage λ, which is tightly repressed by the lambda repressor cI (Ptashne 2004), represents yet another, an ‘alternative’ resolution of the tradeoff. $P_L$ is strong despite its poor homology to the consensus sequence and relatively weak binding of RNAP (Knaus & Bujard 1988). This allows for tight repression. What remains to be understood is what, if not tight binding of RNAP, makes $P_L$ a strong promoter. Bujard and coworkers concluded from their detailed *in vitro* and *in vivo* studies that the sequence determinants of the strength of $P_L$ must be downstream (in the twofold sense: spatially, i.e. downstream of the -10 box, and temporally, i.e. after initial binding) (Knaus & Bujard 1990). If they are located downstream of the -10 box, this may explain why our 36N dataset is unsuitable for identifying alternative promoter signatures, as these downstream regions are likely too large to be covered by the variable part of the 36N library.

We conclude that we should not expect alternative promoters to be abundant in bacterial genomes, as there are other ways to avoid the strength/regulatability tradeoff. Instead, phage genomes may be a better place to look for them, and repressible strong promoters are prime candidates.

## *4.4 Materials and Methods*

### 4.4.1 Plasmid cloning

We used plasmid pMS9_4 as a starting point for plasmid and library construction. The only difference between pMS9_4 and the other pMS9 plasmids from chapter 3, is a different 36 nt sequence upstream of the *gfp* reporter. pMS9_4 was initially part of the project described in chapter 3, but was later abandoned due to technical issues. The reference plasmid pMS9_PL was built from pMS9_4 using a Q5 site directed mutagenesis kit and primers PL_f and PL_r. The plasmid changes were verified by Sanger sequencing.

### 4.4.2 Creation of the 36N library

The 36N plasmid DNA library was generated using a Q5 site directed mutagenesis kit (NEB) in a 20 µL reaction. For amplification, we used plasmid pMS9_4 as a template and two pools of primers with a constant 3' end and an 18N random 5' end (18N_f and 18N_r). We transformed 5 µL of the KLD reaction mix into 50 µL of chemically competent NEB5α cells. After 1 h outgrowth in 1 mL LB, we plated 100 µL of the culture on 10 LB kan agar plates and 5 g/L sterile charcoal. Based on plating dilutions of the same culture, the total number of colonies plated in this way is ~ $2 \times 10^4$ cells. After overnight incubation, colonies from the 10 plates were scraped off and resuspended in LB kan. Suspensions were vortexed vigorously and diluted to an $OD_{600}$ of ~1. Aliquots of the 36N library were then frozen at -80 °C (100 µL cell suspension with 40 µL glycerol (50%).

### 4.4.3 FACS-sorting

Prior to sorting, cells were grown in freshly filtered (0.22µm) M9 minimal medium with 0.2% CAS, 0.2% Glucose and 50 µg/mL kanamycin. Frozen aliquots of the 36N library and the reference plasmid were diluted 1:10 and grown overnight. Prior to sorting, overnight cultures were diluted again 1:100 and grown for 3 h to reach exponential phase.

FACS-sorting was performed on an FACS Aria III flow cytometer (BD Biosciences, San Jose, CA) with a 70 µm nozzle for droplet formation. A 488 nm laser was used to detect forward scatter (FSC) and side scatter (SSC) with a 488/10 band-pass filter. The same laser was used for excitation of GFP (FITC channel, emission filters 502LP, 530/30). We chose the FITC channel voltage such that the median fluorescence of a plasmid-free auto-fluorescence control sample (AF) is between 0 and 100 on the FITC axes. The flow rate was set to 1.0 and samples were diluted to obtain a cell count of approximately 5000 events/second. Cells for sorting were manually gated on the densest population in an FSC/SSC scatter plot, which comprised 95.5% of all events exceeding a threshold of 1000 on the SSC axis. Twelve sorting gates were set on the FITC axes as follows: The upper boundary of the lowest gate (B1) corresponded to the median of an autofluorescence control sample (plasmid-free cells). The lower boundary of the highest gate (B12) was set to $2 \times 10^4$. Distances between the remaining intermediate nine gate boundaries defining B2 to B11 were chosen with a constant multiplication factor of 1.85, i.e. gates were of equal size on the log-scale FITC histogram (Table 7). Prior to sorting, we recorded $10^6$ events. The number of cells to be sorted into each of the twelve bins B1-B12 then corresponded to the number of cells previously recorded in each of the bins and can be found in Table 7. Cells were sorted into 24-well plates with 500 µL sorting medium / well. After completion of sorting, 1000 cells of the culture with the reference plasmid pMS9_PL were sorted into each of the wells holding cells from one bin. The recipient plate was cooled to 4 °C to halt growth while sorting to other wells was still going on.

*Table 7. Number of cells sorted (top) and bin boundaries (bottom).*

| | *B1* | *B2* | *B3* | *B4* | *B5* | *B6* | *B7* | *B8* | *B9* | *B10* | *B11* | *B12* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *pMS9_36N (cells)* | 306183 | 170570 | 233624 | 138623 | 57313 | 33512 | 21488 | 14381 | 8857 | 4644 | 2716 | 2197 |
| *pMS9_PL (cells)* | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 |
| *Upper boundary (FACS units)* | 42 | 78 | 144 | 267 | 495 | 917 | 1698 | 3146 | 5827 | 10796 | 20000 | - |

After completion of sorting, 1000 cells of the culture with the control plasmid pMS9_L were added into each of twelve wells. Sorted cells were spun down in a cooled centrifuge and resuspended in 1 mL medium. We then plated a dilution from each well on LB Kan (around 100 cells / plate), quantified colony fluorescence using the macroscope to estimate the frequency of mis-sorting (mean outlier frequency over 12 bins was 4.2%, standard deviation 0.5%, outlier classification using ROUT with Q=1%). Finally, the cells from each bin were grown overnight.

### 4.4.4 Plasmid library isolation and barcoding PCR

We isolated plasmid from the twelve culture pools and quantified DNA concentration using a Nanodrop spectrophotometer. Given the number of cells sorted and the plasmid pool concentrations, every sorted cell is expected to contribute 600 plasmid molecules or more to

1 ng of plasmid pool DNA, which is the amount of template we used in the subsequent PCR amplification step.

For barcoding PCR products containing the mutagenized region, we created primers mutseq_f1-12 and mutseq_r1-12. They contain a 3’ constant region, a bin-specific barcode of 5 nt and a constant 5’ tail of 5 nt.

PCRs were performed using Q5 high fidelity polymerase and 1 ng of the plasmid pools as template in a 50 µL reaction. We first performed five cycles with an annealing temperature calculated for the constant 3’ part of the primers, followed by 25 cycles using an annealing temperature matched to each of the full length primer pairs.

PCR products were column-purified (Zymo research, Irvine, CA) and eluted in 30 µL, of which 2 µL were run on an agarose gel for relative product quantification based on band fluorescence. PCR products were finally pooled to reach approximately equimolar concentrations of the twelve reaction products.

### 4.4.5 Illumina sequencing

We sent ~1 µg of pooled PCR product to sequencing by GATC biotech (Konstanz, Germany) on an Illumina sequencer (125 bp paired end).

### 4.4.6 Characterizing a reference set of clones

We picked 8 colonies plated from each of the 12 bins and quantified $OD_{600}$-normalized fluorescence of single replicate exponential cultures using an H1 platereader (Biotek, Vinooski, Vermont) with a GFP filter. Platereader fluorescence was found to correlate linearly with the median of the FACS signal of the bin clones were derived from. The variable region of the clones was identified using Sanger sequencing. After filtering out clones which could not be sequenced or which had rearrangements, we obtained a dataset of 78 unique clones.

### 4.4.7 Debiasing read distributions and calculating a fluorescence proxy

The sequencing raw data was processed by Srdjan Sarikas. For our analysis, we only used reads with matching barcodes in the forward and reverse primers and a variable region of length 34-38 bp. For subsequent analysis, we only used mutant sequences with ≥10 reads. This threshold excludes presumable artifact sequences resulting from molecular sequencing noise that can be recognized by being almost identical to other unique sequences with a much higher read count. These and additional filters yield 15492 unique sequences, each with a distribution over expression bins (‘raw read distributions’). Raw read distributions are then pruned to reduce the impact of sequencing noise (the detailed method of this pruning were developed by Srdjan Sarikas) and debiased by dividing by the number of reads of the reference sequence (Table 8). Finally, we took the geometric mean of debiased distributions as fluorescence proxy. The previous steps were validated by comparing fluorescence proxies against platereader fluorescence data of 78 clones.

*Table 8. Distribution of control sequence reads across bins*

| *Bin* | *B1* | *B2* | *B3* | *B4* | *B5* | *B6* | *B7* | *B8* | *B9* | *B10* | *B11* | *B12* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *# of control sequence reads* | 2152 | 3188 | 3048 | 7054 | 22973 | 51575 | 33671 | 73805 | 103077 | 169988 | 293402 | 381119 |

### 4.4.8 Model inference

Thermodynamic models were inferred by Srdjan Sarikas, with the support of Gašper Tkačik and will be published in detail elsewhere. Here, I only outline important points and the differences to the work in chapter 3 of this thesis. Instead of using only binding energy, we work with the full thermodynamic model based on equation (1). This means that chemical potential and, in case of a flexible spacer, spacer penalties are inferred from the data. Also, instead of using a published energy matrix, matrix entries of two binding regions of the RNAP ‘feet’ are inferred from the data. For initialization, the energy matrix values of the *lac* promoter are used. Instead of using fluorescence proxies directly, datapoints are binned by their fluorescence proxy, into the 12 bins previously used in FACS sorting. For inference, we use logistic regression between log(binding probability, $P_{on}$) and log(fluorescence). The maximized value during training is a log likelihood estimator. Since the abundance of clones in each bin decreases with higher fluorescence, observations are weighted by the inverse of unique sequence counts in each bin. After obtaining fitted parameters, the energy scale and the chemical potential, but not matrix entries and spacer penalties, are refitted to achieve a linear fit with slope 1 between log($P_{on}$) and log(fluorescence), as this is the expected relationship between the two quantities. For refitting the energy scale and the chemical potential, data from the lowest and highest bins are excluded. Scatterplots in Figure 23 and reported $R^2$ refer to refitted models and weighted data.

### 4.4.9 List of primers

Cloning primers

| | |
|---|---|
| PL_f | CTGGCGGTGATACTGAGCTGTGCATACAGATTGAGTAATGG |
| PL_r | CTGGCGGTGATACTGAGCTGTGCATACAGATTGAGTAATGG |
| 18N_f | NNNNNNNNNNNNNNNNNNTGTGCATACAGATTGAGTAATG |
| 18N_r | NNNNNNNNNNNNNNNNNNAATAGCATTGAGATTTGAAGC |

Barcoding primers

For primers mutseq_f/r1-6 see chapter 3.

| | |
|---|---|
| mutseq_f7 | AAGCTGACACTCGTCTTCACCTCGAGCAC |
| mutseq_f8 | AAGCTTCGTATCGTCTTCACCTCGAGCAC |
| mutseq_f9 | AAGCTCCAATTCGTCTTCACCTCGAGCAC |
| mutseq_f10 | AAGCTTGGTGTCGTCTTCACCTCGAGCAC |
| mutseq_f11 | AAGCTGCTATTCGTCTTCACCTCGAGCAC |
| mutseq_f12 | AAGCTTGACCTCGTCTTCACCTCGAGCAC |
| mutseq_R7 | CGTACGACACTTCTCCTTTACTCATATGTATATCT |
| mutseq_R8 | CGTACTCGTATTCTCCTTTACTCATATGTATATCT |
| mutseq_R9 | CGTACCCAATTTCTCCTTTACTCATATGTATATCT |
| mutseq_R10 | CGTACTGGTGTTCTCCTTTACTCATATGTATATCT |
| mutseq_R11 | CGTACGCTATTTCTCCTTTACTCATATGTATATCT |
| mutseq_R12 | CGTACTGACCTTCTCCTTTACTCATATGTATATCT |

# 5 Conclusions

In this thesis, we have seen how the evolution of gene expression in bacteria depends on sequence context at two levels: at the level of genes in their chromosomal context, and at the level of nucleotides in the context of a promoter sequence. At both levels, we have seen how the influence of context can be dramatic, but also how it can be dealt with: by identifying the important determinants in the context (chapter 2 / chromosomal neighborhood) or by using models that are inferred from highly diverse contexts and thus are less prone to being overfitted to any particular context (chapter 4 / promoter context).

Overall, this gives us an optimistic outlook that context dependency in the evolution of gene expression is not an unpredictable beast, but that it can be tamed. Of course, the specific way in which we have done this here needs verification in a wider set of contexts. For chromosome neighborhood effects, this would entail testing whether our expectations of adaptive potential hold for additional chromosomal loci and, beyond *E. coli*, for different bacterial species. For context effects on promoter strength, this would entail testing our model in particular for longer variable regions, known strong promoters, other flanking regions, and chromosomal integrations instead of plasmid systems. Some of these tests are currently under way.

If our results hold more generally, they should be useful for a number of important questions:

- Can we predict the likelihood of rapid adaptation such as the evolution of drug resistance from genomic data?
- How much is horizontal gene transfer constrained by the need to evolve proper expression patterns?
- Is there *de novo* evolution of bacterial genes?

# References

Adler, M. et al., 2014. High Fitness Costs and Instability of Gene Duplications Reduce Rates of Evolution of New Genes by Duplication-Divergence Mechanisms. *Molecular Biology and Evolution*.

Akhtar, W. et al., 2013. Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell*, 154(4), pp.914–927.

Anderson, P. & Roth, J., 1981. Spontaneous tandem genetic duplications in Salmonella typhimurium arise by unequal recombination between rRNA (rrn) cistrons. *Proceedings of the National Academy of Sciences*, 78(5), pp.3113–3117.

Andersson, D.I., 1998. Evidence That Gene Amplification Underlies Adaptive Mutability of the Bacterial lac Operon. *Science*, 282(5391), pp.1133–1135.

Andersson, D.I. & Hughes, D., 2009. Gene amplification and adaptive evolution in bacteria. *Annual Review of Genetics*, 43, pp.167–195.

ar-Rushdi, A. et al., 1983. Differential expression of the translocated and the untranslocated c-myc oncogene in Burkitt lymphoma. *Science*, 222(4622), pp.390–393.

Baba, T. et al., 2006. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, 2, p.2006.0008.

Barnes, S.L. et al., 2018. Mapping DNA sequence to transcription factor binding energy in vivo. *bioRxiv*.

Barrick, J.E. et al., 2014. Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq. *BMC Genomics*, 15, p.1039.

Belliveau, N.M. et al., 2018. Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. *Proceedings of the National Academy of Sciences*, 115(21), pp.E4796–E4805.

Bendtsen, K.M. et al., 2011. Direct and indirect effects in the regulation of overlapping promoters. *Nucleic Acids Research*, 39(16), pp.6879–6885.

Bintu, L. et al., 2005. Transcriptional regulation by the numbers: models. *Current Opinion in Genetics & Development*, 15(2), pp.116–124.

Blank, D. et al., 2014. The predictability of molecular evolution during functional innovation. *Proceedings of the National Academy of Sciences*, 111(8), pp.3044–3049.

Blount, Z.D. et al., 2012. Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature*, 489(7417), pp.513–518.

Boyd, E.F. & Hartl, D.L., 1997. Nonrandom location of IS1 elements in the genomes of natural isolates of Escherichia coli. *Molecular Biology and Evolution*, 14(7), pp.725–732.

Brenner, S., 2009. Sequences and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537), pp.207–212.

Browning, D.F. & Busby, S.J.W., 2004. The regulation of bacterial transcription initiation. *Nature Reviews Microbiology*, 2(1), pp.57–65.

Bryant, J.A. et al., 2014. Chromosome position effects on gene expression in Escherichia coli K-12. *Nucleic Acids Research*, 42(18), pp.11383–11392.

Cardinale, C.J. et al., 2008. Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in E. coli. *Science*, 320(5878), pp.935–938.

Carlson, S.M., Cunningham, C.J. & Westley, P.A.H., 2014. Evolutionary rescue in a changing world. *Trends in Ecology & Evolution*, 29(9), pp.521–530.

Carroll, S.B., 2000. Endless forms: the evolution of gene regulation and morphological diversity. *Cell*, 101(6), pp.577–580.

Ciampi, M.S., 2006. Rho-dependent terminators and transcription termination. *Microbiology*, 152(9), pp.2515–2528.

Coderre, J.A. & Beverley, S.M., 1983. Overproduction of a bifunctional thymidylate synthetase-dihydrofolate reductase and DNA amplification in methotrexate-resistant Leishmania tropica. *Proceedings of the National Academy of Sciences*, 80, pp.2132–2136.

Cole, S.P. et al., 1992. Overexpression of a transporter gene in a multidrug-resistant human lung cancer cell line. *Science*, 258(5088), pp.1650–1654.

Conrad, T.M., Lewis, N.E. & Palsson, B.Ø., 2011. Microbial laboratory evolution in the era of genome-scale science. *Molecular Systems Biology*, 7.

Conway, T. et al., 2014. Unprecedented High-Resolution View of Bacterial Operon Architecture Revealed by RNA Sequencing. *mBio*, 5(4), pp.e01442–14.

Core Team, R.D., 2012. *R: A language and environment for statistical computing*, Vienna, Austria.

Craig, N.L., 1997. Target site selection in transposition. *Annual Review of Biochemistry*, 66, pp.437–474.

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), pp.6640–6645.

David, L.A. & Alm, E.J., 2010. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*, 469(7328), pp.93–96.

de Boer, C. et al., 2018. Deciphering cis-regulatory logic with 100 million synthetic promoters. *bioRxiv*.

Deuschle, U. et al., 1986. Promoters of Escherichia coli: a hierarchy of in vivo strength indicates alternate structures. *The EMBO Journal*.

Devonshire, A.L. & Field, L.M., 1991. Gene amplification and insecticide resistance. *Annual Review of Entomology*, 36, pp.1–23.

Dickson, R.C. et al., 1975. Genetic regulation: the Lac control region. *Science*, 187(4171), pp.27–35.

Djordjevic, M. & Bundschuh, R., 2008. Formation of the Open Complex by Bacterial RNA Polymerase—A Quantitative Model. *Biophysical journal*, 94(11), pp.4233–4248.

Dobrindt, U. et al., 2004. Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology*, 2(5), pp.414–424.

Ehrenberg, M. et al., 2003. Systems biology is taking off. *Genome Research*, 13(11), pp.2377–2380.

Elliott, K.T., Cuff, L.E. & Neidle, E.L., 2013. Copy number change: evolving views on gene amplification. *Future Microbiology*, 8(7), pp.887–899.

Ellison, C.E. & Bachtrog, D., 2013. Dosage compensation via transposable element mediated rewiring of a regulatory network. *Science*, 342(6160), pp.846–850.

Fay, M.P. & Shaw, P.A., 2010. Exact and Asymptotic Weighted Logrank Tests for Interval Censored Data: The interval R package. *Journal of Statistical Software*, 36(2), p.i02.

Feklistov, A. et al., 2006. A Basal Promoter Element Recognized by Free RNA Polymerase σ Subunit Determines Promoter Recognition by RNA Polymerase Holoenzyme. *Molecular cell*, 23(1), pp.97–107.

Foster, P.L. et al., 2013. On the mutational topology of the bacterial genome. *Genes, Genomes, Genetics*, 3(3), pp.399–407.

Freddolino, P.L., Goodarzi, H. & Tavazoie, S., 2012. Fitness Landscape Transformation through a Single Amino Acid Change in the Rho Terminator J. Zhang, ed. *PLoS Genetics*, 8(5), p.e1002744.

Gajduskova, P. et al., 2007. Genome position and gene amplification. *Genome Biology*, 8(6), p.R120.

Gama-Castro, S. et al., 2016. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, 44(D1), pp.D133–D143.

Garcia, H.G. et al., 2012. Operator Sequence Alters Gene Expression Independently of Transcription Factor Occupancy in Bacteria. *CELREP*, 2(1), pp.150–161.

Gottesman, M.E. & Weisberg, R.A., 2004. Little Lambda, Who Made Thee? *Microbiology and Molecular Biology Reviews*, 68(4), pp.796–813.

Graña, D., Gardella, T. & Susskind, M.M., 1988. The effects of mutations in the ant promoter of phage P22 depend on context. *Genetics*, 120(2), pp.319–327.

Green, L. et al., 1984. Distribution of DNA insertion element IS5 in natural isolates of Escherichia coli. *Proceedings of the National Academy of Sciences*, 81(14), pp.4500–4504.

Griswold, K.E. et al., 2003. Effects of codon usage versus putative 5'-mRNA structure on the expression of Fusarium solani cutinase in the Escherichia coli cytoplasm. *Protein expression and purification*, 27(1), pp.134–142.

Grosso, A.R. et al., 2015. Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. *eLife*, 4.

Haldimann, A. & Wanner, B.L., 2001. Conditional-Replication, Integration, Excision, and Retrieval Plasmid-Host Systems for Gene Structure-Function Studies of Bacteria. *Journal of Bacteriology*, 183, pp.6384–6393.

Harish, A., Tunlid, A. & Kurland, C.G., 2013. Rooted phylogeny of the three superkingdoms. *Biochimie*, 95(8), pp.1593–1604.

Haugen, S.P. et al., 2006. rRNA Promoter Regulation by Nonoptimal Binding of σ Region 1.2: An Additional Recognition Element for RNA Polymerase. *Cell*, 125(6), pp.1069–1082.

Haugen, S.P., Ross, W. & Gourse, R.L., 2008. Advances in bacterial promoter recognition and its control by factors that do not bind DNA. *Nature Reviews Microbiology*, 6(7), pp.507–519.

Hawley, D.K. & McClure, W.R., 1983. Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Research*, 11(8), pp.2237–2255.

Heyduk, E. & Heyduk, T., 2014. Next Generation Sequencing-Based Parallel Analysis of Melting Kinetics of 4096 Variants of a Bacterial Promoter. *Biochemistry*, 53(2), pp.282–292.

Hook-Barnard, I.G. & Hinton, D.M., 2007. Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. *Gene regulation and systems biology*, 1, pp.275–293.

Hudson, R.E. et al., 2002. Effect of chromosome location on bacterial mutation rates. *Molecular Biology and Evolution*, 19(1), pp.85–92.

Huerta, A.M. & Collado-Vides, J., 2003. Sigma70 Promoters in Escherichia coli: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals. *Journal of Molecular Biology*, 333(2), pp.261–278.

Ishihama, A., 2000. Functional modulation of Escherichia coli RNA polymerase. *Annual Review of Microbiology*, 54, pp.499–518.

Jacob, F. & Monod, J., 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3), pp.318–356.

James, K., Cockell, S.J. & Zenkin, N., 2017. Deep sequencing approaches for the analysis of prokaryotic transcriptional boundaries and dynamics. *Methods*, 120, pp.76–84.

Jesse, F., 2017. *The lac operon in the wild.*

Kawano, M., 2005. Detection of low-level promoter activity within open reading frame sequences of Escherichia coli. *Nucleic Acids Research*, 33(19), pp.6268–6276.

Keilty, S. & Rosenberg, M., 1987. Constitutive function of a positively regulated promoter reveals new sequences essential for activity. *Journal of Biological Chemistry*, 262(13), pp.6389–6395.

Kingsford, C.L., Ayanbule, K. & Salzberg, S.L., 2007. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biology*, 8(2), p.R22.

Kinney, J.B. et al., 2010. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*, 107(20), pp.9158–9163.

Knaus, R. & Bujard, H., 1988. PL of coliphage lambda: an alternative solution for an efficient promoter. *The EMBO Journal*, 7(9), pp.2919–2923.

Knaus, R. & Bujard, H., 1990. Principles governing the activity of E. coli promoters. *Nucleic Acids and Molecular Biology*, (4).

Koonin, E.V. & Wolf, Y.I., 2010. Constraints and plasticity in genomeand molecular-phenome evolution. *Nature Reviews Genetics*, 11(7), pp.487–498.

Kreamer, N.N. et al., 2015. Predicting the impact of promotervariability on regulatory outputs. *Scientific Reports*, pp.1–13.

Kuhlman, T.E. & Cox, E.C., 2010. Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Research*, 38(6), p.e92.

Kumar, A. et al., 1993. The minus 35-recognition region of Escherichia coli sigma 70 is inessential for initiation of transcription at an "extended minus 10" promoter. *Journal of Molecular Biology*, 232(2), pp.406–418.

Kurtz, S. et al., 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*, 29(22), pp.4633–4642.

Lamberte, L.E. et al., 2017. Horizontally acquired AT-rich genes in Escherichia coli cause toxicity by sequestering RNA polymerase. *Nature microbiology*, 2, p.16249.

Lanzer, M. & Bujard, H., 2007. Promoters largely determine the efficiency of repressor action. *Proceedings of the National Academy of Sciences*, 85(23), pp.8973–8977.

Lawrence, J.G. & Roth, J.R., 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics*, 143(4), pp.1843–1860.

Lederberg, E.M. & Lederberg, J., 1953. Genetic Studies of Lysogenicity in Escherichia Coli. *Genetics*, 38(1), pp.51–64.

Levis, R., Hazelrigg, T. & Rubin, G.M., 1985. Effects of genomic position on the expression of transduced copies of the white gene of Drosophila. *Science*, 229(4713), pp.558–561.

Li, X.-Z., Plésiat, P. & Nikaido, H., 2015. The Challenge of Efflux-Mediated Antibiotic Resistance in Gram-Negative Bacteria. *Clinical Microbiology Reviews*, 28(2), pp.337–418.

Lind, P.A., Farr, A.D. & Rainey, P.B., 2015. Experimental evolution reveals hidden diversity in evolutionary pathways. *eLife*, 4, p.e07074.

Lisser, S. & Margalit, H., 1993. Compilation of E. coli mRNA promoter sequences. *Nucleic Acids Research*, 21(7), pp.1507–1516.

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

Mahillon, J. & Chandler, M., 1998. Insertion sequences. *Microbiology and Molecular Biology Reviews*, 62(3), pp.725–774.

Malan, T.P. et al., 1984. Mechanism of CRP-cAMP activation of lac operon transcription initiation activation of the P1 promoter. *Journal of Molecular Biology*, 180(4), pp.881–909.

Maniatis, T. et al., 1975. Recognition sequences of repressor and polymerase in the operators of bacteriophage lambda. *Cell*, 5(2), pp.109–113.

Martin, G. et al., 2013. The probability of evolutionary rescue: towards a quantitative comparison between theory and evolution experiments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1610), p.20120088.

Martinez, J.L. & Baquero, F., 2000. Mutation frequencies and antibiotic resistance. *Antimicrobial Agents and Chemotherapy*, 44(7), pp.1771–1777.

Masel, J., 2013. Q&A: Evolutionary capacitance. *BMC biology*, 11, p.103.

McClure, W.R. et al., 1983. DNA determinants of promoter selectivity in Escherichia coli. *Cold Spring Harbor symposia on quantitative biology*, 47 Pt 1, pp.477–481.

McLysaght, A. & Guerzoni, D., 2015. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678), p.20140332.

Miroslavova, N.S. & Busby, S.J.W., 2006. Investigations of the modular structure of bacterial promoters. *Biochemical Society symposium*, (73), pp.1–10.

Mitchell, J.E. et al., 2003. Identification and analysis of "extended -10" promoters in Escherichia coli. *Nucleic Acids Research*, 31(16), pp.4689–4695.

Moran, C.P. et al., 1982. Nucleotide-Sequences That Signal the Initiation of Transcription and Translation in Bacillus-Subtilis. *Molecular & general genetics : MGG*, 186(3), pp.339–346.

Navarre, W.W., 2006. Selective Silencing of Foreign DNA with Low GC Content by the H-NS Protein in Salmonella. *Science*, 313(5784), pp.236–238.

Neme, R. & Tautz, D., 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife*.

Notebaart, R.A. et al., 2014. Network-level architecture and the evolutionary potential of underground metabolism. *Proceedings of the National Academy of Sciences*, 111(32), pp.11762–11767.

Palmer, A.C. & Kishony, R., 2014. Opposing effects of target overexpression reveal drug mechanisms. *Nature Communications*, 5, pp.1–8.

Panyukov, V.V. & Ozoline, O.N., 2013. Promoters of Escherichia coli versus Promoter Islands: Function and Structure Comparison M. Isalan, ed. *PloS ONE*, 8(5), p.e62601.

Peterman, N. & Levine, E., 2016. Sort-seq under the hood: implications ofdesign choices on large-scale characterizationof sequence-function relations. *BMC Genomics*, pp.1–17.

Peterson, B.C. & Rownd, R.H., 1985. Drug resistance gene amplification of plasmid NR1 derivatives with various amounts of resistance determinant DNA. *Journal of Bacteriology*, 161(3), pp.1042–1048.

Peterson, M.L. & Reznikoff, W.S., 1985. Properties of lac P2 in vivo and in vitro. An overlapping RNA polymerase binding site within the lactose promoter. *Journal of Molecular Biology*, 185(3), pp.535–543.

Pleška, M., 2017. *Biology of restriction-modification systems at the single-cell and population level.*

Pósfai, G. et al., 2006. Emergent properties of reduced-genome Escherichia coli. *Science*, 312(5776), pp.1044–1046.

Pribnow, D., 1975. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proceedings of the National Academy of Sciences*, 72(3), pp.784–788.

Ptashne, M., 2004. *A genetic switch. Phage lambda revisited* 3rd ed., New York: Cold Spring Harbor Laboratory Press.

Rastogi, C. et al., 2018. Accurate and sensitive quantification of protein-DNA binding affinity. *Proceedings of the National Academy of Sciences*, 115(16), pp.E3692–E3701.

Reams, A.B. & Neidle, E.L., 2004. Selection for gene clustering by tandem duplication. *Annual Review of Microbiology*, 58(1), pp.119–142.

Rohs, R. et al., 2010. Origins of Specificity in Protein-DNA Recognition. *Annual Review of Biochemistry*, 79(1), pp.233–269.

Romero, D. & Palacios, R., 1997. Gene amplification and genomic plasticity in prokaryotes. *Annual Review of Genetics*, 31, pp.91–111.

Rouillard, J.M. et al., 2004. Gene2Oligo: oligonucleotide design for in vitro gene synthesis. *Nucleic Acids Research*, 32(Web Server), pp.W176–W180.

Ruff, E., Record, M., Jr. & Artsimovitch, I., 2015. Initial Events in Bacterial Transcription Initiation. *Biomolecules*, 5(2), pp.1035–1062.

Sawyer, S.A. et al., 1987. Distribution and abundance of insertion sequences among natural isolates of Escherichia coli. *Genetics*, 115, pp.51–63.

Schlötterer, C., 2015. Genes from scratch – the evolutionary fate of de novo genes. *Trends in Genetics*, 31(4), pp.215–219.

Schuster, S.C., 2007. Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1), pp.16–18.

Seaton, S.C. et al., 2011. Genome-wide selection for increased copy number in Acinetobacter baylyi ADP1: locus and context-dependent variation in gene amplification. *Molecular Microbiology*, 83(3), pp.520–535.

Sharan, S.K. et al., 2009. Recombineering: a homologous recombination-based method of genetic engineering. *Nature Protocols*, 4(2), pp.206–223.

Skordalakes, E. & Berger, J.M., 2003. Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading. *Cell*.

Snyder, L. & Champess, W., 2007. *Molecular Genetics of Bacteria* 3rd ed., Washington: ASM Press.

Soo, V.W.C., Hanson-Manful, P. & Patrick, W.M., 2011. Artificial gene amplification reveals an abundance of promiscuous resistance determinants in Escherichia coli. *Proceedings of the National Academy of Sciences*, 108(4), pp.1484–1489.

Stefano, J.E. & Gralla, J.D., 1982. Spacer mutations in the lac ps promoter. *Proceedings of the National Academy of Sciences*, 79(4), pp.1069–1072.

Steinrueck, M. & Guet, C.C., 2017. Complex chromosomal neighborhood effects determine the adaptive potential of a gene under selection. *eLife*, 6.

Stenström, C.M. & Isaksson, L.A., 2002. Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3' side. *Gene*, 288(1-2), pp.1–8.

Stoebel, D.M. et al., 2009. Compensatory Evolution of Gene Regulation in Response to Stress by Escherichia coli Lacking RpoS. *PLoS Genetics*, 5(10), p.e1000671.

Stormo, G.D., 2000. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1), pp.16–23.

Studnicka, G.M., 1988. Escherichia coli promoter -10 and -35 region homologies correlate with binding and isomerization kinetics. *The Biochemical journal*, 252(3), pp.825–831.

Sulavik, M.C. et al., 2001. Antibiotic Susceptibility Profiles of Escherichia coli Strains Lacking Multidrug Efflux Pump Genes. *Antimicrobial Agents and Chemotherapy*, 45(4), pp.1126–1136.

Thomason, L.C. et al., 2014. Recombineering: genetic engineering in bacteria using homologous recombination. *Current Protocols in Molecular Biology*, 106, pp.1.16.1–39.

Thomason, M.K. et al., 2014. Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in Escherichia coli R. L. Gourse, ed. *Journal of Bacteriology*, 197(1), pp.18–28.

Touchon, M. et al., 2009. Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths J. Casadesús, ed. *PLoS Genetics*, 5(1), p.e1000344.

Tuğrul, M., 2016. *Evolution of transcriptional regulatory sequences*.

Tuğrul, M. et al., 2015. Dynamics of Transcription Factor Binding Site Evolution J. C. Fay, ed. *PLoS Genetics*, 11(11), p.e1005639.

van der Meide, P.H. et al., 1982. Regulation of the expression of tufA and tufB, the two genes coding for the elongation factor EF-Tu in Escherichia coli. *FEBS letters*, 139(2), pp.325–330.

Vind, J. et al., 1993. Synthesis of proteins in Escherichia coli is limited by the concentration of free ribosomes. Expression from reporter genes does not always reflect functional mRNA levels. *Journal of Molecular Biology*, 231(3), pp.678–688.

Wagner, A., 2006. Cooperation is fleeting in the world of transposable elements. *PLoS Computational Biology*, 2(12), p.e162.

Wahl, G.M., Robert de Saint Vincent, B. & DeRose, M.L., 1984. Effect of chromosomal position on amplification of transfected genes in animal cells. *Nature*, 307(5951), pp.516–520.

Wilson, B.A. et al., 2017. Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. *Nature Publishing Group*, 1(6), pp.0146–0146.

Wittkopp, P.J., Haerum, B.K. & Clark, A.G., 2004. Evolutionary changes in cis and trans gene regulation. *Nature*, 430(6995), pp.85–88.

Wolf, L., Silander, O.K. & van Nimwegen, E., 2015. Expression noise facilitates the evolution of gene regulation. *eLife*, 4.

Wray, G.A., 2007. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, 8(3), pp.206–216.

Xiao, H. et al., 2008. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science*, 319(5869), pp.1527–1530.

Yomtovian, I. et al., 2010. Composition bias and the origin of ORFan genes. *Bioinformatics*, 26(8), pp.996–999.

Yona, A.H., Alm, E.J. & Gore, J., 2018. Random sequences rapidly evolve into de novo promoters. *Nature Communications*, 9(1), p.1530.

Zaslaver, A. et al., 2006. A comprehensive library of fluorescent transcriptional reporters for Escherichia coli. *Nature Methods*, 3(8), pp.623–628.

Zhang, Z. & Saier, M.H., Jr, 2009. A Novel Mechanism of Transposon-Mediated Gene Activation. *PLoS Genetics*, 5(10), p.e1000689.

# 6 Appendix

## *6.1 Figure Supplements of Chapter 2*

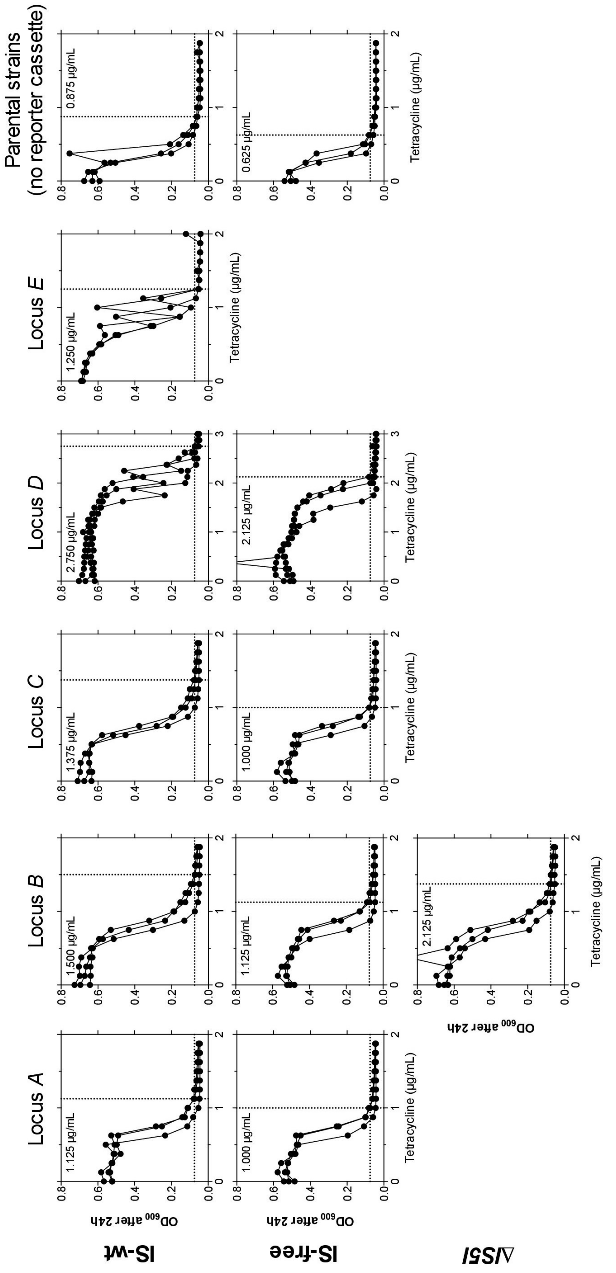


***Figure 1 - Figure Supplement 1. Fine-scale determination of MICs of tetracycline for ancestor strains used in experimental evolution.*** *$OD_{600}$ (platereader units) after 24 hr is shown across tetracycline concentrations (triplicates). Panel columns = integration loci of the reporter cassette, panel rows = genetic background. Note the different scaling of the x-axis for D strains. We define MIC (dashed vertical lines and inset values) as the lowest concentration that restricts growth to $OD_{600}$ ≤0.075 (= $OD_t$, plate reader units, dashed horizontal lines) in all three replicates. We regard the highest replicate value of strain E at 2 µg/mL as an outlier uninformative about ancestral drug sensitivity, as this culture showed highly increased CFP fluorescence indicative of reporter cassette amplification. The selective conditions in evolution experiments (i.e., tetracycline concentrations) were adjusted according to strain-specific MICs to make results more comparable between strains. Without such an adjustment in tetracycline concentrations, different MICs would cause large differences in population sizes and consequently in the probability of acquiring beneficial mutations.*

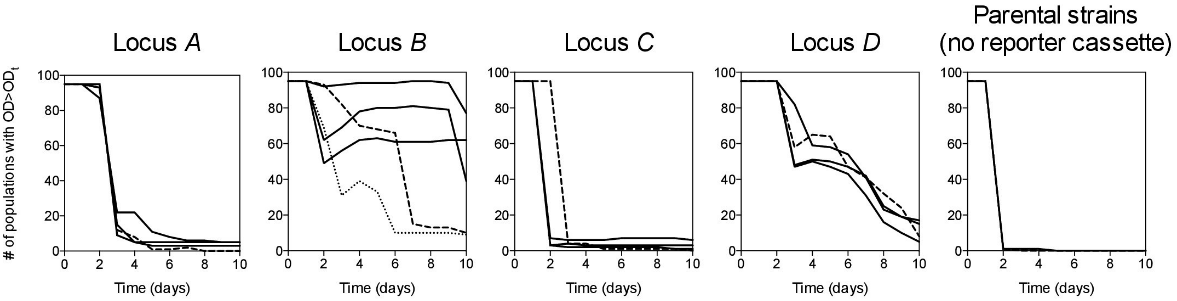


***Figure 2 - Figure Supplement 1. Survival curves of 95 populations in evolution experiments.*** *OD$_t$ = threshold OD$_{600}$. Solid lines = IS-wt genetic background, dashed lines = IS-free genetic background, dotted line in Locus B panel = strain BΔIS5I. Triple solid lines for loci A-D represent replicate sets of evolution experiments. Local minima in the number of populations are due to populations that fell below OD$_t$ only transiently.*

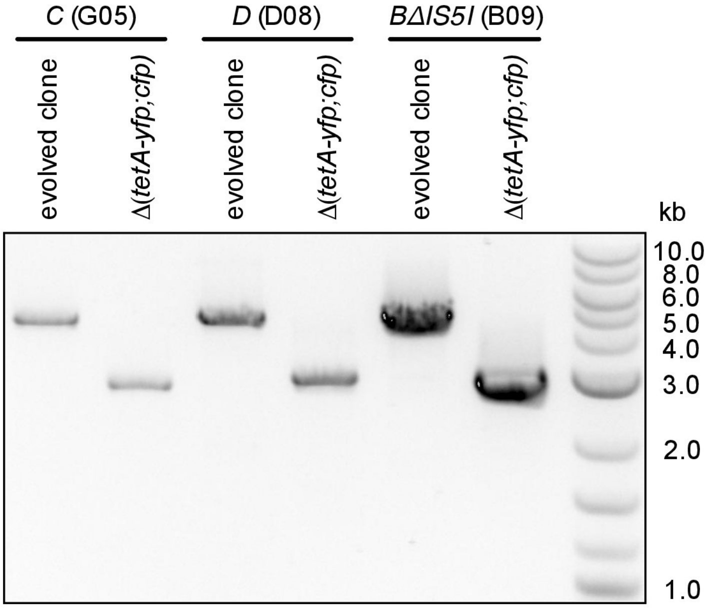


***Figure 2 - Figure Supplement 2. PCR products confirming the deletion of reporter cassette genes in clones shown in Figure 2B.*** *Colony PCR was performed with primers flanking integration loci.*

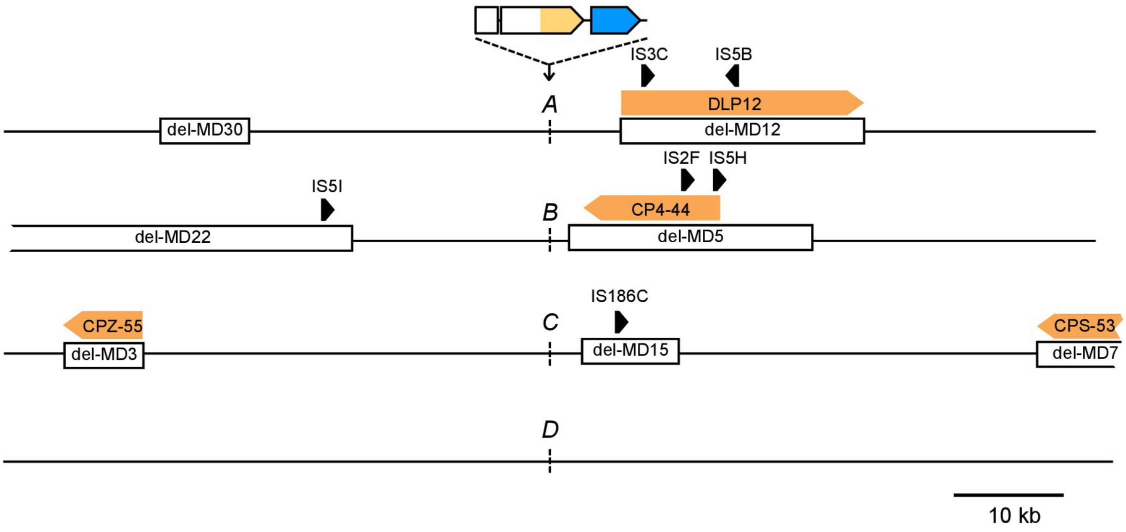


***Figure 3 - Figure Supplement 1. Differences in the chromosomal neighborhood (100 kb) of loci A-D between IS-wt and IS-free strains.****White boxes = regions deleted in the IS-free strains derived from strain MDS42 (Pósfai et al., 2006). Orange arrows = prophages, black arrows = insertion sequences. Chromosomal neighborhoods of loci B, C, and D are shown reversed with respect to conventional chromosome coordinates, so that the orientation relative to the reporter cassette is shown in the same way for all four loci. Reporter cassette genes are not drawn to scale.*

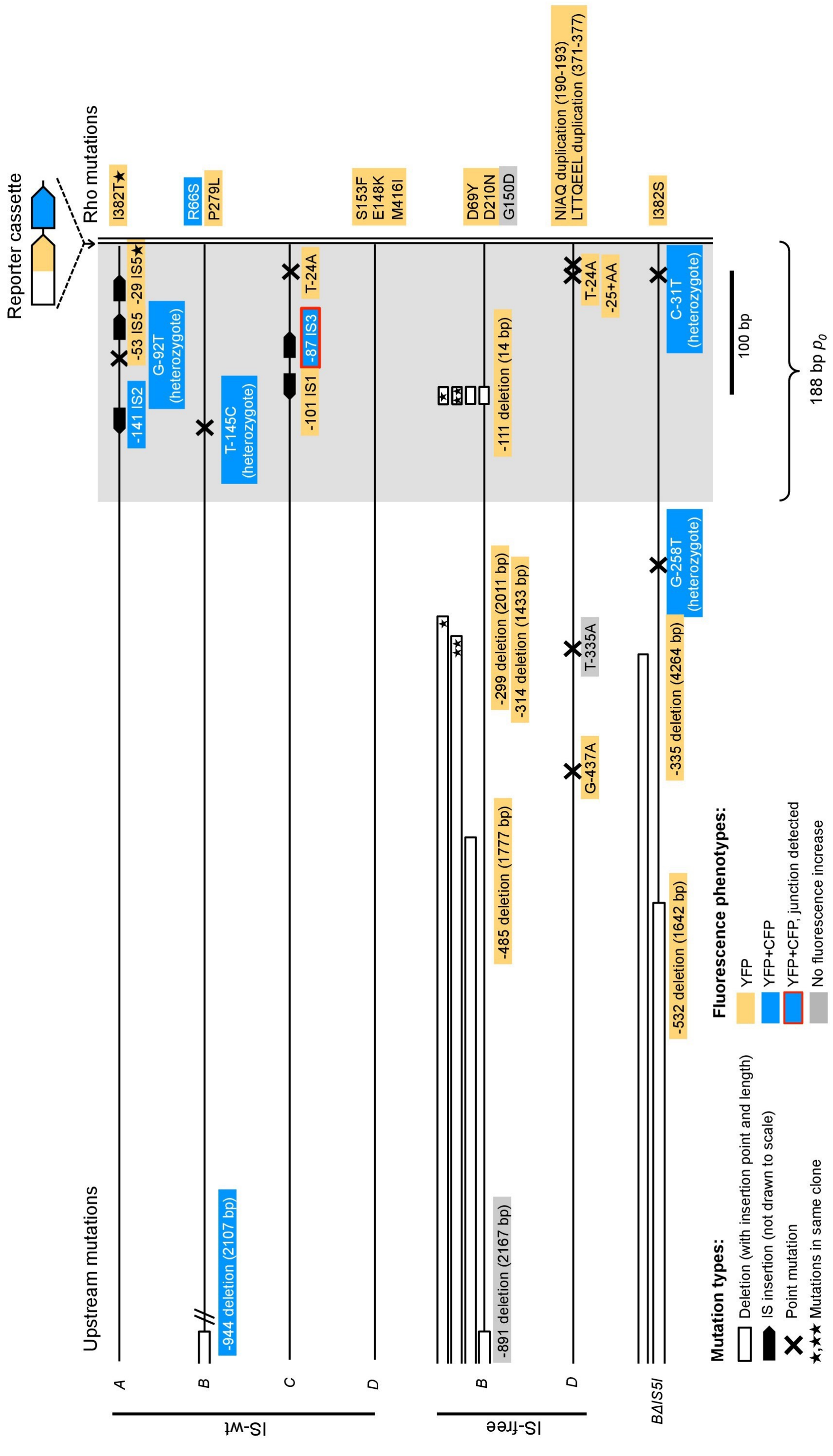


***Figure 3 - Figure Supplement 2. Graphical overview of mutations identified by sequencing.*** *All mutations found within 1 kb DNA upstream of tetA-yfp are labeled with their distance to the tetA-yfp start codon and the fluorescence phenotype of the population they were found in (YFP or YFP+CFP). Grey-shaded area indicates the 188 bp random DNA sequence common to all strains. Trans mutations in the Rho protein are shown at the right edge. IS-free A and C strains are not shown as they did not give any survivors. 'Heterozygote' indicates overlapping peaks in the sequence chromatogram, which suggests that the mutation is present only in some copies contained in the amplification. Red box frame indicates that for this amplification, we showed by PCR that the junction of this amplification was at the breakpoint between the newly inserted IS3 copy and a second IS3 copy downstream of locus C. Mutations identified in additional replicate experiments are not shown.*

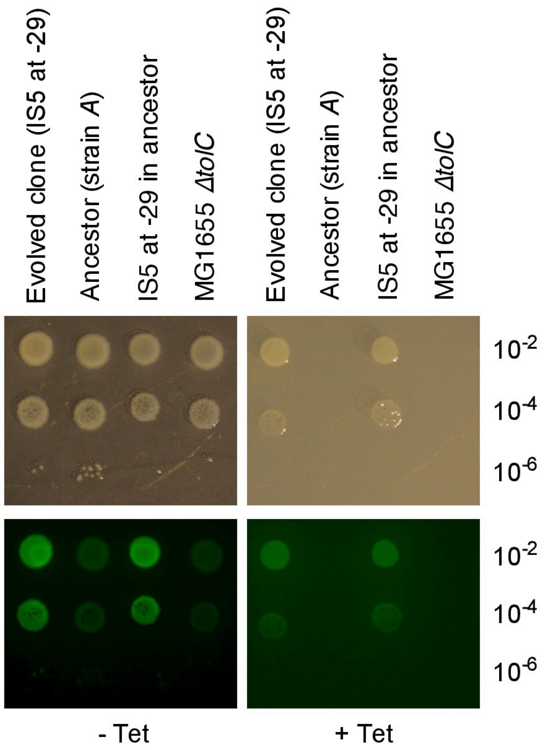


***Figure 3 - Figure Supplement 3. An upstream IS5 insertion in the chromosomal context of the reporter cassette confers resistance to tetracycline and increases tetA-yfp fluorescence.*** *Pictures show brightfield (top) and YFP-fluorescence (bottom) images of cultures spotted at different dilutions on solid medium with and without tetracycline (2.25 µg/mL). We used an evolved clone isolated from population A09 of strain A in which IS5 was found inserted 29 bp upstream of the TetA-YFP start codon as a donor for P1 transduction of the reporter cassette with upstream IS5. MG1655 ΔtolC, the cassette-free parent strain of strain A was used as recipient strain for the transduction.*

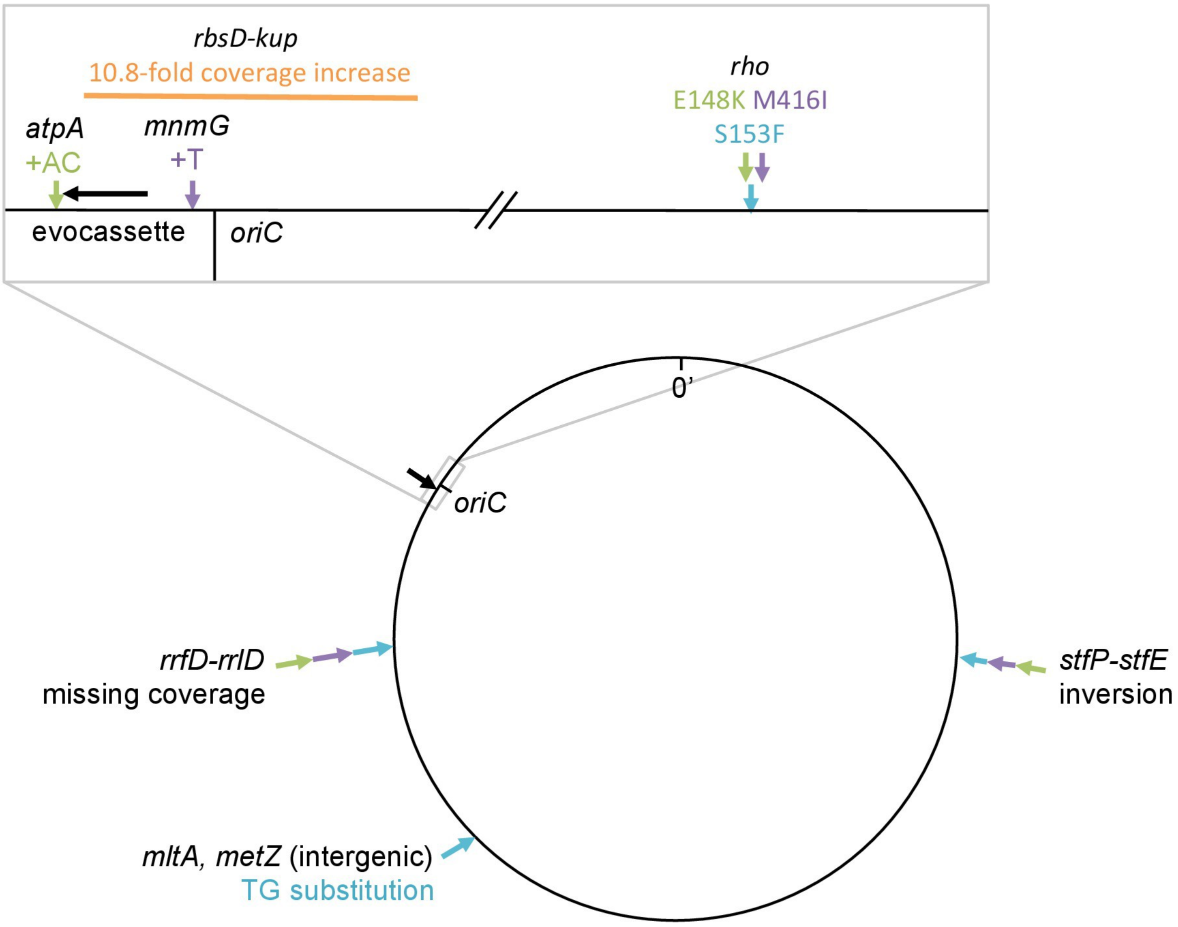


***Figure 3 - Figure Supplement 4. Mutations identified by whole genome sequencing of clones from four rescued populations of IS-wt strain D.*** *All mutations identified by the breseq pipeline (Barrick et al., 2014) in reference to the strain D ancestral genome are shown. Black arrow in magnified box: reporter cassette. Mutations from the same clone are indicated by the same color. Orange (source population C10): 11-fold amplification of a region including tetA-yfp and half of the cfp gene, explaining why we did not observe increased CFP fluorescence. Notably, this amplification included the origin of replication. Blue, purple, green: Mutations found in sequenced clones of the three remaining populations (blue = A11, purple = D08, green = C08), in which we consider non-synonymous substitutions in rho as main adaptive mutations. We interpret missing coverage in one of the rRNA operons as an assembly artifact related to the multiplicity of rRNA operons, rather than as a deletion, as no corresponding junction was detected. The inversion found in the stfP-stfE region of prophage e14 is catalyzed by the e14-encoded Pin recombinase (van de Putte et al., 1984), possibly expressed as a secondary effect of rho mutations (Cardinale et al., 2008). We thus assume that the same inversion was found three times not because it was adaptive, but because it was the consequence of an adaptive mutation in rho. Mutations at other sites were not tested for their fitness effect. Fastq files are deposited online: http://dx.doi.org/10.15479/AT:ISTA:65.*

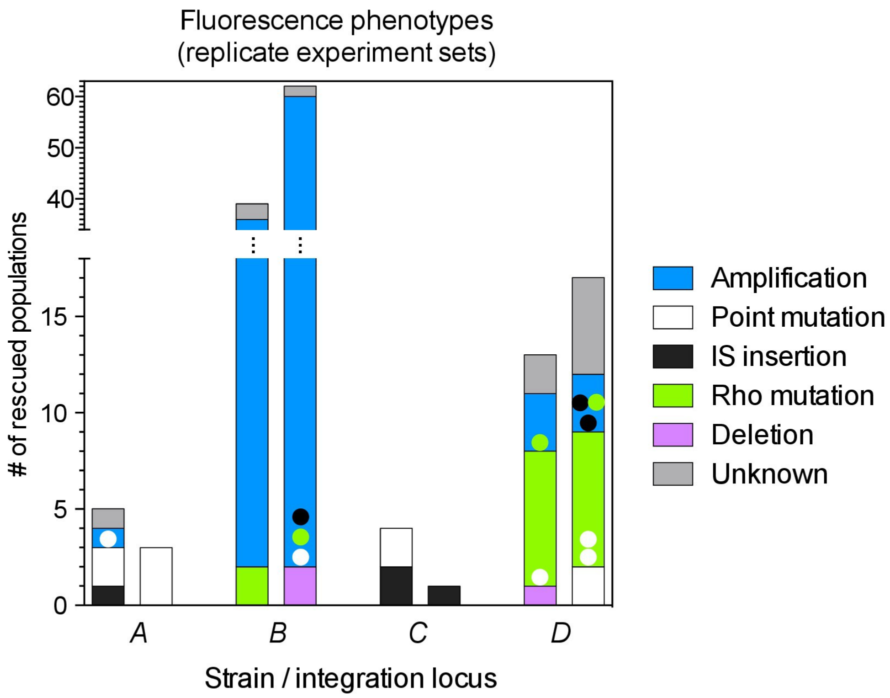


***Figure 3 - Figure Supplement 5. Numbers of rescued populations by mutation type in two additional replicate sets of evolution experiments.*** *Each bar represents the number of rescued populations out of 95 started populations per experiment.*

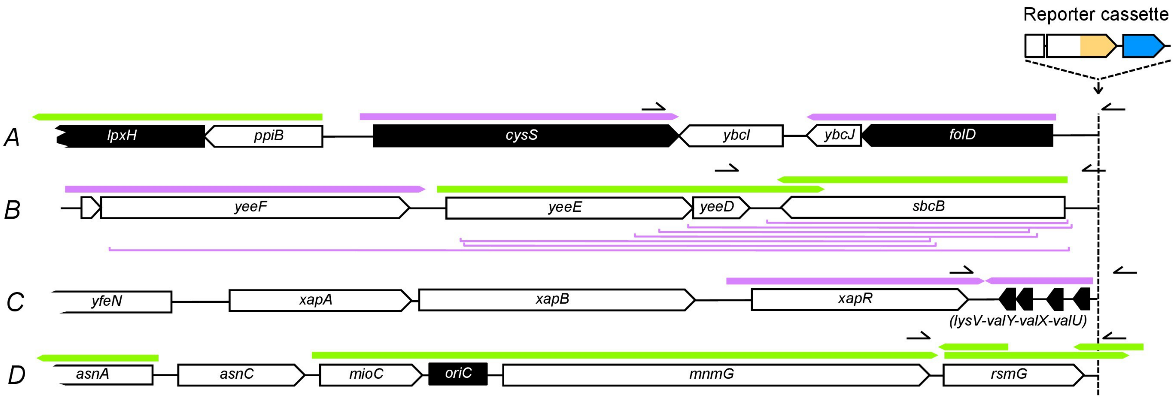


***Figure 4 - Figure Supplement 1. Fully annotated genes and putatively expressed transcripts of either orientation upstream of the reporter cassette insertion loci.*** *Genes and transcripts upstream of loci A, B, C, and D. Black arrows = essential genes (see Methods), white arrows = non-essential genes, purple arrows = intrinsically terminated transcripts, green arrows = Rho-terminated transcripts, purple brackets = deletions. Start- and endpoints of expressed transcripts and termination mode (intrinsic or factor-dependent) were taken from a recent dataset (Conway et al., 2014), for which RNA from E. coli cells grown in minimal glucose medium was sequenced at base pair resolution. Pointers on the right = position of PCR products shown in Figure 4C.*

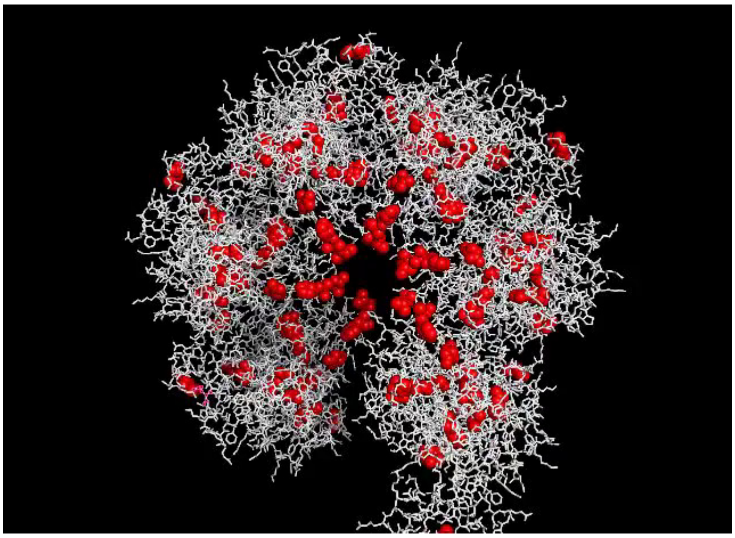


***Figure 4 - Video 1. Animated structure of the Rho hexamer with mutated residues highlighted.*** *Mutations were mapped on the previously published structure of Rho (Skordalakes and Berger, 2003). – This video can be played at https://doi.org/10.7554/eLife.25100.021*

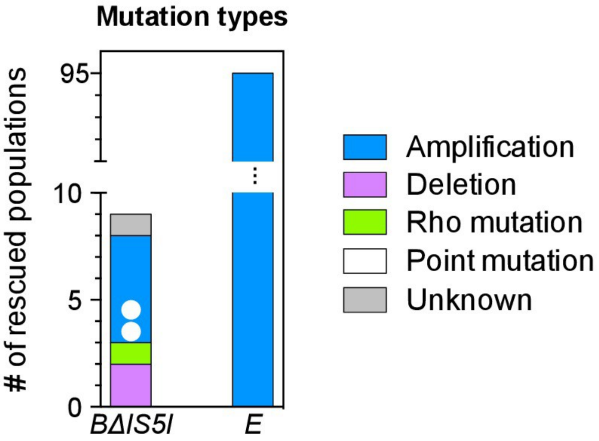


***Figure 5 - Figure Supplement 1. Rescued populations of strains BΔIS5I and E by mutation type.*** *Number of rescued populations out of 95 replicates, shown by mutation type. Colored dots = later mutations occurring on top of earlier mutations.*

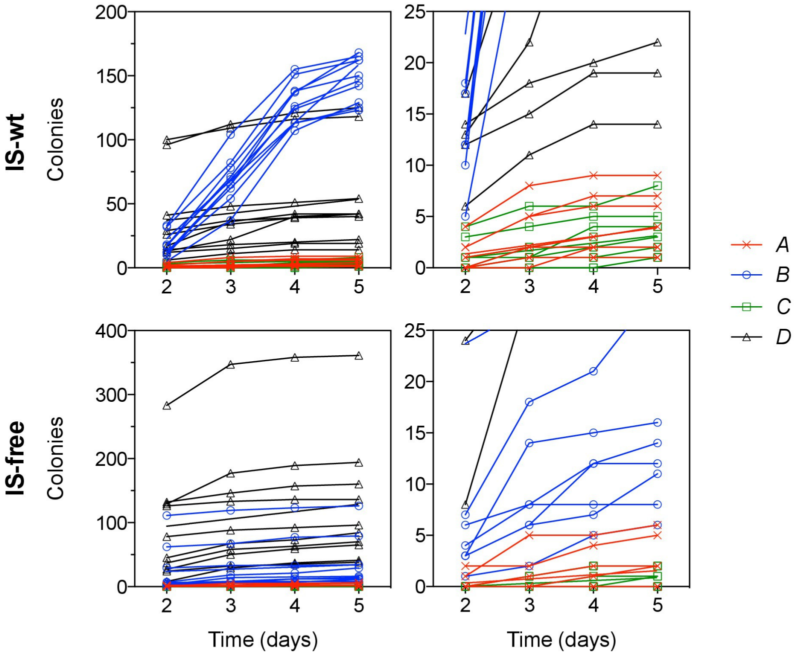


***Figure 6 - Figure Supplement 1. Colony appearance over time in plating experiments.*** *Each line represents the number of colonies on one of 10 replicate plates per strain. Right panels show the same data as on the left with different y-axis scaling.*

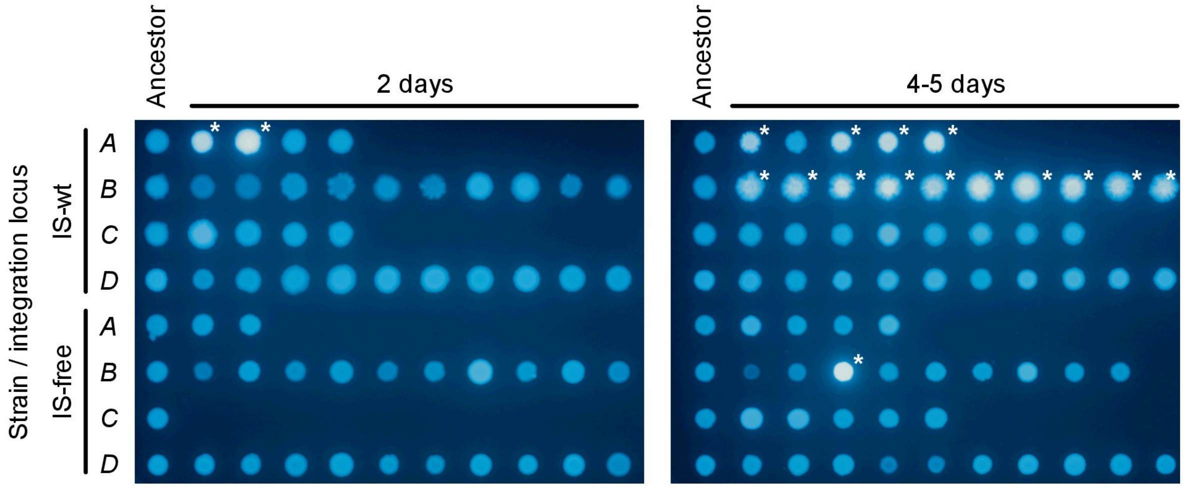


***Figure 6 - Figure Supplement 2. CFP-fluorescence of cultures spotted on non-selective medium used to obtain pie-chart data in Figure 6.*** *The leftmost column of spots on each picture is derived from colonies of the ancestor strain plated on non-selective medium. The other spots are derived from colonies that appeared on day 2 (left picture) or on days 4–5 (right picture) after plating. One colony was picked from every replicate plate on which at least one colony had appeared in the respective time interval. White asterisks indicate spots with CFP fluorescence intensity greater than 6 standard deviations above the mean fluorescence intensity of all ancestor spots.*

## *6.2 List of source data files pertaining to Chapter 2 (available online)*

**Figure 1—source data 1.** Chromosomal coordinates of reporter cassette insertion loci.

https://doi.org/10.7554/eLife.25100.003

**Figure 1—source data 2.** Source data for Figure 1E.

Mean and standard deviation of chromosomal *tetA-yfp* copy number (qPCR) and final CFP fluorescence (plate reader data).

https://doi.org/10.7554/eLife.25100.004

**Figure 3—source data 1.** Source data for Figure 3C–E.

$OD_{600}$-normalized fluorescence values measured in exponential phase (six replicates).

https://doi.org/10.7554/eLife.25100.012

**Figure 4—video 1—source data 1.** Rho mutations from all replicate evolution experiments.

28 unique mutations (substitutions at 22 different amino acid residues, two internal duplications and one upstream insertion) were found in 31 rescued populations. Affected amino acid residues of Rho are highlighted in red in Figure 4—video 1.

https://doi.org/10.7554/eLife.25100.022

**Figure 6—figure supplement 2—source data 1**

Mean fluorescence intensity values of culture spots and thresholding for identification of colonies with extensive amplifications.

https://doi.org/10.7554/eLife.25100.032

**Figure 7—source data 1**

Extended legend of Figure 7A explaining each arrow and what loci are affected by respective interactions.

https://doi.org/10.7554/eLife.25100.034

**Figure 7—source data 2**

List of *E. coli* genes included in the analysis shown in Figure 7B and their assignment to the three sets shown by colored circles.

https://doi.org/10.7554/eLife.25100.035

## *6.3 List of additional files pertaining to Chapter 2 (available online)*

**Supplementary file 1.** Population trajectories.

Set of 96-panel figures showing OD and OD-normalized fluorescence values for each population in each of 18 evolution experiments.

https://doi.org/10.7554/eLife.25100.036

**Supplementary file 2.** Source data populations.

Excel table containing information on survival, fluorescence phenotypes, sequences, and mutation types of every experimental population, as well as information on which populations where used for further investigation (plasmid reconstruction etc.). This contains the source data of Figures 2AC, 3AB, 5CD, Table 1, and respective Figure Supplements.

https://doi.org/10.7554/eLife.25100.037

**Supplementary file 3.** Strains, plasmids, oligonucleotides.

Excel table with all strains, plasmids and oligonucleotides used in this study.

https://doi.org/10.7554/eLife.25100.038

**Source code 1**

Compressed file containing Matlab scripts and OD/YFP/CFP plate-reader raw data files of evolution experiments.

https://doi.org/10.7554/eLife.25100.039