THE EVOLUTION OF GENE EXPRESSION BY COPY NUMBER AND POINT MUTATIONS

by

Isabella Tomanek-Leithner

August, 2020

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

The dissertation of Isabella Tomanek-Leithner, titled *The Evolution Of Gene Expression By Copy Number And Point Mutations* is approved by:

Supervisor: Calin Guet, IST Austria, Klosterneuburg, Austria

Signature: _____

Committee Member: Nick Barton, IST Austria, Klosterneuburg, Austria

Signature: _____

Committee Member: Jonathan Bollback, University of Liverpool, UK

Signature: _____

Committee Member: Laurence Hurst, University of Bath, UK

Signature: _____

Exam Chair: Georgios Katsaros, IST Austria, Klosterneuburg, Austria

Signature: _____

signed page is on file

IST Austria Thesis, ISSN: 2663-337X

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: _____

Isabella Tomanek-Leithner

October 16, 2020

Signed page is on file

Abstract

Mutations are the raw material of evolution and come in many different flavors. Point mutations change a single letter in the DNA sequence, while copy number mutations like duplications or deletions add or remove many letters of the DNA sequence simultaneously. Each type of mutation exhibits specific properties like its rate of formation and reversal.
Gene expression is a fundamental phenotype that can be altered by both, point and copy number mutations. The following thesis is concerned with the dynamics of gene expression evolution and how it is affected by the properties exhibited by point and copy number mutations. Specifically, we are considering i) copy number mutations during adaptation to fluctuating environments and ii) the interaction of copy number and point mutations during adaptation to constant environments.

Acknowledgments

I am truly grateful to my supervisor, Călin Guet for the scientific freedom I enjoyed all these years, for the time I was given to wander off my original project and explore things in my own pace. Also, for his constant mentorship, for encouraging me to look beyond canonical answers, to think outside boxes and for the many uplifting conversations.

I want to thank my co-supervisor, Jon Bollback, above all, for equipping me with a basic appreciation of population genetics such that I could interpret my experiments based on a solid foundation.

I want to also thank Gasper Tkačik, for his guidance and help, for his scrutinizing look on data of any kind, and for serving as a tireless consultant when we approached him with data-model issues.

I am very grateful to the members of my thesis committee, Laurence Hurst and Nick Barton for insightful comments on the manuscript, feedback and interesting discussions.

I want to thank OMV for sponsoring the first three years of my PhD.

I am grateful that I could spend these last years around fantastic people at IST. I especially want to thank Mato Lagator for his encouragement, mentorship and for very valuable lessons on data presentation and storytelling; Rok Grah, for a truly fun collaboration; Anna Andersson for finding lovely peaks in extremely messy data; Tobias Bergmiller, for introducing me to the dark arts of microfluidics, and for knowing quick and dirty shortcuts to any lab protocol; Remy Chait, for his skillful tinkering around the lab, especially for building the macroscope, which was essential to study beasts as fleeting as gene amplifications; Magdalena Steinrueck, for laying such solid foundations I could build upon; Kathrin Tomasek for being our lab´s all-time magician; Maros Pleska for critical comments on our manuscript and for sharing his creative ways of masterfully screwing up experiments. Kirti Jain, Bor Kavcic, Michael Lukesch and Anna Staron, for comments on the manuscript, coffee breaks, shared cookies and a lot of fun.

My final thanks go to my family. What more can a person ask from life than having someone to share one´s enthusiasm with? I want to thank Alexander Leithner, my favorite biologist, for his endless support and love.

As my career as a biologist started early with jars full of tadpoles, leeches, snails and salamanders (that would not always be confined to said jars), I am thankful to my parents, for letting me explore. I am also thankful to my daughter, for reminding me, that the nature I´m trying to understand mostly resides outside a test tube.

About the Author

Isabella Tomanek completed a BSc in Microbiology and Genetics and a MSc in Molecular Microbiology and Immunobiology at the University of Vienna, spending one semester also at the University of Bergen, Norway. She joined IST Austria in 2013 and rotated in the labs of Jon Bollback and Tobias Bollenbach before joining the laboratory of Calin Guet, jointly supervised by Jon Bollback. Starting off with a project aimed at studying the evolution of complex bacterial promoters, Isabella established a dual selection system that would allow directed evolution of gene regulation. When selecting for increased gene expression in bacteria it is hard not to encounter, and become fascinated by, copy number mutations, which were appearing in all of her evolution experiments. During the second year of her PhD Isabella spent the summer as teaching assistant at the Hopkins Microbiology course at the University of Stanford, CA. There, Isabella met the godfather of gene amplifications, John Roth, and started thinking about the ecological and evolutionary implications of copy number mutations that some people, including John, view more as 'regulatory state' rather than a mutation. Having a dual selection system at hand allowed to directly test this idea. This way, Isabella´s main PhD project moved from promoter evolution to studying how 'regulation' can work in the absence of any proper promoter. Eventually, her thesis work returned to promoters, using them as a convenient tool to detect point mutations, thereby testing a second implication of the idea that gene amplifications act as a transient 'regulatory state': Amplifications could hinder the evolution by point mutation, granting adaptation to transient stress while acting as a genomic buffer against less reversible adaptations.

List of Publications Appearing in Thesis

1. Tomanek, I., Grah, R., Lagator, M. *et al.* Gene amplification as a form of population-level gene expression regulation. *Nat Ecol Evol* **4,** 612–625 (2020). https://doi.org/10.1038/s41559-020-1132-7

Table of Contents

List of Figures

List of Tables

List of Symbols/Abbreviations

AMGET	amplification-mediated gene expression tuning
bp	base pairs
CNV	copy number variation
DOG	2-deoxy-galactose
GDA	gene duplication and amplification
IAD	innovation amplification divergence
indel	insertion or deletion of bases
IS	insertion sequence
kbp	kilo base pairs
LOF	loss of function
OD_{600}	optical density measured at 600 nm
p0	random promoter-sequence
PCR	polymerase chain reaction
SNP	single nucleotide polymorphism
WGD	whole genome duplication

Preface

The year 2020 has so far been a –not too gentle– reminder for humanity that it is the tiniest of replicators, the unseen majority of life, which really rule this planet.

I am hopeful, however, that eventually this year will also be a reminder of just how powerful a certain way of perceiving the world is, a way which we call the scientific method: asking a question, coming up with an answer and some implications that answer must hold. Then trying honestly and hard to prove the answer wrong with a reality check of its implications. If the answer survives the test, it must be true – of course we will never know for sure. Instead, we can ask more questions and come up with more answers and predictions they imply. While the scientific method really is nothing else than a way of letting nature challenge our ideas, it is still the best we can do to understand the world around us given the limits of our five senses.

I have always perceived it as an immense privilege that going to work every day means making visible the unseen majority and asking them some questions, even if the answer I get more often than not is: “Next time, label your tubes correctly!”.

1 Introduction

Mutations are the raw material of evolution, upon which selection or drift act to shift population genotypes. While both experiments and theory often only consider point mutations, that is, changes to a single nucleotide in the DNA sequence, nature is full of different bigger-scale mutations, such as transpositions, duplications, deletions and inversions. These bigger-scale mutations can span stretches of DNA ranging from only a few base pairs up to half a bacterial genome (Anderson and Roth, 1977; Darmon and Leach, 2014). In organisms, which harbor more than one chromosome, bigger-scale mutations can also encompass the duplication, deletion (Smith and Sheltzer, 2018) or fusion of entire chromosomes (Fan *et al.*, 2002; Mizuno *et al.*, 2013).

All of these mutations have different general properties, such as their rates of formation and reversal, but also their mechanism of formation, which may be tied to certain sequence features. For instance, deletions and duplications - here collectively referred to as copy number mutations - form at high frequency via *recA*-dependent homologous recombination of genomic regions flanked by direct repeat sequences, such as mobile genetic elements or *rRNA* genes (Reams and Roth, 2015). Depending on the length and distance of these direct repeats, duplications or deletions occur with a frequency ranging from 10^{-6} up to 10^{-2} per cell per generation in bacteria. Unlike duplications, inversions form in the presence of inverse repeats, which can be positioned as far as on opposite points of a circular bacterial chromosome (Cui *et al.*, 2012; Slager, Aprianto and Veening, 2018). In addition, the process of illegitimate recombination creates deletions or duplications in the complete absence of repeats, but generally at a lower frequency than homologous recombination. Importantly, once a duplication occurred, and irrespective of its formation mechanism, it becomes highly unstable as the long stretch of sequence homology is prone for high rates of homologous recombination (Roth *et al.*, 1988; Andersson and Hughes, 2009; Reams and Roth, 2015). In other words, a duplication is a mutation activated for either deletion or further duplication, also referred to as amplification.
Another type of sequence-feature associated with high rates of mutations, which does not involve homologous recombination, are micro-repeats of up to five base pairs in length. These repeats expand and contract at a frequency of 10^{-4} per cell per generation in bacteria, a mechanism that allows to alter gene expression (Vinces *et al.*, 2009) and serves as the basis for antigenic variation in a variety of pathogens (Darmon and Leach, 2014).

The general properties of mutations such as their rate of formation and reversal might influence the evolutionary dynamics in important ways, but are rarely considered. For instance, mutations to the copy number of genes or genomic regions are orders of magnitude more frequent than point mutations (Drake *et al.*, 1998; Elez *et al.*, 2010). Moreover, copy number mutations not only differ from point mutations in their frequency

of occurrence, but also in their reversibility. Their intrinsic rate of deletion, in combination with a sometimes significant fitness cost (Bergthorsson, Andersson and Roth, 2007; Mats E Pettersson *et al.*, 2009; Reams *et al.*, 2010), has one important implication: observing more than two copies of a certain gene or genomic region means they are most probably an adaptation, or at least have been one in the recent evolutionary past. Point mutations can fix via drift, but as we will see in chapters two and four of this thesis, gene amplifications remain polymorphic and thus hardly ever fix even under strong positive selection. This feature has been of practical medical relevance and copy number variation (CVN) in human cancers is used to identify causative oncogenes (Albertson, 2006).

1.1 The effect of mutations on gene expression

Gene expression is a fundamental phenotype that can be altered by all mutations mentioned above. Point mutations or small duplications or deletions in the promoter region of a gene can change the sequence motifs recognized by the σ70 subunit of RNA polymerase, like the −10 element (consensus: TATAAT) and the −35 element (consensus: TTGACA), thereby altering the level of transcription (Barnard, Wolfe and Busby, 2004; Yona, Alm and Gore, 2018).

Copy number mutations of the entire gene will, as a first approximation, increase its expression by means of elevated gene dosage (Elde *et al.*, 2012; Näsvall *et al.*, 2012; Belikova *et al.*, 2020; Todd and Selmecki, 2020). Not surprisingly, due to their high rate of formation, gene amplifications are found as an adaptation to any situation where fast increases in gene expression are needed: resistance to antibiotics, pesticides or drugs via the over-expression of some resistance determinant (Prody *et al.*, 1989; Albertson, 2006; Bass and Field, 2011b; Nicoloff *et al.*, 2019). Examples of animals and plants illustrate the short time scales on which copy number mutation can increase gene expression levels. For instance, in species of *Xenopus* and *Drosophila* programmed gene amplification during development provides a mechanism to drastically increase the expression of certain gene products required at high levels, such as rRNA (Kafatos, Orr and Delidakis, 1985; Claycomb and Orr-Weaver, 2005). An even more intriguing form of gene amplification of multiple loci occurs in flax (*Linum usitatissimum*) with heritable copy number variation occurs within a single generation in a manner that is responsive to the current environmental conditions (Cullis, 2005).

In general, elevating a pre-existing low level of gene expression seems a relatively easy task from an evolutionary standpoint. Not only gene amplifications lead to fast increases in expression, constitutive gene expression also readily evolves from a random sequence by point mutations (Wolf, Silander and van Nimwegen, 2015; Steinrueck and Guet, 2017; Yona, Alm and Gore, 2018). The underlying reason seems to be the binding flexibility of the σ70 subunit of RNA polymerase (Lagator *et al.*, 2020), which means that as many as 10% of random sequences are sufficiently close to the canonical binding site to be expressed.

Another 60% of all random sequences are only one point mutation away from the consensus binding site (Yona, Alm and Gore, 2018).
For regulated promoters, which even in bacteria exhibit complex architectures consisting of several transcription factor-binding sites, the situation is more complicated. On the one hand, genomic studies hint at the rapid evolutionary turnover of promoter sequences (Perez and Groisman, 2009; Matus-Garcia, Nijveen and van Passel, 2012a; Nijveen, Matus-Garcia and van Passel, 2012; Oren et al., 2014; van Passel, Nijveen and Wahl, 2014) that is complemented by anecdotal evidence demonstrating instances of rapid regulatory rewiring of promoter sequences (Anderson and Roth, 1978; Kloeckener-Gruissem and Freeling, 1995; Zinser and Kolter, 2004; Blount *et al.*, 2012; Taylor *et al.*, 2015). On the other hand, evolutionary models seem to indicate that the evolution of regulated promoters by point mutations is very slow (Tuğrul *et al.*, 2015). Mutations other than point mutations may solve this paradox (Surguchov, 1991; Matus-Garcia, Nijveen and van Passel, 2012b; Tuğrul *et al.*, 2015). However, to find out which kinds of mutations are involved in promoter evolution, we need to catch regulatory evolution as it happens. Chapter three will discuss an experimental system potentially useful for answering this question, while the second chapter will explore an evolutionary strategy that can occur in the absence of an evolved complex promoter.

1.2 Questions addressed in this thesis

The main overarching question of this thesis is concerned with the fate of copy number and point mutations in fluctuating, as well as in constant environments that both select for altered levels of gene expression: How does gene expression change in response to selection? What kinds of adaptive mutations occur? What kind of evolutionary dynamics do we observe if copy number and point mutations both have adaptive value?
Our general approach to address these questions is experimental evolution. While many studies of laboratory evolution are focused primarily on understanding the molecular basis of adaptive phenotypes (Lang and Desai, 2014), we are trying to use our experimental system to abstract away from any specific “molecular solutions” we observe. Instead, we are interested in finding general patterns that may have been overlooked.
As an idealized concept, evolutionary dynamics means the study of evolution as a process without reference to the specific phenotypes in question. In reality, of course, the biological details set the scene for all evolutionary dynamics (Cvijović, Nguyen Ba and Desai, 2018). By choosing simple experimental systems with relatively general phenotypes, such as the level of gene expression, there is hope that we can abstract from a given specific fitness landscape, yet still learn some general rules.
Our simple experimental system consists of one gene, *galK* and the phenotype we are interested in is simply its expression level. Since we are not interested in *galK* itself, we could have chosen virtually any other gene, as long as it has the potential to be transcribed

and exhibits the discernible phenotype of an expression level. We chose *galK* for one special property and that is the fact that we can experimentally control whether or not expression is beneficial (selected for) or detrimental (selected against). In order to study the evolution of gene expression by point mutations and copy number mutations we attached two different fluorophores to *galK*.

The second chapter of this thesis deals with the evolutionary dynamics of copy number mutations under fluctuating selection. There, the main question is whether gene amplification can serve as a crude solution to the problem of regulating gene expression in the absence of canonical gene regulatory systems. A few simple experiments serve as a proof-of-principle that this is indeed the case. As this chapter is the result of a lot of collaborative effort, the experiments are complemented by a simple model, which tries to generalize the idea that unstable tandem duplications allow populations to respond to fluctuating selection also beyond the realm of our experimental system.

The third chapter is a methodological one, elaborating on the fluctuating selection system introduced in the second chapter. It briefly discusses the general use of dual selection systems and its potential use in experimental evolution studies of gene regulation. This chapter explores potential improvements to the selection system, aiming to reduce the occurrence of escape mutants, which strongly limit its use in any long-term experiment.

The fourth chapter is focused on the early dynamics of adaptation that arise between point mutations and copy number mutations. The considerations are relevant for the evolution of paralogs, the process of duplication-divergence. However, instead of experimentally evolving paralogs, we are simply selecting for increased expression of an existing, but weakly expressed function. To do so, we make use of our *galK* selection and dual reporter system that allows us to *phenotypically* distinguish copy number and point mutations. We focus on adaptation according to the innovation amplification divergence (IAD) model (Bergthorsson, Andersson and Roth, 2007; Näsvall *et al.*, 2012), which posits that under selection for a weakly expressed biological function, gene amplification provides an initial adaptation until eventually point mutations refine the selected function and additional gene copies are lost again. We test one important and widely cited, but never directly tested, assertion of the IAD model, namely that amplification increases the chance for point mutations to fix.

2 Gene amplification as form of population-level gene expression regulation

This chapter was published as **Tomanek I**∗, Grah R∗, Lagator M, Andersson AMC, Bollback JP, Tkačik G, Guet CC. Gene amplification as a form of population-level gene expression regulation. *Nature Ecology & Evolution*. 4(4):612- 625, 2020.
∗ These authors contributed equally.

Some changes have been made to the text in order to integrate it into this thesis.
Supplementary Notes, Figures and Tables can be found at the Appendix of the thesis.

Contributions:
Tomanek I. has done all experiments.
Grah R. has constructed the model and has done model analysis.
Andersson AMC. has analyzed single cell time trace data.
Tomanek I. and Grah R. have analyzed and interpreted the data.

2.1 Abstract

Organisms cope with change by employing transcriptional regulators. However, when faced with rare environments, the evolution of transcriptional regulators and their promoters may be too slow. We ask whether the intrinsic instability of gene duplication and amplification provides a generic alternative to canonical gene regulation. By real-time monitoring of gene copy number mutations in E. coli, we show that gene duplications and amplifications enable adaptation to fluctuating environments by rapidly generating copy number, and hence expression level, polymorphism. This 'amplification-mediated gene expression tuning' occurs on timescales similar to canonical gene regulation and can deal with rapid environmental changes. Mathematical modeling shows that amplifications also tune gene expression in stochastic environments where transcription factor-based schemes are hard to evolve or maintain. The fleeting nature of gene amplifications gives rise to a generic population-level mechanism that relies on genetic heterogeneity to rapidly tune expression of any gene, without leaving any genomic signature.

2.2 Introduction

Natural environments change periodically or stochastically with frequent or very rare fluctuations and life crucially depends on the ability to respond to such changes. Gene regulatory networks have evolved into an elaborate mechanism for such adjustments as populations were repeatedly required to cope with specific environmental changes (Savageau, 1974; Moxon *et al.*, 1994; Gerland and Hwa, 2009). Gene regulation requires many dedicated components – transcription factors and promoter sequences on the DNA –

for information processing to occur. However, due to low single base-pair mutation rates, complex promoters cannot easily evolve on ecological time scales (Berg, Willmann and Lässig, 2004; Tuğrul *et al.*, 2015).

Gene copy number mutations might provide a fundamentally different adaptation strategy, which neither depends on existing regulation nor requires regulation to evolve. Gene duplications arise by homologous or illegitimate recombination between sister-chromosomes. Depending on the genomic locus, duplication rates (k_{dup}) can vary between 10^{-6} and 10^{-2} per cell per generation in bacteria (Anderson and Roth, 1981; Mats E. Pettersson *et al.*, 2009; Reams *et al.*, 2010; Sun *et al.*, 2012). This means that a typical bacterial population will contain at any given time a large fraction of cells with a duplication somewhere on the chromosome (Segall, Mahan and Roth, 1988; Sun *et al.*, 2012). Due to the long stretches of homology, duplications are highly unstable: at rates (k_{rec}) between 10^{-3} and 10^{-1} per cell per generation (Mats E. Pettersson *et al.*, 2009; Reams *et al.*, 2010) *recA*-dependent unequal crossover of the repeated sequence leads to deletion of the second copy – restoring the ancestral state – or to further amplification (Fig. 1a). If a gene is under selection for increased expression (Albertson, 2006; Bass and Field, 2011a; Nicoloff *et al.*, 2019), the process of gene duplication and amplification (GDA) can dramatically increase organismal fitness by increasing gene copy numbers. Due to their high rates of formation, amplifications provide fast adaptation and facilitate the evolution of functional innovation (Andersson and Hughes, 2009). In contrast, their high rate of loss makes amplifications transient and difficult to study (Andersson and Hughes, 2009). Surprisingly, until recently it has not been appreciated how this high loss rate impacts the distribution of copy numbers and associated expression levels in the population, a phenomenon causing antibiotic heteroresistance (Hjort, Nicoloff and Andersson, 2016; Nicoloff *et al.*, 2019). Moreover, amplifications have been studied only under constant selection for increased expression (Elde *et al.*, 2012; Näsvall *et al.*, 2012), while natural environments are rarely ever constant. While a large body of work suggests that phenotypic heterogeneity serves as an adaptation to fluctuating environments (Kussell and Laibler, 2005; Veening, Smits and Kuipers, 2008), it is not known how the genetic heterogeneity resulting from copy number mutations impacts survival in fluctuating environments.

Here, we ask whether the intrinsic genetic instability of gene amplifications allows bacterial populations to tune gene expression in the absence of evolved regulatory systems. To test this idea experimentally we devised a system of fluctuating environmental selection, which selects for the regulation of a model gene. In this fluctuating environment, we track, in real time, copy number mutations in populations as well as single cells of Escherichia coli. Using this system, we test the ability of GDA to effectively tune gene expression levels on ecological timescales, when environmental perturbations occur at rates far too fast for transcriptional gene regulation to emerge de novo.

2.3 Results

2.3.1 Amplification-mediated gene expression tuning (AMGET) occurs in fluctuating environments

To test whether GDA can act as a form of gene regulation at the population level, we experimentally introduced environmental fluctuations, such that a given level of expression of a model gene is advantageous in one, but detrimental in another environment. As the model gene, we used the dual selection marker *galK*, encoding galactokinase. Expression of *galK* is necessary for growth on galactose, but deleterious in the presence of its chemical analogue, 2-deoxy-galactose (DOG)(Barkan, Stallings and Glickman, 2011). Using *galK* with an arabinose-inducible promoter, we mapped the relationship between *galK* expression level and growth in (i) galactose, which selects for high *galK* expression levels and which we refer to as the 'high expression environment'; and in (ii) DOG, which selects for low *galK* expression and which we refer to as the 'low expression environment' (Fig. 2.1b). In order to establish a strong selective tradeoff between high and low expression, we used 0.1 % galactose for the high expression environment and 0.0001% DOG for the low expression environment in all experiments.

We then constructed a reporter gene cassette to monitor expression and copy number changes of *galK* (Fig. 2.1c) based on a previously described construct (Steinrueck and Guet, 2017). In this construct, *galK* is not expressed from a promoter but harbors p0, a randomized 188 bp nucleotide sequence matching the average GC content of E. coli instead (Steinrueck and Guet, 2017). This allowed for the selection of increased expression of *galK*. The reporter cassette harbors two fluorophores that allowed us to distinguish the two principal ways of increasing *galK* expression in evolving populations: promoter mutations and copy number mutations (Fig 1c). The promoterless *galK* gene is transcriptionally fused to a yellow fluorescence protein (*yfp*) gene, which reports on *galK* expression. Directly downstream, but separated by a strong terminator sequence, an independently transcribed cyan fluorescence protein (*cfp*) gene provides an estimate of the copy number of the whole cassette (Fig. S1a). We inserted this cassette into the bacterial chromosome, close to the origin of replication (*oriC*) – a location with an intermediate tendency for GDA (Steinrueck and Guet, 2017). However, our results also hold for a second locus, which is flanked by two identical insertion sequence (IS) elements and has a much higher tendency for GDA(Steinrueck and Guet, 2017) (Fig. S4).

The ancestral strain carrying the promoterless *galK* construct does not visibly grow in the high expression environment. After one week of cultivation at 37°C, mutants with increased *galK* expression appeared (Fig. S1b). We randomly selected one evolved clone with increased CFP fluorescence ('the amplified strain') and analyzed it in detail (see methods) to confirm its amplification. This amplified strain was then used for further experiments in alternating environments (Fig. 2.2a-c).

In all three alternating regimes, which change on a daily timescale, mean CFP levels of 60 replicate populations of the amplified strain tracked the environments for the full duration of the experiments. The adaptive change in *galK* copy number (Fig. 2.2b) occurred within the imposed ecological timescale, rapidly enough to maintain population growth given the daily dilution bottleneck under all three alternating selection regimes (Fig. S3a). We confirmed the observed changes in copy number using whole genome sequencing (Fig S2b). To understand these population-level observations, we monitored changes in expression of *galK* and *cfp* at the single cell level for two consecutive environmental switches (Fig. 2.2c). Expression of *galK-yfp* (Fig. S3b) was tightly correlated with the observed changes in gene copy number (Fig. S3c), indicating that gene expression was effectively tuned by GDA. We refer to this phenomenon as amplification-mediated gene expression tuning (AMGET).

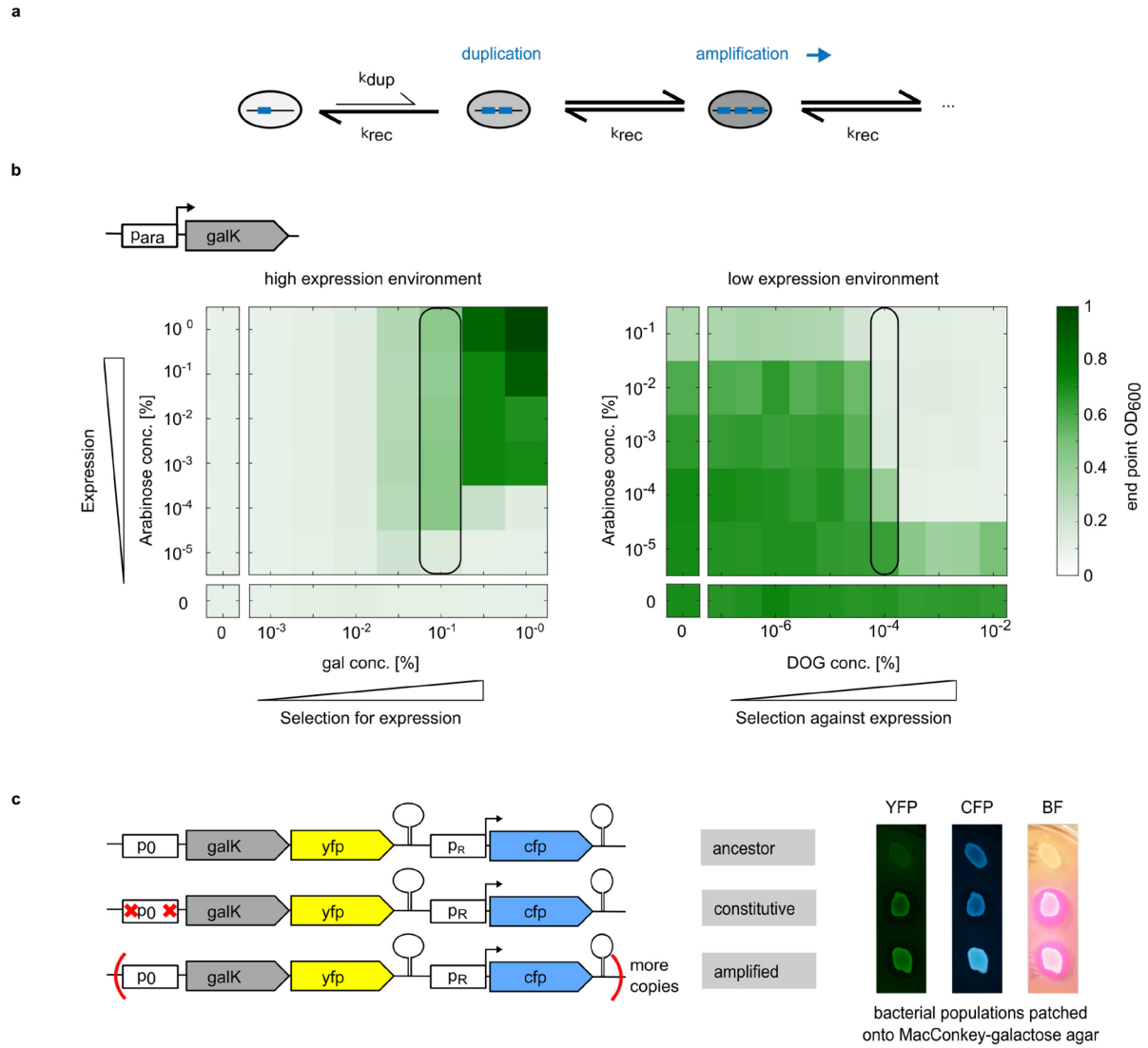


Figure 2.1 . ***An experimental system for monitoring gene copy number under fluctuating selection in real time.***

a*, Gene duplication and amplification (GDA). Genomic loci duplicate at rate (k_{dup}) 10^{-6} - 10^{-2} per cell per generation. The two gene copies oriented in tandem provide long stretches of identical sequence allowing for homologous recombination at rate (k_{rec}) 10^{-4} -*

10^{-1} per cell per generation with recA-dependent unequal crossover leading to further duplication (amplification) or deletion. Grey shading of cells symbolizes the amount of gene product made: increases in copy number result in increased gene expression.
***b**, top panel: Schematic of chromosomal cassette used. Expression of the selection marker, galK, is driven by an arabinose-inducible promoter (para). bottom panel: Growth (as measured by end point OD_{600}) in a 2D gradient of arabinose with galactose (high expression environment) or DOG (low expression environment), respectively. Boxes mark concentrations of 0.1% galactose and 0.0001% DOG, which result in a strong selective tradeoff between high and low expression and were used for further experiments. **c**, Schematic showing galK reporter cassette (p0 = random sequence/'non-promoter', pR = strong constitutive lambda promoter, terminator sequences downstream of yfp and cfp, respectively) and genetic changes of strains evolved in the high expression environment with resulting phenotypes on MacConkey galactose agar. Both evolved strains show increased galK-yfp expression over the ancestral strain (YFP) and the ability to grow on galactose (BF = bright field image, white versus pink colonies). The amplified strain shows increased CFP fluorescence (CFP) over the ancestral and the constitutive strain, indicating a gene copy number increase.*

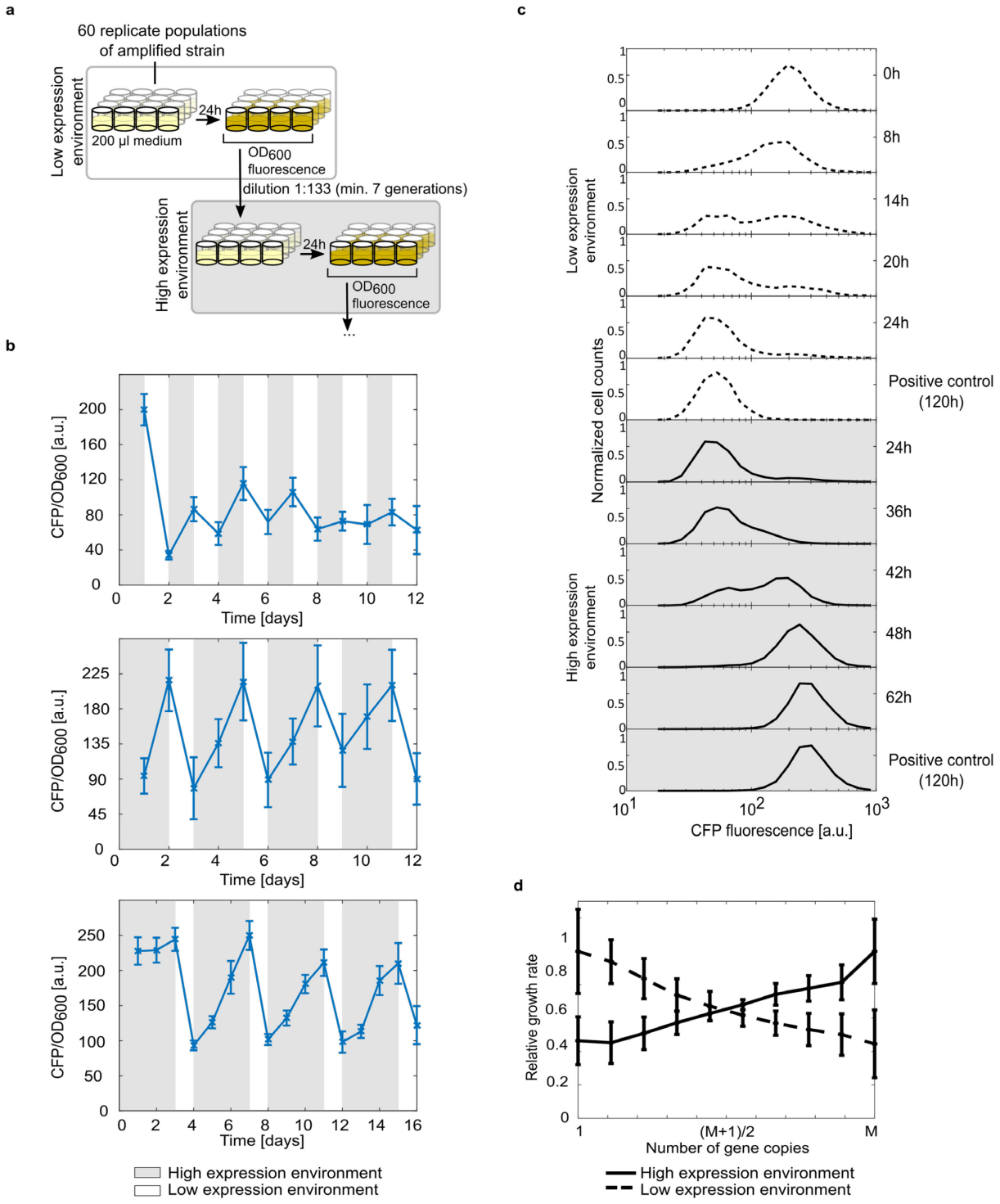


Figure 2.2 ***Amplification-mediated gene expression tuning (AMGET) occurs in fluctuating environments.***

a*, Experimental design of alternating selection in 96-well plate batch cultures, with a daily dilution of 1:133. A minimal duration of 24h per environmental condition (no shading = low expression environment, grey shading = high expression environment) allows measuring* OD_{600} *and fluorescence in populations that have reached stationary phase after dividing at least seven times after their last dilution.* ***b****, Alternating selection of 1 day - 1 day, 2 days – 1 day and 3 days – 1 day in high and low expression environment, respectively. Normalized*

CFP fluorescence as proxy for gene copy number of 60, 48 and 60 populations of the amplified strain. Error bars represent standard deviation (SD) over all populations. ***c****, Flow cytometry histograms (one of six replicates from two independent experiments; see d. for an overview of the full dataset) following the adaptation of an amplified bacterial population to low and high expression environments. Positive controls represent populations grown in respective environment for 5 days****.***
d*, Fitness as a function of copy number in the two environments. Growth rates relative to those of maximally adapted populations (positive controls in* ***c****) as a proxy for fitness were calculated from the population´s shift in CFP fluorescence over time (see Methods). M denotes the maximum copy number, which we estimate to be approximately 10 (see bulk measurements of M in Fig. S1a and Fig. S2a, and single cell-based measurements in Fig. S5b). Note that results do not depend on the precise value of M). Error bars represent the standard deviation of six replicates from two independent experiments.*

2.3.2 AMGET depends on selection acting on a gene copy number polymorphism

The rapid population dynamics observed during environmental switches (Fig. 2.2c) might simply be explained by selection acting on gene copy numbers with different fitness (Fig. 2.2d; Supplementary Note). We therefore hypothesized that AMGET occurs because of the intrinsic genetic instability of gene amplifications, which continuously and rapidly generate copy number polymorphisms that selection can act on. Re-streaking a single bacterial colony of the amplified strain resulted in colonies with different CFP levels, sometimes with sectors of different CFP expression levels within individual colonies (Fig. 2.3a), demonstrating the intrinsic genetic instability of the amplification. Importantly, this genetic instability is dependent on homologous recombination, as a *ΔrecA* derivative of the amplified strain failed to show a decrease in CFP fluorescence (and thus copy number) in response to increasing concentrations of DOG (Fig. S3d). Similarly, *ΔrecA* populations were not able to track fluctuating environments as their *recA wild-type* counterparts did (Fig. S3e).

To determine the rate at which copy number polymorphisms are generated in an amplified population, we followed individual bacteria over ~40 generations in a mother-machine microfluidic device (Wang *et al.*, 2010; Bergmiller *et al.*, 2017) and monitored their CFP levels. Mutations in copy number were clearly visible as changes in CFP fluorescence of the mother cell. In approximately 35% of cases, these changes were accompanied by a reciprocal fold-change of fluorescence in the daughter cell (Fig. 2.3b, Table S1) as expected from unequal crossover (Reams and Roth, 2015).

In order to quantify the combined rate of copy number gain and loss events by homologous recombination, we analyzed the fluorescence time trace of 1089 mother cells. 55% of traces exhibit constant levels of CFP fluorescence (Fig. 2.3c – panel 1) indicating stable inheritance of copy number. In about 7% of traces, the constant level of CFP is interrupted by a sudden

decrease or increase (Fig. 2.3c – panel 2-3). The corresponding fold-changes of fluorescence are consistent with gains or losses of entire copies of *cfp*. We estimated the lower bound for the average number of copy number mutations, k_{rec}, to be 2.7×10^{-3} per cell per generation, by automatically selecting only clear step-wise transitions in fluorescence, which are indicative of single copy-number mutation events (Methods, Fig. S5, Table S1). Interestingly, 34% of all traces (Fig. S5c) exhibit more complex behaviors (Fig. 2.3c – panel 4) and cannot be explained in terms of single step transitions.

Complex traces are expected to contain more than one duplication or deletion event even under the expectation that copy number variations are independent events (Fig. S5d). In addition, it is conceivable that copy number mutations are not independent, i.e., an increased probability exists for a second mutation after the first copy number increase occurred. However, we cannot exclude the possibility that most of the complex traces are due to expression noise (variance around the mean expression) of one or both fluorophores, especially since CFP expression variance increases with copy number. Moreover, microfluidics experiments showed transient growth defects visible as filamentation (Table S1). Given that the amplification includes the origin of replication (*oriC*), complex traces might in part result from replication issues. Transiently stalled replication forks could result in an overproduction of CFP relative to mCherry, which is located at phage attachment site attP21, almost opposite on the *E.coli* chromosome. Thus using only single clear step-wise transitions provides a very conservative lower bound for the rate of copy number mutations.

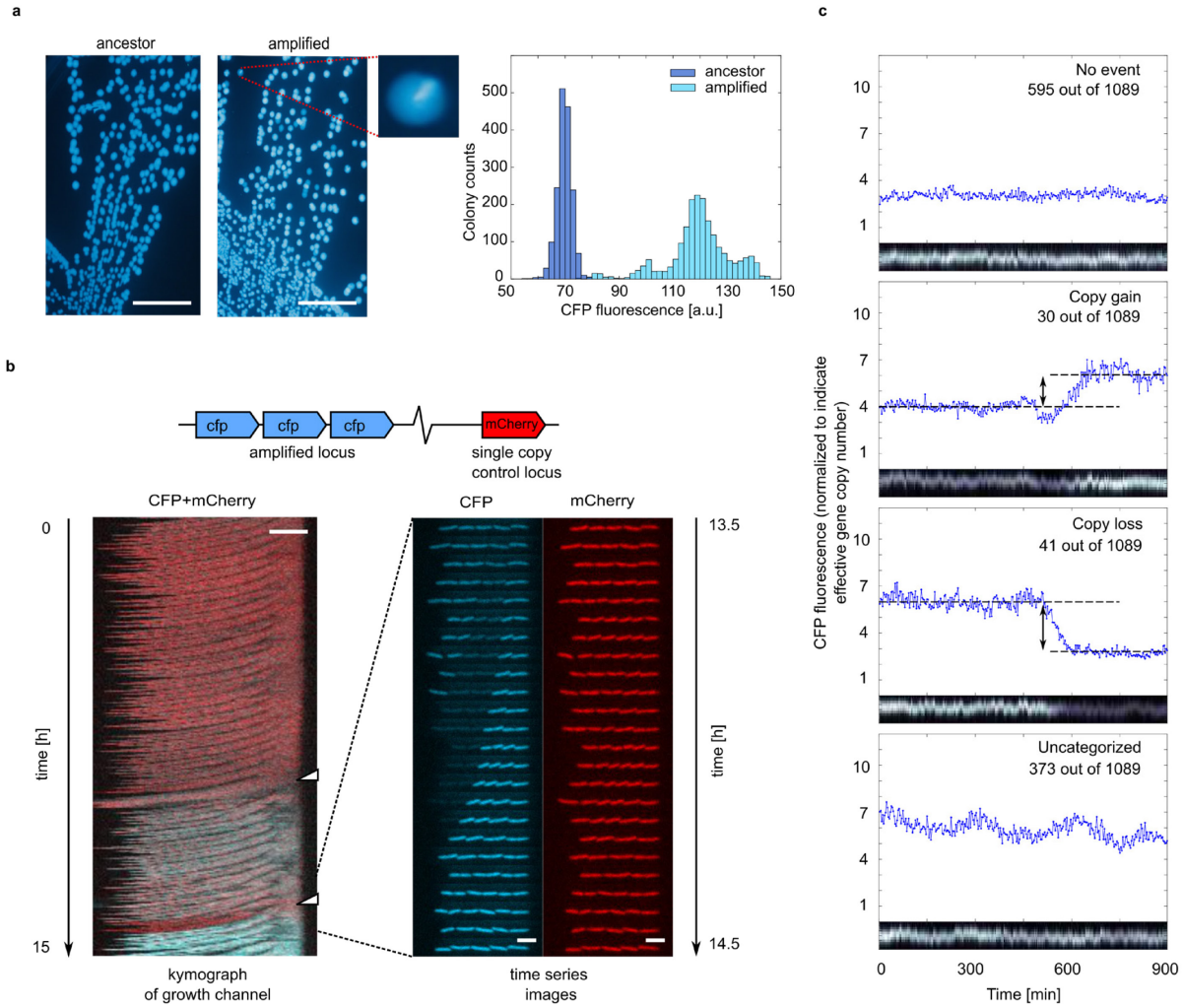


Figure 2.3 ***High-frequency deletion/duplication events in the amplified locus create gene copy number polymorphism in populations.***

a*, Re-streaks of a single bacterial colony on nonselective agar. Ancestral strain bearing a single copy of cfp (left), amplified strain (middle) colonies display sectors of different CFP fluorescence (inset). Scale bars, 10 mm. Histogram of single-colony mean CFP intensities obtained by resuspending and diluting five ancestral and amplified colonies, respectively (right).* ***b****, The amplified strain carrying a single copy of mCherry in a control locus (top) was grown in a microfluidics device to allow tracking of cell lineages in the absence of selection. Overlay of kymographs of CFP and mCherry fluorescence for one microfluidics growth channel (left). Two recombination events are visible as pronounced changes in CFP relative to mCherry fluorescence (white arrows). Time series images of CFP and mCherry fluorescence (right) of the same channel during the second amplification event. An increase in CFP fluorescence of the mother cell (rightmost position in the growth channel) occurs concomitantly with reciprocal loss of CFP fluorescence in its first daughter cell. As mother and daughter cell divide again, their altered level of CFP fluorescence is inherited by their respective daughter cells. mCherry fluorescence of the control locus stays constant during the recombination event. Scale bars, 5µm.* ***c****, Examples of single-cell time traces (kymographs and CFP fluorescence sampled from the mother cell) for four representative behaviors: constant expression, stepwise increase and decrease in expression, and complex expression*

changes. Frequencies of each behavior across 1089 channels from three independent experiments are shown in figure panels.

2.3.3 AMGET requires continual generation of gene copy number polymorphisms

Because the mechanism behind AMGET is selection acting on copy number polymorphism, we asked whether it differs from selection acting on single nucleotide polymorphisms (SNPs). To do so, we artificially created a polymorphic population comprised of an equal ratio of two strains — the ancestral strain with no detectable *galK-yfp* expression and a strain with two SNPs in p0 (Fig. 2.1c) resulting in constitutive expression of *galK* (Fig. 2.4a). Importantly, this 'co-culture' contained standing variation in *galK* expression, but because it is not due to amplification, variation is not replenished at high rates. While the 'co-culture' population tracked short-term environmental fluctuations in a manner similar to the amplified population (Fig. 2.4b), the long-term dynamics of the two populations were crucially different. Despite being grown from a single cell, the amplified population was able to respond to environmental change rapidly after being maintained in a constant high expression environment for increasingly longer periods (Fig. 2.4c). The 'co-culture' population, in stark contrast, progressively lost the ability to respond to sudden environmental change (Fig. 2.4d). While standing variation in the 'co-culture' provided some ability for a population to adapt in the short run, it is only replenished at the rate of point mutations. Hence, this variation – as well as the ability to adapt - is depleted by prolonged selection as the genotype with higher fitness goes to fixation in the population.

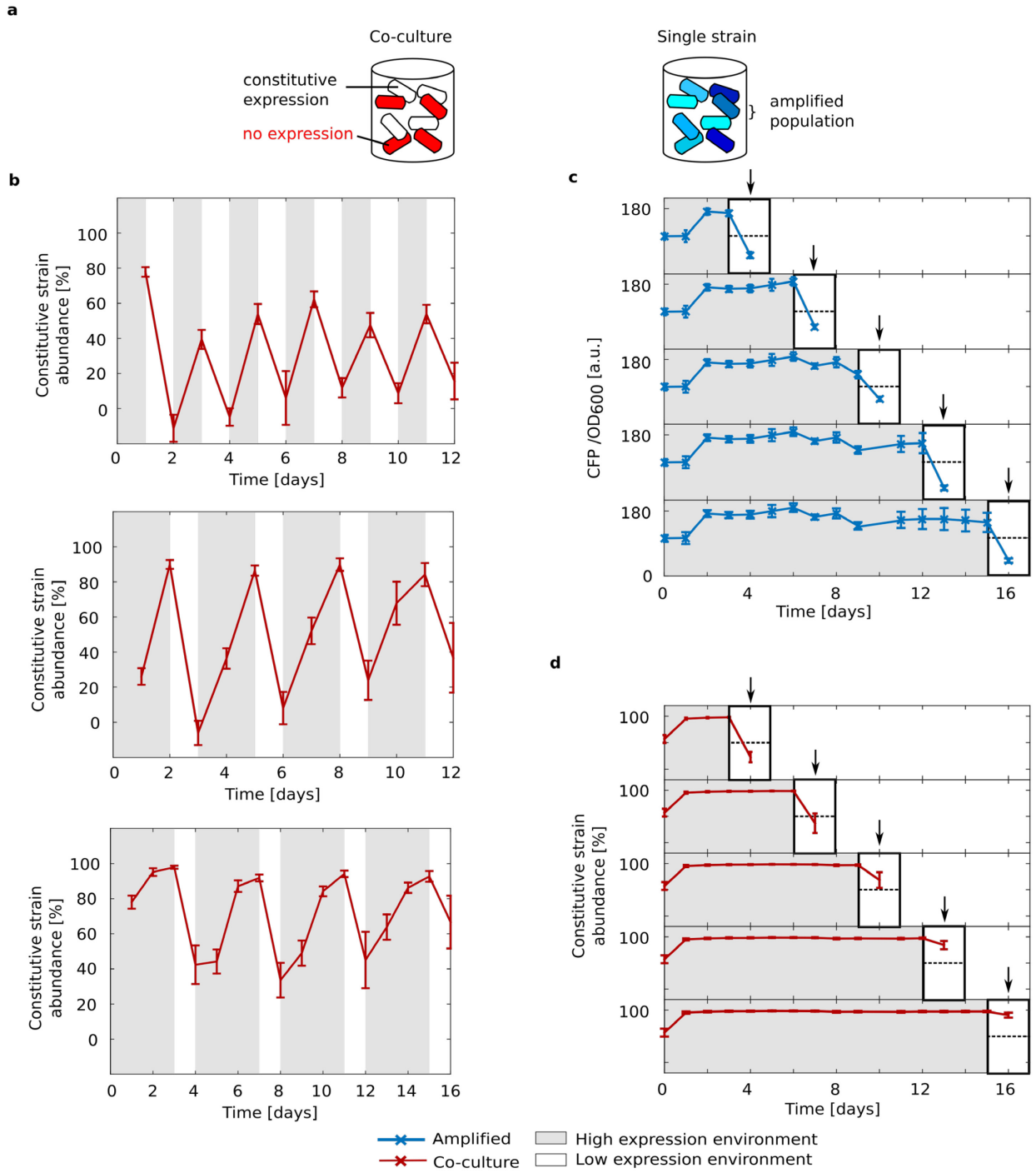


Figure 2.4 ***AMGET requires continual generation of gene copy number polymorphisms****.*
a*, Schematic of a co-culture composed of the ancestral strain without galK expression and a strain with two SNPs in* p_0 *(Fig. S1C) resulting in high galK expression (left). Fluorescently labeling the ancestor allows monitoring relative strain abundance (Methods). A population consisting of a single amplified strain (right) contains cells with different galK copy numbers and, accordingly, expression levels.* ***b****, Alternating selection following the scheme 1 day - 1 day, 2 days – 1 day and 3 days – 1 day in high and low expression environment, respectively. Constitutive strain abundance of 18 co-culture populations tracks environments, with the non-expressing strain being abundant in the low expression environment and the constitutive strain being abundant in the high expression environment. Error bars represent*

the SD of 18 replicates. ***c-d****, To estimate a population´s ability to respond to a change in the environment, periods of increasing length spent in the high expression environment are followed by one day in the low expression environment.* ***c****, Copy number of amplified populations as measured by CFP fluorescence is adjusted to the low expression environment (black arrows) even after prolonged growth in the high expression environment.* ***d****, In contrast, response of the co-culture to the low expression environment after prolonged growth in the high expression environment decreases with time spent in the high expression environment. The mean response on day 16 (1.11 for co-culture, 4 for amplified) differs significantly ($p<10^{-3}$, two-sided t-test) between populations of co-culture (**d**) and amplified (**c**) (see Methods). Error bars represent the SD of 36 replicates.*

2.3.4 AMGET is a general and robust mechanism

The experimental results have qualitatively shown that both, gene copy number polymorphism and selection acting on it, are necessary for AMGET to occur. Using population genetics theory, we developed a generic mathematical model to quantitatively predict the observed experimentally observed population dynamics (Fig. 2.2b). The model describes how gene copy number changes over time in a population under selection. Each copy number is treated as a distinct state, and these states differ with respect to growth rates in each of the two environments. Duplication and amplification events are the only source of transition between states. Importantly, all model parameters (the strength of selection and the rate at which the copy-number polymorphism is introduced as shown in Fig.1a) are obtained from independent measurements (Table S2). Thus, without specifically fitting any parameters, the generic model fully captured the experimentally observed dynamics of AMGET (Fig. 2.5a, Fig. S6a). The good fit between model and experimental data meant that we could use the model to expand the understanding of the basic conditions under which AMGET can act as an efficient de facto mechanism of population-level gene regulation.

Qualitatively, the model revealed that for a population to respond to environmental change at all, two conditions must be met: (i) constant introduction of gene copy number variation (i.e. non-zero duplication/recombination rate), and (ii) selection acting on it. If either of these are not present, the population is not able to maintain any long-term response to environmental change.

In order to more quantitatively examine the environmental conditions under which a population can respond to environmental change through AMGET, we defined the response R as the maximum fold change in gene expression before and after an environmental change.

We used the model to expand the range of environmental durations beyond those tested in experiment. In periodic environments, we find a sharp, switch-like transition from no response to full response for environments that switch typically on a day or longer timescale (Fig. 2.5b). In stochastically fluctuating environments, the transition is more gradual (Fig.

2.5c), yet no less effective. Furthermore, AMGET maintains its efficiency to tune gene expression in bacterial populations over order-of-magnitude variations in the duplication and recombination rates, as well as for any fitness cost of expression (Fig. S7).

2.3.5 AMGET tunes gene expression levels when transcription factor-based schemes are hard to evolve or maintain

Canonical gene regulation is unlikely to evolve or be maintained when a population is exposed to an almost constant environment that is sporadically interrupted by a rare environmental perturbation (Gerland and Hwa, 2009). We tested if AMGET might provide a generic mechanism of regulating expression under such conditions, by asking how long a population that is fully adapted to one environment needs for responding to a step-like environmental change (Fig. 2.5b top and side part of heat map; Fig. S6b). Our model results showed very rapid responses to step-like environmental changes on the order of one to six days, for all biologically relevant parameter values of amplification and duplication rates, as well as fitness cost of expression (Fig. 2.5d; Fig. S6c-e). AMGET is also a viable mechanism for practically any population size, especially for typical bacterial ones, although its efficiency drops for small populations (Fig. S6f). Therefore, AMGET efficiently tunes gene expression levels across a wide range of environments where transcription factor-mediated regulation would take prohibitively long to evolve (Berg, Willmann and Lässig, 2004; Tuğrul *et al.*, 2015).

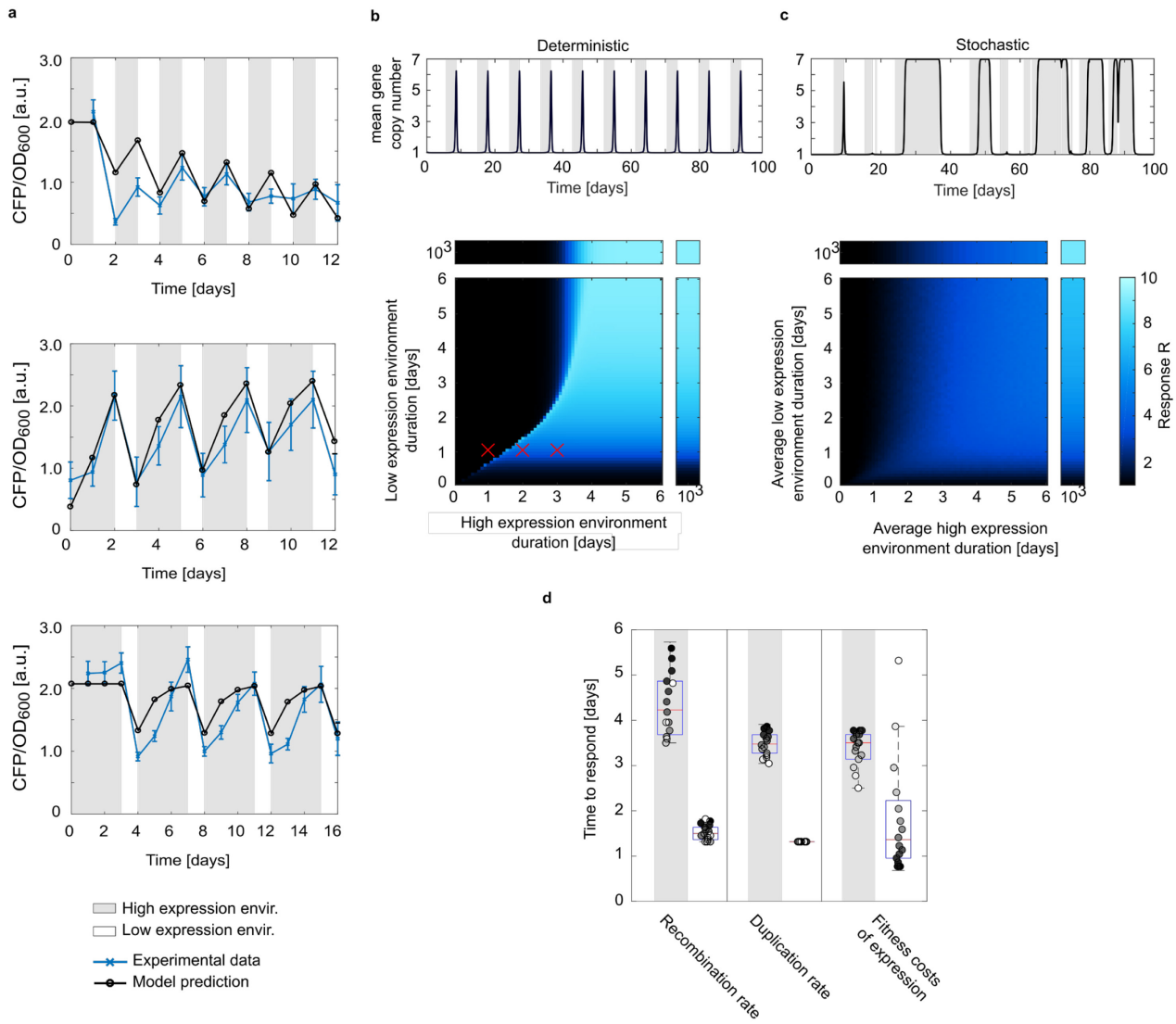


Figure 2.5 ***AMGET is a robust strategy for population level gene expression tuning across a range of environments****.*

a*, Comparison of model predictions (with all parameters derived from independent calibration experiments; see Methods) and experimental data for three different environmental durations. Pearson correlation between data and model: 0.72 (top), 0.92 (middle), 0.87 (bottom). See Fig. S6a for parameter sensitivity. Error bars represent standard deviation (SD) over of 60, 48 and 60 bacterial populations, respectively.* ***b-c****, Top: example of gene expression time trace for deterministic (****b****) and stochastic (****c****) environment durations. Bottom: response R (maximum expression fold change before and after the environmental change), shown in color, as a function of the two environment durations. Red crosses in* ***b*** *mark environments shown in* ***a****. The gradual increase in response in* ***c*** *occurs because of averaging across responses, which are deterministic for each individual environmental transition (****c*** *top).* ***d****, Variation of response time when uniformly sampling sets of parameters (black circles) in the range of 10^{-4} - $5x10^{-2}$, 10^{-5} - 10^{-3}, and 0.1 - 1 for recombination rate, duplication rate, and fitness costs of expression, respectively (Fig. S6c-e). The plot shows the median (red line) with the 25^{th} and 75^{th} percentile (blue box). In all plots, when not varied, we use recombination and duplication rates $k_{rec}^{0}=1.34x10^{-2}$ and $k_{dup}=10^{-4}$, respectively. All rates have units of cell^{-1} generation^{-1}. In our setup, one-day timescale is equivalent to*

between 10 and 23 generations (lower and upper bound, respectively; the bounds are estimated from the minimum and maximum growth rate of the least and best adapted copy number types, Table S2, Fig. 2.2d).

2.4 Discussion

Biology often relies on messy solutions, be it due to physical limitations or because evolution proceeds by opportunistic tinkering (Jacob, 1977; Tawfik, 2010). For organisms living in constantly fluctuating environments even the crudest form of gene regulation (Troein *et al.*, 2007) or gene expression heterogeneity (Wolf, Silander and van Nimwegen, 2015) increases fitness compared to not having any regulation at all. Here, we showed that the intrinsic instability of gene amplifications rapidly tunes gene expression levels when gene regulation is required but no other molecular regulatory mechanism is in place.

Despite resembling canonical gene regulation when observing populations as a whole (Fig. 2.2b), AMGET does not allow all single cells to change their gene expression concurrently. Instead, only a fraction of the population grows after the environment changes (Table 1). Thus, AMGET may effectively work by allowing bacterial populations to 'hedge their bets' for expression levels that could be required in a future environment. Unlike traditional descriptions of bet-hedging, where genetically identical individuals show variability in their phenotypic states (Veening, Smits and Kuipers, 2008), AMGET populations differ in their genotype due to the intrinsic instability of gene amplifications, thus passing on the adaptive state with high probability. Moreover, bet-hedging is typically characterized by switching between a small number of alternative phenotypic states (Veening, Smits and Kuipers, 2008), while in an amplified locus, expression can adopt a graded response due to a wide range of copy numbers.

Table 2.1. ***Comparison of regulation, amplification, adaptation and bet-hedging strategies.***

	regulation	**amplification**	**adaptation (rewiring via point mutations)**	**bet-hedging strategies**
mechanism	hard-wired response of individual cells	mutation	mutation	phenotypic differences between genetically identical cells
rate ON	1	10^{-6} - 10^{-2} cell^{-1} gen.$^{-1}$ (Anderson and Roth, 1981; Mats E. Pettersson *et al.*, 2009; Reams *et al.*, 2010; Sun *et al.*, 2012)	10^{-9} bp^{-1} cell^{-1} gen.$^{-1}$ (Drake, 1991; Elez *et al.*, 2010)	>10^{-5} variants per total cells (Bayliss, 2009)
rate OFF	1	10^{-3} - 10^{-1} cell^{-1} gen.$^{-1}$ (Mats E. Pettersson *et al.*,	10^{-9} bp^{-1} cell^{-1} gen.$^{-1}$ (Drake, 1991; Elez *et al.*, 2010)	

		2009; Reams *et al.*, 2010)		
active sensing machinery required	yes	no	no	no
can substitute for regulation on ecological time scales	-	yes	no	yes
expression state genetically heritable	no	yes	yes	no
tuning (allows graded expression)	typically not	yes	yes, but very long timescales	typically not
High reversibility (rate OFF > rate ON)	yes	yes	no	yes
suitable for rare stresses	no	yes	probably not, due to slow reversibility	depends on cost and rate

Because AMGET enables rapid dynamics and at the same time graded responses, it can be thought of as a form of primitive gene expression regulation at the population level (Anderson and Roth, 1977). Mechanistically, AMGET bears no resemblance to canonical gene regulation, which employs sensory machinery to alter gene expression in the course of just a single generation. Yet, despite the mechanistic difference, AMGET operates on the time scales of days and thus closer to those of canonical gene regulation, compared to the process of transcriptional rewiring by point mutations, which occur several orders of magnitude less frequently (Table 2.1).

It is interesting to consider AMGET in the context of evolutionary rescue, which describes a scenario of adaptive evolutionary change that restores the growth of declining populations under challenging selective conditions (Bell, 2013; Carlson, Cunningham and Westley, 2014). Whether or not evolutionary rescue occurs depends heavily on the population size, standing genetic variation, the strength of selection and the rate at which it changes (Barton and Partridge, 2000; Bell, 2013; Uecker and Hermisson, 2016). In practice, evolutionary rescue (i.e. adaptation by natural selection) and phenotypic plasticity (adaptation by existing regulatory mechanisms) often result in similar population dynamics. In cases where the basis of adaptation (genetic versus phenotypic) cannot easily be determined, this similarity complicates the interpretation of results (Bell, 2013; Carlson, Cunningham and Westley, 2014; McDermott, 2019). AMGET, being caused by adaptive copy number mutations is an example for evolutionary rescue. However, given its reversible nature, it -in some sense- resembles phenotypic plasticity.

AMGET may be one of several ways by which populations can make use of variation in expression levels to rapidly adapt to environmental change thereby allowing evolutionary rescue. While point mutations occur at lower rates, regulatory rewiring can still be surprisingly fast (Taylor *et al.*, 2015), especially when there is pre-existing variation in the precise architecture of regulatory networks. Moreover, noise propagation within gene regulatory networks can create an abundance of different expression levels, which are – in

principle – tunable by selection (Wolf, Silander and van Nimwegen, 2015). However, as the results of our co-culture experiment (Fig. 2.4) show, pre-existing variation can be easily depleted from a population if under strong selection. While it was previously shown that variation can be maintained in the form of multiple plasmid copies (Rodriguez-Beltran *et al.*, 2018), our results highlight that multiple copies of a genomic region actively regenerate heterogeneity due to the high recombination rate. Due to this property, AMGET provides a means of tuning expression to rare environmental fluctuations, where canonical gene regulation cannot evolve or be maintained (Gerland and Hwa, 2009).

AMGET is fast in bacteria because their generation times are short and their population sizes are usually large. However, our model results show that AMGET is in principle applicable to any other organism, but would take much longer time in relatively small populations (Fig. S6f). A compelling example for the "up-regulation" of a gene on relatively short evolutionary time-scales is that of the salivary amylase in humans, where variation in AMY1 copy number correlates with dietary starch content of human populations (Perry *et al.*, 2007).

Because any genomic region can be potentially amplified, AMGET can act on essentially any bacterial gene, providing regulation when the promoter is lacking altogether or when the existing promoter is not adequately regulated (Latorre *et al.*, 2005; Gil *et al.*, 2006). For instance, horizontally transferred genes tend to be poorly regulated, as their integration into endogenous gene regulatory networks can take millions of years (Pál, Papp and Lercher, 2005; Lercher and Pál, 2008). At the same time, they are enriched in mobile genetic elements (Dobrindt *et al.*, 2004; Juhas *et al.*, 2009), providing repetitive sequences for duplication by homologous recombination (Pettersson *et al.*, 2005; Andersson and Hughes, 2009). Indeed, genes with a recent history of horizontal transfer are often amplified (Hooper and Berg, 2003; Gusev *et al.*, 2014; Eme *et al.*, 2017).

Similarly, gene amplifications can confer resistance to antibiotics and pesticides, but they are often accompanied by a fitness cost in the absence of the compound (Nguyen *et al.*, 1989). In fact, heteroresistance caused by copy number polymorphisms is much more prevalent than previously thought and can lead to antibiotic treatment failure (Nicoloff *et al.*, 2019). Repeated use of antibiotics or pesticides can therefore create alternating selection regimes (Gladman *et al.*, 2015), where AMGET might play an important, yet previously overlooked, role in bacterial adaptation.

In spite of their ubiquity, GDA has been underappreciated (Andersson and Hughes, 2009; Elliott, Cuff and Neidle, 2013). In principle, fixed amplifications can easily be detected in next generation sequence data by an increase in coverage and mismatches corresponding to the duplication junctions (Fig. S2, Methods). However, duplications revert to the single copy state at high rate without leaving any traces in the genome (Fig. S2a). This implies that populations have to be kept under selection prior to sequencing, a condition that may not typically be met, especially not for environmental isolates (Eydallin *et al.*, 2014). However, despite this challenge, there are many reports of cases where amplified genes have been detected in the sequences of environmental strains and were found associated with

adaptation to environmental conditions (Gil *et al.*, 2006; Gusev *et al.*, 2014; Greenblum, Carr and Borenstein, 2015).

The notion that GDA "might be thought of as a rather crude regulatory mechanism" (Anderson and Roth, 1977) is more than 40 years old. However, so far almost all experimental work has focused on the benefits of amplification in constant, stable environments, thereby selecting for increased expression only (Näsvall *et al.*, 2012; Dhar, Bergmiller and Wagner, 2014). Here, we demonstrated how flexible GDA is in rapidly altering gene expression levels of populations in response to a wide range of environmental fluctuations. AMGET is thus a critical, and a critically underappreciated, mechanism of bacterial survival.

2.5 Methods

2.5.1 Bacterial strain background construction

Except when noted otherwise, all changes to the *E.coli* chromosome were introduced by pSIM6-mediated recombineering (Datta, Costantino and Court, 2006). All recombinants were selected on either 25µg/ml kanamycin or 10µg/ml chloramphenicol, to ensure single-copy integration. All resistance markers introduced by recombineering were flipped by transforming plasmid pCP20 and streaking transformants on LB at the non-permissive temperature of 37°C (Datsenko and Wanner, 2000). We used strain MG1655 for all experiments, except for testing galactose and DOG concentrations (Fig.2.1c). For that purpose, we placed *galK* under control of the pBAD promoter and used strain BW27784, which allows relatively linear induction of the pBAD promoter over a 1000 fold range of arabinose concentration (Khlebnikov *et al.*, 2001). In both strain backgrounds the genes *galK*, *mglBAC* and *galP* were altered in order to allow galactose- and DOG-selection.

Endogenous *galK* was deleted by P1-transduction of *galK*::*kan* from the Keio-collection (Baba *et al.*, 2006). The *mglBAC* operon was deleted to avoid selective import of galactose but not DOG (Nagelkerke and Postma, 1978). To express *galP* for DOG to be imported in the absence of galactose, its endogenous promoter was replaced by constitutive promoter J23100 (Zhou *et al.*, 2017). For this, the fragment BBa_K292001 (available at the Registry of Biological Parts, http://parts.igem.org/Part:BBa_K292001) was cloned into pKD13 (Datsenko and Wanner, 2000) yielding plasmid pMS1 with FRT-*kan*-FRT upstream of J23100. The cassette FRT-kan-FRT-J23100 was used for recombineering.

2.5.2 Assembly of the chromosomal gene cassettes

The chromosomal reporter gene cassette used for experimental evolution (*p0-RBS-galK-RBS-yfp-pR-cfp*; Fig. 2.1c) was assembled on plasmid pMS6* using standard cloning techniques. Plasmid pMS6* is based on plasmid pMS7, which contains the 'evo-cassette' (*p0-RBS-tetA-yfp-pR-cfp*) (Steinrueck and Guet, 2017). To obtain pMS6* we replaced the translational fusion of *tetA-yfp* on pMS7 with *galK* from MG1655 in a transcriptional fusion

with *yfp venus*, originally derived from pZA21-*yfp* (Lutz and Bujard, 1997). In addition, XmaI and XhoI restriction sites were added directly upstream and downstream of p0 by two consecutive inverse PCRs.

The chromosomal gene cassette for testing galactose and DOG concentrations (*pBAD-galK,* Fig. 2.1b) was assembled on plasmid pIT07, which was obtained by cloning *galK-yfp* as well as a chloramphenicol resistance flanked by FRT sites from pMS6* into pBAD24 (Guzman *et al.*, 1995). Gene cassettes were integrated into chromosomal loci 1 and 2 (corresponding to locus D and E in (Steinrueck and Guet, 2017)) by recombineering (Datta, Costantino and Court, 2006) and checked by PCR with flanking primers and sequencing of the full-length construct.

2.5.3 Strain modification for microfluidics

The amplification of locus 1 was moved from the evolved strain IT028-EE1-D8 to the ancestral background (IT028) by P1 transduction to isolate it from the effect of other potential mutations in the evolved background, including a sticky phenotype, which clogged the microfluidic devices. In order to obtain a single copy control locus *pR-mCherry* from our lab collection was introduced into the phage 21 attachment site (*attP21*) by P1-transduction (Bergmiller *et al.*, 2017).

2.5.4 RecA deletion in amplified strain locus 1 (Fig. S3d,e)

RecA was deleted in the amplified strain by replacing it with the kanamycin cassette from pKD13 (Datsenko and Wanner, 2000). In order to maintain the amplified state, recombinants were selected on M9 0.1% galactose medium supplemented with 25µg/ml kanamycin and verified by sequencing.

2.5.5 Culture conditions

All experiments were conducted in M9 medium supplemented with 2 mM MgSO4, 0.1 mM CaCl2 and different carbon sources (all Sigma-Aldrich, St. Louis, Missouri). For evolution experiments 0.1% galactose (high expression environment) or 1% glycerol combined with 0.0001% 2-deoxy-d-galactose (DOG) (low expression environment), respectively, were added as carbon sources. For microfluidics experiments M9 medium was supplemented with 0.2% glucose and 1% casein hydrolysate and 0.01% Tween20 (Sigma-Aldrich, St. Louis, Missouri) was added as surfactant prior to filtering the medium (0.22 µm).

All bacterial cultures were grown at 37°C. Growth and fluorescence measurements in liquid cultures were performed in clear flat-bottom 96-well plates using a Biotek H1 platereader (Biotek, Vinooski, Vermont).

2.5.5.1 Mapping the relationship between *galK* expression level and growth

For the 2D gradients of arabinose and galactose or DOG (Fig. 2.1b), respectively, an overnight culture of the test-cassette strain was diluted 1:200 into 96-well plates containing 200 µl of M9 supplemented with carbon sources, DOG and the inducer arabinose, as indicated in Fig. 2.1b. Cultures were grown in the platereader with continuous orbital shaking.

2.5.5.2 Evolution experiments

For all evolution experiments (1. experimental evolution of the amplified strains in the high expression environment and 2. alternating selection experiments), cultures were grown in 200µl liquid medium in 96-well plates and shaken in a Titramax plateshaker (Heidolph, Schwabach, Germany, 750 rpm). Populations were transferred to fresh plates using a VP407 pinner (V&P SCIENTIFIC, INC., San Diego, California) resulting in a dilution of ~ 1:133.

2.5.5.2.1 Evolution of the amplified strains in the high expression environment

To obtain the amplified strains of locus 1 and 2, respectively, an overnight culture inoculated from a single colony of the ancestral strain carrying the reporter gene cassette in the respective loci (IT028; Fig. S1b-c) or 2 (IT030; Fig. S4b) was started in LB-medium. Cells were pelleted, washed twice and diluted 1:100 into M9 0.1% galactose (locus 1) or M9 0.1% galactose supplemented with 0.1% casamino acids (locus 2). For locus 1, the timing of each dilution into fresh medium (~1:133) was chosen such as to maximize the number of rescued populations and to minimize the amount of time spent in stationary phase for grown populations. The transfers happened at days 10, 13, 15, 17, 18 and 19 (Fig. S1c). The first signs of growth were detected in several wells only after approximately one week of cultivation in minimal galactose medium (Fig S1b). The evolving populations were monitored by spotting them onto MacConkey galactose agar in 128 x 86mm omnitray plates prior to transfer. For locus 2, the evolving populations were transferred daily (~1:133, corresponding to seven generations) and spotted on to LB plates supplemented with 0.5% charcoal (Fig. S4b) to improve fluorescence quantification. Colony fluorescence of all experiments was recorded using a custom-made macroscope set-up (https://openwetware.org/wiki/Macroscope)(Chait *et al.*, 2010). For the isolation of clones, evolved populations were streaked twice for purification on LB agar and grown in M9 galactose medium prior to freezing. For both locus 1 and 2, respectively, all further experiments were started from the original freezer stock of the amplified strain. This was done for two practical reasons: i) to save the time needed for duplications (and higher order amplifications) to evolve (one week in M9 galactose medium used for locus 1 and one day in M9 medium supplemented with casaminoacids used for locus 2), and more importantly, ii)

to allow interpretation and reproducibility of the fluorescence data of the alternating selection experiments. As the reporter gene cassette allows selecting for increased *galK* expression but not for amplification itself, it is necessary to screen mutants with increased *galK* expression for increased CFP fluorescence. During amplification the initial duplication step is rate-limiting and break-points differ between evolving populations. We therefore limited ourselves to two amplified strains (locus 1 and 2), which we analyzed in detail. Amplified populations were thus started from single colonies, which were grown non-selectively on LB (Lennox) agar by streaking the original freezer stock. Due to the high rate of recombination, any given streak of the original amplified freezer stock contains colonies with a single copy of *galK* (Fig. 2.3a, right panel). In order to pick only amplified colonies, we examined CFP fluorescence using the macroscope.

We characterized evolved amplified strains by Sanger sequencing of the p0 region, amplification junctions and the rho gene, which was found mutated in a previous study using the same locus (Steinrueck and Guet, 2017). For the strain amplified in locus 1 (IT028-EE1-D8), increased *galK* expression is achieved by increased *galK* copy number as evident from increased CFP fluorescence (Fig. 2.1c), as well as through a missense mutation in the termination factor rho (S265>A), allowing for baseline-expression via transcriptional read-through from the upstream rsmG into *galK* (Steinrueck and Guet, 2017). The amplified region spans 16 kb from atpB at the left replicore over the origin of replication to rbsD into the right replicore.
For the strain amplified in locus 2 (IT030-EE11-D4), *galK* expression comes solely from the increase in copy number (no mutations in p0 were detected). In this case, inverse PCR and sequencing confirmed that two identical IS elements (IS1B and IS1C) form the junction of the amplified segment(Steinrueck and Guet, 2017). Whole genome sequencing of both amplified strains confirmed amplification junctions and the rho mutation detected with PCR and Sanger sequencing and revealed two additional single nucleotide changes in the amplified strain locus 1 (coaA, pos. 4174770, C>T, resulting in R>H; wcaF, pos. 2128737, C>A, resulting in G>V).

2.5.5.2.2 Alternating selection experiments

For the experiments in Fig. 2.2b, a pre-culture of the amplified strain (IT028-EE1-D8) was grown in M9 0.1% galactose overnight, which was then inoculated 1:200 into the medium as indicated. For the experiment alternating two days in high and one day in low expression environment (Fig. 2.2b – middle panel), populations were first subjected to a scheme of daily alternating selection for six days prior to switching to the 2-1 scheme.

For the co-culture experiments (Fig. 2.4), a pre-culture of the amplified strain (IT028-EE1-D8) was grown in M9 0.1% galactose overnight. In parallel, the ancestral strain carrying a single silent copy of *galK* in locus 1 (IT028) and a strain constitutively expressing *galK* in locus 1

(IT028-H5r), were grown overnight in M9 1% glycerol and mixed in a 1:1 ratio. We labeled the ancestral strain by transduction of *attP21::pR-mCherry* (IT034). The constitutive strain was obtained by oligo-recombineering two point mutations into p0 of the ancestral strain and selecting recombinants on M9 0.1% galactose agar. These two point mutations (-29 A>T and -37 G>T) have initially evolved in parallel to the amplified strain and result in a similar level of *galK* expression (Fig. 2.1c).
To quantify the relative abundance of the two strains in the co-culture, we calculated the expression ratio of the two strains, using an exchange rate between CFP and mCherry units from the ancestral strain expressing both fluorophores (IT034).

2.5.6 Whole genome sequencing

We isolated gDNA from overnight cultures of single clones of i) the ancestral strains ii) the amplified strains after initial selection in the high expression environment (galactose) as well as iii) the amplified strains after overnight selection in the low expression environment (DOG), for Locus 1 and Locus 2, respectively. In all cases overnight cultures were inoculated from colonies grown non-selectively on LB agar. For the overnight culture M9 1% glycerol was used for the ancestral and DOG-selected clones, while M9 0.1% galactose was used for the galactose-selected clones. A whole genome library was prepared and sequenced by Microsynth AG (Balgach, Switzerland) on an Illumina Next.Seq (with a mean read length of 75 bp). Fastq files were assembled to the MG1655 genome (Genbank accession number U00096.3) using the Geneious alignment algorithm with default options of the software Geneious Prime version 2019.2.1. SNPs were analyzed using the variant finding tool of Genious.

2.5.7 Flow Cytometry

Three colonies of the amplified strain and the constitutive control strain, respectively, were inoculated into culture tubes with 2ml M9 0.1% galactose (high expression environment) and grown for three days with transfers every 24h. This population was inoculated into M9 + 1% glycerol + 0.0001% DOG (low expression environment). OD_{600} was monitored to assure continuous exponential growth by regular dilutions. Samples for flow cytometry were frozen at the indicated time points (Fig. 2.2c). After 24h in the low expression environment, the populations were transferred back to the high expression environment with dilution and sampling occurring in the same manner. In parallel, the positive controls were grown for five days in both selection environments, respectively, with transfers occurring every 24h. Fluorescence was measured using a BD FACSCanto™ II system (BD Biosciences, San Jose, CA) equipped with FACSDiva software. Fluorescence from the Pacific Blue channel (CFP) was collected through a 450/50nm band-pass filter using a 405nm laser. Fluorescence of the FITC channel (YFP) was collected through a 510/50 band-pass filter using a 488nm laser. The

bacterial population was gated on the FSC and SSC signal resulting in approximately 6000 events analyzed per sample, out of 10,000 recorded events.

2.5.8 Microfluidics experiments

For the microfluidics experiments, a single colony of the amplified strain was picked and grown overnight in nonselective LB (Lennox) medium.

Microfluidics devices were prepared as described previously (Bergmiller *et al.*, 2017). Briefly, devices had dimensions 23 µm × 1.3 µm × 1.3 µm (l, w, h) for the growth channels with 5 µm spacing along a trench for growth medium. Devices were fabricated by curing degassed polydimethylsiloxane (Sylgard 184, 1:10 catalyst:resin) inside epoxy replicate master molds produced from primary wafer-molded devices. Microscopy was performed on an inverted Nikon Ti-Eclipse microscope and with a previously described set-up (Bergmiller *et al.*, 2017). Per experiment, multiple positions of a single mother machine were imaged using a 60x 1.4 NA oil immersion objective lens. To image constitutive mCherry, the green LED (549+/-15nm) was used at a light intensity of 670µW and an exposure time of 170-200ms. To image CFP, the cyan LED (475+/-28nm) at a light intensity of 270µW and an exposure time of 90-100ms was used.

2.5.9 Analysis of microfluidics data

The mother machine allowed tracing of mother cells for ~ 38 divisions, thereby following the fate of arising copy number mutations in the absence of selection. In three experiments, we analyzed 336, 369 and 384 mother cell lineages, respectively, equaling a total of approximately 40.000 cell divisions (with a division time of 23.6 (+- 1.5) min as determined by counting septation lines in growth channel kymographs).

Microfluidics data analysis was based on mother cell time traces (Fig. 2. 4c). To this end, we used Fiji/ImageJ to create kymographs, by laying a line through the middle of mother cells perpendicular to the growth channel using the built-in Multi-Kymograph tool with a pixel width of 9. Kymographs of CFP and mCherry were then analyzed using MATLAB.

2.5.9.1 Determining what data to include

To minimize the influence of three unknown factors (maturation rate and bleaching of the two fluorophores, and the degree of bleedthrough between channels on the microfluidic chip), we were restrictive with the colonies we included.

1. We excluded all fluorescence changes that occured when the cells were dying. Only colonies (mother cell lineages) that continuously grew until the end of the experiment were included. Specifically, the last 10 frames of mean mCherry fluorescence of mother cells needed to exceed the background threshold (68%, 76%, 82% of total colonies included, respectively, for the three experiments).

2. Some colonies exhibited a large variation in growth rate, due to temporary slowdown and/or filamentation. In the kymographs this was seen as a large variance in the constitutive mCherry channel. We excluded colonies with a variance > 1.5 times the mCherry experiment-wide variance (thus including 96%, 96%, 96% of total colonies included for the three experiments, respectively).
3. In some cases there was significant bleedthrough between adjacent colonies. To avoid double counting transitions, the colony that was less bright was removed from the data set if two adjacent colonies had a correlation of 0.6 or higher (99%, 98%, 98% of total colonies included, respectively, for the three experiments).

For the identified colonies the maximum fluorescence value per time point was extracted for both, mCherry and CFP channels. These were plotted against each other and a rectangular area, bounded by a manually selected max and min for each channel was chosen such as to include all but extreme outliers (Fig. S5a). Accordingly, 99% of data points were included in all three experiments.

2.5.9.2 **Normalization**

To correct for slow temporal drift in the signal of CFP and mCherry, a time average over all colonies was taken and a 7th degree polynomial fitted. All time points were divided by the corresponding polynomial estimates.
Furthermore, mCherry fluorescence was flat-field corrected based on the expectation that mCherry is roughly constant across all colonies. To do so, a line was fitted to the coordinate to get an estimate of the backround of each location. The data was divided by the corresponding estimated value.

2.5.9.3 **Probability density function**

For the probability density function (PDF) in Fig. S5b we normalized for differential growth rate by dividing the CFP fluorescence by the constitutively expressed mCherry fluorescence. To reduce noise, a median filter (MATLAB medfilt1) was applied to the ratio of CFP and mCherry over 20 data points.
To get an estimate of the PDF of the CFP/mCherry single cell fluorescence, we used a kernel density estimation (KDE) (MATLAB function ksdensity). To estimate a proxy for copy numbers, we found points where the first and second derivative of the PDF is zero. These points were set as initial conditions for a pairwise fitting of peak mean and variance. All but the first and the last peak had two estimates for mean and variance. For the mean, the average of the two was taken and for the variance the smaller one was chosen. To assign boundaries for states, the estimated variance was halved. For plotting, the height of each peak was set to match the peak height. No weight was fitted. The mean inter-peak distance for each PDF was used as a proxy of copy numbers for plotting in Fig. 2.4c.

2.5.9.4 Estimation of nS2R2 for classification of single cell traces

We have classified the single cell traces using a normalized R squared, the proportion of variance explained, which we call nS2R2. In this adjustment, each element in both the residual and the total sum of squares is normalized by the predicted value:
nS2R2 = 1-Snormres/Snormtotal, where Snormres = ∑i (yi-fi)2/fi2, Snormres = ∑i (yi-y0)2/fi2, where yi, fi, and y0 represent measurements, fitted/predicted values, and mean of the measurements, respectively. This normalization takes into account that the intrinsic noise increases with expression and thus penalizes it less. Next, the algorithm fits one constant to the start and one constant to the end value of the CFP/mCherry trace, and reports this estimation parameter (nS2R2) based on which it classifies traces as shown in the pie charts of Fig. S5c. Clear transitions exhibit an nS2R2 score of >0.5 and were verified by eye analyzing microfluidics movies in detail (Table S1). The algorithm classifies no-events ("flat lines") if the nS2R2 score lies between 0 and 0.5. Traces, which cannot be classified unambiguously neither as clear transition nor as a clear no-event, i.e. with nS2R2 below 0, are classified as "complex traces". This occurs if the start and end of CFP/mCherry trace values are similar but vary significantly in between.

2.5.10 Quantitative PCR

For qPCR, DNA was isolated using Wizard Genomic DNA purification kit (Promega, Madison, Wisconsin) from 50 ul of frozen samples from different time points (1,4,9,10,11, gal 10, single copy control, DOG 8, DOG 10) of one flow cytometry experiment grown for 4-5 generations in LB. To quantify fluorescence, the same cultures were patched onto LB agar supplemented with 0.5% charcoal and imaged using the macroscope.
We performed qPCR using Promega qPCR 2x Mastermix (Promega, Madison, Wisconsin) and a C1000 instrument (Bio-Rad, Hercules, California). To quantify the copy number of samples of an evolving population, we designed one primer within *cfp* (target) and used one primer within rbsB as a close reference, which lies outside the amplified region. We compared the ratios of the target and the reference loci to the ratio of the same two loci in the single copy control. Using dilution series of one of the gDNA extracts as template, we calculated the efficiency of primer pairs to be 89.01% and 92.57%, for *cfp* and *rbsB*, respectively. We quantified the copy number of *cfp* in each sample employing the Pfaffl method, which takes amplification efficiency into account (Pfaffl, 2001). qPCR was done in three technical replicates.

2.5.11 Measurement of colony fluorescence (Fig. S1c, Fig. S4b, Fig. 2.3a)

Colonies were grown without selection and imaged using the macroscope set up.

To obtain mean colony CFP fluorescence intensity, a region of interest was determined using the ImageJ plugin 'Analyze Particles' (settings: 200px-infinity, 0.5-1.0 roundness) to identify colonies on 16-bit images with threshold adjusted according to the default value. The region of interest including all colonies was then used to measure intensity.

2.5.12 Mathematical model

A simple mathematical model recapitulates the change in *galK* copy number of the amplified population (Fig. 2.5a). Importantly, the parameters for the model were estimated purely from calibration measurements (growth rates, fitness in the two environments with respect to copy number (flow cytometry experiments), number of generations spent in each environment, and recombination rate, k_{rec}) and the literature (k_{dup}, (Andersson and Hughes, 2009)). Their values are listed in Table S2. No parameter was fit to reproduce the measurements in Fig. 2.5a.

The model describes the time evolution of a population where cells with different gene copy numbers are represented by distinct states. The duplication and amplification events are the only source of transition between states. The time evolution proceeds iteratively; with discrete times representing synchronous cell divisions in the population. The size of subpopulation N_j of cells with gene copy number j at time $t+1$ equals:

$$N_j(t+1) = \overbrace{\underbrace{(1-k_{rec}s_j)N_j(t)}_{\text{no duplication of amplification event}}}^{\text{daughter 1}} + \overbrace{\underbrace{(1-k_{rec}-k_{dup}\delta_{j,1})s_jN_j(t)}_{\text{no duplication of amplification event}}}^{\text{daughter 2}} + \underbrace{\sum_{k=2}^{M} k_{rec}P_{kj}s_kN_k(t)}_{\text{amplification event}} + \underbrace{k_{dup}s_1N_1(t)\delta_{j,2}}_{\text{duplication event}} \qquad (1)$$

where s_j is the relative growth rate of the subpopulation with j gene copies in the given environment (taken from Fig. 2.2d), δ_{jk} a Kronecker delta which equals 1 if $j=k$ and 0 otherwise. The equation for single and double gene copy numbers (j=1 or j=2, respectively) has an additional term to reflect duplication events. As we assume that the rate of recombination per copy is constant, the overall recombination is proportional to the number of gene copies k; $k_{rec}=k\ k_{rec}^{0}$ (Mats E. Pettersson *et al.*, 2009). P_{kj} represents the transition probabilities given an amplification event and is computed in the following way: assuming a homologous recombination between sister chromosomes occurs somewhere in the gene, we computed all possible combinations of how genes can be recombined to form different number of gene copies between the two daughter cells. P_{kj} then represents the probability that, given a recombination event, a daughter cell obtains j gene copies with its mother having k of them before the event. For example, starting with three gene copies, there is 22% probability to obtain four gene copies, or 22% probability to have one copy in the daughter (Fig. S6h). We have observed in microfluidics experiments that most (65%) copy number changes happen only in the mother cell while the daughter cell remains unchanged. Therefore, we do not model recombination as a reciprocal event.

Based on platereader bulk experiments, observations indicated an upper limit for the copy number a cell can have (see Supplementary Note 1). Thus, in our model, a cell can have up

to *M* gene copies; if that number is exceeded, the cell stops dividing. This upper limit for gene copy number was confirmed in microfluidics and qPCR experiments, indicating to be between 6 and 12. Our single cell analysis showed that *M*=10 is a good estimate (Fig. S5b, according to number of states in the probability density function, see Analysis of the microfluidics data). However, the results of the mathematical model do not depend on the precise value within the measured range, as all results remain qualitatively the same for any value in the range of 6 and 12. Fig. S6g shows that relative growth rates, obtained from flow cytometry experiments, are independent of *M*.

2.5.12.1 Measurements of model parameters (Table S2)

T1 & T2, generations per day in 96 well plates

In order to model the alternating selection experiment (Fig. 2.5a), we needed to know the maximal growth rate of the amplified strain (IT028-EE1-D8) in the high and low expression environments, respectively. Because the exact details of cultivation (such as culture volume, shaking speed and temperature fluctuations) strongly affected growth rate, we were unable to measure growth curves while keeping cultures under the conditions of the original experiment. Hence, we estimated growth rate indirectly without perturbing the experiment, by determining the maximal number of generations possible in 24h (number of generations = 24[h] * growth rate[1/h]/log(2)) from a dilution series experiment. Populations pre-adapted to the respective environment were grown to carrying capacity of the respective medium and diluted by a factor of approximately 2*n* (with *n* ranging between 7 and 28). We sought the maximal dilution that could still be compensated by growth (by requiring after 24h of growth the OD_{600} to reach the OD_{600} of the stationary phase). All dilutions of equal to or less than 1:222 and 1:223 were able to reach stationary phase in the high and low expression environment, respectively, yielding model parameters T1=22 and T2=23 for the maximal possible number of generations. Hence, by adding an upper bound on the number of generation the model implicitly captures the experimental reality of cells running out of nutrients.

T10 & T20, generations per day in culture tubes

Parameters T10 and T20 were necessary for obtaining the fitness landscape in Fig. 2.2d (and the resulting relative growth rates s_j). T10 and T20 generations per day, measured under the exact conditions of the flow cytometry experiment (Fig. 2.2c), namely exponential growth in culture tubes with 2ml volume of M9 0.1% galactose or M9 1% glycerol + 0.0001% DOG, respectively. We measured OD_{600} with a WPA Biowave spectrophotometer (Biochrom, UK).

Determining fitness landscape and relative growth rates s_j

The relative growth rates for each genotype (copy number state) in the high and low expression environments, respectively, were computed from flow cytometry time series experiments assuming exponential growth with no duplication/amplification event (k_{dup}=0,

k_{rec}=0). This is a valid approximation as long as the two rates are small enough, such that the population structure consists of all copy number types, i.e., that each subpopulation is much larger than the additional cells created by a single amplification or duplication event.
The flow cytometry measurements of the distribution of CFP expression at different times were split in *M* equal-width bins. The lowest and highest bins were set according to the equilibrium fluorescence distribution in DOG and galactose, respectively. For the lowest bin, we took the values of fluorescence <85, while for the high bin we took the mode fluorescence values of the measured distributions, corresponding to >160 for the first, and >245 for the second set of flow cytometry experiments. Each bin represents a given gene copy number. The distributions between different times were then compared using iterative exponential growth model:

$$N_j(t_2) = (1+s_j)^{(t_2-t_1)/t_{1/2}} N_j(t_1) \qquad (2)$$

where N_j is the population size with *j* gene copy number, $t_{1/2}$ is the timescale of exponential growth, t_1 and t_2 are two measurement times, and s_j represents the relative growth of cells with *j* gene copies. The population distributions for all time points were obtained from the flow cytometry data given the binning described above. Using this model, we obtained growth rates s_j for each pair of consecutive distributions at times t_i and t_{i+1} in the following way: given population distribution at time *i*, we predicted the new distribution given Eq. (2). We found such s_j values that minimize the Euclidian difference between the predicted and observed population distribution at time *i*+1. We repeated this for all pairs of consecutive distributions (at times t_i and t_{i+1}) and different replicates to obtain a set of solutions for s_j. Using this approach, we acquired only relative growth rates, which still allowed constants to be added to the growth rates. To tackle this, we added such constants to each growth rates in order to i) minimize the $\chi 2$ of the differences between each growth rate solution and the mean of all solutions, which optimally removes the replicate-to-replicate variability (error bars in Fig 2.2d) on the inferred relative growth rates but does not affect their mean value; and ii) force the average growth rate of the adapted state to be 1 (i.e., for *j*=1 in low expression environment and *j*=*M* is high expression environment, s_j=1) by adding a term to the χ2 error function of the form (adapted state expression - 1)^2. Fixing *s* to be 1 in a reference environment is a convention that mathematically will not affect any subsequent results.
The absolute maximal growth rates in the two environments were measured in populations grown in high and low expression environments for 120h, respectively. Thus, they represent the growth rates of populations with the highest and lowest possible copy number (Fig. 2.2c, positive controls). The estimated fitness values for both high expression environment (s_{jHEE}) and low expression environment (s_{jLEE}) can be found in Table S2.

Estimation of recombination rate k_{rec} from microfluidics data

We obtained a conservative estimate for the lower bound for the average number of copy number mutations from single step transitions in the pie charts (Fig. S5c). Out of 72 mother cell time traces classified as clear transition events, we verified 67 by detailed analysis of microscopy images (Table S1). We accordingly calculated the lower bound for the mutation rate as 67 events/1089 lineages/22.7 generations yielding $k_{rec} = 2.7*10^{-3}$ (+-$7.4*10^{-4}$) per cell per generation.

To estimate the mean recombination rate to be used in the model, two corrections have to be made: i) because our model assumes that the recombination rate is proportional to the number of gene copies (Mats E. Pettersson *et al.*, 2009), we had to take into account that cells with higher initial gene copy number are more likely to undergo a recombination event; and ii) as our experimental setup only allowed us to see if there has been a change in gene copy numbers or not, we had to take into account that there are some recombination events that do not change the gene copy number.

To account for i), we first computed the probability distribution that a given number of independent recombination events occur (Fig. S5d): given the assumed independence of recombination events, the probability of observing a certain number of recombination events for a given cellular trace is approximately Poisson distributed, with the parameter being the expected number of events per microfluidic experiment duration (i.e., the effective recombination rate times the number of generations). The total number of observed generations was: 37.7, 36.3, and 41.3 for the three microfluidics experiments, respectively. Our approach is an approximation, namely it assumes a constant effective recombination rate for each trace throughout the experiment, which can be violated if more than one recombination event occurs. For example, the first recombination event can change the gene copy number, which in turn changes the probability of subsequent recombination events happening. While it is in principle possible to take this into account, it substantially complicates the inference of the recombination rate from data and makes it strongly model dependent.

As per our model assumption, the effective recombination rate is equal to the initial number of gene copies times the basal recombination rate. Therefore, we used all single cell traces to estimate a starting gene copy distribution. To do this, we averaged the normalized fluorescence (as a proxy for the starting effective gene copy number, see Fig. 2.3c) over the time points 20 through 50. Next, we computed a Poisson probability distribution of obtaining k events (k=0,1,…) in the time of the experiment for each individual trace, with the basal recombination rate multiplied with the starting gene copy number (Fig. S5d). For example, if a single cell trace started with 4 gene copies, the expected number of events per experiment would be 4 times the basal recombination rate times the number of generations. Next, we averaged over all computed Poisson probability distributions, obtained from all single cell traces. This effectively means obtaining a total probability distribution for seeing 0, 1, or more recombination events over all recorded single-cell traces, taking into account point i).

Next, we consider point ii), taking into account the effect of recombination events that do not change the gene copy. We know from the P_{kj} matrix that the probability of keeping the gene copy numbers is the reciprocal of the initial gene copy number. Therefore, we took into account all events that would be seen as zero or single events (but are not) and adjusted the probability distributions. For this, we defined two probability distributions: the distribution of observed events, $p_{observed}$, which we are trying to find; and the distribution of "actual" number of events, p_{actual}, which we computed as described above. For example, in the observed distribution that is compared with experimental data, we classified as single events all double events where one of the recombination events leaves the copy number unchanged, all triple events where two events keep the copy numbers unchanged, etc. Therefore, the probability of observed events also includes the actual probability from states with k>0 in which recombination did not change the copy number: $p_{observed(k=0)} = p_{actual(k=0)} + \sum j\ p_{actual(j)} / \varepsilon_0^j$, for all j>0, with $p(j)$ being the probability of having j recombination events, and ε0 being the initial gene copy number in the given single cell trace (estimated from experimental single cell traces). The $(1/\varepsilon_0)^j$ represents the probability of having j consecutive recombination events, all of which leave the gene copy number unchanged. Analogously, the observed probability for a single event (k=1) to occur is: $p_{observed(k=1)} = p_{actual(k=1)} + \sum j\ (j\text{-}1)\ p_{actual(j)} / \varepsilon_0^j\text{-}1$, for all j>1. The prefactor (j-1) comes from the number of different possibilities of having events that keep the gene copy number unchanged. For example, having 3 recombination events, there are 3 different ways of having two events that keep the gene copy number unchanged while one event changes it.

After taking both corrections into account, we obtain a probability distribution of observing k recombination events (Fig. S5d). The estimate of the basal recombination rate, k_{rec}^0, is based on the proportion of traces classified by our algorithm as no mutation events. We looked for such a recombination rate that best matched the number of no-events in the probability distribution (Fig. S5c-d). We obtained k_{rec}^0 as 0.01434 per cell per generation, which is approximately 5x larger than the conservative lower bound.

2.5.12.2 Model comparison with experimental data

For comparison of the model with the experimental data (Fig. 2.5a), we simulated the full experimental protocol (for parameter values, see Table S2):
We exposed a single copy, ancestral population to a week of high expression environment, driving the population structure close to equilibrium. This mimicked the evolution of the amplified strain in the high expression environment such that both experimental and simulated population started with the same degree of copy number polymorphism.
The population spent one day in the low environment (for details on procedure in each day, see below).
For the experiment shown in Fig. 2.5a top panel, the population was additionally exposed to three daily oscillations between high and low expression environment.

The population was exposed to the environments indicated in Fig. 2.5a.
For every experiment, bacterial culture was diluted by a factor of *D*=133 every day, thus limiting growth. This growth limitation was enforced by multiplying all growth rates by $g(c) = (1 - \min(c/133,0))^{0.01}$, with c being the number of cells, relative to the number of cells after each dilution. The exponent 0.01 was chosen such that *g(c)* was smooth but nearly a step function.
To compare the units of experimental and simulated data, we obtained a common reference point. We took this to be the expression value after one week in the high expression environment, when the population has already equilibrated. We aligned these two points to have the same expression value. This value varies between different experiments.

The simulation of one day consisted of (for parameter values see Table S2):
Given the recombination rate and number of states *M*, we computed the transition matrix P_{kj} (see Eq. 1) in the following way: given *k* copy numbers, the probability of going from *k* to *j*<*k* copy numbers equals j/k^2, while probability for *k* to *j*>= equals $(2k\text{-}j)/k^2$ (Mats E. Pettersson *et al.*, 2009). Furthermore, we assumed that no transitions that increase copy numbers beyond *M* are allowed. We implemented this by setting all probabilities that go over *M* gene copies to zero.
Next, to update the current population structure following Eq. 1, we used the current population structure, N_j, selection on the states (growth rates) in the given environment, s_j (Fig. 2.2d), transition matrix, P_{kj} (probability of having *j* copies given *k* copies), the duplication and recombination rate (k_{dup} and k_{rec}, respectively), and the dilution factor *D*. First, we computed the total population growth since the last dilution, i.e., the ratio of population size of current time point and the size after last dilution. Second, we computed *g(c)* (taking into account the saturation of the population) and multiplied it with each of the selection values s_j in Eq. 1. Then, we used these new values to compute N_j at the new time point.
We repeated the step 2 for 23 or 22 times for low or high expression environment, respectively. These numbers represent the number of cell divisions per day and were determined experimentally. Steps 2-3 represent time evolution of the population over the period of one day.
We diluted the population by a factor of *D*=133.
We repeated the steps 2-4 according to the environment the population is exposed at on the new day (selection different between the two environments). With this step, we simulate different days, diluting after each (step 4).
For each time point, we computed expression as the average gene copy number: $E=\sum j w_j$, where w_j is the proportion of cells with *j* gene copies and sum goes over all gene copy numbers.
At the end, we returned the population distribution and expression at each time point.

For simulation of the stochastic environmental durations, we followed the same procedure as for the deterministic ones, except that the environment durations here were randomly drawn from an exponential distribution.

2.5.12.3 Finite size population model

To compute the response times for a finite size population (Fig. S6f), we used the Wright-Fisher model where the population size is kept constant. The procedure was:

Given all parameters of the system and using the infinite size population model (Eq. 1), we obtained the equilibrium distribution of the population in the starting environment. We computed the equilibrium distribution of copy numbers in the infinite population size limit by computing the eigenvector corresponding to the largest eigenvalue of the transition matrix (obtained from r.h.s. of Eq. 1), and obtained the starting finite population as a multinomial draw of N individuals from this equilibrium distribution.

After the environmental transition, we updated the distribution after each division. The new distribution was computed using the Eq. 1.

We computed the new population, as a multinomial draw of N individuals, randomly drawn from the new population distribution.

After each division, we computed the expression of the population.

We repeated steps 3-5 until response $R=M/2$ has been reached. The number of generations until this point represents the time to response. We define response as the ratio of mean copy numbers before and after the environmental switch.

Fig. S6f shows the response time as the average over 100 replicate simulations of the algorithm above.

2.5.13 Quantification and Statistical Analysis

Statistical details of individual experiments, including number of replicate experiments, mean values, and standard deviations, are described in the figure legends and indicated in the figures.

For the t-test in Fig. 2.4c-d we computed the response as the fold change between mean expression of days 1-15 in the high expression environment and mean expression in the low expression environment on day 16 for amplified populations (Fig. 2.4c). For the co-culture populations (Fig. 2.4d), we analogously computed the response as fold change between mean constitutive strain abundance of days 1-15 in the high expression environment and mean constitutive strain abundance in the low expression environment on day 16.

We used a two-sided t-test (Matlab function ttest2) to compute the p-value ($2.6*10^{-68}$) for the difference in mean response between amplified (Fig. 2.4c) and co-culture populations (Fig. 2.4d).

For measuring the linear dependence between the experimental data and model prediction in Fig. 2.5a, we computed the Pearson correlation coefficient using the inbuilt Matlab function corrcoef.

2.5.14 Data availability

Experimental data that support the findings of this study have been deposited in IST DataRep and are publicly available at https://doi.org/10.15479/AT:ISTA:7016.

3 An improved experimental system to create fluctuating environments

In this chapter, we briefly discuss the utility of dual selection systems for the directed evolution of a regulated promoter and describe how the *galK* selection system can be improved to reduce the emergence of escape mutants.

3.1 Introduction

Gene regulation is ubiquitous in all organisms and allows alleviating selective tradeoffs which exist for a given gene under different conditions (Troein *et al.*, 2007; Poelwijk, de Vos and Tans, 2011). In other words, the widespread existence of gene regulation hints at the fact that the expression of most genes is favorable only under certain conditions, not others. Understanding the evolution of gene regulation is an ongoing effort (Poelwijk, de Vos and Tans, 2011; Tuğrul *et al.*, 2015; Wolf, Silander and van Nimwegen, 2015; Friedlander *et al.*, 2016; Igler *et al.*, 2018). While genomic studies show that promoter regions evolve more rapidly than coding regions (Villar, Flicek and Odom, 2014; Villar *et al.*, 2015), theoretical models indicate that the evolution of regulated promoter sequences (unlike constitutive ones (Yona, Alm and Gore, 2018)) is exceedingly slow (Tuğrul *et al.*, 2015), (Grah , Lagator , Guet, Tkačik G. Evolving complex promoters for complex phenotypes. Manuscript in preparation). To resolve this apparent paradox, it has been hypothesized that mutations other than point mutations, such as transposition, deletion or duplication, play an important role in the evolution of complex promoter sequences and drive the fast turn-over observed in *cis*-regulatory sequences (Surguchov, 1991; Matus-Garcia, Nijveen and van Passel, 2012b; Nijveen, Matus-Garcia and van Passel, 2012; Oliver and Greene, 2012; van Passel, Nijveen and Wahl, 2014; Yona, Alm and Gore, 2018). However, experimental evidence of this so-called “promoter-shuffling” has remained anecdotal (Anderson and Roth, 1978; Kloeckener-Gruissem and Freeling, 1995; Zinser and Kolter, 2004; Blount *et al.*, 2012; Taylor *et al.*, 2015), as few studies have attempted to evolve gene regulation via directed evolution.
While studies involving large libraries of promoter sequences helped to map the immediate mutational space surrounding both evolved and random promoter sequences (Kinney *et al.*, 2010; Lagator *et al.*, 2017, 2020; Igler *et al.*, 2018), we are still left completely in the dark as to whether promoter regions in fact evolve step-by-step via single point mutations. This question can only be answered if we observe the evolution of promoter sequences while it is ongoing. Thus, evolution experiments could give insights into the type(s) of mutations involved in the process of promoter evolution and the resulting evolutionary dynamics.

Experimental evolution has offered exciting insights into adaptive mutations underlying traits like metabolic innovation, antibiotic resistance or cellular mobility (Blount and Lenski; Poelwijk, de Vos and Tans, 2011; Näsvall *et al.*, 2012; Taylor *et al.*, 2015; Steinrueck and Guet, 2017; MacLean and Millan, 2019). Yet, it has been of limited use for the study of the evolution of gene regulation. Studies are hampered by the lack of experimental conditions,

which select for the regulation of a single gene/operon of interest. To achieve this, expression of the gene must be tied to an experimentally controllable fitness benefit and cost, respectively, in two specific experimental conditions. However, few genes are known for which these conditions can be controlled experimentally.

In one experimental evolution study, Poelwijk et al. use experimental evolution to show that fluctuating selection can change the regulatory logic of an existing bacterial promoter (Poelwijk, de Vos and Tans, 2011). The study powerfully demonstrates just how fast existing regulatory logic can change, however, there are two drawbacks to the methodology used. First, mutations are introduced artificially, constraining the evolutionary trajectory to adaptation that occurs via single SNPs. Second, their experimental system for alternating selection makes use of a synthetic operon consisting of two separate selection markers: *sacB* for selecting against expression in a sucrose environment, and *cmR* selecting for expression in a chloramphenicol environment. As both genes in the operon are under different selection pressures, a frequent route of escape from selection is the inactivation of the gene *sacB* used for counter selection. Loss of function (LOF) mutations of *sacB* do not affect growth in chloramphenicol and are thus beneficial in both environments. In fact, LOF mutations are an intrinsic problem of any two separate selection markers and render long-term evolution experiments impossible, as escape mutants will arise early on and spread to fixation, abolishing the counter-selection part of the two environments.

One obvious improvement over selection cassettes containing two genes are dual selection markers, which consist of only a single gene. In *E.coli*, three genes are known whose expression can be selected for and against: *galK* (using DOG and galactose), *tetA* (using fusaric acid and tetracycline) and *pyrF* (using 5-fluoroorotic acid and a medium without uracil). All three dual selection markers have been used extensively for the construction of scar-less deletion mutations in diverse organisms (Reyrat *et al.*, 1998). There, the insertion of a gene of interest along with a dual selection marker into a genomic position of choice is selected for using the positive selection condition. Then, the removal of the dual selection marker can be achieved selecting against its presence (expression). Single gene dual selection markers are desirable, as one frequent route towards escape from negative selection, namely LOF mutations, are not viable during positive selection. Hence, in principle, escape mutants should not be possible by simple LOF mutations. Instead, the only way of adaptation should be by evolving gene regulation or increased substrate specificity to discriminate against the negative selection component.

In the following, we aimed to optimize the dual selection system based on *galK*.

3.2 Results

In *E.coli* two major routes exist for the import of galactose, the high affinity D-galactose/methyl-D-galactoside uptake system, *mglBAC,* and the galactose permease *galP* (Henderson, Giddens and Jones-Mortimer, 1977; Nagelkerke and Postma, 1978; Csiszovszki *et al.*, 2011). In order to constrain the possible ways of importing galactose without

importing DOG into the cell, we deleted *mglBAC.* Furthermore, we put *galP* under the control of a constitutive promoter, as galactose, but not DOG acts as an inducer for the galactose regulon. Using this genetic background, we tested the response of a strain constitutively expressing *galK* to alternating galactose and DOG selection (Fig. 3.1). The first two of the daily OD_{600} measurements confirmed the strain´s capability of growing in minimal medium supplemented with 0.1% galactose, but not in minimal medium containing 1% glycerol and 0.0001% DOG. However, already on the second day spent in DOG medium, most replicate populations showed increased OD_{600} values. On the third day spent in DOG, populations exhibited an OD_{600} even greater than in galactose.

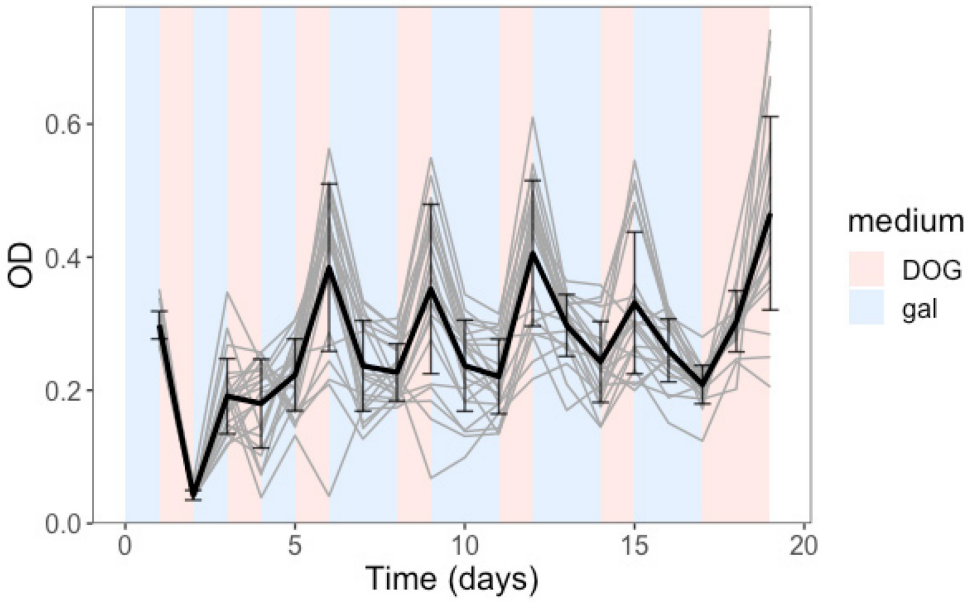


Figure 3.1. ***Day-wise OD_{600} of 24 populations of strain IT028-H5r constitutively expressing galK in fluctuating selection.***
Intervals of selective conditions vary between one and two days and are indicated in the plot: DOG=2-deoxy-galactose, gal=galactose.

We wanted to understand whether these populations could grow in DOG due to phenotypic acclimation or whether they evolved DOG-resistance. Therefore, we compared the growth of the evolved populations with their ancestors in liquid 0.0001% DOG medium and on 0.001% DOG agar plates, where we plated only 10^5 cells. Figure 3.2 summarizes the results of these experiments: the ancestral populations grew in the liquid DOG medium, however, with a considerable lag phase of 20 hours as compared to the populations of the same strain, which evolved in alternating selection for 19 days. Importantly, while the evolved populations formed visible colonies on the DOG agar plate after 24h, the ancestral strain did not grow on the DOG agar plate even after 1 week of incubation. The observation that 10^5

cells were too few to form a visible colony on DOG agar indicates that mutation, and not acclimation is the reason behind DOG resistance.

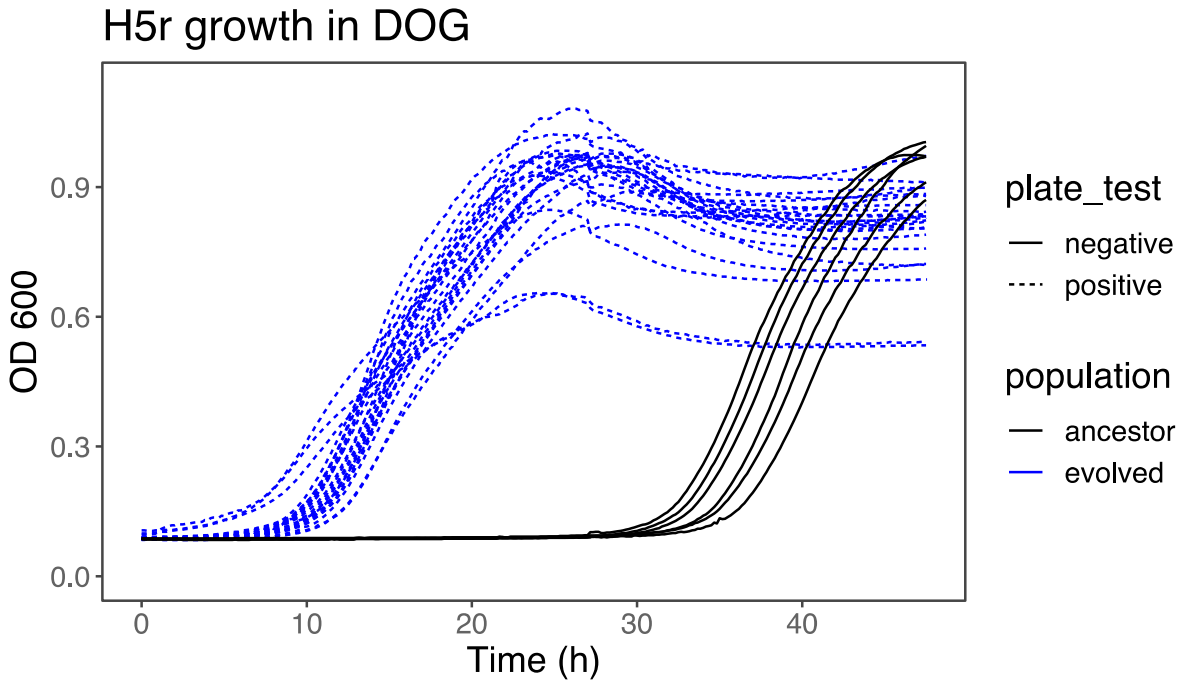


Figure 3.2 ***Growth of six ancestral populations (light green) and populations evolved in alternating selection (Fig. 3.1) (dark green) in 0.0001% liquid DOG medium.***
10^5 cells of the same populations were also plated on 0.001% DOG agar plates and growth was scored after one week of incubation (plate test).

To confirm this result, we quantified the spontaneous rate for DOG resistance evolution by plating 10^8 cells on a 0.001% DOG agar plate, where approximately 50 colonies formed from single cells, yielding a mutation rate towards DOG resistance of $5*10^{-7}$ per cell per generation.
However, the important question for alternating selection is, whether DOG resistance occurred by losing the ability to grow on galactose, or by evolving the means to grow in both environments. We hypothesized that the evolved populations growing in both environments might in fact consist of two specialist subpopulations, one growing only in galactose and the other growing only in DOG due to LOF mutations in *galK*.
To test this hypothesis, we assayed whether the evolved colonies isolated from the DOG agar plate were also able to grow on 0.1% galactose agar. Surprisingly, all of the colonies grew in both environments, thus representing true escape mutants. Sequencing of the promoter region of several evolved clones showed an ancestral *galK* promoter region (i.e. the *galK* gene was still constitutively expressed). We also did not find mutations in *galK* itself. Therefore, we sequenced the whole genome of one escape mutant and found a 12 bp deletion in the galactose permease, *galP*. Sequencing of *galP* from nine further escape mutants showed mutations in all of them (Table 3.1) that were either SNPs resulting in

nonsynonymous or frameshift mutations and deletions. Judging from these sequencing we can only be certain that frameshift mutations lead to LOF of GalP. However, given the plethora of other *galP* mutations we found, we assumed that all of them are likely abolish *galP* function. This result was puzzling, given that GalP should have been the only cellular entry point for both DOG and galactose in our *mglBAC*- strain background.

Table 3.1. ***Mutations of the galactose permease, galP, in 10 evolved clones of strain IT028-H5r, constitutively expressing galK****.* NS = nonsynonymous.

clone	Type	start	end	Length
17_I>F	NS	1075	1075	1
19_G>V	NS	944	944	1
12_ins_FS	frameshift	877	877	1
11_ins_FS	frameshift	862	862	1
18_L>P	NS	464	464	1
15_G>R	NS	457	457	1
16_G>V	NS	347	347	1
WGS_del	deletion	197	208	12
20_del	deletion	197	208	12
13_del	deletion	197	208	12

The explanation behind the *galP*- mutants had to be that another route of import exists for galactose, which does not import DOG. Indeed, Kornberg and Riordan showed already in 1976 that this is the case: if present in concentrations up to 2mM, galactose can diffuse through the phosphotransferase system via PtsG (Kornberg and Riordan, 1976). To test whether this explanation is in agreement with our experimental observations, we compared the growth of three different strains in different concentrations of galactose and DOG: i) one without *galK* expression, ii), one constitutively expressing *galK and* iii) an escape mutant, constitutively expressing *galK* with a *galP* deletion (Fig.3.3). First, the experiment shows that the strain expressing *galK* (H5) and a strain without *galK* expression (IT028) show diametrically opposite patterns of growth. This highlights the strong tradeoff, which exists for *galK* expression in the range of galactose and DOG concentrations tested here. The results also show that the escape mutant (H5-*galP* -) is capable of growing in 0.0001% DOG (the concentration we previously used for alternating selection experiments) just as well as the ancestor lacking *galK* expression. At the same time, the escape mutant is not impacted by the *galP* deletion in the highest galactose concentration (Fig. 3.3A - upper left panel). Importantly, however, the escape mutant fails to grow in 0.01% galactose, a concentration that is conducive to the growth of the constitutive strain with functional *galP* (Fig.3.3A – upper right panel). In summary, the experiment confirms that in the absence of *mglBAC, galP* is necessary to take up galactose present in low concentrations.

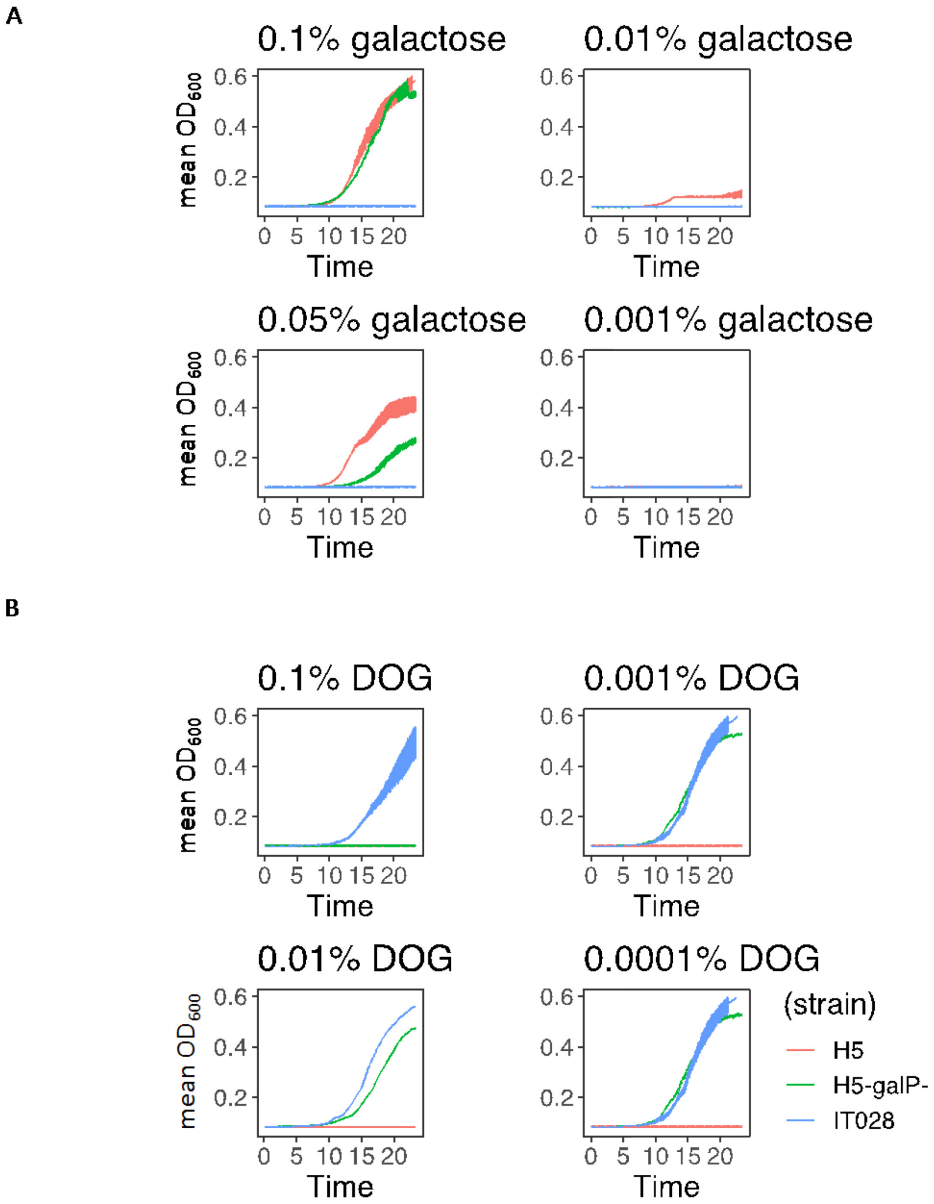


Figure 3.3. ***Mean growth curves of three biological replicates in different concentrations of galactose (A) and DOG (B), respectively.***
H5=strain H5r, constitutively expressing galK. H5-galP-=strain H5r without functional galP, IT028=strain with low baseline galK expression driven by p0.

Our follow-up experiments confirmed that the concentrations of galactose and DOG used for alternating selection in fact allow for a frequent escape route via simple LOF-mutations of *galP*. Knowing, however, that *galP* is indeed required for galactose uptake when present in low concentrations, we wanted to find a concentration of galactose and DOG not permissive to the growth of *galP*- strains, but permissive to the growth of constitutive or ancestral strain, respectively. Figure 3.4. summarizes the growth rates for the three strains in all different environments. In 0.01% galactose, only the constitutive strain can grow, but

neither ancestor nor escape mutant. In 0.1% DOG, only the ancestor can grow, but neither the constitutive nor the escape mutant. Using a combination of 0.01% galactose and 0.1% DOG in alternating selection should thus abolish the fixation of simple LOF escape mutants, as *galP* is required in either of them.

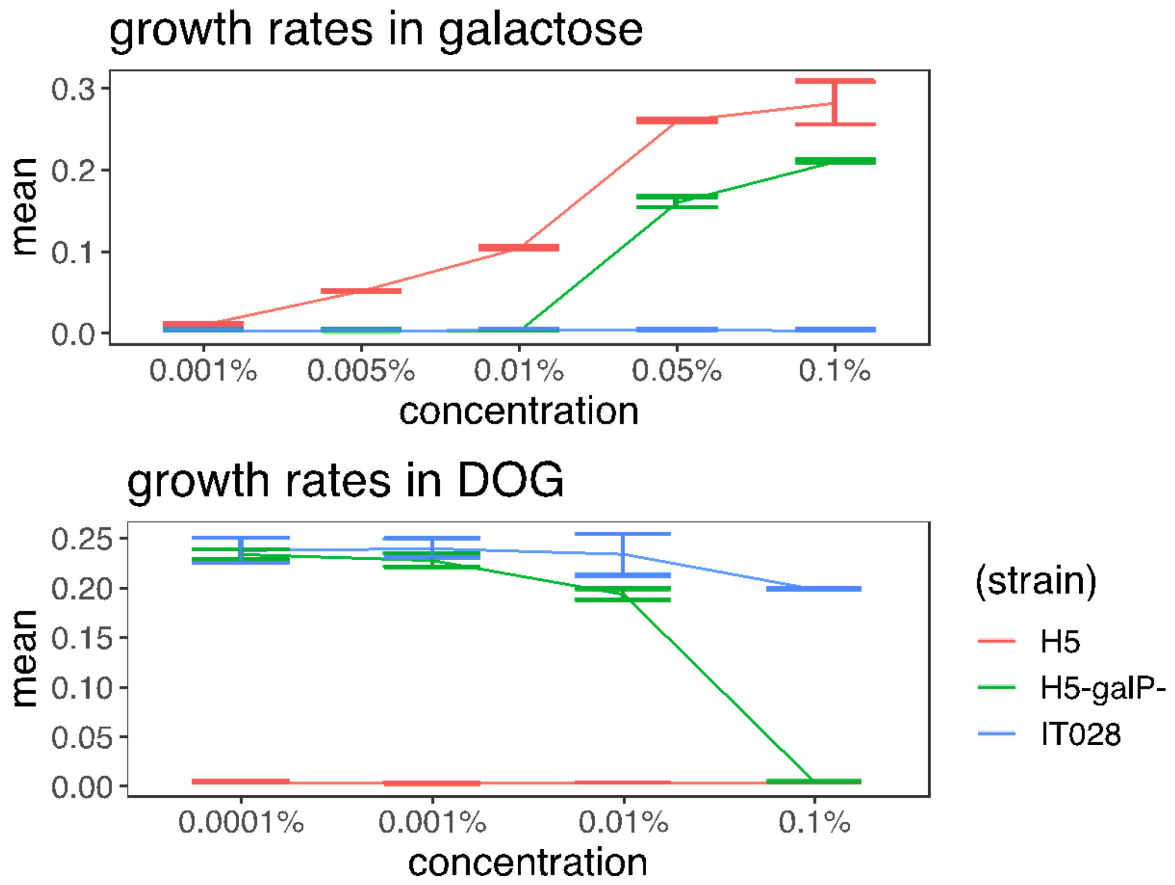


Figure 3.4. ***Finding a combination of galactose and DOG concentrations detrimental to escape mutants.***
Mean maximal growth rates of three biological replicates in different concentrations of galactose (upper panel) and DOG (lower panel). H5=Strain H5r, constitutively expressing galK. H5-galP,=Strain H5r without functional galP, IT028=ancestor with low baseline galK expression.

3.3 Discussion

This chapter explored the only route of escaping galactose and DOG selection we have so far encountered: LOF mutations in *galP* no longer allow import of DOG into the cell. At the same time, galactose can still enter the cell through facilitated diffusion via the phosphotransferase system if present in concentrations greater than 2mM (Kornberg and

Riordan, 1976). Importantly, *galP*- strains are not completely resistant to DOG. Unlike *wt galP* strains that lack *galK* expression like the ancestral strain, they cannot grow in 0.1% DOG. Presumably this is because DOG, like galactose, can enter the cell via some other route if present in high concentrations. Accordingly, *galP* LOF mutants are neither viable in 0.01% galactose, nor in 0.1% DOG. When trying to select for the evolution of a regulated promoter from a random sequence, however, care must be taken that the environments are roughly symmetric in terms of the number of cell divisions sustained by the respective growth medium. Low concentrations of galactose sustain only a small population of cells reflected in the low final OD_{600} (for instance, 0.01%: Fig.3.3A – upper right panel). Given that the overall fitness in fluctuating environment is the product of the fitness in either environment (Yi and Dean, 2013), fitness in galactose loses its relevance if cells spend fewer generations there as compared to DOG. Therefore, one simple way to increase overall fitness in these two environments would be to ”gnore” fitness in galactose by losing *galK* or *galP* and becoming DOG resistant. It may therefore be advisable to increase the number of generations by simply growing populations in galactose for two or more consecutive transfers. As an alternative, *ptsG*- strains could be used to avoid facilitated diffusion of galactose into the cell in the absence of *galP*.

3.4 Methods

All strains and experimental conditions, including the alternating selection experiment, are described in chapter two (Tomanek et al., 2020).

We used a Biotek H1 platereader (Biotek, Vinooski, Vermont) to measure OD_{600} in M9 minimal medium supplemented with different concentrations (as mentioned in the text and figures) of galactose and DOG plus 1% glycerol. To obtain growth rates in the different galactose and DOG concentrations we analyzed growth curves using a custom R script (R Development Core Team, 2005), where a linear model (function lm) is fitted to a sliding window of 20 data points (corresponding to 200 minutes) of $\log(OD_{600})$ over time. The steepest of these tangents is selected for every growth curve, respectively, and plotted (Fig.3.4).

To sequence the whole genome of one DOG-resistant clone, we grew a colony in LB medium and extracted genomic DNA using a Wizard Genomic DNA purification kit (Promega, Madison, Wisconsin). Illumina paired end 2*125bp sequencing was carried out by GATC biotech (Konstanz, Germany).

4 Gene copy number mutation can hinder the evolution by point mutation

4.1 Introduction

In this chapter, we are exploring the adaptive evolution of increased enzyme *activity* that can occur via two broad classes of mutations: i) single point and small insertion or deletion (indel) mutations in the promoter or coding region of a gene and ii) mutations to the copy number of the whole gene.

Here, we consider the different biological properties of the two mutational classes and ask which dynamics arise when these two types of mutations occur as a combination. In fact, the combination of these two mutations forms the basis for most evolutionary innovations in gene functions (Conant and Wolfe, 2008; Andersson, Jerlström-Hultqvist and Nasväll, 2015). It seems likely that the plethora of highly specific and efficient present-day enzymes evolved from a much smaller set of proto-enzymes that existed some 4000-3500 million years ago. These ancestors of present-day enzyme super-families were characterized by a low substrate-specificity and therefore catalyzed a wider range of reactions, albeit with low efficiency for each individual substrate (Kacser and Beeby, 1984). In essence, duplication provided the source for the adaptive radiation of specialized present-day enzymes from a smaller number of promiscuous ancestral enzymes (Kacser and Beeby, 1984; Lynch and Conery, 2000). After duplication of an ancestral gene, the second gene copy is free to accumulate point mutations, which may in rare cases eventually lead to functional novelty, a process referred to as neofunctionalization (Ohno, 1970). Since Ohno´s monograph, approximately a dozen models have been put forward to describe the fate of gene pairs after a duplication mutation occurred. Each model differs with respect to the time point point mutations occur and whether or not (and at which point) directional selection acts on the pair of paralogous genes (Conant and Wolfe, 2008; Innan and Kondrashov, 2010). In general, a lot of attention has been focused on understanding the process of divergence (Lynch and Conery, 2000; Teufel, Masel and Liberles, 2015) as opposed to the short-lived dynamics associated with the initial duplication itself. This is due to the technical difficulties of studying transient copy number variation (Andersson and Hughes, 2009; Lauer and Gresham, 2019; Belikova *et al.*, 2020; Tomanek *et al.*, 2020) on the one hand, and the plethora of long-term evolutionary data documenting the sequence divergence of paralogs on the other hand (“attention is shifted to where the data are”) (Kondrashov, 2012).

The lack of interest in the initial stage of paralog evolution, i.e. the copy number mutation itself, is especially puzzling, since the canonical model of paralog evolution put forward by Ohno rests on one important assumption: gene duplication is considered a neutral mutation, where the second, redundant copy is freed from purifying selection and can accumulate point mutations (Ohno, 1970). However, several problems exist with the assumption of neutrality.

First, neutral evolution means that the fate of duplicates will most often be nonfunctionalization as the duplicate gene copy accumulates loss-of-function mutations in

the absence of purifying selection. Hence, nonfunctionalisation is expected to be the most common fate of neutral duplications (Ohno, 1970; Lynch and Conery, 2000; Bershtein and Tawfik, 2008). However, the rate of homologous recombination of a duplicated fragment is even higher than the rate of LOF point mutations (leading to nonfunctionalization) (Mats E Pettersson *et al.*, 2009; Reams *et al.*, 2010; Tomanek *et al.*, 2020). Therefore, in the absence of selection, the most common fate for any arising duplication should be its reversal to a single copy by a deletion mutation, not its pseudogenization. Accordingly, the fraction of duplicates, which eventually fixed has survived degrading mutations either due to drift or because of a selective advantage. Comparative genomics data suggests that in eukaryotes the most frequent long-term evolutionary fate of any fixed duplicate pair is subfunctionalization (Force *et al.*, 1999; Lynch and Force, 2000). After the duplicate pair fixes via drift both copies need to be maintained by purifying selection. This maintenance is best explained by complementary degenerative mutations: each gene in the pair loses one part of the ancestral gene function, such that both copies become necessary to maintain the ancestral function (Force *et al.*, 1999; Lynch and Force, 2000). An alternative model for subfunctionalization invokes adaptive mutations to resolve a tradeoff between two ancestral gene functions, which cannot simultaneously be optimized by natural selection. In practice, it may be hard to distinguish between passive and adaptive subfunctionalization even in the presence of experimental data on enzyme functions (Conant and Wolfe, 2008; Innan and Kondrashov, 2010).

Directly related to the properties of duplication, another problem exists with the core assumption of neutrality in Ohno´s model. Duplications are known to be rarely neutral, but often carry a significant fitness cost (Kondrashov *et al.*, 2002; Kondrashov and Kondrashov, 2006; Bergthorsson, Andersson and Roth, 2007; Bershtein and Tawfik, 2008; Mats E Pettersson *et al.*, 2009; Reams *et al.*, 2010; Adler *et al.*, 2014). In bacteria, the fitness cost of duplication has been estimated for duplications ranging 8 –1.246 kbp in size. In these, the average cost (selection coefficient) for each duplicated kbp ranges between $0.05*10^{-3}$ and $1.5*10^{-3}$ (Mats E Pettersson *et al.*, 2009; Reams *et al.*, 2010; Adler *et al.*, 2014). In *Drosophila melanogaster*, mutation accumulation experiments have estimated that 99% of all spontaneous duplications are deleterious and hence are ten times more likely to be removed by purifying selection than nonsynonymous point mutations (Schrider *et al.*, 2013). The cost of duplication can arise from various sources including i) the cost of superfluous expression, ii) consequences of novel sequences arising at the duplication junctions (for instance, the amplification of locus one described in chapter two has one junction inside ATPase subunit B (Tomanek *et al.*, 2020)) iii) dosage imbalance between the genes in the duplicated fragment and the rest of the genes remaining in single copy (Papp, Pál and Hurst, 2003; Katju and Bergthorsson, 2013), iv) the titration of transcription factors. In addition, more specific reasons may underlie further costs of duplication. For instance, the amplification of locus one described in chapter two extends over the origin of replication, *oriC*, likely interfering with the process of DNA-replication (Tomanek *et al.*, 2020). If the cost

of a duplication is substantial, positive selection is needed for it to be maintained in the presence of these costs (Adler *et al.*, 2014).

Countless examples exist of adaptive duplication and amplification that occurs in response to novel environmental pressures, i.e. in the presence of positive selection. Many of these examples include organisms responding to selection pressures created by humans, such as insecticides (Bass and Field, 2011b), herbicides (Tranel, 2017), antibiotics (Sandegren and Andersson, 2009) or chemotherapeutics (Albertson, 2006). Other striking examples include the adaptation to starch-rich diets, which occurred independently in several mammalian lineages including humans through the amplification of the salivary amylase gene AMY1 (Perry *et al.*, 2007; Pajic *et al.*, 2019) or the amplification of a horizontally transferred gene in the insect *P. vanderplanki* that underlies its extreme desiccation tolerance (Gusev *et al.*, 2014). While the evidence for selective duplication and amplification as the driving force for paralogization still remains anecdotal, it is increasingly recognized as an alternative or at least complementary model to the canonical model of neutral evolution (Moore and Purugganan, 2003; Kondrashov and Kondrashov, 2006; Shiu *et al.*, 2006; Conant and Wolfe, 2008; Kondrashov, 2012).

Ohno´s original model included neutral evolution in order to explain how an assumed tradeoff between ancestral and novel enzyme function can be overcome. However, the degree of tradeoff between old and new enzymatic function varies widely in known enzymes (Bershtein and Tawfik, 2008). If mutations towards the new gene function partially compromise the ancestral function, selection may favor additional gene copies, which increase the fitness of evolutionary intermediates (Bershtein and Tawfik, 2008).

As we have seen above, the model of paralog evolution benefits from invoking selection for increased dosage in three ways: i) it guards enzymes against LOF mutations (Bershtein and Tawfik, 2008), ii) allows to fix duplications despite their cost (Katju and Bergthorsson, 2013; Adler *et al.*, 2014) and iii) dosage-selection is fully consistent with the exaptation of pre-existing secondary enzymatic functions through duplication-divergence (Kacser and Beeby, 1984; Conant and Wolfe, 2008).

The exaptation of such pre-existing, secondary enzymatic functions (also referred to as moon-lighting functions) (Tawfik, 2010; Copley, 2017) is conceptualized in the Innovation-Amplificaton-Divergence (IAD) model of paralog evolution (Bergthorsson, Andersson and Roth, 2007), which was later validated by evolution experiments (Elde *et al.*, 2012; Näsvall *et al.*, 2012).

The IAD model posits that selection for increased dosage of an enzyme´s moonlighting-function (referred to as the “innovation” despite being an existing function as it becomes “innovative” only in a novel environmental condition) leads to copy number increases (amplification), which eventually result in the fixation of a point mutation that improves this moonlighting function (divergence): a new protein function is born from an old one. After the new (stronger) function is present, superfluous additional gene copies will be lost again, leaving only the copies of the two (ancestral and evolved) paralogs (Fig. 4.1a). Crucially, IAD differs from the neutral model of neofunctionalization in that selection not only occurs after

the diverged function occurred, but is already driving the fixation of the duplication in the first place (Bergthorsson, Andersson and Roth, 2007; Conant and Wolfe, 2008).

After its proposal, IAD has mainly been regarded in the microbiology community, partly due to the use of different terminology ("gene amplification" versus "duplication" or "copy number variation (CNV)") (Conant and Wolfe, 2008; Kondrashov, 2012), and potentially also because adaptive duplication and amplification is frequently observed by microbiologists exposing bacteria to a variety of strong selection pressures in the laboratory (Roth *et al.*, 1988; Sandegren and Andersson, 2009). This is slowly changing (Conant and Wolfe, 2008; Kondrashov, 2012).

One prevalent interpretation of the IAD model is that amplification plays a dual role: it not only occurs as an adaptation to selection for increased gene dosage, but it also "facilitates functional innovation". It has been hypothesized that initial adaptation by amplification increases the target size for point mutations, either directly given multiple gene copies or, indirectly, by allowing survival of a population under stressful environmental conditions until further adaptation occurs (Andersson and Hughes, 2009). Since this idea is intuitively obvious, several experimental studies interpreted their observations of adaptation according to the IAD model in the light of "amplification as a facilitator of functional innovation" (Sandegren and Andersson, 2009; Song *et al.*, 2009; Pränting and Andersson, 2011; Elde *et al.*, 2012; Näsvall *et al.*, 2012). However, none of the studies showing the occurrence of IAD provide a direct test of the hypothesis that amplification speeds up adaptation.

The original idea that amplification increases the chance for further adaptation by point mutations comes from a classic paper (Andersson, Slechta and Roth, 1998). Their *amplification-mutagenesis hypothesis* was originally developed as strong evidence against the adaptive mutagenesis hypothesis proposed by Cairns and others (Cairns, Overbaugh and Miller, 1988; Cairns and Foster, 1991). Instead of directing mutations to the site of need as suggested by Cairns, adaptive amplification (in this case also in combination with induction of the error prone polymerase *dinB*) results in a higher mutation rate in the locus under selection (Hendrickson *et al.*, 1995; Roth and Andersson, 2004).

Support for the "amplification mutagenesis hypothesis" prediction that increased copy number accelerates the speed of divergence comes from a paper by San Millan et al. The authors show that when compared to chromosomal single copy genes, genes residing on multi-copy plasmids potentiate the evolution of paralogs (San Millan *et al.*, 2017). However, despite confirming that more gene copies indeed speed up their evolution, this experimental evolution scenario does not directly reflect the early dynamics during GDA formation. Similarly, even in a study, which explicitly looks at "the initial stages of duplicate evolution" the authors simply installed a tandem duplication onto a plasmid to study its fate. Moreover, they used a strain lacking the *recA* gene, thereby omitting further duplication or deletion by homologous recombination, a frequent mutation for tandem duplications (Reams *et al.*, 2010; Reams and Roth, 2015; Tomanek *et al.*, 2020) and one certainly relevant for the IAD model.

All else being equal, more copies indeed mean more target for mutations (San Millan *et al.*, 2017). However, all else is not necessarily equal, as the evolutionary dynamics may differ strongly between an organism that can increase copy number as an adaptation and an organism that cannot. Given that in asexual populations (typically used for evolution experiments), any two clones with adaptive mutations compete with each (Gerrish and Lenski, 1998), it is puzzling that nobody has considered that clonal interference could occur between point and copy number mutations in the context of the IAD model.

Interestingly, one experimental evolution study with barcoded yeast populations indeed finds strong clonal interference between different independently formed CNV lineages including some with secondary adaptive point mutations (Lauer et al., 2018). However, despite this finding the authors conclude that their results are in line with the amplification mutagenesis hypothesis, namely that copy number mutations speed up adaptation by increasing the target site for point mutations.

Another surprising finding comes from an experimental evolution study subjecting yeast populations to a regime of increased temperature (Yona *et al.*, 2012). When abruptly shifting the temperature, adaptation happens via a CNV in the form of an aneuploidy. In contrast, slowly increasing the temperature results in adaptation by point mutation. In the abruptly shifted population, eventually, the CNV disappears as adaptive point mutations fix. While these point mutations appeared only very late in the CNV populations, the authors still conclude that the CNV facilitated adaptation by point mutations.

Despite presenting findings that could be interpreted to suggest otherwise (Yona *et al.*, 2012; Lauer *et al.*, 2018), consensus in the recent literature seems to be that amplification (CNVs) not only serve as a first step in the "relay race of adaptation" (Yona, Frumkin and Pilpel, 2015), but that they also facilitate the evolution by point mutations, either indirectly by providing a first 'crude' adaptation to cope with stress until more refined adaptation occur by point mutations, or directly by increasing the target size for point mutations (Elde et al., 2012; Yona, Frumkin and Pilpel, 2015; Cone et al., 2017; Bayer, Brennan and Geballe, 2018; Lauer et al., 2018; Todd and Selmecki, 2020).

However, to our knowledge, no experimental study looking at the early dynamics of naturally occurring duplication mutations (Elde et al., 2012; Näsvall et al., 2012), has *directly* tested the hypothesis that amplification not only plays a role in but also increases the chance for functional innovation. Here, we test this simple hypothesis by comparing the speed at which two *E.coli* strains with high and low amplification ability diverge.

4.2 Results

An evolution experiment conducted in a locus exhibiting high rates of gene amplification (Steinrueck and Guet, 2017), which seemingly failed to produce any evolved clones with point mutations led us to hypothesize that amplification may hinder evolution by point mutations under certain conditions conducive to clonal interference. We will refer to this

hypothesis as "*amplification hindrance hypothesis*" in the following and suggest it could be an extension to the IAD model.

4.2.1 The amplification hindrance hypothesis

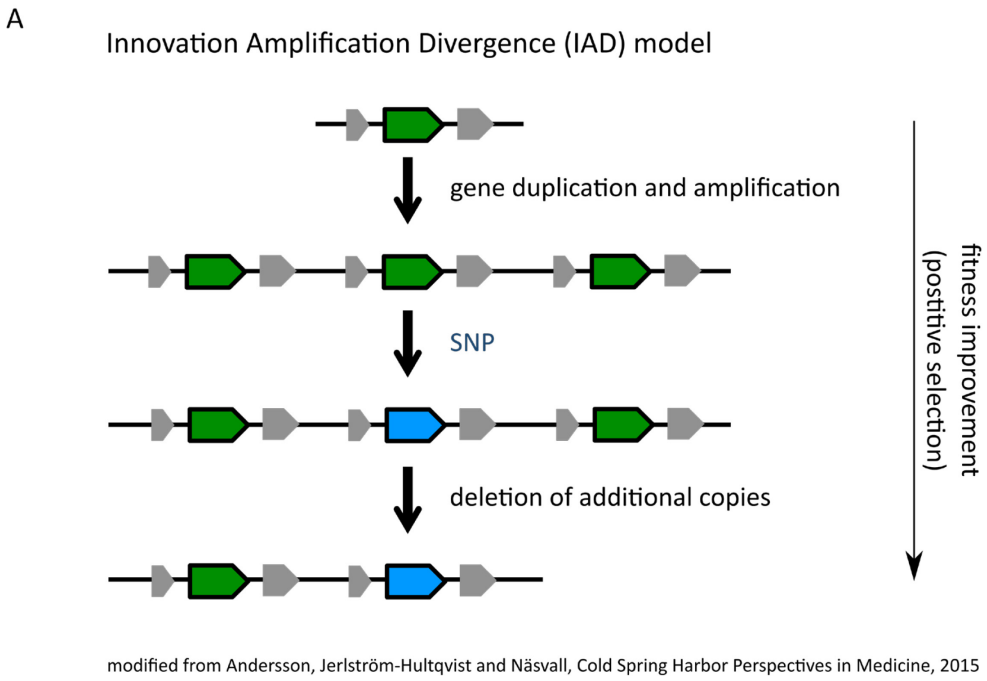
The IAD model describes how two mutations *consecutively* occur in a population (Fig. 4.1a) (Bergthorsson, Andersson and Roth, 2007; Näsvall *et al.*, 2012; Andersson, Jerlström-Hultqvist and Näsvall, 2015). First, copy number mutations improve a given biological function, such as a moonlighting-activity of an enzyme, by increasing its level of expression. Then, point mutations further increase this moon-lightning activity by improving the catalytic mechanism itself (Tawfik, 2010). As refining point mutations render the additional gene copies redundant, they are eventually lost due to their fitness cost, resulting in the fixation of the two paralogs.

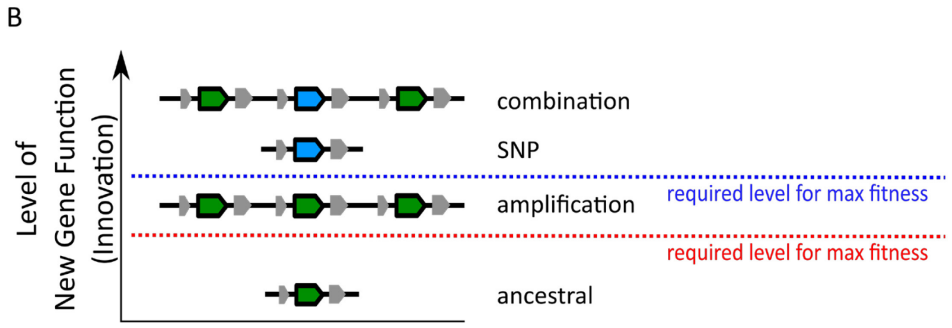
However, any two mutations need not necessarily occur consecutively in the *same* genetic background but may instead appear in *different* genetic backgrounds. In asexual organisms such as bacteria, this competition between two clones with adaptive mutations slows down the fixation of either mutation and is referred to as clonal interference (Gerrish and Lenski, 1998). The reason IAD does not include clonal interference, is because it assumes two consecutive selective sweeps: first the amplification sweeps the population and once it is fixed a point mutation occurs on top of it and sweeps again to fixation, ultimately leading to the fixation of only two divergence copies (paralogs) (Fig. 4.1a).

Hence an implicit assumption is needed for the amplification mutagenesis hypothesis to work within the IAD model: sufficient "room for improvement" is necessary for two consecutive selective sweeps to occur. However, selective environments might differ in their requirement for "improvement" or expression of a biological function (Fig. 4.1b). The scheme in Figure 4.1b shows two such thresholds for an expression requirement. In line with the IAD model, in the condition with a high expression requirement (blue dashed line), only the final point mutation confers the level of expression needed to reach maximal growth. However, in the condition with low expression requirement (red dashed line), a copy number mutation alone might confer sufficient expression needed for maximal growth. These two different expression requirements for maximal fitness give rise to two genotype-phenotype maps (fitness landscapes) with different topology (Fig. 4.1c). For the high expression requirement ("a lot of room for improvement"), adaptation occurs in a step-wise manner according to the IAD model (Fig. 4.1c - upper panel). For the low expression requirement ("less room for improvement"), there is a fitness plateau that hinders the second selective sweep, namely that of a point mutation being gained on top of an amplification (Fig. 1c - lower panel). The plateau simply arises from the fact that both, the amplification, and the combination of amplification and point mutation have an expression level above the required threshold shown in Fig. 1b (red dashed line). This means that a point mutation occurring on top of the amplification will not add any additional fitness

benefit. In this case, the point mutation has to sweep to fixation in a regime of drift, which is clearly slower than evolution under selection.
In both environments, the last step of IAD – deletion of additional copies resulting in a diverged gene- is driven by the selective benefit of losing costly gene copies. The difference between the two environments, and ultimately between the amplification hindrance or mutagenesis hypothesis, is simply whether or not a point mutation on top of the amplification will sweep to (near) fixation by drift or selection.

Importantly, the concept of an expression requirement threshold is not part of the original IAD model. However, we can safely assume that such maximum level of gene expression or enzyme activity exits for any enzymatic function. Increasing the enzyme abundance or activity beyond this maximum level will not add any selective benefit. According to our hypothesis, whether or not amplification speeds up evolution or could actually slow it down by clonal interference between amplified and combined mutants depends on how large this room for improvement is.
If a novel enzymatic function evolves from an existing moon-lighting function that is many orders of magnitude weaker (Tawfik, 2010), it is reasonable to assume that there is indeed a lot of room for improvement i.e. two (or even more) selective sweeps can occur consecutively. If so, we do not expect clonal interference to occur between copy number and point mutations and amplifications should indeed speed up adaptation as predicted by amplification mutagenesis hypothesis (Andersson, Slechta and Roth, 1998) implied in the IAD model (Bergthorsson, Andersson and Roth, 2007; Näsvall *et al.*, 2012).
However, situations may exist, where several rounds of amplification yield an expression level that is at the maximum beneficial level. While additional amplifications and point mutations may increase expression, they would not increase fitness (Fig. 4.1c – bottom panel). Clonal interference can happen with any two mutations, but gene copy number mutations are special in that they can be orders of magnitude more frequent than point mutations. Therefore, according to the amplification hindrance hypothesis, competing with amplifications may be very hard for point mutations.





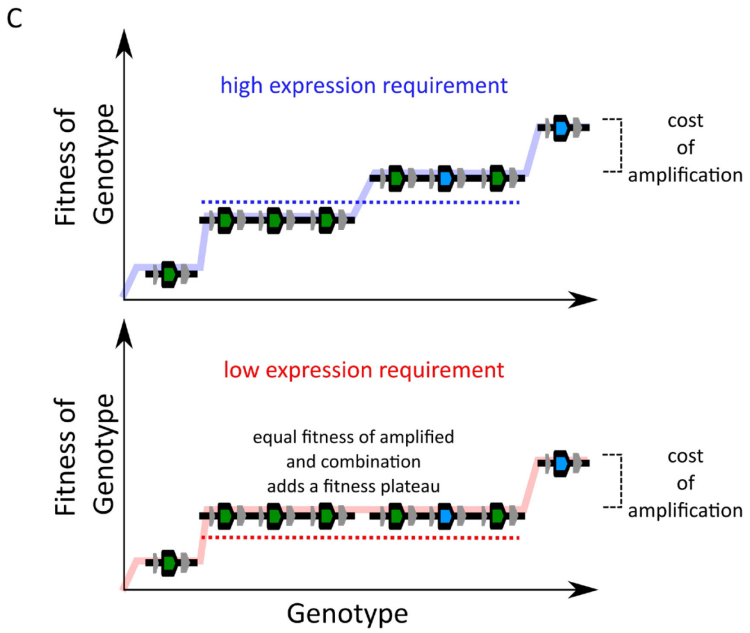


Figure 4.1. ***The amplification hindrance hypothesis****.*

***A**. Scheme showing the Innovation Amplification Divergence (IAD) model describing duplication and divergence under continuous selection for increased gene dosage, modified from (Andersson and Jerlstro, 2015).*
***B**. Expression level of the selected biological function is shown for all genotypes occurring during IAD. Two selective environments may differ in the expression level for which maximal fitness is achieved (dashed lines).*
***C**. Genotype-phenotype map showing the trajectory of adaptation. When high expression is required the trajectory (blue trace) allows adaptation to proceed as according to IAD (upper panel). When low expression is required, the adaptive trajectory (red trace) includes a fitness plateau resulting from the fact that expression of both amplification and combination mutants exceeds the required expression level and does not add any additional fitness benefit (lower panel). The last adaptive step in both trajectories is losing costly copy mutations with only the diverged copy remaining.*

4.2.2 An experimental system that allows phenotypically distinguishing copy number and point mutations in strains with *localized* differences in duplication rate

To test our amplification hindrance hypothesis in the context of the innovation amplification divergence model, we experimentally select for the increase in an enzymatic function in two strains of *E.coli* that differ in their rate of forming duplications.
It is important to note here that while we are interested in the evolutionary dynamics of amplification and divergence that play a role in paralog evolution we are not looking at a process of paralog evolution itself. Instead of studying aspects of divergence after duplication as it is commonly done, we are interested in the initial stage of paralog evolution, including the duplication itself. We therefore employ an experimental system that allows us to phenotypically distinguish copy number and point mutations. This is possible by specifically selecting for the increased dosage of a barely expressed, but otherwise intact endogenous *galK* gene of *E.coli*. The barely expressed *galK* gene corresponds to the innovation, which gets amplified and finally diverges by point mutations in the promoter region (see Fig. 4.1a).
Being part of a chromosomal reporter gene cassette, *galK* is transcriptionally fused to a *yfp* gene, such that adaptive point mutations in its promoter region can be detected as increases in YFP expression. Mutations to the copy number of the *galK* locus can be detected by an independently transcribed *cfp* gene (Steinrueck and Guet, 2017; Tomanek et al., 2020) (Figure 4.2a).
As the amplification hindrance hypothesis predicts copy number mutations to interfere with the fixation of point mutations we compare the divergence of the *galK* promoter region in two strains of *E.coli* that differ in their ability to form duplications of *galK*. Duplication rate can be manipulated, for instance, by deleting the *recA* gene involved in homologous recombination (Reams *et al.*, 2010). However, given its role in DNA repair, the comparison of strains with and without homologous recombination would be strongly influenced by the

growth defects such a mutation entails. To not have to consider pleiotropic effects caused by a difference in the genome-wide duplication rate, we compare two identical strains whose difference in duplication rate is restricted to a single genomic locus. To this end we are taking advantage of a chromosomal location that is characterized by high rates of duplication and amplification due to homologous recombination occurring between two endogenous identical insertion sequences (IS) elements that flank this locus. By deleting one copy of IS1, we created two otherwise isogenic strains of *E.coli* that differ in solely in the presence of one insertion sequence (IS) element approximately 10 kb downstream of *galK* (Fig. 4.2b), yet are predicted to show strong differences in their rates of duplication formation in this locus. We will refer to these strains as IS+ and IS- strains in the following.

4.2.3 Different sugar concentrations result in different enzyme expression requirements

We started by testing whether two basic premises of the amplification hindrance hypothesis are fulfilled in the context of our experimental system: First, do environments exist that differ in the level of *galK* expression required for maximal fitness (approximated here by bacterial growth rate) as shown in Fig. 4.1b?

Our experimental environment consists of liquid minimal medium containing amino acids as a basic carbon and energy source such that cells can grow even in the absence of *galK* expression (Fig. 4.2c – grey line). Adding galactose to this basic medium renders *galK* expression beneficial. By inducing *galK* expression stepwise using an arabinose-inducible promoter, we confirmed that growth rate increased along with *galK* expression and saturates at a certain expression level (Fig. 4.2c). Importantly, this expression level required for maximum growth is different for different galactose concentrations. This property makes it possible for us to study adaptation in different galactose concentrations representing regimes of different expression requirements.

Second, do point mutation(s) exists, which confer a fitness benefit greater than that of a gene amplification (as suggested by Figure 4.1c)? To test this assumption, we evolved gene amplifications for seven days in the three different galactose concentrations (1%, 0.1% and 0.01%) and compared their growth to promoter mutants, which we created by introducing either a single (strain D8c) or two (strain H5r) SNPs into the ancestral random p0 sequence, known confer *galK-yfp* expression (Tomanek *et al.*, 2020). Indeed, in all three galactose concentrations, both promoter mutants show higher growth rates than the amplified strains, which evolved in these respective environments for seven days (Fig. 4.2d).

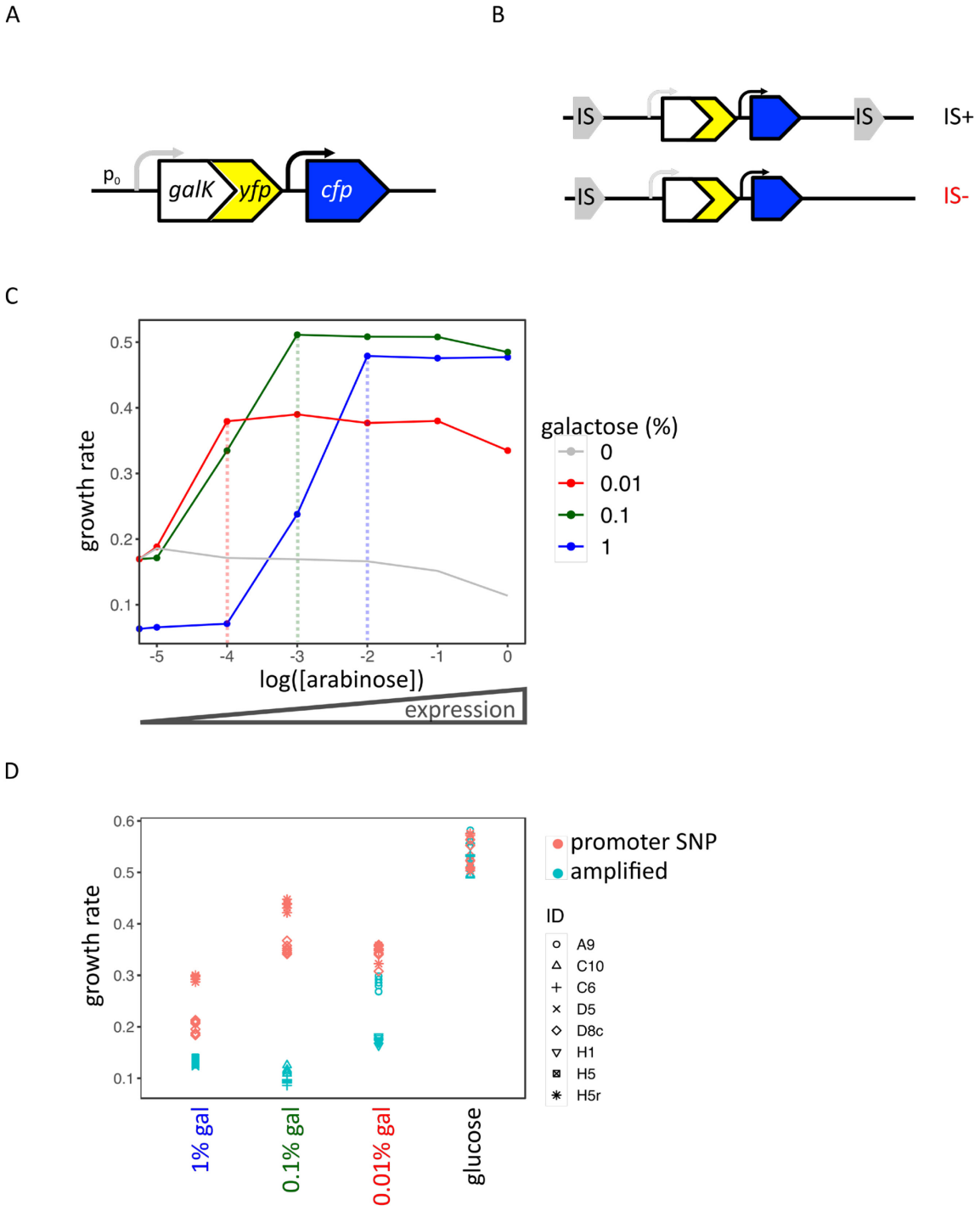


Figure 4.2 ***An experimental system to study the duplication and divergence in strains with different duplication rate.***

A. Scheme showing the chromosomal selection and reporter cassette. GalK does not conatin functional promoter; instead a random sequence (p0) drives very low levels of baseline gene expression.

B. *IS+ and IS- strain differ with respect to a copy of IS1C 12kb downstream of the selection and reporter cassette.*

***C**. Growth rate as a function of galK expression in four minimal media supplemented with 0.1% casaminoacids and different concentrations of galactose, respectively. Expression of p_{ara}-galK is induced by adding arabinose.*
***D**. Growth rate of strains with galK expression being conferred by copy number (amplified) and point (promoter SNP) mutations, respectively. Promoter mutants exhibit a significantly greater growth rate than amplified in all three galactose concentrations (p<0.0001, one-sided t-test).*

4.2.4 Evolution of *galK* expression in the IS+ and IS- strain

To directly test whether amplification speeds up the evolution by adaptive point mutation as predicted by the amplification mutagenesis hypothesis, we evolved 96 replicate populations of our IS+ and IS- strains in minimal medium containing only amino acids (control) or supplemented with three different galactose concentrations for twelve days, respectively (Fig. 4.3a). First, the evolution experiment confirmed that the strain lacking two flanking IS elements (IS-) indeed shows a strong reduction in the ability to undergo *galK* amplification: In contrast to the IS+ strain very few IS- populations evolved increased CFP expression (Fig. 4.3a – left panels). Interestingly, the maximum CFP fluorescence attained by IS+ populations differs for all three galactose concentrations (Figure 4.3b). While populations evolve increases in CFP fluorescence within a single day, they maintain this level relatively stably for the duration of the experiment. This difference in the number of *galK* copies might reflect the expression level required for the respective environments (Fig. 4.2c) and confirms that amplifications are a efficient way of tuning gene expression (Tomanek *et al.*, 2020).

To get an overview of the nature of adaptive mutations that occurred during evolution in the three galactose concentrations, we plotted individual populations´ YFP fluorescence as a proxy for *galK* expression against their CFP fluorescence as a proxy for *galK* copy number for all time points (Fig. 4.3c). Interestingly, populations in the YFP-CFP plot exhibit different patterns of distribution in all four environments, indicating that adaptation occurs via different routes.
In the lowest galactose concentration (0.01%), data points occupy the space along a diagonal axis between YFP and CFP fluorescence indicative of gene copy number mutations (“amplified fraction”, Fig. 4.3c – right upper panel). In the high galactose concentration (1%), data points occupy a second space, where YFP is increased relative to CFP (“mixed fraction”, Figure.3c - left upper panel). Based on these population-level data, we hypothesized that this phenotypic space is occupied either by a population of mixed mutants carrying a combination of point and copy mutations or by a mixed population consisting of cells with promoter mutations and cells with copy number mutations (Figure 4.4a). While SNPs and amplifications occurring in different genomic backgrounds are indicative of clonal interference and in agreement with the amplification hindrance hypothesis (“mutually exclusive” mutations), combinations of point and copy number mutations are predicted by

IAD. Hence, knowing the single cell genotype is crucial for distinguishing between the two alternative hypotheses. We will revisit this question below.

Interestingly, in the intermediate galactose concentration (0.1%) data points occupy a third space next to the two described above ("YFP+ fraction", Fig. 4.3c – lower left panel). These populations exhibit increased YFP fluorescence and ancestral CFP fluorescence indicative of promoter mutants. Intriguingly, only IS- populations occupy this space. To confirm that these populations with strongly increased YFP fluorescence ("YFP+ populations") are indeed promoter mutants, we sequenced the p0 region upstream of *galK* of five evolved clones. To our surprise, all of them harbored an ancestral p0 sequence, with only one of the purified clones showing a mixed signal in one position of the Sanger electropherogram indicative of ancestral and mutated sequences occurring in the same clone and, hence, in multiple copies. As re-streaking individual colonies on agar plates produced colonies of different YFP fluorescence intensity, we hypothesized that they carried an unstable amplification extending over *galK-yfp* but does not include *cfp* . Quantitative real-time PCR confirmed our suspicion (Fig. 4.4b). Since the IS- strain cannot undergo the frequent duplication between the two flanking IS elements, adaption might occur via a duplication that happens with lower frequency and has its junction somewhere between *yfp* and *cfp* (Fig. 4.4c). Unfortunately, this mutation hijacks our system to reliably detect copy number mutations. Moreover, the finding that amplifications exist even if a CNV-detection system fails to detect them tells a cautionary tale, namely that copy number mutations may underlie many adaptations, but unless we search for them directly, we will never know.

Given that the YFP+ fraction occurred only in the IS- strain and only in 0.1% galactose, we proceeded with our analysis nonetheless, having learned that only sequencing can confirm promoter mutations.

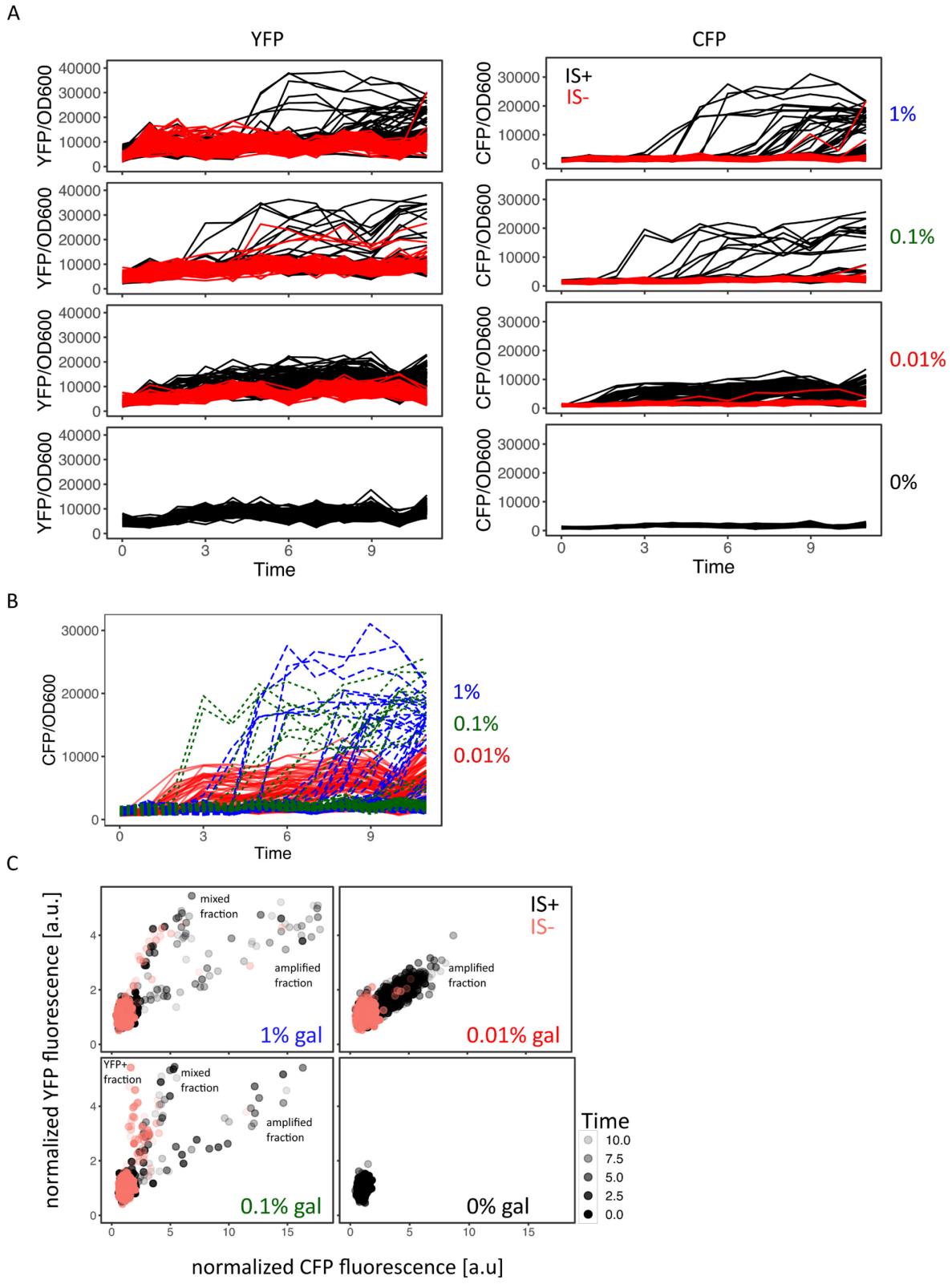


Figure 4.3 ***Evolutionary dynamics in different galactose concentration.***
A*. Daily measurements of normalized YFP and CFP fluorescence of 96 populations of IS+ (black) and IS- (red) strain growing in four different galactose concentrations (indicated to the right), respectively. Populations were grown in volumes of 200 µl, respectively, with daily dilution of 1:820 after the fluorescence measurement.*

***B**. Comparison of normalized CFP fluorescence for the IS+ strain in all three galactose environments (re-plotted from A).*
***C**. YFP - CFP plot of populations during evolution in different galactose concentrations (% indicated in the plot) normalized to the mean fluorescence of all populations from the control experiment in 0% galactose (re-plotted from A). Time points indicated by shading.*

4.2.5 Combination mutations occur in intermediate and high galactose concentrations

Next, we wanted to understand whether copy number and point mutation are mutually exclusive or occur as a combination after evolution in intermediate (0.1%) and high (1%) galactose. To this end, we determined the single-cell fluorescence of all mixed fraction populations (Fig. 4.3c) using flow cytometry. Overall, single cell fluorescence recapitulated the population measurements in that the YFP-CFP phenotype fell into three distinct fractions (Fig. 4.4d). It is worth noting that after twelve days of evolution cells with ancestral YFP and CFP fluorescence were still present in every single amplified population. In other words, while some populations consisted of a high fraction of cells with elevated CFP fluorescence, mutants did not yet spread to complete fixation in any of them. This is puzzling, given that once an increase in CFP was reached, populations showed relatively stable fluorescence for many days (Fig. 4.3b).
Flow cytometry results showed that IS+ populations of the mixed fraction from 0.1% galactose consisted of a single type with increased YFP/CFP fluorescence (Fig. 4.5b-c, left panels). If a population consisted of two mutually exclusive mutants, we would expect cells to fall into two distinct clusters. Moreover, YFP fluorescence of the mixed fraction cells was greater than YFP for pure amplification mutants, which falls along diagonal axis (Fig. 4.3c - right panel) again indicating a combined amplification and promoter mutation. Sequencing of three amplified (increased CFP) and single copy (ancestral CFP) colonies, respectively, from mixed fraction populations evolved in 0.1% galactose revealed that all of them harbor a SNP (T>A) in p0 only in the amplified clones, confirming the combination mutation (Fig. 4.5a).
Like in intermediate galactose, IS+ populations from the high (1%) galactose mixed fraction harbored cells with the combination mutation phenotype and, in addition, cells with pure amplifications (Fig.4.5b-c, right panels).

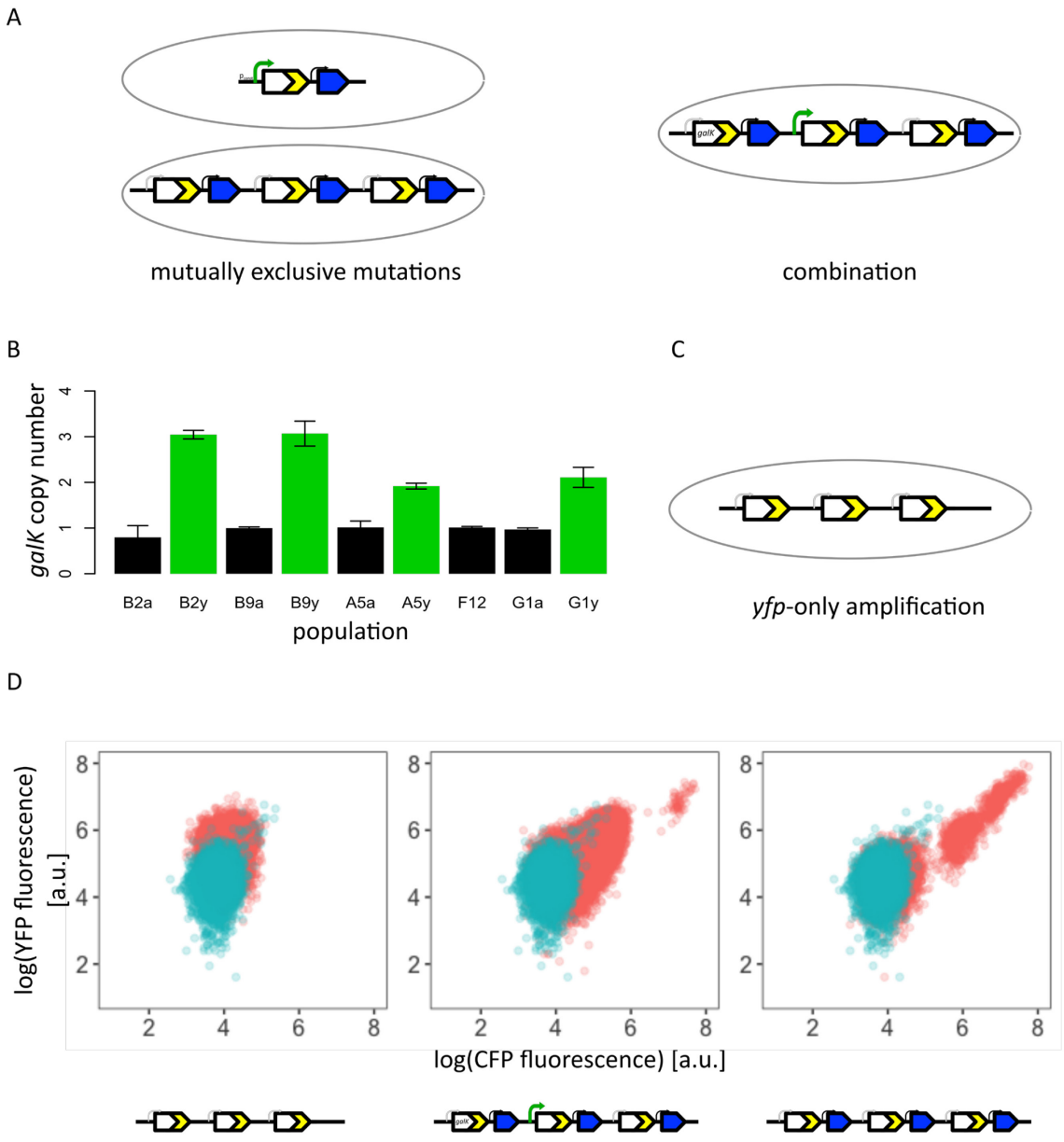


Figure 4.4. ***Genotypes of evolved clones.***
A. *Scheme of genotypes of a mixed population ("mutually exclusive mutations") or a mixed mutant ("combination").*
B. *GalK copy number of YFP+ IS- populations evolved in 0.1% galactose as estimated by qPCR. For each population, genomic DNA of one colony with ancestral ("a") and one with increased YFP ("y", green bars) fluorescence was analyzed.*
C. *Scheme of yfp-only amplification with a duplication junction upstream of cfp.*
D*. Representative flow cytometry plots showing single-cell YFP and CFP fluorescence of a population from the YFP+ (left), mixed (middle) and amplified (right) fraction of Fig.3C.*

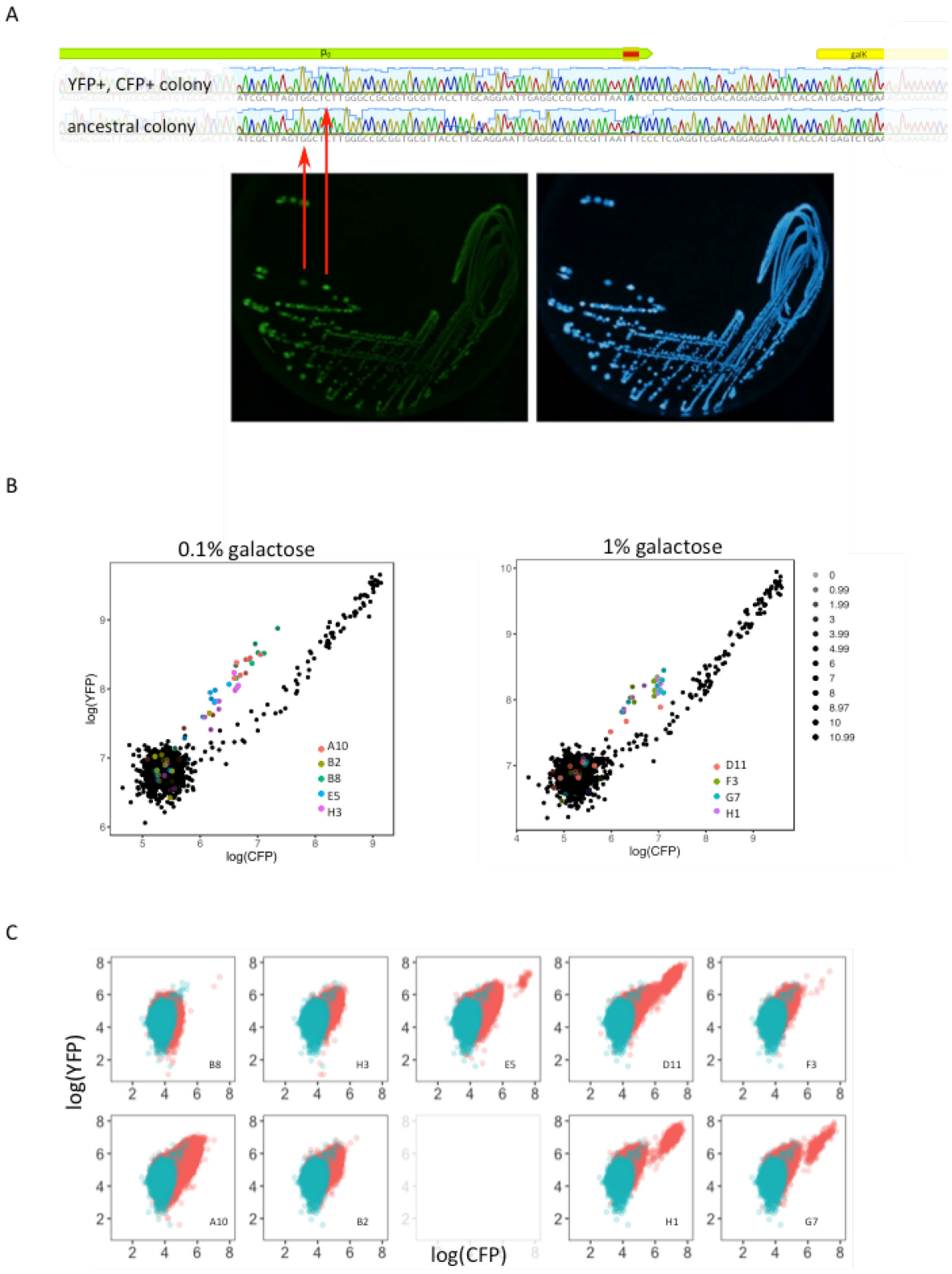


Figure 4.5 ***Confirming combined copy number and point mutations in intermediate and high galactose.***

A*. Sanger sequencing electropherogram of a p0 sequence with a T>A point mutation in an amplified but not an ancestral colony of IS+ population A10 evolved for 12 days in 0.1% galactose. Photographs show YFP (left) and CFP (right) fluorescence of population A10 streaked on LB agar. Red arrows indicate sequenced colonies.*

***B**. Log YFP-CFP plot of IS+ populations from 0.1% and 1% galactose, respectively. Colored points identify mixed fraction populations, which exhibit single-cell fluorescence phenotype of combination mutations in flow cytometry data (**C**).*

4.2.6 Mutually exclusive mutations occur in low galactose concentration

Next, we analyzed the single cell fluorescence of cells from low (0.01%) galactose. Consistent with population measurements, most IS + populations adapted via gene amplification (Fig. 4.6a – left panels). Interestingly, some populations did not follow this general trend (Fig. 4.6a – middle panels). Instead they showed an increase in YFP without concomitant increase in CFP not visible in the population measurements in liquid cultures (Fig.4.6b – right panel). However, intriguingly, elevated YFP fluorescence was also visible after patching populations onto LB agar, which alleviates any changes in fluorescence related to growth-rate (Fig. 4.6b - left panel). We examined population B1 more carefully by re-streaking it on agar. Consistent with flow cytometry results (Fig. 4.6a - right panel) we found three types of colonies: ancestral fluorescence, YFP+, and a small subpopulation of amplified cells. Sequencing of the amplified colony confirmed it is a *bona fide* amplification without additional SNPs. Sequencing of the YFP+ colony revealed two SNPs in p0. Given that we detected this promoter mutant only after careful analysis, fluorescence measurements in evolving populations (e.g. Fig. 4.6b - right panel) might not be sufficiently sensitive to detect point mutations in a single copy background.
Since even slightly increased YFP might indicate point mutations, we sequenced YFP+ clones from IS- populations. We found SNPs in some but not others. At the same time, qPCR confirmed single-copy states in all of the clones with slightly increased YFP. Therefore, some of the slight increases in YFP might be due to mutations outside (further upstream of) p0. In order to quantify the fraction of promoter SNPs evolved in IS- versus IS+ strains despite these complications, we want make use of amplicon deep sequencing of p0 in all evolved populations. Results are being generated while writing.

Encouraged by the fact that we fail to see combination mutants (i.e. a mixed fraction) in population measurements from low galactose (Fig. 4.3c), we used the agar patches to screen for IS+ populations with elevated YFP but not CFP (judging by eye). Restreaking, sequencing and flow cytometry analysis revealed that like population B1 (Fig.4.6a) all other populations harbored either a mixed population of few *bona fide* amplified cells and a majority of promoter mutants or only promoter mutants (Table 4.1). As opposed to high and intermediate galactose, we did not find a single population with combined mutants in low galactose. The fact that mutations were mutually exclusive within populations was also mirrored when analyzing the fate of different populations over time. Cursory analysis of different time points of patched populations (see Fig. 4.6b –left and middle panel for d12) suggested YFP+ colonies never became amplified. Quantitative analysis of the fluorescence intensity of patched populations using an automated script (Fig. 4.6c) confirmed that populations with a significant frequency of promoter mutants (i.e. visibly YFP+ in the agar patch) did not become amplified. As one exception, population F6 gained the YFP+

phenotype early but became dominated by a *bona fide* amplification by the end of the experiment (Fig.4.6c – bottom right panel, blue triangle). A second outlier from this apparent rule, population C9, was found to have been incorrectly classified as YFP+ (Table 4.1). Conversely, YFP+ populations evolved exclusively from those with ancestral phenotype, no single amplified population gained a promoter within the time of the experiment.

Table 4.1. ***Sequencing and phenotypic analysis of IS+ populations evolved in 0.01% galactose.*** *Increase in fluorescence relative to the ancestral (anc.) phenotype indicated by Y+(YFP) and C+(CFP). Day 12 populations were characterized unless otherwise noted (d4, d8).*

Population	**seq (Y+)**	**FACS phenotype**	**agar streak**	**comment**
A6	T>A	Y+, v.f.C+	Y+, v. f. C+	
B1	T>A,C>T ("H5r")	Y+, C+ (two populations)	f. Y+, f.C+, mixed pop	
B2	T>A	Y++	Y+, v.f.C+	
C1	T>A	Y+ (d12)	Y+, v.f.C+	
C9	?	clearly ancestral Y (d8), only C+(d12)	?	incorrectly classified as Y+
D2	T>A	Y+ (d12)	Y+ only	
D9	anc	Y+ (d8,d12)	Y+ only	
E9	anc	ancestral	?	
E10	T>A	Y+ (d12)	Y+ only	
F6	?	Y+ (d4), C+(d12), no combined mutant	?	Y+ at d8, then amplified population
F10	T>A	Y+,C+, anc	Y+,C+,mixed pop	
G1	T>A	Y+(d4-8),C+ (d12)	Y+, v.f.C+	
G12	T>A	Y+ (d8), C+ (d12)	Y+, only	FACS C+ carry-over

We conclude by saying that data analysis is complicated by the fact that it requires a lot of low-throughput work, for instance, to determine the genotype of ambiguous cases where every method used has their advantages and drawbacks: i) population patch image analysis can be automated and profits from equalized growth rates but can be biased by colony morphology or reflection artifacts, ii) flow cytometry allows determining sub-populations but is biased by small growth rate differences/time delays when processing many samples; iii) Sanger-sequencing is very low throughput and too expensive and labor-intensive if doing for all sub-populations, iv) quantitative real-time PCR is very low throughput (few populations per day) but is the best way to determine *bona fide* copy number mutations.

While we are yet to analyze amplicon deep sequencing data of p0 to compare the level of divergence for the IS+ versus IS- strain, we can already observe two interesting differences between evolution low and high/intermediate galactose concentrations.

In low galactose we see mutually exclusive mutations within populations and on the level of population phenotypes over time; we failed to detect a single instance of a combined mutation as for intermediate or high galactose, indicating that clonal interference occurs between copy number and point mutations. This finding is in line with our amplification hindrance hypothesis, which predicts combinations of copy and point mutations to occur when there is a lot of room for improving gene expression, and either or mutation to occur when there is little room for improving gene expression.

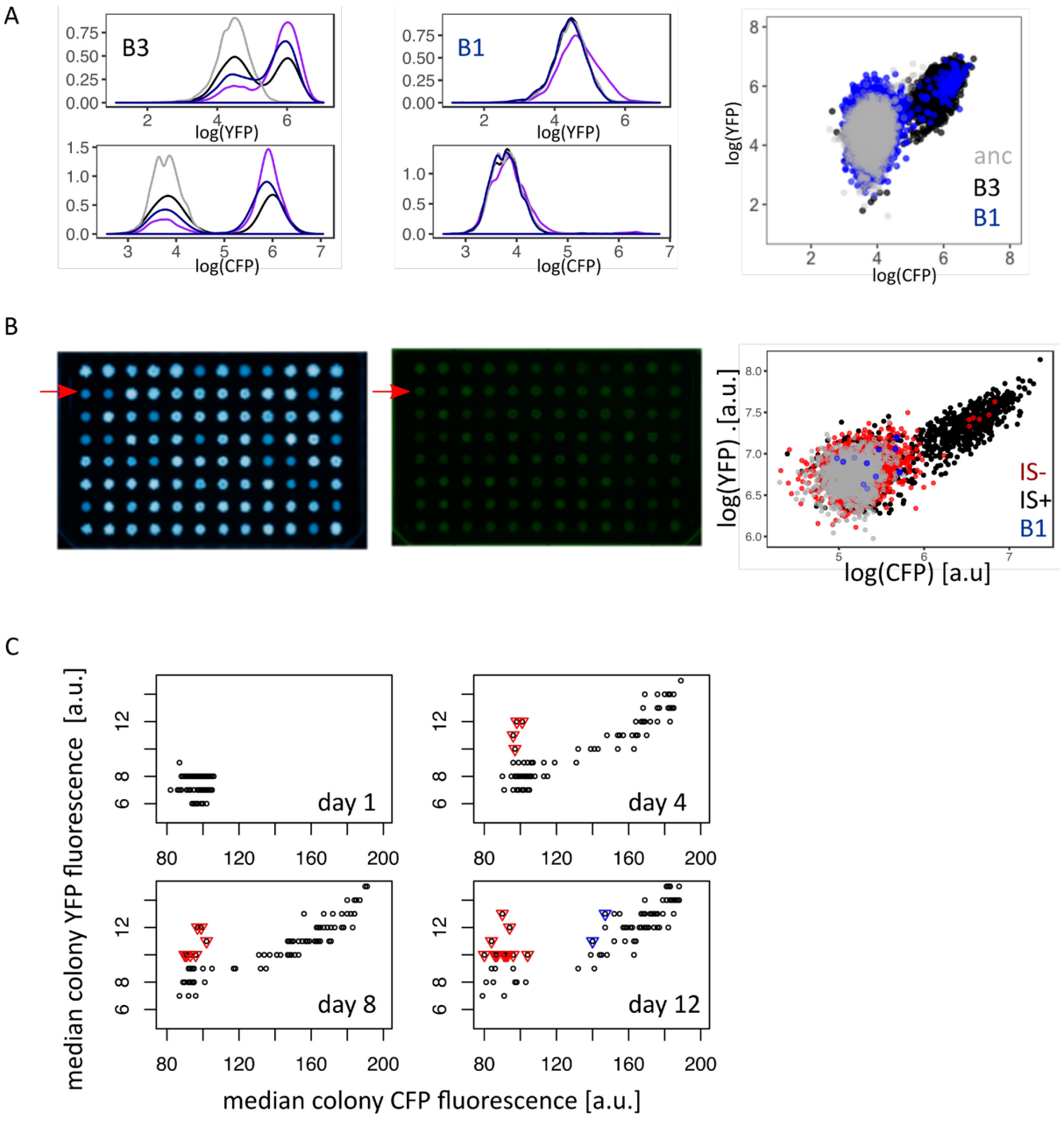


Figure 4.6 ***Confirming mutually exclusive mutations in low galactose.***
A. *Flow cytometry data showing log(YFP) fluorescence (upper left and middle panel) and log(CFP) fluorescence (lower left and middle panel) of population B3 (left panels) and B1*

(middle panels) over time (grey-ancestral, black–d4, dark blue-d8, purple-d12). Right panel shows the same data for population B3 and B1 in a log(YFP) versus log(CFP) plot to visualize two distinct sub-populations in B1 (blue).
B. *CFP (left panel) and YFP (middle panel) fluorescence of populations patched onto LB agar supplemented with charcoal. Red arrows indicate population B1, which exhibits increased YFP but ancestral CFP fluorescence. The same population, B1, is highlighted in the log(YFP) versus log(CFP) plot measured on evolving populations.*
C*. Automatic classification of YFP+ fraction from patched populations such as those shown in (B) using pixel intensities and a threshold based on ancestral fluorescence values: red triangles represent populations that kept the YFP+ phenotype they exhibited on earlier time points or evolved YFP+ from an ancestral phenotype. Blue triangles: previously classified as YFP+ populations now classified as amplified (note that flow cytometry-analysis showed that population C9 was incorrectly classified as YFP+ based on colony fluorescence.)*

4.2.7 Evolutionary dynamics differ for different random p0 sequences

Given the paucity of point mutations we observed for the evolution of our random p0 sequence (one evolved clone with two SNPs identical to the “H5r" mutation from Fig. 4.2c, otherwise all evolved clones had a single T>A SNP, also part of the H5r mutation) and the relatively small phenotypic effect in terms of YFP+ increase, we wondered whether higher number of different mutations could be obtained when using a different random p0 sequence as a starting point for evolution. Therefore, we repeated our evolution experiment in intermediate (0.1%) galactose with three additional random promoter sequences (p0-1, p0-2, p0-3).

In our ten-day evolution experiment only two out of four random p0 sequences, evolved increased *galK-yfp* expression (Fig. 4.7a). This is roughly consistent with the fact that approximately 60% of random sequences are one point mutation away from a relatively strong constitutive promoter (Yona, Alm and Gore, 2018; Lagator *et al.*, 2020). Interestingly, p0-1 and p0-3 did not even gain gene duplications or amplifications. On a first glance, this drastic difference in amplification was not expected, since the IS+ strains only differ in their p0 sequence, not in their duplication rate. However, random sequences differ in their ability to recruit RNA-polymerase and their resulting baseline expression level (Yona, Alm and Gore, 2018; Lagator *et al.*, 2020). Given that a plateau exists in the expression-growth relation for low levels of expression (Fig. 4.2c), the initial expression level conferred by p0-1 and p0-3 might be too low to yield a selective benefit upon duplication alone. According to this hypothesis, these random (non-)promoters are not only two (or more) point mutations away from a beneficial sequence, but also two (or more) copy number mutations.

For p0, the evolution experiment in intermediate galactose reproduced our previous findings, namely a mixed and amplified fraction for IS+ populations and a YFP+ fraction for IS- populations (compare Fig. 4.7a with Fig. 4.3c - lower left panel). Like for the previous experiment, sequencing of six evolved clones from different IS- populations with strong increases in YFP/CFP fluorescence (Fig. 4.7b – upper right panel) revealed an ancestral p0 (Table 4.2), suggesting that the underlying genotype is an amplification of YFP, but not CFP.

For p0-2, evolutionary dynamics differed drastically to p0. In the IS+ strain, almost every single population evolved amplifications within in the first two days of the evolution experiment (Fig. 7a – compare upper and lower panels in the left column). Interestingly, only two fractions are visible in the YFP-CFP plots of p0-2. The first fraction is occupied by YFP+ populations carrying a single copy of CFP. The second fraction along the diagonal between YFP and CFP is most likely occupied by amplified populations without point mutations. As expected, this amplified fraction is smaller in the IS- strain. Moreover, it is shifted towards higher values of YFP/CFP, suggesting that p0-2 exhibits a higher baseline expression level than all three other random promoter sequences. Accordingly, this "advanced starting position" in the expression-growth relation might explain the rapid amplification dynamics of p0-2 populations. Further experiments are needed to confirm our suspicion that the initial expression level correlates with the ability to evolve *galK* expression.

Intriguingly, those p0-2 IS+ populations that failed to evolve amplifications show an increase in YFP/CFP early in the evolution experiment (Fig.4.7b – lower left panel). This result combined with the idea that p0-2 exhibits a relatively high baseline expression level and the fact that mixed fractions are missing for p0-2 (Fig. 4.7a) suggests that increases in expression evolve either via amplification or point mutation. In other words, because initial *galK* expression is high in p0-2, the expression requirement can be reached with few amplifications or one point mutation resulting in a situation of mutually exclusive mutations (Fig. 4.4a).

In contrast to the IS+ strain where only six populations evolved increased YFP/CFP fluorescence only within the first three days of the experiment, IS- populations were evolving increased YFP/CFP fluorescence throughout the experiment (Fig. 4.7b –lower right panel). We were curious whether the increase in YFP/CFP in both IS+ and IS- populations was due to promoter mutations. Sequencing revealed that in most cases (4/6 for IS+, 5/6 for IS-) clones with increased YFP/CFP indeed harbored a mutation in p0-2 (Table 4.2). We confirmed that the 12bp deletion mutation, the 13bp deletion mutation and the SNP were in fact adaptive, by reconstituting these mutations into the ancestral p0-2 strain, where they conferred the ability to grow on 0.1% galactose agar.

Table 4.2 ***Sequencing of p0 and p0-2 of clones of IS+ and IS- populations evolved in 0.1% galactose.***

IS + clones		**IS - clones**	
p02-A11	13bp deletion	**p02-G1**	ancestral
p02-E9	ancestral	**p02-F2**	12 bp deletion
p02-B10	12bp deletion	**p02-A7**	C>T
p02-F4	C>T	**p02-H12**	C>T
p02-E1	seq failed	**p02-C3**	C>T
p02-C4	ancestral	**p02-C5**	p0-2 almost fully deleted (19bp left)
p02-F4	C>T, poor quality read	**p02-C11**	seq failed
p0-C6	mix of CFP deletion + anc.	**p0-B5**	ancestral

p0-B8	ancestral
p0-C8	ancestral
p0-H4	ancestral
p0-F2	ancestral
p0-H5	ancestral

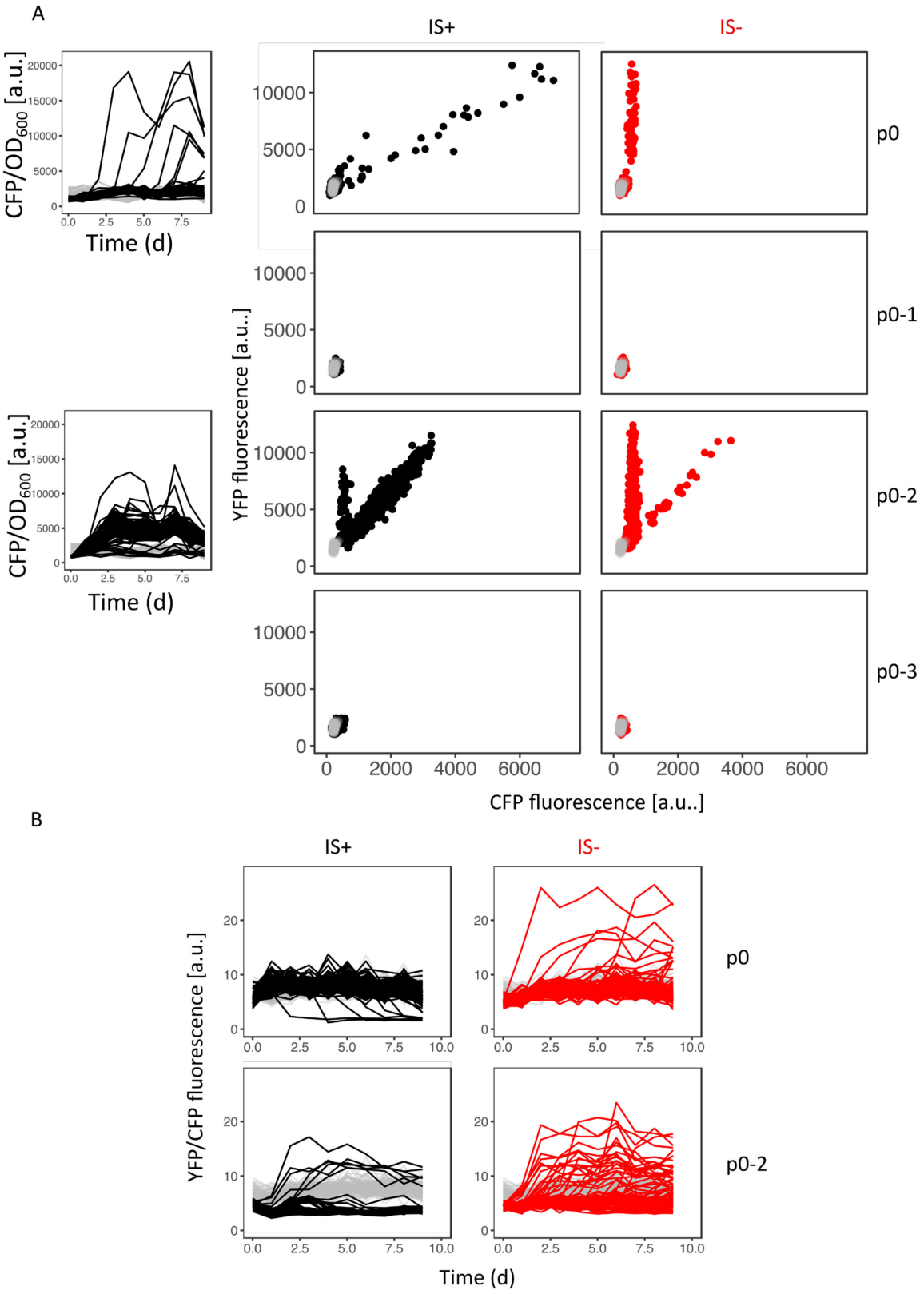


Figure 4.7. ***Evolutionary dynamics for different random p0 sequences in 0.1% galactose.*** ***A.*** *Left panels: CFP/OD_{600} plotted over the course of the evolution experiment for control (grey), IS+/p0 (upper panel) and IS+/p0-2 (lower panel) populations. Middle and right panels*

show YFP - CFP summary plot of population-level measurements of all time points of during evolution of four different promoter sequences (p0, p0-1, p0-2, p0-3, indicated to the right) of IS+ (black) and IS- (red) strains.
***B**. YFP/CFP fluorescence over time for p0 and p0-2 populations of IS+ (black) and IS- (red) (re-plotted from A). Gene amplifications are visible as decrease in YFP/CFP relative to the 0% galactose control (grey).*

4.3 Discussion

Here, we introduced the amplification hindrance hypothesis as an extension to the IAD model and as an antithesis to the amplification mutagenesis hypothesis implied in the IAD model and widely cited by literature (Elde et al., 2012; Yona, Frumkin and Pilpel, 2015; Cone et al., 2017; Bayer, Brennan and Geballe, 2018; Lauer et al., 2018; Todd and Selmecki, 2020) concerned with the adaption by copy number mutations.

While the *galK* selection and dual fluorescence reporter cassette lends itself to study the dynamics of point and copy number mutations, this project also revealed two shortcomings of the experimental system. First, point mutations in p0 may increase *galK-yfp* expression not sufficiently strongly to be detected by population-level measurements conducted with a platereader. Second, duplications and amplifications can occur between *yfp* and *cfp*, mimicking the phenotype of a promoter mutation. Both shortcomings make it hard to unequivocally count the number of adaptive point mutations in the IS+ and IS- strain, which would be the one metric to falsify either the amplification mutagenesis or amplification hindrance hypothesis. To overcome these limitations, we are planning to analyze deep sequencing data from p0 and p0-2 of all populations. We expect that the resulting data will allow to simply score which strain background (IS+ or IS-) carries more point mutations in its p0 region.

However, despite the limitations of the experimental system and the fact that this chapter is still work in progress, we can draw several conclusions from existing results.

First, the evolutionary dynamics of evolving gene expression strongly depends on the sugar concentration used in selection experiments (providing different fitness landscapes – Fig. 4.2c), as well as on the local rate of duplication (IS+ versus IS- strain) and the nature of the random p0 sequence upstream of *galK*.

Second, the kind of mutations observed in populations of the IS+ strain depends on the expression requirement of the galactose concentration used. Consistent with the IAD model, we observe cells with combinations of point and copy number mutations in high and intermediate levels of galactose. However, in low-level galactose we observe mutually exclusive mutations, that is, populations consisting of cells carrying either gene amplifications or point mutations, but not both. This is in agreement with the amplification hindrance hypothesis, which predicts clonal interference between the two mutation types in a regime that only requires low levels of expression for maximal growth.

We found more direct evidence for the amplification hindrance hypothesis when selecting for increased expression starting from a different random p0 sequence. IS+ and IS- strains

carrying p0-2 differed in their response to selection in that IS+ populations almost exclusively evolved amplifications within the first two days of the experiment. Intriguingly, those populations that did not evolve amplifications, showed increased YFP fluorescence that was in most cases conferred by adaptive point and small-scale deletion mutations in p0-2. The fact that IS+ populations carrying p0-2 either fix promoter or amplification mutations when evolving in intermediate galactose (Fig. 4.7b – lower left panel) mirrors the situation of IS+ populations carrying p0 and evolving in low galactose (Fig. 4.6). However, the most direct evidence for amplification hindrance to occur in our experiments comes from comparing the number of YFP+ populations between p0-2 of IS+ and IS- strains. Importantly, unlike p0, in p0-2 most IS- clones with increased YFP fluorescence harbored adaptive point and small scale deletion mutations in their p0 sequence (Table 4.2).
It is tempting to speculate that these adaptive small-scale deletions may be sufficiently frequent to change the dynamics from what is observed for p0. These promoter mutations may be more frequent than the *yfp*-only amplifications observed in p0 IS- populations and more frequent than point mutations. If they occur with a frequency comparable to gene amplifications (formed between IS elements), we would expect to see strong clonal interference – which seems to be the case in IS+ populations that evolve either amplification or point mutation. However, at this point, caution is warranted as we still need to confirm, that i) most other YFP+ populations in IS- are in fact mutations in the promoter and ii) that the amplifications in the IS+ strain do not contain promoter mutations on top of their amplifications.

4.4 Methods

4.4.1 Bacterial strain construction

To construct the IS- strain, we replaced the second copy of IS1 downstream of the selection and reporter cassette in IT030 (Tomanek *et al.*, 2020) with a kanamycin cassette using pSIM6-mediated recombineering (Datta, Costantino and Court, 2006). Recombinants were selected on 25µg/ml kanamycin to ensure single-copy integration.
To create the additional random promoters sequences p0-1, p0-2 and p0-3, we generated 189 nucleotides using the "Random DNA sequence generator" (https://faculty.ucr.edu/~mmaduro/random.htm) with the same GC content as p0 (55%). We ordered these three sequences as gBlocks (integrated DNA technology, BVBA, Leuven, Belgium) with attached XmaI and XhoI restriction sites, which we used to clone p0-1, p0-2 and p0-3 into plasmid pMS6* (Tomanek et al., 2020) by replacing p0. We used pMS6* with the respective p0 sequence as a template to amplify the selection and reporter cassette and integrate it into MS022 as described previously (Tomanek et al., 2020).

>p0
ACCGGAAAGACGGGCTTCAAAGCAACCTGACCACGGTTGCGCGTCCGTATCAAGATCCTCTTAATAA
GCCCCCGTCACTGTTGGTTGTAGAGCCCAGGACGGGTTGGCCAGATGTGCGACTATATCGCTTAGTG
GCTCTTGGGCCGCGGTGCGTTACCTTGCAGGAATTGAGGCCGTCCGTTAATTTCC

>p0_1
GTAGGCCCGCACGCAAGACAAACTGCTGGGGAACCGCGTTTCCACGACCGGTGCACGATTTAACTT
CGCCGACGTGACGACATTCCAGGCAGTGCCTCCGCCGCCGGACCCCCCTCGTGATCGGGTAGCTGG
GCATGCCCTTGTGAGATATAACGAGAGCCTGCCTGTCTAATGATCTCACGGCGAAAG

>p0_2
TCGGGGGGACAGCAGCGGCTGCAGACATTATACCGCAACAACACCAAGGTGAGATAACTCCGTAGT
TGACTACGCGTCCCTCTAGGCCTTACTTGACCGGATACAGTGTCTTTGACACGTTTGTGGGCTACAGC
AATCACATCCAAGGCTGGCTATGCACGAAGCAACTCTTGGGTGTTAGAATGTTGA

>p0_3
CCCCTGTATTTGGGATGCGGGTAGTAGATGAGCGCAGGGACTCCGAGGTCAAGTACACCACCCTCT
CGTAGGGGGCGTTCCAGATCACGTTACCACCATACCATTCGAGCATGGCACCATCTCCGCTGTGCCC
ATCCTGGTAGTCATCATCCCTATCACGCTTTCGAGTGTCTGGTGGCGGATATCCCC

Table 4.3 ***List of strains used***

Strain name	**Genotype**	**purpose**	**Source**
MG1655	F^- λ^- ilvG- rfb-50 rph-1	strain background for all evolution experiments	lab collection
IT013-TCD	BW27784, JA23100::*galP* , *mglBAC* ::FRT, *galK* ::FRT, locus1::pBAD-*galK*	strain with pBAD-*galK* for testing expression-growth relation	Tomanek et al., 2020
BW25142	lacIq rrnB3 (lacZ4787 hsdR514 DE(araBAD)567 DE(rhaBAD)568 (phoBR580 rph-1 galU95 (endA9 uidA((MluI)::pir-116 recA1	host for pir plasmid pMS6*	Khlebnikov et al., 2001
MS022	MG1655, JA23100::*galP* , *mglBAC* ::FRT, *galK* ::FRT	IS+ background for ancestor strain construction	lab collection
IT030	MS022 locus2::p0-RBS-*galK*-RBS-*yfp*-FRT-pR-*cfp*	IS+ ancestor strain	Tomanek et al., 2020
IT049	MS022 deleted for IS1C	IS- background for ancestor strain construction	this study
IT049-p0	IT049 locus2::p0-RBS-*galK*-RBS-*yfp*-FRT-pR-*cfp*	IS- ancestor strain p0	this study
IT049-p01	IT049 locus2::p0-1-RBS-*galK*-	IS- ancestor strain p0-1	this study

	RBS-*yfp*-FRT-pR-*cfp*		
IT049-p02	IT049 locus2::p0-2-RBS-*galK*-RBS-*yfp*-FRT-pR-*cfp*	IS- ancestor strain p0-2	this study
IT049-p03	IT049 locus2::p0-3-RBS-*galK*-RBS-*yfp*-FRT-pR-*cfp*	IS- ancestor strain p0-3	this study
IT030-H5r	MS022 locus2::pconst-RBS-*galK*-RBS-*yfp*-FRT-pR-*cfp*	strain with constiutive *galK* expression conferred by two SNPs in p0	Tomanek et al., 2020
IT030-D8c	MS022 locus2::pconst-RBS-*galK*-RBS-*yfp*-FRT-pR-*cfp*	strain with constiutive *galK* expression conferred by one SNP in p0	Tomanek et al., 2020

Table 4.4 ***List of primers used***

Name	**Sequence**	**Purpose**
E_flank_f	GCTGGAGCCACTTGTAGCC	cassette integration test locus 2, sequencing p0s
E_flank_r	TCCTTGCTGAATCATTTTGTTC	cassette integration test locus 2
p0_check_Fw	GTGTGAGTGGCAGGGTAG	sequencing p0s
qPCR_*galK*_Fw	GCTACCCTGCCACTCACA	estimating *galK* copy number
qPCR_*galK*_Rv	CGCAGGGCAGAACGAAAC	estimating *galK* copy number
rbsB_qPCR_Fw	GGCACAAAAATTCTGCTGATTAA	qPCR control locus
rbsB_qPCR_Rv	GCAGCTCGATAACTTTGGC	qPCR control locus
P1_p01	GCCTTAGTTGTAAGTGTCTACCATGTCCCCGAACAAGTGTTCACTATGTCTAGGCCCGCACGCAAGAC	integration of the selection and reporter cassette with p01 (Fw primer)
P1_p02	GCCTTAGTTGTAAGTGTCTACCATGTCCCCGAACAAGTGTTCACTATGTCTCGGGGGGACAGCAGCG	integration of the selection and reporter cassette with p02 (Fw primer)
P1_p03	GCCTTAGTTGTAAGTGTCTACCATGTCCCCGAACAAGTGTTCACTATGTCTGTATTTGGGATGCGGGTAGTAGA	integration of the selection and reporter cassette with p03 (Fw primer)
E_int_Rv	TCGGAAGGGAAGAGGGAGTGCGGGAAATTTAAGCTGGATCACATATTGCCGAGGCCTTATGCTAGCTTC	integration of the selection and reporter cassette (Rv primer)
E_int_Fw	GCCTTAGTTGTAAGTGTCTACCATGTCCCCGAACAAGTGTTCACTATGTCACCGGA-AAGACGGGCTTC-----	integration of the selection and reporter cassette with p0 (Fw primer)

4.4.2 Evolution experiments

Evolution experiments were inoculated with ancestral colonies of IS+ and IS- strain grown in 3 ml of LB medium over night, after two washing steps in M9 buffer and a dilution of 1:200. All evolution experiments were conducted in M9 medium supplemented with 2 mM MgSO4, 0.1 mM CaCl2, 0.1% casaminoacids and a carbon source at the indicated concentration (all Sigma-Aldrich, St. Louis, Missouri). Bacterial cultures were grown in 200µl liquid medium in 96-well plates and shaken in a Titramax plateshaker (Heidolph, Schwabach, Germany, 750 rpm). Every day, populations were transferred to fresh plates using a VP408 pin replicator (V&P SCIENTIFIC, INC., San Diego, California) resulting in a dilution of ~ 1:820 (Steinrueck and Guet, 2017). Immediately after the transfer, growth and fluorescence measurements were performed in in the overnight plates using a Biotek H1 platereader (Biotek, Vinooski, Vermont).

4.4.3 Flow cytometry experiments

Frozen evolved populations (-80°C, 15% glycerol) from day 4, day 8 or day 12 (as indicated in the figures) were pinned (1:820) into M9 buffer and put on ice until the measurement. Fluorescence was measured using a BD FACSCanto™ II system (BD Biosciences, San Jose, CA) equipped with FACSDiva software. Fluorescence from the Pacific Blue channel (CFP) was collected through a 450/50nm band-pass filter using a 405nm laser. Fluorescence of the FITC channel (YFP) was collected through a 510/50 band-pass filter using a 488nm laser. The bacterial population was gated on the FSC and SSC signal resulting in approximately 6000 events analyzed per sample, out of 10,000 recorded events

4.4.4 Quantitative real-time PCR

For qPCR, gDNA was isolated from overnight cultures grown in the respective evolution medium inoculated by single evolved colonies using Wizard Genomic DNA purification kit (Promega, Madison, Wisconsin). We performed qPCR using Promega qPCR 2x Mastermix (Promega, Madison, Wisconsin) and a C1000 instrument (Bio-Rad, Hercules, California). To quantify the copy number of samples of an evolving population, we designed one primer pair within *galK* (target) and one primer within *rbsB* as a reference, which lies outside the amplified region. We compared the ratios of the target and the reference loci to the ratio of the same two loci in the single copy control. Using dilution series of one of the gDNA extracts as template, we calculated the efficiency of primer pairs and quantified the copy number of *galK* in each sample employing the Pfaffl method, which takes amplification efficiency into account (Pfaffl, 2001). qPCR was done in three technical replicates.

4.4.5 Measurement of colony fluorescence

Evolving populations were pinned onto LB agar supplemented with 1% charcoal and imaged using the macroscope set up. To obtain median colony YFP and CFP fluorescence intensity, a region of interest was determined using the ImageJ plugin 'Analyze Particles' (settings: 200px-infinity, 0.5-1.0 roundness) to identify colonies on 16-bit images with threshold adjusted according to the default value. The region of interest including all colonies was then used to measure intensity.

5 Conclusions

This thesis focused on the transient evolutionary dynamics intrinsic to copy number mutations.

Chapter two was concerned with the transient dynamics of duplication and deletion that may play out before the evolution of canonical gene regulation. The fact that gene duplications are several orders of magnitude more frequent than point mutations (10^{-6} – 10^{-2} per cell per generation) in bacteria and highly reversible renders adaptation by copy number mutation transient by nature (Andersson and Hughes, 2009). Indirect measurements have estimated that once a duplication has formed it is deleted or further amplified at a rate between 10^{-3} and 10^{-1} per cell per generation (Mats E. Pettersson *et al.*, 2009; Reams *et al.*, 2010). Our microfluidics-based results corroborate this finding. While both, the high rates of formation and the genetic instability of amplifications were widely appreciated before, our evolution experiment was the first one to demonstrate the resulting evolutionary dynamics in fluctuating environments. In our manuscript we speculated about the fact that AMGET may be found in nature and especially clinical settings, where harsh and potentially fluctuating selection is prevalent. A very recent publication showed that this is in fact the case: clinical isolates of *Staphyloccocus aureus* (Uhlemann *et al.*, 2014) use gene amplification to tune the expression level of *csa1*, a gene with immunostimulatory capacity, during infection and thereby modulates the host immune response (Belikova et al., 2020).

Given the high rates of duplication and deletion and the fact that gene expression changes along with copy number, it should not be surprising that amplification-mediated gene expression tuning is actually occurring in nature. In fact, the list of ways in which bacteria use “specialized” mutations to change their phenotypes in a stochastic or more “programmed” manner is long and fascinating (Moxon *et al.*, 1994; Moxon, Bayliss and Hood, 2006). The expression of pili in *E.coli* is controlled by an enzymatically catalyzed inversion of a promoter fragment (Hung et al., 2014). Clinical isolates of *Streptococcus pneumoniae* and *Staphylococcus aureus* switch the expression of virulence factors via frequently occurring large-scale inversions with defined break points on opposite arms of the chromosome (Cui et al., 2012; Slager, Aprianto and Veening, 2018). *Streptomyces* species carry linear chromosomes, which are highly unstable. Incidentally, most genes required for the production of antibiotics and other secreted compounds are confined to the unstable chromosome arms (Leblond and Decaris, 1994; Chen et al., 2002; Traxler and Kolter, 2015). Other examples include pathogens which carry an increased copy number of virulence factor only when isolated from their human host: this is true for the cholera toxin of *Vibrio cholerae*, (Goldberg & Mekalanos, 1986) and the *cap* locus of *Haemophilus influenzae* (Moxon et al, 1994). Worldwide isolates of *H. influenza* carry an asymmetric tandem duplication of the *cap* locus, which activate it for duplication (Moxon *et al.*, 1994). *Cap* codes for the production of capsular polysaccharides and is amplified during invasive infection but not in commensal isolates (Corn *et al.*, 1993; Davis *et al.*, 2011).
In the above examples, second-order selection seems to have shaped the genomic loci for an increased mutation rate, allowing the establishment of polymorphic populations for first-order selection to act on. For instance, contingency loci such as *cap* are thought to have evolved adaptively in pathogenic bacteria (Moxon *et al.*, 1994) to allow switching between

genotypes at rates between 10^{-5} to 10^{-2} per generation (Moxon, Bayliss and Hood, 2006).
In principle, it is conceivable that the genome-wide rate of homologous recombination might be tuneable by second-order selection to increase evolvability (Nelson and Masel, 2018). The feasibility of the adaptive evolution of evolvability has been shown mathematically (King and Masel, 2007) and the activity of *recA*, necessary for homologous recombination in bacteria, can be subject to evolutionary changes. Indeed, we know that in eukaryotes the rates of meiotic recombination catalysed by *recA*-homologs differ in different species (Webster and Hurst, 2012; Stapley *et al.*, 2017). Thus, given enough time, second-order selection could in principle act to increase the rate of recombination. However, in our experiments there is not sufficient time for this to happen; the rate of amplification is already sufficiently high to produce an excess of copy number polymorphism in the population on the first day of the experiment. The results of chapter two show how our fluctuating selection regime maintains even a standing expression level-polymorphism (Fig. 2.4b), which is only depleted after more than a week of growing in a constant environment (Fig. 2.4d). This means fluctuating first-order selection has many individuals with different expression levels to pick from. Hence, second-order selection is unlikely to act to increase the rate of recombination in this situation. Moreover, in bacteria RecA is the most conserved protein involved in DNA metabolism (Bell and Kowalczykowski, 2016). Thus, given its crucial role in DNA damage repair, *recA* is expected to be under stringent functional constrains. Accordingly, at least in the context of our fluctuating selection experiments, the genome-wide rate of homologous recombination seems an unlikely target for secondary selection to act on.

We can conclude by repeating what was said in Chapter three: canonical gene regulation is ubiquitous even in the simplest of organisms. Moreover, *cis*-regulatory sequences even seem to evolve faster than the coding regions they regulate (Surguchov, 1991; Matus-Garcia, Nijveen and van Passel, 2012b; Nijveen, Matus-Garcia and van Passel, 2012; Oliver and Greene, 2012; van Passel, Nijveen and Wahl, 2014; Yona, Alm and Gore, 2018). However, between these two "classical ways" of changing gene expression –regulation (highly non-random and fast) and adaptive mutation (random and slow) - a third "category" seems to exist in bacteria. There, gene expression is changed on rather short evolutionary timescales and while in some cases partially hard-wired (frequent chromosome break points, catalyzed inversion) still relies on mutations. Importantly, our synthetic example - *galK* copy number mutations in an artificial fluctuating environment - relies entirely on random mutations. At any rate, the list of ways in which bacteria change gene expression by frequent mutations is growing and provides exciting examples of transient evolutionary dynamics.

Generally speaking, transient evolutionary dynamics are hard to study. As the name suggests, they do not leave traces in the record of genomic sequence data (and to repeat a quote from Chapter four – "attention is shifted to where the data are", (Kondrashov, 2012)). We have seen in this thesis that a dedicated experimental system is needed to observe and appreciate the transient dynamics pertaining to copy number mutations. Here we were using a dual reporter system to monitor mutations on the single cell and population level in real time.

In addition, precisely because transient dynamics do not leave direct traces in the record of genomic sequence data, their detailed study may not seem warranted – after all they are just transient. To prove this intuition wrong, chapter four seeks to uncover a potentially long-lasting effect resulting from the transient dynamics associated with copy number mutations: if adaptation by amplification is the fastest and sufficient, other –less frequent- mutations may not have a chance to compete. However, this means that after the transient adaptation is over, no change remains. In other words, amplifications could effectively act as buffer against long-lasting point mutations.

Another viewpoint tells us that studying (transient) evolutionary dynamics is warranted. We now know that the spectrum of adaptive mutations depends critically on the degree of clonal interference. Hence, whether or not large or small effect mutations have a chance to fix depends on population size and mutation rate (Cvijović, Nguyen Ba and Desai, 2018). We have seen from chapter four that –if available - frequent mutations such as amplifications and to a lesser degree small-scale deletions in the promoter region dominate adaptation. Given the high rates of amplification, it is not surprising that we found signatures of clonal interference in our experiments. Recent technical advances such as barcoding of bacterial populations for evolution experiments helped to uncover a degree of clonal interference previously not appreciated (Cvijović, Nguyen Ba and Desai, 2018; Lauer *et al.*, 2018).

Importantly, on sufficiently long time- scales, the transient dynamics that play out before fixation of mutations may ultimately shape entire genomes (Cvijović, Nguyen Ba and Desai, 2018).

5.1 Considerations on the generalizability of the results of this thesis

The results of this thesis pertain to the tandem duplication and amplification of a model gene with well-defined function. Expression of *galK* can be readily measured, such that we are able equate increased gene copy-number with increased expression. This result is in line with many reported cases of adaptive gene amplification that occur in response to selection for increased expression in prokaryotic and eukaryotic species (Albertson, 2006; Perry *et al.*, 2007; Sandegren and Andersson, 2009; Bass and Field, 2011b; Gusev *et al.*, 2014; Selmecki *et al.*, 2015; Tranel, 2017; Pajic *et al.*, 2019). However, more copies do not necessarily always mean more expression. This simple relation may not hold for genes whose expression is controlled by auto-feedback regulation or in situations, where transcriptional activators (supplied *in trans*) become limiting as copy number increases.

In this thesis experimental evolution in *E.coli* was used to explore two main implications of unstable gene copy number mutations: the fact that they (i) allow to increase and decrease gene expression on short time scales (chapter two) and (ii) may to some degree hinder the evolution by point mutations (chapter four). In principle, these results should be generalizable to most organisms from the bacterial and archaeal domain, which are similarly characterized by large effective population sizes, the absence of sexual reproduction and concise genome architectures (Lynch, 2007; Hanage, 2016). However, it is also interesting to

consider the generalizability of the findings presented here to eukaryotes, especially those with sexual reproduction and small effective population sizes; The latter implies that non-adaptive processes play a bigger role during evolution (Lynch, 2007).

Observing the rapid dynamics of AMGET (i.e. gene expression changes within the timescale of a single day), described in chapter two, certainly requires organisms with short generation times and large population sizes. Nonetheless, adaptive amplification does occur even in elephants, where copy number expansions of the tumor suppressor gene TP53 compensate for the increased cancer risk inherent to large body sizes (Sulak *et al.*, 2016; Tollis, Schneider-Utaka and Maley, 2020). As laid out in chapter two, the timescales of AMGET will depend on the effective population size, the strength of selection and the rate of spontaneous duplication (k_{dup}) as well as the rate of deletion and further amplification (k_{rec}). While k_{rec} has only been estimated in bacteria (Mats E. Pettersson *et al.*, 2009; Reams *et al.*, 2010; Tomanek *et al.*, 2020), the rate of spontaneous duplication formation may differ between prokaryotes and eukaryotes. Direct measurements conducted in *Salmonella enterica* found a duplication rate between 10^{-6} – 10^{-2} per locus per cell per generation (Andersson and Hughes, 2009). In eukaryotes, mutation accumulation experiments estimated the rate of duplication and deletion to be around 10^{-5} per locus per generation in *Daphnia pulex* (Keith *et al.*, 2016) and 10^{-7} per locus per generation *Caenorhabditis elegans* and *Drosophila melanogaster* (Lipinski *et al.*, 2011; Schrider *et al.*, 2013) and 10^{-6} per locus per generation *Saccharaomiyces cerivisiae* (Lynch *et al.*, 2008), respectively. As long as rates of copy number mutation are higher than those of point mutation, the former should in principle allow more rapid adaptive changes to gene expression.

It seems that, not only this thesis but the prokaryotic literature in general, is more concerned with the rate of spontaneous duplication (Andersson and Hughes, 2009), whereas eukaryotic studies focus on the rate of duplicate retention over longer evolutionary timescales. This may be reflective of a fundamentally different evolutionary role duplication plays in prokaryotes and eukaryotes (Treangen and Rocha, 2011). In prokaryotes, duplication seems to mostly confer adaptive increases in gene expression, while gene families tend to expand via horizontal gene transfer rather than duplication. In contrast, paralogs are the main source of functional novelty in eukaryotes (Treangen and Rocha, 2011).

A *bona fide* difference between prokaryotes and eukaryotes (especially such with small population sizes) is likely the role of drift for the fixation of duplicates. Large effective population sizes ensure that purifying selection removes even mildly detrimental duplicates, which are otherwise able to fix via drift (Lynch, 2007). However, irrespective of organism, the fate of most duplicates should be their pseudogenization or deletion. Only those which are fixed – by drift or selection or a combination of the two- can be seen in sequence data. Therefore, comparative genomics studies can inform us about the kinds of genes that are more likely to be retained as duplicates. For instance, rapidly evolving genes tend to be retained as duplicates more frequently (O'Toole, Hurst and McLysaght, 2018), refuting the idea that duplication itself creates conditions of relaxed selection that allows for rapid

divergence (Lynch and Conery, 2000). Moreover, genes whose products participate in protein-protein complexes need to be balanced in dosage and are thus more recalcitrant to single gene duplications. However, dosage-sensitive genes can be retained as duplicates (referred to as ohnologs) after a whole genome duplication (WGD) (Papp, Pál and Hurst, 2003). Exceptions to the balance hypothesis (Papp, Pál and Hurst, 2003) in fact confirm the rule, as the duplication of dosage-sensitive ohnologs was found to underlie a variety of human diseases (Makino and McLysaght, 2010; McLysaght *et al.*, 2014).
In general, any consideration of gene dosage is more complex in diploid or polyploid eukaryotes. While prokaryotes are haploid, during rapid growth the presence of multiple replication forks leads to increased dosage of genes proximal to the origin of replication relative to genes proximal to the terminus of replication (Couturier and Rocha, 2006). Fast growing species may harbor up to a dozen replication forks simultaneously. Hence, gene dosage is a function of both the growth rate and the position on the circular chromosome (Couturier and Rocha, 2006). Therefore, even in haploid bacterial or archaeal species, this means that the relative effect of any duplication depends on its position on the chromosome.

Chapter four introduces the idea that gene copy number mutations might hinder the evolution by adaptive point mutations. Given the importance of clonal interference in this process, the results of chapter four are most relevant to asexually reproducing microorganisms. However, it is interesting to consider duplication and its effect on gene dosage in sexually reproducing organisms, specifically in the context of sex chromosome evolution. In mammals, for instance, degradation of the Y-chromosome decreases dosage for X-linked genes in males by one half (Gu and Walters, 2017). One simple idea of how to compensate for the missing gene dosage of a single X-chromosome is gene duplication, especially since transcriptional traffic jams are thought to limit the maximal per-promoter expression (Hurst *et al.*, 2015). However, at least in the case of the mammalian X-chromosome this does not necessarily seem to be the case. Rather, the relatively high number of duplicates seen on the X (versus autosomes) can be explained by the fact that the retention of duplicates (which later become subfunctionalized for tissue-specific expression patterns) is biased towards genes, which are not deleterious when duplicated. Hence, it seems rather than dosage-selection, passive or adaptive subfunctionalization is driving the retention of duplications in the X-chromosome (Hurst *et al.*, 2015).
What this example illustrates more than anything is that simple patterns of gene duplication and amplification in response to dosage selection may be a distinct feature of bacteria and that matters are likely more complicated in eukaryotes. However, even bacterial genomic sequence data are largely devoid of tandem duplications, and yet amplification is a frequent adaptation seen in experiments and natural settings with strong dosage selection. We should therefore not yet dismiss simple dosage-selection as an early driver for the formation of paralogs even in eukaryotes. In order to fully understand how duplications shape the

genomes of all life forms we need both, an understanding of their transient evolutionary dynamics as well as their long-term fate.

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*, 31(6), pp. 1526–35. doi: 10.1093/molbev/msu111.

Albertson, D. G. (2006) 'Gene amplification in cancer', *Trends in Genetics*, 22(8), pp. 447–455. doi: 10.1016/j.tig.2006.06.007.

Anderson, P. and 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. doi: 10.1073/pnas.78.5.3113.

Anderson, R. P. and Roth, J. R. (1977) 'Tandem Genetic Duplications in Phage and Bacteria', *Annual Review of Microbiology*, 31(1), pp. 473–505. doi: 10.1146/annurev.mi.31.100177.002353.

Anderson, R. P. and Roth, J. R. (1978) 'Tandem chromosomal duplications in Salmonella typhimurium: Fusion of histidine genes to novel promoters', *Journal of Molecular Biology*, 119(1), pp. 147–166. doi: 10.1016/0022-2836(78)90274-7.

Andersson, D. I. and Hughes, D. (2009) 'Gene Amplification and Adaptive Evolution in Bacteria', *Annual Review of Genetics*, 43(1), pp. 167–195. doi: 10.1146/annurev-genet-102108-134805.

Andersson, D. I. and Jerlstro, J. (2015) 'Evolution of new functions *de novo* and from preexisting genes', *Cold Spring Harbor Perspectives in Biology*, (1932). doi: 10.1101/cshperspect.a017996.

Andersson, D. I., Jerlström-Hultqvist, J. and Nasväll, J. (2015) 'Evolution of new functions de novo and from preexisting genes', *Cold Spring Harbor Perspectives in Biology*, 7(6), pp. 1–19. doi: 10.1101/cshperspect.a017996.

Andersson, D. I., Jerlström-Hultqvist, J. and Näsvall, J. (2015) 'Evolution of new functions de novo and from preexisting genes.', *Cold Spring Harbor perspectives in biology*, 7(6), pp. a017996-. doi: 10.1101/cshperspect.a017996.

Andersson, D. I., Slechta, S. E. and Roth, J. R. (1998) 'Evidence That Gene Amplification Underlies Adaptive Mutability of the Bacterial lac Operon', *Science*, 282(5391), pp. 1133–1135. doi: 10.1126/science.282.5391.1133.

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. doi: 10.1038/msb4100050.

Barkan, D., Stallings, C. L. and Glickman, M. S. (2011) 'An improved counterselectable marker system for mycobacterial recombination using galK and 2-deoxy-galactose', *Gene*, 470(1–2), pp. 31–36. doi: 10.1016/j.gene.2010.09.005.

Barnard, A., Wolfe, A. and Busby, S. (2004) 'Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes.', *Current opinion in microbiology*, 7(2), pp. 102–8. doi: 10.1016/j.mib.2004.02.011.

Barton, N. and Partridge, L. (2000) 'Limits to natural selection', *BioEssays*, 22(12), pp. 1075–1084. doi: 10.1002/1521-1878(200012)22:12<1075::AID-BIES5>3.0.CO;2-M.

Bass, C. and Field, L. M. (2011a) 'Gene amplification and insecticide resistance.', *Pest management science*, 67(8), pp. 886–90. doi: 10.1002/ps.2189.

Bass, C. and Field, L. M. (2011b) 'Gene amplification and insecticide resistance', *Pest Management Science*, 67(8), pp. 886–890. doi: 10.1002/ps.2189.

Bayer, A., Brennan, G. and Geballe, A. P. (2018) 'Adaptation by copy number variation in monopartite viruses', *Current Opinion in Virology*, 33, pp. 7–12. doi: 10.1016/j.coviro.2018.07.001.

Bayliss, C. D. (2009) 'Determinants of phase variation rate and the fitness implications of differing rates for bacterial pathogens and commensals', *FEMS Microbiology Reviews*, 33(3), pp. 504–520. doi: 10.1111/j.1574-6976.2009.00162.x.

Belikova, D. *et al.* (2020) '"Gene accordions" cause genotypic and phenotypic heterogeneity in clonal populations of Staphylococcus aureus', *Nature Communications*, 11(1), p. 3526. doi: 10.1038/s41467-020-17277-3.

Bell, G. (2013) 'Evolutionary rescue and the limits of adaptation', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1610), pp. 1–6. doi: 10.1098/rstb.2012.0080.

Bell, J. C. and Kowalczykowski, S. C. (2016) 'RecA: Regulation and Mechanism of a Molecular Search Engine', *Trends in Biochemical Sciences*, 41(6), pp. 491–507. doi: 10.1016/j.tibs.2016.04.002.

Berg, J., Willmann, S. and Lässig, M. (2004) 'Adaptive evolution of transcription factor binding sites.', *BMC evolutionary biology*, 4, p. 42. doi: 10.1186/1471-2148-4-42.

Bergmiller, T. *et al.* (2017) 'Biased partitioning of the multidrug efflux pump AcrAB-TolC underlies long-lived phenotypic heterogeneity.', *Science (New York, N.Y.)*, 356(6335), pp. 311–315. doi: 10.1126/science.aaf4762.

Bergthorsson, U., Andersson, D. I. and Roth, J. R. (2007) 'Ohno's dilemma: evolution of new genes under continuous selection.', *Proceedings of the National Academy of Sciences of the United States of America*, 104(43), pp. 17004–17009. doi: 10.1073/pnas.0707158104.

Bershtein, S. and Tawfik, D. S. (2008) 'Ohno's model revisited: Measuring the frequency of potentially adaptive mutations under various mutational drifts', *Molecular Biology and Evolution*, 25(11), pp. 2311–2318. doi: 10.1093/molbev/msn174.

Blount, Z. D. *et al.* (2012) 'Genomic analysis of a key innovation in an experimental Escherichia coli population.', *Nature*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 489(7417), pp. 513–8. doi: 10.1038/nature11514.

Blount, Z. D. and Lenski, R. E. 'Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli', *Proc Natl Acad Sci U S A*, pp. 513–518. doi: 10.1073.

Cairns, J. and Foster, P. L. (1991) 'Adaptive reversion of a frameshift mutation in Escherichia coli', *Genetics*, 128(4), pp. 695–701. Available at:

https://pubmed.ncbi.nlm.nih.gov/1916241.

Cairns, J., Overbaugh, J. and Miller, S. (1988) 'The origin of mutants', *Nature*, 335(6186), pp. 142–145. doi: 10.1038/335142a0.

Carlson, S. M., Cunningham, C. J. and Westley, P. A. H. (2014) 'Evolutionary rescue in a changing world', *Trends in Ecology and Evolution*, 29(9), pp. 521–530. doi: 10.1016/j.tree.2014.06.005.

Chait, R. *et al.* (2010) 'A differential drug screen for compounds that select against antibiotic resistance', *PLoS ONE*. Edited by P. J. Planet. Public Library of Science, 5(12), p. e15179. doi: 10.1371/journal.pone.0015179.

Chen, C. W. *et al.* (2002) 'Once the circle has been broken: Dynamics and evolution of Streptomyces chromosomes', *Trends in Genetics*, 18(10), pp. 522–529. doi: 10.1016/S0168-9525(02)02752-X.

Claycomb, J. M. and Orr-Weaver, T. L. (2005) 'Developmental gene amplification: Insights into DNA replication and gene expression', *Trends in Genetics*, 21(3), pp. 149–162. doi: 10.1016/j.tig.2005.01.009.

Conant, G. C. and Wolfe, K. H. (2008) 'Turning a hobby into a job: How duplicated genes find new functions', *Nature Reviews Genetics*, 9(12), pp. 938–950. doi: 10.1038/nrg2482.

Cone, K. R. *et al.* (2017) 'Emergence of a Viral RNA Polymerase Variant during Gene Copy Number Amplification Promotes Rapid Evolution of Vaccinia Virus', *Journal of Virology*, 91(4). doi: 10.1128/jvi.01428-16.

Copley, S. D. (2017) 'Shining a light on enzyme promiscuity', *Current Opinion in Structural Biology*, 47, pp. 167–175. doi: 10.1016/j.sbi.2017.11.001.

Corn, P. G. *et al.* (1993) 'Genes involved in Haemophilus influenzae type b capsule expression are Amplified', *Journal Of Infectious Diseases*, 167(2), pp. 356–364. Available at: http://www.jstor.org/stable/30113084.

Couturier, E. and Rocha, E. P. C. (2006) 'Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes.', *Molecular microbiology*, 59(5), pp. 1506–18. doi: 10.1111/j.1365-2958.2006.05046.x.

Csiszovszki, Z. *et al.* (2011) 'Structure and Function of the D -Galactose Network in Enterobacteria', *mBio*, 2(4), pp. 1–8. doi: 10.1128/mBio.00053-11.Editor.

Cui, L. *et al.* (2012) 'Coordinated phenotype switching with large-scale chromosome flip-flop inversion observed in bacteria', *Proceedings of the National Academy of Sciences of the United States of America*, 109(25), pp. 5–9. doi: 10.1073/pnas.1204307109.

Cullis, C. A. (2005) 'Mechanisms and control of rapid genomic changes in flax', *Annals of Botany*, 95(1), pp. 201–206. doi: 10.1093/aob/mci013.

Cvijović, I., Nguyen Ba, A. N. and Desai, M. M. (2018) 'Experimental Studies of Evolutionary Dynamics in Microbes', *Trends in Genetics*, 34(9), pp. 693–703. doi: 10.1016/j.tig.2018.06.004.

Darmon, E. and Leach, D. R. F. (2014) 'Bacterial genome instability.', *Microbiology and*

molecular biology reviews : MMBR, 78(1), pp. 1–39. doi: 10.1128/MMBR.00035-13.

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

Datta, S., Costantino, N. and Court, D. L. (2006) 'A set of recombineering plasmids for gram-negative bacteria', *Gene*, 379, pp. 109–115. doi: 10.1016/j.gene.2006.04.018.

Davis, G. S. *et al.* (2011) 'Use of bexB to detect the capsule locus in Haemophilus influenzae', *Journal of Clinical Microbiology*. American Society for Microbiology (ASM), 49(7), pp. 2594–2601. doi: 10.1128/JCM.02509-10.

Dhar, R., Bergmiller, T. and Wagner, A. (2014) 'INCREASED GENE DOSAGE PLAYS A PREDOMINANT ROLE IN THE INITIAL STAGES OF EVOLUTION OF DUPLICATE TEM-1 BETA LACTAMASE GENES', *Evolution*, 68(6), pp. 1775–1791. doi: 10.1111/evo.12373.

Dobrindt, U. *et al.* (2004) 'Genomic islands in pathogenic and environmental microorganisms.', *Nature reviews. Microbiology*, 2(5), pp. 414–24. doi: 10.1038/nrmicro884.

Drake, J. W. (1991) 'A constant rate of spontaneous mutation in DNA-based microbes.', *Proceedings of the National Academy of Sciences*, 88(16), pp. 7160–7164. doi: 10.1073/pnas.88.16.7160.

Drake, J. W. *et al.* (1998) 'Rates of spontaneous mutation', *Genetics*, 148(4), pp. 1667–1686.

Elde, N. C. *et al.* (2012) 'Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses.', *Cell*. Elsevier, 150(4), pp. 831–41. doi: 10.1016/j.cell.2012.05.049.

Elez, M. *et al.* (2010) 'Seeing Mutations in Living Cells', *Current Biology*, 20(16), pp. 1432–1437. doi: 10.1016/j.cub.2010.06.071.

Elliott, K. T., Cuff, L. E. and Neidle, E. L. (2013) 'Copy number change: evolving views on gene amplification', *Future Microbiology*, 8(7), pp. 887–899. doi: 10.2217/fmb.13.53.

Eme, L. *et al.* (2017) 'Lateral Gene Transfer in the Adaptation of the Anaerobic Parasite Blastocystis to the Gut', *Current Biology*. Elsevier Ltd., 27(6), pp. 807–820. doi: 10.1016/j.cub.2017.02.003.

Eydallin, G. *et al.* (2014) 'The nature of laboratory domestication changes in freshly isolated Escherichia coli strains', *Environmental Microbiology*, 16(3), pp. 813–828. doi: 10.1111/1462-2920.12208.

Fan, Y. *et al.* (2002) 'Gene content and function of the ancestral chromosome fusion site in human chromosome 2q13-2q14.1 and paralogous regions', *Genome Research*, 12(11), pp. 1663–1672. doi: 10.1101/gr.338402.

Force, A. *et al.* (1999) 'Preservation of duplicate genes by complementary, degenerative mutations', *Genetics*, 151(4), pp. 1531–1545.

Friedlander, T. *et al.* (2016) 'Intrinsic limits to gene regulation by global crosstalk', *Nature Communications*, 7. doi: 10.1038/ncomms12307.

Gerland, U. and Hwa, T. (2009) 'Evolutionary selection between alternative modes of gene

regulation.', *Proceedings of the National Academy of Sciences of the United States of America*. NATL ACAD SCIENCES, 2101 CONSTITUTION AVE NW, WASHINGTON, DC 20418 USA, 106(22), pp. 8841–6. doi: 10.1073/pnas.0808500106.

Gerrish, P. J. and Lenski, R. E. (1998) 'The fate of competing beneficial mutations in an asexual population', *Genetica*, 102, pp. 127–144.

Gil, R. *et al.* (2006) 'Plasmids in the aphid endosymbiont Buchnera aphidicola with the smallest genomes. A puzzling evolutionary story', *Gene*, 370, pp. 17–25. doi: 10.1016/j.gene.2005.10.043.

Gladman, S. L. *et al.* (2015) 'Large tandem chromosome expansions facilitate niche adaptation during persistent infection with drug-resistant Staphylococcus aureus', *Microbial Genomics*, 1(2), p. e000026. doi: 10.1099/mgen.0.000026.

Greenblum, S., Carr, R. and Borenstein, E. (2015) 'Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species', *Cell*. Elsevier Inc., 160(4), pp. 583–594. doi: 10.1016/j.cell.2014.12.038.

Gu, L. and Walters, J. R. (2017) 'Evolution of sex chromosome dosage compensation in animals: A beautiful theory, undermined by facts and bedeviled by details', *Genome Biology and Evolution*, 9(9), pp. 2461–2476. doi: 10.1093/gbe/evx154.

Gusev, O. *et al.* (2014) 'Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge.', *Nature communications*, 5, p. 4784. doi: 10.1038/ncomms5784.

Guzman, L. M. *et al.* (1995) 'Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter.', *Journal of bacteriology*, 177(14), pp. 4121–30. doi: 10.1128/jb.177.14.4121-4130.1995.

Haldimann, A. and Wanner, B. L. (2001) 'Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria.', *Journal of bacteriology*, 183(21), pp. 6384–93. doi: 10.1128/JB.183.21.6384-6393.2001.

Hanage, W. P. (2016) 'Not so simple after all: Bacteria, their population genetics, and recombination', *Cold Spring Harbor Perspectives in Biology*, 8(7). doi: 10.1101/cshperspect.a018069.

Henderson, P. J., Giddens, R. a and Jones-Mortimer, M. C. (1977) 'Transport of galactose, glucose and their molecular analogues by Escherichia coli K12.', *The Biochemical journal*, 162(2), pp. 309–20. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1164603&tool=pmcentrez&rendertype=abstract.

Hendrickson, H. *et al.* (1995) 'Adaptive reversion of an episomal frameshift mutation in Escherichia coli requires conjugal functions but not actual conjugation.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 92(12), pp. 5487–90. doi: 10.1073/pnas.92.12.5487.

Hjort, K., Nicoloff, H. and Andersson, D. I. (2016) 'Unstable tandem gene amplification generates heteroresistance (variation in resistance within a population) to colistin in Salmonella enterica', *Molecular Microbiology*, 102(2), pp. 274–289. doi:

10.1111/mmi.13459.

Hooper, S. D. and Berg, O. G. (2003) 'Duplication is more common among laterally transferred genes than among indigenous genes.', *Genome biology*, 4(8), p. R48. doi: 10.1186/gb-2003-4-8-r48.

Hung, M. *et al.* (2014) 'Modulating the frequency and bias of stochastic switching to control phenotypic variation', *Nature Communications*. Nature Publishing Group, 5(May), p. 4574. doi: 10.1038/ncomms5574.

Hurst, L. D. *et al.* (2015) 'The Constrained Maximal Expression Level Owing to Haploidy Shapes Gene Content on the Mammalian X Chromosome', *PLoS Biology*, 13(12), pp. 1–48. doi: 10.1371/journal.pbio.1002315.

Igler, C. *et al.* (2018) 'Evolutionary potential of transcription factors for gene regulatory rewiring', *Nature Ecology & Evolution*, 2(10), pp. 1633–1643. doi: 10.1038/s41559-018-0651-y.

Innan, H. and Kondrashov, F. (2010) 'The evolution of gene duplications: Classifying and distinguishing between models', *Nature Reviews Genetics*, 11(2), pp. 97–108. doi: 10.1038/nrg2689.

Jacob, F. (1977) 'Evolution and tinkering', *Science*, 196(4295), pp. 1161–1166. doi: 10.1126/science.860134.

Juhas, M. *et al.* (2009) 'Genomic islands: tools of bacterial horizontal gene transfer and evolution.', *FEMS microbiology reviews*, 33(2), pp. 376–93. doi: 10.1111/j.1574-6976.2008.00136.x.

Kacser, H. and Beeby, R. (1984) 'Evolution of catalytic proteins - On the origin of enzyme species by means of natural selection', *Journal of Molecular Evolution*, 20(1), pp. 38–51. doi: 10.1007/BF02101984.

Kafatos, F. C., Orr, W. and Delidakis, C. (1985) 'Developmetally regulated gene amplification', *Trends in Genetics*. Elsevier, 1, pp. 301–306. doi: 10.1016/0168-9525(85)90119-2.

Katju, V. and Bergthorsson, U. (2013) 'Copy-number changes in evolution: Rates, fitness effects and adaptive significance', *Frontiers in Genetics*.

Keith, N. *et al.* (2016) 'High mutational rates of large-scale duplication and deletion in Daphnia pulex', *Genome Research*, 26(1), pp. 60–69. doi: 10.1101/gr.191338.115.

Khlebnikov, A. *et al.* (2001) 'Homogeneous expression of the P(BAD) promoter in Escherichia coli by constitutive expression of the low-affinity high-capacity AraE transporter.', *Microbiology (Reading, England)*, 147(Pt 12), pp. 3241–7. doi: 10.1099/00221287-147-12-3241.

King, O. D. and Masel, J. (2007) 'The evolution of bet-hedging adaptations to rare scenarios', *Theoretical Population Biology*, 72(4), pp. 560–575. doi: 10.1016/j.tpb.2007.08.006.

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*

of the United States of America, 107(20), pp. 9158–63. doi: 10.1073/pnas.1004290107.

Kloeckener-Gruissem, B. and Freeling, M. (1995) 'Transposon-induced promoter scrambling: a mechanism for the evolution of new alleles.', *Proceedings of the National Academy of Sciences of the United States of America*, 92(6), pp. 1836–40. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=42377&tool=pmcentrez&rendertype=abstract (Accessed: 11 September 2014).

Kondrashov, F. A. *et al.* (2002) 'Selection in the evolution of gene duplications.', *Genome biology*, 3(2), pp. 1–9. doi: 10.1186/gb-2002-3-2-research0008.

Kondrashov, F. A. (2012) 'Gene duplication as a mechanism of genomic adaptation to a changing environment', *Proceedings of the Royal Society B: Biological Sciences*, 279(1749), pp. 5048–5057. doi: 10.1098/rspb.2012.1108.

Kondrashov, F. A. and Kondrashov, A. S. (2006) 'Role of selection in fixation of gene duplications', *Journal of Theoretical Biology*, 239(2), pp. 141–151. doi: 10.1016/j.jtbi.2005.08.033.

Kornberg, H. L. and Riordan, C. (1976) 'Uptake of galactose into Escherichia coli by facilitated diffusion', *Journal of General Microbiology*, 94(1), pp. 75–89. doi: 10.1099/00221287-94-1-75.

Kussell, E. and Laibler (2005) 'Phenotypic Diversity, Population Growth, and Information in Fluctuating Environments', *Science*, 309(5743), pp. 2075–2078. doi: 10.1126/science.1114383.

Lagator, M. *et al.* (2017) 'Regulatory network structure determines patterns of intermolecular epistasis', *eLife*, 6, pp. 1–22. doi: 10.7554/eLife.28921.

Lagator, M. *et al.* (2020) 'Structure and Evolution of Constitutive Bacterial Promoters', *bioRxiv*, 21(1), p. 2020.05.19.104232. doi: 10.1101/2020.05.19.104232.

Lang, G. I. and Desai, M. M. (2014) 'The spectrum of adaptive mutations in experimental evolution.', *Genomics*, 104(6 Pt A), pp. 412–6. doi: 10.1016/j.ygeno.2014.09.011.

Latorre, A. *et al.* (2005) 'Chromosomal stasis versus plasmid plasticity in aphid endosymbiont Buchnera aphidicola', *Heredity*, 95(5), pp. 339–347. doi: 10.1038/sj.hdy.6800716.

Lauer, S. *et al.* (2018) 'Single-cell copy number variant detection reveals the dynamics and diversity of adaptation', *PLoS Biology*, p. 381590. doi: 10.1371/journal.pbio.

Lauer, S. and Gresham, D. (2019) 'An evolving view of copy number variants', *Current Genetics*, pp. 1287–1295. doi: 10.1007/s00294-019-00980-0.

Leblond, P. and Decaris, B. (1994) 'New insights into the genetic instability of streptomyces', *FEMS Microbiology Letters*, 123(3), pp. 225–232. doi: 10.1111/j.1574-6968.1994.tb07229.x.

Lercher, M. J. and Pál, C. (2008) 'Integration of horizontally transferred genes into regulatory interaction networks takes many million years.', *Molecular biology and evolution*, 25(3), pp. 559–67. doi: 10.1093/molbev/msm283.

Lipinski, K. J. *et al.* (2011) 'High spontaneous rate of gene duplication in Caenorhabditis

elegans', *Current Biology*. NIH Public Access, 21(4), pp. 306–310. doi: 10.1016/j.cub.2011.01.026.

Lutz, R. and 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. 1203–10. doi: 10.1093/nar/25.6.1203.

Lynch, M. (2007) 'The frailty of adaptive hypotheses for the origins of organismal complexity', *In the Light of Evolution*, 1(Table 1), pp. 83–103. doi: 10.17226/11790.

Lynch, M. *et al.* (2008) 'A genome-wide view of the spectrum of spontaneous mutations in yeast.', *Proceedings of the National Academy of Sciences of the United States of America*, 105(27), pp. 9272–9277. doi: 10.1073/pnas.0803466105.

Lynch, M. and Conery, J. S. (2000) 'The Evolutionary Fate and Consequences of Duplicate Genes', *Science*, 290(5494), pp. 1151–1155. doi: 10.1126/science.290.5494.1151.

Lynch, M. and Force, A. (2000) 'The probability of duplicate gene preservation by subfunctionalization', *Genetics*, 154(1), pp. 459–473.

MacLean, R. C. and Millan, A. S. (2019) 'The evolution of antibiotic resistance', *Science*, 365(6458), pp. 1082–1083. doi: 10.1126/science.aax3879.

Makino, T. and McLysaght, A. (2010) 'Ohnologs in the human genome are dosage balanced and frequently associated with disease', *Proceedings of the National Academy of Sciences of the United States of America*, 107(20), pp. 9270–9274. doi: 10.1073/pnas.0914697107.

Matus-Garcia, M., Nijveen, H. and van Passel, M. W. J. (2012a) 'Promoter propagation in prokaryotes.', *Nucleic acids research*, 40(20), pp. 10032–40. doi: 10.1093/nar/gks787.

Matus-Garcia, M., Nijveen, H. and van Passel, M. W. J. (2012b) 'Promoter propagation in prokaryotes.', *Nucleic acids research*, 40(20), pp. 10032–40. doi: 10.1093/nar/gks787.

McDermott, A. (2019) 'Probing the limits of "evolutionary rescue" Could species threatened by climate change and other stresses avoid extinction through rapid evolution?', *Proceedings of the National Academy of Sciences of the United States of America*, 116(25), pp. 12116–12120. doi: 10.1073/pnas.1907565116.

McLysaght, A. *et al.* (2014) 'Ohnologs are overrepresented in pathogenic copy number mutations', *Proceedings of the National Academy of Sciences of the United States of America*, 111(1), pp. 361–366. doi: 10.1073/pnas.1309324111.

Mizuno, K. *et al.* (2013) 'Recombination-restarted replication makes inverted chromosome fusions at inverted repeats', *Nature*, 493(7431), pp. 246–249. doi: 10.1038/nature11676.

Moore, R. C. and Purugganan, M. D. (2003) 'The early stages of duplicate gene evolution', *Proceedings of the National Academy of Sciences*, 100(26), pp. 15682–15687. doi: 10.1073/pnas.2535513100.

Moxon, E. R. *et al.* (1994) 'Adaptive evolution of highly mutable loci in pathogenic bacteria', *Current Biology*, 4(1), pp. 24–33. doi: 10.1016/S0960-9822(00)00005-1.

Moxon, R., Bayliss, C. and Hood, D. (2006) 'Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation.', *Annual review of genetics*, 40, pp. 307–33.

doi: 10.1146/annurev.genet.40.110405.090442.

Nagelkerke, F. and Postma, P. W. (1978) '2-Deoxygalactose, a specific substrate of the Salmonella typhimurium galactose permease: Its use for the isolation of galP mutants', *Journal of Bacteriology*, 133(2), pp. 607–613.

Näsvall, J. *et al.* (2012) 'Real-time evolution of new genes by innovation, amplification, and divergence.', *Science (New York, N.Y.)*, 338(6105), pp. 384–7. doi: 10.1126/science.1226521.

Nelson, P. and Masel, J. (2018) 'Evolutionary Capacitance Emerges Spontaneously during Adaptation to Environmental Changes', *Cell Reports*. ElsevierCompany., 25(1), pp. 249–258. doi: 10.1016/j.celrep.2018.09.008.

Nguyen, T. N. *et al.* (1989) 'Effects of carriage and expression of the Tn10 tetracycline-resistance operon on the fitness of Escherichia coli K12.', *Molecular biology and evolution*, 6(3), pp. 213–25. doi: 10.1093/oxfordjournals.molbev.a040545.

Nicoloff, H. *et al.* (2019) 'The high prevalence of antibiotic heteroresistance in pathogenic bacteria is mainly caused by gene amplification', *Nature Microbiology*, 4(3), pp. 504–514. doi: 10.1038/s41564-018-0342-0.

Nijveen, H., Matus-Garcia, M. and van Passel, M. W. J. (2012) 'Promoter reuse in prokaryotes.', *Mobile genetic elements*. Landes Bioscience, 2(6), pp. 279–281. doi: 10.4161/mge.23195.

O'Toole, Á. N., Hurst, L. D. and McLysaght, A. (2018) 'Faster Evolving Primate Genes Are More Likely to Duplicate', *Molecular Biology and Evolution*, 35(1), pp. 107–118. doi: 10.1093/molbev/msx270.

Ohno, S. (1970) *Evolution by Gene Duplication*, *Evolution by Gene Duplication*. doi: 10.1007/978-3-642-86659-3.

Oliver, K. R. and Greene, W. K. (2012) 'Transposable elements and viruses as factors in adaptation and evolution: an expansion and strengthening of the TE-Thrust hypothesis.', *Ecology and evolution*, 2(11), pp. 2912–33. doi: 10.1002/ece3.400.

Oren, Y. *et al.* (2014) 'Transfer of noncoding DNA drives regulatory rewiring in bacteria.', *Proceedings of the National Academy of Sciences of the United States of America*, 111(45). doi: 10.1073/pnas.1413272111.

Pajic, P. *et al.* (2019) 'Independent amylase gene copy number bursts correlate with dietary preferences in mammals', *eLife*, 8, pp. 1–22. doi: 10.7554/eLife.44628.

Pál, C., Papp, B. and Lercher, M. J. (2005) 'Adaptive evolution of bacterial metabolic networks by horizontal gene transfer', *Nat Genet*, 37(12), pp. 1372–1375. doi: 10.1038/ng1686.

Papp, B., Pál, C. and Hurst, L. D. (2003) 'Dosage sensitivity and the evolution of gene families in yeast', *Nature*, 424(6945), pp. 194–197. doi: 10.1038/nature01771.

van Passel, M. W. J., Nijveen, H. and Wahl, L. M. (2014) 'Birth, death, and diversification of mobile promoters in prokaryotes.', *Genetics*, 197(1), pp. 291–9. doi: 10.1534/genetics.114.162883.

Perez, J. C. and Groisman, E. A. (2009) 'Evolution of transcriptional regulatory circuits in bacteria.', *Cell*, 138(2), pp. 233–44. doi: 10.1016/j.cell.2009.07.002.

Perry, G. H. *et al.* (2007) 'Diet and the evolution of human amylase gene copy number variation', *Nature Genetics*, 39(10), pp. 1256–1260. doi: 10.1038/ng2123.

Pettersson, M. E. *et al.* (2005) 'The Amplification Model for Adaptive Mutation', *Genetics*, 169(2), pp. 1105–1115. doi: 10.1534/genetics.104.030338.

Pettersson, Mats E. *et al.* (2009) 'Evolution of new gene functions: simulation and analysis of the amplification model.', *Genetica*, 135(3), pp. 309–324. doi: 10.1007/s10709-008-9289-z.

Pettersson, Mats E *et al.* (2009) 'Evolution of new gene functions: simulation and analysis of the amplification model', *Genetica*. Oxford, UK: Oxford University Press, 135(3), pp. 309–324. doi: 10.1007/s10709-008-9289-z.

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

Poelwijk, F. J., de Vos, M. G. J. and Tans, S. J. (2011) 'Tradeoffs and optimality in the evolution of gene regulation.', *Cell*, 146(3), pp. 462–70. doi: 10.1016/j.cell.2011.06.035.

Pränting, M. and Andersson, D. I. (2011) 'Escape from growth restriction in small colony variants of Salmonella typhimurium by gene amplification and mutation.', *Molecular microbiology*, 79(2), pp. 305–15. doi: 10.1111/j.1365-2958.2010.07458.x.

Prody, C. A. *et al.* (1989) 'De novo amplification within a "silent" human cholinesterase gene in a family subjected to prolonged exposure to organophosphorous insecticides.', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 86(2), pp. 690–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2911599 (Accessed: 20 July 2016).

R Development Core Team (2005) 'R: A language and environment for statistical computing'. Vienna, Austria. Available at: http://www.r-project.org.

Reams, A. B. *et al.* (2010) 'Duplication Frequency in a Population of Salmonella enterica Rapidly Approaches Steady State With or Without Recombination', *Genetics*, 184(4), pp. 1077–1094. doi: 10.1534/genetics.109.111963.

Reams, A. B. and Roth, J. R. (2015) 'Mechanisms of gene duplication and amplification.', *Cold Spring Harbor perspectives in biology*, 7(2), p. a016592. doi: 10.1101/cshperspect.a016592.

Reyrat, J. M. *et al.* (1998) 'Counterselectable markers: untapped tools for bacterial genetics and pathogenesis.', *Infection and immunity*, 66(9), pp. 4011–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=108478&tool=pmcentrez&rendertype=abstract (Accessed: 26 January 2015).

Rodriguez-Beltran, J. *et al.* (2018) 'Multicopy plasmids allow bacteria to escape from fitness trade-offs during evolutionary innovation', *Nature Ecology & Evolution*, 2(5), pp. 873–881. doi: 10.1038/s41559-018-0529-z.

Roth, J. R. *et al.* (1988) 'Rearrangements of the Bacterial Chromosome: Formation and Applications', *Science*, 241(4871), pp. 1314–1318. Available at: http://ecosal.org/.

Roth, J. R. and Andersson, D. I. (2004) 'Amplification-mutagenesis - How growth under selection contributes to the origin of genetic diversity and explains the phenomenon of adaptive mutation', *Research in Microbiology*, 155(5), pp. 342–351. doi: 10.1016/j.resmic.2004.01.016.

San Millan, A. *et al.* (2017) 'Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria', *Nature Ecology & Evolution*, 1(1), p. 0010. doi: 10.1038/s41559-016-0010.

Sandegren, L. and Andersson, D. I. (2009) 'Bacterial gene amplification: implications for the evolution of antibiotic resistance.', *Nature reviews. Microbiology*. Nature Publishing Group, 7(8), pp. 578–88. doi: 10.1038/nrmicro2174.

Savageau, M. A. (1974) 'Genetic regulatory mechanisms and the ecological niche of Escherichia coli.', *Proceedings of the National Academy of Sciences of the United States of America*, 71(6), pp. 2453–5. doi: DOI 10.1073/pnas.71.6.2453.

Schrider, D. R. *et al.* (2013) 'Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster', *Genetics*, 194(4), pp. 937–954. doi: 10.1534/genetics.113.151670.

Segall, A., Mahan, M. and Roth (1988) 'Rearrangements of the Bacterial Chromosome: Formation and Applications', *Science*, 241(4871), pp. 1314–1318. Available at: http://ecosal.org/.

Selmecki, A. M. *et al.* (2015) 'Polyploidy can drive rapid adaptation in yeast', *Nature*. Nature Publishing Group, 519(7543), pp. 349–351. doi: 10.1038/nature14187.

Shiu, S. H. *et al.* (2006) 'Role of positive selection in the retention of duplicate genes in mammalian genomes', *Proceedings of the National Academy of Sciences of the United States of America*, 103(7), pp. 2232–2236. doi: 10.1073/pnas.0510388103.

Slager, J., Aprianto, R. and Veening, J. W. (2018) 'Deep genome annotation of the opportunistic human pathogen Streptococcus pneumoniae D39', *Nucleic Acids Research*, 46(19), pp. 9971–9989. doi: 10.1093/nar/gky725.

Smith, J. C. and Sheltzer, J. M. (2018) 'Systematic identification of mutations and copy number alterations associated with cancer patient prognosis', *eLife*, 7, pp. 1–26. doi: 10.7554/eLife.39217.

Song, S. *et al.* (2009) 'Contribution of gene amplification to evolution of increased antibiotic resistance in Salmonella typhimurium', *Genetics*, 182(4), pp. 1183–1195. doi: 10.1534/genetics.109.103028.

Stapley, J. *et al.* (2017) 'Variation in recombination frequency and distribution across eukaryotes: patterns and processes.', *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*. The Royal Society, 372(1736), p. 20160455. doi: 10.1098/rstb.2016.0455.

Steinrueck, M. and Guet, C. C. (2017) 'Complex chromosomal neighborhood effects determine the adaptive potential of a gene under selection', *eLife*, 6, pp. 1–26. doi: 10.7554/eLife.25100.

Sulak, M. *et al.* (2016) 'TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants', *eLife*, 5(September2016), pp. 1–30. doi: 10.7554/eLife.11994.

Sun, S. *et al.* (2012) 'Genome-Wide Detection of Spontaneous Chromosomal Rearrangements in Bacteria', *PLoS ONE*. Edited by M. Watson. Public Library of Science, 7(8), p. e42639. doi: 10.1371/journal.pone.0042639.

Surguchov, A. (1991) 'Migration of promoter elements between genes: a role in transcriptional regulation and evolution', *Biomedical science*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/1680476 (Accessed: 23 July 2014).

Tawfik, D. S. (2010) 'Messy biology and the origins of evolutionary innovations', *Nature Chemical Biology*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 6(10), pp. 692–696. doi: 10.1038/nchembio.441.

Taylor, T. B. *et al.* (2015) 'Evolutionary resurrection of flagellar motility via rewiring of the nitrogen regulation system', *Science*, 347(6225), pp. 1014–1017. doi: 10.1126/science.1259145.

Teufel, A. I., Masel, J. and Liberles, D. A. (2015) 'What fraction of duplicates observed in recently sequenced genomes is segregating and destined to fail to fix?', *Genome Biology and Evolution*, 7(8), pp. 2258–2264. doi: 10.1093/gbe/evv139.

Todd, R. T. and Selmecki, A. (2020) 'Expandable and reversible copy number amplification drives rapid adaptation to antifungal drugs', *eLife*, 9, p. e58349. doi: 10.7554/eLife.58349.

Tollis, M., Schneider-Utaka, A. K. and Maley, C. C. (2020) 'The Evolution of Human Cancer Gene Duplications across Mammals', *Molecular Biology and Evolution*, 37(10), pp. 2875–2886. doi: 10.1093/molbev/msaa125.

Tomanek, I. *et al.* (2020) 'Gene amplification as a form of population-level gene expression regulation', *Nature Ecology & Evolution 2020*, pp. 1–14. doi: 10.1038/s41559-020-1132-7.

Tranel, P. (2017) 'Herbicide-resistance mechanisms: gene amplification is not just for glyphosate', *Pest Management Science*. doi: 10.1002/ps.4679.

Traxler, M. F. and Kolter, R. (2015) 'Natural products in soil microbe interactions and evolution', *Natural Product Reports*, 32(7), pp. 956–970. doi: 10.1039/c5np00013k.

Treangen, T. J. and Rocha, E. P. C. (2011) 'Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes', *PLoS Genetics*, 7(1). doi: 10.1371/journal.pgen.1001284.

Troein, C. *et al.* (2007) 'Is transcriptional regulation of metabolic pathways an optimal strategy for fitness?', *PloS one*. PUBLIC LIBRARY SCIENCE, 185 BERRY ST, STE 1300, SAN FRANCISCO, CA 94107 USA, 2(9), p. e855. doi: 10.1371/journal.pone.0000855.

Tuğrul, M. *et al.* (2015) 'Dynamics of Transcription Factor Binding Site Evolution', *PLOS Genetics*. Edited by J. C. Fay. Public Library of Science, 11(11), p. e1005639. doi: 10.1371/journal.pgen.1005639.

Uecker, H. and Hermisson, J. (2016) 'The role of recombination in evolutionary rescue',

Genetics, 202(2), pp. 721–732. doi: 10.1534/genetics.115.180299.

Uhlemann, A. C. *et al.* (2014) 'Molecular tracing of the emergence, diversification, and transmission of S. aureus sequence type 8 in a New York community', *Proceedings of the National Academy of Sciences of the United States of America*, 111(18), pp. 6738–6743. doi: 10.1073/pnas.1401006111.

Veening, J.-W., Smits, W. K. and Kuipers, O. P. (2008) 'Bistability, Epigenetics, and Bet-Hedging in Bacteria', *Annual Review of Microbiology*, 62(1), pp. 193–210. doi: 10.1146/annurev.micro.62.081307.163002.

Villar, D. *et al.* (2015) 'Enhancer Evolution across 20 Mammalian Species', *Cell*, 160(3), pp. 554–566. doi: 10.1016/j.cell.2015.01.006.

Villar, D., Flicek, P. and Odom, D. T. (2014) 'Evolution of transcription factor binding in metazoans — mechanisms and functional implications', *Nature Reviews Genetics*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 15(4), pp. 221–233. doi: 10.1038/nrg3481.

Vinces, M. D. *et al.* (2009) 'Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability', *Science*, 324(5931), pp. 1213 LP – 1216. Available at: http://science.sciencemag.org/content/324/5931/1213.abstract.

Wang, P. *et al.* (2010) 'Robust Growth of Escherichia coli', *Current Biology*, 20(12), pp. 1099–1103. doi: 10.1016/j.cub.2010.04.045.

Webster, M. T. and Hurst, L. D. (2012) 'Direct and indirect consequences of meiotic recombination: implications for genome evolution.', *Trends in genetics : TIG*, 28(3), pp. 101–9. doi: 10.1016/j.tig.2011.11.002.

Wolf, L., Silander, O. K. and van Nimwegen, E. (2015) 'Expression noise facilitates the evolution of gene regulation', *eLife*, 4, pp. 1–48. doi: 10.7554/eLife.05856.

Yi, X. and Dean, A. M. (2013) 'Bounded population sizes, fluctuating selection and the tempo and mode of coexistence.', *Proceedings of the National Academy of Sciences of the United States of America*, 110(42), pp. 16945–50. doi: 10.1073/pnas.1309830110.

Yona, A. H. *et al.* (2012) 'Chromosomal duplication is a transient evolutionary solution to stress.', *Proceedings of the National Academy of Sciences of the United States of America*, 109(51), pp. 21010–5. doi: 10.1073/pnas.1211150109.

Yona, A. H., Alm, E. J. and Gore, J. (2018) 'Random sequences rapidly evolve into de novo promoters', *Nature Communications*, 9(1), p. 1530. doi: 10.1038/s41467-018-04026-w.

Yona, A. H., Frumkin, I. and Pilpel, Y. (2015) 'A Relay Race on the Evolutionary Adaptation Spectrum.', *Cell*. Elsevier Inc., 163(3), pp. 549–59. doi: 10.1016/j.cell.2015.10.005.

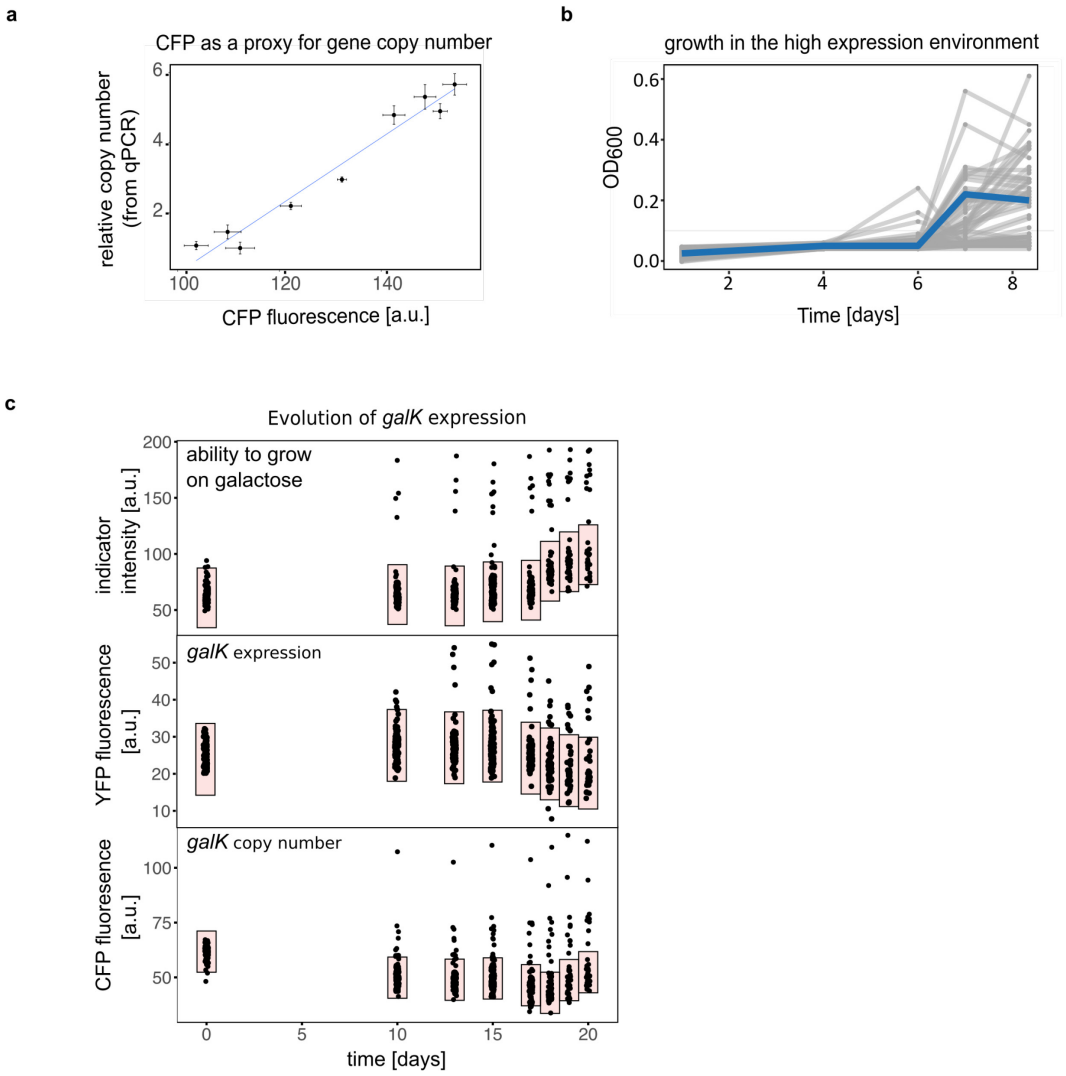
Zhou, L. *et al.* (2017) 'Chromosome engineering of Escherichia coli for constitutive production of salvianic acid A', *Microbial Cell Factories*, 16(1), p. 84. doi: 10.1186/s12934-017-0700-2.

Zinser, E. R. and Kolter, R. (2004) 'Escherichia coli evolution during stationary phase.', *Research in microbiology*, 155(5), pp. 328–36. doi: 10.1016/j.resmic.2004.01.014.

A. Appendix: Supplementary Information for Chapter Two

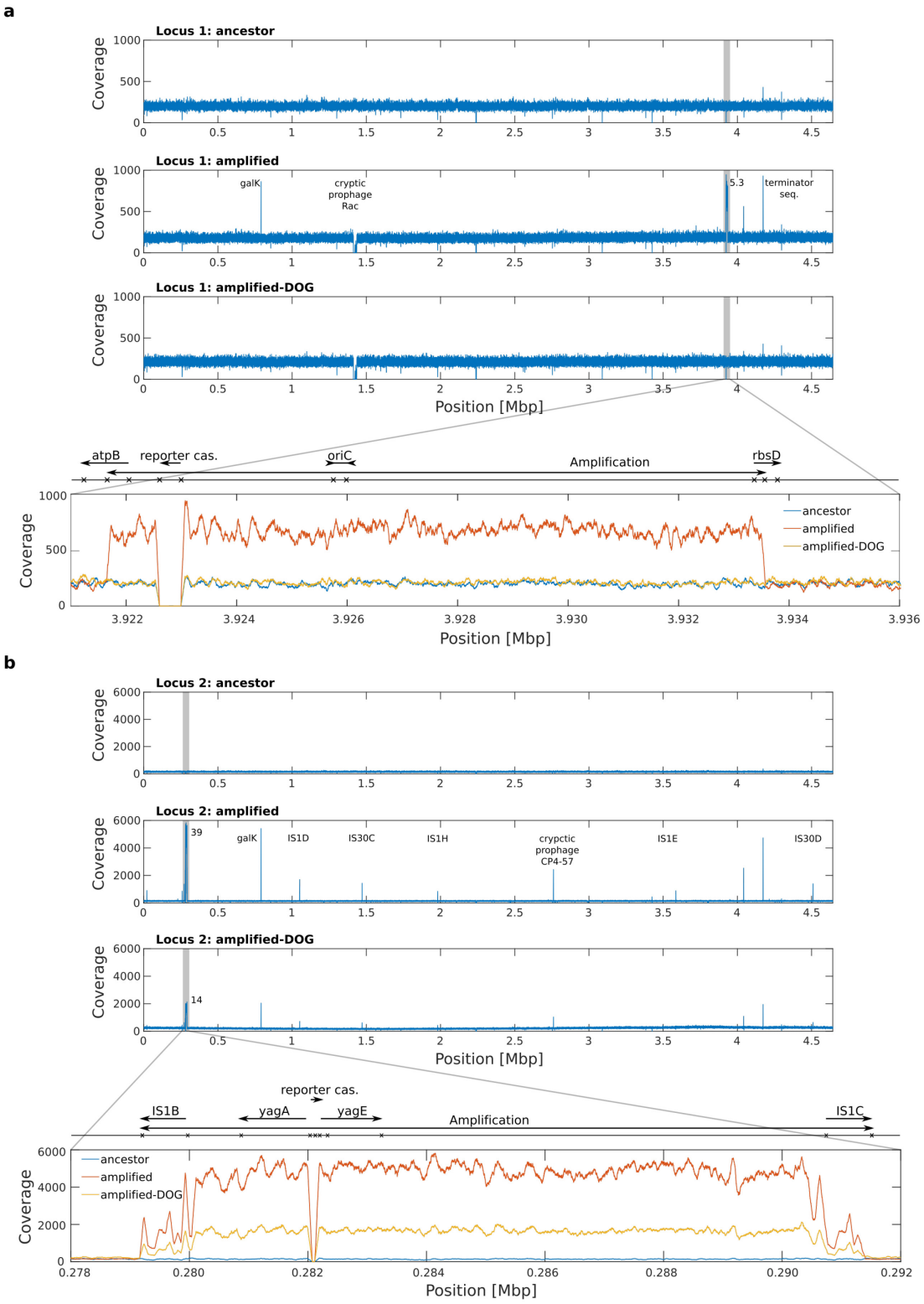
Supplementary Note. An upper limit for copy number exists in locus 1.

CFP levels of the amplified strain stabilize in populations following prolonged exposure to the high selection environment (Fig. 2. 2c, positive control), indicating that there is a cost to increasing copy number above a certain point. Indeed, microfluidics experiments revealed that increasing copy number beyond the maximum attainable level of CFP fluorescence generally led to cell death, which lead to the exclusion of all such lineages from our analysis (Methods). Both, microfluidics experiments (Fig. S5b) and qPCR (Fig. S1a), consistently estimate a maximum copy number between six and ten. This upper limit to copy number might be due to the fact that the origin of replication lies within the amplified segment (see Methods) and could thus be specific to the strain we are using. This is corroborated by the fact that the copy number of the strain amplified in locus 2 is estimated to be 39 according to read-depth (Fig. S2b). If there is a strict limit to copy number in locus 2, it is much higher than in locus 1.



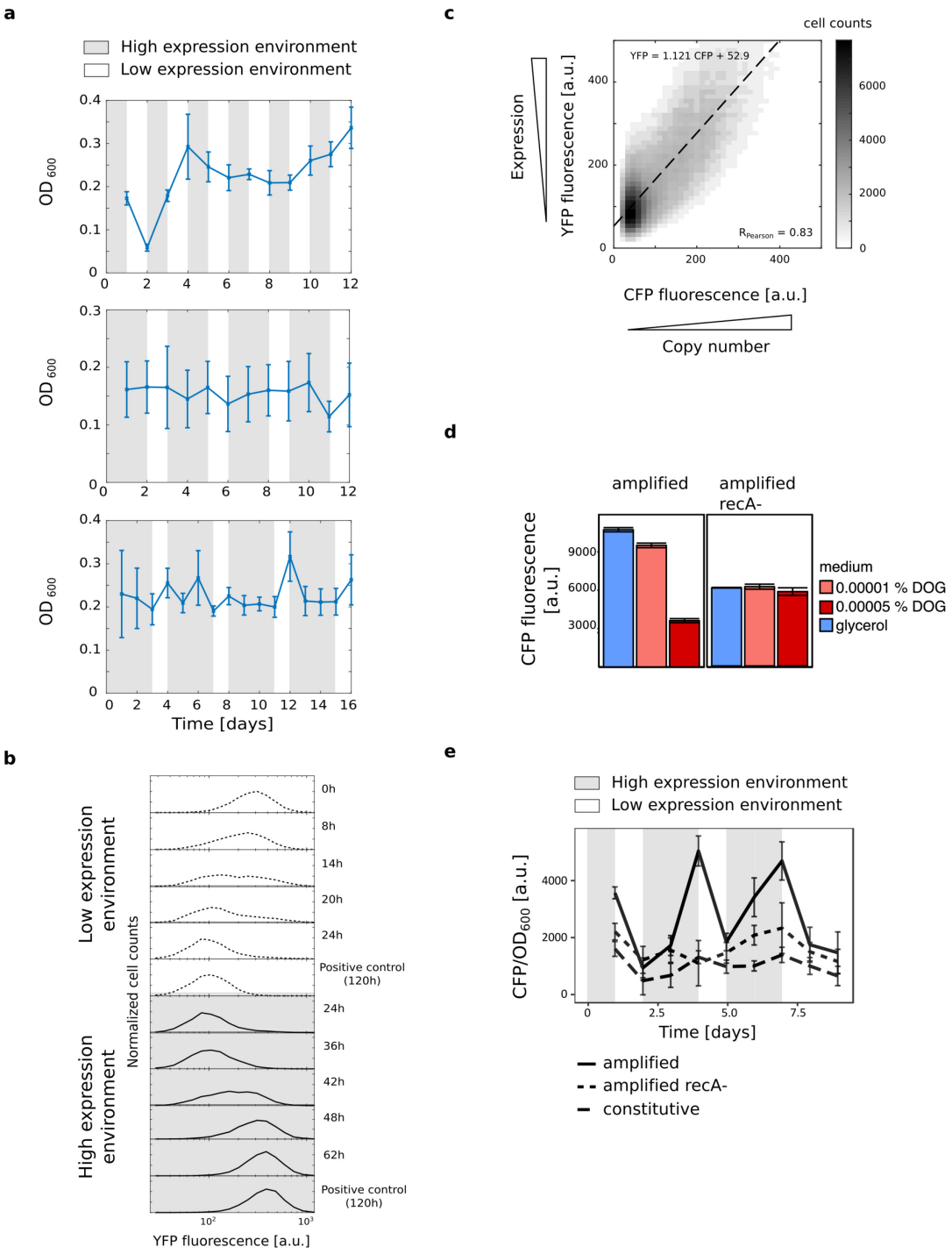
Supplementary Figure 1. Experimental evolution of *galK* expression. a, CFP fluorescence of bacterial colonies as a proxy for copy number. Copy number relative to a single copy control strain as determined by qPCR is plotted for eight populations with varying levels of CFP fluorescence. Error bars represent the standard deviation of three and four replicates for copy number and CFP fluorescence, respectively. Linear fit: Adjusted R-squared=0.9558, p-value=3.352e-06. **b**, OD_{600} of 95 replicate populations of the ancestral strain each evolving in 200µl minimal galactose medium (high expression environment). Plot shows the initial continuous cultivation phase of the evolution experiment prior to the first transfer to fresh medium. Blue line shows the population of the amplified strain. **c**, MacConkey agar pins (as shown in Fig. 2. 1b, – right part) of the 95 replicate populations shown in **b** during 21 days of evolution in the high expression environment. Evolving populations were pinned onto MacConkey agar at the beginning of the evolution experiment and prior to each transfer into fresh medium to monitor their phenotypic changes: ability to grow on galactose (apparent from pH indicator color shift to pink) - top panel, colony YFP fluorescence (as a proxy for *galK* expression) - middle panel and colony CFP fluorescence (as a proxy for *galK*

copy number) - bottom panel. Area shaded in red corresponds to population median +/- 3xSD of the ancestral population.



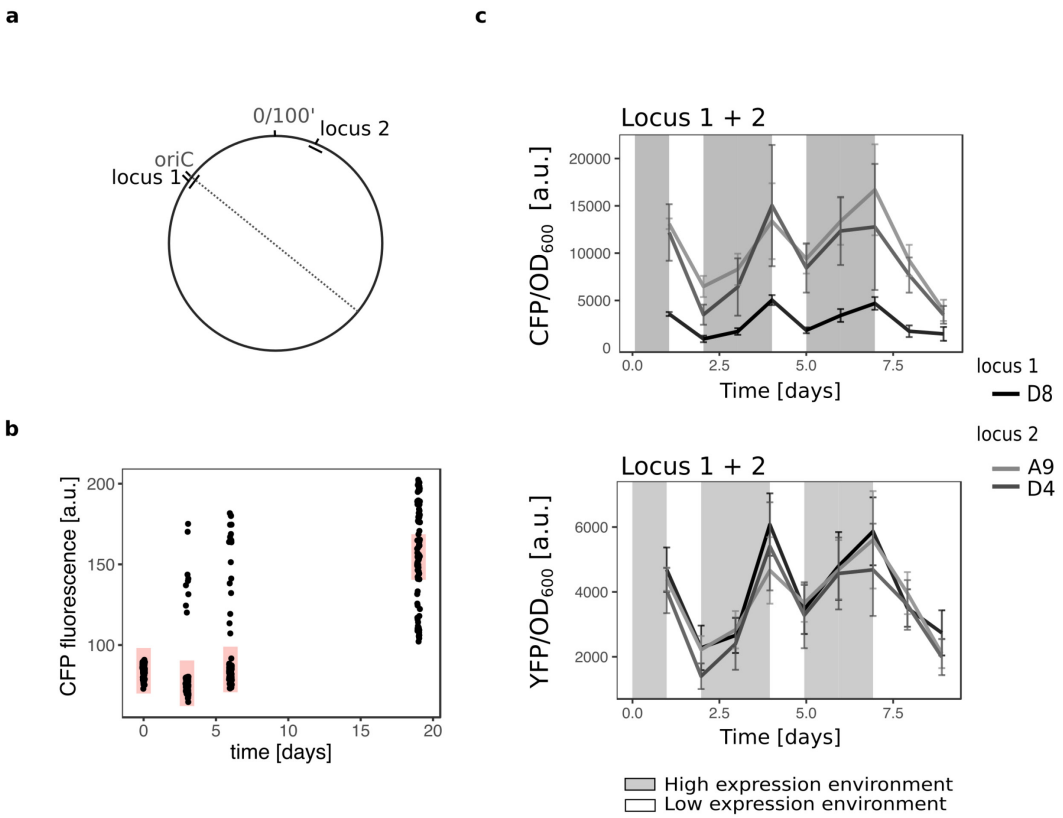
Supplementary Figure 2. Coverage plot of ancestral and evolved strains of locus 1 and locus 2. Read-depth is shown for the whole genome of **a**, Locus 1: (**top**) ancestral strain, (**middle**) amplified strain isolated after evolution in the high expression environment (Fig.

S1c), (**bottom**) amplified strain after 24h in the low expression environment (clone from experiment shown in Fig. 2. 2c). **b**, Locus 2 (**top**) ancestral strain, (**middle**) amplified strain isolated after evolution in the high expression environment (Fig. S4b), (**bottom**) amplified strain after 24h in the low expression environment. The number next to the amplified region indicates the fold change in coverage as compared to the respective ancestral strain. Additional regions with increased coverage (labeled in the middle panels of a) are caused by sequence reads of the synthetic reporter cassette mapping to homologous sequences within the *E.coli* genome: endogenous *galK* , terminators downstream of *yfp* and *cfp* (~4.1 and 4.2 Mbp, resp.). Prophage Rac is absent in the evolved strains of locus 1. For locus 2, additional regions of increased coverage (labeled in the middle panel of b) are caused by homologies with the amplified region, especially insertion sequence (IS) element 1.

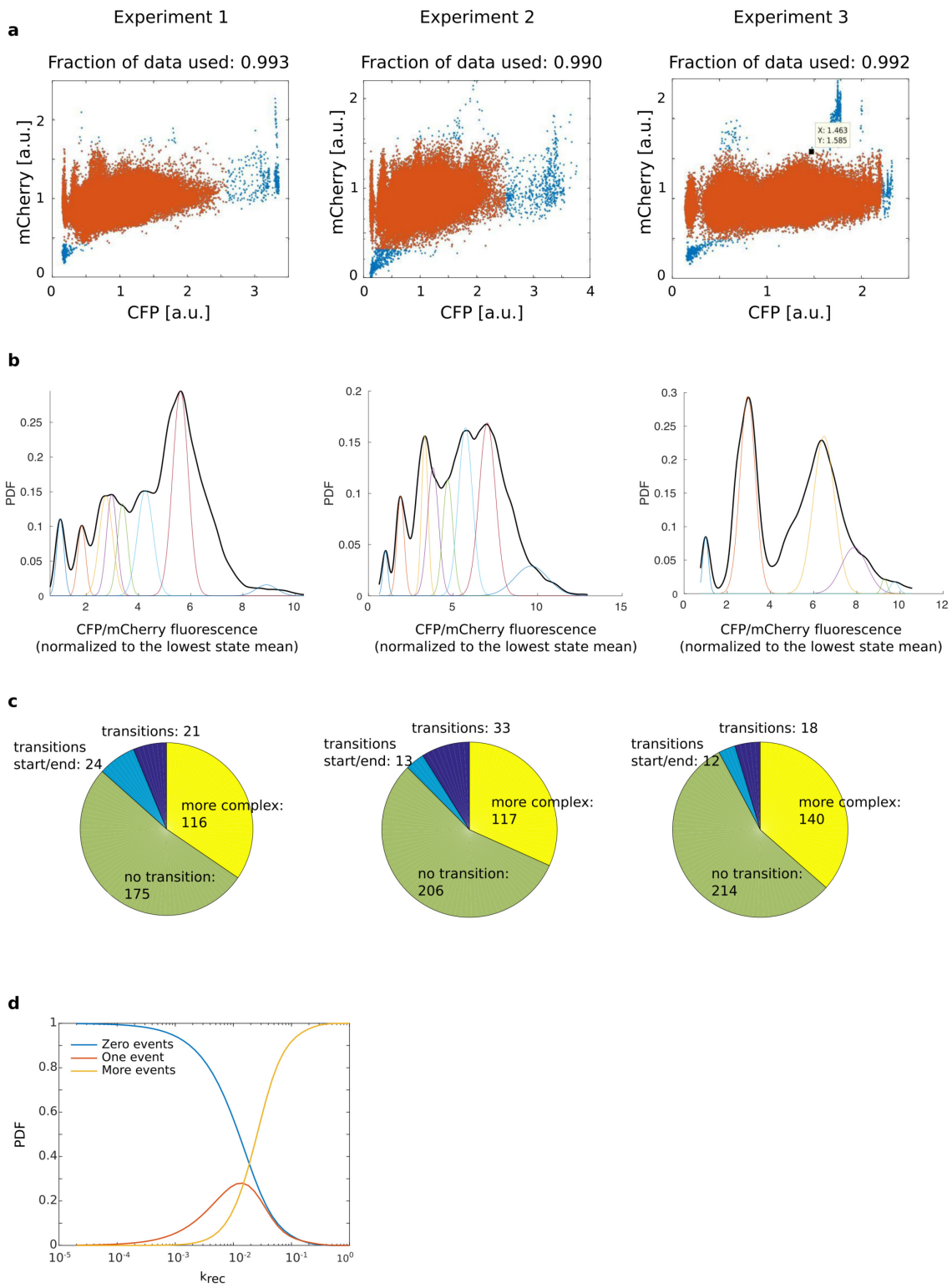


Supplementary Figure 3. Amplification-mediated gene expression tuning (AMGET) allows growth in alternating environments and is dependent on *recA*. a, Growth of the amplified strain during alternating selection (see also Fig. 2.2b). OD_{600} is shown for alternating selection following the scheme of 1 day - 1 day, 2 days – 1 day and 3 days – 1 day in high and low expression environment, respectively. Error bars represent the standard deviation (SD) of 60 populations.

b, Flow cytometry histogram (one of six replicates from two independent experiments) following the adaptation of an amplified bacterial population to low and high expression environments. Population was inoculated from a single colony and selected for two days in the high expression environment prior to the two transitions shown here. When switched from high to low expression environment, YFP fluorescence as a proxy for *galK* expression is decreasing within 24h to reach the steady state level of the same population after 5 days in the low environment (positive control). When shifted back to the high expression environment, the amplified population increases in CFP fluorescence to the level reached by the same population after 5 days in the high expression environment (positive control). **c**, Plot shows CFP fluorescence as a proxy for *galK* copy number and YFP fluorescence as a proxy for *galK* expression of the evolving population (data from the experiment shown Fig. 2.2c and Fig. S3b, respectively). **d**, Mean steady state CFP fluorescence of amplified populations with (left) and without (right) functional *recA* allele grown in 0%, 0.00005% and 0.00001% DOG. **e**, During alternating selection, CFP levels of the amplified strain tracks fluctuating environments. CFP levels of neither the *recA*- derivative of the amplified strain nor a constitutive, single-copy derivative of the amplified strain follows the environments. The constitutive strain evolved serendipitously in an overnight culture as a clone that lost its amplification but gained a point mutation in p0 of the chromosomal cassette allowing for *galK* expression in the absence of amplification.

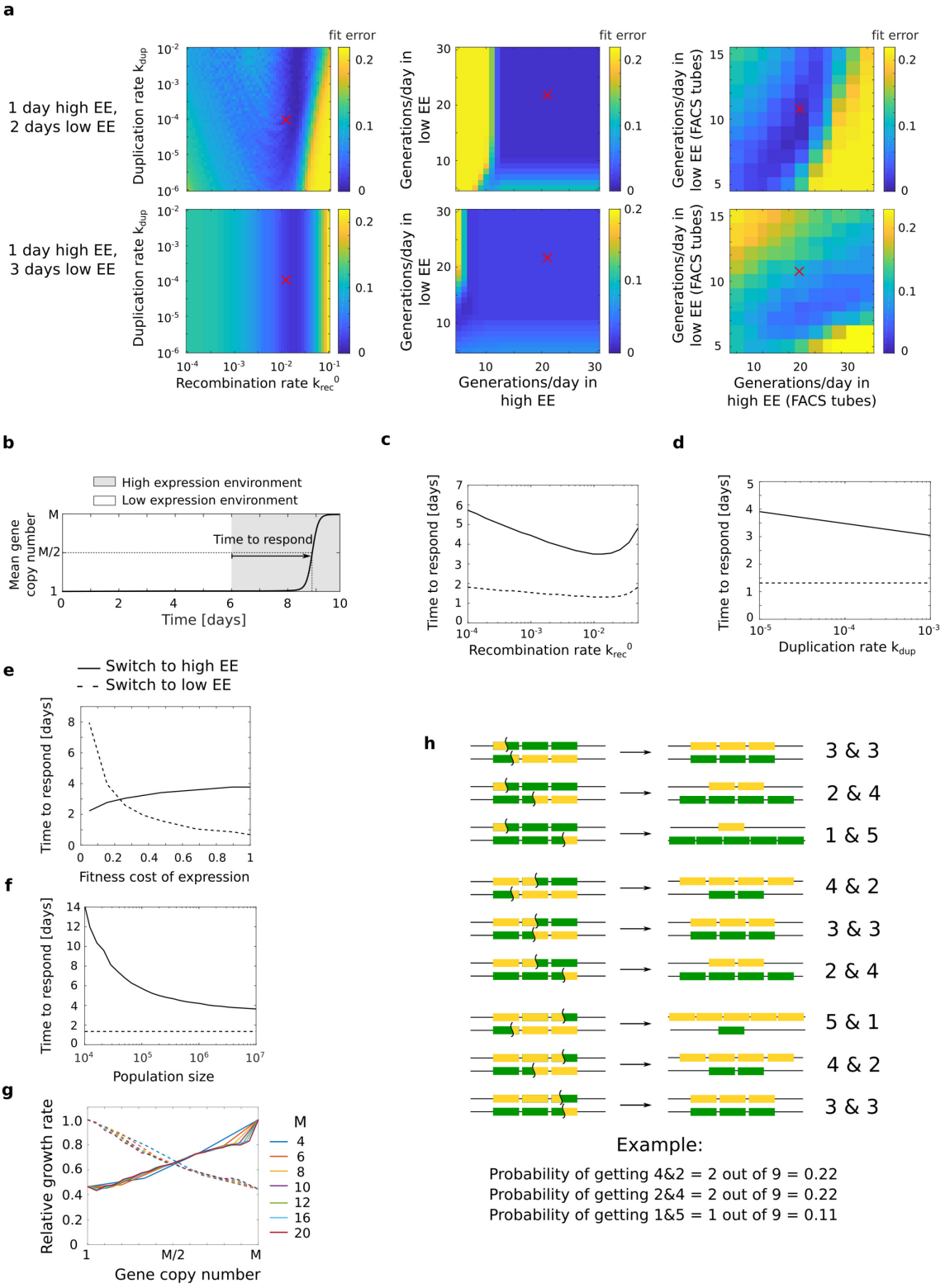


Supplementary Figure 4. AMGET occurs at a different genomic locus. a, *E. coli* chromosome map showing positions of locus 1 (downstream of *rsmG*) and locus 2 (inside cryptic prophage CP4-6 and flanked by two identical IS elements) relative to the origin of replication (*oriC*). **b**, Amplifications readily evolve in locus 2. Colony CFP fluorescence as a proxy for gene copy number of 95 replicate populations pinned onto agar before and during evolution in the high expression environment. Red shaded area represents the median +/- 3xSD of the ancestral population. **c**, Normalized CFP fluorescence of strains with gene amplification in locus 2 ("A9", "D4") tracks fluctuating environments like the strain with a gene amplification in locus 1 ("D8"). Although absolute CFP levels are higher in locus 2 than locus 1 (top panel), fold change of CFP and YFP is similar between both loci (bottom panel).



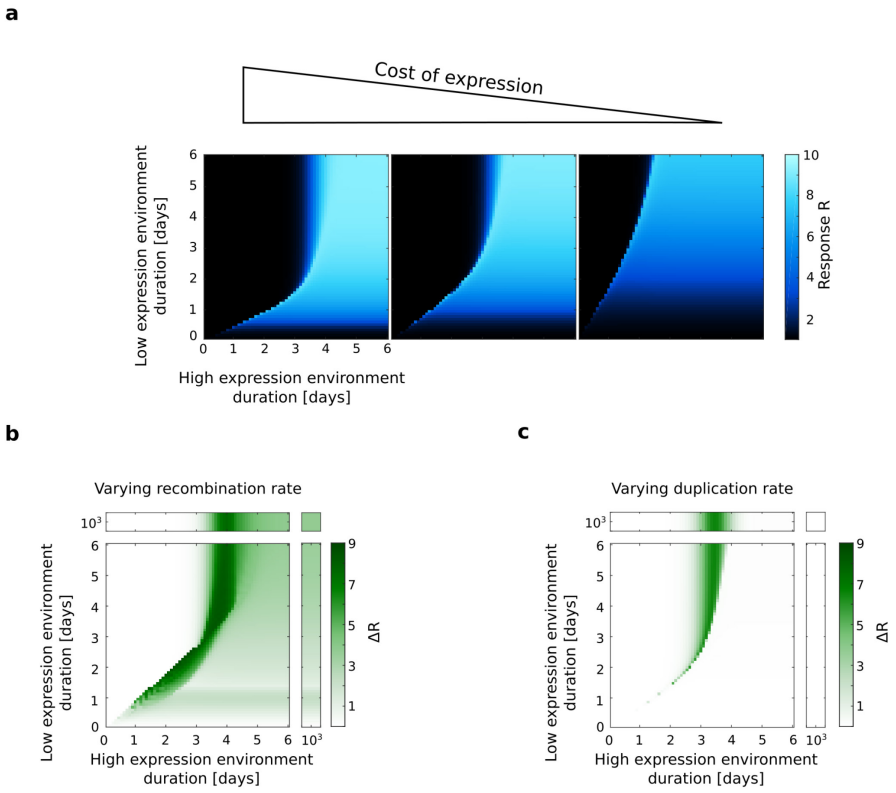
Supplementary Figure 5. Microfluidics data analysis. a, Scatter of fluorescence in mother cell for constitutive mCherry and copy number marker CFP in three replicate experiments. Orange data points are included in the further analysis, whereas blue points were manually excluded (for further details, see Methods). **b**, Probability density function of orange data points in a are shown in black. Colored lines represent gene copy number estimates that

were calculated using a Gaussian mixture model (for further details, see Methods). **c**, The time series of the amplification marker fluorescence (growth normalized) for each mother cell was automatically classified into four categories. **Green** - no transition. **Light blue** – transition, but the transition was too close to the start or end of the experiment in order to determine if it was transient or not. **Dark blue** – transition considered to be stable. This number was verified by inspecting microfluidics movies (Table S1) and used to calculate the lower bound of the recombination rate. **Yellow –** more complex behavior, multiple fast transitions, oscillations. **d**, Probability distribution of observing zero, one, or more independent recombination events, which lead to a change in copy number (see Methods).



Supplementary Figure 6. Mathematical model is not very sensitive to experimentally measured parameters. a, Error of fitting for varying different parameters: gene amplification and duplication rate (left); growth rates, shown as generations per day, in high- and low expression environment (high EE, low EE) (middle), and in FACS tubes (right).

When a set of two parameters is varied, all other parameters remain fixed. Error of fit of two different experiments is shown in top and bottom. The error of fitting is defined as the average squared difference between experimental and simulated data point. Values that we measured in independent experiments and are used in our simulations are marked by a red x. **b**, An example of a rare environment where the preceding environment is long enough such that the gene copy number distribution does not change. The time to response is defined as the time needed by the population after environmental switch to achieve response R=M/2=5. **c-f**, Time to respond as a function of amplification- c, and duplication d, rate, fitness costs of expression e, and population size f, for either switching from low to high expression environment (full line), or from high to low expression environment (dashed line). **g**, Relative growth rate for different choices of maximum number of gene copies, *M*, for low expression environment (dashed lines), and high expression environment (full line). **h**, An example of all combinations of two sister chromosomes undergoing homologous recombination and splitting six gene copies among themselves. In all plots, unless stated otherwise, we use relative growth rates as shown in Fig. 2.2d, and amplification and duplication rates of k_{rec}^{0}=1.34x10^{-2} and k_{dup}=10^{-4} per cell per generation, respectively.



Supplementary Figure 7. Robustness of AMGET with respect to varying model parameters. **a**, The response *R* as the function of the two environment durations for three different expression costs (from left to right: 0.8, 0.5, 0.2). With decreasing cost of expression, AMGET effectively slows down and increases the environmental durations required to observe a visible response increase. This behavior leads to a predictable outcome in the limit of vanishing expression cost, where the population remains in the high expression state forever and thus no regulation via AMGET is needed. **b-c**, Population response generated by AMGET is robust to large variations in the recombination and duplication rate. Maximal variation in response (color scale), defined as ΔR=max(R)-min(R), for varying recombination rate b, and duplication rate c, for a set of environmental durations. We densely sample the parameter ranges for basal recombination rate, k_{rec}^{0} (see Methods), in the range of 10^{-4} – $5x10^{-2}$ and duplication rate in the range of 10^{-5}-10^{-3}, and find the largest and smallest response within this range to compute ΔR. The recombination rate mostly affects *R* around the narrow range of environment durations near the switch from no response to full response. For shorter environment durations, amplifications do not have enough time to sweep through a population and hence no response is achieved for any realistic recombination rate. Conversely, for longer durations enough time has passed in each environment that a response will always be maximal, except for recombination rates above 10^{-2}, which dampen the response as mutation decreases the efficacy of selection. The duplication rate only affects the response for environmental durations close to the switch

from no response to full response, and for low expression environments of a long duration. This is because the emergence of new duplications becomes rate-limiting after the low expression environment switches back to the high expression environment. In all plots, unless stated otherwise, we use recombination and duplication rates k_{rec}^{0}=1.34x10^{-2} and k_{dup}=10^{-4}, respectively. All rates have units per cell per generation. In our setup, one-day timescale is equivalent to between 10 and 23 generations (lower and upper bound, respectively; the bounds are estimated from the minimum and maximum growth rate of the least and best adapted copy number types, Table S2, Fig. 2.2d).

Supplementary Table 1. Verification of amplification events by detailed analysis of time-lapse microscopy images.

kymograph ID			nS2R2	trace description	event description
EXPERIMENT # 20170613					
S2	L	347	0.899	step down	deletion ~fr. 148
S7	R	191	0.800	step down	deletion along with short filamentation of daughter ~ fr. 190
S8	L	295	0.853	step up	amplification along with filamentation of daughter, mother stops dividing (EXCLUDED)
S8	R	189	0.84	step up	amplification along with transient filamentation of mother
S7	L	492	0.839	step down	deletion along with short filamantation of daughter ~ fr. 116
S9	L	476	0.836	step down	deletion ~fr. 133 along with short filamentation of mother and daughter, followed by deletion ~ fr. 146
S9	L	118	0.794	(small) step up	amplification fr. 128 followed by filamentation of daughter followed by another amplification ~ fr. 190
S4	L	268	0.786	step up	amplification with reciprocal deletion in daughter at fr. 116
S8	R	496	0.766	high peak	amplification fr. 226 another amplification fr. 278;
S6	L	647	0.735	step up	amplification fr. 107
	L	219	0.722	step down	filamentation ~ fr. 200, not a deletion (EXCLUDED)
S9	R	400	0.683	peak	amplification fr. 200, fr. 230, mother transiently filaments fr. 260
S6	R	337	0. 682	step down	deletion along with filamentation of daughter ~ fr. 107
S7	L	696	0.669	(small) step down	deletion ~ fr. 164; preceded by daughter amplification ~fr. 68
S7	L	185	0.661	step down	deletion in daughter 2+3 in fr. 114, transient filamentation of mother at fr. 179, after 1 division deletion fr. 211 with reciprocal amplification in daugther
S4	L	679	0.653	step down	transient mother filament fr. 92, deletion visible fr. 123
S9	L	297	0.629	(small) step up	amplification with reciprocal deletion in daughter at fr. 185
S4	L	704	0.621	step up	amplification with reciprocal deletion at fr. 184
S1	L	181	0.612	2 steps up	duplication with transient filamentation at fr. 184, duplication with reciprocal deletion at fr. 273
S6	R	235	0.610	step up	duplication with reciprocal deletion at fr. 213
S6	L	468	0.573	peak up, step down	duplication at fr. 115, few divisions, deletion at fr. 143
(19/21 verified amplification events, 6 reciprocal, 6 double events)					
EXPERIMENT # 20170614					
S4	R	498	0.963	step down	deletion with reciprocal fluorescence gain in daughter cell fr. 80-94

S8	R	350	0.938	step up high CFP	mother stops growing (EXCLUDED)
S6	R	905	0.927	step down	deletion at fr. 125 with v. transiently elongated first daughter
S11	L	198	0.927	step down	deletion at fr. 122 with v. transiently elongated first daughter
S6	L	33	0.891	step up	amplification at fr. 195, mother stops growing, but daughters grow
S3	R	86	0.887	step down	deletion at fr. 187, transiently elongated first daughter
S11	R	140	0.878	step up	amplification with reciprocal deletion fr. 117, normal cell division
S6	R	442	0.875	step down	deletion at fr. 155 with transiently filamenting mother and first daughter cell
S11	R	37	0.865	step down	deletion with reciprocal duplication at fr. 188
S5	L	366	0.858	step down	deletion around fr. 57 with transiently filamenting mother cell
S11	R	344	0.847	step up	amplification around fr. 189
S10	L	45	0.840	small step down	deletion with reciprocal amplification of daughter at fr. 188
S4	L	390	0.792	step up, then down	amplification with reciprocal deletion at fr. 141, replacement of mother by daughter, second deletion around fr. 180
S5	R	214	0.774	step up	amplification at fr. 164; another at fr. 236 with transiently filamenting first daughter, mother elongated and divides slowly, daughters normal
S9	R	551	0.756	step up	mother filaments and stops growing/diving (EXCLUDED)
S6	R	494	0.750	step down	deletion at fr. 102 (reciprocally brighter last daughter)
S5	R	242	0.757	step up	amplification with reciprocal deletion at fr. 185, amplification again at fr. 210, very bright first daughter filamentous
S8	R	530	0.695	step down	deletion with reciprocal daughter amplification at fr. 164
S11	R	472	0.695	step down	deletion around fr. 203 with transiently filamenting mother cell
S1	L	891	0.693	step down	amplification in first daughter at fr. 147, deletion in mother cell around fr. 184
S3	R	318	0.684	step down	deletion preceded by filamentous daughter cell fr. 197
S3	R	292	0.675	step down	deletion with reciprocal amplification in daughters at fr. 49
S3	R	370	0.656	step up	amplification with reciprocal deletion fr. 129
S1	R	456	0.646	step up	amplification with reciprocal deletion at fr. 219, further amplification at fr. 250, mother cell stops growing eventually
S5	R	499	0.642	step up	transient filamentation of the mother cell at fr. 143 prior to amplification
S1	R	148	0.631	step up	amplification with reciprocal deletion at fr. 62
S10	L	482	0.584	step up	amplification ~fr. 220
S1	R	662	0.567	step up	amplification with reciprocal deletion ~fr. 183 (transient displacement of mother cell)
S10	L	611	0.550	step down	transient filamentation of daugther cell at fr. 171 prior to visible deletion at ~fr. 200
S10	R	683	0.545	step down	deletion at fr. 32, another deletion fr. 183

S3	L	293	0.522	step down	filamenting daughter at 45; mother not fully in channel, normal growth
S3	L	729	0.516	step down	deletion with reciprocal amplification in daughter at fr. 148
S10	R	119	0.505	step down	not clear (EXCLUDED)

(30/ 33 verified amplification events, 14 reciprocal, 5 double events)

EXPERIMENT # 20170718

S8	L	375	0.907	(small) step down	deletion with reciprocal amplification in daughter at fr. 202
S9	L	36	0.904	(small) step down	deletion with filamentation of daughter cell at fr. 196
S7	L	935	0.875	step down	deletion with reciprocal amplification in daughter cell at fr. 152
S6	L	802	0.868	step down	deletion at fr. 116
S7	R	860	0.834	step up	amplification at fr. 232; mother gets brighter (2nd amplification?), eventually filaments, daughters grow
S10	R	292	0.822	step down	deletion at fr. 178
S6	L	955	0.817	step down	deletion at fr. 96
S10	L	595	0.792	step down	deletion along with transiently filamenting mother cell at fr. 242
S6	L	879	0.761	step up	amplification along with transiently filamenting mother cell at fr. 169
S4	R	826	0.755	(small) step down	deletion along with transiently filamenting mother at fr. 165; cells partially stacked due to wide channel
S7	L	884	0.747	step up with peak	amplification of the daughter cell at fr. 191, with simultaneous copy gain of mother cell
S7	L	348	0.747	(small) step down	deletion with reciprocal copy gain at fr. 65
S7	L	92	0.642	step down	deletion at fr. 74
S5	R	107	0.592	step down	deletion at fr. 139; cells transiently stack in the channel shortly afterwards
S6	R	110	0.545	rugged trace up	amplification at fr. 220, another at 257, imperfect focus, cells stacked
S5	R	439	0.542	step up	amplification with reciprocal deletion at fr. 119; cells partially stack in wide channel
S4	R	852	0.524	step up	amplification at fr. 153; cells partially stacked in wide channel
S5	R	923	0.522	step up	amplification at fr. 74

(18/18 verified amplification events, 4 reciprocal, 0 double events)

Supplementary Table 2. Model Parameter Values. LEE – low expression environment, HEE – high expression environment. Reported values represent the mean of triplicate experiments.

Parameter values	symbol	Value	obtained from
Max number of copies	M	10	qPCR & microfluidics (methods)
LEE FACS tubes generation per day	T20	10.4	growth rate in culture tube
HEE FACS tubes generations per day	T10	14.7	growth rate in culture tube
LEE generations per day	T2	23	dilution series in 96-well plates
HEE generations per day	T1	22	dilution series in 96-well plates
recombination rate ($cell^{-1}$ $gen.^{-1}$)	k_{rec}^{0}	0.0134	microfluidics experiment (methods)
duplication rate ($cell^{-1}$ $gen.^{-1}$)	k_{dup}	0.0001	(Anderson and Roth, 1981; Mats E. Pettersson *et al.*, 2009; Reams *et al.*, 2010; Sun *et al.*, 2012)
Relative growth rates in HEE	s^{HEE}_{1}	0.46	flow cytometry experiment (fitness landscape; Fig. 2.2d) in combination with growth rate measurement of the fittest copy number (S^{HEE}_{10}), (methods)
	s^{HEE}_{2}	0.45	
	s^{HEE}_{3}	0.51	
	s^{HEE}_{4}	0.57	
	s^{HEE}_{5}	0.62	
	s^{HEE}_{6}	0.68	
	s^{HEE}_{7}	0.74	
	s^{HEE}_{8}	0.78	
	s^{HEE}_{9}	0.81	
	s^{HEE}_{10}	1	
Relative growth rates in LEE	s^{LEE}_{1}	1	flow cytometry experiment (fitness landscape; Fig. 2.2d) in combination with growth rate measurement of the fittest copy number (S^{LEE}_{1}), (methods)
	s^{LEE}_{2}	0.94	
	s^{LEE}_{3}	0.84	
	s^{LEE}_{4}	0.74	
	s^{LEE}_{5}	0.67	
	s^{LEE}_{6}	0.62	
	s^{LEE}_{7}	0.57	
	s^{LEE}_{8}	0.53	
	s^{LEE}_{9}	0.50	
	s^{LEE}_{10}	0.44	

Supplementary Table 3. Oligonucleotides

Name	Sequence	Purpose
mgl_keio_f	TTTATGACCGAATGCGGACCACATTCACATCATTTCT TACGCGCGTATTTTGTAGGCTGGAGCTGCTTCG	mglBAC deletion
mgl_keio_r	AGCATTTATCTCAAGCACTACCCTGCATAAGAAAAA CCGGAGATACCATGATTCCGGGGATCCGTCGACC	mglBAC deletion
galP_replace_p_f	CAATAACATCATTCTTCCTGATCACGTTTCACCGCAG ATTATCATCACAATGTAGGCTGGAGCTGCTTCG	replace pgalP with constitutive promoter
galP_replace_p_r	AAAAACGTCATTGCCTTGTTTGACCGCCCCTGTTTTT TAGCGTCAGGCATGGTACCTTTCTCCTCTTTAATC	replace pgalP with constitutive promoter
D_int_f	AATTTTTATCAAAAAAATCATAAAAAATTGACCGGT TAGACTGTTAACAAGGAAAGACGGGCTTCAA	integration of reporter gene cassette for evolution into locus 1
D_int_r	ATACGGTGCGCCCCCGTGATTTCAAACAATAAGTAG CCAAAAGGTGAATACGAGGCCTTATGCTAGCT	integration of reporter gene cassette for evolution into locus 1
E_int_f	CCATGTCCCCGAACAAGTGTTCACTATGTCCCCGGA CCGTACACCCCAAACCGGAAAGACGGGCTTC	integration of reporter gene cassette for evolution into locus 2
E_int_r	TCGGAAGGGAAGAGGGAGTGCGGGAAATTTAAGC TGGATCACATATTGCCGAGGCCTTATGCTAGCTTC	integration of reporter gene cassette for evolution into locus 2
D_H1P1	AATTTTTATCAAAAAAATCATAAAAAATTGACCGGT TAGACTGTTAACAACCCGCCATTCAGAGAAGAAA	integration of cassette for testing into locus 1
D_H2P2	ATACGGTGCGCCCCCGTGATTTCAAACAATAAGTAG CCAAAAGGTGAATATTGTCTCATGAGCGGATACA	integration of cassette for testing into locus 1
D_flank_f	TTCACTCCTGCTCCCTTC	cassette integration test locus 1
D_flank_r	GCGAACGTCATCTGGTGGTG	cassette integration test locus 1
E_flank_f	GCTGGAGCCACTTGTAGCC	cassette integration test locus 2
E_flank_r	TCCTTGCTGAATCATTTTGTTC	cassette integration test locus 2
pMS*_add_XhOI_5 UTR_F	AAAACTCGAGGTCGACAGGAGGAATTCACCATG	plasmid construction, adding XhoI into pMS
pMS*_add_XhoI_p 0_R	AAAACTCGAGGGAAATTAACGGACGGCCTC	plasmid construction, adding XhoI into pMS
pMS*_XmaI_add_p 0_F	AAAAACCCGGGACCGGAAAGACGGGCTTCAAAGC	plasmid construction, adding XmaI into pMS
pMS*_XmaI_add_o ri_R	AAAAACCCGGGGCGGCCGCAAGATCCGG	plasmid construction, adding XmaI into pMS
5delRecA	TGACTATCCGGTATTACCCGGCATGACAGGAGTAA AAGGGGATCCGTCGACCTGCAGTT	deletion of recA
3delRecA	AAGGGCCGCAGATGCGACCCTTGTGTATCAAACAA GACGATGTAGGCTGGAGCTGCTTC	deletion of recA
rbsB_qPCR_Fw	GGCACAAAAATTCTGCTGATTAA	qPCR control locus
rbsB_qPCR_Rv	GCAGCTCGATAACTTTGGC	qPCR control locus
CFP_qPCR2_Fw	AGCATTGAACACCCCAGG	qPCR amplified locus

CFP_qPCR2_Rv	CTGTTTACTGGTGTGGTTCCTA	qPCR amplified locus
D_ujunFind_F	GGGGCTTATTAAGAGGATCT	finding junction of amplified region locus 1
D_ujunFind_3R	TCCATCCTGCAACGTTAT	finding junction of amplified region locus 1
D_ujunFind_2R	GGCATCCTCGATTCCTTC	finding junction of amplified region locus 1
D_ujunFind_1R	GTGGATGAAGCTGCTAAC	finding junction of amplified region locus 1
D_djunFind_R	CATAAGGCCTCTATTCACCT	finding junction of amplified region locus 1
D_djunFind_3F	ATGCAGTAAGTCCAGGATC	finding junction of amplified region locus 1
D_djunFind_2F	TTGATAAGTACATGCTGGAGA	finding junction of amplified region locus 1
D_djunFind_1F	TAAAACATGGTGATTGCCTC	finding junction of amplified region locus 1
IS1C_flank	TACAAAGGTGGAGGCAAACC	finding junction of amplified region locus 2
IS1B_flank	GATTCACCGGCTCATTCAGT	finding junction of amplified region locus 2
rho_seq_f	TCCTGCCATACCATTCACAA	rho sequencing
rho_seq_r	TAACATGCCAGCAAATTCCA	rho sequencing

Supplementary Table 4. Plasmids.

Name	Purpose	Source
pZA21-*yfp*	source for *yfp* in pMS6*	(Lutz and Bujard, 1997)
pKD13	kan template for recombineering	(Datsenko and Wanner, 2000)
pMS7	starting point for construction of the gene cassette for evolution, pir dependent replication	(Steinrueck and Guet, 2017)
pMS6*	gene cassette template for recombineering, pir dependent replication, pir dependent replication	this study
pMS1	template for constitutive *galP* promoter (fragment J23100) based on pKD13	lab collection
pBAD24	basis for pIT07	(Guzman *et al.*, 1995)
pIT07	gene cassette template for recombineering, based on pBAD24	this study

Supplementary Table 5. Bacterial Strains.

Strain name	Genotype	purpose	Source
MG1655	$F^- \lambda^-$ ilvG- rfb-50 rph-1	strain background for all experiments except testing experiment	lab collection
BW27784	lacIq rrnB3 ΔlacZ4787 hsdR514 DE(araBAD)567 DE(rhaBAD)568 DE(araFGH) Φ(ΔaraEp PCP4A–araE)	background for testing experiment (Fig. 2.1b) strain construction	(Khlebnikov *et al.*, 2001)
IT013	BW27784, JA23100::*galP* , *mglBAC* ::FRT, *galK* ::FRT	background for testing experiment (Fig. 2.1b) strain construction	this study
IT013-TCD	BW27784, JA23100::*galP* , *mglBAC* ::FRT, *galK* ::FRT, locus1::pBAD-*galK*	strain background for testing experiment (Fig. 2.1b)	this study
MS022	MG1655, JA23100::*galP* , *mglBAC* ::FRT, *galK* ::FRT	background for evolution experiment strain construction	lab collection
JW0740-3	F-, Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ-, ΔtolC732::kan, rph-1, Δ(rhaD-rhaB)568, hsdR514 Δ*galK* 729::kan	source for *galK* deletion	(Baba *et al.*, 2006)
BW25142	lacIq rrnB3 (lacZ4787 hsdR514 DE(araBAD)567 DE(rhaBAD)568 (phoBR580 rph-1 galU95 (endA9 uidA((MluI)::pir-116 recA1	host for pir plasmids pMS6* and pMS7	(Haldimann and Wanner, 2001)
IT028	MS022 locus1::p0-RBS-*galK*-RBS-*yfp*-FRT-pR-*cfp*	ancestor strain for evolution experiment (Fig. S1b,c)	this study
IT030	MS022 locus2::p0-RBS-*galK*-RBS-*yfp*-FRT-pR-*cfp*	ancestor strain for evolution experiment (Fig. S4b)	this study
IT028-EE1-D8	IT028 dup(atpB-rsbD), rho (S265>A)	amplified strain locus 1, evolved in evolution experiment (Fig. S1b,c)	this study
IT028-EE1-D8-recA	IT028 dup(atpB-rsbD), rho (S265>A), ΔrecA	ΔrecA-stabilzed version of amplified strain IT028-EE1-D8 (Fig. S3d,e)	this study
IT028-EE11-D4	IT030 dup(IS1B-IS1C)	amplified strain locus 2, evolved in evolution experiment (Fig. S4b)	this study
IT028-EE1-D8-pRmCherry	MS022 dup(locus1::p0-RBS-*galK*-RBS-*yfp*-FRT-pR-*cfp*), attP21::pR-mCherry	amplified strain locus 1, for microfluidics (Fig. 2.3b,c)	this study
IT034	IT028 attP21::pR-mCherry	ancestral strain in co-culture experiments (Fig. 2.4b,d)	this study
IT028-H5r	MS022 locus1::pconst-RBS-*galK*-RBS-*yfp*-FRT-pR-*cfp*	constitutive strain in co-culture experiments (Fig. 2.4b,d)	this study