Tampere University

VATSALA CHAUHAN

Bacterial Regulatory Mechanisms of Gene Expression During Stress Responses with Focus on Closely Spaced Promoters

Tampere University Dissertations 886

Tampere University Dissertations 886

VATSALA CHAUHAN

Bacterial Regulatory Mechanisms of Gene Expression During Stress Responses with Focus on Closely Spaced Promoters

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Medicine and Health Technology
of Tampere University,
for public discussion in the Jarmo Visakorpi auditorium
of the Arvo building, Arvo Ylpön katu 34, Tampere,
on 7 Nov 2023, at 12 o'clock.

ACADEMIC DISSERTATION
Tampere University, Faculty of Medicine, and Health Technology
Finland

Responsible supervisor and Custos	Professor Andre S. Ribeiro Tampere University Finland	
Pre-examiners	Professor Tobias Bollenbach University of Cologne Germany	Professor Ulrike Endesfelder University of Bonn Germany
Opponent	Professor Remus T. Dame Leiden University The Netherlands	

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Cover design: Roihu Inc.

ISBN 978-952-03-3111-5 (print)
ISBN 978-952-03-3112-2 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-3112-2

HIILINEUTRAALI PAINOTUOTE
ClimateCalc CC-000025/FI
PunaMusta Printing

Carbon dioxide emissions from printing Tampere University dissertations have been compensated.

PunaMusta Oy – Yliopistopaino
Joensuu 2023

Dedicated to my kind parents

ACKNOWLEDGEMENTS

This Ph.D. study, conducted under the Doctoral Program of Biomedical Sciences and Engineering, was carried out in the Laboratory of Biosystem Dynamics (LBD) of the Faculty of Medicine and Health Technology (MET) of Tampere University, under the supervision of Professor Andre Ribeiro.

Andre, I would not be writing this preface if it wasn’t for you. It has been an extraordinary six-year odyssey, one that I could not have made without your support. Thank you for teaching me all that I know today. Whatever caliber scientist I may become in the future, I will owe that to you.

My sincere thanks to all the funders who invested in my research. First, the EDUFI fellowship, which supported the budding year of my Ph.D. studies. Then, the Tampere University Graduate Program which supported the middle years. Finally, the project grants of ASR which supported the final year, specifically, the Jane and Aatos Erkko Foundation and the Sigrid Jusélius Foundation.

Next, I would like to thank the pre-examiners, Professor Tobias Bollenbach of the University of Cologne, and Professor Ulrike Endesfelder of the University of Bonn, for their insightful comments on my thesis. Also, I would like to thank Professor Remus T. Dame from Leiden University for acting as the opponent during the public examination of the thesis.

I also thank Professor Fred G. Biddle for being my follow-up group member and accessing the progress of my Ph.D., especially during the initial years of my Ph.D.

I would like to acknowledge the assistance provided by the administrative staff members of the faculty. Especially, Dr. Henna Mattila, I thank you for your support during my Ph.D. submission process.

I am thankful to the LBD members that I encountered when joining, who I look up to, and whose support has been instrumental in supporting my grant applications. Dr. Vinodh Kandavalli, thank you for teaching me most of the things in the lab. Dr. Sofia Startceva, thank you for your insights on my progress every year as my Follow-up group member. Dr. Samuel Oliveira, thank you for answering my questions with patience. Dr. Ramakanth Neeli, Dr. Nadia Goncalves, Dr. Marco Minoia, and Dr. Mahendra Prajapat, thank you for the scientific discussions.

I am even more thankful to my colleagues who recently finished their Ph.D. at the LBD, with whom I have had the privilege to grow and spend the longest time of my stay in Finland. Dr. Cristina Palma, your art of rationale approach has helped me to navigate ideas. Thank you for all your support and cheering for each other. Dr. Mohammed Bahrudeen, thank you for your wonderful collaboration on two main projects. Working side by side with you has been an incredible experience. Bilena Almeida, your hard-working nature has inspired me greatly.

Finally, to my fellow colleagues who are all not just hardworking colleagues but also great friends. Ines Baptista, your kindness, smartness, and friendship have been one of the biggest sources of strength and courage. You make the world, and my life a better place. Your commitment to work has been awe-inspiring. Suchintak, you made the lab experience amazing and truly worthwhile. I have no doubt that you will carry the lab legacy forward with sincerity. Rahul Jagadeesan, it has been a pleasure to know you, share office, and see you work diligently. Thank you for being my guide in the gym after work, these sessions have added so much joy in the finishing phase of my Ph.D. and you have been truly inspirational.

Lastly, I would like to express appreciation to my family, and friends. Their belief in my abilities has been a constant source of motivation. Dr. Surabhi has been my cheerleader and an example of intelligence for the last 10 years. Dr. Swaleha, thanks for your love and prayers for me. Nikhat, Raishal, Dipi, Dr. Suniti, Dr. Devansh, and Dr. Shreya, thanks for your constant support and friendship. They have kept me going in the dark Finnish winters.

To the friends that I found in Finland. Sheba, you were an example of bravery and smartness. Karan, the wise friend whom I go to when in need of enlightenment. I am glad that I had you both as my family away from home, whom I could trust with everything. Next, the Mishra family (Krishna, Shradha, and a little boss Aarna) thank you! You were always a family that I could go to after work. Shambhavee, I found a sister in you. That homely comfort feeling, and the sleepovers, made Finland more of a home. Elisa, you were the best neighbor one could ever find in Finland. Lingyu, Suvi, Suji, Sanni, and Ankita thank you for the love and care.

I thank my parents eternally, for sending me to good schools and colleges even when it was at par with their financial abilities. I would not have the courage to dream the dreams that I carry, without the faith and the confidence that you have in me. Sid, thank you for being the big brother, and for the support, utmost care, and trust that you have in me. Komal, thank you for bringing so much joy to our family. What fabulous uncles, aunts, and cousins I have been given who constantly reached out to me. Bua, thank you for being my second mother, whom I could trust with everything. Amma and Nani, I wish to be like you when I am old. Nanaji and Bauji, I hope that you are proud of me in your heavenly abode.

Vinodh, నా జీవితం యొక్క ప్రేమ మరియు కాంతి, your brain, and kindness have inspired me every single day in this journey. What a delight it has been to be able to share the highs and lows of this journey with you. You have been my biggest cheerleader and my most honest critic. Thank you for feeding me the best meals of my life. Thank you for standing the test of time, distance, and patience with me. I, too, look forward to the journey of forever by your side.

It is the journey (with its highs and lows) that makes you - the destination is incidental, and an outcome of the choices made.

Vatsala, Tampere 23’

ABSTRACT

Bacteria are exposed to changing environments and other stresses, such as antibiotics. Some of these events can even be lethal. Their phenotypic adaptations to these stresses are driven by internal mechanisms of gene regulation that, therefore play a fundamental role in their survivability. The core mechanism of gene regulation is arguably the promoter regions. These are DNA sequences that largely determine whether a gene is sensitive to specific transcription factors, supercoiling fluctuations, and other regulatory factors and events.

In this thesis, we used *Escherichia coli* as a model organism to study bacterial mechanisms of genome-wide expression regulation during stresses, focusing on the promoters in tandem formation. For that, we started by developing a novel method to determine single-cell distributions of RNA numbers from flow cytometry data. This method can predict the moments of the distribution of the single-cell RNA numbers from the moments of the distribution of total fluorescence of cells expressing the proteins that the RNAs code for. This greatly facilitates the study of transcription dynamics using large numbers of cells as a source of data.

Next, we used a large strain library of tagged genes controlled by tandem promoters. From the single-cell distributions of their protein levels under different stress conditions, we dissected the main features controlling the kinetics of overlapping tandem promoters. Specifically, we identified the distance between start sites and the dynamics of the transcription initiation at each promoter, as the main factors.

Finally, we designed and constructed a strain library of synthetic genes controlled by non-overlapping tandem promoters. We used them to validate, by proof of concept, that they can be used to engineer genes with predictable dynamics. Moreover, we identified a key variable controlling these constructs, namely, the strength of the downstream promoter, which acts as the main limiting factor of the overall transcription rate. Overall, this study dissected important regulatory features of tandem promoters. The findings facilitate their use as building blocks of future synthetic circuits.

CONTENTS

ABBREVIATIONS

aTc	Anhydrotetracycline
bp	Base pair
CC	Closed complex
DEG	Differentially expressed gene
DNA	Deoxyribonucleic acid
d_{TSS}	Distance between transcription start sites
E. coli	*Escherichia coli*
EF	Elongation Factor
FITC	Fluorescein isothiocyanate
FSC	Forward scatter
GFP	Green Fluorescent Protein
GR	Global regulator
HILO	Highly inclined and laminated optical sheet
IF	Initiation Factor
IPTG	Isopropyl β-D-1-thiogalactopyranoside
mRNA	Messenger RNA
OC	Open complex
PMT	Photomultiplier tube
RBS	Ribosome Binding Site
RFP	Red Fluorescent Protein
RNA	Ribonucleic acid
RNAP	RNA polymerase
RPc	RNA polymerase in closed complex formation
RPo	RNA polymerase in open complex formation
rRNA	Ribosomal RNA
TEC	Transcriptional elongation complex
TF	Transcription factor
TFN	Transcription factor network

TI	Transcriptional interference
TIRF	Total internal reflection fluorescence
tRNA	Transfer RNA
TSS	Transcription Start Site
YFP	Yellow Fluorescent Protein

ORIGINAL PUBLICATIONS

This thesis is a compilation of four publications. This thesis refers to them as **Publications I, II, III, and IV**. The publications are reproduced with permission from the publishers.

Publication I **V. Chauhan***, M.N.M. Bahrudeen*, C.S.D. Palma, S.M.D. Oliveira, V. Kandavalli, A.S. Ribeiro (2019) Estimating RNA numbers in single cells by RNA fluorescent tagging and flow cytometry. *Journal of Microbiological Methods*, 166, 105745, 2019. DOI: 10.1016/j.mimet.2019.105745. *Equal contributions.

Publication II B Almeida, **V Chauhan***, MNM Bahrudeen*, S Dash*, V Kandavalli, A Häkkinen, J Lloyd-Price, CSD Palma, ISC Baptista, A Gupta, J Kesseli, E Dufour, O-P Smolander, M Nykter, P Auvinen, HT Jacobs, SMD Oliveira, and AS Ribeiro (2022) The transcription factor network of *E. coli* steers global responses to shifts in RNAP concentration. *Nucleic Acids Research,* 50(12), 6801-6819. DOI: 10.1093/nar/gkac540. *Equal contributions.

Publication III **V. Chauhan***, M.N.M. Bahrudeen*, C.S.D. Palma, I. Baptista, B.L.B. Almeida, S. Dash, V. Kandavalli, and A.S. Ribeiro (2022) Analytical kinetic model of native tandem promoters in *E. coli. PLOS Computational Biology*, 18(1). DOI: 10.1371/journal.pcbi.1009824. *Equal contributions.

Publication IV **V. Chauhan,** I.S.C Baptista, R. Jagadeesan, S. Dash, and A.S. Ribeiro (2023) Using synthetic tandem promoter formations to tailor genes with desired dynamics (Submitted manuscript).

* The authors contributed equally and can list their names first as described in the publications.

The following is a description of the contribution to each publication from the author of this thesis.

In **Publication I**, the author conceived the study with M.N.M. Bahrudeen and A.S. Ribeiro. The author was responsible for planning and performing all laboratory experiments. The author contributed to the analysis of the results. Finally, the author contributed to the writing of the manuscript.

In **Publication II**, the author contributed to the planning and execution of the flow-cytometry, microscopy, western blotting, growth curves, and RNA-seq experiments. The author actively participated in the discussion of the results. In addition, the author contributed to the writing of the experimental methods and the revision of the manuscript.

In **Publication III**, the author conceived the study with M.N.M. Bahrudeen and A.S. Ribeiro. The author proposed, designed, and standardized all the experimental protocols and then conducted all laboratory measurements. Furthermore, the author actively participated in the analysis and discussion of the results. Finally, the author co-wrote the manuscript.

In **Publication IV**, the author conceived the study with A.S. Ribeiro. The author designed all the synthetic constructs and then standardized all laboratory measurements. The author analyzed and discussed the results. Finally, the author co-wrote the manuscript.

Publications I and **II** have been included in the Ph.D. dissertation of M.N.M. Bahrudeen and B.L.B Almeida. **Publication III** has been included in the Ph.D. dissertations of C.S.D Palma and M.N.M. Bahrudeen. **Publication IV** is not included in any other thesis.

1 INTRODUCTION

Bacterial genes are subject to strict regulation (Stoebel, Hokamp, Last, & Dorman, 2009), in order for the cells to be able to control when a specific gene will be activated or repressed. Regulation can also be exerted on the noise in expression rates (Elowitz, Levine, Siggia, & Swain, 2002; Kaern, Elston, Blake, & Collins, 2005). These evolved regulation mechanisms play a critical role in the ability of cells to carry out complex transcriptional programs that make possible their adaptation to stresses (Kussell & Leibler, 2005; Stoebel et al., 2009).

Studies have shown that transcription initiation is a major checkpoint in the process of regulating the expression of a gene (Browning & Busby, 2004, 2016; McLeod & Johnson, 2001). Several rate-limiting steps occur during transcription initiation (for a review see (Häkkinen & Ribeiro, 2016)). The rate of these steps differs with factors such as the promoter sequence in general (for a review see (McClure, 1985)), promoter specificity to σ factors (Kandavalli, Tran, & Ribeiro, 2016), promoter configuration, supercoiling, global regulators, specific transcription factors, among others (for a review see (Häkkinen & Ribeiro, 2016)). By tuning each these factors, the rate-limiting steps can be controlled, which, in turn, affects both mean and noise in Ribonucleic acid (RNA) levels (Kaern et al., 2005).

While mean expression levels could be measured using techniques such as northern blotting (Alwine, Kemp, & Stark, 1977), quantification of noise in gene expression was first made possible by the use of fluorescent proteins and fluorescence microscopy (and currently flow-cytometry). The measurements showed how much protein numbers can differ even between sister cells (Elowitz et al., 2002; Kaern et al., 2005). These methods, along with the presently vast libraries of tagged fluorescent proteins and synthetic probes, currently allow the *in vivo* counting and tracking of the majority of natural proteins of *Escherichia coli* (*E. coli*) (Baba et al., 2006; Endesfelder, 2019; Taniguchi et al., 2010; Yu, Xiao, Ren, Lao, & Xie, 2006; Zaslaver et al., 2006). There is, however, measurement noise associated with each of these techniques that need to be considered, including in flow-cytometry (Steen, 1992). For instance, fluorescent proteins can be out of focus. This makes collecting precise data difficult, particularly in time-lapse microscopy measurements (Golding, Paulsson, Zawilski, & Cox, 2005; Häkkinen & Ribeiro, 2014).

While several regulatory mechanisms of gene expression have been discovered, the ones that are best characterized are those based on transcription factors (TFs), including global regulators (GRs) (Razo-Mejia et al., 2018). Meanwhile, there are many mechanisms directly embedded in the Deoxyribonucleic acid (DNA) sequence, such as pause sequences, closely spaced promoters, and highly supercoiling sensitive sequences (Palma et al., 2020; Sneppen et al., 2005). While these are known to exist, the quantification of how their specific DNA sequences affect the dynamics remains challenging. Importantly, they are likely to play major roles, for example, during genome-wide stress responses, potentially differentiating the response of their hundreds of downstream genes from that of other genes of *E. coli*.

Using *E. coli* as a model organism, this thesis focuses on the investigation of promoters' arrangement in tandem formations as a structural regulatory mechanism of gene expression. For this, we subject cells to stresses that include quick changes in RNAP levels and antibiotics. We expect that our results will help engineering synthetic genetic circuits with kinetics predictable from their structure.

2 LITERATURE REVIEW

2.1 *Escherichia coli* and its use as a Model Organism

Escherichia coli (*E. coli*) (Figure 1), which is a rod-shaped and gram-negative bacterium, was discovered by Theodor Escherich (Escherich, 1988) in 1886, and is found in the normal gut flora of human beings. A typical *E. coli* cell measures between 2 and 4 μm in length and 0.5 to 0.8 μm in width (Volkmer & Heinemann, 2011), and can divide once every 30 minutes in optimal conditions.

There are a number of advantages in using *E. coli* as a model organism, including their simple nutritional requirements, the rapid growth rate, and the ability to survive both with and without oxygen. In agreement, *E. coli* is easy to maintain and breed in laboratories. Another advantage is their well-established genetics (Ullmann, 2011). For these reasons, *E. coli* is the most common choice for researchers to study life-sustaining biological processes in bacteria (Elowitz et al., 2002; Golding et al., 2005; Hufnagel, Depas, & Chapman, 2015; Lukačišinová, Fernando, & Bollenbach, 2020; Muthukrishnan et al., 2012; Zaslaver et al., 2006). Moreover, its extensive use has made possible the development of techniques to genetically clone modified strains of *E. coli* (Lukačišinová et al., 2020) to better understand biological processes and to test new concepts using genetic engineering.

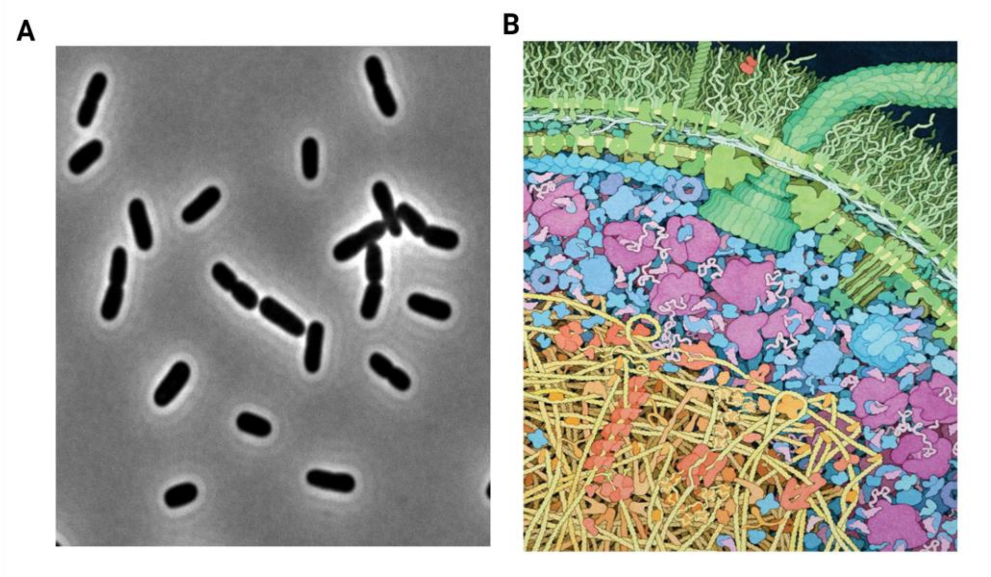


Figure 1. *E. coli*. **(A)** Image of *E. coli* cells taken by phase contrast when observed under microscope using an 100x objective. **(B)** An illustration of a cross-section of a small portion of an *E. coli* cell. The cell wall and transmembrane proteins are shown in green. Nucleoid region, DNA, and enzymes are shown in yellow, orange, and blue, respectively. The illustration was created by D.S. Goodsell (Goodsell, 2012).

E. coli cells contain a couple of thousands of protein-coding sequences (Blattner et al., 1997). It also contains several plasmids in the natural state. These plasmids are relatively small DNA sequences that are not located in the chromosome and can replicate independently. The plasmids typically carry genes for specialized functions, such as antibiotic resistance (Eliasson, Bernander, Dasgupta, & Nordström, 1992; Russo & Johnson, 2003). We made use of plasmids to introduce specific genes and fluorescent proteins in *E. coli* cells.

2.2 Gene Expression Machinery of *E. coli*

2.2.1 Promoters

Commonly, bacterial genes are composed of a promoter, followed by an RNA coding region, downstream of the promoter. The promoters are DNA segments to which RNAP can bind and initiate transcription at a transcription start site (TSS)

(Santos-Zavaleta et al., 2019). Aside from a TSS, promoters also contain operators. They can be used by transcription factors (TFs) to regulate transcription rates. Usually, TFs bind to the operators to either block (repressor) or stimulate transcription (activator). Usually, when a promoter is active, several RNAs copies are produced. From each RNA, several proteins are translated.

We made use of synthetic promoters $P_{Lac/ara-1}$, and $P_{LtetO-1}$, to control the production of RNA target for MS2-GFP, in order to estimate how their number in single cells relates to corresponding single cell protein numbers. For this, we used the same promoters to control the expression of fluorescent proteins. The first promoter, $P_{Lac/ara-1}$, is the result of combining the operator regions of the Arabinose promoter with the transcription start site and operator regions of the Lac promoter (R. Lutz & Bujard, 1997).

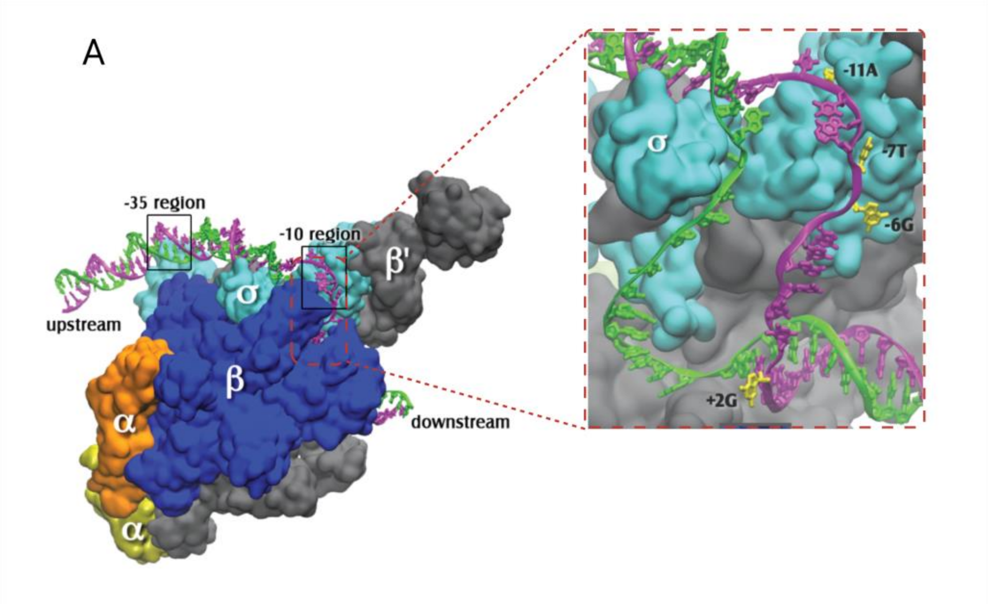
Meanwhile, we studied natural tandem promoters, using the YFP strain library (Taniguchi et al., 2010), that includes 102 strains where YFP is expressed by a tandem promoter. Of these we used 30 because in the other strains, the tandem promoters suffer from clear interferences by other gene coding regions or promoters. Finally, we made use of modified versions of the natural promoters $P_{Lac,}$ $P_{tetA,}$ and $P_{BAD.}$ Specifically, we designed new sets of tandem promoters, whose dynamics we predicted from the dynamics of the individual promoters.

2.2.2 Transcription

During transcription, the DNA sequence determines the future sequence of RNA produced (also known as a transcript). Transcription in bacteria has three main steps: initiation, elongation, and termination (Alberts et al., 2002). The enzyme responsible for RNA synthesis is RNA polymerase (RNAP). Roger D. Kornberg was awarded the Nobel Prize in 2006 for the structure of RNAP performing transcription (Cramer, Bushnell, & Kornberg, 2001). The RNAP has five subunits: two α subunits, a β subunit, a β' subunit, and an ω subunit, which forms the RNAP core enzyme (Haugen, Ross, & Gourse, 2008), which has a crab claw-like structure (Figure 2). The α subunit (36.5 kDa) is involved in RNAP assembly. The β (150 kDa) and β' subunits (155 kDa) form the jaws of the RNAP (Tagami, Sekine, & Yokoyama, 2011). This structure that these two jaws form acts as the DNA binding clamp. The smallest subunit of RNAP is ω (10.2 kDa). This subunit is the only one that was classified as non-essential for cell growth (W. Ross et al., 1993). Nevertheless, studies suggest that it acts as a catalyst, maintains the structure of the β' subunit, and responds to ppGpp (W. Ross et al., 1993; So et al., 2011; Weiss et al., 2017). As such, it has regulatory capabilities.

Due to being a global regulator, perturbing RNAP numbers should have genome-wide consequences. We studied this using RNA-seq technology**.** However, in addition to average concentrations in cell populations, we also considered that their single-cell numbers as well as their spatial distributions could be significant in our studies. As such, we made use of the RL1314 strain of *E. coli*, where β' subunit is tagged with GFP.

The RL1314 strain was engineered by (Bratton, Mooney, & Weisshaar, 2011), and generously offered to us, with the main purpose of measuring RNAP::GFP in single cells in four different LB media richness conditions. That was very helpful in our studies, where changes in RNAP::GFP levels were the key perturbations of the gene networks. In recent works, this strain has been used for several purposes by several teams, including to study the spatial distribution and diffusion of RNA Polymerase (Bratton et al., 2011), how RNA polymerase is redistributed to support cell growth (Fan et al., 2023), and novel models of genetic circuits (Barajas, 2022).



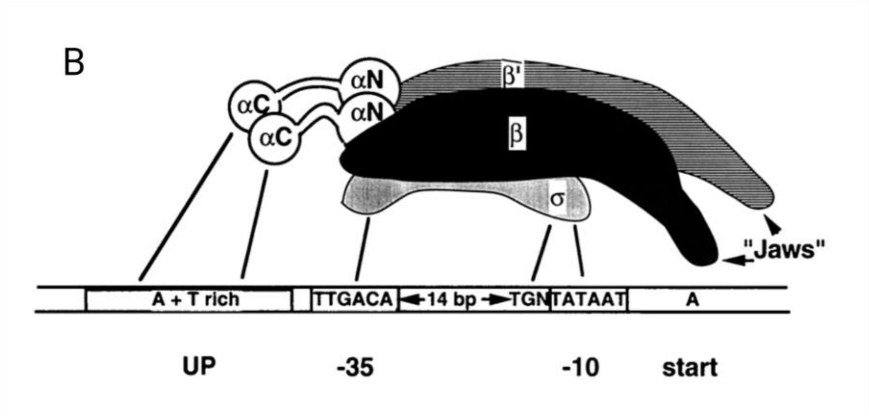


Figure 2. An illustration of the RNAP holoenzyme. **(A left)** Structure of an RNAP holoenzyme with its subunits including the sigma factor when interacting with the promoter region (-35 and -10 positions). **(A right)** An enlarged version of a small region of the A left picture, without the β subunit to show the transcription bubble. Picture taken from (Karpen & deHaseth, 2015) with permission. **(B)** A cartoon of the RNA polymerase interacting with a promoter region. Shown are the consensus sequences for the −10 and −35 regions (boxed) of the promoter. The jaws of the RNAP are shown on the right bottom region, and they allow the RNA polymerase to 'grab' the DNA. Picture was taken from (deHaseth Pieter L., Zupancic Margaret L., & Record M. Thomas, 1998) with permission.

Transcription initiates when the RNAP core enzyme binds to the DNA, with a σ factor, and forms the RNAP holoenzyme (Figure 3) (Murakami & Darst, 2003). In this formation, the RNAP occupies ~35 bp of DNA (deHaseth Pieter L. et al., 1998). We used this information to decipher the effects of tandem promoters distanced sufficiently close to allow for RNAP interference on gene regulation.

σ^{70} is the most commonly needed factor (Feklístov, Sharon, Darst, & Gross, 2014). σ factors recognize the promoter and ensure that the RNAP holoenzyme binds to the promoter. In order for RNAP to recognize a sequence, two specific sequences are needed. These are shown in the -35 and -10 'boxes', respectively (Figure 2B). These sequences of DNA define the exact position where transcription of a gene begins (referred to as transcription start site, TSS, which is numbered '+1'). Small variations in the sequences of these regions suffice to block recognition by RNAP (Alberts et al., 2002). When these regions are recognized, the RNAP holoenzyme and DNA form a complex, known as the closed promoter complex (RPc). It covers about 60 bp of DNA (Alberts et al., 2002). Upstream and downstream are towards the 5' and 3' ends, respectively, of the coding strand of the DNA. Next, the RNAP holoenzyme unwinds the double helix of the promoter, which leads to the formation of a structure known as the open promoter complex (RPo). The σ factor now dissociates from the holoenzyme, converting it into a core enzyme. The first phosphodiester bond of RNA is synthesized at this moment (Alberts et al., 2002).

Several recent studies have focused on the genome-wide consequences of changes in σ factor numbers. (John et al., 2022) investigated genome-wide assembly of RNAP at promoters while (Britton et al., 2002) studied the role of sigma H in the phase transition of *B. subtilis* to stationary. Interestingly, binding sites for σ factors constitute 9.1% of all known binding sites listed in PRODORIC (Dudek & Jahn, 2021). We quantified the role of changes in σ factor numbers in the genome-wide responses to shifts in RNAP concentrations.

Transcription elongation begins when RNAP slides along the DNA molecule, unspooling the DNA helix as it advances. As it does this, it attaches ribonucleotides to the 3' end of the growing RNA molecule (Figure 3). A transcriptional elongation complex (TEC) is formed by RNAP, DNA sequence, and RNA molecule. The TEC occupies ~35 bp of DNA (Alberts et al., 2002).

Once transcription elongation has started, only a limited region of DNA is melted at any given time. This region of the DNA forms a transcription bubble, which occupies 12-14 bp (Korzheva et al., 2000; Saecker, Record, & Dehaseth, 2011). Studies showed that there are variations in the rate of transcription as an RNAP moves along the DNA. Occasionally, the RNAP slows down, pauses, and reaccelerates (Fujita, Iwaki, & Yanagida, 2016; Greive & von Hippel, 2005; Lewin, 2008). The rate of occurrence of these events depends on the interactions between TEC, DNA, RNA, and other regulatory molecules (Greive & von Hippel, 2005). Pauses usually last from seconds to a couple of minutes (Herbert et al., 2006). Pauses can also be due to collisions between RNAPs, e.g., in closely spaced promoters (Epshtein & Nudler, 2003). The elongation factor NusG acts as a down regulator of both backtracked and non-backtracked pausing (Yakhnin et al., 2023). We investigated the potential role of pause sequences in the dynamics of transcription under the control of tandem promoter formations.

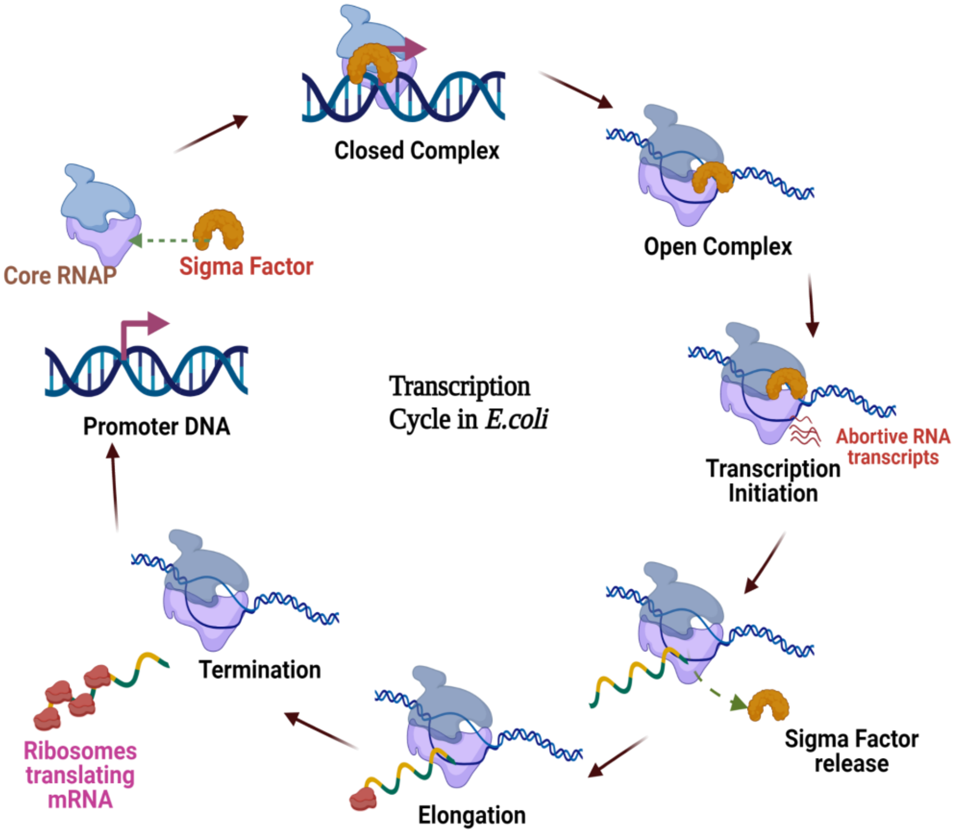


Figure 3. Illustration of the transcription cycle in *E. coli*. It has three main steps: initiation, elongation, and termination. First, a σ factor (yellow) binds to the core RNAP to form an RNAP holoenzyme. The holoenzyme then binds to the promoter and forms a closed complex. Upon binding, the enzyme opens a bubble in the DNA, named open complex, from which transcription is initiated. Abortive transcription events, when occurring, produce short RNAs. Once the enzyme starts sliding along the DNA, the elongation phase begins and the σ factor is released. Finally, RNAP reaches a terminal sequence of DNA where it detaches and produces RNA transcripts. Once ribosomes recognize the RNA transcript, translation begins. Picture adapted from (Stracy & Kapanidis, 2017). Modified using Biorender.

Similar to initiation, transcription termination occurs at specific positions after the end of the gene coding region (Alberts et al., 2002). The features of these positions are not yet fully understood. However, it is known that termination in *E. coli* can be intrinsic or Rho-dependent. In intrinsic terminations, the terminators are complimentary palindromes, forming base pairing between and within the strands of the double helix, and also within the RNA transcript. This base pairing results in a cruciform or a stem-loop structure, which is an essential cause for the termination. It is hypothesized that the RNAPs pause just after the stem-loop structure forms, causing a break in base pairing, which terminates transcription (Martin & Tinoco, 1980; Wilson & von Hippel, 1994).

The Rho-dependent termination occurs in the presence of the Rho protein. When the RNAP pauses at the stem-loop structure, the Rho protein disrupts the base pairing between the DNA template and RNA, terminating transcription (Greive & von Hippel, 2005). The RNAP dissociates from the DNA and RNA is released. Recent studies suggest that the RNAP can stay associated with DNA or unbind, after which it can begin a new round of transcription (Harden et al., 2020; Song et al., 2022). The RNA transcript formed is ready for translation or acts as the end-product of gene expression.

2.2.3 Translation

Ribosomes synthesize proteins by the translation process. George Palade discovered ribosomes in 1955 (Palade, 1955). They consist of ribosomal proteins and rRNA (ribosomal RNA). The rRNA has two subunits in *E. coli*: a small 30S subunit, made of 16S RNA, and a bigger 50S subunit, is made of 5S and 23S RNAs (Ramakrishnan, 2002).

There are three stages in translation: initiation, extension, and termination (Figure 4). *E. coli* initiates translation by binding an mRNA to the small subunit of the ribosome. Translation starts at a specific position of the RNA, the ribosome binding site (RBS). These sequences are called Shine-Dalgarno sequences, which bind with the small subunit of the ribosome (Ramakrishnan, 2002).

As the 30S subunit migrates, it encounters the AUG codon. An initiation complex then forms around this codon, which marks the start of translation initiation. This complex is formed by the pairing between the 30S subunit of mRNA and an aminoacylated tRNA (transfer RNA). Translation initiation in *E. coli* requires three initiation factors (IF). First, IF1 and IF3 mediate the dissociation of ribosomes into smaller subunits. Meanwhile, IF3 is involved in the recognition of the ribosome binding site. Moreover, IF2 is responsible for attaching tRNA and binding the initiation complex to GTP, which provides energy for translation (Alberts et al., 2002).

During translation elongation, the tRNA enters the A-site, which is formed by the hydrolysis of the GTP molecule. This requires elongation factors (EF-Tu and EF-Ts). The ribosome slides along the mRNA and amino acids are linked, one by one, causing the polynucleotide chain to grow. When the stop codon (UAA, UAG, or UGA) enters the A-site, the translation terminates. Upon entry into the A-site, release factors dissociate the polypeptide from the tRNA. In general, Polypeptide chains then fold into tertiary structures, at which point they become functional (Alberts et al., 2002).

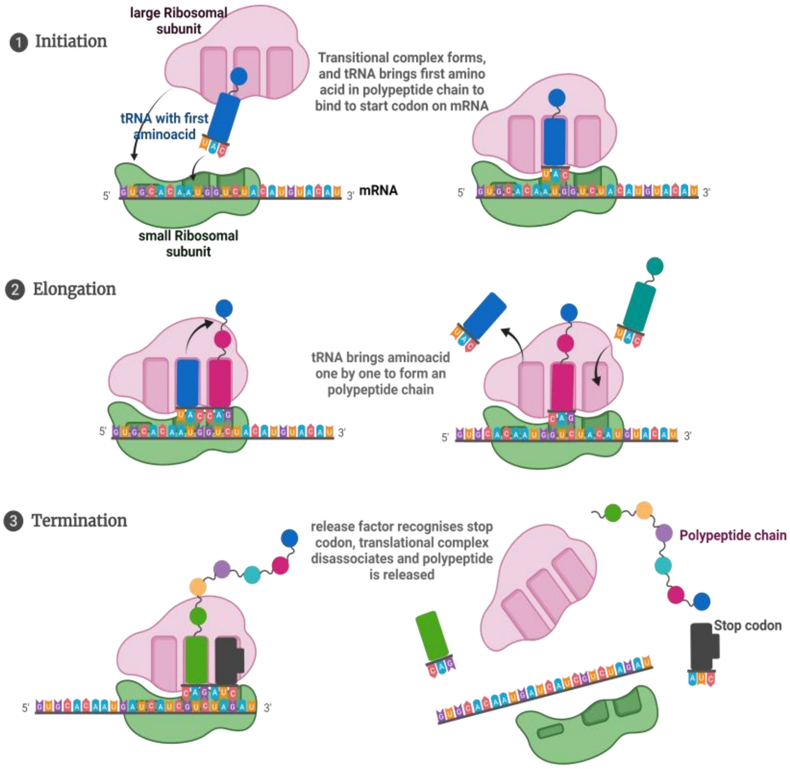

1 Initiation
large Ribosomal subunit
tRNA with first aminoacid
mRNA
5'
3'
small Ribosomal subunit
Transitional complex forms, and tRNA brings first amino acid in polypeptide chain to bind to start codon on mRNA
2 Elongation
tRNA brings aminoacid one by one to form an polypeptide chain
3 Termination
release factor recognises stop codon, translational complex disassociates and polypeptide is released
Polypeptide chain
Stop codon

Figure 4. A cartoon illustrating translation in *E. coli*. Translation initiates with the formation of an initiation complex, which includes an mRNA sequence, a small ribosomal subunit, and an aminoacylated tRNA. During elongation, the ribosome slides along the mRNA, and amino acids are linked one by one, causing the polynucleotide chain to grow. Termination occurs when the release factor finds the stop codon and dissociates the polypeptide from the tRNA. The figure is generated using Biorender.

Since transcription and translation are mechanically linked in bacteria (Alberts et al., 2002), they are also expected to be dynamically correlated, which has been confirmed. Because of the mechanical link, we used measurements of protein levels (translation level) from a strain library with fluorescently tagged proteins to track transcription (Taniguchi et al., 2010; Zaslaver et al., 2006). These measurements confirmed the results from RNA-seq (transcription level) in our publications.

2.3 Regulation of Gene Expression

An organism appropriate responsiveness to environmental changes is heavily influenced by gene regulation (Browning & Busby, 2004, 2016). Jacob and Monod laid the foundation of the basic understanding of gene expression in 1961, with the (now classic) example of the lactose (lac) operon. The lac operon has three genes for lactose metabolism (lacZ, lacY, and lacA), and LacI is the regulator gene. Aside from these, lac operon has three operator sites (O_1, O_2 and O_3).

In *E. coli*, transcription initiation is the main step most influenced by gene expression regulation (Browning & Busby, 2004, 2016; Chamberlin, 1974; McClure, 1985). This bacteria has evolved multiple regulators of gene expression such as σ factors, RNAP, TFs, and Gyrase (Browning & Busby, 2004, 2016; Sneppen et al., 2005). Also, influential factors, but not subject to regulation, are the promoter sequence and location in the DNA.

2.3.1 Transcription Factors

Transcription factors (TFs) are regulatory proteins whose presence can influence transcription rates (Browning & Busby, 2004). Most TFs are either activators or repressors (Figure 5) (which can differ between promoters). There are more than 300 transcription factors in *E. coli*. Some control a large number of genes, while others control one to a few genes (Hochschild & Dove, 1998; Madan Babu & Teichmann, 2003; Martínez-Antonio & Collado-Vides, 2003).

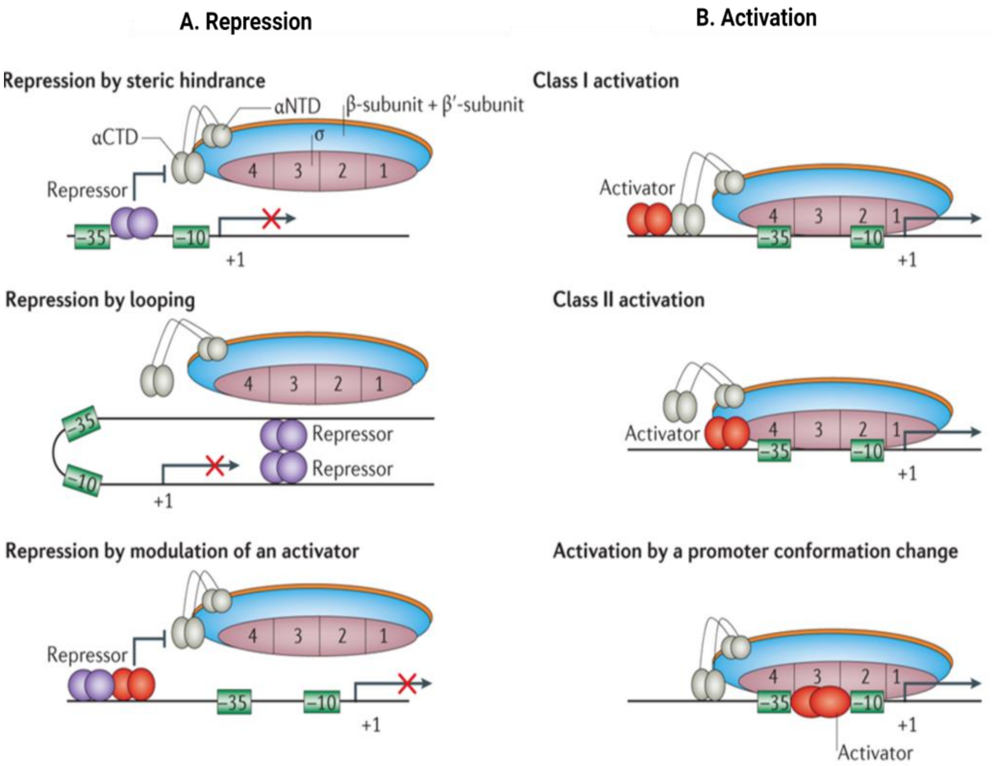


Figure 5. TF regulation of gene expression. **(A)** Repression. In the case of repression due to steric hindrance, RNA polymerase is blocked from binding to the promoter by the binding of a TF to a site that overlaps the promoter. In the case of repression by looping, protein–protein interactions form between repressors that bind to sites upstream and downstream of the promoter, which prevents the recognition of promoter elements by RNA polymerase. Repressors can also modulate activators to prevent recruitment of RNA polymerase. **(B)** Activation. In class I activation, the activator binds to a site upstream of the promoter and recruits RNA polymerase to the promoter. In class II activation, the activator binds to a site in the promoter adjacent to (or overlapping with) the −35 element, where it recruits RNA polymerase through direct interactions with the sigma factor. Some activators induce a conformational change in the promoter DNA to activate transcription. These activators bind at, or near to, the core RNA polymerase recognition elements of the promoter and often realign the −10 and −35, hence enabling the recruitment of RNA polymerase to the promoter and activation of transcription. Picture taken from (Browning & Busby, 2016) with permission.

There are various mechanisms of transcription repression. The most common mechanism for inhibiting transcription is 'steric hindrance', by the binding of repressors to operator sites, which limits the RNAP from accessing the -10 or -35 regions of the DNA (Garcia & Phillips, 2011; Oehler, Eismann, Krämer, & Müller-Hill, 1990). Another mechanism is the occupation of operator sites by TFs, if close to the RNAP binding region, causing the DNA to bend, which forms a loop and causes repression (Müller, Oehler, & Müller-Hill, 1996; Swint-Kruse & Matthews, 2009).

There are three common mechanisms by which TF-mediated activation occurs (Browning & Busby, 2004, 2016). In the first, a TF binds to an operator site and helps RNAP recruiting by interacting with the subunits of the RNAP. By binding the TF to its operator site, it activates an operator site upstream of the -35 element (Ebright, 1993). In another form of activation, a TF attached to the DNA, influences the affinity of RNAP holoenzyme (Dove, Darst, & Hochschild, 2003). In the last type of TF-mediated activation, a TF binds with the DNA conformationally changes the promoter's spacing between the elements of consensus sequences. Through this conformational change, the promoter DNA is positioned optimally to enhance RNAP binding to -35 and -10 elements.

Additionally, TFs themselves are subject to regulatory mechanisms. For example, the TFs' affinity to bind with DNA can be altered by binding of specific molecules whose concentration can vary with the environment (Browning & Busby, 2004). In laboratory conditions, the use of such inducers has been one of the most common methods to control TF binding to DNA (Garcia & Phillips, 2011).

We use inducers Isopropyl ß-D-1-thiogalactopyranoside (IPTG), L-arabinose, and anhydrotetracycline (aTc) as chemical inducers for the regulation of the activity of the promoters of $P_{Lac/ara-1}$, $P_{LtetO-1}$, P_{Lac}, P_{tetA}, and P_{BAD}. We also investigated the effects of perturbations on TF-gene transcription regulation at the genome-wide level. In addition, we studied the relationship between genes controlled by tandem promoters and their input and output TFs.

The transcription factor network (TFN) of *E. coli* is widely mapped (Santos-Zavaleta et al., 2019). It has information on approximately 4700 transcription factors (TFs) interactions in approximately 4500 genes. It also informs on their activators and repressors. We used this information to quantify the response of genes to the shifts in RNAP concentrations.

2.3.2 σ Factors regulation

σ factors play specific roles in different stages of transcription initiation (Figure 6). These roles include the recognition of promoter elements to create a closed complex (CC), the stabilization of the open complex (OC), and interactions with transcription activators (Hengge-Aronis, 2002a; Saecker et al., 2011).

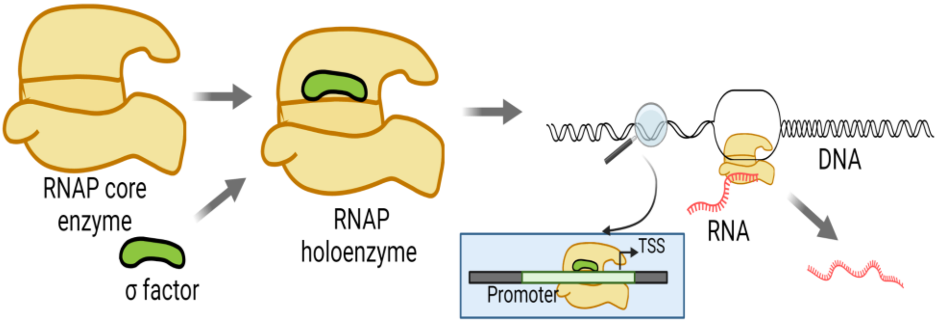


Figure 6. Illustration of σ factors driven transcription. When an RNAP core enzyme binds with a σ factor, it forms the RNAP holoenzyme. This holoenzyme can recognize the promoter region of a gene with the help of the σ factor. After promoter recognition, transcription of that gene occurs (assembling an RNA). *E. coli* uses different types of σ factors to recruit different sets of genes to adapt to different perturbations. Created using Biorender.

There are seven σ factors in *E. coli*, named in accordance with their molecular weights (kDa) (Table 1). These σ factors are σ^{70}, σ^{38}, σ^{54}, σ^{24}, σ^{32}, σ^{19}, and σ^{28}. The first, σ^{70} (also known as σ^{D}), is an important housekeeping factor that controls many promoters in *E. coli* (Tripathi, Zhang, & Lin, 2014).

Table 1. σ factors in *E. coli*, the genes producing them, and their function. For a review, see (Ishihama, 2000).

Type of σ factor	Gene	Function
σ^{70}	rpoD	House keeping
σ^{38}	rpoS	Respond to stress
σ^{54}	rpoN	Respond to nitrogen stress
σ^{32}	rpoH	Respond to heat shock
σ^{19}	fecI	Uptake of ferric citrate
σ^{24}	rpoE	Extracytoplasmic function
σ^{28}	fliA	Flagellar synthesis

Changes in σ factor numbers are a main cause for changes in the rate of transcription of genes with different preference for σ factors. As such, most σ factors are only expressed in unfavorable conditions. For example, σ^{32} is expressed in response to heat shock (Hengge-Aronis, 2002a, 2002b). σ^{38} (also referred to as RpoS) is expressed in response to nutrient deprivation, which may cause cells to enter the stationary growth phase (Battesti, Majdalani, & Gottesman, 2011; Ishihama, 2000). σ^{54} expresses in response to nitrogen limitation. σ^{24} (also called RpoE) is an extreme heat-responsive factor. σ^{32} (also called RpoH) expresses when the bacteria are exposed to heat. Finally, σ^{19} (also called FecI) transports and metabolizes iron. σ^{28} (also called RpoF/FliA) is responsible for flagellar synthesis and chemotaxis. Consequently, genes responsible for responding to these stresses only become active in these specific conditions.

Evidence suggests that the time length of the closed complex formation relative to the open complex formation of a promoter affects the promoter's responsiveness to changes in σ factor numbers (Kandavalli, Tran, & Ribeiro, 2016). Meanwhile, a recent study investigated the dynamics of genes with preference for both σ^{38} and σ^{70}, found that they are upregulated in stationary growth, and proposed a sequence dependent model of that process (Baptista et al., 2022). Also, both studies suggest that the strength of the regulation exerted by σ factors differs between genes. Another recent study (Van Brempt et al., 2020) used promoter libraries specific to *E. coli* σ70 and *B. subtilis* σ^{B}, σ^{F}, and σ^{W} to develop prediction models. These models predicted the transcription initiation frequency (TIF) of promoters associated with σ factors promoters. Another recent work (Balakrishnan et al., 2022) showed that the components that regulate σ factors such as Rsd controls the RNAP accessibility for promoters.

We measured RpoS levels using an MGmCherry strain coding for RpoS tagged with mCherry (Patange et al., 2018). We also measured the RNA fold changes of the genes expressing each of the seven σ factors following the shifts in media dilution. However, their influence was not found to be significant in that context.

2.3.3 Other Factors

In addition to TFs, σ factors, and closely spaced promoters, several other intracellular factors influence gene expression in *E. coli* (for a review see (Bervoets & Charlier, 2019)). Small molecules, such as guanosine 3', 5'-diphosphate (ppGpp), inhibit the expression of genes responsible for stress responses by interacting with the RNAP (Wilma Ross, Vrentas, Sanchez-Vazquez, Gaal, & Gourse, 2013). Studies show that ppGpp inhibits transcription initiation (Artsimovitch & Henkin, 2009; Wilma Ross et al., 2013). Moreover, ppGpp enhances the expression of genes that are responsible for amino acid synthesis (Wilma Ross et al., 2013).

Meanwhile, some global regulators (GRs) are responsible for DNA compaction, such as H-NS, Fis, and HU proteins (Dillon & Dorman, 2010). This is known to influence gene expression. H-NS, for instance, binds to the AT-rich region of DNA and acts as a repressor global, eventually leading to the inhibition of global transcription activity(Browning & Busby, 2016; Navarre et al., 2006).

DNA supercoiling is another factor that influences gene expression and, thus, is under regulation. Under standard conditions, the DNA of *E. coli* is negatively supercoiled (Vinograd, Lebowitz, Radloff, Watson, & Laipis, 1965). Studies have shown that increases in supercoiling levels can regulate gene expression, by slowing down or dissociating the RNAP (Chong, Chen, Ge, & Xie, 2014).

Environmental factors, such as temperature, also highly influence gene expression, in some cases in complex ways, in both bacteria and eukaryotic cells (Charlebois, Hauser, Marshall, & Balázsi, 2018; Richter, Haslbeck, & Buchner, 2010). It is known that bacteria trigger an evolved genome-wide cold shock response (Phadtare & Inouye, 2004), while a rise in temperature also initiates a heat-shock response program (Craig & Schlesinger, 1985; Neidhardt, VanBogelen, & Lau, 1983).

2.3.4 The Stochastic Nature of Gene Expression

Gene expression in *E. coli* is a multi-step process (Jones, Brewster, & Phillips, 2014; Rolf Lutz, Lozinski, Ellinger, & Bujard, 2001; Saecker et al., 2011), with each step involving complex biochemical reactions. In a few steps, there is only a small number of reactive molecules that, by being regulated, allow for regulating transcription rates. This also contributes to making gene expression a noisy process (Kaern et al., 2005). As a result of this noise, clonal cell populations display phenotypic diversity (Bury-Moné & Sclavi, 2017; Eldar & Elowitz, 2010; Elowitz et al., 2002). The noise in gene expression can be both “intrinsic” as well as “extrinsic” (Elowitz et al., 2002).

Intrinsic noise arises from the stochastic nature of biochemical reactions along with the small number of some of the molecules involved in those reactions composing transcription. Meanwhile, extrinsic noise is caused by differences in the number of cellular components between cells (including proteins) (Elowitz et al., 2002). Furthermore, asymmetries in RNA and proteins partitioning during cell division also contribute to extrinsic noise in a cell population (for a review see (Baptista & Ribeiro, 2020)).

2.3.5 Genome-wide Stresses

In their natural environment, gut bacteria such as *E. coli* are subject to fluctuations in pH, temperature, nutrients availability, and osmolarity, among others. These fluctuations cause stresses that perturb hundreds of genes. E.g., approximately 60-90% of genes in *E. coli* are responsive to changes in the conditions of growth (Sanchez-Vazquez, Dewey, Kitten, Ross, & Gourse, 2019). These responses can be in the short-, mid- and/or long-term (Mitosch, Rieckh, & Bollenbach, 2019).

Transcription factors (TFs) play crucial roles in influencing those stress responses (Brooks et al., 2014; Côté et al., 2016; Fang et al., 2017; Urchueguía et al., 2021) and may partially explain such large percentages of responsive genes. In addition, the TF network (TFN) topological features may further influence the coordination of the stress responses. This poses a challenge in understanding bacteria genome-wide stress responses. We studied the influence of the topology of the TFN of E. coli on the genome-wide responses to shifting RNAP concentrations.

2.3.6 Closely Spaced Promoters

E. coli has hundreds of genes controlled by more than one promoter (Santos-Zavaleta et al., 2019). Usually, two such promoters separated by less than 1kb distance between their transcription start sites (TSSs) are defined as being closely spaced in the DNA (Trinklein et al., 2004). They exist in convergent, divergent, or tandem configurations (Figure 7) (for a review, see (McClure, 1985)).

In the divergent configuration, the two promoters are arranged in opposite orientations (Figure 7A). In the convergent configuration, the promoters are arranged in face-to-face orientations (Figure 7B). Because of this, in both cases, the two promoters control the transcription of different genes. Meanwhile, in tandem configurations, the promoters are placed in the same orientation, one being upstream of the other. This causes them to control the transcription of the same gene(s) (Figure 7C).

Closely spaced promoters are expected to be subject to transcriptional interference (TI) (Eszterhas, Bouhassira, Martin, & Fiering, 2002), which is a reduction of the transcriptional activity of one promoter, due to the transcriptional activity of the other promoter (Shearwin, Callen, & Egan, 2005). This phenomenon can be particularly strong in convergent promoter configurations (Eszterhas et al., 2002).

TI can involve promoter occlusion, where transcription initiation events at one promoter physically obstruct RNAPs attempting to bind to the TSS of the other promoter (Figure 8). In detail, an RNAP occupies approximately 35 bp, when in OC formation (Greive & von Hippel, 2005). Therefore, if the distance between the TSSs of two promoters (d_{TSS}) is less than 35 bp, an RNAP occupying one of the promoters will occlude the other promoter (Sneppen et al., 2005).

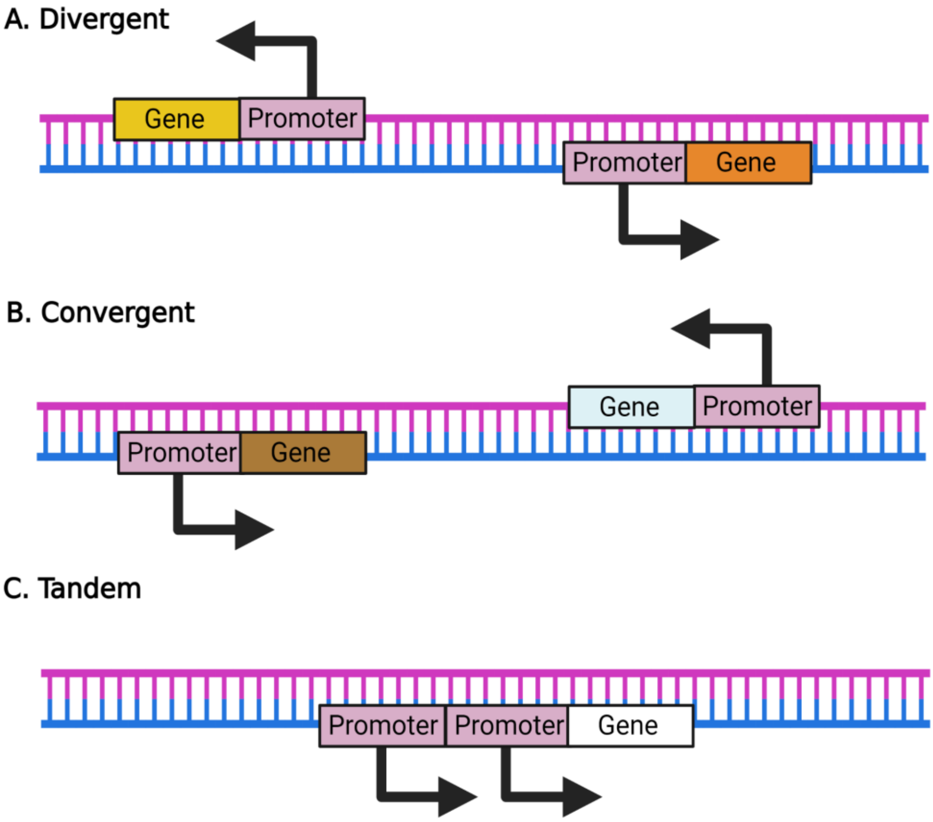


Figure 7. Configurations of closely spaced promoters. **(A)** Divergent promoters. **(B)** Convergent promoters. **(C)** Tandem promoters.

Meanwhile, if the distance between the two promoters is greater than 35 bp, the RNAPs elongating from one promoter can bump into RNAPs occupying the other promoter. Such collisions can cause the RNAPs to fall off (Figure 10) (Callen, Shearwin, & Egan, 2004; Hoffmann, Hao, Shearwin, & Arndt, 2019; Ponnambalam & Busby, 1987; Sneppen et al., 2005). Moreover, variables such as the RNAP binding affinity to a downstream promoter, are expected to influence the outcome of those events. However, models using realistic parameter values suggest that such collisions are rare (Häkkinen, Oliveira, Neeli-Venkata, & Ribeiro, 2019; Martins et al., 2012; Sneppen et al., 2005).

Another factor that influences closely spaced genes are the time-lengths of rate-limiting steps in transcription (Häkkinen et al., 2019). To test this, *in vivo* single-cell, single-RNA measurements on gene pairs that share promoter elements have been conducted using genes controlled by promoters with a head-to-head configuration. The results support the hypothesis. Based on this, the authors suggested that closely located genes could evolve compatible configurations in their rate-limiting steps to achieve specific degrees of cooperation.

Another work by (Yeung et al., 2017) studied transcriptional interference with a focus on supercoiling in convergent, divergent, and tandem genes. They used modeling to simulate the effects of supercoiling on the transcriptional rates of these genes. They also conducted experiments using synthetic constructs to validate their findings. They postulate that supercoiling is responsible for the observed differences in the expression levels of different configurations of closely spaced genes. More recently (Johnstone & Galloway, 2022) characterized the influence of supercoiling in two-gene systems in Zebrafish. The two-gene system consisted of a reporter gene and an inducible gene in convergent, divergent, or tandem orientation.

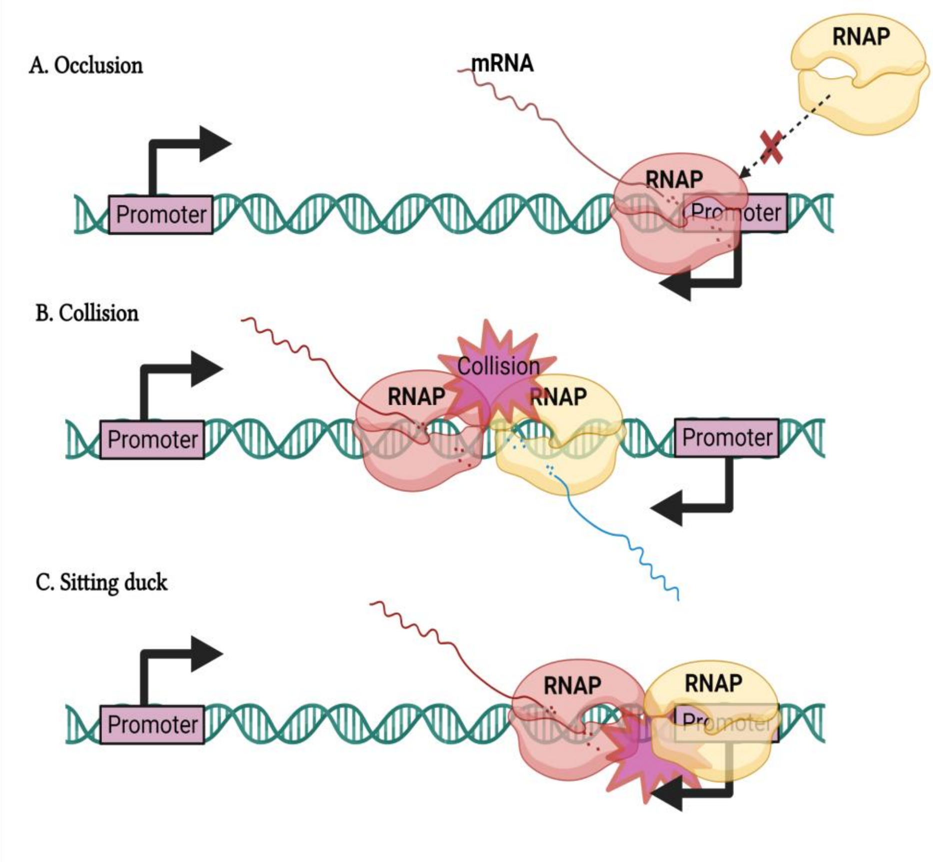


Figure 8. Illustration of the phenomenon of transcription interference. **(A)** Occlusion: An RNAP blocks another RNAP's attempt to bind to the promoter. **(B)** Collision: Two elongating RNAPs undergo collision, potentially causing one or more fall-offs. **(C)** Sitting duck: An elongating RNAP is stopped by an RNAP bound to a promoter. Creating using Biorender.

We investigated the dynamics of native genes controlled by promoters in tandem configuration. We also designed synthetic constructs in tandem configuration. We then used them to study the effects of transcription-targeting antibiotics on the dynamics of tandem promoters.

2.4 Models of Bacterial Gene Expression

Numerous models of prokaryotic gene expression have been proposed over the years, with the aim of predicting, mimicking, or better visualizing the dynamics of the genetic circuits and the cellular functions that they control (for reviews see (de

Jong, 2002; Ribeiro, 2010a)). These models, once validated by empirical data, can also serve as a framework for testing new hypotheses. The development of live cell imaging, cloning, and genetic engineering has led to a better understanding of gene expression, which in turn allowed the development of models. Since transcription initiation is the major regulation checkpoint of bacterial gene expression (Browning & Busby, 2004; Djordjevic & Bundschuh, 2008), most gene expression models include it explicitly.

Since bacterial transcription and translation are mechanically coupled, some models of prokaryotic gene expression assume that RNA and proteins are produced in a single step. The degradation of RNA and proteins are also typically modeled as one-step process, not subject to complex regulation (Munsky & Khammash, 2006; Peccoud & Ycart, 1995). Stochastic models are usually implemented using chemical reaction systems. A basic chemical reaction model is shown in 2.4.1:

$$\mathrm{A} + \mathrm{B} \xrightarrow{k} \mathrm{C} \tag{2.4.1}$$

This reaction involves two reactants, A and B, which react to form a product, C. The rate constant, k, determines the reactivity between A and B. Assuming this modeling approach, transcription can be modeled as follows (Ribeiro, 2010a):

$$RNAP + Pro \xrightarrow{k} Pro + RNAP + RNA \tag{2.4.2}$$

In this reaction, RNAP represents the RNA polymerase holoenzyme, Pro is a free promoter, and *k* is the rate constant of the reaction. Since RNAP and the promoter are not consumed in the reaction, they are also products of the reaction. In this model, no reversible steps or rate-limiting steps are included.

Transcription initiation in *E. coli*, however, is a more complex process, since it involves a few rate-limiting steps that significantly influence mean kinetics and noise (Browning & Busby, 2016; Muthukrishnan et al., 2012). The following model of the set of reaction events accounts for the most rate-limiting steps (Walter, Zillig, Palm, & Fuchs, 1967):

$$RNAP + Pro \xleftrightarrow{k_x} RP_{cc} \xrightarrow{k_i} RP_{oc} \xrightarrow{\infty} RNA \tag{2.4.3}$$

In this reaction, RNAP binds with the promoter, at a rate (*k_X*), and forms a closed complex (RP_{cc}). The closed complex is then isomerized and forms an open complex (RP_{oc}) at a rate (*k_i)*. RNAP then reaches termination and releases the RNA. Since elongation usually takes much less time than the rate-limiting steps in initiation, it is usually modeled as an (infinitely) fast step, for simplicity (Walter et al., 1967). Meanwhile, the first step is set to be reversible, to account for the chemical instability of the closed complex (for a review see (Ribeiro, 2010b)).

Using these reactions, one can model the dynamics of two closely spaced genes (gene 1 and gene 2) sharing promoter elements in the head-to-head configuration as follows (Häkkinen et al., 2019):

$$P_0 \xrightarrow{k_1} I_1 \xrightarrow{k_3} P_0 + X \qquad 2.4.4$$

$$P_0 \xrightarrow{k_2} I_2 \xrightarrow{k_4} P_0 + Y \qquad 2.4.5$$

Here, P_0 refers to a free promoter, while I_1 and I_2 are the intermediate transcription complexes from gene 1 and gene 2, respectively. Also, X and Y refer to the mRNA from genes 1 and 2, respectively, k_1 and k_2 are the rates of closed complex formation respectively, and k_3 and k_4 represent the rates of open and elongation complex formations, respectively.

We proposed new transcription models for genes regulated by tandem promoters. For that, we modeled independent promoters, that then can interfere with each other depending on the distance and the rates of occupation of each TSS.

2.5 Gene Expression Measurement Techniques

2.5.1 Fluorescent Labeling, MS2-GFP Tagging System, and DNA Integration

Green fluorescent protein (GFP) was isolated from jellyfish, Aequorea (Morise, Shimomura, Johnson, & Winant, 1974; Shimomura, Johnson, & Saiga, 1962; Ward, Cody, Hart, & Cormier, 1980). GFP, when expressed in *Escherichia coli*, produces a stable fluorescence and has since been established as a tool for monitoring gene expression in live cells (Chalfie, Tu, Euskirchen, Ward, & Prasher, 1994). Identifying GFP from Aequorea has been described as a revolutionary step in cell biology. In subsequent years, many new FPs were engineered, for quantifying gene expression in live cells (Day & Davidson, 2009). In comparison to other fluorescent markers, fluorescent proteins have several advantages (Hayashi-Takanaka, Stasevich, Kurumizaka, Nozaki, & Kimura, 2014; Schneider & Hackenberger, 2017).

Modified fluorescent proteins tagged to viral proteins can be fused with RNA sequences, to observe RNAs in live cells (Golding et al., 2005). Also, because of the unique structure of fluorescent proteins, mutations have a considerable chance to alter their fluorescent properties (Patterson, Knobel, Sharif, Kain, & Piston, 1997). As a result, presently there is a wide range of different emission colors of fluorescent proteins (Kremers, Gilbert, Cranfill, Davidson, & Piston, 2011).

Fluorescent proteins also have limitations. These include fluctuations in the fluorescent intensity following some changes in the environment (Shaner et al., 2004). For example, most fluorescent proteins are temperature-sensitive, while yellow fluorescent proteins (YFPs) are pH- and chloride-sensitive. This hampers their use in quantifying changes in protein levels between conditions. Fluorescent proteins can also be toxic (Shaner et al., 2004) .

Several versions of fluorescent proteins (Shaner et al., 2004) have been developed to improve their original properties. For example, there are new GFP proteins that have much-improved fluorescence intensity at 37°C than the wild-type proteins (Cormack, Valdivia, & Falkow, 1996). In addition, many fluorescent proteins can be found as dimers or trimers. In many cases, this oligomerization causes toxic effects (Shaner et al., 2004). A protein can also become indirectly toxic if a certain wavelength is needed to excite it. Microorganisms, for example, are likely to be adversely affected by exposure to UV light (Jagger, 1976; Kramer & Ames, 1987).

In general, when genetically engineering a fluorescent protein for a bacterium, it is recommended to perform toxicity tests (Turkowyd et al., 2017). Furthermore, the fluorescence signal needs to be higher than the autofluorescence that originates from the cellular background. There should also be minimal crosstalk between the emission spectrum and the excitation spectrum. Finally, when the fluorescent protein is fused with the target protein, it should have the least effect possible on the native protein's functionality (Shaner et al., 2004). We made use of several fluorescent proteins in this thesis.

A special case of fluorescent labeling is the MS2-GFP tagging system which we utilized in our studies. The MS2-GFP system allows for quantifying RNA molecules in single cells (Golding et al., 2005) by tagging multiple GFPs to a single RNA. Robert Singer first developed this method for eukaryotic cells, but it was later modified for use in bacteria (Bertrand et al., 1998; Golding et al., 2005).

The MS2-GFP system that we used consists of two components that need to be expressed simultaneously in the same cell (Golding 2005): (i) a fluorescent protein fused with an MS2 coat protein, which enables binding specifically, and (ii) RNA that carries repetitive sequences to which MS2 proteins can bind to. Figure 9 illustrates these components. The original MS2 coat proteins are from bacteriophages, which use them to protect their RNA (Bernardi & Spahr, 1972). The binding pathway to a specific RNA sequence inhibits RNA replication and causes phage packaging (Peabody, 1993; Querido & Chartrand, 2008). Molecular scientists have developed fluorescent fusion proteins based on MS2 coat proteins that bind to RNA with stem-loop sequences, as the original (Fusco et al., 2003; Golding et al., 2005; Lenstra, Rodriguez, Chen, & Larson, 2016).

The MS2-GFP system allows investigation individual transcription events in live cells (Golding et al., 2005; Mäkelä et al., 2013). Before producing target RNA, a reporter plasmid coding for a fluorescent protein coupled to the MS2 coat protein must be highly expressed. The high concentration of MS2-GFP ensures its binding to target RNAs with multiple binding sites for MS2, as soon as they are produced in a cell. By binding multiple MS2-GFP proteins to one target RNA, the RNA becomes brighter than its surrounding with only freely diffusing, unbound MS2-GFPs. (Mäkelä et al., 2013; Muthukrishnan et al., 2012). In a confocal microscope, target RNAs bound by MS2-GFP appear as bright spots (Golding et al., 2005; Mäkelä et al., 2013). Lastly, the degradation of the target RNA is very delayed as a result of the coating by MS2-GFP proteins (Fusco et al., 2003; Tran, Oliveira, Goncalves, & Ribeiro, 2015).

However, quantifying RNA by MS2d-GFP from microscopy data is difficult (Häkkinen, Muthukrishnan, Mora, Fonseca, & Ribeiro, 2013; Häkkinen & Ribeiro, 2014, 2016). One common problem, easily visible in time-lapse microscopy, is the intermittent disappearance of MS2-GFP tagged RNAs. Also, the precision of estimating the number of tagged RNAs in an 'RNA spot' decreases with the number of RNAs in that spot (Golding et al., 2005). Other viral proteins have also been used for the same purpose, including PP7 proteins derived from the PP7 bacteriophage, which operate similarly to the MS2-GFP system (Larson, Zenklusen, Wu, Chao, & Singer, 2011; Lenstra et al., 2016).

The above systems are carried by plasmids, which, in turn, require DNA integration. The techniques to integrate DNA sequences and to create combinations of different genetic sequences have been developed and improved for almost four decades now, following the discovery of DNA ligase and restriction enzymes (H. O. Smith & Wilcox, 1970)(for a review, see (Kiermer, 2007)). In 1973, the construction of the first bacterial plasmid vector was achieved using the restriction enzyme EcoRI, which generated fragments from two plasmids carrying sequences that provide cells with two different antibiotic resistances. These fragments were joined using DNA ligase. The transformants resulted in bacterial cells resistant to both antibiotics by carrying a single plasmid (for a review see (Kiermer, 2007))

The further development of these techniques made possible the construction of the first artificial genome (*Mycoplasma genitalium*) in which many DNA fragments have been synthesized independently, then assembled, and then transferred into a host (Gibson et al., 2009). Nowadays it is possible to construct synthetic genomes using advanced molecular biology software such as Snapgene and engineering techniques such as *de novo* synthesis, recombinant DNA technology, Gibson Assembly, and, more recently, CRISPR-Cas9 (Deltcheva et al., 2011; Gibson et al., 2009; Jinek et al., 2012).

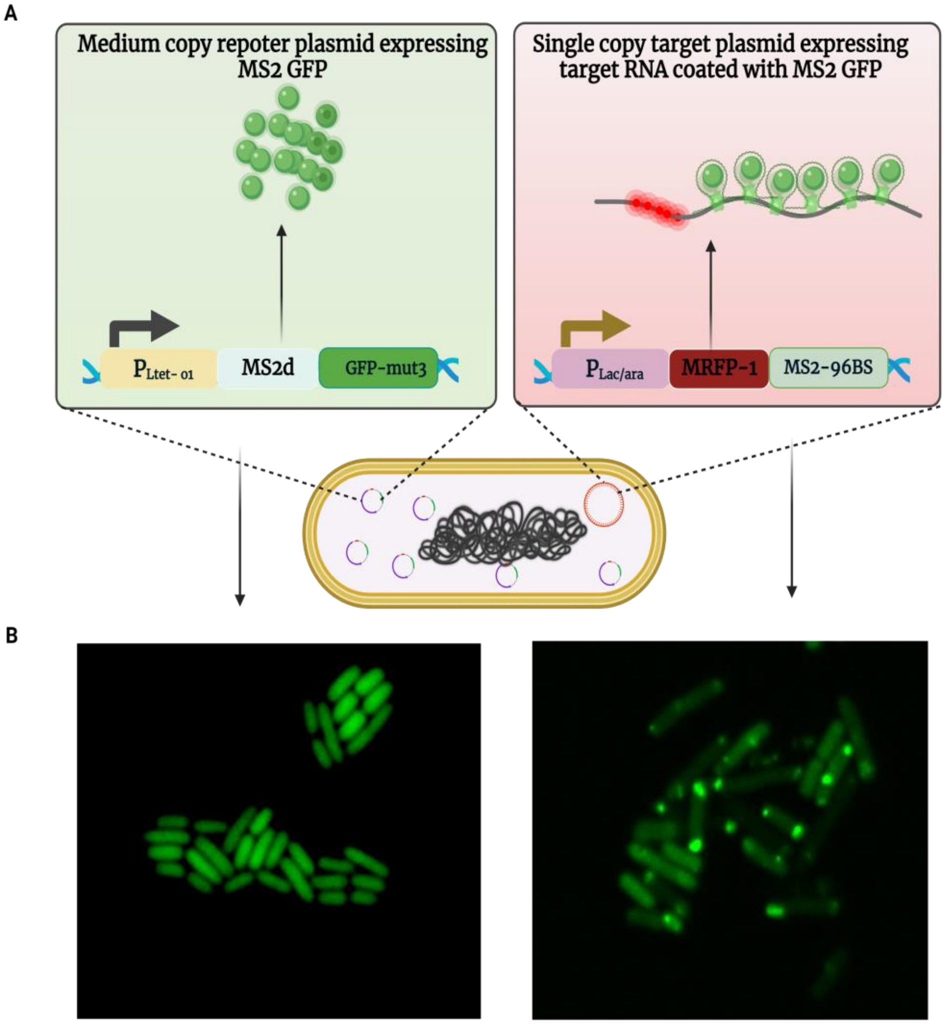

A
Medium copy repoter plasmid expressing MS2 GFP
$P_{Ltet-01}$
MS2d
GFP-mut3
Single copy target plasmid expressing target RNA coated with MS2 GFP
$P_{Lac/ara}$
MRFP-1
MS2-96BS
B

Figure 9. Figure illustrating the MS2-GFP system. **(A)** Cells produce multiple MS2-GFP reporter proteins, under the control of $P_{Ltet-O1}$, while the production of RNAs target for MS2-GFP is under the control of $P_{Lac/ara-1}$. MS2-GFP proteins accumulate in the cytoplasm and bind to the target RNA upon its production. Meanwhile, the mRFP region of the target RNA is translated into proteins that glow red after translation. **(B)** Example images of *E. coli* cells when expressing only MS2-GFP molecules (left) and cells expressing RNA molecules appearing as fluorescent spots inside the cells (right). The final figure is created using Biorender.

We we used molecular cloning techniques for the transformation of plasmids into host strain *E. coli*. The transformation was based on the selection of recombinants depending on the antibiotic-resistance gene of the plasmid DNA. We also designed and assembled the synthetic constructs encoding mCherry in tandem formation using Snapgene software. We also made use of the chemically competent cells for the transformation of plasmids into host strain.

2.5.2 Microscopy Imaging

Monitoring gene expression using fluorescent proteins is frequently carried out using fluorescence microscopy (for a review see (Stephens & Allan, 2003)). By illuminating the sample with the wavelength of the excitation and then capturing the light that is emitted from the fluorescently tagged molecules, a fluorescence microscope can detect the light emitted by fluorescently tagged molecules. Moreover, confocal microscopy eliminates the need to expose the entire sample to a light source, partially solving the background noise problem.

Confocal microscopy uses point illumination and point detection to examine a small sample area at a time (Pawley, 2022). As a result, background fluorescence is eliminated, improving optical resolution. In detail, a pinhole is used to ensure that no light other than the illuminated section of the sample reaches the detector. A photomultiplier tube (PMT) then collects the fluorescence and develops the complete image, which is then processed by a computer into a two-dimensional image (Elliott, 2020). Nevertheless, this approach has its limitations, such as that it can only capture a small section of the sample at a time. This time interval can be reduced by a spinning disk (Castellano-Muñoz, Peng, Salles, & Ricci, 2012; Nakano, 2002). We made use of confocal microscopy of fluorescent proteins and of RNA tagged with MS2-GFP.

Meanwhile, to study the morphological characteristics of cells, phase-contrast microscopy is commonly used (Murphy, Oldfield, Schwartz, & Davidson, n.d.). This technique was developed by Fritz Zernike (Zernike, 1942). He was awarded a Nobel Prize for this development in 1953. Phase contrast allows detecting the phase shift when light is scattered, eventually converting it into the contrast of the image. The principle of phase contrast microscopy involves the introduction of a phase shift in the light passing through different parts of the sample. As the transmitted light passes through the sample, it encounters regions with varying refractive indices due to the differences in sample thickness and composition. These variations in the refractive index cause changes in the phase of the transmitted light. The phase contrast technique greatly enhances the visibility of cellular structures, which are often difficult to observe with traditional microscopy. We used phase contrast microscopy to study the morphological characteristics of the strains in varying conditions of growth, temperatures, and inductions.

2.5.3 Single-cell Data Acquisition Using Flow-cytometry

Flow cytometers can measure the overall protein fluorescence levels inside individual cells orders of magnitude faster than microscopes. As cells pass through the laser beam, detectors collect several optical signals (e.g., forward scatter and side scatter) (Figure 10). The intensity of these signals provides information on cell size and the internal structures of the cell. Meanwhile, FITC (fluorescein isothiocyanate) is used to quantify the green fluorescence intensity, for example.

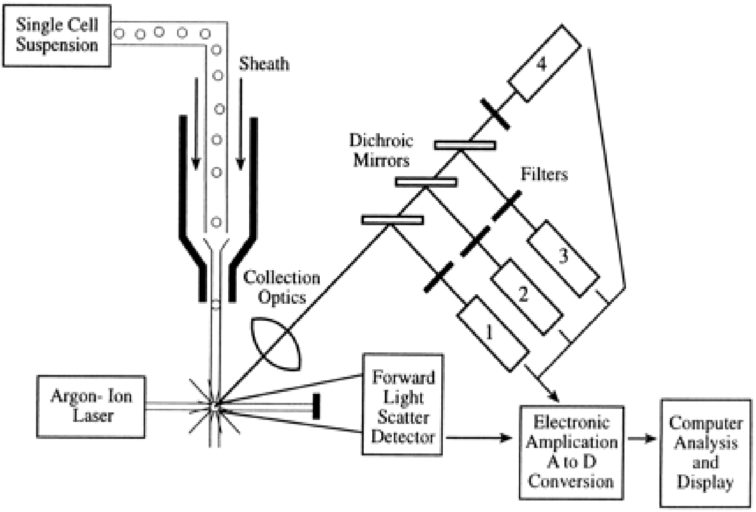

Single Cell Suspension
Sheath
Dichroic Mirrors
Filters
4
3
2
1
Collection Optics
Argon- Ion Laser
Forward Light Scatter Detector
Electronic Amplication A to D Conversion
Computer Analysis and Display

Figure 10. Schematic diagram of a flow cytometer. A single-cell suspension is focused on the light source (laser). A detector collects and amplifies the forward light scatter and converts it into digital signals, which can then be used to perform further analyses. This picture is taken from (Brown & Wittwer, 2000) with permission.

Using flow cytometry, thousands of cells can be analyzed per minute. In addition, the single-cell data is in a format that can be processed quickly while, in comparison, microscopy data requires image analysis. However, similar to microscopy, when studying protein levels by flow cytometry, cellular autofluorescence should be considered (Galbusera, Bellement-Theroue, Urchueguia, Julou, & van Nimwegen, 2020). An ACEA Novocyte flow-cytometer equipped with a blue and a yellow laser was used for single-cell protein quantification.

2.5.4 Transcriptome Quantification using RNA-sequencing

Understanding the functioning of genomes requires measuring the transcriptome (defined as 'all transcripts in a cell at a certain point in time'). Transcriptomes, therefore, include mRNAs, siRNAs, and non-coding RNAs (Wang, Gerstein, & Snyder, 2009). In the last decade, efforts have been made to develop and improve transcriptome quantification technologies (Emrich, Barbazuk, Li, & Schnable, 2007; Lister et al., 2008).

Transcriptome research has improved dramatically with the advent of high-throughput sequencing (Emrich et al., 2007; Lister et al., 2008), following the demonstration that next-generation sequencing technologies can be used to sequence complementary DNA (cDNA) through RNA sequencing. There are many advantages in next-generation sequencing, such as the ability to resolve single base pairs with very low background signal, the ability to capture a wide range of expression dynamics, and the high reproducibility of results (Cloonan et al., 2008; Nagalakshmi et al., 2008; Wang et al., 2009). The methodology of RNA sequencing used is illustrated in Figure 11.

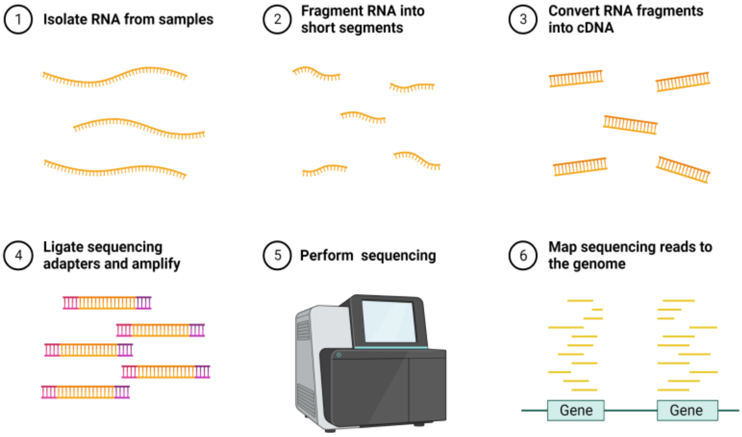

1 Isolate RNA from samples
2 Fragment RNA into short segments
3 Convert RNA fragments into cDNA
4 Ligate sequencing adapters and amplify
5 Perform sequencing
6 Map sequencing reads to the genome
Gene
Gene

Figure 11. Step-by-step workflow of RNA sequencing. The figure is adapted from "RNA Sequencing" in BioRender.com (2023).

Several variations exist in the analysis of RNA-seq data following sequencing. Many tools are available to support different aspects of RNA-seq analysis (Han, Gao, Muegge, Zhang, & Zhou, 2015). However, most of these tools are intended for eukaryotic genomes, which poses a challenge for bacterial RNA-seq.

Bacterial RNA-seq presents different challenges from eukaryotic RNA-seq (for a review, see (Stark, Grzelak, & Hadfield, 2019)). These include overlapping genes, which makes distinguishing gene transcripts a more challenging task. Typical procedures of RNA-seq analysis are therefore quality control checks, aligning reads with a reference genome, and analyzing alignment results. In addition, quality control involves the assembly of overlapping reads without referring to a reference genome (*de novo* assembly), generating raw counts, and determining gene expression differences based on these raw counts.

3 AIMS OF THE STUDY

In this thesis, we examined some of the regulatory mechanisms that govern the global transcription dynamics of the bacterium *Escherichia coli* under stress conditions. The aims of the study were:

1. Develop a method to quantify RNA numbers on cells expressing RNA tagged with MS2-GFP from flow-cytometry data (**Publication I**).

2. Investigate how individual gene responses to genome-wide stresses targeting the transcription machinery can be influenced by the TFN (**Publication II**).

3. Investigate how promoter transcription interference affects the dynamics of genes controlled by natural tandem promoters following shifts in RNAP concentration (**Publication III**).

4. Engineer novel, synthetic tandem promoter arrangements to tune gene responses to transcription-targeting antibiotics (**Publication IV**).

4 MATERIALS AND METHODS

4.1 Strains and Growth Conditions

In **Publication I**, to quantify the RNA expression at the single-cell level using fluorescent tags, we used a multi-copy reporter plasmid responsible for producing MS2d-GFP proteins, controlled by a $P_{LtetO-1}$ promoter that can be activated by aTc. Meanwhile, the target RNAs were produced from a single-copy plasmid under the control of a $P_{lac-ara-1}$ promoter. As such, the target RNA can be expressed by adding IPTG and L-Arabinose. From this system, one produces RNAs that encode for both a fluorescent protein (mRFP1) as well as for ~96 RNA binding sites for MS2-GFP. The plasmids were transformed into the DH5α-PRO cells, which already produce the necessary regulatory proteins (LacI, TetR, and AraC) required for the regulation of the target and reporter plasmids (R. Lutz & Bujard, 1997). Specifically, LacI and AraC repress the Lac and Ara1 promoters respectively, while TetR represses $P_{LtetO-1}$. Note that $P_{LtetO-1}$ is a viral promoter that was altered to be repressed by TetR, by binding to the operator O1 (R. Lutz & Bujard, 1997).

Overnight bacterial cultures containing the plasmids were inoculated in fresh Luria-Bertani (LB) medium with an optical density ($O.D._{600}$) of 0.03. The cultures were then incubated at 37 °C with continuous shaking at 250 rpm. When the cells reached the mid-exponential phase ($O.D._{600}$ 0.3), aTc was introduced in the media at a concentration of 100 ng/ml to induce the expression of the $P_{LtetO-1}$ promoter. Simultaneously, L-arabinose was added at a concentration of 0.1% to pre-activate the target promoter $P_{Lac/ara-1.}$ Finally, we waited for 50 minutes to allow the cells to intake the inducers as well as to allow the cells to accumulate sufficient MS2-GFP.

After the 50 minutes, IPTG was added at varying concentrations (0, 6.25, 50, 100, 200, 300, 500, and 1000 μM). This induction initiated the production of the RNA target for MS2d-GFP. After an additional hour of incubation, the cells were then examined through microscopy or flow cytometry to quantify both tagged RNA and mRFP1 protein levels. The settings for these two measurements are described in subsequent sections.

In **Publication II**, we used MG1655 cells to measure the transcriptome and RL1314 cells with a RpoC subunit endogenously tagged with GFP to measure single-cell RNAP levels (generously given by Robert Landick). In addition, we used strains endogenously tagged with YFP to measure protein levels (YFP strain library) (Taniguchi et al., 2010). We further used a strain carrying the rpoS::mCherry gene to measure the distribution of rpoS in single cells (generously given by James Locke). Finally, we used a low-copy plasmid fusion library coding for GFP to track promoter activity (Zaslaver et al., 2006).

Cells were streaked on LB agar plates (supplemented with appropriate antibiotics of the antibiotic resistance gene of the strain grown), and a single colony was picked. The colony was inoculated into LB medium (with necessary antibiotics) and allowed to grow overnight with aeration at 250 rpm. These cultures were further diluted into tailored LB media in a 1:1000 ratio, at 37 °C with aeration until reaching O.D.$_{600}$ of 0.4.

The tailored LB media (mx, for 100 ml) was prepared using *m* grams of tryptone, *m*/2 grams of yeast extract, and 1 g NaCl in accordance with the protocol described in (Lloyd-price et al., 2016). As an example; LB$_{0.75x}$ (100 ml) has 0.75 g tryptone, 0.375 g yeast extract, and 1 g NaCl. This media, whose preparation consists of a "partial dilution" of LB media, is essential to cause a reduction in intracellular RNAP concentration (Lloyd-price et al., 2016).

In **Publication III,** we used strains of the YFP fusion library (Taniguchi et al., 2010) to study the single-cell protein expression of genes with tandem promoters. We also studied the influence of their transcriptome on genes with tandem promoters by RNA-seq. The protein expression measurements were performed in M9 media (with 0.4% glucose as the carbon source) by flow cytometry and microscopy. The cell growth protocol was the same as in **Publication II**.

The M9 media used for these experiments had the following components: 1x M9 Salts, 0.1 mM $CaCl_2$, 2 mM MgSO4, 5x M9 Salts with 34 g/L Na_2HPO4, 15 g/L KH_2PO4, 2.5 g/L NaCl, 5 g/L NH_4Cl, 0.2% Casamino acids, and 100X vitamins. Diluted M9 media was also used. The media was diluted using autoclaved distilled water to 0.5X.

Finally, in **Publication IV,** we used synthetic constructs tagged with mCherry. The promoters used in this study were Lac0_30_1, TetA, and BAD. These promoters were induced by 1mM IPTG, 15ng aTc, and 0.1% arabinose, respectively. The media used in this study was M9 and the cellular preparation protocol and growth conditions were the same as in **Publication III**. The control strain which was used to transform the plasmids was DH5α-PRO as in **Publication I**.

4.2 Single-cell Gene Expression Microscopy Measurements

To conduct single-cell gene expression measurements, we used microscopy (supported by image analysis) as well as flow-cytometry. In general, in **Publications I-III**, we obtained 'big data' (thousands of cells) by flow-cytometry, and then observed a few hundred cells in detail, by microscopy. The latter observations, commonly, were to confirm that there were no events that could invalidate the conclusions taken from the flow-cytometry.

In all the microscopy experiments, we captured confocal images of the cells using a C2+ (Nikon) microscope. The system uses the point-scanning confocal technique. In general, once the cells reached measurement timing/phase, they were sandwiched between the coverslip and 2% agarose gel pad and then visualized using a Ti-E Nikon inverted microscope, using a 100× objective. The fluorescence intensity of the GFP was measured with an argon ion laser of 488 nm and a filter of 514/30 nm. The morphology of the cells was studied from phase-contrast images, taken with a CCD camera (DS-Fi2, Nikon). Finally, NIS-Elements software was used to capture the images.

In **Publication I,** microscopy data on single-cell RNA levels was collected for more than 300 cells per condition using confocal microscopy (Figure 1 of **Publication I**). Meanwhile, in **Publication II,** we quantified the levels of single-cell RNAP from confocal microscopy images while, at the same time, phase-contrast images were also acquired to learn the above cell morphology (for example, see **Figure S1** of **Publication II**). Additionally, MG1655 cells were imaged in various LB media dilutions (LB$_{1.0x}$, LB$_{0.75x}$, LB$_{0.5x}$, LB$_{0.25x}$, LB$_{1.5x}$, LB$_{2.0x}$, and LB$_{2.5x}$) to measure cell size. Moreover, MG1655 cells in LB$_{1.0x}$ were imaged during stationary growth. Furthermore, cells from the YFP strain library were imaged to evaluate their morphology and physiological state, to ensure that we observed healthy cells.

In **Publication III,** we performed confocal microscopy experiments on the protein levels of genes with tandem promoters (see example Figure 4B in **Publication III**), using cells of the YFP strain library, whose genes were endogenously tagged with YFP. We also performed confocal microscopy experiments of MG1655 cells (not expressing YFP) for background subtraction. Phase contrast images were collected to obtain information on cell morphology. The confocal and phase contrast images were acquired simultaneously.

4.3 Microscopy Image Analysis

In **Publications I-III**, the microscopy images were analyzed by the '*CellAging*' software (Häkkinen et al., 2013). In detail, *CellAging* applies automated image analysis to extract information from the images. First, the cell segmentation was carried out using the Gradient path labeling algorithm (Mora, Vieira, Manivannan, & Fonseca, 2011). This accurately identified and pre-segmented individual cells in the images. Next, to enhance segmentation results and minimize over-segmentation, classifiers were employed to merge relevant segments and discard unnecessary ones, such as unwanted artifacts. These classifiers were constructed using the Classification and Regression Trees algorithm (Breiman, Friedman, Olshen, & Stone, 1984). To ensure accurate classifier training, an expert manually trained the system using many example images. Additionally, to achieve precise segmentation, the software further allows for manual corrections, which were applied when necessary.

Following segmentation, *CellAging* aligned confocal images with corresponding phase-contrast images, employing a semi-automated approach. To perform this alignment, the thin-plate spline interpolation technique was used for the registration transform. By manually selecting landmark points, the cell masks were adjusted to accurately match the corresponding cells in the confocal images. This alignment process ensures that the two types of images are properly registered, facilitating further analysis and comparison. Overall, this software combined automated segmentation algorithms with manual adjustments and registration techniques to provide robust and accurate analysis of cell images.

4.4 Flow-cytometry Measurements

Flow-cytometry experiments were conducted using an ACEA NovoCyte flow cytometer with a blue and a yellow laser (**Publications I-IV).** For this, the cells growing in bacterial cultures were diluted into 1 ml PBS. This mixture was vortexed for 10 s. In each measurement, 50.000 events were collected with a flow rate of 14 μl/min. These events were collected for 3 biological replicates in each condition. To remove the events due to particles smaller than *E. coli*, the forward scatter (FSC)-H threshold was set to 5000. For GFP and YFP detection, the FITC-H channel was used (488 nm excitation and 530/30 nm emission). For mCherry and mRFP1 detection, the PE-Texas Red-H channel was used (561 nm excitation, 615/20 nm emission). YFP and GFP were excited using a blue laser (488 nm) and detected using the fluorescein isothiocyanate (FITC)-H channel.

Control cells (not expressing any fluorescence protein) were measured to obtain an average background fluorescence that was then subtracted to the other data. Prior to each experiment, QC (quality control) was performed as per the recommendations of the manufacturer. Finally, the data was collected using the ACEA NovoExpress software in .csv file format.

In **Publication I**, induction curves were obtained using data from flow-cytometry of cells with target and reporter plasmids. For protein detection, the PMT voltage of PE-Texas Red was set to 584. Unsupervised gating (Razo-Mejia et al., 2018) was applied to the data to remove doublets and data produced from debris. The events that did not produce fluorescence were also removed by the application of a minimum threshold. Overall, less than 5.000 events were discarded in any of the measurements.

In **Publication II,** time-lapse flow-cytometry was performed of RL1314 (RpoC subunit tagged with GFP) cells in $LB_{1.0x}$, $LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$. The time-series was performed for 210 min, with flow-cytometry data acquisition being conducted at every 30 min. Aside from time-series, flow-cytometry data was also acquired at 180 min in $LB_{1.0x}$, $LB_{0.75x}$, $LB_{0.5x}$, $LB_{0.25x}$, $LB_{1.5x}$, $LB_{2.0x}$, and $LB_{2.5x}$. Flow-cytometry data was further acquired for many strains from YFP strain library (Taniguchi et al., 2010) and the crl gene from the promoter fusion library (Zaslaver et al., 2006) in various media conditions. The PMT voltage was always set at 550 for GFP detection. Finally, flow-cytometry data was also acquired for RpoS levels using the

MGmCherry strain in exponential and stationary phases, in various media. The PMT voltage was set to 561 nm for detection of mCherry.

In **Publication III**, flow-cytometry data was collected for 30 genes controlled by tandem promoters from the YFP fusion library (Taniguchi et al., 2010). These measurements were performed in exponential growth phase in two conditions of M9 media (1X and 0.5X). Flow-cytometry of MG1655 cells (control strain for background subtraction) was also performed in same conditions. Flow-cytometry data was also acquired for RNAP levels (using RL1314 strain, where RpoC subunit is fused with GFP) and W3110 strain (for background subtraction). Three biological replicates were performed for each condition. For the removal of debris and doublets, unsupervised gating was applied to the data (Razo-Mejia et al., 2018). Further, secondary gating was also applied to the distributions to remove outliers.

In **Publication IV**, flow-cytometry data was obtained for synthetically engineered inducible tandem promoters and their component promoters. The individual promoters were P_{lac} (inducible by IPTG), P_{tetA} (inducible by aTc), P_{BAD} (inducible by arabinose), and the tandem promoters were P_{lac}-P_{tetA}, P_{tetA}-P_{lac}, and P_{tetA}-P_{BAD}-P_{lac}. These promoters are tagged with mCherry. The flow-cytometry distributions were obtained during full induction of each promoter and non-induction cases (for control) in M9 media after 180 min of induction. The data was collected for 3 biological replicates in each condition. Outliers were removed by discarding 1% events with the highest fluorescence intensities.

Finally, in all experiments above, three biological replicates were performed for all conditions. Moreover, data was also collected in the same conditions for the control cells (strains not tagged with any fluorescence protein) to subtract background fluorescence. Finally, abnormalities in the data, such as bimodal distributions, were carefully analyzed and, in some cases, were repeated, to ensure that it was not due to the presence of many unhealthy cells.

4.5 Spectrophotometry

In this thesis, a BioTek Synergy HTX Multi-Mode Microplate Reader equipped with Gen5 software was used to measure the optical densities and fluorescence intensities of cell cultures over time. This spectrophotometer has inbuilt incubation with shaking, and temperature control that allows obtaining data in live cells, which can reach up to 50C temperature. We also made use of a spectrophotometer (Ultrospec 10; GE Healthcare), which is only capable of capturing one time-point absorbance measurement.

From overnight cultures, cells were diluted in fresh medium and then allowed to grow in a shaking incubator until reaching an OD_{600} of 0.3. Next, the cells were aliquoted into 24-96 micro-well plates and grown in appropriate media, while keeping the temperature constant and shaking consistent. To measure cellular GFP fluorescence intensities over time, we used the excitation filter of wavelength 485/20 nm and the emission filter of wavelength 525/20 nm. For mCherry detection, the excitation and emission wavelengths used were 575/15 nm and 620/15 nm, respectively, with a gain of 50.

In **Publication I**, we used spectrophotometry to measure the fluorescence of i) cells carrying a multi-copy reporter plasmid for producing MS2d-GFP proteins, under the control of $P_{LtetO-1}$, inducible by aTc. Also, present was a single-copy target plasmid producing an RNA coding for mRFP1 upstream of a 96 MS2 binding site array, controlled by the promoter $P_{Lac/ara-1}$, inducible by IPTG and L-Arabinose. ii) cells with only the reporter plasmid. This time series experiment was performed for 10 hours, at an interval of 10 mins, with 6 technical replicates.

In **Publication II**, we used spectrophotometry to measure the growth curves ($O.D._{600}$) of cells with RL1314 strain growing in several media conditions ($LB_{1.0x}$, $LB_{0.75x}$, $LB_{0.5x}$, $LB_{0.25x}$, $LB_{1.5x}$, $LB_{2.0x,}$ and $LB_{2.5x}$). For this, the overnight cultures were diluted into respective media with the starting $O.D._{600}$ 0.05 and aliquoted into 24-well transparent plates. The growth was monitored for 10 hours, every 10 minutes with continuous shaking.

In **Publication III**, we used spectrophotometry to measure growth curves ($O.D._{600}$) of MG1655 cells in 0.25X, 0.50X, and 1X in M9 media. The overnight cultures were diluted into respective fresh M9 media and $O.D._{600}$ was measured for 450 min, while recording the measurement every 30 min with three biological replicates for 450 min while recording every 30 min. This was repeated for three biological replicates.

In **Publication IV**, we used spectrophotometry to measure the optical density ($O.D._{600}$) to measure the cell growth rates of the DH5α-PRO cells and DH5α-PRO cells carrying different plasmids (Figure 2 of **Publication IV**). The growth was measured for 720 min and the measurement was recorded every 20 min. We also measured the fluorescence of these cells having plasmids tagged with mCherry after full inductions and after the application of antibiotics targeting transcription (rifampicin and ofloxacin). The measurements were recorded every 20 min for a period of 650 min with three biological replicates.

4.6 Protein Quantification using Western Blotting

In Publication II, we used the western blot technique to quantify the relative RNAP levels in MG1655 and RL1314 strains. The overnight cultures were inoculated into respective fresh media (LB_{1x}, $LB_{0.75x}$, $LB_{0.5x}$, $LB_{0.25x}$) and were allowed to grow until $O.D._{600}$ 0.4. The cultures were next subjected to centrifugation (8000 rpm for 5 min) and the cell pellet obtained was lysed using bacterial protein extraction reagent. To this, a protease inhibitor was added at room temperature for 10 min. After this, centrifugation was done at 14000 rpm for 10 min and the supernatant was collected. This supernatant was mixed in 4X Laemmli buffer with β-mercaptoethanol and this mixture was boiled at 95C for 5 min. Next, the samples with total soluble proteins were loaded on TGX stain-free precast gels. These proteins were separated by electrophoresis and transferred on PVDF membrane. The membrane was blocked with 5% non-fat milk at room temperature for 60 min and then tagged with primary antibodies (1:2000 ratio) at 4C overnight.

Next, HRP-secondary antibody (1:5000) treatment was performed for 60 min (RpoC antibodies for MG1655 and GFP antibodies for RL1314). Excess antibodies were removed by buffer wash. The membrane was then treated with chemiluminescence reagent, and the bands were detected. Images were obtained by the Chemidoc XRS (Biorad). The quantification of protein bands was performed using the Image Lab software. The images show three bands that are directly associated with the molecular weight of RNAP-GFP, RNAP and GFP, respectively. Consequently, they allowed determining, e.g., the fraction of RNAP and GFP molecules that are bound.

4.7 Genome-wide RNA-sequencing

In **Publication II**, we performed RNA seq measurements in several media richness conditions ($LB_{1.0x}$, $LB_{0.75x}$, $LB_{0.5x}$, $LB_{0.25x}$, $LB_{1.5x}$, $LB_{2.0x}$, and $LB_{2.5x}$). We used MG1655 cells and performed three biological replicates in each condition. RNA-seq data was collected for the following conditions:

1. $LB_{0.75x}$, $LB_{0.5x}$, $LB_{0.25x}$ at 180 min relative to $LB_{1.0x}$
2. $LB_{0.5x}$ at 60 and 125 min relative to $LB_{1.0x}$
3. $LB_{1.5x}$, $LB_{2.0x}$, and $LB_{2.5x}$ at 180 min relative to $LB_{1.0x}$

For RNA extraction, first, the cells were treated with RNA protect bacteria reagent (Qiagen, Germany). This was done to prevent RNA degradation. The RNA was then extracted using RNeasy kit by Qiagen. The RNA was treated with DNAase (Turbo DNA-free kit, Ambion) and quantified using Qubit Fluorometer assay. Total RNA abundance was determined by gel electrophoresis (using 1% agarose gel, SYBR safe stain). Chemidoc XRS imager (Biorad) was used to detect RNA.

The sequencing of $LB_{0.75x}$, $LB_{0.5x}$, $LB_{0.25x}$ at 180 min relative to $LB_{1.0x}$ (point 1 above) was performed by Acobiom (Montpellier, France). The RNA integrity number (RIN) of the samples was determined using the 2100 Bioanalyzer. To remove ribosomal RNA, the Ribo-Zero removal kit (Illumina) was used. Following the Illumina protocol, RNA-seq libraries were constructed. The samples were sequenced on an Illumina MiSeq instrument in a single-index (1×75 bp single-end configuration, 10M reads per library).

For conditions 2 and 3, the sequencing was performed by GENEWIZ (Leipzig, Germany). The RNA integrity number (RIN) of the samples was determined using the Agilent 4200 TapeStation. Ribosomal RNA depletion was carried out using the Ribo-Zero Gold Kit (Illumina). NEB Next Ultra RNA Library Prep Kit was used to construct the RNA-seq libraries.

For conditions $LB_{1.5x}$, $LB_{2.0x}$, and $LB_{2.5x}$ at 180 min relative to $LB_{1.0x}$, the Illumina HiSeq 4000 instrument was used (2 × 150 bp paired-end configuration, 350M reads per lane). For conditions $LB_{0.5x}$ at 60 and 125 min relative to $LB_{1.0x,}$ the Illumina NovaSeq 6000 instrument was used (2 × 150 bp paired-end configuration, 10M reads per lane). Raw sequence data was converted into fastq files using Illumina bcl2fastq v.2.20 software.

In **Publication III**, RNA seq experiments were performed using the *E. coli* MG1655 strain. The cells were grown in different growth-phase conditions and at time points: 0, 20, and 180 mins. RNA was isolated from three independent biological replicates using the RNeasy kit (Qiagen) as in the **Publication II**. The sequencing was performed by GENEWIZ (Leipzig, Germany) using Illumina HiSeq instrument, and the raw files were converted into fastq files using Illumina bcl2fastq v.2.20.

4.8 Genetic Engineering

In **Publication IV**, we designed and engineered synthetic constructs using Snapgene Software (GSL Biotech). For this, we used well-known individual synthetic promoters (P_{lac}, P_{BAD}, and P_{tetA}) (R. Lutz & Bujard, 1997; B. R. Smith & Schleif, 1978). These individual promoters were then arranged in tandem formations: 'P_{lac}-P_{tetA}', 'P_{tetA}-P_{lac}', and 'P_{tetA}-P_{BAD}-P_{lac}' (going upstream to downstream from left to right) using Snapgene software such that the distance between the TSSs of promoters is always between 150 and 200 bps. This distance was produced by the insertion of random sequences. The absence of hairpin loops and other translational products in the spacer sequences was confirmed. These biopart sequences were then assembled in Integrated DNA Technology, Iowa, U.S.A. A single-copy pBAC plasmid was used as then vector backbone where the bioparts were integrated (Lee et al., 2016). This vector also contains a coding region for mCherry supported by a strong RBS, to ensure visualization. Moreover, the maturation time of mCherry is 15 min which allows easy mapping of RNA numbers from the fluorescence intensity. Finally, the proteins are non-toxic and stable to reduce influences from perturbations of the phenotype (Lambert, 2019).

This fully formed construct, with each biopart integrated in the vector backbone, is named 'device'. These devices were transformed into a host strain of *E. coli* (DH5α-PRO). Critically, this strain already carries the repressors for tight negative regulation of the promoters P_{lac}, P_{BAD}, and P_{tetA}, in the absence of external activation. For their transformation, first, chemically competent DH5α-PRO cells (CC) were prepared. Plasmid DNA and DH5α-PRO CC were mixed in 1:10 ratio, accompanied with 30 min of ice incubation. This mixture was then given a heat shock (42 °C) in a water bath for 1 min. 800 μL of LB media was then added to the mixture and allowed to shake at 37 °C and 250 RPM. After 1 hour, 200 μL of the mixture was plated on agar plates supplemented with 34 μg/mL of chloramphenicol. These plates were kept overnight and the colonies containing the transformants (each device in DH5α-PRO cells) were harvested the next morning.

5 RESULTS

5.1 Quantifying the Single-cell Distributions of RNA molecules Using MS2-GFP tagging System from flow-cytometry data (Publication I)

In **Publication I**, we showed that it is possible to quantify with accuracy the single-cell distributions of number of RNA molecules (tagged with multiple MS2-GFPs) in live bacterial cells, from flow-cytometry data. The method proposed overcomes what likely is the major limitation of the MS2-GFP system of RNA counting, based on microscopy. Specifically, using microscopy, it is necessary to apply image analysis in order to extract RNA numbers from live cells. Unfortunately, this analysis is more complex than standard cell segmentation (which is already a arduous process), because it is further necessary to segment individual fluorescent spots and then quantify how many MS2-GFP tagged RNAs compose each spot (Häkkinen et al., 2013; Häkkinen & Ribeiro, 2014, 2016). Consequently, most studies using this technology limited their data to less than a few hundred (Golding et al., 2005) to a few thousand cells (Lloyd-Price et al., 2016).

Because of this, we set out to develop a method to automatically extract single-cells numbers of MS2-GFP tagged RNA molecules from flow-cytometry data. We first considered that one important feature of the MS2-GFP system is that a single RNA tagged with 96 MS2-GFP (Golding et al., 2005) is fluorescent enough for its signal to not be lost due to cellular background nor due to the background signal from individual, free-floating MS2-GFP proteins (unbound to any RNA) (Tran et al., 2015). Moreover, we observed by spectrophotometry when new RNA molecules are tagged by MS2-GFPs, the total cell fluorescence increases (Figure 12). While we have not established the reason for this, one explanation could be the 'immortalization' of the MS2-GFP proteins once bound to a target RNA. That immortalization would cause bacteria producing tagged RNA to carry more MS2-GFP, on average, than bacteria without target RNA (Figure 1 of **Publication I**).

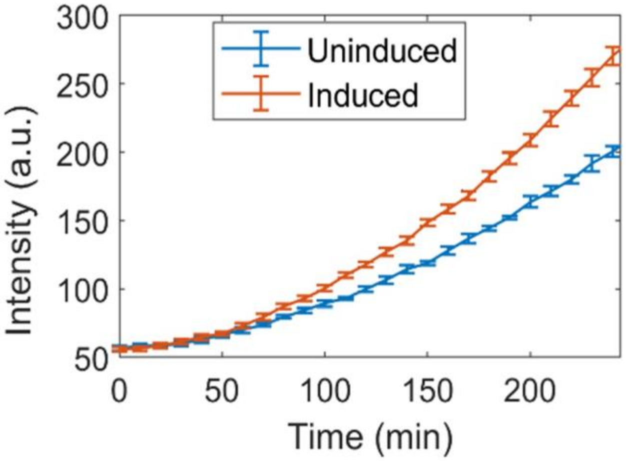


Figure 12. Time series using spectrophotometry. Fluorescence intensity (in arbitrary units) of cell populations over time, was obtained from cells with target and reporter plasmids induced (light brown line) and from cells with only the reporter plasmid induced (blue line). Figure taken from **Publication I.**

To test for other explanations (e.g., that the inducer adds fluorescence), we further determined the mean fluorescence of cells containing only the reporter gene by flow-cytometry. We observed that it was not affected by adding an inducer of the target gene (IPTG), as expected (Figure 2 of **Publication I**). In contrast, the mean cell fluorescence of cells carrying both target and reporter genes increases with IPTG (Figure 2 of **Publication I**). This implies that RNAs, once with MS2-GFP tags, increase the total cell fluorescence. In support, we observed the same using microscopy images.

Next, we measured how the total cell fluorescence, when measured by flow cytometry, scale with the changes in RNA numbers as measured by microscopy and image analysis (Figure 4 of **Publication I**). Unfortunately, we observed that it is not feasible to quantify with precision the number of MS2-GFP tagged RNAs in individual cells from flow-cytometry data. However, it is possible to quantify the single-cell distributions of these numbers. Visibly, there is a clear correlation between this distribution when obtained by microscopy with when obtained by flow-cytometry (Figure 13). Therefore, it should be possible to "calibrate" the flow-cytometry to the microscopy data, so as to directly extract single-cell distributions of RNA numbers from the flow-cytometry data.

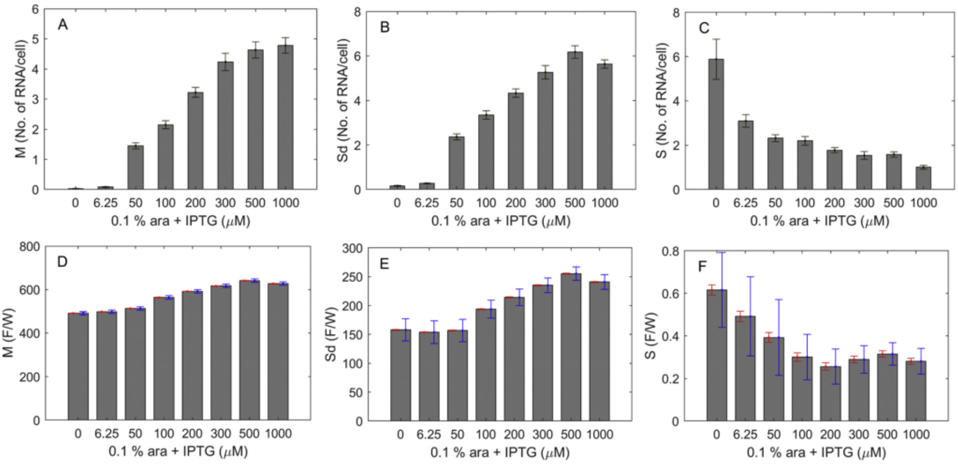


Figure 13. (A, B, and C) Single-cell moments of distribution (mean, standard deviation, and skewness) of RNA numbers by microscopy with increase in IPTG concentration. **(D, E, and F**) Single-cell moments of distribution (mean, standard deviation, and skewness) of F/W values by flow cytometry with increase in IPTG concentration. The figure is taken from **Publication I.**

The calibration should be possible by using two or more data points from flowcytometry and corresponding data points from microscopy. We performed this calibration independently for the distributions of the mean, standard deviation, and skewness for different inductions strengths (Figure 7 of **Publication I)**. We found that the best fitting line of the calibration could not be distinguished from the ideal line, which proves the accuracy of the method.

Overall, we expect this technique to be able to contribute to studies of transcription at the single cell level, by facilitating the collection of data. Moreover, this method should also be applicable to other technologies relying on single-cell imaging such as FISH. Increased data should allow dissecting regulatory mechanisms that only weakly affect gene expression dynamics and, due to that, may have not yet been discovered.

5.2 Transcription Factor Network Dynamics Following the Shifts in Media Richness (Publication II)

In **Publication II**, we studied how the transcription factor network (TFN) of bacteria responds to stresses. Lack of environmental nutrition is one such stress. From previous studies (Lloyd-price et al., 2016), we knew that it leads to a quick decrease in the numbers of RNA polymerases inside *E. coli*. Hence, it should cause a great disturbance in the TFN. On the other hand, cells adapt quickly to this stress, and thus the TFN responses are likely beneficial. We thus studied the TFN dynamics following quick shifts from rich to poor media (Figure 2 of **Publication II)** and vice versa (Figure 6 of **Publication II**).

We observed shifts of 25-50% in RNAP levels at ~ 75 min which stabilized at ~ 165 min (Figure 2 of **Publication II**). Moreover, we found that 180 min measurement time sufficed to detect RNA changes due to the changes in RNAP and due to changes in the numbers of direct input TFs. Consequently, the data can be used to study the influence of TF numbers on the genome-wide responsiveness.

We then investigated if and how the initial gene responses propagated to other genes, because of the TFN. To make this possible, we extracted the known TFN of *E. coli* from the RegulonDB database (Santos-Zavaleta et al., 2019). We also studied the functional role (activation or repression) of each TF. By confronting the data on gene expression responses, with the data on gene-gene interactions via transcription factors, we found that, on average, the difference in the numbers of activator and repressor input TFs of a gene is strongly correlated to the response strength of that gene (Figure 14A).

Moreover, genes separated by a path length of two or more TFs in the TFN did not have correlated response strengths (Figure 14B). This suggests that the information on the stress did not propagate beyond the nearest neighbor genes in the TFN (during the measurement time of 3 hours). As such, the results above are mostly due to the interactions between nearest neighbors in the TFN.

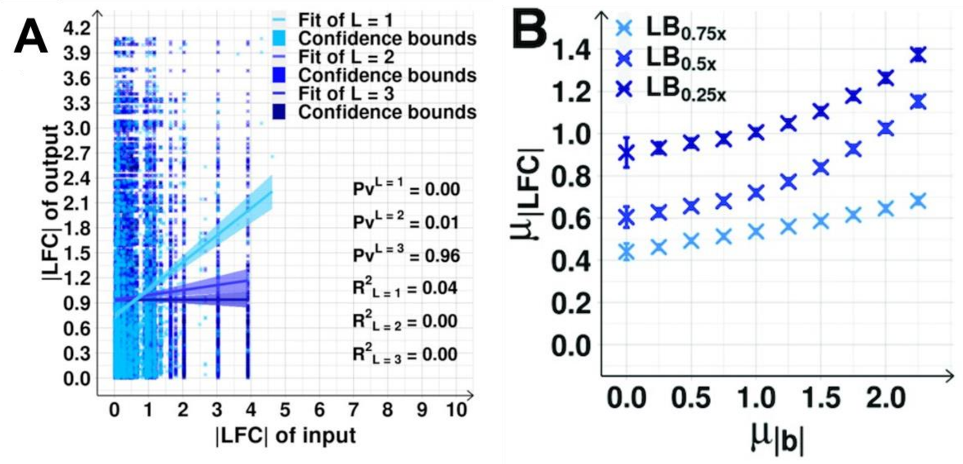


Figure 14. (A) Correlation plot between LFC of input and output genes distanced by minimum path length of 1, 2, and 3**. (B)** Mean changes in LFC as a function of the mean bias of the effects of input TFs. The figure is taken from **Publication II.**

Finally, GRs, σ factors, non-coding RNAs, (p)ppGpp, and structural parameters of the TFN (e.g., betweenness, stress centrality, clustering coefficient, eccentricity, and out-degree) had little to no influence on the main results. Overall, the results constitute empirical evidence that the TFN (topology and logic) can influence global stress responses of *E. coli*. Potentially, they may also influence the responses to antibiotics.

5.3 d_{TSS} and RNAP- promoter Occupancy Times play a major role in the Dynamics of Genes Controlled by Tandem Promoters (Publication III)

In **Publication III**, we studied the dynamics of genes controlled by tandem promoters as a function of the nucleotide distance (d_{TSS}) and the time intervals during which RNAP binds with the downstream promoter. Tandem promoters were defined as two closely spaced promoters controlling the transcription of the same gene in the same direction, in agreement with (Shearwin et al., 2005). Our aim was to find rules controlling the dynamics of tandem promoters to guide the design of future synthetic genetic circuits.

The *E. coli* genome contains 831 genes controlled by more than one promoter in tandem configuration (Santos-Zavaleta et al., 2019). We began by identifying which genes are controlled solely by their tandem promoters (rather than other nearest-neighbor promoters). We identified 102 such pairs of genes (Section 'Selection of natural genes controlled by tandem promoters', S1 Appendix, **Publication III**).

We collected their single-cell protein expression levels (limited to the 30% of those genes that are represented in the YFP protein fusion library) (Taniguchi et al., 2010). We found that the dynamics of tandem promoters, when close enough for their TSSs to be distanced by less than 35 bp, differ significantly from when the distance is larger than 35 bp. From this, we concluded that, if less than 35 bp, RNAPs in one TSS most likely can block other RNAPs from binding to other TSS. In agreement with increased expression levels, we also found that CV^2 and skewness decreased with d_{TSS} (Figure 15).

Exponential 1, Exponential 2, and Step functions were then used to model the relationship between protein expression levels and $d_{TSS.}$ The equations are shown in Table 2. The result in Figure 15 shows that that the Step function best fits the empirical data (R^2 equal to 0.36 in mean optimal condition), which we validated using the data from 0.5X M9 media richness. This can be explained by the occurrence of occlusion of one TSS due to the occupation of the other TSS by RNAP. As such, the time length that the RNAPs occupied the promoter during transcription initiation was a key variable, as it will influence the propensity for the RNAPs to occupy a TSS at any given time.

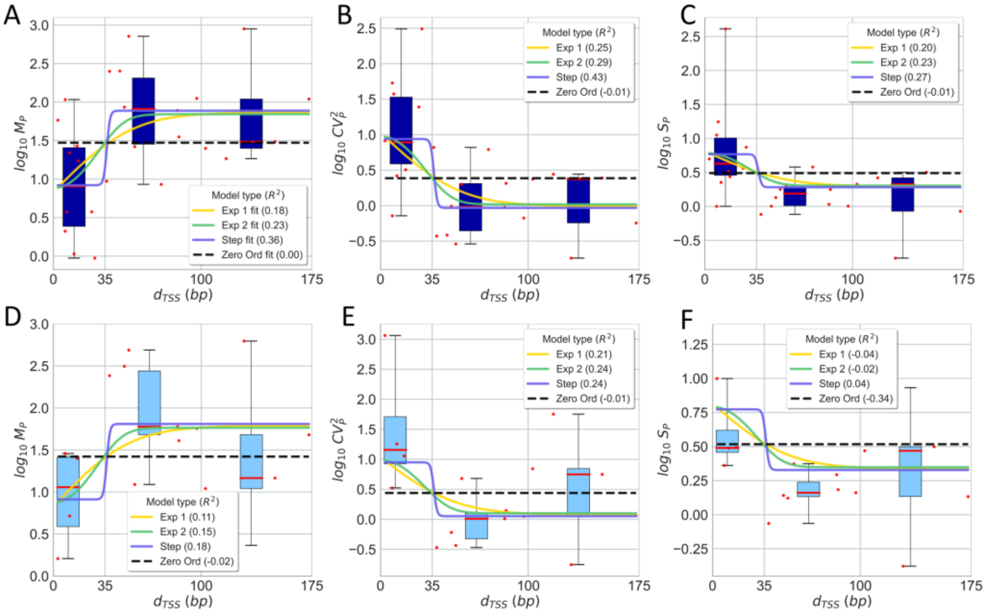


Figure 15. Influence of d_{TSS} on the single-cell protein numbers of genes controlled by tandem promoters and the analytical model. (A) Mean, (B) CV², and (C) S of protein numbers in the 1X media as a function of d_{TSS}. (D), (E), and (F) show the same for the 0.5X media, respectively. Red dots are the mean values from 3 biological repeats. These dots are grouped in boxes based on their d_{TSS}. In each box, the red line is the median. The vertical bars are the range between the minimum and maximum of the red dots. The insets show the R² for each model fit and prediction. Figure is taken from **Publication III.**

Table 2. Models of transcriptional interference due to promoter occlusion. Taken from **Publication III**.

Interference by occlusion	***I(dTSS)***	***kocclusion***
Exponential 1 ("Exp1")	$e^{-(b_1 \cdot d_{TSS})}$	$k_{ocl}^{\max} \cdot e^{-(b_1 \cdot d_{TSS})} \cdot \omega$
Exponential 2 ("Exp2")	$e^{-\left(b_1 \cdot d_{TSS} + b_2 \cdot d_{TSS}^2\right)}$	$k_{ocl}^{\max} \cdot e^{-\left(b_1 \cdot d_{TSS} + b_2 \cdot d_{TSS}^2\right)} \cdot \omega$
Step ("Step")	$1 - \dfrac{1}{1 + e^{-m \cdot \left(d_{TSS} - L\right)}}$	$k_{ocl}^{\max} \cdot \left(1 - \dfrac{1}{1 + e^{-\left(d_{TSS} - L\right)}}\right) \cdot \omega$, for m = 1 bp^{-1}
Zero order ("ZeroO")	k	$k_{ocl}^{\max} \cdot \omega$

Here, b_1 represents the first order exponential decay constant, $k_{ocl}^{\max}$ is the maximum occlusion possible, ω represents fraction of time that 'other' promoter is occupied, b_2 stands for the second-order exponential decay constant, L is the length of DNA (in bp), k is Interference by occlusion is constant irrespective of d_{TSS}, and m stands for the steepness of the step.

Based on the data on the single-cell distributions of protein numbers, we proposed and validated a new analytical model for the dynamics of promoters in tandem formation (Figure 16). The model aimed to account for known mechanisms of transcriptional interference in closely spaced promoters, and was based on past models proposed in (Callen et al., 2004; Shearwin et al., 2005; Sneppen et al., 2005). Thus, occupancy times of promoter regions are key parameters (**Publication III**). In conclusion, the model suggests that RNAP-promoter occupancy times and d_{TSS} are the main determinants of interference.

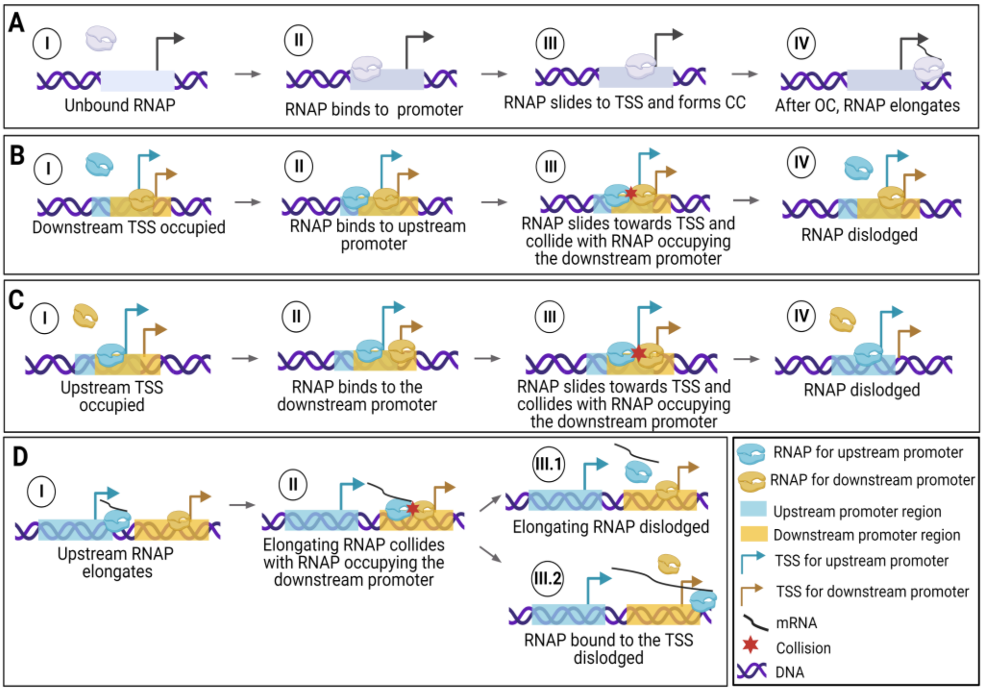

A
I
Unbound RNAP
II
RNAP binds to promoter
III
RNAP slides to TSS and forms CC
IV
After OC, RNAP elongates
B
I
Downstream TSS occupied
II
RNAP binds to upstream promoter
III
RNAP slides towards TSS and collide with RNAP occupying the downstream promoter
IV
RNAP dislodged
C
I
Upstream TSS occupied
II
RNAP binds to the downstream promoter
III
RNAP slides towards TSS and collides with RNAP occupying the downstream promoter
IV
RNAP dislodged
D
I
Upstream RNAP elongates
II
Elongating RNAP collides with RNAP occupying the downstream promoter
III.1
Elongating RNAP dislodged
III.2
RNAP bound to the TSS dislodged
RNAP for upstream promoter
RNAP for downstream promoter
Upstream promoter region
Downstream promoter region
TSS for upstream promoter
TSS for downstream promoter
mRNA
Collision
DNA

Figure 16. Transcriptional interference between tandem promoters. **(A)** The sequence of events in transcription in isolated promoters. A similar set of events occurs in tandem promoters if only one RNAP interacts with them at any given time. **(B / C)** Interference due to the occlusion of the downstream / upstream promoter by a bound RNAP, which will impede the incoming RNAP from binding to the TSS. **(D)** Interference of the activity of the RNAP incoming from the upstream

promoter by the RNAP occupying the downstream promoter. One of these RNAPs will be dislodged by the collision. Figure is taken from **Publication III.**

Overall, our findings revealed that the dynamics of tandem promoters can be predicted by a model of the transcription initiation dynamics of the component promoters when not closely spaced, along with the nucleotide distance between them when closely spaced. This suggests that a method could be implemented, where from a library of promoters whose individual dynamics is known and from a pre-selected distance, one could assemble tandem promoters with novel, and yet predictable dynamics. Such could assist the design of novel circuits.

5.4 Engineering Genes with Predictable Dynamics Using Synthetic Tandem Promoters (Publication IV)

In **Publication IV**, we reported on our new genetic constructs of tandem promoters and findings using them. As first discussed in Publication 3, in Publication 4 we started by hypothesizing that it should be possible to engineer genes with predictable dynamics from promoters with known dynamics. Specifically, based on the results in **Publication III**, it should be possible, by placing one promoter in an upstream position from another promoter, to create a combined dynamics that is stronger than the dynamics of the downstream promoter alone, although not as strong as if the two promoters were fully independent. I.e., the interference due to the tandem formation should lead to new dynamics that ought to be tunable.

To test this, as a proof-of-concept, we designed synthetic constructs in tandem formation based on the well-characterized promoters P_{lac}, P_{tetA}, and P_{BAD} (Figure 17). The engineered tandem constructs are (P_{lac}-P_{tetA}), (P_{tetA}-P_{lac}), and (P_{tetA}-P_{BAD}-P_{lac}). These tandem constructs are in single-copy plasmids and are distanced by a 150-200 bp random sequence between them (for a detailed description, see Figure 17). Having measured their dynamics, first, we showed that the synthetic tandem promoters have the same mean-to-noise relationship (Figure 4 in **Publication IV**) as observed in natural *E. coli* promoters (**Publication III**). This suggests that the synthetic tandem promoters do not differ widely in behavior from natural ones, i.e. in an unpredictable manner.

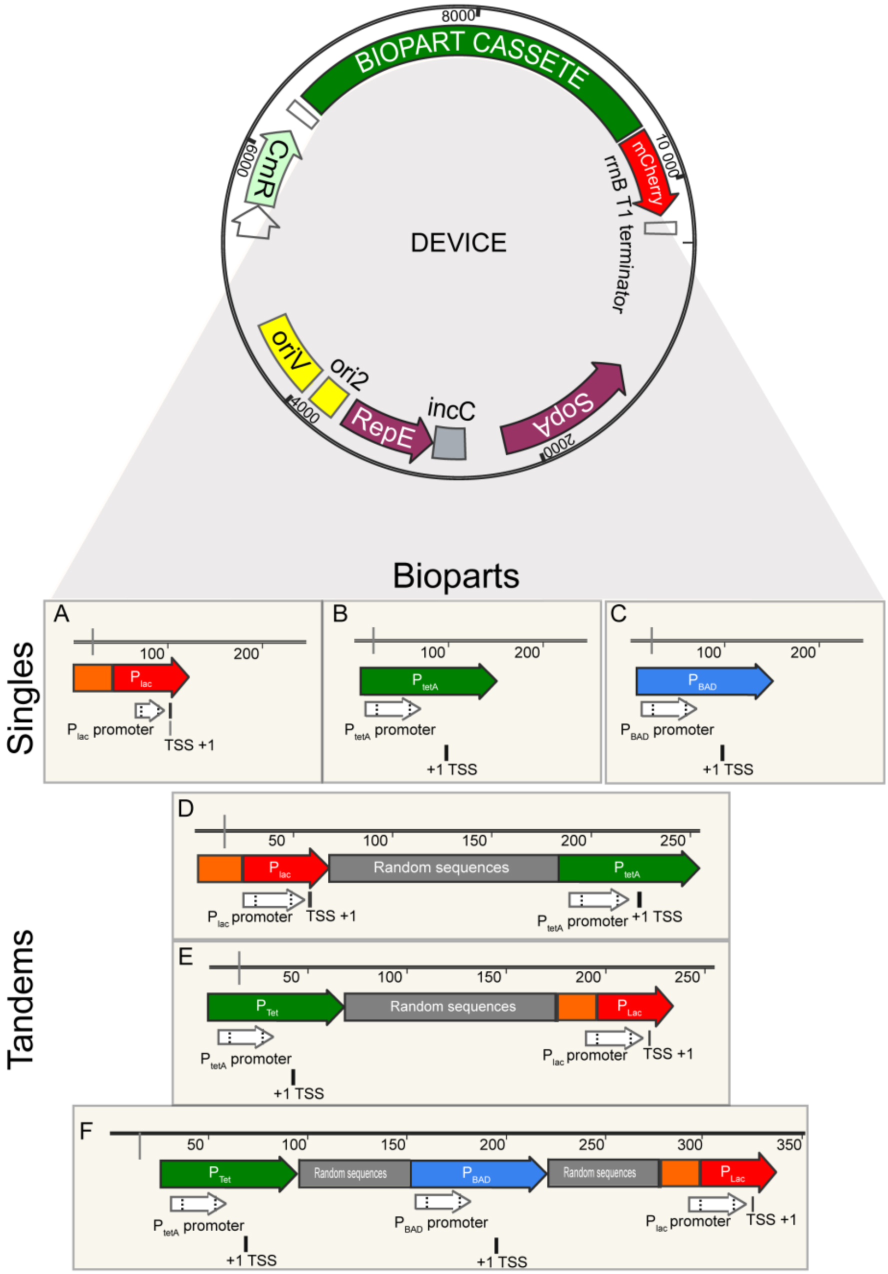


Figure 17. Illustration of the synthetic device and biopart cassettes integrating individual promoters (A, B, and C) into tandem (D, E, and F) configurations. Each biopart is inserted into the biopart cassette (shown in green) forming a device. The tandem bioparts contain the Lac,Tet, and BAD promoters, whose transcription start sites are distanced by 150 bp. In the bioparts D and E, the upstream and downstream promoters are flipped. Figure created with SnapGene and edited using Adobe Illustrator.

Next, by comparing the mean expression rates of the tandem promoter constructs, we showed that the dynamics of the downstream promoter is amplified by the activity of the upstream promoter (Figure 18). Moreover, we showed that this amplification is also a function of the strength of promoters. Based on this result, we proposed a general model to explain the kinetics of synthetic tandem promoters:

$$A_T = A_D + A_U . f(A_U, A_D), \text{ where } f(A_U, A_D) \geq 1 \qquad 5.4.1$$

Here, A_D corresponds to the expression level of the downstream promoter, A_U corresponds to the expression level of the upstream promoter, and A_T stands for the overall expression level of the two promoters when placed in tandem formation.

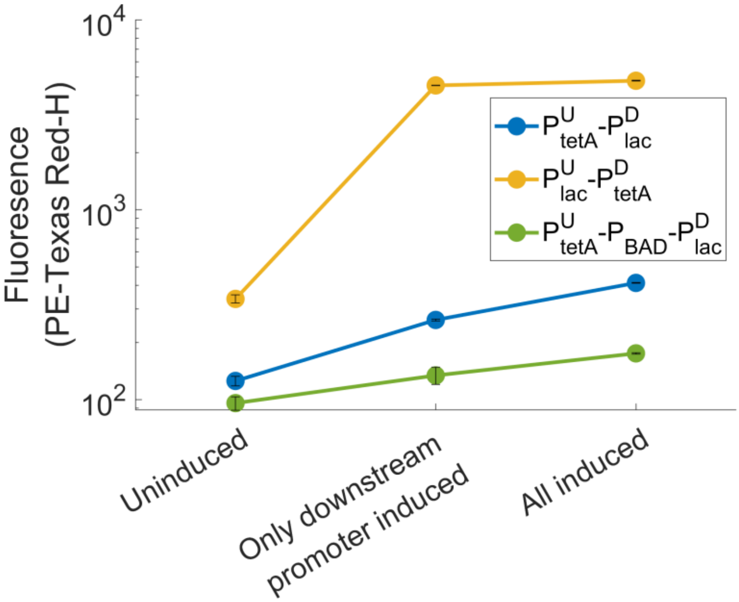


Figure 18. Mean Protein levels of tandem promoters when: a) uninduced; b) only the downstream promoter is induced, and c) both promoters are fully induced. The figure is modified from **Publication IV.**

Next, using time-lapse spectrophotometry (Figures S2 and S3 in **Publication IV**), we studied the responses of the new constructs to stresses caused by antibiotics known to target the transcription machinery. These antibiotics, rifampicin, and ofloxacin, act by blocking the main regulatory elements of the transcription machinery (Campbell et al., 2001). In detail, first, Rifampin inhibits RNA polymerase. This occurs through the blocking of the elongating RNA's pathway

(Kurepina, Chudaev, Kreiswirth, Nikiforov, & Mustaev, 2022). On the other hand, ofloxacin inhibits DNA gyrase and topoisomerase IV, both of which are type II topoisomerases (Shen et al., 1989). Specifically, ofloxacin binds to DNA-bound DNA gyrase, which increases the rate of double-stranded breaks in the DNA (Todd & Faulds, 1991). We expected that the effects of these antibiotics should depend on the strength of the downstream promoter. Interestingly, we found that for P^U_{tetA}-P^D_{lac}, the antibiotics attenuate the synergy levels, while for P^U_{lac}-P^D_{tetA}, the antibiotics amplify the synergy (Figure 19).

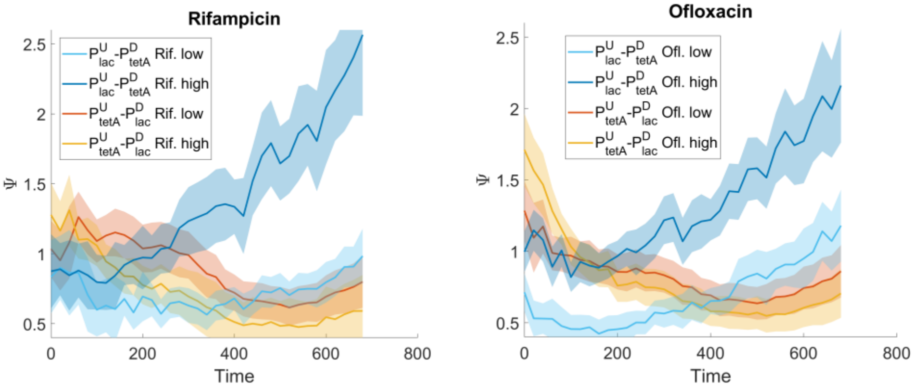


Figure 19. Fluorescence levels normalized by the sum of the individual component promoters, relative to the normalized fluorescence in control conditions. The population fluorescence levels are measured by spectrophotometer over time (in minutes). The shaded areas are the standard error of the mean of 3 biological replicates. The figure is modified from **Publication IV.**

Here, synergy (denoted as 'Ψ' in Figure 19) is assumed to occur when the overall expression is higher than the sum of the expression levels of the two promoters, when not in tandem formation. These results mean that the new constructs have complex responses to the antibiotics. The results also suggest that the tandem constructs can achieve a wide range of dynamical responses. Thus, they can be used to enrich the library of promoters with unique dynamics which can be used in future synthetic genetic circuits.

6 DISCUSSION

Bacteria face several environmental stresses during their lifetime, such as nutrient scarcity, pH shifts, and non-optimal temperatures (Chung, Bang, & Drake, 2006). Overcoming these stresses requires phenotypic adaptations (Jozefczuk et al., 2010; Patange et al., 2018), such as the adjustment of the growth rate accordingly (Jozefczuk et al., 2010). Underlying these adaptations are genome-wide modifications. Some modifications involve one to a few genes, while others require the down- or upregulation of many genes (Hengge-Aronis, 2002a; Jozefczuk et al., 2010).

To perform complex modifications, bacteria have evolved transcriptional programs involving hundreds of genes. These programs are made possible by having multiple genes that share similar internal features (e.g., similar promoter arrangements or supercoiling sensitivity) or TFs. Some TFs are known to control a few hundred genes (Martínez-Antonio & Collado-Vides, 2003).

To dissect why some genes, but not others, respond similarly to certain genome-wide stresses, one needs knowledge of the internal features of the TFN. For this reason, we used *E. coli* as the model organism in our study. Specifically, its TFN is probably one of the best characterized, due to more than 25 years of studies (Fang et al., 2017). Here, we used this knowledge to study the correlations between genes' response to stresses as a function of their known interactions with other genes via TFs (including the global regulator σ^{38}) and promoters' proximity. Our findings provide direct evidence that these interactions influence the genome-wide response dynamics to the stresses studied. For example, we observed that, following changes in RNAP concentration, the response strength of pairs of genes that interact via TFs is more correlated than between randomly selected pairs of genes. In the same study, we further observed that, e.g., the response strength of each gene was correlated to the number of TFs regulating that gene.

Our presently reported observations were restricted to stresses caused by changes in media richness affecting RNAP concentration and/or growth rates. Nevertheless, we expect similar future observations following responses to other genome-wide stresses. Specifically, we expect that many stresses will initially activate a specific set of genes.

We also expect that responsive genes will either share a similar feature, which provides them their similar responsiveness, and/or will be under the control of the same TFs. Moreover, in most cases, following the initial up/down-regulation of a gene cohort, the information will likely be sent through the TFN to other genes, which will then originate a longer-term response. We expect this because, e.g., for many stresses, the effect of a TF (activation or repression of other gene(s)) should be relatively independent from the nature of the stress.

Finding the properties that determine which genes respond to specific stress remains complex (Larsen, Röttger, Schmidt, & Baumbach, 2019). E.g., in general, and as noted above, it is yet not possible to predict gene responses based solely on the promoter sequence, due to the interference from other features of the natural genomes. However, we found that this can become more feasible by reducing the search to a set of genes sharing properties and focusing on a specific stress or a small set of similar stresses. By focusing on a reduced set of promoters with the same σ factor dependency, we were able to establish that the promoter sequence of that gene cohort partially explained those genes' response to a specific change in conditions (from exponential to stationary growth). Another method, employed in (Garcia & Phillips, 2011), is to engineer all possible sequences of promoters (within a restricted range of nucleotides) and measure the response of each synthetic promoter. This methodology should assist in verifying predictions made from the responses of native promoters to the same stress.

We also provided evidence that subjecting *E. coli* to specific genome-wide stresses is an effective strategy to identify gene regulatory mechanisms. Specifically, it allows the partitioning of the genes into cohorts of responsive and non-responsive (Weber, Polen, Heuveling, Wendisch, & Hengge, 2005) or quickly and slowly responsive (Bhatia, Kirit, Predeus, & Bollback, 2022; Dash et al., 2022).

The differences in the dynamics of each cohort can then be confronted with known specificities of the component genes (e.g., sequence and TFs). For example, we studied a cohort of tandem promoters without neighboring genes. Meanwhile, there are cases where one of the promoters overlaps with a neighboring gene (Santos-Zavaleta et al., 2019). Comparing the mean behavior of the two cohorts should provide insight into the role of the overlapping. In general, provided enough natural cases, it should be possible to remove the influence from other variables.

Also noteworthy, in our study of a reduced set of tandem promoters, we observed differences in the dynamics of tandem promoters due to two variables, d_{TSS} and RNAP-promoter occupancy time. In other promoter arrangements, other factors are influential (Bordoy, Varanasi, Courtney, & Chatterjee, 2016; Eszterhas et al., 2002; Meyer & Beslon, 2014). A promising strategy for dissecting how these and other regulatory mechanisms act on transcription is the combined use of large-scale perturbations, to observe the behavior of the natural circuits. Then, having established hypotheses of what are the controls of the natural systems, we hypothesized that it should be possible to use synthetic biology to engineer sufficient constructs to test the hypotheses. Following these hypotheses, we designed and engineered synthetic genetic constructs controlled by tandem promoters differing in strength from their component promoters.

In the future, our research could be expanded by considering convergent and divergent arrangements, as well as additional variables, such as supercoiling buildup as a function of the location of the promoter in the DNA. For example, the current model does not account for interference from neighboring gene's elongation events, nor due to transcription factor bindings. The future models could be expanded by considering these factors. Further, small libraries of genetic constructs could be designed with different arrangements to support the findings using the models. Similarly, we could study the effects of supercoiling, global regulators, temperature, and antibiotics. These constructs could, later on, become valuable components of future synthetic gene circuits.

Overall, as the dissection of GRN structures expands to more organisms, it should be possible to use strategies as the ones applied in this thesis and study their genome-wide, transcriptional response programs to stresses. This could provide better insights into how such complex programs evolved in bacteria. Such programs are likely responsible for capabilities that we are yet to decipher how they are achieved, such as persistence to antibiotics.

7 CONCLUSIONS

When I started my Ph.D. studies, transcription dynamics in *E. coli* was being studied using MS2-GFP tagging of RNA molecules. That led to several developments, such as the characterization of time intervals between RNA production events at the single-cell level. However, this was greatly limited by the small number of cells that could be followed and then segmented from microscopy images.

To overcome this limitation, I worked on the development of a method to measure the numbers of tagged RNAs in individual cells (**Publication I**). This was found to be partially possible. Particularly, while we discovered that we could not easily quantify the numbers of RNAs in individual cells, surprisingly we were able to match single-cell distributions of RNA numbers in individual cells. This finding makes it possible to replace microscopy with flow-cytometry as a means to quantify single-cell distributions of RNA numbers. This has the highly significant advantage of allowing the use of hundreds of thousands of cells to characterize transcription, instead of a few hundred cells due to the labor intensiveness of conducting microscopy and image analysis. Potentially, if combined with microfluidics, this may revive the use of MS2-GFP or FISH in the future to characterize transcription in bacteria.

Perhaps more interesting, due to this finding, we could characterize transcription by observing the levels of fast-maturing proteins by flow-cytometry. Specifically, since the first images using electron microscopy of transcription and translation in bacteria, it is well known that these two processes are mechanically and thereby dynamically linked and highly correlated, respectively. However, protein maturation times and events in cell division, among others, are expected to be sufficient to "de-correlate" RNA and protein numbers significantly. However, if the proteins maturated rapidly and efficiently, the correlation could remain high. In agreement, I observed that while it is not possible to match the RNA and protein numbers of individual cells with accuracy, the single-cell distributions of RNA and fluorescent proteins the RNA codes for are very well correlated. Thus, in subsequent studies, I made use of single-cell distributions of fast-maturing YFP proteins to study transcription (Taniguchi et al., 2010).

Subsequently, because of the findings above, the next main study used YFP fluorescent tags in the chromosome as the main method. Specifically, we observed single-cell distributions of YFP levels, from strains where these tags are attached to pairs of promoters in tandem formation. We focused on the phenomenon of interference between closely spaced promoters in tandem formation as a function of the distances between the promoter sequences and their strengths.

Prior to this thesis, the behavior of tandem promoters had only been explored using synthetic constructs. Therefore, the knowledge of the behavior of natural genes controlled by such promoters was relatively speculative. Because we had already linked protein numbers with RNA numbers in single cells and because I was already experienced in using the YFP strain library, I set out to use it to the study tandem promoters as a regulatory mechanism in normal and in stress conditions (**Publication III**).

Our main conclusion was that in natural *E. coli* cells, promoters with tandem formation that are spaced by less than 35 bp are subject to a strong phenomenon of interference due to promoter occlusion by RNAP occupancy of one of the promoters. Moreover, the degree of occlusion can differ with the stress condition. Based on this, I hypothesize that closely spaced promoters in tandem formation are evolved forms of gene regulatory mechanisms in bacteria and can be used to reliably control synthetic circuits.

In our final work, we present a proof of concept of the finding about the natural tandem promoters (**Publication IV**). Shortly, our novel synthetic promoter arrangements exhibit the expected dynamics. I.e., they exhibit predictable interference, which influences their behavior in optimal conditions and also their responses to antibiotics targeting the core regulators of transcription and supercoiling. This suggests that transcription interference between closely spaced promoters could help engineering genes with novel dynamics.

Overall, these studies reached the aims listed in Chapter 3. They also allow setting aims for future studies. I expect that it will prove to be valuable to synthetic genetic engineering studies and the further exploration of the regulatory mechanisms of closely spaced promoters in convergent and divergent geometries. These future studies should consider the influence of other factors affecting transcription dynamics such as temperature, supercoiling, global regulators, transcription factors, and interference due to neighboring genes, among others.

The main value is likely derived from their dynamics being subject to regulation due to not only their proximity, but also by the timing of the open and closed complex formation as well as by the timing of binding of transcription factors, for example. Such transcription factors could be chosen e.g., based on our observations of the TFN response to RNAP (**Publication II**). Specifically, those observations may be of use to estimate the strength that TFs have on their output genes and, thus, select those that best fit the desired synthetic genetic circuits.

8 REFERENCES

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular Biology of the Cell.* Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK21054/

Alwine, J. C., Kemp, D. J., & Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5350–5354. doi:10.1073/pnas.74.12.5350

Artsimovitch, I., & Henkin, T. M. (2009). In vitro approaches to analysis of transcription termination. *Methods* , *47*(1), 37–43. doi:10.1016/j.ymeth.2008.10.006

Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., … Mori, H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, *2*, 2006.0008. doi:10.1038/msb4100050

Balakrishnan, R., Mori, M., Segota, I., Zhang, Z., Aebersold, R., Ludwig, C., & Hwa, T. (2022). Principles of gene regulation quantitatively connect DNA to RNA and proteins in bacteria. *Science*, *378*(6624), eabk2066. doi:10.1126/science.abk2066

Baptista, I. S. C., Kandavalli, V., Chauhan, V., Bahrudeen, M. N. M., Almeida, B. L. B., Palma, C. S. D., … Ribeiro, A. S. (2022). Sequence-dependent model of genes with dual σ factor preference. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1865*(3), 194812. doi:10.1016/j.bbagrm.2022.194812

Baptista, I. S. C., & Ribeiro, A. S. (2020). Stochastic models coupling gene expression and partitioning in cell division in Escherichia coli. *Bio Systems*, *193–194*(104154), 104154. doi:10.1016/j.biosystems.2020.104154

Barajas, C. (2022). *Modeling and controlling resource loading in bacterial genetic circuits* (Massachusetts Institute of Technology). Retrieved from https://dspace.mit.edu/handle/1721.1/147210?show=full

Battesti, A., Majdalani, N., & Gottesman, S. (2011). The RpoS-mediated general stress response in Escherichia coli. *Annual Review of Microbiology*, *65*, 189–213. doi:10.1146/annurev-micro-090110-102946

Bernardi, A., & Spahr, P. F. (1972). Nucleotide sequence at the binding site for coat protein on RNA of bacteriophage R17. *Proceedings of the National Academy of Sciences of the United States of America*, *69*(10), 3033–3037. doi:10.1073/pnas.69.10.3033

Bertrand, E., Chartrand, P., Schaefer, M., Shenoy, S. M., Singer, R. H., & Long, R. M. (1998). Localization of ASH1 mRNA particles in living yeast. *Molecular Cell*, *2*(4), 437–445. doi:10.1016/s1097-2765(00)80143-4

Bervoets, I., & Charlier, D. (2019). Diversity, versatility and complexity of bacterial gene regulation mechanisms: opportunities and drawbacks for applications in synthetic biology. *FEMS Microbiology Reviews*, *43*(3), 304–339. doi:10.1093/femsre/fuz001

Bhatia, R. P., Kirit, H. A., Predeus, A. V., & Bollback, J. P. (2022). Transcriptomic profiling of Escherichia coli K-12 in response to a compendium of stressors. *Scientific Reports*, *12*(1), 8788. doi:10.1038/s41598-022-12463-3

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., … Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science*, *277*(5331), 1453–1462. doi:10.1126/science.277.5331.1453

Bordoy, A. E., Varanasi, U. S., Courtney, C. M., & Chatterjee, A. (2016). Transcriptional Interference in Convergent Promoters as a Means for Tunable Gene Expression. *ACS Synthetic Biology*, *5*(12), 1331–1341. doi:10.1021/acssynbio.5b00223

Bratton, B. P., Mooney, R. A., & Weisshaar, J. C. (2011). Spatial distribution and diffusive motion of RNA polymerase in live Escherichia coli. *Journal of Bacteriology*, *193*(19), 5138–5146. doi:10.1128/JB.00198-11

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall, CRC.

Britton, R. A., Eichenberger, P., Gonzalez-Pastor, J. E., Fawcett, P., Monson, R., Losick, R., & Grossman, A. D. (2002). Genome-wide analysis of the stationary-phase sigma factor (sigma-H) regulon of Bacillus subtilis. *Journal of Bacteriology*, *184*(17), 4881–4890. doi:10.1128/JB.184.17.4881-4890.2002

Brooks, A. N., Reiss, D. J., Allard, A., Wu, W.-J., Salvanha, D. M., Plaisier, C. L., … Baliga, N. S. (2014). A system-level model for the microbial regulatory genome. *Molecular Systems Biology*, *10*(7), 740. doi:10.15252/msb.20145160

Brown, M., & Wittwer, C. (2000). Flow cytometry: principles and clinical applications in hematology. *Clinical Chemistry*, *46*(8 Pt 2), 1221–1229. doi:10.1093/clinchem/46.8.1221

Browning, D. F., & Busby, S. J. W. (2004). The regulation of bacterial transcription initiation. *Nature Reviews. Microbiology*, *2*(1), 57–65. doi:10.1038/nrmicro787

Browning, D. F., & Busby, S. J. W. (2016). Local and global regulation of transcription initiation in bacteria. *Nature Reviews. Microbiology*, *14*(10), 638–650. doi:10.1038/nrmicro.2016.103

Bury-Moné, S., & Sclavi, B. (2017). Stochasticity of gene expression as a motor of epigenetics in bacteria: from individual to collective behaviors. *Research in Microbiology*, *168*(6), 503–514. doi:10.1016/j.resmic.2017.03.009

Callen, B. P., Shearwin, K. E., & Egan, J. B. (2004). Transcriptional Interference between Convergent Promoters Caused by Elongation over the Promoter. *Molecular Cell*, *14*(5), 647–656. doi:10.1016/j.molcel.2004.05.010

Campbell, E. A., Korzheva, N., Mustaev, A., Murakami, K., Nair, S., Goldfarb, A., & Darst, S. A. (2001). Structural mechanism for rifampicin inhibition of bacterial rna polymerase. *Cell*, *104*(6), 901–912. doi:10.1016/s0092-8674(01)00286-0

Castellano-Muñoz, M., Peng, A. W., Salles, F. T., & Ricci, A. J. (2012). Swept field laser confocal microscopy for enhanced spatial and temporal resolution in live-cell imaging. *Microscopy and Microanalysis: The Official Journal of Microscopy Society of America, Microbeam Analysis Society, Microscopical Society of Canada*, *18*(4), 753–760. doi:10.1017/S1431927612000542

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W., & Prasher, D. C. (1994). Green fluorescent protein as a marker for gene expression. *Science*, *263*(5148), 802–805. doi:10.1126/science.8303295

Chamberlin, M. J. (1974). The Selectivity of Transcription. *Annual Review of Biochemistry*, *43*(1), 721–775. doi:10.1146/annurev.bi.43.070174.003445

Charlebois, D. A., Hauser, K., Marshall, S., & Balázsi, G. (2018). Multiscale effects of heating and cooling on genes and gene networks. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(45), E10797–E10806. doi:10.1073/pnas.1810858115

Chong, S., Chen, C., Ge, H., & Xie, X. S. (2014). Mechanism of transcriptional bursting in bacteria. *Cell*, *158*(2), 314–326. doi:10.1016/j.cell.2014.05.038

Chung, H. J., Bang, W., & Drake, M. A. (2006). Stress response of Escherichia coli. *Comprehensive Reviews in Food Science and Food Safety*, *5*(3), 52–64. doi:10.1111/j.1541-4337.2006.00002.x

Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., … Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, *5*(7), 613–619. doi:10.1038/nmeth.1223

Cormack, B. P., Valdivia, R. H., & Falkow, S. (1996). FACS-optimized mutants of the green fluorescent protein (GFP). *Gene*, *173*(1 Spec), 33–38. doi:10.1016/0378-1119(95)00685-0

Côté, J.-P., French, S., Gehrke, S. S., MacNair, C. R., Mangat, C. S., Bharat, A., & Brown, E. D. (2016). The Genome-Wide Interaction Network of Nutrient Stress Genes in Escherichia coli. *MBio*, *7*(6). doi:10.1128/mBio.01714-16

Craig, E. A., & Schlesinger, M. J. (1985). The Heat Shock Respons. *Critical Reviews in Biochemistry and Molecular Biology*, *18*(3), 239–280. doi:10.3109/10409238509085135

Cramer, P., Bushnell, D. A., & Kornberg, R. D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science*, *292*(5523), 1863–1876. doi:10.1126/science.1059493

Dash, S., Palma, C. S. D., Baptista, I. S. C., Almeida, B. L. B., Bahrudeen, M. N. M., Chauhan, V., … Ribeiro, A. S. (2022). Alteration of DNA supercoiling serves as a trigger of short-term cold shock repressed genes of E. coli. *Nucleic Acids Research*. doi:10.1093/nar/gkac643

Day, R. N., & Davidson, M. W. (2009). The fluorescent protein palette: tools for cellular imaging. *Chemical Society Reviews*, *38*(10), 2887–2921. doi:10.1039/b901966a

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *9*(1), 67–103. doi:10.1089/10665270252833208

deHaseth Pieter L., Zupancic Margaret L., & Record M. Thomas. (1998). RNA Polymerase-Promoter Interactions: the Comings and Goings of RNA Polymerase. *Journal of Bacteriology*, *180*(12), 3019–3025. doi:10.1128/JB.180.12.3019-3025.1998

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., … Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, *471*(7340), 602–607. doi:10.1038/nature09886

Dillon, S. C., & Dorman, C. J. (2010). Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nature Reviews. Microbiology*, *8*(3), 185–195. doi:10.1038/nrmicro2261

Djordjevic, M., & Bundschuh, R. (2008). Formation of the open complex by bacterial RNA polymerase--a quantitative model. *Biophysical Journal*, *94*(11), 4233–4248. doi:10.1529/biophysj.107.116970

Dove, S. L., Darst, S. A., & Hochschild, A. (2003). Region 4 of sigma as a target for transcription regulation. *Molecular Microbiology*, *48*(4), 863–874. doi:10.1046/j.1365-2958.2003.03467.x

Dudek, C.-A., & Jahn, D. (2021). PRODORIC: state-of-the-art database of prokaryotic gene regulation. *Nucleic Acids Research*, *50*(D1), D295–D302. doi:10.1093/nar/gkab1110

Ebright, R. H. (1993). Transcription activation at Class I CAP-dependent promoters. *Molecular Microbiology*, *8*(5), 797–802. doi:10.1111/j.1365-2958.1993.tb01626.x

Eldar, A., & Elowitz, M. B. (2010). Functional roles for noise in genetic circuits. *Nature*, *467*(7312), 167–173. doi:10.1038/nature09326

Eliasson, Å., Bernander, R., Dasgupta, S., & Nordström, K. (1992). Direct visualization of plasmid DNA in bacterial cells. *Molecular Microbiology*, *6*(2), 165–170. doi:10.1111/j.1365-2958.1992.tb01997.x

Elliott, A. D. (2020). Confocal Microscopy: Principles and Modern Practices. *Current Protocols in Cytometry / Editorial Board, J. Paul Robinson, Managing Editor ... [et Al.]*, *92*(1), e68. doi:10.1002/cpcy.68

Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science (New York, N.Y.)*, *297*(5584), 1183–1186. doi:10.1126/science.1070919

Emrich, S. J., Barbazuk, W. B., Li, L., & Schnable, P. S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research*, *17*(1), 69–73. doi:10.1101/gr.5145806

Endesfelder, U. (2019). From single bacterial cell imaging towards in vivo single-molecule biochemistry studies. *Essays in Biochemistry*, *63*(2), 187–196. doi:10.1042/EBC20190002

Epshtein, V., & Nudler, E. (2003). Cooperation between RNA polymerase molecules in transcription elongation. *Science*, *300*(5620), 801–805. doi:10.1126/science.1083219

Escherich, T. (1988). The intestinal bacteria of the neonate and breast-fed infant. 1884. *Reviews of Infectious Diseases*, *10*(6), 1220–1225. doi:10.1093/clinids/10.6.1220

Eszterhas, S. K., Bouhassira, E. E., Martin, D. I. K., & Fiering, S. (2002). Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position. *Molecular and Cellular Biology*, *22*(2), 469–479. doi:10.1128/MCB.22.2.469-479.2002

Fan, J., El Sayyed, H., Pambos, O. J., Stracy, M., Kyropoulos, J., & Kapanidis, A. N. (2023). RNA polymerase redistribution supports growth in E. coli strains with a minimal number of rRNA operons. *Nucleic Acids Research*. doi:10.1093/nar/gkad511

Fang, X., Sastry, A., Mih, N., Kim, D., Tan, J., Yurkovich, J. T., … Palsson, B. O. (2017). Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(38), 10286–10291. doi:10.1073/pnas.1702581114

Feklístov, A., Sharon, B. D., Darst, S. A., & Gross, C. A. (2014). Bacterial sigma factors: a historical, structural, and genomic perspective. *Annual Review of Microbiology*, *68*, 357–376. doi:10.1146/annurev-micro-092412-155737

Fujita, K., Iwaki, M., & Yanagida, T. (2016). Transcriptional bursting is intrinsically caused by interplay between RNA polymerases on DNA. *Nature Communications*, *7*, 13788. doi:10.1038/ncomms13788

Fusco, D., Accornero, N., Lavoie, B., Shenoy, S. M., Blanchard, J.-M., Singer, R. H., & Bertrand, E. (2003). Single mRNA molecules demonstrate probabilistic movement in living mammalian cells. *Current Biology: CB*, *13*(2), 161–167. doi:10.1016/s0960-9822(02)01436-7

Galbusera, L., Bellement-Theroue, G., Urchueguia, A., Julou, T., & van Nimwegen, E. (2020). Using fluorescence flow cytometry data for single-cell gene expression analysis in bacteria. *PloS One*, *15*(10), e0240233. doi:10.1371/journal.pone.0240233

Garcia, H. G., & Phillips, R. (2011). Quantitative dissection of the simple repression input–output function. *Proceedings of the National Academy of Sciences*, *108*(29), 12173–12178. doi:10.1073/pnas.1015616108

Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., 3rd, & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, *6*(5), 343–345. doi:10.1038/nmeth.1318

Golding, I., Paulsson, J., Zawilski, S. M., & Cox, E. C. (2005). Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell*, *123*(6), 1025–1036. doi:10.1016/j.cell.2005.09.031

Goodsell, D. S. (2012). Putting proteins in context: scientific illustrations bring together information from diverse sources to provide an integrative view of the molecular biology of cells. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *34*(9), 718–720. doi:10.1002/bies.201200072

Greive, S. J., & von Hippel, P. H. (2005). Thinking quantitatively about transcriptional regulation. *Nature Reviews. Molecular Cell Biology*, *6*(3), 221–232. doi:10.1038/nrm1588

Häkkinen, A., Muthukrishnan, A.-B., Mora, A., Fonseca, J. M., & Ribeiro, A. S. (2013). CellAging: a tool to study segregation and partitioning in division in cell lineages of Escherichia coli. *Bioinformatics* , *29*(13), 1708–1709. doi:10.1093/bioinformatics/btt194

Häkkinen, A., Oliveira, S. M. D., Neeli-Venkata, R., & Ribeiro, A. S. (2019). Transcription closed and open complex formation coordinate expression of genes with a shared promoter region. *Journal of the Royal Society, Interface / the Royal Society*, *16*(161), 20190507. doi:10.1098/rsif.2019.0507

Häkkinen, A., & Ribeiro, A. S. (2014). Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data. *Bioinformatics* , *31*(1), 69–75. doi:10.1093/bioinformatics/btu592

Häkkinen, A., & Ribeiro, A. S. (2016). Characterizing rate limiting steps in transcription from RNA production times in live cells. *Bioinformatics* , *32*(9), 1346–1352. doi:10.1093/bioinformatics/btv744

Han, Y., Gao, S., Muegge, K., Zhang, W., & Zhou, B. (2015). Advanced Applications of RNA Sequencing and Challenges. *Bioinformatics and Biology Insights*, *9*(Suppl 1), 29–46. doi:10.4137/BBI.S28991

Harden, T. T., Herlambang, K. S., Chamberlain, M., Lalanne, J.-B., Wells, C. D., Li, G.-W., … Gelles, J. (2020). Alternative transcription cycle for bacterial RNA polymerase. *Nature Communications*, *11*(1), 448. doi:10.1038/s41467-019-14208-9

Haugen, S. P., Ross, W., & Gourse, R. L. (2008). Advances in bacterial promoter recognition and its control by factors that do not bind DNA. *Nature Reviews. Microbiology*, *6*(7), 507–519. doi:10.1038/nrmicro1912

Hayashi-Takanaka, Y., Stasevich, T. J., Kurumizaka, H., Nozaki, N., & Kimura, H. (2014). Evaluation of chemical fluorescent dyes as a protein conjugation partner for live cell imaging. *PloS One*, *9*(9), e106271. doi:10.1371/journal.pone.0106271

Hengge-Aronis, R. (2002a). Recent insights into the general stress response regulatory network in Escherichia coli. *Journal of Molecular Microbiology and Biotechnology*, *4*(3), 341–346. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11931567

Hengge-Aronis, R. (2002b). Stationary phase gene regulation: what makes an Escherichia coli promoter sigmaS-selective? *Current Opinion in Microbiology*, *5*(6), 591–595. doi:10.1016/s1369-5274(02)00372-7

Herbert, K. M., La Porta, A., Wong, B. J., Mooney, R. A., Neuman, K. C., Landick, R., & Block, S. M. (2006). Sequence-resolved detection of pausing by single RNA polymerase molecules. *Cell*, *125*(6), 1083–1094. doi:10.1016/j.cell.2006.04.032

Hochschild, A., & Dove, S. L. (1998). Protein-protein contacts that activate and repress prokaryotic transcription. *Cell*, *92*(5), 597–600. doi:10.1016/s0092-8674(00)81126-5

Hoffmann, S. A., Hao, N., Shearwin, K. E., & Arndt, K. M. (2019). Characterizing Transcriptional Interference between Converging Genes in Bacteria. *ACS Synthetic Biology*, *8*(3), 466–473. doi:10.1021/acssynbio.8b00477

Hufnagel, D. A., Depas, W. H., & Chapman, M. R. (2015). The Biology of the Escherichia coli Extracellular Matrix. *Microbiology Spectrum*, *3*(3). doi:10.1128/microbiolspec.MB-0014-2014

Ishihama, A. (2000). Functional Modulation of Escherichia Coli RNA Polymerase. *Annual Review of Microbiology*, *54*(1), 499–518. doi:10.1146/annurev.micro.54.1.499

Jagger, J. (1976). Effects of near-ultraviolet radiation on microorganisms. *Photochemistry and Photobiology*, *23*(6), 451–454. doi:10.1111/j.1751-1097.1976.tb07279.x

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, *337*(6096), 816–821. doi:10.1126/science.1225829

John, J., Jabbar, J., Badjatia, N., Rossi, M. J., Lai, W. K. M., & Pugh, B. F. (2022). Genome-wide promoter assembly in E. coli measured at single-base resolution. *Genome Research*, *32*(5), 878–892. doi:10.1101/gr.276544.121

Johnstone, C. P., & Galloway, K. E. (2022). Supercoiling-mediated feedback rapidly couples and tunes transcription. *Cell Reports*, *41*(3), 111492. doi:10.1016/j.celrep.2022.111492

Jones, D. L., Brewster, R. C., & Phillips, R. (2014). Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, *346*(6216), 1533–1536. doi:10.1126/science.1255301

Jozefczuk, S., Klie, S., Catchpole, G., Szymanski, J., Cuadros-Inostroza, A., Steinhauser, D., … Willmitzer, L. (2010). Metabolomic and transcriptomic stress response of Escherichia coli. *Molecular Systems Biology*, *6*, 364. doi:10.1038/msb.2010.18

Kaern, M., Elston, T. C., Blake, W. J., & Collins, J. J. (2005). Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews. Genetics*, *6*(6), 451–464. doi:10.1038/nrg1615

Kandavalli, V. K., Tran, H., & Ribeiro, A. S. (2016). Effects of σ factor competition are promoter initiation kinetics dependent. *Biochimica et Biophysica Acta*, *1859*(10), 1281–1288. doi:10.1016/j.bbagrm.2016.07.011

Karpen, M. E., & deHaseth, P. L. (2015). Base flipping in open complex formation at bacterial promoters. *Biomolecules*, *5*(2), 668–678. doi:10.3390/biom5020668

Kiermer, V. (2007). The dawn of recombinant DNA. *Nature Reviews. Genetics*, *8*(1), S6–S6. doi:10.1038/nrg2246

Korzheva, N., Mustaev, A., Kozlov, M., Malhotra, A., Nikiforov, V., Goldfarb, A., & Darst, S. A. (2000). A structural model of transcription elongation. *Science*, *289*(5479), 619–625. doi:10.1126/science.289.5479.619

Kramer, G. F., & Ames, B. N. (1987). Oxidative mechanisms of toxicity of low-intensity near-UV light in Salmonella typhimurium. *Journal of Bacteriology*, *169*(5), 2259–2266. doi:10.1128/jb.169.5.2259-2266.1987

Kremers, G.-J., Gilbert, S. G., Cranfill, P. J., Davidson, M. W., & Piston, D. W. (2011). Fluorescent proteins at a glance. *Journal of Cell Science*, *124*(Pt 2), 157–160. doi:10.1242/jcs.072744

Kurepina, N., Chudaev, M., Kreiswirth, B. N., Nikiforov, V., & Mustaev, A. (2022). Mutations compensating for the fitness cost of rifampicin resistance in Escherichia coli exert pleiotropic effect on RNA polymerase catalysis. *Nucleic Acids Research*, *50*(10), 5739–5756. doi:10.1093/nar/gkac406

Kussell, E., & Leibler, S. (2005). Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, *309*(5743), 2075–2078. doi:10.1126/science.1114383

Lambert, T. J. (2019). FPbase: a community-editable fluorescent protein database. *Nature Methods*, *16*(4), 277–278. doi:10.1038/s41592-019-0352-8

Larsen, S. J., Röttger, R., Schmidt, H. H. H. W., & Baumbach, J. (2019). E. coli gene regulatory networks are inconsistent with gene expression data. *Nucleic Acids Research*, *47*(1), 85–92. doi:10.1093/nar/gky1176

Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A., & Singer, R. H. (2011). Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science*, *332*(6028), 475–478. doi:10.1126/science.1202142

Lee, J. W., Gyorgy, A., Cameron, D. E., Pyenson, N., Choi, K. R., Way, J. C., … Collins, J. J. (2016). Creating Single-Copy Genetic Circuits. *Molecular Cell*, *63*(2), 329–336. doi:10.1016/j.molcel.2016.06.006

Lenstra, T. L., Rodriguez, J., Chen, H., & Larson, D. R. (2016). Transcription Dynamics in Living Cells. *Annual Review of Biophysics*, *45*(1), 25–47. doi:10.1146/annurev-biophys-062215-010838

Lewin, B. (2008). *Genes IX* (9th ed). Sudbury, Mass: Jones and Bartlett Publishers.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, *133*(3), 523–536. doi:10.1016/j.cell.2008.03.029

Lloyd-price, J., Kandavalli, V., Chandraseelan, J. G., Goncalves, N., Oliveira, S. M. D., Häkkinen, A., & Ribeiro, A. S. (2016). Dissecting the stochastic transcription initiation process in live Escherichia coli. *DNA Research*, *23*(March), 203–214. doi:10.1093/dnares/dsw009

Lloyd-Price, J., Startceva, S., Kandavalli, V., Chandraseelan, J. G., Goncalves, N., Oliveira, S. M. D., … Ribeiro, A. S. (2016). Dissecting the stochastic transcription initiation process in live Escherichia coli. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *23*(3), 203–214. doi:10.1093/dnares/dsw009

Lukačišinová, M., Fernando, B., & Bollenbach, T. (2020). Highly parallel lab evolution reveals that epistasis can curb the evolution of antibiotic resistance. *Nature Communications*, *11*(1), 3105. doi:10.1038/s41467-020-16932-z

Lutz, R., & Bujard, H. (1997). Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research*, *25*(6), 1203–1210. doi:10.1093/nar/25.6.1203

Lutz, Rolf, Lozinski, T., Ellinger, T., & Bujard, H. (2001). Dissecting the functional program of Escherichia coli promoters: the combined mode of action of Lac repressor and AraC activator. *Nucleic Acids Research*, *29*(18), 3873–3881. doi:10.1093/nar/29.18.3873

Madan Babu, M., & Teichmann, S. A. (2003). Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Research*, *31*(4), 1234–1244. doi:10.1093/nar/gkg210

Mäkelä, J., Kandhavelu, M., Oliveira, S. M. D., Chandraseelan, J. G., Lloyd-Price, J., Peltonen, J., … Ribeiro, A. S. (2013). In vivo single-molecule kinetics of activation and subsequent activity of the arabinose promoter. *Nucleic Acids Research*, *41*(13), 6544–6552. doi:10.1093/nar/gkt350

Martin, F. H., & Tinoco, I. (1980). DNA-RNA hybrid duplexes containing oligo(dA:rU) sequences are exceptionally unstable and may facilitate termination of transcription. *Nucleic Acids Research*, *8*(10), 2295–2300. doi:10.1093/nar/8.10.2295

Martínez-Antonio, A., & Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology*, *6*(5), 482–489. doi:10.1016/j.mib.2003.09.002

Martins, L., Mäkelä, J., Häkkinen, A., Kandhavelu, M., Yli-Harja, O., Fonseca, J. M., & Ribeiro, A. S. (2012). Dynamics of transcription of closely spaced promoters in Escherichia coli, one event at a time. *Journal of Theoretical Biology*, *301*, 83–94. doi:10.1016/j.jtbi.2012.02.015

McClure, W. R. (1985). Mechanism and control of transcription initiation in prokaryotes. *Annual Review of Biochemistry*, *54*, 171–204. doi:10.1146/annurev.bi.54.070185.001131

McLeod, S. M., & Johnson, R. C. (2001). Control of transcription by nucleoid proteins. *Current Opinion in Microbiology*, *4*(2), 152–159. doi:10.1016/s1369-5274(00)00181-8

Meyer, S., & Beslon, G. (2014). Torsion-mediated interaction between adjacent genes. *PLoS Computational Biology*, *10*(9), e1003785. doi:10.1371/journal.pcbi.1003785

Mitosch, K., Rieckh, G., & Bollenbach, T. (2019). Temporal order and precision of complex stress responses in individual bacteria. *Molecular Systems Biology*, *15*(2), 1–15. doi:10.15252/msb.20188470

Mora, A. D., Vieira, P. M., Manivannan, A., & Fonseca, J. M. (2011). Automated drusen detection in retinal images using analytical modelling algorithms. *Biomedical Engineering Online*, *10*(1), 59. doi:10.1186/1475-925X-10-59

Morise, H., Shimomura, O., Johnson, F. H., & Winant, J. (1974). Intermolecular energy transfer in the bioluminescent system of Aequorea. *Biochemistry*, *13*(12), 2656–2662. doi:10.1021/bi00709a028

Müller, J., Oehler, S., & Müller-Hill, B. (1996). Repression of lac promoter as a function of distance, phase and quality of an auxiliary lac operator. *Journal of Molecular Biology*, *257*(1), 21–29. doi:10.1006/jmbi.1996.0143

Munsky, B., & Khammash, M. (2006). The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, *124*(4), 044104. doi:10.1063/1.2145882

Murakami, K. S., & Darst, S. A. (2003). Bacterial RNA polymerases: the wholo story. *Current Opinion in Structural Biology*, *13*(1), 31–39. doi:10.1016/s0959-440x(02)00005-2

Murphy, D. B., Oldfield, R., Schwartz, S., & Davidson, M. W. (n.d.). Introduction to Phase Contrast Microscopy | MicroscopyU. Retrieved from MicroscopyU website: https://www.microscopyu.com/techniques/phase-contrast/introduction-to-phase-contrast-microscopy

Muthukrishnan, A.-B., Kandhavelu, M., Lloyd-Price, J., Kudasov, F., Chowdhury, S., Yli-Harja, O., & Ribeiro, A. S. (2012). Dynamics of transcription driven by the tetA promoter, one event at a time, in live Escherichia coli cells. *Nucleic Acids Research*, *40*(17), 8472–8483. doi:10.1093/nar/gks583

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, *320*(5881), 1344–1349. doi:10.1126/science.1158441

Nakano, A. (2002). Spinning-disk confocal microscopy -- a cutting-edge tool for imaging of membrane traffic. *Cell Structure and Function*, *27*(5), 349–355. doi:10.1247/csf.27.349

Navarre, W. W., Porwollik, S., Wang, Y., McClelland, M., Rosen, H., Libby, S. J., & Fang, F. C. (2006). Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. *Science*, *313*(5784), 236–238. doi:10.1126/science.1128794

Neidhardt, F. C., VanBogelen, R. A., & Lau, E. T. (1983). Molecular cloning and expression of a gene that controls the high-temperature regulon of Escherichia coli. *Journal of Bacteriology*, *153*(2), 597–603. doi:10.1128/jb.153.2.597-603.1983

Oehler, S., Eismann, E. R., Krämer, H., & Müller-Hill, B. (1990). The three operators of the lac operon cooperate in repression. *The EMBO Journal*, *9*(4), 973–979. doi:10.1002/j.1460-2075.1990.tb08199.x

Palade, G. E. (1955). A small particulate component of the cytoplasm. *The Journal of Biophysical and Biochemical Cytology*, *1*(1), 59–68. doi:10.1083/jcb.1.1.59

Palma, C. S. D., Kandavalli, V., Bahrudeen, M. N. M., Minoia, M., Chauhan, V., Dash, S., & Ribeiro, A. S. (2020). Dissecting the in vivo dynamics of transcription locking due to positive supercoiling buildup. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1863*(5), 194515. doi:10.1016/j.bbagrm.2020.194515

Patange, O., Schwall, C., Jones, M., Villava, C., Griffith, D. A., Phillips, A., & Locke, J. C. W. (2018). Escherichia coli can survive stress by noisy growth modulation. *Nature Communications*, *9*(1), 5333. doi:10.1038/s41467-018-07702-z

Patterson, G. H., Knobel, S. M., Sharif, W. D., Kain, S. R., & Piston, D. W. (1997). Use of the green fluorescent protein and its mutants in quantitative fluorescence microscopy. *Biophysical Journal*, *73*(5), 2782–2790. doi:10.1016/S0006-3495(97)78307-3

Pawley, J. B. (2022). *Handbook of Biological Confocal Microscopy* (p. 28). doi:10.1007/978-0-387-45524-2

Peabody, D. S. (1993). The RNA binding site of bacteriophage MS2 coat protein. *The EMBO Journal*, *12*(2), 595–600. doi:10.1002/j.1460-2075.1993.tb05691.x

Peccoud, J., & Ycart, B. (1995). Markovian Modeling of Gene-Product Synthesis. *Theoretical Population Biology*, *48*(2), 222–234. doi:10.1006/tpbi.1995.1027

Phadtare, S., & Inouye, M. (2004). Genome-wide transcriptional analysis of the cold shock response in wild-type and cold-sensitive, quadruple-csp-deletion strains of Escherichia coli. *Journal of Bacteriology*, *186*(20), 7007–7014. doi:10.1128/JB.186.20.7007-7014.2004

Ponnambalam, S., & Busby, S. (1987). RNA polymerase molecules initiating transcription at tandem promoters can collide and cause premature transcription termination. *FEBS Letters*, *212*(1), 21–27. doi:10.1016/0014-5793(87)81549-1

Querido, E., & Chartrand, P. (2008). Using fluorescent proteins to study mRNA trafficking in living cells. *Methods in Cell Biology*, *85*, 273–292. doi:10.1016/S0091-679X(08)85012-1

Ramakrishnan, V. (2002). Ribosome structure and the mechanism of translation. *Cell*, *108*(4), 557–572. doi:10.1016/s0092-8674(02)00619-0

Razo-Mejia, M., Barnes, S. L., Belliveau, N. M., Chure, G., Einav, T., Lewis, M., & Phillips, R. (2018). Tuning Transcriptional Regulation through Signaling: A

Predictive Theory of Allosteric Induction. *Cell Systems*, *6*(4), 456-469.e10. doi:10.1016/j.cels.2018.02.004

Ribeiro, A. S. (2010a). Mathematical Biosciences Stochastic and delayed stochastic models of gene expression and regulation. *Mathematical Biosciences*, *223*(1), 1–11. doi:10.1016/j.mbs.2009.10.007

Ribeiro, A. S. (2010b). Stochastic and delayed stochastic models of gene expression and regulation. *Mathematical Biosciences*, *223*(1), 1–11. doi:10.1016/j.mbs.2009.10.007

Richter, K., Haslbeck, M., & Buchner, J. (2010). The heat shock response: life on the verge of death. *Molecular Cell*, *40*(2), 253–266. doi:10.1016/j.molcel.2010.10.006

Ross, W., Gosink, K. K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., … Gourse, R. L. (1993). A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science*, *262*(5138), 1407–1413. doi:10.1126/science.8248780

Ross, Wilma, Vrentas, C. E., Sanchez-Vazquez, P., Gaal, T., & Gourse, R. L. (2013). The magic spot: a ppGpp binding site on E. coli RNA polymerase responsible for regulation of transcription initiation. *Molecular Cell*, *50*(3), 420–429. doi:10.1016/j.molcel.2013.03.021

Russo, T. A., & Johnson, J. R. (2003). Medical and economic impact of extraintestinal infections due to Escherichia coli: focus on an increasingly important endemic problem. *Microbes and Infection / Institut Pasteur*, *5*(5), 449–456. doi:10.1016/s1286-4579(03)00049-2

Saecker, R. M., Record, M. T., Jr, & Dehaseth, P. L. (2011). Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. *Journal of Molecular Biology*, *412*(5), 754–771. doi:10.1016/j.jmb.2011.01.018

Sanchez-Vazquez, P., Dewey, C. N., Kitten, N., Ross, W., & Gourse, R. L. (2019). Genome-wide effects on Escherichia coli transcription from ppGpp binding to its two sites on RNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(17), 8310–8319. doi:10.1073/pnas.1819682116

Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., … Collado-Vides, J. (2019). RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Research*, *47*(D1), D212–D220. doi:10.1093/nar/gky1077

Schneider, A. F. L., & Hackenberger, C. P. R. (2017). Fluorescent labelling in living cells. *Current Opinion in Biotechnology*, *48*, 61–68. doi:10.1016/j.copbio.2017.03.012

Shaner, N. C., Campbell, R. E., Steinbach, P. A., Giepmans, B. N. G., Palmer, A. E., & Tsien, R. Y. (2004). Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nature Biotechnology*, *22*(12), 1567–1572. doi:10.1038/nbt1037

Shearwin, K. E., Callen, B. P., & Egan, J. B. (2005). Transcriptional interference--a crash course. *Trends in Genetics: TIG*, *21*(6), 339–345. doi:10.1016/j.tig.2005.04.009

Shen, L. L., Mitscher, L. A., Sharma, P. N., O'Donnell, T. J., Chu, D. W., Cooper, C. S., … Pernet, A. G. (1989). Mechanism of inhibition of DNA gyrase by quinolone antibacterials: a cooperative drug--DNA binding model. *Biochemistry*, *28*(9), 3886–3894. doi:10.1021/bi00435a039

Shimomura, O., Johnson, F. H., & Saiga, Y. (1962). Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea. *Journal of Cellular and Comparative Physiology*, *59*(3), 223–239. doi:10.1002/jcp.1030590302

Smith, B. R., & Schleif, R. (1978). Nucleotide sequence of the L-arabinose regulatory region of Escherichia coli K12. *The Journal of Biological Chemistry*, *253*(19), 6931–6933. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/357433

Smith, H. O., & Wilcox, K. W. (1970). A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *Journal of Molecular Biology*, *51*(2), 379–391. doi:10.1016/0022-2836(70)90149-x

Sneppen, K., Dodd, I. B., Shearwin, K. E., Palmer, A. C., Schubert, R. A., Callen, B. P., & Egan, J. B. (2005). A mathematical model for transcriptional interference by RNA polymerase traffic in Escherichia coli. *Journal of Molecular Biology*, *346*(2), 399–409. doi:10.1016/j.jmb.2004.11.075

So, L.-H., Ghosh, A., Zong, C., Sepúlveda, L. A., Segev, R., & Golding, I. (2011). General properties of transcriptional time series in Escherichia coli. *Nature Genetics*, *43*(6), 554–560. doi:10.1038/ng.821

Song, E., Uhm, H., Munasingha, P. R., Hwang, S., Seo, Y.-S., Kang, J. Y., … Hohng, S. (2022). Rho-dependent transcription termination proceeds via three routes. *Nature Communications*, *13*(1), 1–12. doi:10.1038/s41467-022-29321-5

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews. Genetics*, *20*(11), 631–656. doi:10.1038/s41576-019-0150-2

Steen, H. B. (1992). Noise, sensitivity, and resolution of flow cytometers. *Cytometry*, *13*(8), 822–830. doi:10.1002/cyto.990130804

Stephens, D. J., & Allan, V. J. (2003). Light microscopy techniques for live cell imaging. *Science*, *300*(5616), 82–86. doi:10.1126/science.1082160

Stoebel, D. M., Hokamp, K., Last, M. S., & Dorman, C. J. (2009). Compensatory evolution of gene regulation in response to stress by Escherichia coli lacking RpoS. *PLoS Genetics*, *5*(10), e1000671. doi:10.1371/journal.pgen.1000671

Stracy, M., & Kapanidis, A. N. (2017). Single-molecule and super-resolution imaging of transcription in living bacteria. *Methods* , *120*, 103–114. doi:10.1016/j.ymeth.2017.04.001

Swint-Kruse, L., & Matthews, K. S. (2009). Allostery in the LacI/GalR family: variations on a theme. *Current Opinion in Microbiology*, *12*(2), 129–137. doi:10.1016/j.mib.2009.01.009

Tagami, S., Sekine, S.-I., & Yokoyama, S. (2011). A novel conformation of RNA polymerase sheds light on the mechanism of transcription. *Transcription*, *2*(4), 162–167. doi:10.4161/trns.2.4.16148

Taniguchi, Y., Choi, P. J., Li, G.-W., Chen, H., Babu, M., Hearn, J., … Xie, X. S. (2010). Quantifying *E. coli* Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science*, *329*(5991), 533–538. doi:10.1126/science.1188308

Todd, P. A., & Faulds, D. (1991). Ofloxacin. A reappraisal of its antimicrobial activity, pharmacology and therapeutic use. *Drugs*, *42*(5), 825–876. doi:10.2165/00003495-199142050-00008

Tran, H., Oliveira, S. M. D., Goncalves, N., & Ribeiro, A. S. (2015). Kinetics of the cellular intake of a gene expression inducer at high concentrations. *Molecular BioSystems*, *11*(9), 2579–2587. doi:10.1039/C5MB00244C

Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P., & Myers, R. M. (2004). An abundance of bidirectional promoters in the human genome. *Genome Research*, *14*(1), 62–66. doi:10.1101/gr.1982804

Tripathi, L., Zhang, Y., & Lin, Z. (2014). Bacterial sigma factors as targets for engineered or synthetic transcriptional control. *Frontiers in Bioengineering and Biotechnology*, *2*, 33. doi:10.3389/fbioe.2014.00033

Turkowyd, B., Balinovic, A., Virant, D., Carnero, H. G. G., Caldana, F., Endesfelder, M., … Endesfelder, U. (2017). A General Mechanism of Photoconversion of Green-to-Red Fluorescent Proteins Based on Blue and Infrared Light Reduces Phototoxicity in Live-Cell Single-Molecule Imaging. *Angewandte Chemie* , *56*(38), 11634–11639. doi:10.1002/anie.201702870

Ullmann, A. (2011). Escherichia coli and the Emergence of Molecular Biology. *EcoSal Plus*, *4*(2). doi:10.1128/ecosalplus.1.1.2

Urchueguía, A., Galbusera, L., Chauvin, D., Bellement, G., Julou, T., & van Nimwegen, E. (2021). Genome-wide gene expression noise in Escherichia coli is condition-dependent and determined by propagation of noise through the regulatory network. *PLoS Biology*, *19*(12), e3001491. doi:10.1371/journal.pbio.3001491

Van Brempt, M., Clauwaert, J., Mey, F., Stock, M., Maertens, J., Waegeman, W., & De Mey, M. (2020). Predictive design of sigma factor-specific promoters. *Nature Communications*, *11*(1), 5822. doi:10.1038/s41467-020-19446-w

Vinograd, J., Lebowitz, J., Radloff, R., Watson, R., & Laipis, P. (1965). The twisted circular form of polyoma viral DNA. *Proceedings of the National Academy of Sciences*, *53*(5), 1104–1111. doi:10.1073/pnas.53.5.1104

Volkmer, B., & Heinemann, M. (2011). Condition-dependent cell volume and concentration of Escherichia coli to facilitate data conversion for systems biology modeling. *PloS One*, *6*(7), e23126. doi:10.1371/journal.pone.0023126

Walter, G., Zillig, W., Palm, P., & Fuchs, E. (1967). Initiation of DNA-Dependent RNA Synthesis and the Effect of Heparin on RNA Polymerase. *European Journal of Biochemistry / FEBS*, *3*(2), 194–201. doi:10.1111/j.1432-1033.1967.tb19515.x

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. doi:10.1038/nrg2484

Ward, W. W., Cody, C. W., Hart, R. C., & Cormier, M. J. (1980). Spectrophotometric identity of the energy transfer chromophores in renilla and aequorea green-fluorescent proteins. *Photochemistry and Photobiology*, *31*(6), 611–615. doi:10.1111/j.1751-1097.1980.tb03755.x

Weber, H., Polen, T., Heuveling, J., Wendisch, V. F., & Hengge, R. (2005). Genome-wide analysis of the general stress response network in Escherichia coli: sigmaS-dependent genes, promoters, and sigma factor selectivity. *Journal of Bacteriology*, *187*(5), 1591–1603. doi:10.1128/JB.187.5.1591-1603.2005

Weiss, A., Moore, B. D., Tremblay, M. H. J., Chaput, D., Kremer, A., & Shaw, L. N. (2017). The ω Subunit Governs RNA Polymerase Stability and Transcriptional Specificity in Staphylococcus aureus. *Journal of Bacteriology*, *199*(2). doi:10.1128/JB.00459-16

Wilson, K. S., & von Hippel, P. H. (1994). Stability of Escherichia coli transcription complexes near an intrinsic terminator. *Journal of Molecular Biology*, *244*(1), 36–51. doi:10.1006/jmbi.1994.1702

Yakhnin, A. V., Bubunenko, M., Mandell, Z. F., Lubkowska, L., Husher, S., Babitzke, P., & Kashlev, M. (2023). Robust regulation of transcription pausing in *Escherichia coli* by the ubiquitous elongation factor NusG. *Proceedings of the National Academy of Sciences*, *120*(24), e2221114120. doi:10.1073/pnas.2221114120

Yeung, E., Dy, A. J., Martin, K. B., Ng, A. H., Del Vecchio, D., Beck, J. L., … Murray, R. M. (2017). Biophysical Constraints Arising from Compositional Context in Synthetic Gene Networks. *Cell Systems*, *5*(1), 11-24.e12. doi:10.1016/j.cels.2017.06.001

Yu, J., Xiao, J., Ren, X., Lao, K., & Xie, X. S. (2006). Probing gene expression in live cells, one protein molecule at a time. *Science*, *311*(5767), 1600–1603. doi:10.1126/science.1119623

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

Zernike, F. (1942). Phase contrast, a new method for the microscopic observation of transparent objects. *Physica*, *9*(7), 686–698. doi:10.1016/S0031-8914(42)80035-X

PUBLICATIONS

PUBLICATION I

Estimating RNA numbers in single cells by RNA fluorescent tagging and flow cytometry

M.N.M. Bahrudeen*, Vatsala Chauhan*, Cristina S.D. Palma, Samuel M.D. Oliveira, Vinodh K. Kandavalli, Andre S. Ribeiro. *Equal contributions.

Journal of Microbiological Methods,166,105745,2019
DOI: 10.1016/j.mimet.2019.105745

Journal of Microbiological Methods 166 (2019) 105745

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Estimating RNA numbers in single cells by RNA fluorescent tagging and flow cytometry

Mohamed N.M. Bahrudeen[a,1], Vatsala Chauhan[a,1], Cristina S.D. Palma[a], Samuel M.D. Oliveira[a,b], Vinodh K. Kandavalli[a], Andre S. Ribeiro[a,*]

[a] *Laboratory of Biosystem Dynamics, BioMediTech, Faculty of Medicine and Health Technology, Tampere University, 33101 Tampere, Finland*
[b] *Department of Electrical and Computer Engineering, Center of Synthetic Biology, Boston University, Boston, USA*

ARTICLE INFO

Keywords:
Flow cytometry
Time-lapse microscopy
MS2d-GFP RNA tagging
Single-cell RNA numbers

ABSTRACT

Estimating the statistics of single-cell RNA numbers has become a key source of information on gene expression dynamics. One of the most informative methods of in vivo single-RNA detection is MS2d-GFP tagging. So far, it requires microscopy and laborious semi-manual image analysis, which hampers the amount of collectable data. To overcome this limitation, we present a new methodology for quantifying the mean, standard deviation, and skewness of single-cell distributions of RNA numbers, from flow cytometry data on cells expressing RNA tagged with MS2d-GFP. The quantification method, based on scaling flow-cytometry data from microscopy single-cell data on integer-valued RNA numbers, is shown to readily produce precise, big data on in vivo single-cell distributions of RNA numbers and, thus, can assist in studies of transcription dynamics.

1. Introduction

Single-cell imaging and fluorescent proteins have become a key source of information on multiple processes in live cells, particularly gene expression (Kærn et al., 2005). Originally, they have been used for, e.g., quantifying cell-to-cell diversity in protein levels (Elowitz et al., 2002; Ozbudak et al., 2002; Pedraza and Van, 2005; Engl, 2018). Subsequent progresses in microscopy and in the engineering of synthetic fluorescent proteins have allowed observing in vivo individual proteins (Yu et al., 2006) and RNA molecules (Fusco et al., 2003; Golding et al., 2005; Trcek et al., 2012; Femino et al., 1998; Raj et al., 2008). This made possible, among other, the quantification of the effects and the identification of sources of transcriptional bursting (Golding et al., 2005; Yu et al., 2006; Chong et al., 2014).

While there are several methods to quantify RNA, such as RT-qPCR (Saiki et al., 1985)(Higuchi et al., 1993), microarrays (Bumgarner, 2013), RNA seq (Tang et al., 2009), and UMI-based single-cell RNA-seq (Kivioja et al., 2012; Islam et al., 2014), among other, only a few can visualize individual RNAs, such as RNA Fluorescence In Situ Hybridization (Singer and Ward, 1982), RNA aptamers (Bunka and Stockley, 2006), and MS2-GFP RNA tagging (Golding et al., 2005). The latter allows observing individual RNAs using a synthetic protein, MS2d-GFP, and a synthetic target RNA, coding for multiple binding sites for the MS2d capsid protein (Peabody, 1993). Due to the rapid and stable binding of multiple MS2d-GFP proteins to the several binding sites in a single RNA, time-lapse imaging detects individual RNAs as these are produced. This facilitates the identification of sources of intrinsic noise in RNA production (Golding et al., 2005), the dissection of rate-limiting steps in active transcription (Lloyd-Price et al., 2016; Kandavalli et al., 2016), and the quantification of propensities for threshold crossing in RNA numbers (Startceva et al., 2019), among other.

The quantification of RNA by MS2d-GFP tagging is not free from measurement noise. For example, in time-lapse confocal microscopy, it is not uncommon that tagged RNAs (Supplementary Fig. S1) intermittently disappear. In addition, the precision of the estimation of the number of tagged RNAs within a given ‘RNA spot’ decreases rapidly with the number of RNAs in the spot (Golding et al., 2005; Häkkinen et al., 2014). Further, it is laborious to collect data, since even when using tailored, state-of-the-art software for segmenting the microscopy images (e.g. (Martins et al., 2018)), it usually still requires manual corrections and, in the worst cases, the necessary information can be absent from the images (e.g. an existing spot might not be captured in the image, e.g. if not within a given z-plane).

One solution to these problems would be to complement the microscopy data on single-cell numbers of MS2d-GFP tagged RNAs with

[*] Corresponding author at: Arvo Ylpön katu 34, P.O.Box 100, 33014, Tampere University, Finland.
E-mail address: andre.sanchesribeiro@tuni.fi (A.S. Ribeiro).
[1] Equal first authors.

https://doi.org/10.1016/j.mimet.2019.105745
Received 14 June 2019; Received in revised form 11 October 2019; Accepted 15 October 2019
Available online 22 October 2019

flow cytometry data. This would allow to rapidly collect much larger amounts of data, and also reduce significantly the uncertainty in the data (e.g. cells that are not entirely imaged can be automatically removed from the dataset, by using combined information from various channels of the flow cytometer, and RNA spots would always be entirely present in each imaged cell). However, flow cytometry lacks spatial information, which so far has been used in the quantification of MS2d-GFP tagged RNAs (Golding et al., 2005; Häkkinen et al., 2014).

Recent approaches have successfully combined Fluorescence in situ hybridization (FISH) for RNA counting with flow cytometry (see e.g. (Arrigucci et al., 2017; Bushkin et al., 2015; Tiberi et al., 2018) for similar aims. However, achieving the same using the MS2d-GFP tagging technique is expected to be more complex because, unlike when using FISH, not only the MS2d-GFP tagged RNAs are fluorescent but also the cells' cytoplasm, due to the need for large numbers of free floating MS2d-GFP to readily detect newly formed RNAs.

To address this, and since MS2d-GFP tagged RNA have been shown to have constant fluorescence for a few hours following their formation (Tran et al., 2015; Lloyd-Price et al., 2016; Oliveira et al., 2016), we hypothesized that cells with tagged RNAs have, on average, higher fluorescence than otherwise (since the MS2d-GFP proteins attached to the RNA are ‘immortalized’). As such, the total fluorescence of a cell should increase with the number of tagged RNAs that it accumulates. If so, it should be possible, from flow cytometry data on cells expressing MS2d-GFP tagged RNAs, to estimate the statistics of single-cell distributions of RNA numbers. Here we validate these hypotheses and show that flow cytometry data can be used to extract the mean, standard deviation, and skewness of single-cell distributions of RNA numbers that match those observed using microscopy.

2. Materials and methods

2.1. Chemicals

Measurements were performed in Luria-Bertani (LB) medium. The chemicals were: Tryptone and sodium chloride from Sigma Aldrich. Yeast extract was from Lab M (Topley House, Bury, Lancashire, UK). Antibiotics used are kanamycin and chloramphenicol, from Sigma-Aldrich. Inducers isopropyl β-D-1-thiogalactopyranoside (IPTG), anhydrotetracycline (aTc) and L-Arabinose (ara) were purchased from Sigma-Aldrich. For preparing microscopic gel pads we used agarose from Sigma-Aldrich.

2.2. Strains and plasmids

The *E. coli* strain used was DH5α-PRO, identical to DH5αZ1. Its genotype is deoR, endA1, gyrA96, hsdR17 (rK- mK+), recA1, relA1, supE44, thi-1, Δ(lacZYA-argF)U169, Φ80δlacZΔM15, F-, λ-, PN25/tetR, PlacIq/lacI and SpR. This strain produces the regulatory proteins required for tightly regulating the genetic constructs used (LacI, TetR and AraC).

The two genetic constructs used in this strain are: i) a multi-copy reporter plasmid responsible for producing MS2d-GFP proteins, controlled by the promoter $P_{LtetO-1}$, inducible by aTc; ii) A single-copy target F-plasmid is responsible for producing an RNA coding for mRFP1 up-stream of a 96 MS2 binding site array, controlled by the promoter $P_{Lac/ara-1}$, inducible by IPTG and L-Arabinose, ($P_{Lac/ara-1}$-mRFP1-96BS, Supplementary Fig. S2). We also used the *E. coli* DH5α-PRO strain carrying only the reporter plasmid. The plasmids were transferred into the host strain by standard molecular cloning techniques (Alberts et al., 2002).

The high number of binding sites for MS2d and the high affinity of each site with MS2d proteins cause each target RNA, when tagged, to appear as a bright ‘spot’ (Fig. 1B and Supplementary section 1.2), soon after being transcribed (in < 1 min) and to exhibit constant fluorescence intensity for a long period of time (mean half-lives of ~140 min (Tran et al., 2015)). Finally, it has been shown that, in these cells, the protein expression level of the target gene is not affected by MS2d-GFP tagging and follows the RNA numbers (Startceva et al., 2019).

2.3. Growth media and induction of the reporter and target genes

From a glycerol stock (−80 °C), cells were streaked on a LB agar plate and incubated at 37 °C overnight. From this plate, a single colony was picked and inoculated into a fresh LB medium supplemented with appropriate antibiotics (35 μg/ml kanamycin and 34 μg/ml chloramphenicol) and grown overnight at 30 °C with aeration. From the overnight cultures, cells were diluted into fresh LB medium to an optical density (OD_{600}) of 0.03, and grown at 37 °C, 250 rpm. Once the cells reach the OD_{600} 0.3, aTc (100 ng/ml) was added to induce $P_{LtetO-1}$ for MS2d-GFP production. L-Arabinose (0.1%) was also added, at the same time, for pre-activation of the target promoter $P_{Lac/ara-1}$. After 50 min, IPTG was added (0, 6.25, 50, 100, 200, 300, 500, or 1000 μM) to activate the production of the RNA target for MS2d-GFP. Following 1 h, cells were observed to quantify RNA and proteins (microscopy or flow cytometry).

2.4. Spectrophotometry

Fluorescence intensities were measured by using a BioTek Synergy HTX Multi-Mode Microplate Reader with Gen5 software. From the overnight culture, cells were diluted to 1:1000 times in fresh LB medium and incubated at 37 °C with shaking, until an OD_{600} of 0.3. Afterward, cells were aliquoted into 96 well microplates, and allow them to grow while maintaining the same temperature and shaking. Following induction of the reporter and target genes (see Section 2.3), mean fluorescence intensities were recorded for 10 h at an interval of 10 min, using the excitation (485/20 nm) and emission (525/20 nm) filters. We performed 6 technical replicates for each condition. We found weak variability between replicates. Results are the averages and standard error of the means.

2.5. Microscopy and image analysis

A few μl of cells were sandwiched between the coverslip and an agarose gel pad (2%), and visualized by a 488 nm argon ion laser (Melles–Griot) and an emission filter (HQ514/30, Nikon), using a Nikon Eclipse (Ti-E, Nikon) inverted microscope with a 100× Apo TIRF (1.49 NA, oil) objective. Fluorescence images were acquired by C2+ (Nikon), a point scanning confocal microscope system. The laser shutter was open only during exposure time to minimize photo bleaching. Simultaneously with the confocal images, phase contrast images were also captured by a CCD camera (DS-Fi2, Nikon). All images were acquired with NIS-Elements software (Nikon). Microscopy images were analysed using the software ‘CellAging’ (Häkkinen et al., 2013). For details, see Supplementary Materials and Methods, Sections 1.1 and 1.2.

2.6. Flow cytometry and gating

Cells carrying the target and reporter genes were grown and induced as described in Section 2.3. For this, from 5 ml of the bacterial culture, cells were diluted 1:10000 into 1 ml PBS and vortexed for 10 s. In each measurement, 50,000 events were recorded using an ACEA NovoCyte Flow Cytometer (ACEA Biosciences Inc., San Diego, USA), equipped with a blue (488 nm) and a yellow laser (561 nm) for excitation. For detection of MS2d-GFP and RNA-MS2d-GFP, we used the fluorescein isothiocyanate (FITC) detection channel (530/30 nm filter) for emission, with a PMT voltage setting of 417. For detection of red fluorescence proteins (mRFP1), we used the PE-Texas Red fluorescence detection channel (615/20 nm) for emission, with a PMT voltage setting of 584. We set a flow rate of 14 μl/min and a core diameter of

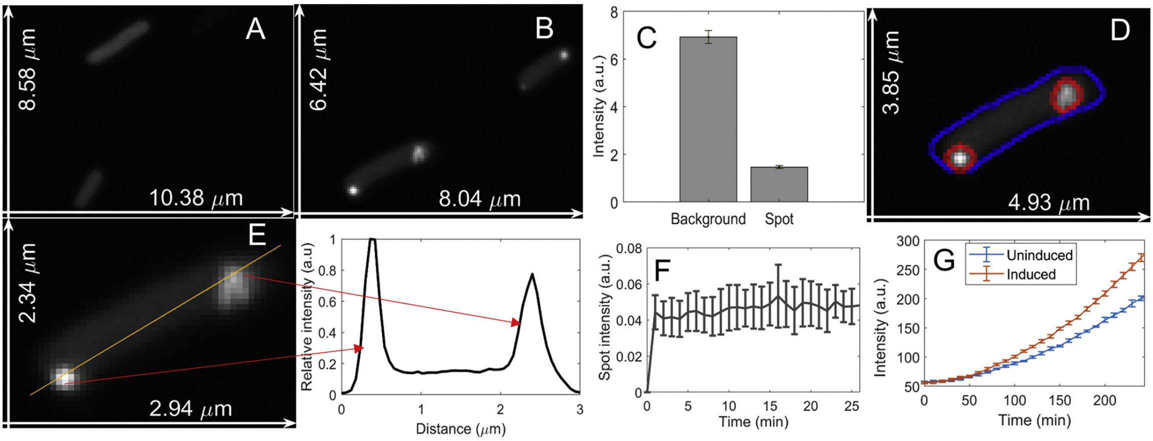


Fig. 1. A) Example microscopy image of cells carrying the reporter gene coding for MS2d-GFP, prior to the production of target RNAs. The cells are visible due to carrying a large amount of MS2d-GFP proteins; B) Example microscopy image of cells carrying the reporter gene coding for MS2d-GFP, after the production of target RNAs. The RNAs tagged with MS2d-GFP are visible as bright spots; (C) Mean total cell background fluorescence intensity and mean total fluorescence intensity of all RNA spots in individual cells (in arbitrary units), as measured by confocal microscopy (Methods, Section 2.5). Data from > 300 cells. The error bars are the standard error of mean. (D) Example image of a cell, along with the results of the segmentation of the cell border (blue line) and of the RNA spots within (red circles) using the tailored software 'SCIP' (Martins et al., 2018) (Methods, Section 2.5). (E) Left: example image of a cell along with a yellow line, manually introduced to obtain a fluorescence intensity profile using imageJ (Abramoff et al., 2004). Right: pixel intensity (in arbitrary units) along the yellow line shown on the left image. The peaks correspond to the regions where the two spots (tagged MS2d-GFP RNAs) are located. (F) Mean fluorescence intensity of individual tagged RNA molecules over time since first appearing. 10 tagged RNAs were tracked, all from cells with only one RNA. Also shown is the standard error of the mean (vertical bars). (G) Total fluorescence intensity (in arbitrary units) of cell populations over time, as measured by spectrophotometry, obtained from cells with target and reporter plasmids induced (brown line) and from cells with only the reporter plasmid induced (blue line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

7.7 μM. To avoid background signal from particles smaller than bacteria, the detection threshold was set to 5000 in FSC-H analysis. Data were extracted using the ACEA NovoExpress software.

We applied unsupervised gating (Razo-Mejia et al., 2018) to the flow cytometry data. We set the fraction of single-cell events whose data is used in the analysis (α) to 0.99, as it was sufficient to remove noncell events produced by debris, cell doublets, cell fragments, clump of cells, and other undesired events. Reducing α further did not change the results qualitatively. In addition, we removed events that did not exhibit fluorescence from free-floating MS2d-GFP by applying (manually) a minimum threshold (Supplementary Fig. S3, Right). Also, we removed < 0.01% of the events with highest FITC-H normalized by pulse width (F/W) values. Similarly, we removed < 0.01% of the events with highest R/W values. In all measurements by flow cytometry followed by data filtering, > 40,000 single-cells events were analysed per condition.

Finally, the total cell fluorescence differs with cell size (Supplementary Fig. S4, Right), while the concentration of MS2d-GFP does not (Supplementary Fig. S4, Left). To account for this, we normalized the FITC-H signal by the pulse Width (which differs with cell size (Cunningham, 1990; Traganos, 1984) denoted by F/W. Likewise, we also normalized the PETexasRedH signal by the pulse Width, denoted by R/W. For this reason, throughout the results section, we only refer to F/W and R/W when referring to flow cytometry data.

2.7. Mean, standard deviation, and skewness of single-cell distributions of RNA and protein numbers

We calculated the mean (M), Variance (Var), standard deviation (Sd), 3rd moment and skewness (S), of the distribution of single-cell RNA numbers (obtained from microscopy), and of the single-cell distributions of F/W, and R/W (obtained from flow cytometry), as shown in Table 1:

Table 1
Mean (M), Variance (Var), standard deviation (Sd), 3rd moment and skewness (S), of a distribution of observed values of the sample items, X, where ⟨.⟩ stands for average.

Feature	M	Var	Sd	3rd moment	S
Definition	$\langle X\rangle$	$\langle (X-\langle X\rangle)^2\rangle$	$\sqrt{\langle (X-\langle X\rangle)^2\rangle}$	$\langle (X-\langle X\rangle)^3\rangle$	$\frac{\langle (X-\langle X\rangle)^3\rangle}{Sd^3}$

The standard error of M is calculated from $\frac{Sd(X)}{\sqrt{N}}$, where N is the sample size of X. Meanwhile, the standard error (SE) of Var, Sd, 3rd moment and S is estimated using a non-parametric bootstrap method (Carpenter and Bithell, 2000; DiCiccio and Efron, 1996), by performing 10^3 random resamples with replacement, to obtain the bootstrapped distributions of Var, Sd, 3rd moment and S.

3. Results and conclusions

3.1. Time-course cell fluorescence in the presence and absence of RNA target for MS2d-GFP

We performed time-lapse microscopy measurements of *E. coli* cells carrying a gene coding for RNA target for MS2d-GFP, under the control of the Lac/Ara-1 promoter ($P_{Lac/ara-1}$). The cells also produce MS2d-GFP proteins from a multi-copy plasmid controlled by the $P_{LtetO-1}$ promoter (Materials and Methods, Section 2.2).

For RNAs target for MS2d-GFP to be readily detected, the cells need to contain multiple MS2d-GFP proteins (Golding et al., 2005). Due to this, their background is green fluorescent (Fig. 1A) and each target RNA appears as a bright spot in < 1 min after being produced (Tran et al., 2015) (Fig. 1B). In general, using these constructs and conditions, the cells produce from one to a few target RNAs during their lifetime (Häkkinen and Ribeiro, 2016).

In the absence of MS2d-GFP tagged RNAs, the total cell background fluorescence (i.e. the sum of the intensity of all pixels covering the cell area) is nearly only due to free floating MS2d-GFPs (Supplementary Fig. S5). In addition, this total background fluorescence is higher than the fluorescence of single MS2d-GFP tagged RNA spots (Fig. 1C). Nevertheless, MS2d-GFP tagged RNAs are clearly visible to the Human eye (Fig. 1D) and detectable by image analysis (Martins et al., 2018), as the fluorescence intensity of pixels with a spot is much higher than in near-neighbour pixels (Fig. 1E). Thus, using spatial information, RNA spots can be segmented, e.g., by kernel density estimation with a Gaussian kernel (Häkkinen and Ribeiro, 2015). In addition, the variability in fluorescence intensity of pixels where spots are absence is much smaller than the difference in fluorescence intensity between pixels with and without a spot (Fig. 1E). Due to this, one can subtract the mean background fluorescence from a spot's total fluorescence to obtain a 'corrected' spot intensity, without risk of wrongly adding a 'false' RNA spot or removing a 'true' RNA spot. Unfortunately, these methods cannot be applied to flow cytometry data, as it only informs on total cell fluorescence.

Even though the spots' fluorescence is weaker than the total cell fluorescence, we hypothesized that the production of an RNA target for MS2d-GFP increases the total cell fluorescence, since the binding of MS2d-GFP to the target RNA will protect bound MS2d-GFP proteins from degradation or loss of fluorescence intensity (Tran et al., 2015). This is due to the weak disassociation rate constant of MS2d from the specific target RNA sequence (Dolgosheina et al., 2014), and the high stability and long lifetime of the fluorescence intensity of MS2d-GFP tagged RNAs. In particular, Fig. 1F shows that the RNA-MS2d-GFP complexes have a weak mean fluorescence decay rate of $\sim 8 \times 10^{-5}\,s^{-1}$, which correspond to long mean half-lives of ~140 min, in agreement with past reports (Tran et al., 2015; Golding and Cox, 2004; Golding et al., 2005). Consequently, following the production of an MS2d-GFP tagged RNA, as a cell produces more MS2d-GFP, a new equilibrium in the number of MS2d-GFP in the cytoplasm is expected to be reached, causing the total cell fluorescence to become higher.

To validate this hypothesis, we measured by spectrophotometry the cells' fluorescence over time, when and when not inducing the target gene with L-Arabinose and IPTG. Also, we measured cell grow rates. From Supplementary Fig. S6, the cell growth rate does not differ between the conditions. Meanwhile, from Fig. 1G, the activation of the target gene, as time progresses and tagged RNAs accumulate, causes a continuous increase in the mean cell fluorescence.

Next, we subjected cells with the target gene controlled by $P_{Lac/ara-1}$ (responsible for producing the RNA target for MS2d-GFP) to various IPTG concentrations (Methods). As a control, we performed the same measurements on the strain without the target gene (Methods). We measured by flow cytometry the single-cell fluorescence intensity relative to cell size, so as to account for differences in cell size (Cunningham, 1990; Traganos et al., 1984). In particular, we calculated the 'FITC-H' signal relative to the 'pulse Width', here onwards referred to as F/W (Methods, Section 2.6).

As a control, we further verified by microscopy that cells do not differ significantly in morphology, for different IPTG concentrations, by comparing their mean length along the major axis. We found no significant differences between conditions (Supplementary Fig. S7).

From Fig. 2, while both strains are subject to the inducers, only cells carrying the gene coding for the RNA target for MS2d-GFP show increased F/W for increasing IPTG, which is consistent with the increase in RNA numbers as measured by microscopy (Fig. 3A and D). It is also consistent with the results by spectrophotometry (note that, at 1 mM IPTG, the total cell fluorescence of cells of the strain carrying the target is also approximately 30% higher, as in Fig. 1G). Given this and all of the above, we conclude that the increase in F/W with increasing IPTG is solely due to the appearance of RNAs tagged with MS2d-GFP.

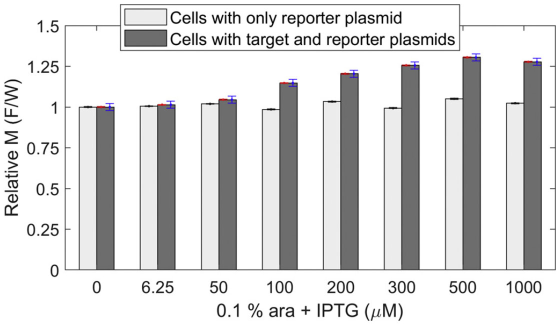


Fig. 2. (Light grey bars) Mean F/W values of the strain carrying only the multi-copy plasmid carrying the reporter gene, at various IPTG concentrations (x-axis), relative to its mean F/W value at the 0 μM IPTG condition. Its black error bars are the standard error of mean, estimated from the cells in each condition (Methods, section 2.7), relative to its mean F/W value at the 0 μM IPTG condition. (Dark grey bars) Mean F/W values of the strain with both the single-copy F-plasmid with the target gene and the multi-copy plasmid with the reporter gene at various IPTG concentrations (x-axis), relative to its mean F/W value at the 0 μM IPTG condition. The red error bars are the standard error of mean, estimated from the cells in each condition (Methods, section 2.7), relative to its mean F/W value at the 0 μM IPTG condition. The blue error bars result from the standard error of mean, relative to its mean F/W value at 0 μM IPTG condition, after adding the empirical variability between all measurements using cells with only the reporter gene. This estimation is explained in Supplementary Methods, section 1.6. In all conditions, cells were given 0.1% of L-Arabinose (Methods, Section 2.3).

3.2. Relationship between the statistics of single-cell RNA numbers obtained by confocal microscopy and single-cell F/W obtained by flow cytometry

We next measured by microscopy and image analysis (Methods, section 2.5) the RNA numbers produced by our gene of interest, under the control of $P_{Lac/ara-1}$, for different concentrations of IPTG. Fig. 3A-3C show the mean, standard deviation, and skewness of the single-cell distribution of these numbers (Methods, Section 2.7) as a function of IPTG, respectively.

Next, we extracted the same three statistics of the single-cell distribution of F/W values obtained in the same conditions by flow cytometry. Results are shown in Fig. 3D-3F. Supplementary Fig. S8 (Left) shows the probability density functions of the single-cell F/W values, for each condition.

Given this, we investigated the relationship between the statistics for F/W and the statistics for RNA numbers per cell. Results in Supplementary section 1.5, show that there is a linear fit between the two Means, the two Variances and, the two third moments, respectively.

Given these linear relationships, to evaluate whether the moments of single cell RNA numbers from microscopy and single cell F/W values of flow cytometry are correlated, we plotted the results of Mean (M), Variance (Var) and the third moment obtained by microscopy against the results of M, Var and the third moment obtained by flow cytometry in scatter plots (Fig. 4A-C). Next, we did a linear fit to the data, which was performed using the linear regression fitting method explained in Supplementary Methods, section 1.4. The adjusted R^2 values and corresponding *p*-values of the linear fit are shown in Supplementary Table S1. We find that Mean, Var and the third moment are well fitted by a line (in Fig. 4).

Hence, we conclude that there is a good linear fit between the Mean, Var and the 3rd moment of the single-cell distributions of RNA numbers obtained by microscopy, and the Mean, Var, and the 3rd moment of the single-cell distributions of F/W values obtained by flow cytometry, respectively.

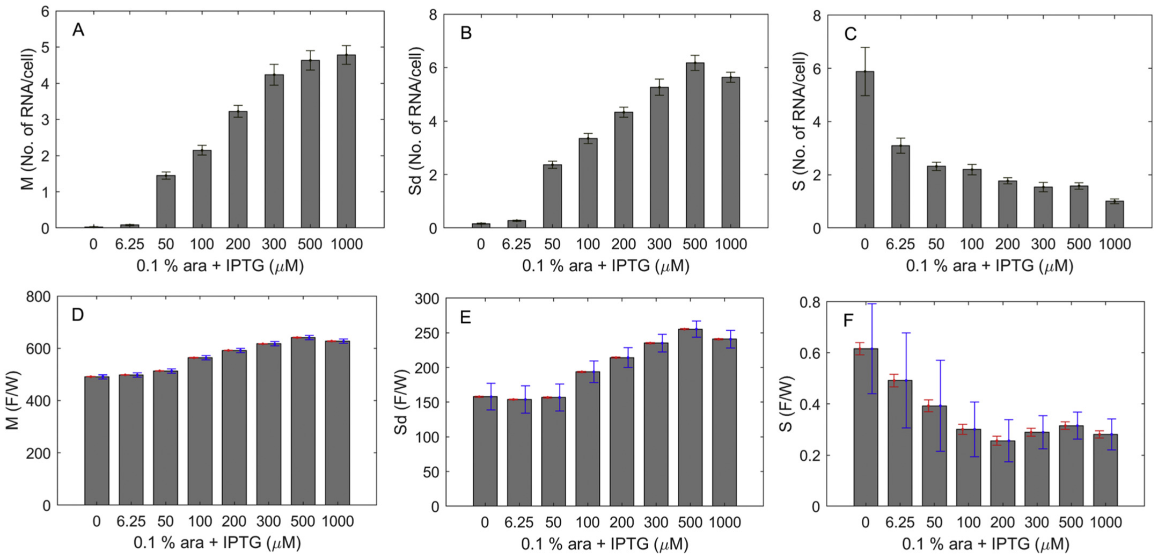


Fig. 3. (A) Mean, M, (B) Standard deviation, Sd, and (C) Skewness, S, of single-cell distributions of integer-valued RNA numbers obtained by microscopy, as a function of IPTG concentration (x-axis). The standard error of M, Sd and S of RNA numbers was estimated as described in Methods, section 2.7. (D) Mean, (E) Standard deviation, and (F) Skewness of the single-cell distribution of F/W values obtained by flow cytometry. The red error bars are standard errors of the statistics (Methods, Section 2.7). The blue error bars are the standard error of the statistics after adding variability estimated from eight technical replicates of cells carrying only the reporter gene (Supplementary Methods, Section 1.6). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

These results are, as expected, dependent on degree of background noise, produced by MS2d-GFP (random motion and measurement error generate spatial heterogeneity). This noise could differ, e.g., in different environments or if different plasmids were used to produce MS2d-GFP. We thus tested the effects of increased background noise on our estimation of M, Var, and 3rd moment in cells observed by Flow Cytometry. For this, we modelled increasing background noise by adding increasingly higher Gaussian noise to the total cell fluorescence (F/W) obtained by Flow Cytometry. These added noises are shown in Supplementary Fig. 10A.

The consequences of adding the increasingly higher noise on the mean, variance, and the 3rd moment of the single-cell fluorescence distributions, as measured by Flow Cytometry, are shown, respectively, in Supplementary Fig. 10B, 10C, 10D.

Visibly, from Supplementary Fig. 10B, the addition of Gaussian noise to the single-cell F/W distribution at different IPTG concentrations (Noise corrupted F/W distribution), does not perturb the mean. In particular, even though the Gaussian noise was gradually increased from σ = 0 to 400, the best fitting lines between the mean of the noise corrupted F/W distribution and mean RNA numbers per cell obtained from microscopy are all identical to the best fitting line when using the original F/W distribution (Supplementary Fig. S10B).

Meanwhile, we expect increasing variance in the noise-corrupted F/W distributions. However, the best fitting line between variance of the noise corrupted F/W distributions and variance of the single-cell RNA numbers distribution shows that only the intercept changes, not the slope (Supplementary Fig. S10C). As such, one can reliably quantify the variance of RNA numbers from the variance of noise corrupted F/W distributions.

Finally, the third moment of noise is not expected to change with increasing Gaussian noise. This can be seen at low noise levels (0 to 200), as the best fitting line between the third moment of noise corrupted F/W distributions and the third moment of the distribution of RNA numbers per cell is almost the same. However, at higher noise levels (300 and 400), the best fitting line shifted slightly (Supplementary Fig. S10D). This may be because the standard

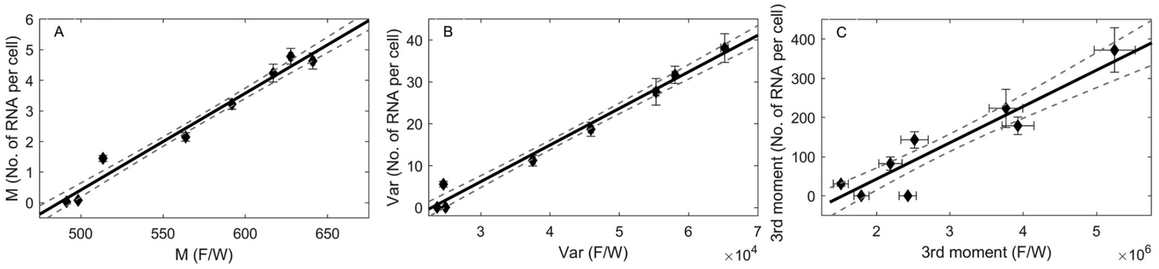


Fig. 4. Scatter plots between (A) Mean (M), (B) Variance (Var), (C) 3rd Moment of the single-cell distributions of F/W values obtained by flow cytometry against M, Var and 3rd Moment of the single-cell distributions of RNA numbers in individual cells obtained by Microscopy for various induction strengths (0, 6.25, 50, 100, 200, 300, 500, and 1000 μM IPTG). The error bars of the points on x and y directions are standard errors estimated as in Methods, section 2.7. In each plot, we obtained the best linear fit (black straight line) as described in Supplementary Methods, section 1.4. The dotted lines are the standard error of the fitted line.

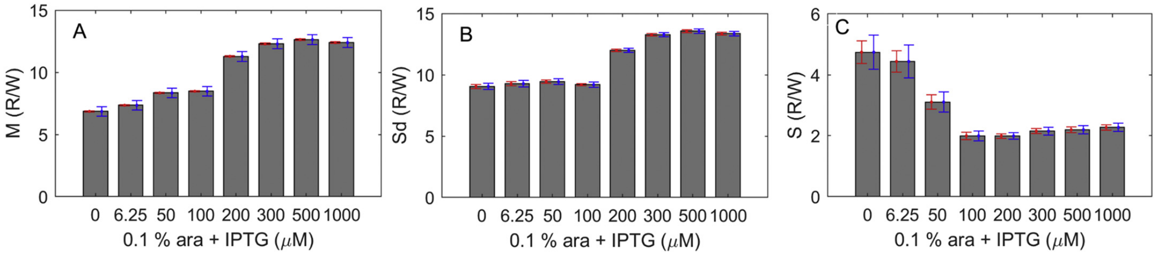


Fig. 5. (A) Mean, M, (B) Standard deviation, Sd, and (C) Skewness, S, of single-cell distributions of R/W values, when subject to various IPTG concentrations. The red error bars are standard errors (Methods, section 2.7). The blue error bars are the standard error of the statistics after adding empirical variability estimated from cells carrying only reporter gene (Supplementary section 1.6 and Supplementary Fig. S12). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

deviation, at the lower induction levels, is smaller than the added noise. In that regime, the parameters of the best fitted line start being sensitive to Gaussian noise, which increases the error of the estimation of the third moment.

3.3. Validation of the quantification of MS2d-GFP tagged RNAs from flow cytometry data

If the signal detected by Flow cytometry is produced by MS2d-GFP tagged RNAs, one should be able to detect the corresponding proteins produced from these RNAs (in particular, mRFP1 red fluorescent proteins, see Methods). To test this, from the same flow cytometry measurements as above, we also extracted the single-cell distribution of PETexasRed-H and normalized these signals by the Pulse Width (denoted as R/W). From the single-cell R/W distribution, we obtained its mean, standard deviation and skewness for each induction strength (Fig. 5A-C). Supplementary Fig. S8 (right) shows the probability density function of R/W for each induction strength.

To assess if the protein statistics (Fig. 5A-C) follows the RNA statistics (Fig. 3D-F), we plotted the values of each statistic in scatter plots (Fig. 6A-C) and fitted with a linear fit. The adjusted R^2 values and corresponding *p*-values of the linear fit are shown in Supplementary Table S2. From the Figures and Table, all three statistics are well fitted by a line. Given the adjusted R^2 values and p-values, we conclude that there is a strong linear fit between those statistics of the single-cell distribution of FITC-H normalized by Pulse width and PETexasRed-H normalized by Pulse width obtained by flow cytometry, respectively. These results confirm that the statistics of distribution of F/W values in Fig. 3D-F should be the result of single-cell distribution of MS2d-GFP tagged RNAs.

3.4. Estimation of mean, standard deviation and skewness of the single-cell distribution of RNA numbers from single-cell F/W values

From the above, it should be possible to estimate the statistics of the distribution of single-cell RNA numbers from the single-cell distribution of F/W values. In particular, it should be possible to 'map' the flow cytometry data to the microscopy data. E.g. one could calibrate two, or more, data points (conditions) of the flow-cytometry data to the corresponding points (conditions) of the microscopy data. Then, we could estimate the RNA numbers statistics of the remaining F/W data points by linear interpolation and/or extrapolation. From this, we can obtain an absolute RNA count scale for estimating the mean, standard deviation, and skewness of the single-cell distribution of RNA numbers from flow-cytometry data.

We start by calibrating the difference between a pair of conditions from flow-cytometry data (e.g. 0 and 1000 μM IPTG) to the difference between the corresponding pair of conditions from microscopy data. This process has to be done independently for the mean, variance, and third moment, but one can use any pair of conditions for each of the moments.

Here we use the data in Fig. 4A-C to obtain the necessary pairs of data points. For this, we started by testing all possible combinations of pairs of data points (Fig. 4A-C contain 8 data points each, and thus, there are 28 possible pairs of data points). Out of these, there are several pairs that provide calibration lines that are consistent between them, and thus can be used to obtain reliable results.

In order to find the largest group of calibration lines that are consistent between, we plotted the y-intercepts against the slopes of the 28 calibration lines. Then we calculated the location of a 'Median point' in that graph whose x-coordinate is the median of the slopes and the y-coordinate is the median of the y-intercepts of the calibration lines (Fig. S9A-S9C).

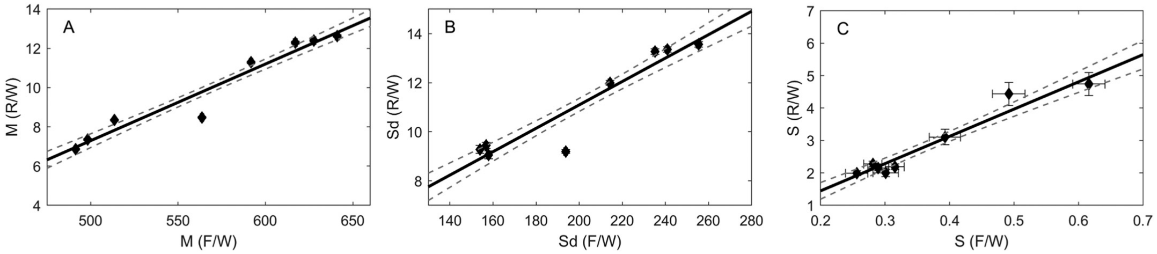


Fig. 6. Scatter plots between (A) Mean (M), (B) Standard deviation (Sd), (C) Skewness (S) of the single-cell distributions of F/W values against M, Sd and S of the single-cell distributions of R/W values for various induction strengths, differing in IPTG concentration (0, 6.25, 50, 100, 200, 300, 500, and 1000 μM IPTG). The error bars of the points are the standard errors (red error bars as in Fig. 3D-F and Fig. 5A-C). In each plot, we obtained the best linear fit (black straight line) as described in Supplementary Methods, section 1.4. The dotted lines are the standard error of the fitted line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

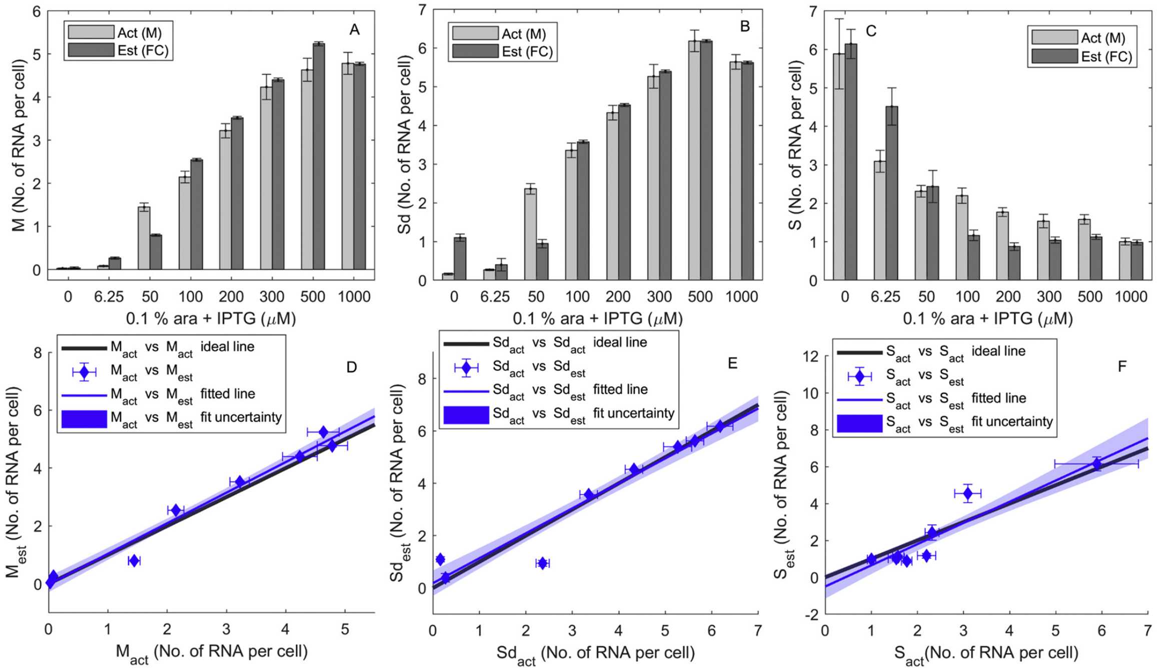


Fig. 7. (A) Mean single-cell RNA numbers estimated from Flow cytometry data, using microscopy data (Mean RNA numbers per cell) in the minimum (0 μM IPTG) and maximum induction (1000 μM IPTG) conditions for the calibration. (B) Standard deviation of single-cell RNA numbers estimated from Flow cytometry data, using the microscopy data (Variance of RNA numbers per cell) in 6.25 μM IPTG and maximum induction conditions (1000 μM IPTG) for the calibration. (C) Skewness of single-cell RNA numbers estimated from Flow cytometry, using the microscopy data (3rd moment of RNA numbers per cell) in 50 and 1000 μM IPTG for the calibration. Light grey bars are the actual values obtained from microscopy data and dark grey bars are the estimated values from flow cytometry data (F/W). (D) Scatter plot between estimated and actual mean values of single-cell RNA numbers. The blue points along with their standard error bars are the estimated mean of single-cell RNA numbers (M_{est}), plotted against the corresponding actual values (M_{act}). Also shown is the best linear fit to the blue points (blue line) along with the uncertainty of the fit (blue area). Finally, it is shown the 'ideal' linear fit (black line). The black line crosses 0 at the y-axis and has an inclination of 1, which would correspond to the estimated values being identical to the actual values. (E) Scatter plot between estimated and actual standard deviations of single-cell RNA numbers. The blue points along with their standard error bars are the estimated standard deviations of singe-cell RNA numbers (Sd_{est}), plotted against the corresponding actual values (Sd_{act}). Also shown is the best linear fit to the blue points (blue line) along with the uncertainty of the fit (blue area). Finally, it is shown the 'ideal' linear fit (black line). The black line crosses 0 at the y-axis and has an inclination of 1, which would correspond to the estimated values being identical to the actual values. (F) Scatter plot between estimated and actual skewness of single-cell RNA numbers. The blue points along with their standard error bars are the estimated skewness of single-cell RNA numbers (S_{est}), plotted against the corresponding actual values (S_{act}). Also shown is the best linear fit to the blue points (blue line) along with the uncertainty of the fit (blue area). Finally, it is shown the 'ideal' linear fit (black line). The black line crosses 0 at the y-axis and has an inclination of 1, which would correspond to the estimated values being identical to the actual values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Next, we found that using the 33% points with smaller Euclidean distance to the Median point, one obtains consistent calibration lines for the mean, variance, and third moment. These lines are shown, respectively, in Supplementary Fig. S9D-S9F. As expected, these set of consistent lines correspond to using pairs of data point that differ significantly between them in the Fig. 4A-4C (e.g. the pair of conditions 0 and 1000 μM IPTG).

Next, using these calibration lines (see Supplementary Section 1.7), we estimated the mean, standard deviation and skewness (along with their standard errors) of the single-cell distribution of RNA numbers from the distribution of F/W values. Fig. 7A shows the estimated mean of the single-cell distribution of RNA numbers from flow cytometry data, for each condition, using the calibration line obtained by using the pair of conditions 0 μM IPTG and 1000 μM IPTG. Fig. 7B shows the estimated standard deviation of the single-cell distribution of RNA numbers from flow cytometry data, for each induction level, using the calibration line obtained using the pair of conditions 6.25 μM IPTG and 1000 μM IPTG. Fig. 7C shows the estimated skewness of the single-cell distribution of RNA numbers from flow cytometry data, for each induction level, using the calibration line obtained using the pair of conditions 50 μM IPTG and 1000 μM IPTG.

To evaluate the accuracy of the estimated mean, standard deviation, and skewness from flow cytometry data, we plotted them against the corresponding actual values, obtained by microscopy (Fig. 7D-F). If the estimations are accurate, one expects the best-fitting line to these points (black lines in Fig. 7D-F) to exhibit a 45-degree inclination and to intercept the y-axis at zero. To test this, we plotted also the 'ideal line' (black lines in Figs. 7D-F). Next, we compared by analysis of covariance (McDonald, 2009) whether the best fitting line and the ideal line could be distinguished in slope and intercept, in a statistical sense. Results of these tests for the mean, standard deviation, and skewness (Supplementary Table S3) show that the best fitting line cannot be distinguished from the ideal line, from which we conclude that the estimations are accurate.

Given the above, we conclude that collecting data using microscopy from two conditions differing in RNA numbers, allows accurate estimations of the mean, standard deviation, and skewness of single cell distributions of RNA numbers from the distribution of total cell fluorescence measured byflow cytometry in multiple conditions differing in induction strength, using MS2d-GFP tagging of RNA.

Next, as in Section 3.2, we tested for the robustness of these estimations by adding increasingly higher Gaussian noise to the empirical F/W values. Results of these estimations using the noise corrupted F/W values are shown in Supplementary Fig. S11. Visibly, the estimations of the mean and standard deviation are not significantly affected. Meanwhile, the estimations of skewness are only affected for the 3 lowest induction conditions, similar to the results in Section 3.2, for similar reasons.

It is noted that the added Gaussian noise is much above what we expect to observe in real data collected from cells with the MS2d-GFP technology. I.e., the highest artificially added noise is much higher than the observed noise at the lowest induction conditions. Specifically, e.g., in the case at 6.25 μM IPTG induction, the observed standard deviation is ~155, while we added up to σ = 400 artificial Gaussian noise.

4. Discussion

Presently, FISH and MS2d-GFP RNA tagging are two of the preferred technologies for visualizing and quantifying RNA numbers in individual cells (Raj and van Oudenaarden, 2009). While the latter is likely more intrusive, it has some advantages, such as allowing to track the dynamics of RNA production in live cells, which has been used to dissect the underlying kinetic steps of transcription initiation, not possible otherwise (Lloyd-Price et al., 2016). So far, the use of MS2d-GFP RNA tagging has required microscopy and subsequent image analysis, which heavily limits the amount of data that can be produced. Further, image analysis introduces many errors (even with manual corrections). The ability to extract information using this technique from flow cytometry would overcome both limitations.

We have shown that it is possible to perform flow cytometry of cells expressing MS2d-GFP and RNA targets for MS2d-GFP and accurately estimate the mean, standard deviation, and skewness of the single-cell distribution of RNA numbers. Importantly, we have shown that the results cannot be distinguished, in a statistical sense, from those obtained by microscopy followed by manually corrected image analysis. Also, we have shown (Fig. 4) that the estimations of integer valued RNA numbers in individual cells are highly correlated with single-cell fluorescent protein levels, which is strong evidence of the accuracy of the estimations.

Interestingly, the estimations of mean single-cell RNA numbers from flow-cytometry data only exhibit significant discrepancy with the microscopy data when RNA production is weaker (Fig. 7). Past studies using microscopy and image analysis of cells with MS2d-GFP tagged RNAs (Häkkinen et al., 2014; Häkkinen and Ribeiro, 2015, 2016) suggest that these discrepancies arise mostly from errors in the microscopy data, which is based on ~500 cells per condition. In comparison, flow-cytometry data is based on ~40,000 cells (each of which randomly collected from a well-stirred medium). Consequently, the microscopy data is more prone to errors due to small sample size, particularly in weak expression conditions, where it is harder to select images of cells that are good representatives of the population. Nevertheless, regarding the estimations from flow-cytometry data, it is worth noting the unexpected value for the standard deviation at 0 μM IPTG (Fig. 7B), likely due to random biological variability.

In general, our results indicate that estimations of the statistics of single-cell RNA numbers can largely be performed from flow cytometry data and then be complemented by microscopy measurements (with scaling only requiring population images in two conditions differing in mean RNA numbers per cell). The large number of cells that can be observed by flow-cytometry promises precise estimations these statistics. Namely, we note that the estimations of single-cell RNA number statistics performed here are accurate not only in what concerns mean and standard deviation, but also skewness, which is in itself evidence of the accuracy of the estimations. Relevantly, this ensures that this technique can be used to estimate the propensity of a specific transcription kinetics to overcome thresholds in RNA and protein numbers (which is of significance in the context of small genetic circuits, among other). Finally, it is worth noting that, in principle, the method is readily applicable to cells with fluorescently tagged RNA using FISH technology. In this case, the methodology is expected to contribute in decreasing the effects of noise due to auto fluorescence from natural cellular components.

Overall, we expect the methodology proposed here to be useful in studies of in vivo transcription at single-molecule level, by adding more reliability to the conclusions, as these will be based on larger number of cells (by 2 to 3 orders of magnitude when compared to when collecting data by microscopy and image analysis). Also, much more conditions can be tested, due to the incomparably faster speed by which results can be obtained, compared to when using microscopy and image analysis.

Funding

Work supported by Tampere University Graduate Program (Finland) [to M.M.B. and V.C.]; Finnish Academy of Science and Letters [to C.P.]; Pirkanmaa Regional Fund [to V.K.K.]; Academy of Finland [295027 to A.S.R.]; and Jane and Aatos Erkko Foundation [610536 to A.S.R.]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mimet.2019.105745.

References

Abramoff, M.D., Magalhaes, P.J., Ram, S.J., 2004. Image Processing with ImageJ. Biophotonics International 11, 36–42.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2002. Molecular Biology of the Cell. Garland Science, New York.

Arrigucci, R., Bushkin, Y., Radford, F., Lakehal, K., Vir, P., Pine, R., Martin, D., Sugarman, J., Zhao, Y., Yap, G.S., Lardizabal, A.A., Tyagi, S., Gennaro, M.L., 2017. FISH-flow, a protocol for the concurrent detection of mRNA and protein in single cells using fluorescence in situ hybridization and flow cytometry. Nat. Protoc. 12, 1245–1260. https://doi.org/10.1038/nprot.2017.039.

Bumgarner, R., 2013. DNA microarrays: types, applications, and their future. Curr. Protoc. Mol. Biol. Chapter 22, Unit 22, 1. https://doi.org/10.1002/0471142727.mb2201s101.

Bunka, D.H.J., Stockley, P.G., 2006. Aptamers come of age – at last. Nat. Rev. Microbiol. 4, 588–596. https://doi.org/10.1038/nrmicro1458.

Bushkin, Y., Radford, F., Pine, R., Lardizabal, A., Mangura, B.T., Gennaro, M.L., Tyagi, S., 2015. Profiling T cell activation using single molecule-FISH and flow cytometry. J. Immunol. 194, 836–841. https://doi.org/10.4049/jimmunol.1401515.

Carpenter, J., Bithell, J., 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. 19, 1141–1164. https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F.

Chong, S., Chen, C., Ge, H., Xie, X.S., 2014. Mechanism of transcriptional bursting in bacteria. Cell 158, 314–326. https://doi.org/10.1016/j.cell.2014.05.038.

Cunningham, A., 1990. Fluorescence pulse shape as a morphological indicator in the analysis of colonial microalgae by flow cytometry. J. Microbiol. Methods 11, 27–36. https://doi.org/10.1016/0167-7012(90)90044-7.

DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence intervals. Stat. Sci. 11, 189–228. https://doi.org/10.1214/ss/1032280214.

Dolgosheina, E.V., Jeng, S.C.Y., Panchapakesan, S.S.S., Cojocaru, R., Chen, P.S.K., Wilson, P.D., Hawkins, N., Wiggins, P.A., Unrau, P.J., 2014. RNA mango aptamer-fluorophore: a bright, high-affinity complex for RNA labeling and tracking. ACS Chem. Biol. 9, 2412–2420. https://doi.org/10.1021/cb500499x.

Elowitz, M.B., Levine, A.J., Siggia, E.D., Swain, P.S., 2002. Stochastic gene expression in a single cell. Science 297, 1183–1186. https://doi.org/10.1126/science.1070919.

Engl, C., 2018. Noise in bacterial gene expression. Biochem. Soc. Trans. 47, 209–217. https://doi.org/10.1042/bst20180500.

Femino, A.M., Fay, F.S., Fogarty, K., Singer, R.H., 1998. Visualization of single RNA transcripts in situ. Science 280, 585–590. https://doi.org/10.1126/science.280.5363.585.

Fusco, D., Accornero, N., Lavoie, B., Shenoy, S.M., Blanchard, J.M., Singer, R.H., Bertrand, E., 2003. Single mRNA molecules demonstrate probabilistic movement in living mammalian cells. Curr. Biol. 13, 161–167. https://doi.org/10.1016/S0960-

9822(02)01436-7.

Golding, I., Cox, E.C., 2004. RNA dynamics in live Escherichia coli cells. Proc. Natl. Acad. Sci. 101, 11310–11315. https://doi.org/10.1073/pnas.0404443101.

Golding, I., Paulsson, J., Zawilski, S.M., Cox, E.C., 2005. Real-time kinetics of gene activity in individual bacteria. Cell 123, 1025–1036. https://doi.org/10.1016/j.cell.2005.09.031.

Häkkinen, A., Ribeiro, A.S., 2015. Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data. Bioinformatics 31, 69–75. https://doi.org/10.1093/bioinformatics/btu592.

Häkkinen, A., Ribeiro, A.S., 2016. Characterizing rate limiting steps in transcription from RNA production times in live cells. Bioinformatics 32, 1346–1352. https://doi.org/10.1093/bioinformatics/btv744.

Häkkinen, A., Muthukrishnan, A.B., Mora, A., Fonseca, J.M., Ribeiro, A.S., 2013. CellAging: a tool to study segregation and partitioning in division in cell lineages of Escherichia coli. Bioinformatics 29, 1708–1709. https://doi.org/10.1093/bioinformatics/btt194.

Häkkinen, A., Kandhavelu, M., Garasto, S., Ribeiro, A.S., 2014. Estimation of fluorescence-tagged RNA numbers from spot intensities. Bioinformatics 30, 1146–1153. https://doi.org/10.1093/bioinformatics/btt766.

Higuchi, R., Fockler, C., Dollinger, G., Watson, R., 1993. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. Nat. Biotechnol. 11, 1026–1030. https://doi.org/10.1038/nbt0993-1026.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., Linnarsson, S., 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods 11, 163–166. https://doi.org/10.1038/nmeth.2772.

Kærn, M., Elston, T.C., Blake, W.J., Collins, J.J., 2005. Stochasticity in gene expression: from theories to phenotypes. Nat. Rev. Genet. 6, 451–464. https://doi.org/10.1038/nrg1615.

Kandavalli, V.K., Tran, H., Ribeiro, A.S., 2016. Effects of σ factor competition are promoter initiation kinetics dependent. Biochim. Biophys. Acta - Gene Regul. Mech. 1859, 1281–1288. https://doi.org/10.1016/j.bbagrm.2016.07.011.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., Taipale, J., 2012. Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods 9, 72–74. https://doi.org/10.1038/nmeth.1778.

Lloyd-Price, J., Startceva, S., Kandavalli, V., Chandraseelan, J.G., Goncalves, N., Oliveira, S.M.D., Häkkinen, A., Ribeiro, A.S., 2016. Dissecting the stochastic transcription initiation process in live Escherichia coli. DNA Res. 23, 203–214. https://doi.org/10.1093/dnares/dsw009.

Martins, L., Neeli-Venkata, R., Oliveira, S.M.D., Häkkinen, A., Ribeiro, A.S., Fonseca, J.M., 2018. SCIP: a single-cell image processor toolbox. Bioinformatics 34, 4318–4320. https://doi.org/10.1093/bioinformatics/bty505.

McDonald, J.H., 2009. Handbook of Biological Statistics, 2nd ed. Sparky House Publishing, Baltimore, MD.

Oliveira, S.M.D., Häkkinen, A., Lloyd-Price, J., Tran, H., Kandavalli, V., Ribeiro, A.S., 2016. Temperature-dependent model of multi-step transcription initiation in Escherichia coli based on live single-cell measurements. PLoS Comput. Biol. 12, 1–18. https://doi.org/10.1371/journal.pcbi.1005174.

Ozbudak, E.M., Thattai, M., Kurtser, I., Grossman, A.D., Van Oudenaarden, A., 2002. Regulation of noise in the expression of a single gene. Nat. Genet. 31, 69–73. https://doi.org/10.1038/ng869.

Peabody, D.S., 1993. The RNA binding site of bacteriophage MS2 coat protein. EMBO J. 12, 595–600.

Pedraza, J.M., Van, O., 2005. A. Noise propagation in genetic networks. Science 307, 1965–1969. https://doi.org/10.1126/science.1109090.

Raj, A., van Oudenaarden, A., 2009. Single-molecule approaches to stochastic gene expression. Annu. Rev. Biophys. 38, 255–270. https://doi.org/10.1146/annurev.biophys.37.032807.125928.

Raj, A., Van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., Tyagi, S., 2008. Imaging individual mRNA molecules using sets of singly labeled probes. Nat. Methods 5, 877–879. https://doi.org/10.1038/nmeth.1253.

Razo-Mejia, M., Barnes, S.L., Belliveau, N.M., Chure, G., Einav, T., Lewis, M., Phillips, R., 2018. Tuning transcriptional regulation through signaling: a predictive theory of allosteric induction. Cell Syst. 6, 456–469.e10. https://doi.org/10.1016/j.cels.2018.02.004.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., Arnheim, N., 1985. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230, 1350–1354. https://doi.org/10.1126/science.2999980.

Singer, R.H., Ward, D.C., 1982. Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog. Proc. Natl. Acad. Sci. U. S. A. 79, 7331–7335. https://doi.org/10.1073/pnas.79.23.7331.

Startceva, S., Kandavalli, V.K., Visa, A., Ribeiro, A.S., 2019. Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression. Biochim. Biophys. Acta - Gene Regul. Mech. 1862, 119–128. https://doi.org/10.1016/j.bbagrm.2018.12.005.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K., Surani, M.A., 2009. mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods 6, 377–382. https://doi.org/10.1038/nmeth.1315.

Tiberi, S., Walsh, M., Cavallaro, M., Hebenstreit, D., Finkenstädt, B., 2018. Bayesian inference on stochastic gene transcription from flow cytometry data. Bioinformatics 34, 647–655. https://doi.org/10.1093/bioinformatics/bty568.

Traganos, F., 1984. Flow cytometry: principles and applications. I. Cancer investigations. Cancer Investig. 2, 149–163.

Tran, H., Oliveira, S.M.D., Goncalves, N., Ribeiro, A.S., 2015. Kinetics of the cellular intake of a gene expression inducer at high concentrations. Mol. BioSyst. 11, 2579–2587. https://doi.org/10.1039/c5mb00244c.

Trcek, T., Chao, J.A., Larson, D.R., Park, H.Y., Zenklusen, D., Shenoy, S.M., Singer, R.H., 2012. Single-mRNA counting using fluorescent in situ hybridization in budding yeast. Nat. Protoc. 7, 408–419. https://doi.org/10.1038/nprot.2011.451.

Yu, J., Xiao, J., Ren, X., Lao, K., Xie, X.S., 2006. Probing gene expression in live cells, one protein molecule at a time. Science 311, 1600–1603. https://doi.org/10.1126/science.1119623.

Supplementary Material for:

Estimating RNA numbers in single cells by RNA fluorescent tagging and flow cytometry

Mohamed N.M. Bahrudeen, Vatsala Chauhan, Cristina S.D. Palma, Samuel M.D. Oliveira, Vinodh K. Kandavalli, and Andre S. Ribeiro

1. Supplementary Materials and Methods

1.1. Image analysis of microscopy data

Cells are visualized by phase-contrast and fluorescence microscopy, nearly simultaneously. Information from the images is automatically extracted by the software 'CellAging' (Häkkinen et al., 2013).

Cell segmentation is performed by applying the Gradient path labelling algorithm (Mora et al., 2011), and uses classifiers for merging (to reduce over segmentation) and discarding segments (e.g. air bubbles and unwanted artifacts). The classifiers were built by applying the Classification and Regression Trees algorithm (Breiman et al.,1984), and was manually trained by an expert using example images (Queimadelas et al., 2012). When necessary, in the end, we performed manual corrections.

Next, the software aligns confocal images (semi-automatically) with the corresponding phase-contrast images. This is executed by thin-plate spline interpolation for the registration transform (by manual selection of 5-8 landmarks), so as to adjust the cell masks to the corresponding cells in the confocal image. Finally, fluorescent spots (MS2d-GFP tagged RNAs) inside the cells are automatically detected using the Gaussian surface-fitting algorithm (Häkkinen and Ribeiro, 2015) (Figure 1D in main manuscript). The resulting data is used for RNA quantification of fluorescent spots in individual cells (supplementary section 1.2).

1.2. RNA quantification from fluorescent spots

Integer-valued number of MS2d-GFP-tagged mRNA molecules are quantified from microscopy images as in (Golding et al., 2005; Kandavalli et al., 2016; Lloyd-Price et al., 2016; Mäkelä et al., 2017; Oliveira et al., 2016). Following the segmentation of RNA spots of multiple cells in an image (e.g. Figure 1D in main manuscript), the mean cell background fluorescence intensity from unbound MS2d-GFP proteins of each cell (average over all pixels not containing an 'RNA-spot') is subtracted from the intensity of each segmented RNA-spot. From the results from all cells (~4500 cells), we estimated how many tagged RNAs are in each cell from a histogram of total RNA spots intensity per cell (Häkkinen et al., 2015).

For this, we first combined the data on each spot, from all conditions, into a single distribution of single-cell RNA spot intensities (as we found no significant difference in mean fluorescence intensity between cells in the various conditions, which is expected as the reporter is equally induced and the cells are in the same media conditions and temperature). From this distribution, we estimated the parameter values in maximum likelihood sense using a maximum a posteriori classifier to estimate the RNA numbers in each cell, in each condition (Häkkinen et al., 2015).

After RNA quantification, 0.5% or less cells with the highest total spot fluorescence intensities were removed from the analysis, if they were clear outliers (due to errors in imaging or abnormal overexpression of MS2d-GFP). Similarly, a few cells that were visible by phase contrast but did not express MS2d-GFP were also removed from the analysis (Supplementary Figure S3, Left). Interestingly, we found a similar fraction of non-expressing cells when using flow-cytometry (Supplementary Figure S3, Right).

1.3. RNA spots lifetime and temporal fluorescence intensity

For RNA counting by MS2d-GFP tagging to be accurate, both when using microscopy or flow-cytometry, the fluorescence intensity of tagged RNAs has to be constant over time and be largely uniform in cells in the same image. In practice, this implies that the fluorescence of a tagged RNA when first appearing needs to be near identical to subsequent moments, for a significant period of time (e.g. a few hours). Both conditions have been shown to be fulfilled in (Tran et al., 2015; Oliveira et al., 2019; Startceva et al., 2019).

To verify this, we measured the mean and standard error of the mean of the fluorescence intensity of 10 individual, MS2d-GFP tagged RNA molecules. We selected by visual inspection cells that contained only 1 tagged RNA at any given moment of the observation period. Images were taken once per minute, for 108 minutes. To plot their mean fluorescence intensity over time, for simplicity, we synchronized the moments when the tagged RNAs were first observed (Figure 1F). From Figure 1F, the ‘maximum’ RNA spot fluorescence is always reached in less than 1 minute. Afterwards, the spots’ fluorescence intensity remains fairly constant over time, with the main contribution to this standard error of the mean being from spots (rarely) leaving (and then returning to) the focal plane. ‘Bleaching’ of tagged RNAs was not observed in any case.

1.4. Linear regression fitting using Ordinary Least Squares

To perform linear fitting (Figures 4, 6 and 7 in main manuscript), we represent the uncertainty of each of the N empirical data points by m points (each without uncertainty), resulting in $n = N \times m$ points. Each of these n points is obtained by random sampling from a normal distribution whose mean (μ) and standard deviation (σ) equal the mean and error of the empirical data point, respectively. Here, we have set $m = 1000$, as it was sufficient to represent the error bars of the actual data points (obtained from the standard error, see main manuscript, section 2.7).

Using such a large number of points per empirical data point, results in a significant underestimation of the standard error of the fit parameters and of their p-values. To correct for this, we use a multivariate regression model (Alexopoulos, 2010) with k independent variables $x_1, \ldots, x_k$ and one response variable, y, for each of the $i = 1, \ldots, n$ points:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} + \varepsilon \tag{S1}$$

Here, β_0, β_1, … , β_k are regression coefficients and ε is the error. Next, each of the N empirical data points with uncertainty is replaced by the m points without uncertainty, resulting in the n points, with Y_{obs} being the vector of all y_i:

$$Y_{obs} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ X_{obs} = \begin{bmatrix} 1 & x_{11} & .. & x_{1k} \\ 1 & x_{21} & .. & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & .. & x_{nk} \end{bmatrix},\ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},\ E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

which can be written as:

$$Y_{obs} = X_{obs} \cdot \beta + E \tag{S2}$$

The least square estimator of β is

$$\hat{\beta} = \left(X_{obs}^T \cdot X_{obs} \right)^{-1} \cdot X_{obs}^T \cdot Y_{obs} \tag{S3}$$

The residual sum of squares (RSS) is calculated by:

$$RSS = \left(Y_{obs} - X_{obs} \cdot \hat{\beta} \right)^T \cdot \left(Y_{obs} - X_{obs} \cdot \hat{\beta} \right) \tag{S4}$$

The mean squared error (MSE) is calculated by:

$$MSE = \frac{RSS}{DOF} \tag{S5}$$

where *DOF* is the degrees of freedom, which equals: *N - (k+1)*. Meanwhile, to account for the number of points (m) per data point, the standard error of the estimated regression coefficients is calculated by:

$$SE\left(\hat{\beta} \right) = diag\left(\sqrt{\left(\frac{X_{obs}^T \cdot X_{obs}}{m} \right)^{-1} \cdot \left(\frac{MSE}{m} \right)} \right)$$

Thus,

$$SE\left(\hat{\beta} \right) = diag\left(\sqrt{\left(X_{obs}^T \cdot X_{obs} \right)^{-1} \cdot MSE} \right) \tag{S6}$$

where *diag* are the diagonal elements of the matrix. The t-statistic of the estimated regression coefficients are calculated as,

$$t\text{-}statistic\left(\hat{\beta}_i \right) = \frac{\hat{\beta}_i}{SE\left(\hat{\beta}_i \right)} \text{, where i = 0,1,...k} \tag{S7}$$

The p-values of the estimated regression coefficients are calculated using this t-statistic and *DOF*. The *R* squared value of the fit is calculated as:

$$R^2 = 1 - \frac{RSS}{\left(Y_{obs} - \langle Y_{obs} \rangle \right)^T \cdot \left(Y_{obs} - \langle Y_{obs} \rangle \right)} \tag{S8}$$

where $\langle Y_{obs} \rangle$ represents the mean value of Y_{obs}. The adjusted R squared value of the fit is calculated as:

$$R_{adj}^2 = 1 - \left(1 - R^2\right)\left[\frac{N-1}{N-(k+1)}\right] \quad \text{(S9)}$$

For a matrix of predictor variables (X), the estimated response (Y_{est}) is calculated as:

$$Y_{est} = X \cdot \hat{\beta} \quad \text{(S10)}$$

Finally, the standard error of estimated response is calculated as:

$$SE\left(Y_{est}\right) = diag\left(\sqrt{X \cdot \left(\frac{X_{obs}^T \cdot X_{obs}}{m}\right)^{-1} \cdot X^T \cdot \frac{MSE}{m}}\right)$$

Thus,

$$SE\left(Y_{est}\right) = diag\left(\sqrt{X \cdot \left(X_{obs}^T \cdot X_{obs}\right)^{-1} \cdot X^T \cdot MSE}\right) \quad \text{(S11)}$$

To show that the p-values are not underestimated due to increasing the resampling size (m), we created the example vectors x = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] and y = [2, 1, 5, 6, 7, 14, 21, 15, 11, 20]. Assuming m = 5, we then created the vectors X = [x, x, x, x, x] and Y = [y, y, y, y, y], that consist of 5 replicate vectors of x and y, respectively.

Next, using the MATLAB function 'fitlm', we applied ordinary least square fitting on the data x vs y and the data X vs Y, respectively. The outcomes were two different p values (0.0014 and 1.3535×10^{-15}, respectively). This is expected, as the size of the X vs Y data is 5 times larger than the size of the x vs y data. Meanwhile, when applying the fitting method described above (instead of 'fitlm'), we obtained the same p value in both cases (0.0014), showing that the p value is not underestimated due to increased resample size.

Finally, it is worth noting that, to perform the linear fitting, one can also use other fitting methods, for example, total least squares, which in general would be a good option, as it uses orthogonal residuals to obtain the best fit. However, the two variables composing our data have different scales, potentially causing incorrect estimation of the residuals. To overcome this, additional normalization procedures would be required. Therefore, instead, we made use of ordinary least square fitting.

1.5 Relationship between single-cell RNA numbers (Microscopy) and total cell fluorescence (Flow cytometry)

Given the results in Figure 3, the statistics of the single-cell distribution of F/W values is strongly correlated with the statistics of the single-cell distribution of RNA numbers obtained by microscopy. However, this may not always be the case. In this section, we provide argument why one should always expect a linear relationship between the first three central moments of these distributions.

Let 'N_B' be the distribution of number of free-floating MS2d-GFPs in a given cell, while 'N_R' is the distribution of mRNAs per cell, 'BS' is the number of MS2d-GFPs bound to 1 mRNA (assumed to be constant) and, 'I' is the fluorescence intensity of one MS2d-GFP (also assumed to be constant). Then, the distribution of total cell fluorescence intensity (F_T) is:

$$N_B \cdot I + N_R \cdot BS \cdot I = F_T \quad \text{(S12)}$$

where, $F_B = N_B \cdot I$ is the distribution of background MS2d-GFP fluorescence intensity per cell (in the absence of spots), and $F_R = N_R \cdot BS \cdot I$ is the distribution of fluorescence intensity per cell from MS2d-GFP tagged RNAs alone.

Finally, we note that, below, we assume that F$_B$ and F$_R$ are independent, and investigate the relationship between the various central moments.

1.5.1 Mean

From (S12), given that I and BS are constants:

$$M\left(F_T\right) = M(N_B) \cdot I + M(N_R) \cdot BS \cdot I \qquad \text{(S13)}$$

From (S13):

$$M(N_R) = \frac{M\left(F_T\right)}{BS \cdot I} - \frac{M(N_B)}{BS} \qquad \text{(S14)}$$

Given (S14), $M(N_R)$ and $M\left(F_T\right)$ are linearly correlated with a slope $m = \frac{1}{BS \cdot I}$ and an intercept $c = -\frac{M(N_B)}{BS}$.

1.5.2 Variance

From (S12), given that *I* and *BS* are constants:

$$Var\left(F_T\right) = Var(N_B) \cdot I^2 + Var(N_R) \cdot BS^2 \cdot I^2 \qquad \text{(S15)}$$

$$Var(N_R) = \frac{Var\left(F_T\right)}{BS^2 \cdot I^2} - \frac{Var(N_B)}{BS^2} \qquad \text{(S16)}$$

From (S16), there is a linear relationship between $Var(N_R)$ and $Var\left(F_T\right)$ with a slope $m = \frac{1}{BS^2 \cdot I^2}$ and intercept $c = -\frac{Var(N_B)}{BS^2}$.

Importantly, from equation (S16), one can estimate the standard deviation of single-cell RNA numbers, as measured by F/W values from the flow-cytometry can be calculated by:

$$std(N_R) = \sqrt{Var(N_R)} \qquad \text{(S17)}$$

1.5.3 Third moment

Finally, also from (S12), given that *I* and *BS* are constants:

$$\mu_3\left(F_T\right) = \mu_3(N_B) \cdot I^3 + \mu_3(N_R) \cdot BS^3 \cdot I^3 \qquad \text{(S18)}$$

$$\mu_3(N_R)=\frac{\mu_3\left(F_T\right)}{BS^3\cdot I^3}-\frac{\mu_3(N_B)}{BS^3} \tag{S19}$$

From equation (S19), there is a linear relationship between the 3rd moment of RNA numbers per cell ($\mu_3(N_R)$) and the 3rd moment of total cell fluorescence per cell ($\mu_3\left(F_T\right)$) with a slope $m=\frac{1}{BS^3\cdot I^3}$ and intercept $c=-\frac{\mu_3(N_B)}{BS^3}$.

Finally, from equation (S16) and (S19), one can estimate the skewness of single-cell RNA numbers, as measured by F/W values from the flow-cytometry can be calculated by:

$$skew(N_R)=\frac{\mu_3(N_R)}{\left(Var(N_R)\right)^{\frac{3}{2}}} \tag{S20}$$

1.6 Estimation of the variability in F/W statistics using technical replicates of control cells

To estimate the variability between technical replicates, we make use of measurements performed using cells absent of the target gene (eight measurements shown in Figure 2).

First, to obtain the standard error of the mean, we used equation (S21):

$$SE(M)=\sqrt{\left(SE\left(M_{t\arg et}\right)\right)^2+\left(SE\left(M_{control}\right)\right)^2} \tag{S21}$$

where $SE\left(M_{t\arg et}\right)$ is the standard error of mean of F/W values of target cells (with the target gene), estimated as described in Methods, section 2.7 in the main manuscript, and $SE\left(M_{control}\right)$ is the standard deviation of the mean F/W from control cells from multiple conditions.

Meanwhile, the variability in the standard deviation (Sd) of F/W values is estimated from:

$$SE(Sd)=\frac{1}{2}\cdot\sqrt{\frac{\left(SE\left(Var_{t\arg et}\right)\right)^2+\left(SE\left(Var_{control}\right)\right)^2}{Var_{t\arg et}}} \tag{S22}$$

Here, $SE\left(Var_{t\arg et}\right)$ is the standard error of the variance of F/W values of target cells, estimated as described in Methods, section 2.7 in main manuscript, $SE\left(Var_{control}\right)$ is the standard deviation of the variance of F/W values of control cells from multiple conditions, and $Var_{t\arg et}$ is the variance of F/W values of target cells.

Finally, the variability in Skewness (S) of F/W values in target cells is estimated from:

$$SE(S)\approx\frac{\sqrt{\left(SE\left(\mu_3^{t\arg et}\right)\right)^2+\left(SE\left(\mu_3^{control}\right)\right)^2}}{(Sd_{t\arg et}+SE(Sd))^3} \tag{S23}$$

where $SE\left(\mu_3^{t\arg et}\right)$ is the standard error of the third moment of F/W values of target cells, estimated as in Methods, section 2.7 in main manuscript, $SE\left(\mu_3^{control}\right)$ is the standard deviation of the third moment of F/W values of control cells from multiple conditions, $Sd_{t\arg et}$ is the standard deviation of F/W values of target cells, and $SE(Sd)$ is the standard error of the standard deviation in F/W values of target cells, calculated using equation (S22).

In skewed distributions, we expect the variance and the third moment to be correlated. To compensate for this correlation, in equation (S23) we summed the standard deviation to its error in the denominator.

Finally, the variability in M, Sd and S of R/W values, is calculated as above, but replacing the F/W values with R/W values.

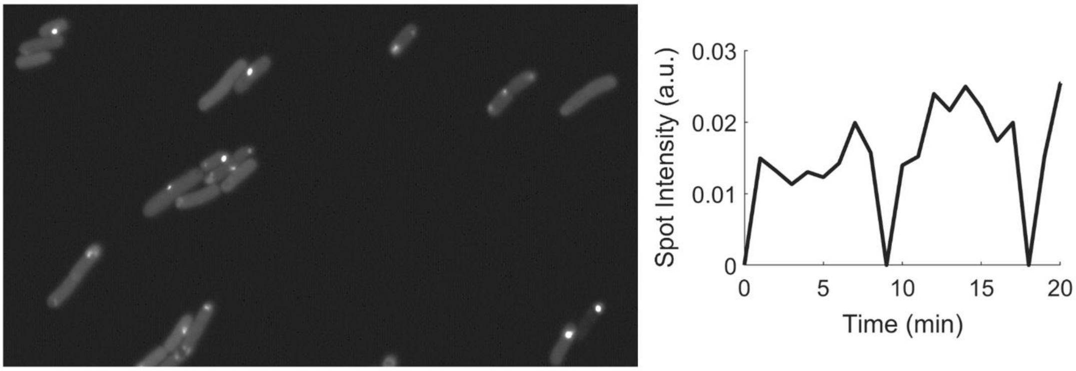
1.7 Estimation of single-cell RNA number statistics and standard error from F/W values from flow cytometry

To quantify mean RNA numbers per cell, we use the calibration line obtained from the mean RNA numbers in two induction conditions using microscopy and the mean F/W values in the same conditions when using flow cytometry. For each induction condition, we obtained a distribution of mean F/W values by bootstrapping. Next, using the calibration line and applying it to the distribution of mean F/W values, we estimated the distribution of mean single-cell RNA numbers, as measured by flow cytometry. In this process, we removed any negative values and calculated the mean and standard deviation of this distribution, which correspond to the mean and standard error of the estimated mean RNA numbers per cell, respectively.

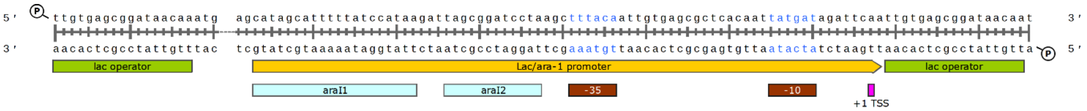
Similarly, to quantify the standard deviation (Sd) of RNA numbers per cell as measured by flow cytometry, we used the calibration line to map the variance of RNA numbers obtained by microscopy to the corresponding variance in F/W values obtained by flow cytometry. Next, we used bootstrapping to obtain the distribution of variances of F/W values in each condition. Next, using the calibration line, we estimated the distribution of variances of RNA numbers per cell, in each condition. Again, we removed any negative values. Finally, the distribution of the Sd values of RNA numbers per cell was obtained by calculating the square root of the values of the distribution of variances of RNA numbers per cell. Next, we calculated the mean and standard deviation of the distribution of the Sd of RNA numbers per cell, which are the mean and standard error of the estimated Sd of RNA numbers per cell, respectively.

To quantify the skewness of RNA numbers per cell, we used the calibration line for the variance and the third moment of RNA numbers (main manuscript, section 3.4). For each induction condition, by bootstrapping, we obtained distributions of variance and third moments of F/W values. Using their respective calibration lines, we calculated the variance and third moment of RNA numbers per cell for each value of the distribution of variances and third moments of F/W values. From this we obtained the distributions of variance and third moment of RNA numbers per cell. Negative variance values and the respective third moment’s values were removed. Next, the values of the distribution of skewness (S) of RNA per cell were obtained using equation S20. Finally, we calculated the mean and standard deviation of the distribution of values of S of RNA numbers per cell, which are the mean and standard error of the estimated S of RNA numbers per cell, respectively.

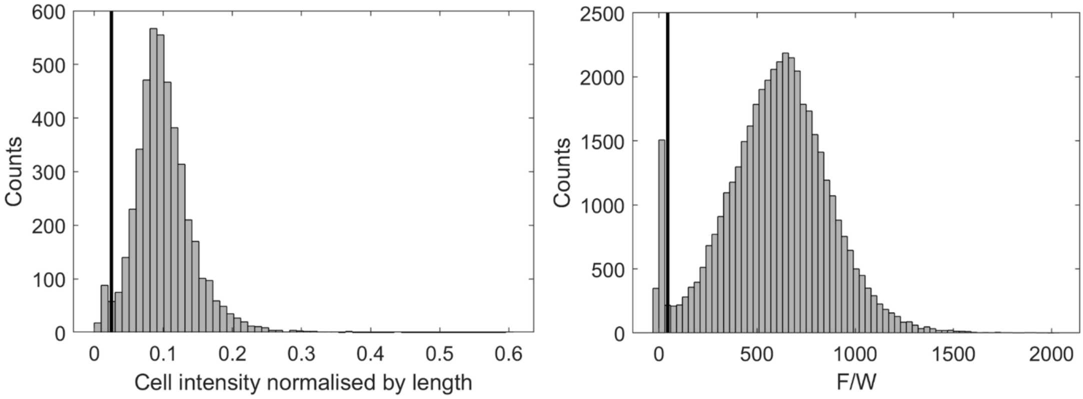
SUPPLEMENTARY FIGURES



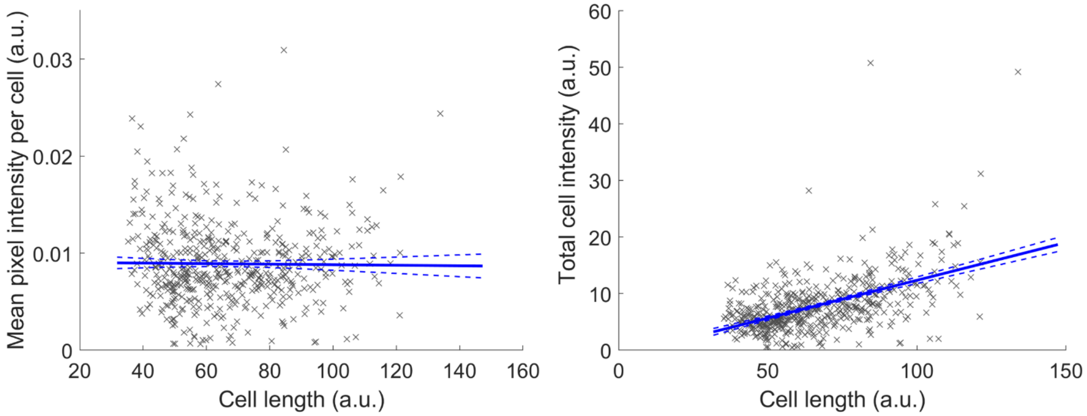
Supplementary Figure S1. Related to section 2.5 in main manuscript. Example microscopy images: (Left) Example image obtained by confocal microscopy. Visible are cells (due to the fluorescent background created by many free-floating MS2d-GFP) and spots within (MS2d-GFP tagged RNAs). (Right) Example of the observed fluorescence intensity over time of an MS2d-GFP RNA spot. Note the significant variability in intensity at two time points, due to its motion along the z axis during a time-lapse microscopy session (specifically, this spot left the focus plane at moments 9 min and 18 min).


5'
ttgtgagcggataacaaatg agcatagcattttt atccataagattagcggatcctaagctttacaattgtgagcgctcacaattatgatagattcaattgtgagcggataacaat
3'
3'
aacactcgcctattgtttac tcgtatcgtaaaaataggtattctaatcgcctaggattcgaaatgttaacactcgcgagtgttaatactatctaagttaacactcgcctattgtta
5'
lac operator
Lac/ara-1 promoter
lac operator
araI1
araI2
-35
-10
+1 TSS

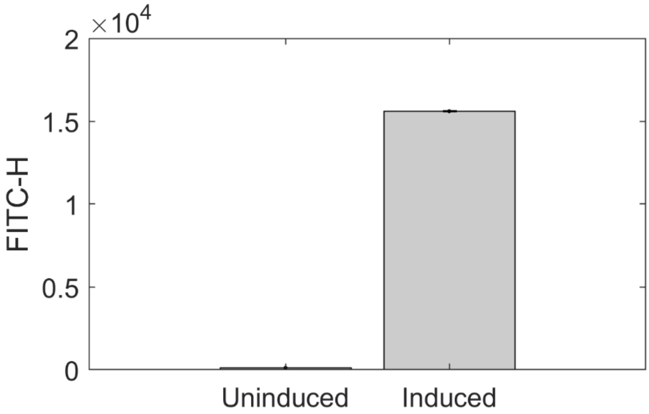
Supplementary Figure S2. Related to section 2.2 in main manuscript. Schematic representation of the target promoter sequence (marked in yellow) of the gene of interest. The -35 and -10 promoter elements are shown in light blue, while the transcription start site (+1 TSS) is marked by a small pink box. Operator sites are marked in green.



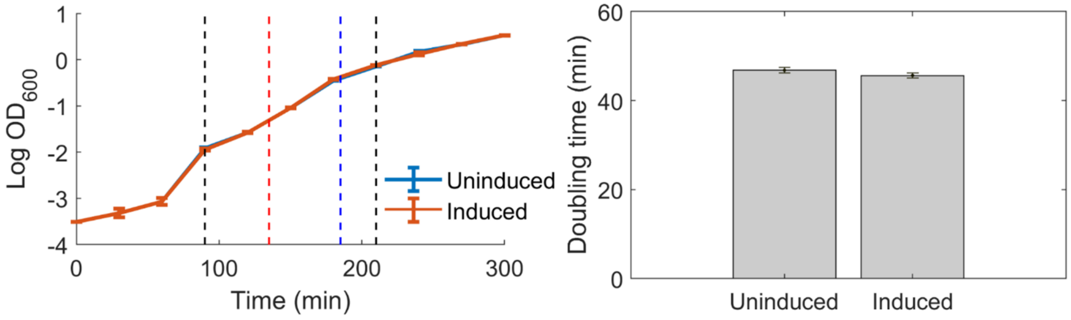
Supplementary Figure S3. Related to section 2.6 in main manuscript. (Left): Distribution of single-cell green fluorescence intensity normalized by cell length, as obtained by microscopy. We (manually) applied a minimum threshold (black vertical line, located at 0.025 in the x axis) to remove from the dataset the few cells not expressing sufficient MS2d-GFP for detecting target RNAs. Approximately 4500 cells were analyzed. (Right): Distribution of single-cell FITC-H values (F) divided by pulse Width (W), obtained by flow-cytometry (more than 40,000 cells were analyzed). To remove from the dataset the few cells lacking sufficient free-floating MS2d-GFP, we applied (manually) a minimum threshold (black vertical line, located at position 45 in the x axis). In addition, we removed the 0.01% or less cells with highest F/W values (not shown in the image).



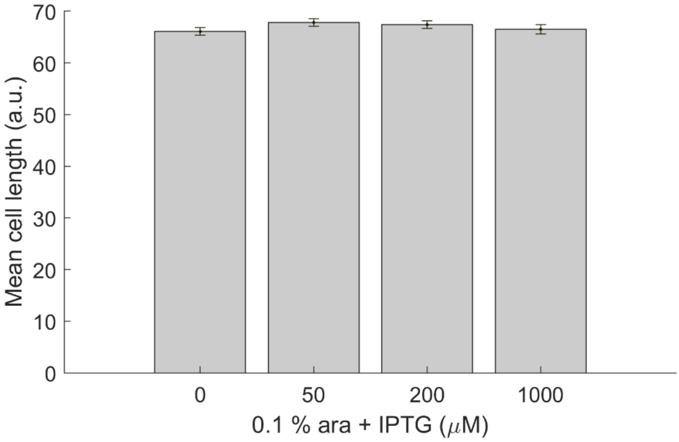
Supplementary Figure S4. Related to section 2.6 in main manuscript. (Left) Scatter plot of the single-cell fluorescence intensity (as estimated by the mean pixel intensity per cell) against the length of the major cell axis, as measured by microscopy. Approximately 500 cells were analyzed. The solid blue line is the best linear fit obtained by linear regression (R^2 value is 0.0002). The linear fit does not reject the null hypothesis that there is no linear correlation, at a significance level of 0.05 (p-value is 0.8). The blue dashed lines are the one standard uncertainty of the fitted line. (Right) Scatter plot of the total cell intensity of individual cells against the cell length along the major cell axis, as measured by microscopy. Approximately 500 cells were analyzed. The solid blue line is the best linear fit obtained by linear regression (R^2 value is 0.27). The null hypothesis that there is no linear correlation is rejected at a significance level of 0.05 (p-value is 10^{-36}). The blue dashed lines are the one standard uncertainty of the fitted line.



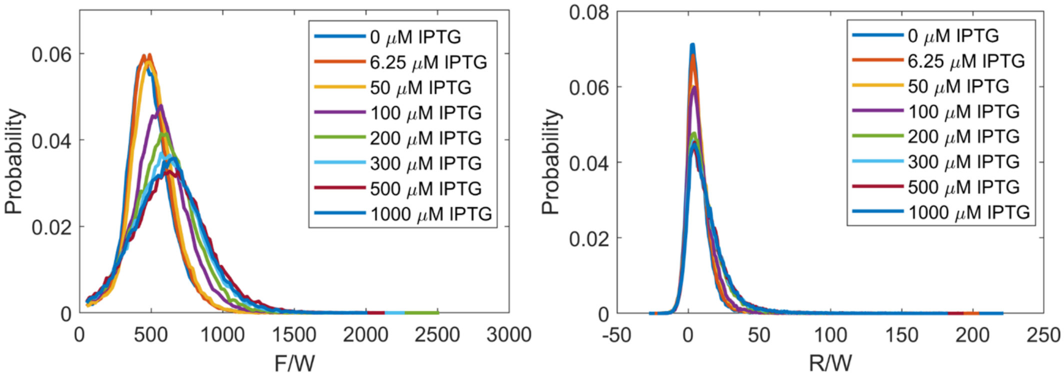
Supplementary Figure S5. Related to section 3.1 in main manuscript. Mean fluorescence intensity of cells, as measured by the FITC-H channel of the flow-cytometer. Measurements of the reporter gene expressing MS2d-GFP under the control of $P_{LtetO-1}$ were performed when induced (100 ng/µl aTc) and when not induced (0 ng/µl aTc). The (small) error bars denote the standard error of the mean. Approximately 40,000 cells were analyzed in each condition.



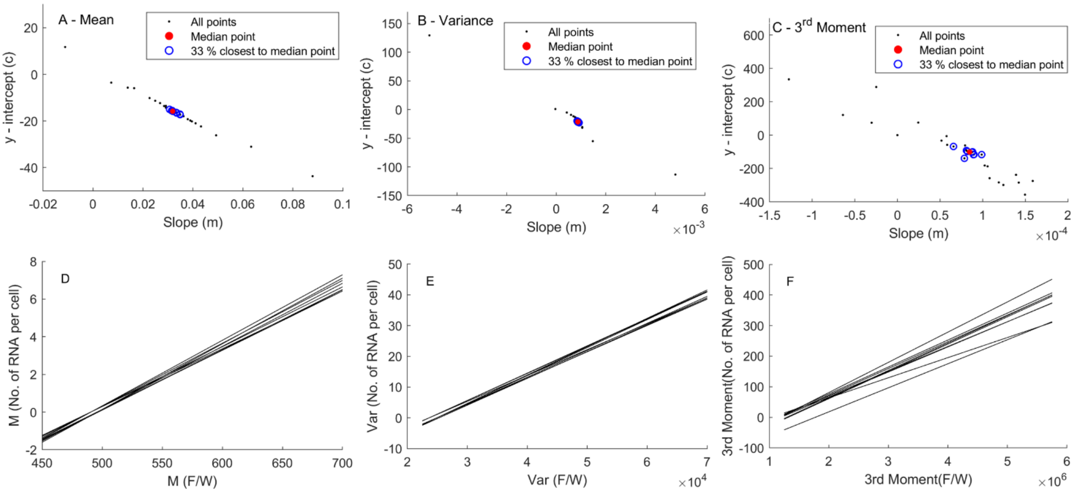
Supplementary Figure S6. Related to Figure 1 in main manuscript. (Left) Optical density (OD) curves of cell populations with the target plasmid in the presence (‘induced’, 1000 µM) and absence (‘uninduced’, 0 µM) of IPTG. Visibly, the two lines overlap. From a -80 °C glycerol stock, cells were streaked on LB agar plates containing 34 µg/ml chloramphenicol and 35 µg/ml kanamycin (Sigma-Aldrich, USA), and incubated overnight at 30 °C. From these plates, a single colony was picked and cultured overnight, with agitation (250 rpm), in LB medium supplemented with the appropriate concentration of antibiotics. From the overnight culture, cells were diluted to an initial OD_{600} of 0.03 in fresh LB medium, and grown at 37 °C. Next, the OD_{600} was measured every 30 minutes for 5 hours. At OD_{600} 0.3, aTc was added to induce the expression of the reporter, MS2d-GFP. Additionally, in cells where the target gene was induced, L-Arabinose was added at the same time as aTc (vertical red line). After 50 mins, IPTG was added to cells where the target gene was induced (vertical blue dashed line). Mean doubling time was estimated for the time period between 90 and 210 minutes (marked by two vertical dashed black lines). (Right) Mean doubling times, as estimated from the measurements in the left figure. Error bars (small) denote the standard error of the mean.



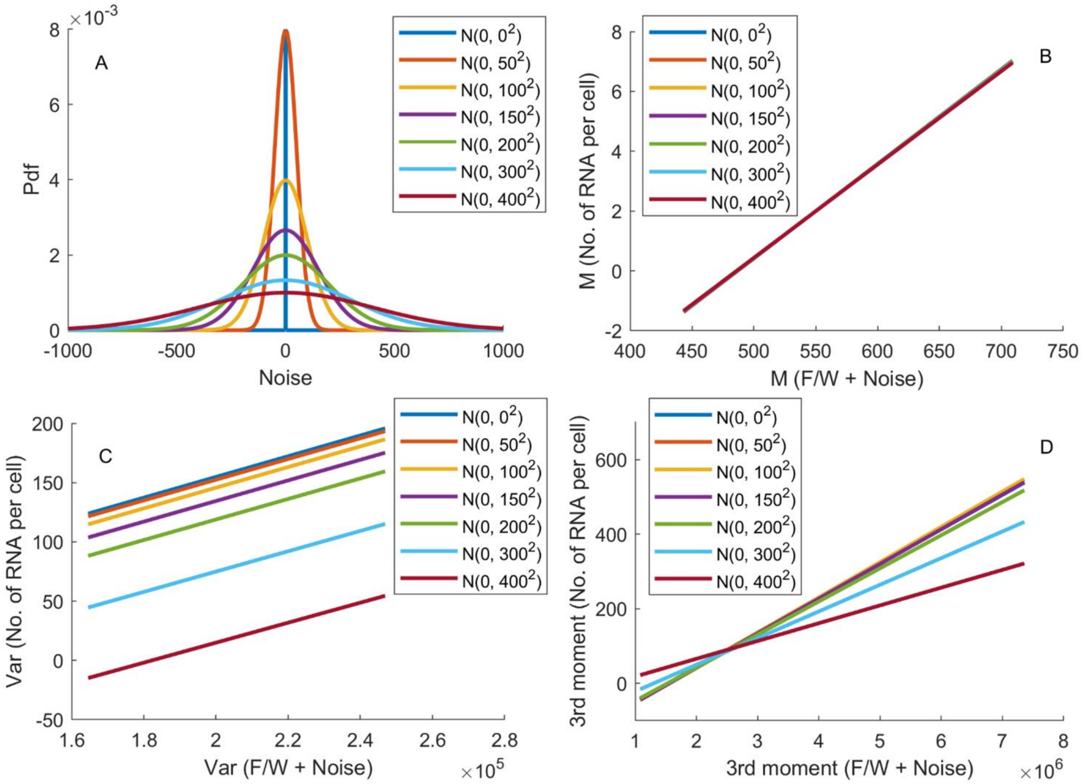
Supplementary Figure S7. Related to Section 3.1 in main manuscript. Mean length of the major axis of cells subject to various IPTG concentrations, as measured by microscopy. On average, approximately 500 cells were analyzed in each condition. The small error bars denote the standard error of the mean. Aside from IPTG, cells were also subjected to aTc and L-Arabinose.



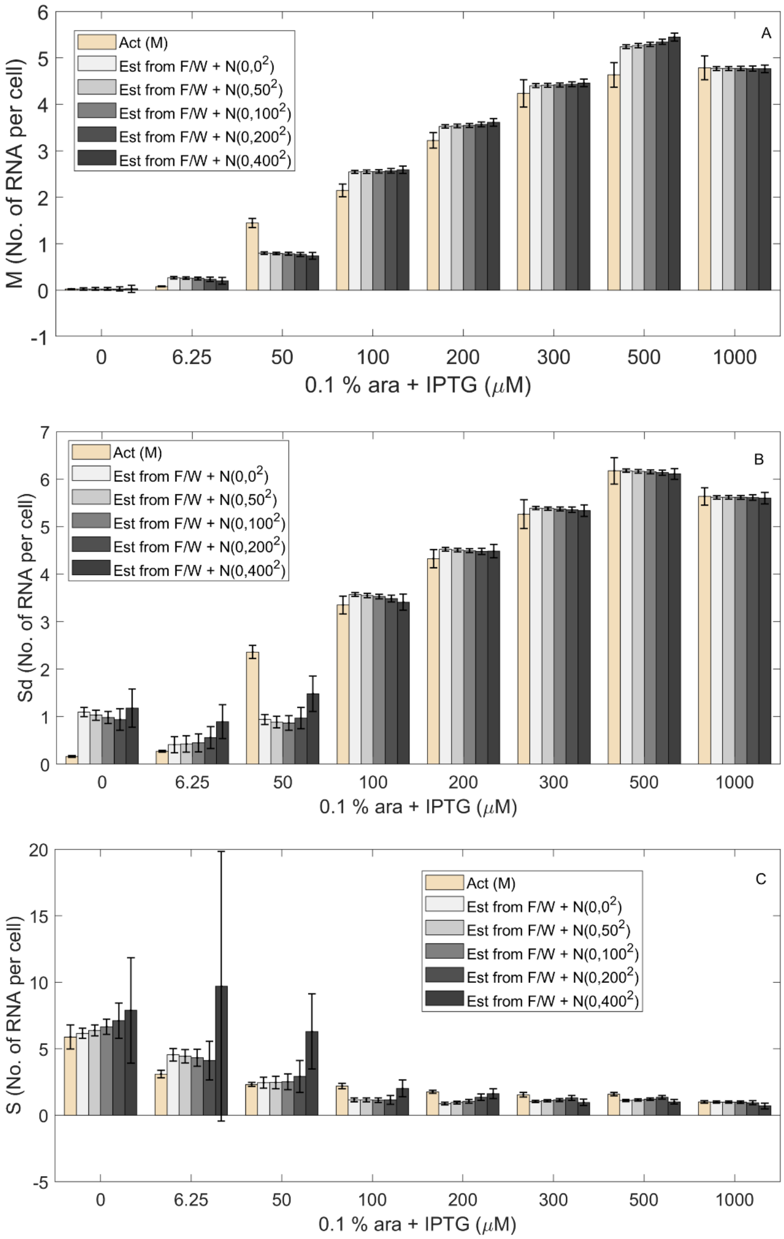
Supplementary Figure S8. Related to Figures 3D-F and 5 in main manuscript. (Left) Probability of single-cell F/W values (bin width = 20 units) for each condition differing in IPTG concentration, after gating (see Section 2.6 in main manuscript). In addition, we removed the 0.01% or less cells with highest F/W values (not shown in the image). (Right) Corresponding probability of single-cell R/W values (bin width = 1 unit), for each condition. In addition, we removed the 0.01% or less cells with highest R/W values (not shown in the image). Approximately 40,000 cells were analyzed in each condition.



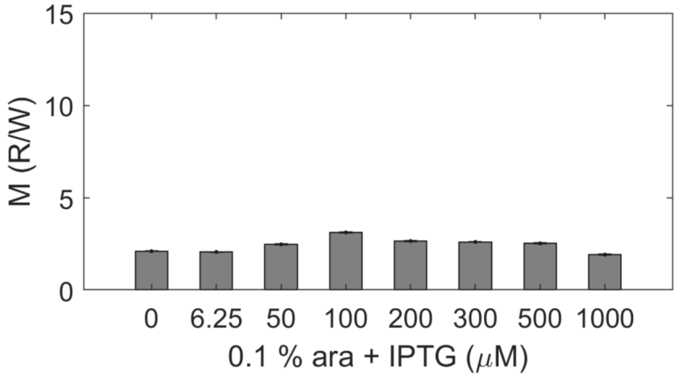
Supplementary Figure S9. Related to Figure 7 in the main manuscript. Slope vs intercept of the calibration lines between single-cell distributions of F/W and single cell distribution of RNA numbers for (A) Mean, (B) Variance and (C) 3rd moment. The small black dots correspond to each of all possible calibration lines (data from Figures 4A-C in main manuscript). The red points are the median point, whose coordinates are the median of the slopes versus the median of the intercepts of all possible calibration lines, respectively. The blue circles identify the black dots that are closer to the median point (the 33 % closest are encircled). Calibration lines of the Mean (D), Variance (E), and 3rd Moment (F) of the single-cell distributions of F/W values and single-cell distributions of RNA numbers. Each line corresponds to one of the black dots identified by a blue circle.



Supplementary Figure S10. (A) Gaussian noises with mean of 0 and increasingly higher std, which ranges from 0 to 400. (B) Best linear fit between the Mean of noise corrupted F/W (F/W from flow cytometry + gaussian noise) and the Mean of single-cell RNA numbers (microscopy data), (C) Best linear fit between the Var of noise corrupted F/W and the Var of single-cell RNA numbers (microscopy data), and (D) Best linear fit between the 3rd Moment of noise corrupted F/W and the 3rd Moment of single-cell RNA numbers (microscopy data).



Supplementary Figure S11. (A) Mean single-cell RNA numbers estimated from noise corrupted, empirical F/W distributions (i.e. with added gaussian noise). (B) Standard deviation of single-cell RNA numbers estimated from noise corrupted F/W distributions, using microscopy data in 6.25 and 1000 µM IPTG conditions for calibration. (C) Skewness of single-cell RNA numbers estimated from noise corrupted F/W, using microscopy data in 50 and 1000 µM IPTG conditions for calibration. The first light yellow bar is the actual value obtained from microscopy data. The rest of the other bars are estimated values from noise corrupted F/W having different levels of gaussian noise 0, 50, 100, 200, 400, respectively. In all cases, the light-yellow bar corresponds to actual (empirical) single-cell RNA numbers statistic as measured by microscopy, for comparison.



Supplementary Figure S12. Mean PETexasRed-H values, normalized by Pulse Width (R/W), of cells carrying only the reporter gene, at various IPTG concentrations (x-axis). The black error bars, barely visible, are the standard error of the mean (Methods, section 2.7). The scale of the y-axis is set to be identical to Figure 5A in the main manuscript, to facilitate comparison.

SUPPLEMENTARY TABLES

Supplementary Table S1. Related to Figure 4. Estimation of the goodness of fit of the linear models to the data in Figure 4 in the main manuscript. Linear fits were done to the scatter plots between the single-cell RNA numbers as measured by microscopy (No. of RNA per cell) and the F/W values obtained by flow-cytometry, for each induction level, using a linear regression fitting method, described in Supplementary Section 1.4. Shown are the adjusted R^2 values and the p-values of the F-statistics versus the constant model. 'M' stands for mean and 'Var' stands for variance.

(No. of RNA per cell) vs (F/W)	R^2	p value
M	0.96	1.8×10^{-5}
Var	0.95	1.9×10^{-5}
3rd moment	0.77	2.5×10^{-3}

Supplementary Table S2. Related to Figure 6. Estimation of the goodness of fit of Linear models to the data in Figure 6 in main manuscript. Fits were done to the scatter plots between R/W and F/W values obtained by flow-cytometry, for each induction level. Fits were obtained by the linear regression fitting method described in Supplementary Section 1.4. Shown are the adjusted R^2 values and the p-values of the F-statistics versus the constant model. 'M' stands for mean, 'Sd' stands for standard deviation, and 'S' stands for skewness.

(R/W) vs (F/W)	R^2	p value
M	0.93	6.3×10^{-5}
Sd	0.87	4.5×10^{-4}
S	0.86	5.2×10^{-4}

Supplementary Table S3. Related to Figure 7. Estimation of goodness of fit of the best linear fit between empirical and estimated values of mean (M), standard deviation (Sd) and skewness (S) of the single-cell distribution of RNA numbers. Shown are the *p* values for the slope and the intercept with the y-axis, assuming the null hypothesis that the empirical and estimated values are the same. In all cases, the test does not reject the null hypothesis that they are the same, at a significance level of 0.05.

	p value	
Empirical vs Estimated	Slope	Intercept
M (No. of RNA per cell)	0.52	0.95
Sd (No. of RNA per cell)	0.70	0.71
S (No. of RNA per cell)	0.53	0.46

Supplementary References

Alexopoulos, E.C., 2010. Introduction to multivariate regression analysis. Hippokratia 14, 23–8.

Breiman, L., Friedman, J., Olshen, R. A., Stone, C. J., 1984. Classification and Regression Trees. Chapman and Hall, CRC.

Carpenter, J., Bithell J., 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. 19, 1141–1164. https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F

DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence intervals. Stat. Sci. 11, 189–228. https://doi.org/10.1214/ss/1032280214

Golding, I., Paulsson, J., Zawilski, S.M., Cox, E.C., 2005. Real-time kinetics of gene activity in individual bacteria. Cell 123, 1025–1036. https://doi.org/10.1016/j.cell.2005.09.031

Häkkinen, A., Muthukrishnan, A.B., Mora, A., Fonseca, J.M., Ribeiro, A.S., 2013. CellAging: A tool to study segregation and partitioning in division in cell lineages of Escherichia coli. Bioinformatics 29, 1708–1709. https://doi.org/10.1093/bioinformatics/btt194

Häkkinen, A., Ribeiro, A.S., 2015. Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data. Bioinformatics 31, 69–75. https://doi.org/10.1093/bioinformatics/btu592

Kandavalli, V.K., Tran, H., Ribeiro, A.S., 2016. Effects of σ factor competition are promoter initiation kinetics dependent. Biochim. Biophys. Acta - Gene Regul. Mech. 1859, 1281–1288. https://doi.org/10.1016/j.bbagrm.2016.07.011

Lloyd-Price, J., Startceva, S., Kandavalli, V., Chandraseelan, J.G., Goncalves, N., Oliveira, S.M.D., Häkkinen, A., Ribeiro, A.S., 2016. Dissecting the stochastic transcription initiation process in live Escherichia coli. DNA Res. 23, 203–214. https://doi.org/10.1093/dnares/dsw009

Mäkelä, J., Kandavalli, V., Ribeiro, A.S., 2017. Rate-limiting steps in transcription dictate sensitivity to variability in cellular components. Sci. Rep. 7, 1–10. https://doi.org/10.1038/s41598-017-11257-2

Mora, A.D., Vieira, P.M., Manivannan, A., Fonseca, J.M., 2011. Automated drusen detection in retinal images using analytical modelling algorithms. Biomed. Eng. Online 10, 59. https://doi.org/10.1186/1475-925X-10-59

Oliveira, S.M.D., Häkkinen, A., Lloyd-Price, J., Tran, H., Kandavalli, V., Ribeiro, A.S., 2016. Temperature-Dependent Model of Multi-step Transcription Initiation in Escherichia coli Based on Live Single-Cell Measurements. PLoS Comput. Biol. 12, 1–18. https://doi.org/10.1371/journal.pcbi.1005174

Oliveira, S.M.D., Goncalves, N.S.M., Kandavalli, V.K., Martins, L., Neeli-Venkata, R., Reyelt, J., Fonseca, J.M., Lloyd-Price, J., Kranz, H., Ribeiro, A.S., 2019. Chromosome and plasmid-borne P LacO3O1 promoters differ in sensitivity to critically low temperatures. Sci. Rep. 9, 1–15. https://doi.org/10.1038/s41598-019-39618-z

Queimadelas, C., Rodrigues, J., Muthukrishnan, A.B., Mora, A., Ribeiro, A.S., Fonseca, J.M., 2012. Segmentation and tracking of Escherichia coli expressing tsr-venus proteins from combined DIC/Fluorescence images. In fifth International Conference on MEDSIP. Liverpool, UK. https://doi.org/10.13140/2.1.3835.3924

Startceva, S., Kandavalli, V.K., Visa, A., Ribeiro, A.S., 2019. Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression. Biochim. Biophys. Acta - Gene Regul. Mech. 1862, 119–128. https://doi.org/10.1016/j.bbagrm.2018.12.005

Tran, H., Oliveira, S.M.D., Goncalves, N., Ribeiro, A.S., 2015. Kinetics of the cellular intake of a gene expression inducer at high concentrations. Mol. Biosyst. 11, 2579–2587. https://doi.org/10.1039/c5mb00244c

PUBLICATION
II

The transcription factor network of E. coli steers global responses to shifts in RNAP concentration

B Almeida, V Chauhan*, MNM Bahrudeen*, S Dash*, V Kandavalli, A Häkkinen, J Lloyd-Price, CSD Palma, ISC Baptista, A Gupta, J Kesseli, E Dufour, O-P Smolander, M Nykter, P Auvinen, HT Jacobs, SMD Oliveira, and AS Ribeiro. *Equal contributions

Nucleic Acids Research 50(12), 6801-6819.
DOI: 10.1093/nar/gkac540.

Published online 24 June 2022 *Nucleic Acids Research, 2022, Vol. 50, No. 12* ***6801–6819***
https://doi.org/10.1093/nar/gkac540

The transcription factor network of *E. coli* steers global responses to shifts in RNAP concentration

Bilena L.B. Almeida[1,*], Mohamed N. M. Bahrudeen[1,†], Vatsala Chauhan[1,†], Suchintak Dash[1,†], Vinodh Kandavalli[2], Antti Häkkinen[3], Jason Lloyd-Price[4], Palma S.D. Cristina[1], Ines S.C. Baptista[1], Abhishekh Gupta[5], Juha Kesseli[6], Eric Dufour[7], Olli-Pekka Smolander[8,9], Matti Nykter[6], Petri Auvinen[9], Howard T. Jacobs[10], Samuel M.D. Oliveira[11] and Andre S. Ribeiro[1,12,*]

[1]Laboratory of Biosystem Dynamics, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland, [2]Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden, [3]Research Program in Systems Oncology, Research Programs Unit, Faculty of Medicine, University of Helsinki, FI-00014 Helsinki, Finland, [4]Google LLC, 111 8th Ave, New York, NY 10010, USA, [5]Center for Quantitative Medicine and Department of Cell Biology, University of Connecticut School of Medicine, 263 Farmington Av., Farmington, CT 06030-6033, USA, [6]Prostate Cancer Research Center, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland; Tays Cancer Center, Tampere University Hospital, Tampere, Finland, [7]Mitochondrial bioenergetics and metabolism, BioMediTech, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland, [8]Department of Chemistry and Biotechnology, Tallinn University of Technology, Tallinn, Estonia, [9]Institute of Biotechnology, University of Helsinki, Viikinkaari 5D, 00790 Helsinki, Finland, [10]Faculty of Medicine and Health Technology, FI-33014 Tampere University, Finland; Department of Environment and Genetics, La Trobe University, Melbourne, Victoria 3086, Australia, [11]Department of Electrical and Computer Engineering, Boston University, Boston, MA, USA and [12]Center of Technology and Systems (CTS-Uninova), NOVA University of Lisbon, 2829-516 Monte de Caparica, Portugal

Received March 07, 2022; Revised June 02, 2022; Editorial Decision June 06, 2022; Accepted June 14, 2022

ABSTRACT

The robustness and sensitivity of gene networks to environmental changes is critical for cell survival. How gene networks produce specific, chronologically ordered responses to genome-wide perturbations, while robustly maintaining homeostasis, remains an open question. We analysed if short- and mid-term genome-wide responses to shifts in RNA polymerase (RNAP) concentration are influenced by the *known* topology and logic of the transcription factor network (TFN) of *Escherichia coli*. We found that, at the gene cohort level, the magnitude of the single-gene, mid-term transcriptional responses to changes in RNAP concentration can be explained by the absolute difference between the gene's numbers of activating and repressing input transcription factors (TFs). Interestingly, this difference is strongly positively correlated with the number of input TFs of the gene. Meanwhile, short-term responses showed only weak influence from the TFN. Our results suggest that the global topological traits of the TFN of *E. coli* shape which gene cohorts respond to genome-wide stresses.

INTRODUCTION

Gene regulatory networks (GRNs) receive, process, act upon and send out information, while being robust to random fluctuations. How signals targeting one to a few genes are processed is relatively well understood (1,2). Meanwhile, many cellular environments fluctuate (sometimes unpredictably) in nutrient availability, pH, temperature, salts, community of other cells or species they live with, etc., which may cause genome-wide stresses. We investigate how GRNs produce chronologically ordered responses to genome-wide perturbations, while robustly maintaining homeostasis.

Evidence suggests that genome-wide stresses initially perturb hundreds to thousands of genes (3) but are quickly

[*]To whom correspondence should be addressed. Tel: +358 2945211; Email: andre.sanchesribeiro@tuni.fi
Correspondence may also be addressed to Bilena L.B. Almeida. Tel: +358 2945211; Email: bilena.limadebritoalmeida@tuni.fi
[†]The authors wish it to be known that, in their opinion, the second, third and fourth authors should be regarded as Joint Second Authors.

processed. As a result, after a transient period, only specific gene cohorts of tens to a few hundred genes (4,5) (usually sharing common feature(s)) participate in the responsive short-, mid- and long-term transcriptional programs (6). For example, when *Escherichia coli* suffers a cold shock, a specific cohort exhibits a fast, short-term response (∼ 70 genes), while another has a longer-term response (∼ 35 genes). Other genes remain relatively passive (7,8). Since cells exhibit predictable, temporally ordered, beneficial phenotypic changes, these response programs have likely been positively selected during evolution.

It has been shown that global regulators (GRs) (9–12), DNA supercoiling (13) and small RNAs (sRNAs) (14), among other, can select large cohorts of stress-specific, responsive genes. It was also reported that 60–90% of *E. coli* genes respond to changing growth conditions following a constant global scaling factor (15). Further, measurements using fluorescent reporters and small circuits (16,17) showed that the effects of RNA polymerase (RNAP), and other GRs, can be separate from the effects of input transcription factors (TFs) during genome-wide responses. Nevertheless, establishing whether and how the global topology and logic of transcription factor networks (TFNs) affect genome-wide responses remains challenging (18), despite successes in establishing that gene-gene interactions generate gene-gene correlated dynamics (19–22).

To investigate the influence of the topology and logic of TFNs on large transcriptional programs, we study what occurs following genome-wide perturbations. We consider that, in *E. coli*, the concentration of the key genome-wide regulator, the RNAP, naturally differs with medium composition (23). Also, it is well established how transcription kinetics differs according to RNAP concentration (24,25). We expect that these changes can have genome-wide effects. We thus obtain increasingly dilute media to alter systematically and rapidly (26,27) the abundance of RNAP (illustrated in Figure 1A), and measure the genome-wide, short- and mid-term changes in transcript abundances.

Decreases in RNAP abundance should, in the short-term, cause quick genome-wide decreases in transcription rates, and thus in RNA abundances (Figure 1B$_1$, Supplementary Results section *Expected effects of shifting RNA polymerase concentration on a gene's transcription dynamics*). Such shifts, likely diverse in magnitudes, should then cause downshifts in the corresponding protein abundances. Thus, in the case of input TFs, their 'output' genes will, later on, be affected as well (Figure 1B$_2$), causing further (here named 'mid-term') changes in RNA abundances (Figure 1B$_4$). Meanwhile, the short-term changes most likely are only affected by the genes' individual features affecting their responsiveness to RNAP, since the protein abundances have not yet changed significantly.

We focus on the mid-term changes in single-gene expression levels. Specifically, we hypothesize that, in the short-term, a reduction in RNAP concentration will be followed by a reduction in most genes' expression levels. Subsequently, in the mid-term, genes with one input TF will have their RNA abundance either further decreased or, instead, increased, depending on whether their input TF is an activator or a repressor, respectively. Meanwhile, the average magnitude of the mid-term response of genes with multiple input TFs should be correlated with the difference between the numbers of their activator and repressor input TFs. This difference is here named 'bias' in the regulatory effect (i.e. activation or repression) of the input TFs of a gene. A schematic and a predictive model of this regulatory mechanism of the genome-wide single-gene mid-term responses are shown in Figure 1B and C, respectively.

At the single-gene level, the magnitude of the mid-term changes in RNA abundances should be influenced by the magnitude of the shift in RNAP and in input TFs concentrations, as well as by the specifics features of each gene and its input TFs (bindings affinities, etc.). However, many of these features are largely unknown. As such, here we only study empirically if the average responses of gene cohorts can be explained by the mean bias in the regulatory effect of their input TFs, along with the magnitude of the shift in RNAP concentration (Figure 1C$_2$). In detail, we interpret the data on the genome-wide kinetics based on the information on the TFN structure (logic and topology) (Figure 1B$_3$).

We use *E. coli* to validate this hypothesis since its gene expression mechanisms have been largely dissected and the kinetics of transcription, translation, and RNA and protein degradation are well known (26,32,33). Also, its TFN is extensively mapped, with RegulonDB (34) informing on ∼ 4700 TF interactions between ∼ 4500 genes (and on their activating or repressing regulatory roles). Consequently, since we know the regulatory network *a priori*, instead of using the data on gene expression for network inference, we use it solely to quantify the genes' responsiveness with respect to the TFN topology and logic. We then investigate whether the mid-term responsiveness to shifting RNAP concentration is in accordance with the *presently known* topology and logic of the TFN, as hypothesized (Figure 1B$_1$-B$_4$). Supplementary Table S1 has a description of the variables used throughout the manuscript.

In summary, here we show that changes in RNAP concentration due to medium dilutions are followed by short- and mid-term genome-wide changes in RNA abundances (obtained by time-lapsed RNA-seq). These RNA changes, globally, cannot be explained by potential influences from GRs, σ factors, (p)ppGpp, non-coding RNAs, or post-translation regulators (e.g. due to lack of correlation with their output genes, lack of RNA changes, etc.). Instead, we find that genes directly linked by TF interactions show correlated changes in RNA abundances. Further, the average magnitude of the mid-term responses of gene cohorts can be explained by the mean bias in the regulatory effect of their genes' input TFs (obtained from RegulonDB and shown to be correlated to the number of input TFs, K_{TF}). Also influential is the magnitude of the shift in RNAP concentration (obtained by flow-cytometry and western blot) and operons and transcription units (TUs) organization (from RegulonDB). Meanwhile, short-term responses are less influenced by TFs. Finally, we show the same phenomenon for opposite shifts in medium concentration that cause the same shifts in RNAP concentration.

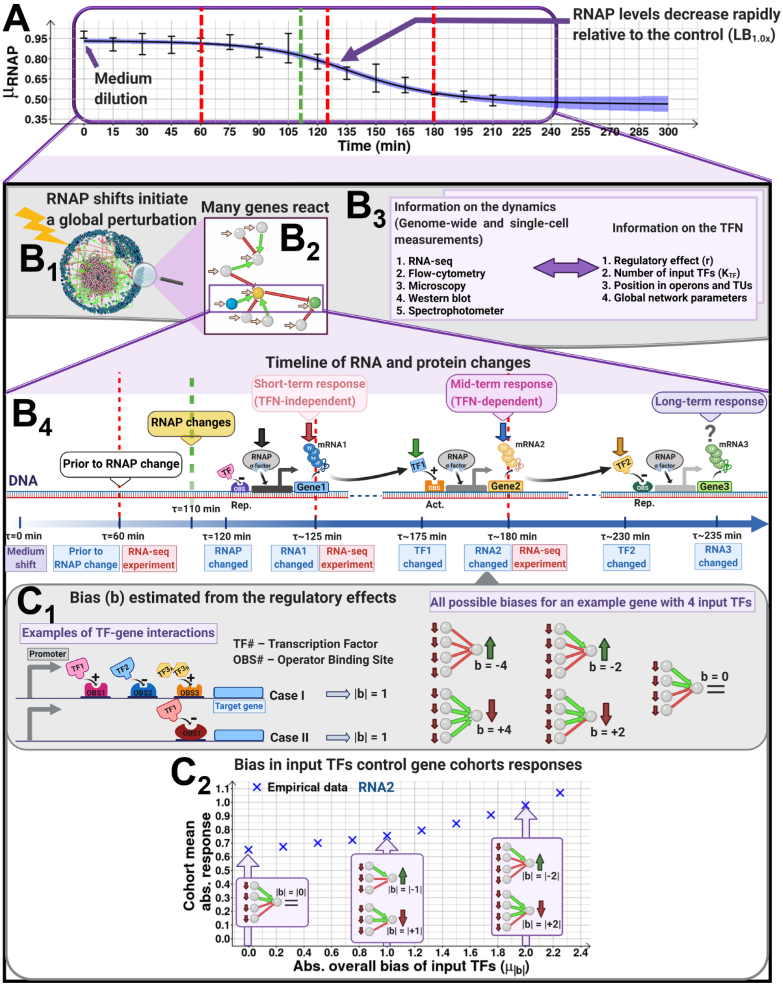


Figure 1. Expected short- and mid-term effects of quick downshifts of the RNAP abundance on the TFN of *E. coli*. (**A**) Example changes in mean RNAP (μ_{RNAP}) and 68% CB (shadow) relative to control ($LB_{1.0x}$) after diluting the medium ($LB_{0.5x}$). Vertical red lines mark when the transcriptome measurements at 60, 125 and 180 min. Given the RNAP levels and the kinetics of RNA and protein abundances, these moments are named 'prior to RNAP changes' and 'short-', and 'mid-term' changes in RNA abundances. (**B$_1$**) Known TF-gene interactions (red and green lines, if repressing and activating, respectively) and genes with (pink) and without (blue) input TFs of *E. coli*. (**B$_2$**) Schematics of the expected effects of a local topology of activating (green) and repressing (red) input TFs on mid-term responses. Genes (balls) are coloured (blue, yellow, and green) according to the events in B$_4$. (**B$_3$**) Data collected on the genome-wide kinetics as well as data collected on the TFN structure. (**B$_4$**) Following a medium dilution, intracellular RNAP concentrations (black arrow) decrease after a time lag, and RNA abundances (red arrow) will decrease accordingly. Compared to when at ∼ 0 min, the RNAP at ∼ 120 min and corresponding RNAs at ∼ 125 min should be lower (25,28). Given translation times (∼ 50 min (29–31)), at ∼ 175 min, the protein abundances, including input TFs, coded by the perturbed RNAs (green arrow) should differ as well. Fluctuations in these input TFs abundances will then propagate to nearest neighbour 'output' genes, further shifting their RNA abundances (blue arrow) depending on whether the input TF is an activator or a repressor. Finally, the yellow arrow represents (not measured) long-term changes (∼ 230 min or longer). We performed RNA-seq at ∼ 60 min (prior to RNAP changes), ∼ 125 min (short-term RNA changes), and ∼ 180 min (mid-term RNA changes, affected by input TFs). Finally, the green dashed line marks when the RNAP level already differs significantly from the control (see example Figure 1A). (**C$_1$**) Predictive model of the expected biases in sets of input TFs of individual genes. Considering TF-gene interactions as either repressions (regulatory effect of -1) or activations (regulatory effect of +1), the overall effect of a set of input TFs during these stresses should be predictable from the sum of the regulatory effect of the input TFs, named 'bias', (*b*). Regulatory effects obtained from RegulonDB. (**C$_2$**) Example average response ($\mu_{|LFC|}$ from RNA-seq) at 180 min of gene cohorts with a given $\mu_{|b|}$ and how they are expected to relate to the biases. Figures created with BioRender.com.

MATERIALS AND METHODS

Bacterial strains, media, growth conditions and curves, and intracellular concentrations

We used wild type MG1655 cells as a base strain to study the transcriptome. In addition, we used an RL1314 strain with RpoC endogenously tagged with GFP (generously provided by Robert Landick) to measure RNAP levels, and 20 YFP fusion strains with genes endogenously tagged with the YFP coding sequence (25) to measure single-cell protein levels (Supplementary Table S4). Further, we used a strain carrying an rpoS::mCherry gene (generously provided by James Locke), shown to track RpoS (35), to measure RpoS levels. In addition, we measured the protein levels of the spoT gene, which is one of the genes responsible for (p)ppGpp synthesis (3), using the YFP fusion library. Finally, we measured single-cell levels of the crl gene using a low-copy plasmid fusion library of fluorescent (GFP) reporter strain (36).

From glycerol stocks (at –80 °C), cells were streaked on lysogeny broth (LB) agar plates with antibiotics and kept at 37 °C overnight. Next, a single colony was picked, inoculated into fresh LB medium and, kept at 30 °C overnight with appropriate antibiotics and aeration at 250 rpm. From overnight cultures (ONC), cells were diluted to 1:1000 in tailored LB media (see below) with antibiotics, incubated at 37 °C with aeration, and allowed to grow until reaching an optical density of ≈ 0.4 at 600 nm (OD_{600}).

Using this protocol, to attain cells with different intracellular RNAP concentration, starting from LB, we used tailored media, denoted as '$LB_{1.0x}$', '$LB_{0.75x}$', '$LB_{0.5x}$', '$LB_{0.25x}$', '$LB_{1.5x}$', '$LB_{2.0x}$' and '$LB_{2.5x}$' specifically, as in (26). Their composition for 100 ml (pH of 7.0) are, respectively: ($LB_{1.0x}$) 1 g tryptone, 0.5 g yeast extract and 1 g NaCl; ($LB_{0.75x}$) 0.75 g tryptone, 0.375 g yeast extract and 1 g NaCl; ($LB_{0.5x}$) 0.5 g tryptone, 0.25 g yeast extract and 1 g NaCl; and ($LB_{0.25x}$) 0.25 g tryptone, 0.125 g yeast extract and 1 g NaCl; ($LB_{1.5x}$) 1.5 g tryptone, 0.75 g yeast extract and 1 g NaCl; ($LB_{2.0x}$) 2 g tryptone, 1 g yeast extract and 1 g NaCl; ($LB_{2.5x}$) 2.5 g tryptone, 1.25 g yeast extract and 1 g NaCl.

To measure cell growth curves and rates, ONC of the RL1314 strain were diluted to an initial optical density at 600 nm (OD_{600}) of ≈ 0.05 into independent fresh media ($LB_{1.0x}$, $LB_{0.75x}$, $LB_{0.5x,}$ $LB_{0.25x}$, $LB_{1.5x}$, $LB_{2.0x}$ and $LB_{2.5x}$). The cultures were aliquoted in a 24-well flat bottom transparent plate and incubated at 37 °C with continuous shaking in a Biotek Synergy HTX Multi-Mode Reader. Growth was monitored every 10 min for 10 h.

Finally, the RNAP and σ factor concentrations were estimated by measuring their average abundances from single-cell fluorescence levels (of RpoC-GFP and rpoS::mCherry, respectively, by flow-cytometry). Next, we divided that abundance by the mean single-cell area (from phase-contrast microscopy images of cell populations in the same condition), used as a proxy for cell volumes, to obtained concentrations in fluorescence intensity per pixel (not shown in the figures). In all cases, for each condition, we obtained images from, on average, 2500 cells (from three biological replicates). The images are provided in Supplementary Data.

Microscopy

To measure single-cell RNAP levels, ONC RL1314 cells were pre-inoculated into $LB_{1.0x}$, $LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$ media. Upon reaching mid-exponential growth phase, cells were pelleted by quick centrifugation (10000 rpm for 1 min), and the supernatant was discarded. The pellet was resuspended in 100 μl of the remaining medium. Next, 3 μl of cells were placed in between 2% agarose gel pad and a coverslip and imaged by confocal microscopy with a 100× objective (example images in Supplementary Figure S1). GFP fluorescence was measured with a 488 nm laser and a 514/30 nm emission filter. Phase-contrast images were simultaneously acquired. MG1655 cells were imaged to measure cell size in $LB_{1.0x}$, $LB_{0.75x}$, $LB_{0.5x}$, $LB_{0.25x}$, $LB_{1.5x}$, $LB_{2.0x}$ and $LB_{2.5x}$ media. Finally, MG1655 cells was also imaged in $LB_{1.0x}$ during stationary growth. Finally, we imaged cells of the YFP strain library to assess if their morphology and physiology were consistent with healthy cells during measurements.

Flow-cytometry

We performed flow-cytometry of RL1314 cells to measure single-cell RNAP over time. ONC were diluted at 1:1000 into respective fresh media ($LB_{1.0x}$, $LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$) and grown as described in Methods section *Bacterial strains, media,growth conditions andcurves, and intracellular concentrations*. Flow-cytometry data was recorded every 30 min (three biological replicates), up to 210 min. Data was also captured in the mid-exponential phase (at 180 min), in the media studied ($LB_{1.0x}$, $LB_{0.75x}$, $LB_{0.5x,}$ $LB_{0.25x}$, $LB_{1.5x}$, $LB_{2.0x}$ and $LB_{2.5x}$), with 3 biological replicates each. We used a similar protocol to perform flow-cytometry of several strains of the YFP library (25) in $LB_{1.0x}$ and $LB_{0.25x}$ (three biological replicates, Supplementary Table S4), including to measure single-cell SpoT levels in $LB_{1.0x}$, $LB_{0.75x}$, $LB_{0.5x}$, $LB_{0.25x}$, $LB_{1.5x}$, $LB_{2.0x}$ and $LB_{2.5x}$ (three biological replicates each).

Meanwhile, we measured single-cell levels of the crl gene in $LB_{0.5x}$ at 0 and 180 min, using a strain from the GFP-promoter fusion library. Further, to measure RpoS levels, we performed flow-cytometry of cells of the MGmCherry strain in $LB_{1.0x}$, $LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$ during the exponential (180 min) and stationary growth phases ($LB_{1.0x}$,14 h after pre-inoculation). In these measurements, as well as the measurements above, we recorded FSC-H, SSC-H and Width, to be used as proxies for cell size and density (i.e. composition), as they are positively correlated with these features (37).

In addition, data from measurements of MG1655 cells were used to discount background fluorescence from cells of the MGmCherry and the YFP strains. Similarly, measurements of the W3110 strain were used to discount the background fluorescence from the RL1314 strain.

For performing flow-cytometry, 5 μl of cells were diluted in 1 ml of PBS, and vortexed. In each condition, 50000 events were recorded. Prior to the experiments, QC was performed as recommended by the manufacturer. Measurements were conducted using an ACEA NovoCyte Flow Cytometer (ACEA Biosciences Inc., San Diego, USA) equipped with yellow and blue lasers.

For detecting the GFP and YFP signals, we used the FITC channel (-H parameter) with 488 nm excitation, 530/30 nm emission, and 14 μl/min sample flow-rate with a core diameter of 7.7 μm. PMT voltage was set to 550 for FITC and kept the same for all conditions. Similarly, to detect the mCherry sinal, we used PE-Texas Red channel (-H parameter) having an excitation of 561 nm and emission of 615/20 nm and sample flow-rate of 14 μl/min, with a core diameter of 7.7 μm. PMT voltage was set to 584 for PE-Texas Red and kept the same for all conditions. To remove background signal from particles smaller than bacteria, the detection threshold was set to 5000. All events were collected by Novo Express software from ACEA Biosciences Inc.

Protein isolation and western blotting

Western blotting was used to quantify relative RNAP levels of MG1655 cells and GFP levels of RL1314 RNAP-GFP cells (Supplementary Figure S2 and Table S2). Briefly, cells were diluted from ONC into respective fresh media and incubated at 37 °C with aeration and grown until reaching an $OD_{600} \approx 0.4$. Next, cells were harvested by centrifugation (8000 rpm for 5 min) and pellets were lysed with B-PER bacterial protein extraction reagent, added with a protease inhibitor for 10 min at room temperature (RT). Following lysis, centrifugation was done at 14 000 rpm for 10 min and the supernatant was collected. Next, the supernatant was diluted in 4X Laemmli buffer with β-mercaptoethanol and samples were boiled at 95 °C for 5 min.

Samples with ∼ 30 μg of soluble total proteins were loaded on 4–20% TGX stain-free precast gels (Biorad). These proteins were then separated by electrophoresis and transferred on PVDF membrane using TurboBlot (Biorad). Next, membranes were blocked with 5% non-fat milk at room temperature (RT) for 1 h and probed with primary antibodies at 1:2000 dilutions (Biolegend) at 4 °C overnight. The antibodies used for the MG1655 strain were against RpoC (β prime subunit of RNAP), while for the RL1314 strain it was used antibodies against GFP. As a control we also subjected MG1655 cells to antibodies against GFP (Supplementary Figure $S2B_2$). HRP-secondary antibody (1:5000) treatment was then done (Sigma Aldrich) for 1 h at RT. Excess antibodies were removed by washing with buffer. The membrane was treated with chemiluminescence reagent (Biorad) for band detection. Images were obtained by the Chemidoc XRS system (Biorad) and band quantification was done using the Image Lab software (v.5.2.1).

RNA-seq

Sample preparation. RNA-seq was performed thrice, for decreasing [$LB_{0.75x}$, $LB_{0.5x}$, and $LB_{0.25x}$, at 180 min; $LB_{0.5x}$ at 60 and 125 min] and for increasing ($LB_{1.5x}$, $LB_{2.0x}$ and $LB_{2.5x}$, at 180 min) medium richness relative to a control ($LB_{1.0x}$) (an independent control was used for each three sets of conditions). Cells from 3 independent biological replicates of MG1655 in each modified medium were treated with RNA protect bacteria reagent (Qiagen, Germany), to prevent degradation of RNA, and their total RNA was extracted using RNeasy kit (Qiagen). RNA was treated twice with DNase (Turbo DNA-free kit, Ambion) and quantified using Qubit 2.0 Fluorometer RNA assay (Invitrogen, Carlsbad, CA, USA). Total RNA abundance was determined by gel electrophoresis, using a 1% agarose gel stained with SYBR safe (Invitrogen). RNA was detected using UV with a Chemidoc XRS imager (Biorad).

Sequencing.

Part 1: For shifts from $LB_{1.0x}$ to $LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$, at 180 min. Sequencing was performed by Acobiom (Montpellier, France). The RNA integrity number (RIN) of the samples was obtained with the 2100 Bioanalyzer (Agilent Technologies, Palo Alto, USA) using Eukaryotic Total RNA 6000 Nano Chip (Agilent Technologies). Ribosomal RNA depletion was performed using Ribo-Zero removal kit (Bacteria) from Illumina. RNA-seq libraries were constructed according to the Illumina's protocol. Samples were sequenced using a single-index, 1 × 75 bp single-end configuration (∼ 10M reads/library) on an Illumina MiSeq instrument. Sequencing analysis and base calling were performed using the Illumina Pipeline. Sequences were obtained after purity filtering.

Part 2: For shifts from $LB_{1.0x}$ to $LB_{1.5x}$, $LB_{2.0x}$ and $LB_{2.5x}$ at 180 min, and from $LB_{1.0x}$ to $LB_{0.5x}$ at 60 and 125 min. Sequencing was performed by GENEWIZ, Inc. (Leipzig, Germany). The RIN of the samples was obtained with the Agilent 4200 TapeStation (Agilent Technologies, Palo Alto, CA, USA). Ribosomal RNA depletion was performed using Ribo-Zero Gold Kit (Bacterial probe) (Illumina, San Diego, CA, USA). RNA-seq libraries were constructed using NEBNext Ultra RNA Library Prep Kit (NEB, Ipswich, MA, USA). Sequencing libraries were multiplexed and clustered on 1 lane of a flow-cell.

For shifts from $LB_{1.0x}$ to $LB_{1.5x}$, $LB_{2.0x}$ and $LB_{2.5x}$ at 180 min, samples were sequenced using a single-index, 2 × 150 bp paired-end (PE) configuration (∼ 350M raw paired-end reads per lane) on an Illumina HiSeq 4000 instrument. Image analysis and base calling were conducted with HiSeq Control Software (HCS). Raw sequence data (.bcl files) were converted into fastq files and de-multiplexed using Illumina bcl2fastq v.2.20. One mismatch was allowed for index sequence identification.

For shifts from $LB_{1.0x}$ to $LB_{0.5x}$ at 60 and 125 min, samples were sequenced using a single-index, 2 × 150 bp paired-end (PE) configuration (∼ 10M raw paired-end reads per lane) on an Illumina NovaSeq 6000 instrument. Image analysis and base calling were conducted with NovaSeq Control Software v1.7. Raw sequence data (.bcl files) were converted into fastq files and de-multiplexed using Illumina bcl2fastq v.2.20. One mismatch was allowed for index sequence identification.

Data analysis. Regarding the RNA-seq data analysis pipeline: (i) RNA sequencing reads were trimmed to remove possible adapter sequences and nucleotides with poor quality with Trimmomatic (38) v.0.36 (for data from sequencing part 1) and v.0.39 (for data from sequencing part 2). (ii) Trimmed reads were then mapped to the reference genome, *E. coli* MG1655 (NC_000913.3), using the

Bowtie2 v.2.3.5.1 aligner, which outputs BAM files (39). (iii) Then, *featureCounts* from the Rsubread R package (v.1.34.7) was used to calculate unique gene hit counts (40). Genes with less than five counts in more than three samples, and genes whose mean counts are less than 10 were removed from further analysis. (iv) Unique gene hit counts were then used for the subsequent differential expression analysis. For this, we used the DESeq2 R package (v.1.24.0) (41) to compare gene expression between groups of samples and calculate p-values and log2 of fold changes (LFC) of RNA abundances using Wald tests (function *nbinomWaldTest*). *P*-values were adjusted for multiple hypotheses testing (Benjamini–Hochberg, BH procedure, (42)) and genes with adjusted *P*-values (False discovery rate (FDR)) < 0.05 were selected to be further tested as being differentially expressed (DE) (Methods section *RNA-seq d*).

For logistical reasons, the sequencing platform for the RNA-seq data in Methods section *RNA-seq b* differ from one another. Consequently, the data sets first mentioned in the Results sections *Genome-wide mid-term responses correlate with shifts in RNAp concentration* and *Further increases in* medium richness do not decrease RNAP concentration and RNA numbers also do not change, respectively, cannot be compared quantitatively nor be used to infer gene-specific conclusions.

Finally, to analyse the data from $\mathrm{LB}_{1.0x}$ and $\mathrm{LB}_{0.5x}$ at 60 and 125 min and compare its results with the results from the data of Methods section *RNA-seq b Part 1* at 180 min, their raw count matrices were merged. Also, only genes that passed the filtering were studied. The filtering removed genes with less than 5 counts in more than 6 samples, and genes whose mean counts were less than 10.

Moreover, we expect the overall sums of LFCs from each perturbation to equal zero since, in DEseq2, the median-of-ratios normalization calculates the normalizing size factors assuming a symmetric differential expression across conditions (i.e. same number of up- and down-regulated genes) (43). Further, it fits a zero-centered normal distribution to the observed distribution of maximum-likelihood estimates (MLEs) of LFCs over all genes (41). Both steps (perhaps related) force the mean LFC to be 0.

LFC criteria for differentially expressed genes. From past methods (44–46), we classified genes as statistically significantly DE following perturbations, by setting a maximum FDR threshold for adjusted *P*-values (Methods section *RNA-seq c*) and a minimum threshold for the absolute LFC of RNA numbers of individual genes (|LFC|).

From the $\mu_{|LFC|}$ of genes whose FDR > 0.05, named $\mu_{|LFC|}(FDR > 0.05)$, we identified DEGs (DE Genes) as those that, in addition to having FDR < 0.05, also have $|\mathrm{LFC}| > \mu_{|LFC|}(FDR > 0.05)$. Specifically, we added the conditions: $|\mathrm{LFC}| > 0.4248$ for $\mathrm{LB}_{0.75x}$, > 0.4085 for $\mathrm{LB}_{0.5x}$, > 0.4138 for $\mathrm{LB}_{0.25x}$, > 0.2488 for $\mathrm{LB}_{1.5x}$, > 0.2592 for $\mathrm{LB}_{2.0x}$, and > 0.2711 for $\mathrm{LB}_{2.5x}$, for accepting a gene as being significantly DE. Meanwhile, for the data in $\mathrm{LB}_{0.5x}$ at 60 and 125 min, we added the conditions: $|\mathrm{LFC}| > 0.2171$ for $\mathrm{LB}_{0.5x}$ 60 min, and > 0.2977 for $\mathrm{LB}_{0.5x}$ 125 min. This allows removing genes whose FDR < 0.05 but have a negligible LFC. Noteworthy, in no condition did we remove from the set of DEG more than 5 genes by applying this rule.

RNA-seq vs Flow-cytometry. RNA and protein abundances are expected to be positively correlated in bacteria, since transcription and translation are mechanically bound (47–49). Further, most regulation occurs during transcription initiation (50), which is the lengthiest sub-process (24).

To validate that this relationship holds during the genome-wide stresses, we randomly selected a set of genes whose LFC's, as measured with RNA-seq, cover nearly the entire spectrum of LFCs observed genome-wide. Next, we measured their LFC in protein levels, using the YFP strain library (25) (Methods section *Bacterial strains, media,growth conditions andcurves, and intracellular concentrations*) and flow-cytometry (Methods section *Flow-cytometry*), at 180 min after shifting the medium. The list of selected genes is shown in Supplementary Table S4. For the fold change of 1/8×, 1/4×, 1/2×, 1×, 2×, $4_{\times}$ and 8×, we selected three genes whose LFC in RNA abundances is closest to that value (except for the 8× fold change, since only two genes were available). This range of values covers nearly the whole LFC spectrum observed by RNA-seq (Supplementary Figure S9).

Transcription factor network of *Escherichia coli*

We assembled a directed graph of the network of TF interactions between the genes present in our RNA-seq data, based on the data in RegulonDB v10.9 (34), as of 28 January 2022. We used all reported TF-gene, TF–TF, TF–operon and TF–TU interactions. These equally contribute to our network of gene-gene directed interactions. In detail, a TF or regulatory protein is a complex protein that activates/represses transcription of a TU upon binding to specific DNA sites. A TU is one or more genes transcribed from a single promoter. Similarly, an operon are one or more genes and associated regulatory elements, transcribed as a single unit.

The TFN graph was analysed using MATLAB (2021b) and Network Analyzer v.3.7.2 plug-in in cytoscape (51) to extract the following network parameters: number of nodes and directed edges, number of connected components, number of isolated nodes and self-loops, and single-gene in- and out-degree, edge-count, clustering coefficient, eccentricity, average minimum path length, betweenness and stress centrality, and neighbourhood connectivity. The statistics considered are shown in Supplementary Tables S5 and S19.

Statistical tests

2-Sample T-test, 2-sample KS-test and 1-sample Z-test. The 2-sample *T*-test evaluates the null hypothesis that the two samples come from independent random samples from normal distributions with equal means and unequal and unknown variances. For this, we have set a significance level of 10% significance level (*P*-value < 0.10) when applying the MATLAB function *ttest2*.

The 2-sample KS-test returns a test decision for the null hypothesis that the data from two data sets are from the same continuous distribution, using the two-sample Kolmogorov-Smirnov test. As above, we have set the null hypothesis at 10% significance level (*P*-value < 0.10).

The one-sample *Z*-test tests for the null hypothesis that the sample is from a normal distribution with mean *m* and a standard deviation σ. In this case, *m* and σ are estimated from the genes with $K_{TF} = 0$. As above, we have set the null hypothesis at 10% significance level (*P*-value < 0.10).

Fisher test. The Fisher test evaluates the null hypothesis that there is no association between the two variables of a contingency table. We reject the null hypothesis at 10% significance level (*P*-value < 0.1), meaning that the variables are significantly associated.

Correlations between data sets. The correlation between two data sets with known uncertainties (standard error of the mean (SEM) in each data point) was obtained by performing linear regression fitting using Ordinary Least Squares. The best fitting line along with its 68.2% confidence interval/bounds (CB) and statistics was obtained as described in Supplementary Materials and Methods 1.4 of (52). In short, the uncertainty of each of the N empirical data points was represented by m points, resulting in $n = N \times m$ points. Each of these points is obtained by random sampling from a normal distribution whose mean (μ) and standard deviation (σ) equal the mean and error of the empirical data point, respectively. It was set $m = 1000$, as it was sufficient to represent the error bars of the actual data points. We obtained the coefficient of determination (R^2) and the root mean square error (RMSE) of the fitted regression line, and the p-values of the regression coefficients. The *P*-value of *x* (*P*-value$_1$) was obtained of a T-test under the null hypothesis that the data is best fit by a degenerate model consisting of only a constant term. If *P*-value$_1$ is smaller than 0.1, we reject the null hypothesis that the line is horizontal, i.e. that one variable does not linearly correlate with the other. When there are more than three data points, we also calculated regression coefficient of x^2 (*P*-value$_2$) of a *T*-test under the null hypothesis that the second order polynomial fit is no better than lower order polynomial fit, i.e. coefficient of $x^2 = 0$. If *P*-value$_2$ is smaller than 0.1, we reject the linear model favouring the quadratic.

To obtain the overall best non-linear fit (and its 68.2% CB) for the empirically measured datasets with uncertainties, Monte Carlo simulations (1000 iterations) were performed. In particular, to obtain Figure 2B, on each iteration, we randomly sampled each data point from a normal distribution whose mean and standard deviation are equal to the mean (actual value) and SEM of the corresponding empirical data point, respectively. Then a sigmoid (logistic) curve fitting (R P (2020). sigm_fit (https://www.mathworks.com/matlabcentral/fileexchange/42641-sigm_fit), MATLAB Central File Exchange. Retrieved 6 August 2020) was used to obtain the best fitting curve and its 68.2% CB for each iteration. Finally, the best fitting curve along with their 68.2% CB is obtained by averaging the respective values from the 1000 iterations.

Finally, to create null-models of how variable X affects variable Y, we performed random sampling without replacement of both X and Y datapoints. The number of samplings and the sampling size (number of samples in each sampling) are set to the maximum array size possible to us (∼ 45980 × 45980, 15.8 GB). The sampling size is set to 5% of the number of datapoints (size_XY) and the number of samplings (*K*) is set according to Max_size/(0.05 × size_XY) where Max_size = 45980/2. Next, for both X and Y, we combine the sampled datapoints in a vector (sample_X, sample_Y) and calculate the correlation between sample_X and sample_Y by linear regression fitting using ordinary least squares. To correct for over-representation of the original datapoints, we corrected the degrees of freedom to be (size_XY – *C*), where *C* is the number of parameters. In detail, for the linear regression fitting, *C* equals to 2 (intercept and slope of best fitting line).

ANCOVA test to evaluate if two lines can be distinguished. To evaluate if two lines are statistically different, we performed the analysis of covariance (ANCOVA) test (53). ANCOVA is an extension of the one-way ANOVA to incorporate a covariate. This allows comparing if two lines are statistically distinct in either slope or intercept. This is done by evaluating the significance of the *T*-test under the null hypothesis that both the slopes and intercepts are equal.

Figures

Figures were produced in R (v.3.6.0) using the packages ‘ggplot2’ (v.3.2.0), ‘pheatmap’ (v.1.0.12), ‘VennDiagram’ (v.1.6.20) along with ‘grid’ (v.3.6.0), ‘gridExtra’ (v.2.3), ‘gplots’ (v.3.0.1.1), ‘R.matlab’ (v.3.6.2), ‘dplyr’ (v.1.0.2), ‘scales’ (v.1.0.0), ‘Metrics’ (v.0.1.4) and ‘fitdistrplus’ (v.1.0–14).

RESULTS

Effects of medium dilution on cell growth, morphology, and RNAP concentration

We first studied how the RNAP concentration changes with medium dilutions. Concentration of RNAP (as well as of other molecular species) was obtained as described in Methods section *Bacterial strains, media, growth conditions and curves, and intracellular concentrations*. From a control medium ($LB_{1.0\times}$), we moved cells to diluted media ($LB_{0.75\times}$, $LB_{0.5\times}$, and $LB_{0.25\times}$, Methods section *Bacterial strains, media,growth conditions andcurves, and intracellular concentrations*). RNAP levels start changing ∼ 75 min later, based on a as yet to be identified mechanism, stabilizing at ∼ 165 min (Figure 2B). Given this timing of events, measurements to assess the effects on the RNA population should be performed after ∼ 165 min.

We also considered that at ∼ 180 min (Figure 2A) the cells are at late mid-log phase. Thus, measuring the effects of changing RNAP should occur *prior* to ∼ 180 min, since leaving the mid-log phase will involve significant, unrelated genome-wide changes in RNA abundances (54–57). From the point of view of cell division, from the moment when the RNAP starts changing, up to the moment when we measure the short- and the mid-term changes in RNA abundances, on average, less than one cell cycle and less than two cell cycles should have passed, respectively.

Interestingly, this time moment (∼ 180 min) matches our predictions of when, on average, RNA abundances have changed due to changes in the abundances of both RNAP as well as direct input TFs. In detail, from the timing of the changes in RNAP (Figure 2B) and from known rates of RNA and protein production and degradation in *E. coli*

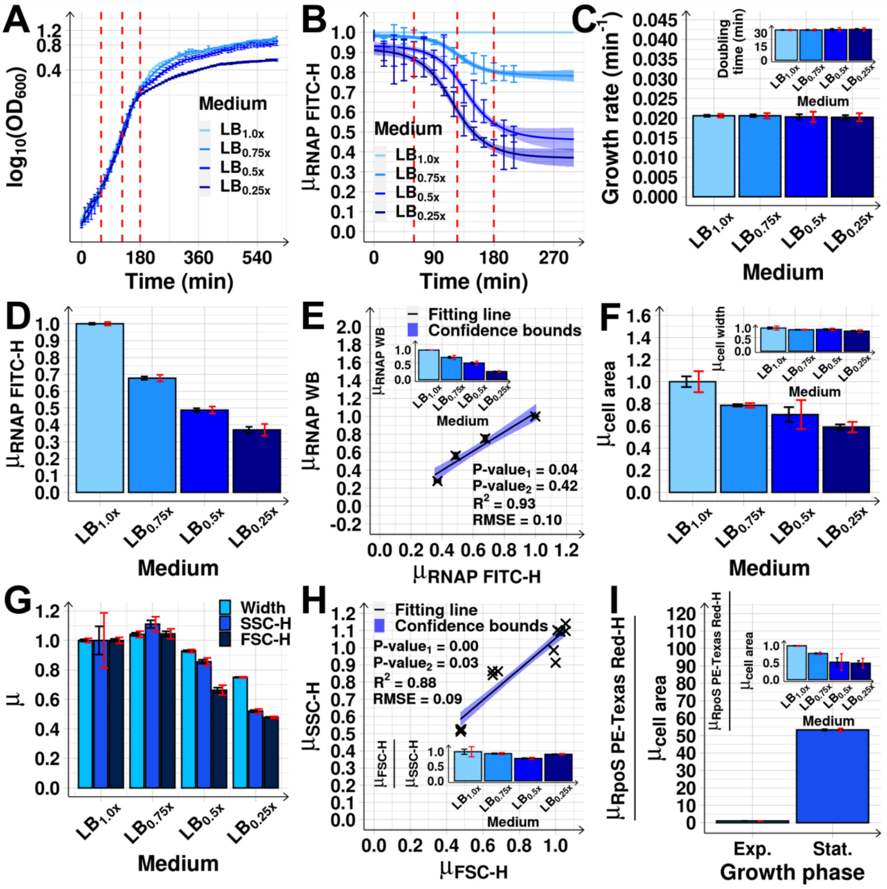


Figure 2. Cell growth and morphology, and RNAP abundance after medium dilutions. (**A**) Growth curves from OD$_{600}$ measured every 10 min (Methods section *Bacterial strains, media,growth conditions andcurves, and intracellular concentrations*). The vertical dashed red lines mark when RNA-seq was performed. After ∼ 180 min, cells subject to different dilutions (LB$_{mx}$) start differing in growth rates. (**B**) Mean single-cell RNAP-GFP fluorescence relative to the control (LB$_{1.0x}$), $\mu_{RNAP\ FITC-H}$, measured every 15 min for 210 min by flow-cytometry (FITC-H channel). The mean cellular background fluorescence in each condition was subtracted (Methods section *Flow-cytometry*). The vertical dashed red lines mark when RNA-seq was performed. (**C**) Growth rates at 180 min after medium dilution. The inset shows the corresponding doubling times. (**D**) Mean single-cell RNAP levels ($\mu_{RNAP\ FITC-H}$) at 180 min relative to the control (Methods section *Flow-cytometry*). (**E**) $\mu_{RNAP\ FITC-H}$ plotted against $\mu_{RNAP\ WB}$ (RNAP levels measured by Western Blot, Methods section *Protein isolation and western blotting*). The inset shows $\mu_{RNAP\ WB}$ alone. (**F**) Mean cell area relative to the control, extracted from phase-contrast images (∼ 2000 cells per condition) (Methods section *Microscopy*). The inset shows the mean cell width relative to the control. (**G**) Mean (relative to the control) Width, FSC-H and SSC-H obtained by flow-cytometry (Methods section *Flow-cytometry*). (**H**) Mean (relative to the control) FSC-H versus SSC-H in each condition, obtained from 3 biological replicates. The inset shows the mean ratio between the relative FSC-H and SSC-H. (**I**) Mean mCherry-tagged RpoS ($\mu_{RpoS\ PE-Texas\ Red-H}$) concentration in the stationary growth phase relative to the exponential growth (set to 1), as measured by mean single-cell fluorescence (PE-Texas Red channel, Methods section *Flow-cytometry*) over mean cell area ($\mu_{cellarea}$) (Methods section *Microscopy*), after subtracting mean background fluorescence(s). The inset shows the same, but after each medium dilution. Measurements in (D)–(I) taken 180 min after medium dilution. Data points are from 3 biological replicates (except for (A) and (B), where 6 replicates were used). μ stands for mean relative to the control. In (A)–(C) error bars represent the SEM. In (B) and (D)-(I), black error bars are the SEM and red error bars are the 95% CB of the SEM. In (C), (E) and (H), the best fitting lines and their 68% CB and statistics (R^2 and RMSE), and P-values at 10% significance level were obtained as described in Methods section *Statistical tests c*.

(25,28–31), widespread heterogenous short-term changes in RNA abundances should occur, on average, at ~ 120–135 min after shifting the medium (at which moment the RNAP has already changed significantly). Changes in the corresponding protein abundances should then occur tenths of minutes later, i.e. at ~ 160–175 min (29–31).

Soon after, we expect additional changes in RNA abundances, now due to changes in direct input TFs abundances. This second stage of events, here classified as 'mid-term', should occur between ~ 165 and 180 min. This is also when cells are in the late mid-log phase (Figure 2A), while cell growth rates do not yet differ between conditions (Figure 2C) and cell sizes only differ slightly (Figure 2F–H and Supplementary Figures S3 and S4, Methods sections *Microscopy* and *Flow-cytometry*). This is relevant since growth rates affect protein concentrations due to dilution in growth and division (58,59).

Finally, at 180 min, the σ^{38} concentration is lower than at 0 min (Figure 2I inset and Supplementary Figure S5), in agreement with previous reports (27,35,60,61), suggesting that the cells are not committed to the stationary growth phase. The same is observed for the Crl protein (Supplementary Figure S6). This protein contributes to the expression of genes whose promoter is recognized by σ^{38} and is known to be at higher abundance during stationary phase (62), as confirmed here (Supplementary Figure S6).

Given the above, to capture the average mid-term effects of RNAP shifts, we measured the transcriptome at 180 min (Figure 2A). This timing should allow discerning the average genes' behaviour under the influence of their local network of TF interactions, albeit the diversity in RNA and protein production and decay kinetics, etc. RNAP levels at that moment are shown in Figure 2D (flow-cytometry data) and 2E (flow-cytometry versus western blot data), Supplementary Figures S2A$_1$ and S2A$_2$ (western blot of RNAP) and Table S2 (absolute values extracted from western blot). Similar RNAP downshifts have been observed in natural conditions (63) and described in (23,26,27).

Finally, we performed an additional western blot to measure RNAP-GFP using antibodies against GFP alone. From Supplementary Figures S2B$_1$ and S2B$_2$ and Table S2, the RNAP-GFP does not appear to be significantly degraded by cleavage, with the strongest bands being observed for molecular weights between 150 and 250 kDa. As such, these strongest bands should correspond to GFP (known to be 27 kDa (64)) fused with the β' unit (known to be 155 kDa (65,66)). Moreover, no clear bands appear in the region between 150 and 250 kDa for the MG1655 strain. Meanwhile, the weak band just above 25 kDa in some samples from RL1314 cells (particularly in LB$_{0.75x}$) might correspond to GFP that has been cleaved off from the chimeric protein but, given that it is only a small fraction compared to the amount of RpoC-GFP in the same cells, one can conclude that its contribution to the total cell fluorescence signals is negligible.

Genome-wide mid-term responses correlate with shifts in RNAP concentration

Transcription rates are expected to follow the free RNAP concentration in a cell, rather than the total RNAP concentration (which is the sum of the free RNAP with the RNAP engaged with the DNA). We here measured the total RNAP concentration. However, within the range of conditions studied, the fractions of free and DNA-bound RNAP remain rather constant (26). Therefore, the total RNAP is a good proxy for the free RNAP. Specifically, using modified strains and plasmids controlled by lac and tet mutant promoters (67–69), whose regulatory mechanisms have been dissected, it was shown that their transcription rates are linearly correlated with the total RNAP concentration (26). From here on, when mentioning RNAP concentration, we refer to the total RNAP concentration.

The increasing medium dilution and corresponding decreases in RNAP concentration (Figure 3A) cause increasingly broad distributions of single-gene LFCs at 180 min (Supplementary Figures S7 and S8A-C and Table S3). Specifically, the mean absolute LFC ($\mu_{|LFC|}$) of the 4045 genes and the number of DEGs increased with medium dilution (Figures 3B and C).

These RNAs changes correlate with subsequent changes in proteins levels (Supplementary Figure S9, Methods sections *Flow-cytometry* and *RNA-seq*). This suggests that no significant translational or post-translational regulation is taking place in between the perturbation and the measurements, that would alter proteins abundances significantly.

Interestingly, while both $\mu_{|LFC|}$ and DEGs numbers follow the RNAP concentration (Supplementary Figures S8D and S10B), these relationships are not strictly linear (p-value of 0.29, Supplementary Figure S8D). This suggests that, in addition to RNAP, the direct input TFs are also influential. In this regard, we note that the assumption of linearity in the absence of the influence of input TFs (observed and discussed in (26)) is only expected to occur within a narrow range of parameter values.

Notably, some of the genes may be also influenced by sources other than RNAP and direct input TFs, such as supercoiling buildup. Also, some input TFs other than the direct input TFs maybe be influential. However, we show evidence below that this does not affect the average results (Figure 4C and Supplementary Figure S18).

We also performed RNA-seq *prior* to when most signals, generated by the shift in RNAP, propagated in the TFN. First, we measured LFCs at 60 min after diluting the medium (Figure 1). From Figure 2B, at this moment, RNAP abundances have not yet changed relative to the control. In agreement, the genome-wide $\mu_{|LFC|}$ is very weak (Figure 3D). We further performed RNA-seq at 125 min. At this moment, RNAP levels have already reduced significantly (Figure 2B), but we do not expect input TF abundances to have changed significantly given protein production times (Figure 1). In agreement, |LFC|s at 125 min are stronger than at 60 min, but much weaker than at 180 min (Figure 3D). We conclude that the mid-term changes in the TFN have not occurred yet (further evidence is provided below). Given this, from here onwards, we focus on the state of the TFN at 180 min.

Influences from regulators other than RNAP

We investigated whether other factors influenced the global response of the TFN. We considered GRs, σ factors,

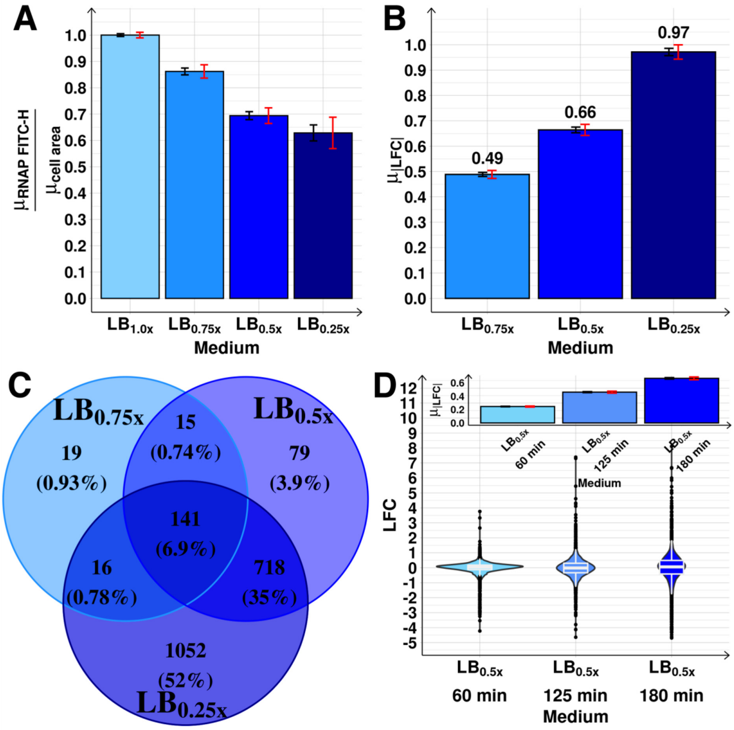


Figure 3. Genome-wide effects on the transcriptome of diluting the medium. (**A**) RNAP concentration estimated from the ratio between the RNAP measured by FITC-H ($\mu_{RNAP\ FITC-H}$) at 180 min (Methods section *Flow-cytometry*), and the mean cell area ($\mu_{cellarea}$, used as a proxy for cell volume) obtained by phase-contrast microscopy (Methods section *Microscopy*). Values relative to the control ($LB_{1.0x}$). (**B**) $\mu_{|LFC|}$ in each medium. (**C**) Venn diagram of the number (and percentage relative to the total number of genes) of DEG. (**D**) Violin plot with the maximum, minimum, median, interquartile ranges, and probability density of the distributions prior to RNAP changes ($LB_{0.5x}$ at 60 min) and the subsequent short- ($LB_{0.5x}$ at 125 min) and mid-term ($LB_{0.5x}$ at 180 min) responses to shifting RNAP. The inset shows $\mu_{|LFC|}$ of the distributions. In (A), (B) and (C), black error bars are the SEM, while red error bars are the 95% CB of the SEM.

(p)ppGpp and non-coding RNAs. We assumed the classification of GR in (70,71) as an input TF that regulates a large number of genes that rarely regulate themselves and participate in metabolic pathways. Meanwhile, we did not account for promoters' close proximity (e.g. tandem formation), since a recent study (72) showed that, under similar stress, while close proximity causes transcription interference, it does not influence the genes' input TF regulation.

First, the RNA-seq shows large numbers of DEGs (> 1000 for the two strongest dilutions (Supplementary Figure S10A)) as well as a linear correlation between these numbers and $\mu_{|LFC|}$ (Supplementary Figure S10C). Thus, we argue that the responsive genes are not constrained to a specific cluster, such as genes responding to a GR other than RNAP (the most influential is, arguably, σ^{70} with 1555 genes recognizing it, while other GRs control less than 510 genes each (34)).

Second, from the RNA-seq, we analysed the relative abundances of GRs, σ factors and of their output genes. From Supplementary Figures S26A and S26C, apart from rpoS (an input TF recognized by 321 genes) and flhC (an input TF recognized by 75 genes), GRs and σ factors did not change significantly (Supplementary Figure S26). Further, those two changes (rpoS and flhC) were positively correlated with the RNAP concentration (Figure 2I inset and Supplementary Figure S5), not allowing to separate their effects. Noteworthy, alternative σ factors did not change significantly relative to σ^{70} (Supplementary Figure S26E), which would have changed the competition for RNAP binding.

We thus failed to find evidence that the σ factors and GRs were influential, globally, in the mid-term responses. Supplementary Table S15 lists the conclusion for each specific GR and σ factor and Supplementary Figure S27 shows these results at 125 min.

We then investigated if (p)ppGpp could be influential since, under some nutrient starvation conditions, they affect ∼ 1000 genes by binding RNAP and altering its affinity for their promoters (3). Reports suggest that the effects are rapid (5–10 min (3)). In agreement, genes responsive

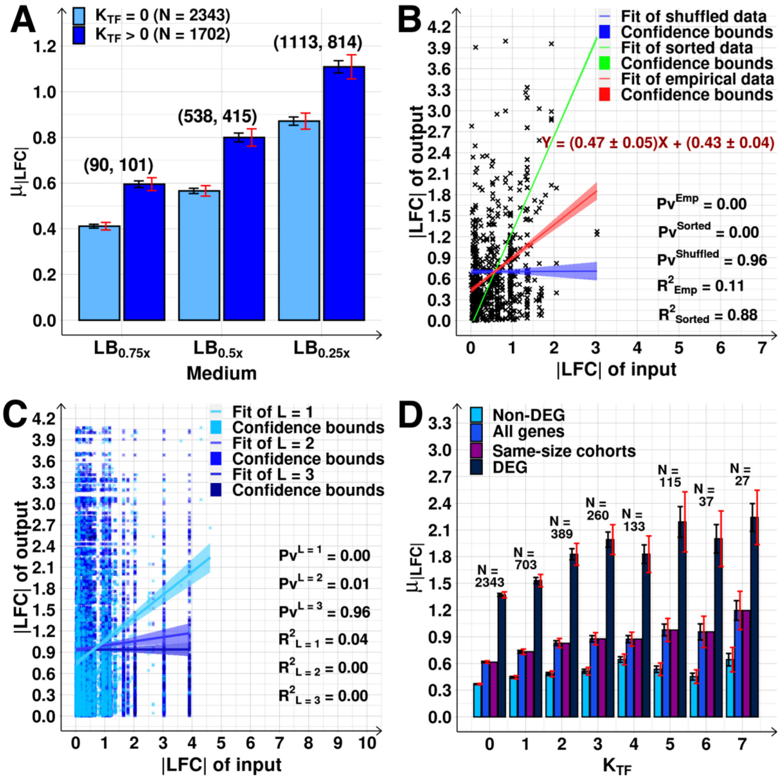


Figure 4. Genome-wide propagation of the effects of shifting RNAP in the TFN. (**A**) $\mu_{|LFC|}$ of N genes with and without input TFs ($K_{\mathrm{TF}} > 0$ and $= 0$, respectively). On the top of each bar is the number of DEGs in each set. (**B**) |LFC| of genes with $K_{\mathrm{TF}} = 1$ versus the |LFC| of genes coding their direct input TFs. Data from the LB$_{0.5x}$ shift. The red line is the best fit. The blue line is the null-model fitting lines and was obtained as described in Methods section *Statistical tests c*. The green line is the best fit after sorting the input-output pair values to maximize the correlation. Shadows are their 68% CB. The equations of the red fitting lines with '±' inform on the standard error of the slope. (**C**) Scatter plots between |LFC| of output and input genes distanced by a minimum path length L of 1, 2 and 3 input TFs (edges) in the TFN, respectively (data from LB$_{0.5x}$). Only for $L = 1$ do the activities of output and input genes correlate (P-value$_1$ > 0 and $R^2 > 0$). (**D**) $\mu_{|LFC|}$ of all genes, DEG, non-DEG, and cohorts of randomly selected genes of the same size ('same sized cohorts') for $K_{\mathrm{TF}} = 0$ to 7, using merged data from all shifts (LB$_{0.75x}$, LB$_{0.5x}$ and LB$_{0.25x}$). Black error bars are the SEM and red error bars are the 95% CB of the SEM. Best fitting lines and 68% CB obtained using *FITLM* (MATLAB). p-values, obtained using the null hypothesis that the data is best fit by a horizontal line, are not rejected at 10% significance level. (B) and (C) do not include a few data points to facilitate visualization. See Supplementary Figures S14 and S18 for complete data.

to (p)ppGpp (3) exhibited abnormal short-term responses (Supplementary Table S20). However, their mid-term responses at 180 min were no longer atypical and, instead, followed the RNAP changes. The expression of spoT, one of the genes responsible for ppGpp synthesis, also followed the RNAP (Supplementary Figure S28). As such, we could not establish a long-lasting global influence from (p)ppGpp in response to growth-medium dilution. Nevertheless, the LFCs of the 14 out of the 22 genes coding for rRNAs listed in RegulonDB did reveal atypical behaviors (Supplementary Table S22).

Finally, we searched for unique behaviors in sRNAs by analyzing the LFC of the 93 sRNAs reported in RegulonDB. Their behavior was not atypical, neither at 180 min after the perturbations (Supplementary Table S21), nor at 125 min. Further, we analysed if their output genes followed their behaviour. We found that the LFCs of genes directly regulated by the sRNAs were not correlated with their input TFs, neither at 125 min, nor at 180 min after the medium shifts. Specifically, of the 93 sRNAs, 37 of them have known output genes (in a total of 145 outputs). The RNA-seq data provided information on the LFC of 40 of the 145 outputs. Finally, we searched for linear correlations between the pairs of LFCs of sRNAs and their output genes, respectively, in the short-term (125 min in LB$_{0.5x}$) and in the mid-term (180 min in LB$_{0.5x}$). We found an R^2 of 0.03 (P value = 0.18) at 125 min and an R^2 of 0.05 (P value = 0.10) at 180 min, respectively. We thus cannot conclude that sRNAs were influential during the short- and mid-term responses to the stresses.

Input TFs influence the transcriptional response

If the TFN influences the genes' mid-term response to the shift in RNAP concentration, this should cause genes with and genes without input TFs to behave differently. Particularly, since the latter should only be affected by the RNAP abundances.

In agreement, genes with input TFs had higher $\mu_{|LFC|}$ than genes without input TFs (Figure 4A, Supplementary Figure S13 and Table S6). Also, the |LFC| of output genes and of genes coding for their direct input TFs correlate statistically (Figure 4B, Supplementary Figure S14 and Table S7). Therefore, on average, TF–gene interactions affected the single-gene, mid-term responses as hypothesized (Figure 1).

Input TFs influence all genes within operons

When considering the TFN topology, we have accounted for TF–gene interactions both between the input TF and the first gene of an operon or TU. Further, we have also accounted for the interactions between the same input TF and the other genes of the operon or TU (illustration of TUs and operons in Supplementary Figure S11B, which follows the standard definition of a group of two or more genes transcribed as a polycistronic unit (1)).

If we had not account for all these interactions, we would have failed to correlate the activities of genes interacting with each other. For example, consider an operon consisting of genes X_1 and X_2. Then, assume that gene A represses X_1 and X_2, by repressing their common promoter. If X_1 is an input TF to gene C, while X_2 is an input TF to gene D, then gene A should indirectly affect both genes C and D. If we had ignored the interaction between A and X_2, because it is not the first gene in its operon, we would explain why A affects C, but fail to explain why A affects D.

Further, many operons contain sets of genes whose RNAs code for subunits of the same protein complex (73,74). However, the opposite is also true. Moreover, the fraction of complexes encoded by proteins from different TU's is higher than those encoded from the same operon (75). Thus, we need to track interactions between input TFs and genes in any position in an operon or TU.

We tested if the positioning of the genes in the operon influenced their responsiveness to their input TFs. As a case study, we considered operons with 3 genes. These account for $\sim$ 21% of all operons with more than 1 gene (34). We found that the genes' positioning did not affect how they relate to the input TFs (Supplementary Figure S16). We obtained similar results for TUs (Supplementary Figure S17). The tests of statistical significance are shown in Supplementary Tables S9-S12.

Genes expressing TFs are correlated with their nearest neighbour output genes

Consider the interval between the shift in RNAP levels and the sampling for RNA-seq (Figure 1). From these, we hypothesized that, on average, at 180 min (i.e. $\sim$ 70 min after the RNAP changed relative to the control), mostly only genes directly linked by input TFs should exhibit correlated responses. Nevertheless, the genome-wide diversity in the kinetics of gene expression and in RNA and protein lifetimes will allow for correlations between genes more distanced in the TFN. The number of such correlations should decrease rapidly with the path length between the gene pairs considered.

Results in Figure 4C support this. Genes distanced by 1 input TF ($L = 1$, i.e. directly linked) have related |LFC|s, while genes distanced by two input TFs in the TFN have much less correlated responses (albeit still statistically significant). Finally, we found no correlations between the |LFC|s, of genes distanced by three input TFs (Supplementary Figure S18).

Noteworthy, the lack of correlation between genes separated by $L > 1$ could also be partially due to interference from the TFs of the 'intermediary' genes between the gene pairs. However, this is only a possibility when all input TFs involved can change in abundance in less than 60 min, which is likely uncommon in *E. coli*. This is supported by the RNA-seq data at 125 min after medium dilution (Supplementary Figure S19), when even direct input TFs and output genes are weakly correlated. This suggests that this shorter time interval was insufficient for most signals to have propagated between nearest neighbours (Supplementary Figure S19).

The number of input TFs, K_{TF}, of a gene correlates to the magnitude of its transcriptional response

We investigated if the genes mid-term responses are sensitive to their K_{TF} (Supplementary Figure S20A). When averaging the results from the three perturbations (Figure 4D), we found that the average of the absolute LFCs, $\mu_{|LFC|}$, increases with K_{TF} (Supplementary Figure S22 shows the goodness of the linear fits). The result was the same whether considering all genes or just the DEGs (Figure 4D, Supplementary Figure S21 and Table S13). Further, it holds true even for non-DEGs (Figure 4D), which justifies also considering these genes when studying the genome-wide effects. In agreement, we found no trend in the fraction of DEGs when plotted against K_{TF} (Supplementary Figure S23). For comparison, neither at 60 min nor 125 min do the genes' response and their K_{TF} correlate (Supplementary Figures S14 and S15 and Tables S7 and S8).

We verified that the relationship between $\mu_{|LFC|}$ and K_{TF} at 180 min is not an artifact caused by a decrease in cohort size with K_{TF}. We used bootstrapping to obtain cohorts of randomly sampled genes with increasing K_{TF} (10000 cohorts). We imposed a cohort size equal to the number of genes with $K_{TF} = 7$ (27 genes). The new, estimated $\mu_{|LFC|}$ was always within the SEM of the $\mu_{|LFC|}$ of the cohorts of all genes (Figure 4D). Finally, we again verified that considering only the first gene of each operon does not affect how $\mu_{|LFC|}$ and K_{TF} relate (Supplementary Figure S25).

The correlation between input and output genes responses decreases with the number of input TFs

Most input TFs discernibly affect the output genes (Supplementary Figure S14), except when $K_{TF} > 5$ (perhaps due to saturation). Nevertheless, the correlation between input and output genes appears to be decreasing with K_{TF}.

Namely, the average slopes of the fitted lines between |LFC| of the output and |LFC| of each input (Supplementary Figure S14) decreased with the K_{TF} of the output gene (Supplementary Figure S20B). Also decreasing was the R^2 between input-output pairs (Supplementary Table S7). This could explain why, when plotting |LFC| against the RNAP concentration, there is a weak trend towards increased slope with K_{TF} (Supplementary Figures S20C and S21).

The variability in single-gene absolute LFC increases with K_{TF}

We also investigated if the variability in |LFC|s, as quantified by its standard deviation $\sigma_{|LFC|}$, relates with $K_{\mathrm{TF.}}$ There should exist (at least) four sources of this variability: (a) RNA-seq measurement noise (76,77); (b) intrinsic and (c) extrinsic noise in gene expression (78,79), and (d) TF and non-TF dependent regulatory mechanisms.

Overall, we observed that, from a genome-wide perspective, $\sigma_{|LFC|}$ increases with K_{TF} (Supplementary Figure S24A) similarly to $\mu_{|LFC|}$, and the two values are also related (Supplementary Figure S24B). Examples of the variability are shown in Supplementary Figures S24C (genes with null K_{TF}), S24F (genes with two GRs, FNR and ArcA) and S24D and S24E (genes controlled by the GRs FIS or CRP) (see also Supplementary Table S14).

Other topological features of the TFN do not influence mid-term responses

Globally, the TFN of *E. coli* has in- and out-degree distributions that are well fit by power laws (Supplementary Figures S12E$_1$, S12E$_2$, S12F$_1$ and S12F$_2$) (80,81). This may explain its relatively short mean path length (Supplementary Figure S12G and Table S5).

Having established a relationship between the response kinetics and the indegree of the TFN, we next searched for correlations between |LFC| and other single-gene topological traits (Methods section *Transcription Factor Network of Escherichia coli*). We considered average shortest path length, betweenness, closeness and stress centrality, clustering coefficient, eccentricity, out-degree, neighbourhood connectivity and edge-count (51). Of these, only the clustering coefficient was statistically correlated with the |LFC| (*P*-value < 0.1) (Supplementary Table S20). However, it should not be influential, since the corresponding R^2 is nearly zero ($R^2 = 0.01$).

The numbers of activating and repressing input TFs differ in most genes

In our original hypothesis, the mid-term response (|LFC|) of a gene should follow from the bias in the numbers of activators and repressors in its set of input TFs (Figure 1B$_4$ and Supplementary Figure S11A). In detail, we predicted that if the sum of the regulatory effect (r) of the input TFs (i.e. bias $b = |\sum r|$) is null (unbiased), then the gene should have weak or zero mid-term LFC. Also, the |LFC| should increase with b.

We tested this hypothesis by extracting information on the input TFs and corresponding r values for each gene from RegulonDB. We set r of an input TF to + 1 if it is activating, to -1 if it is repressing, and to 0 if it is unknown (Supplementary Figure S12B). Then, we obtained the absolute sum of the regulatory effect of the input TFs for each gene: $|b|$.

From the data in RegulonDB, while the gene-TF interactions that are repressions and activations exist in similar numbers, the numbers of repressor TFs exist in larger numbers (Supplementary Figures S12A–C). Also, of the genes with input TFs, most ($\sim 85\%$) have a non-zero $|b|$ (Supplementary Figure S12D and Table S16).

This can explain why so many are mid-term responsive (Figure 3C), even though the genome-wide numbers of *activation* and *repression interactions* are similar (Supplementary Figure S12B). It may also explain why genes with $K_{\mathrm{TF}} \geq 1$ have higher |LFC| than genes with $K_{\mathrm{TF}} = 0$ (Figure 4A).

The bias in the input TFs follows the number of input TFs

Using information from RegulonDB, we found that the mean bias, $\mu_{|b|}$, increases with K_{TF} (Figure 5A, light blue), except for $K_{\mathrm{TF}} > 5$, which includes only $\sim$ 64 out of 4045 genes (Supplementary Table S17). The same is observed if considering only the first gene of each operon (Supplementary Figure S29).

To test if these results were affected by local topological specificities, we employed an ensemble approach (Supplementary Results section *Estimation of the expected*$\mu_{K_{TF}}$*and*$\mu_{|b|}$*using an ensemble approach*), to reduce their influence (82). We sampled genes (with replacement) to form cohorts with a given average $\mathrm{K_{TF}}$ (from 1 to 5, due to insufficient samples for higher K_{TF}). This made the relationship between $\mu_{|b|}$ and K_{TF} more stable (Figure 5A). Thus, from here onwards, we use the ensemble approach to study the influence of the global logical and topological features on the response's dynamics to the RNAP shifts.

The bias of the sets of input TFs can explain the mid-term responses of individual genes

From the data in RegulonDB, using the ensemble approach (Supplementary Results section *Estimation of the expected*$\mu_{K_{TF}}$*and*$\mu_{|b|}$*using an ensemble approach*), we formed random cohorts of genes with an imposed average $|b|$. Next, from the mid-term RNA-seq data, we calculated the average $\mu_{|LFC|}$ of the set of cohorts with a given $\mu_{|b|}$. We found that $\mu_{|LFC|}$ increases with $\mu_{|b|}$ (Figure 5B).

Interestingly, $\mu_{|b|}$ and $\mu_{K_{TF}}$ are strongly correlated in the TFN of *E. coli* (Figure 5C). To assert which one controls $\mu_{|LFC|}$, we assembled cohorts differing in $\mu_{K_{TF}}$, but not in $\mu_{|b|}$. In these, $\mu_{|LFC|}$ does not increase with $\mu_{K_{TF}}$ (Figure 5D). We also assembled cohorts differing in $\mu_{|b|}$, but not in $\mu_{K_{TF}}$. In these, $\mu_{|LFC|}$ increases with $\mu_{|b|}$ (Figure 5E). Thus, the increase of $\mu_{|b|}$ with $\mu_{K_{TF}}$ (Figure 5C), is what explains the increase in $\mu_{|LFC|}$ with $\mathrm{K_{TF}}$ (Figure 4D).

Finally, for comparison, we also investigated the relationship between $\mu_{|b|}$ and $\mu_{K_{TF}}$*prior* to the perturbation and in the short-term (at 60 min and at 125 min after shifting the medium, respectively Figure 1A). From Figure 5F, first, the

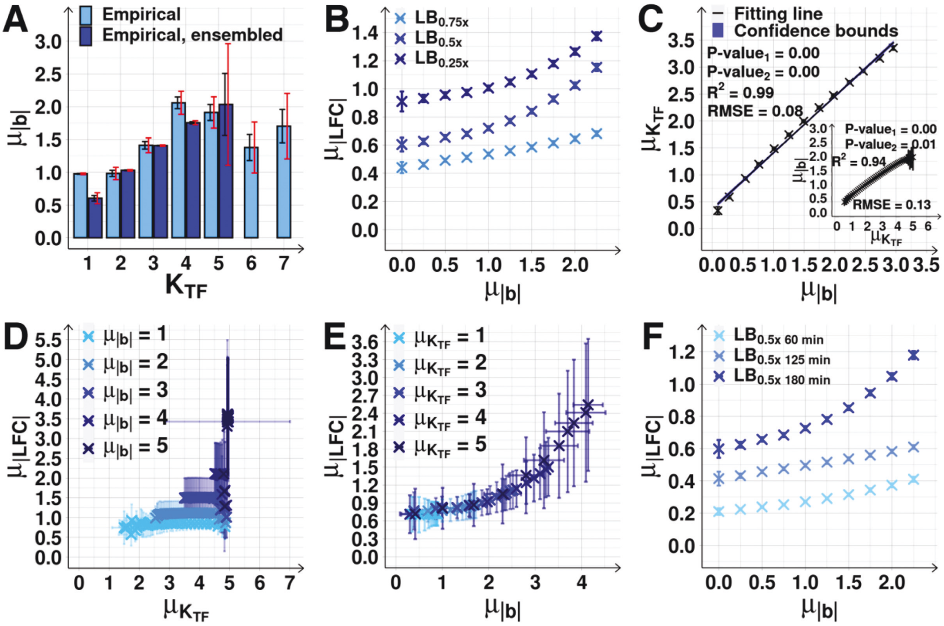


Figure 5. Effect of biases $\mu_{|b|}$ on the magnitude of the response of output genes. (**A**) $\mu_{|b|}$ as a function of $\mathrm{K_{TF}}$ (light blue) of gene cohorts with all genes (light blue) and of gene cohorts assembled using the ensemble approach (dark blue). Supplementary Table S17 shows the fractions of genes with equal $|b|$ and $\mathrm{K_{TF}}$. Black error bars are the SEM, and red error bars are the 95% CB of the SEM. Dark blue bars not shown for $\mathrm{K_{TF}} > 5$ due to small sample sizes. (**B**) Mid-term $\mu_{|LFC|}$ as a function of $\mu_{|b|}$, obtained using the ensemble approach (Supplementary Results section *Estimation of the expected*$\mu_{K_{TF}}$*and*$\mu_{|b|}$*using an ensemble approach*, Supplementary Figures S30 and S31 and Table S18). Each blue cross is the average outcome from up to 24750 cohorts of 10 genes. (**C**) $\mu_{|b|}$ plotted against the corresponding $\mu_{K_{TF}}$, mean of $\mathrm{K_{TF}}$ of the cohorts in (B). The inset shows the inverse correlation plot for the cohorts in Supplementary Figure S30, assembled based on $\mu_{K_{TF}}$ (Supplementary Results section *Estimation of the expected*$\mu_{K_{TF}}$*and*$\mu_{|b|}$*using an ensemble approach*). Shown are best fitting lines and 68% CB (shadow areas, barely visible), R^2, RMSE, and P-value (Methods section *Statistical tests c*). (**D**) $\mu_{|LFC|}$ of gene cohorts with increasing $\mu_{K_{TF}}$, but constant $\mu_{|b|}$ (from 1 to 5) (Supplementary Results section *Estimation of the expected*$\mu_{K_{TF}}$*and*$\mu_{|b|}$*using an ensemble approach*). (**E**) $\mu_{|LFC|}$ of gene cohorts with increasing $\mu_{|b|}$, but constant $\mu_{K_{TF}}$ (from 1 to 5) (Supplementary Results section *Estimation of the expected*$\mu_{K_{TF}}$*and*$\mu_{|b|}$*using an ensemble approach*). (**F**) $\mu_{|LFC|}$ as a function of $\mu_{|b|}$ prior to RNAP changes (60 min) as well as the short-term (125 min) and the mid-term responses (180 min) to RNAP changes when shifting to $\mathrm{LB_{0.5x}}$. Each blue cross is the average outcome from up to 24829 cohorts of 10 genes. In (D) and (E), the data is merged from the three conditions corresponding to (B). In all figures, the error bars are the SEM. Since the three conditions differ slightly in mean values (Figure 5B), the SEM is larger than when observing each condition separately.

$\mu_{|LFC|}$ at 125 min is stronger than at 60 min. This agrees with the expectation that shifts in RNAP suffice to shift the |LFC| of many genes. Second, the $\mu_{|LFC|}$ at 180 min is stronger than at 125 min. This agrees with our expectation that, at 125 min, input TFs numbers have not yet changed significantly in order to enhance the |LFC| of their output genes (Figure 1).

RNA numbers follow the RNAP concentration, not the medium composition

We next increased growth medium richness, instead of diluting it (Methods section *Bacterial strains, media,growth conditions andcurves, and intracellular concentrations*). As before, we limited this to not alter growth rates significantly in the first 180 min (Figure 6A and B), while altering RNAP levels (Figure 6C).

As before (Figure 4C), at mid-term, only genes directly linked by input TFs showed correlation in their |LFC| (Figure 6D$_1$–D$_3$ and Supplementary Figure S32). This supports the previous assumption on the kinetics of transcription, translation, and signals propagation via shifts in input TFs numbers (Figure 1).

Meanwhile, in contrast to above, shifting cells from $\mathrm{LB_{1.0x}}$ to the richer $\mathrm{LB_{1.5x}}$ medium was accompanied by a decrease in the RNAP concentration (Figure 7A). This was followed by substantial alterations in the RNA populations, with a large number of DEGs and high $\mu_{|LFC|}$(Figure 7B and C, respectively). As previously, in the mid-term, genes with input TFs reacted more strongly (Figure 7C).

These results support the initial assumption that the changes in RNA abundances follow the RNAP concentration, rather than the medium richness.

Further increases in medium richness do not decrease RNAP concentration and RNA numbers also do not change

Finally, we further increased growth-medium richness (to $\mathrm{LB_{2.0x}}$ and to $\mathrm{LB_{2.5x}}$). This caused no significant change in RNAP levels and concentration (Figures 6C and 7A). Here, we also did not observe significant changes in DEGs or $\mu_{|LFC|}$ at mid-term, when compared with the $\mathrm{LB_{1.5x}}$ con-

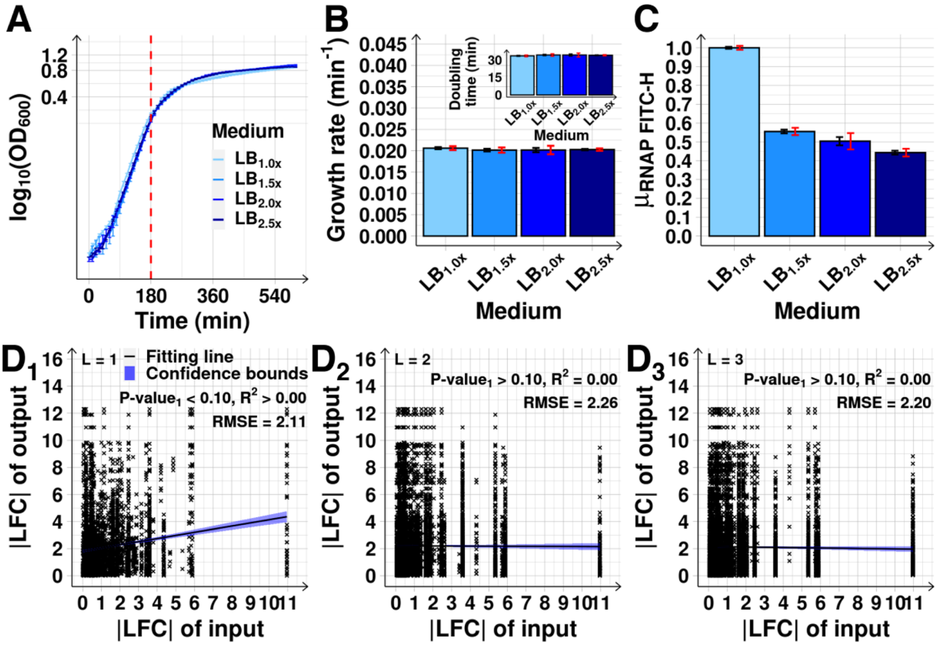


Figure 6. RNAP levels following increasing medium richness and corresponding relationships between |LFC|s of pairs of genes separated by specific path lengths, L. (**A**) Growth curves from OD_{600} assessed every 10 min (Methods section *Bacterial strains, media,growth conditions andcurves, and intracellular concentrations*), following each medium shift. (**B**) Growth rates at 180 min after medium enrichment. The inset shows the corresponding doubling times. (**C**) Mean RNAP levels relative to the control estimated from single-cell RNAP-GFP fluorescence intensities (FITC-H) ($\mu_{RNAP\ FITC-H}$). (**D_1–D_3**) Scatter plots between absolute LFC (|LFC|) of outputs and corresponding input genes distanced by L (path length) of 1, 2 and 3 transcription factors, respectively. Data from the $LB_{2.5x}$ condition. Shown are the best fitting line and its 68% CB (blue shadow), and the R^2 and RMSE of the fitted regression line, along with its P-value at 10% significance level under the null hypothesis that this line is horizontal. From (A) to (C), the black error bars are the SEM and red error bars represent the 95% CB of the SEM.

dition (Figure 7B and C, respectively). This is in agreement with the assumption that the shifts in the RNAP concentration caused the short-term changes in RNA abundances, which then caused the mid-term changes.

Finally, as before, $\mu^{TF}_{|LFC|}$ follows $\mu_{|b|}$ (Figure 7D and Supplementary Figure S33). Moreover, it does so almost identically in the three perturbations, as expected from the original assumptions (Figure 1).

DISCUSSION

We investigated if the mid-term responses to genome-wide perturbations of *E. coli*'s TFN are mediated by its topology and logic. We diluted LB medium since this dramatically and reproducibly affects the RNAP concentration (26,27). The increasingly strong nature of the dilutions facilitated the verification of how the RNAP concentration and single-gene, mid-term |LFC|s related. We focused on mid-term transcriptional responses (Figure 1), since short-term responses are unlikely to have been influenced by the TFN due to protein folding and maturation times, etc. Meanwhile, long-term responses were most likely affected by the TFN. However, dissecting them would have been onerous, due to the complicating effects of loss, backpropagation, and coalescence of possibly dozens of signals from origins other than direct input TFs.

We lack information on the affinity between each gene and their input TFs, on how the input TFs operate, and on how the *de novo* presence of an input TF alters the binding or activity of other input TFs on the same promoter. Thus, we would have failed to predict the behaviour of individual genes with accuracy. Instead, we predicted the responses of gene cohorts, since their behaviour is less influenced by particular single-gene features (other than cohort-specific features), which should average out at the cohort level. Further, as in (18), we were only able to correlate *absolute* LFCs of input and output genes (Figure 4B), likely due to limitations in RNA-seq technology and the analysis, and/or missing information on the TFN. Nevertheless, the present information on input TFs and their regulatory effect sufficed to relate the TFN with the genes' response.

From the RNA-seq data on three time points, we provided evidence that both the TFN and the RNAP affect the results at mid-term (∼ 180 min), and not before that. In addition, while other factors also influenced genes' behaviour at mid-term, including single-gene features, they only had minor, local effects. In detail, first, we could not find evidence of GRs (including σ^{38}) and (p)ppGpp being material in the global mid-term behaviour (although

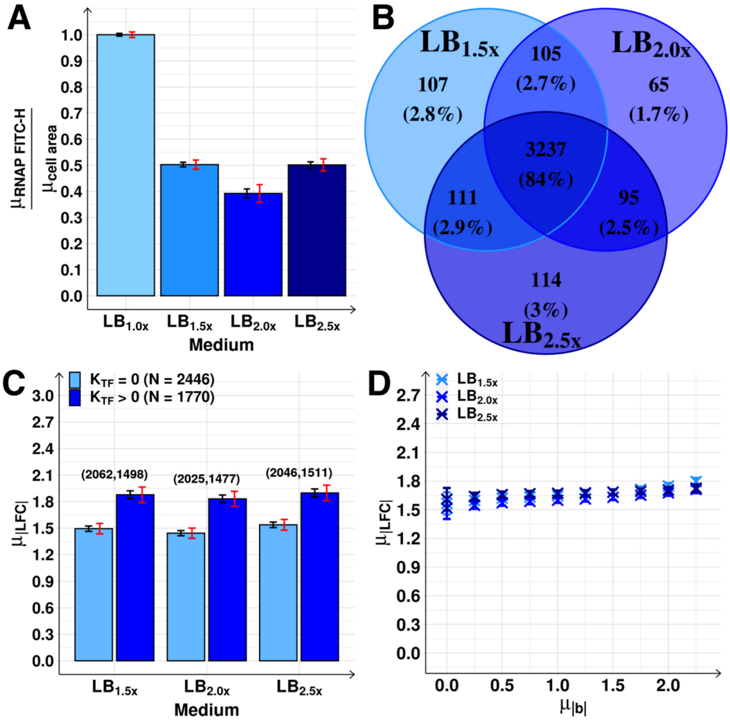


Figure 7. Genome-wide effects of increasing medium richness. (**A**) RNAP concentrations relative to the control, estimated from $\mu_{RNAPFITC-H}$ divided by mean cell area ($\mu_{cellarea}$, used as a proxy for cell volume). (**B**) Venn diagrams of the DEG. (**C**) $\mu_{|LFC|}$ of N genes with K_{TF} equal to and larger than 0, following each medium shift. Above each bar are the number of DEG. (**D**) $\mu_{|LFC|}$ as a function of $\mu_{|b|}$ after the growth-medium shifts. $\mu_{|LFC|}$ obtained using the ensemble approach (Supplementary Results section *Estimation of the expected*$\mu_{K_{TF}}$*and*$\mu_{|b|}$*using an ensemble approach*, Supplementary Figure S33). Each blue cross is the average outcome from up to 24536 cohorts of 10 genes. In (A) and (C), the black error bars are the SEM and the red error bars are the 95% CB of the SEM. In (D), the small error bars are the SEM (most not visible).

(p)ppGpp may be significant in the short-term response). Second, we excluded the medium as directly influencing RNA abundances. Third, we excluded global network parameters, other than K_{TF}, as being influential as well (none of them correlated to single-gene responses). Fourth, we did not find evidence for significant translational or post-translational regulation. Namely, RNA and protein abundances correlated well, and so did the RNA levels of input TFs and of output genes. Finally, sRNAs did not respond atypically to the RNAP shifts neither in the short-term, nor in in the mid-term.

We have made six key observations on the influence of the logic and topology of the TFN on the mid-term response. First, genes without input TFs were less responsive. Second, the |LFC| of input and output genes correlated positively. Thus, we argue that, on average, input TFs enhanced the |LFC| of individual genes. Third, only nearest neighbour genes in the TFN consistently correlated in |LFC|'s. Thus, either the effects of the shift in RNAP only reached nearest neighbour genes or they 'dissipated' beyond that. Since the correlations between nearest neighbours were weaker in the short-term than in the mid-term, the first possibility is more likely. This observation also suggests that there is a degree of genome-wide homogeneity in how long input TF abundance take to change (likely due to physical limitations on the rates constants controlling bacterial gene expression). This agrees with the constraints on timing variability reported in (6). Fourth, the behaviour was orderly (rather than chaotic), with most genes responsive to the weak perturbations also responding to the stronger perturbations. This suggests the existence of features (on genes and/or the TFN) affecting the responsiveness (Supplementary Figure S34). Similarly, there is a good overlap between the sets of genes responsive in the short- and in the mid-term, but weak overlap to those responsive prior to the perturbation (Supplementary Figure S35). Fifth, on average, as K_{TF} increased, the correlation between the input and each output gene decreased. This is likely unavoidable and may be a limiting factor in how many input TFs genes can have. Finally, it is $\mu_{|b|}$ that (partially) controls the genes' responsiveness to the stress, while the apparent relationship between $\mu_{K_{TF}}$ and $\mu_{|LFC|}$ is due to the linear correlation between $\mu_{|b|}$ and $\mu_{K_{TF}}$. Nevertheless, the possible values of $\mu_{|b|}$ are limited by the values of $\mu_{K_{TF}}$.

These observations provide direct empirical evidence that the genome-wide, mid-term, transcriptional stress responsiveness of *E. coli* depends on a global topological feature of its TFN. Namely, by exploiting the known features of the TFN of *E. coli*, we showed that the bias in the regulatory effect of input TFs of gene cohorts, $\mu_{|b|}$, acts as a major determinant of their mean response to a stress. So far, the existence of influence of global topological features has only been supported by theoretical models, e.g. (82,83), and by *indirect* empirical evidence, i.e. by the observation that the abundance of input TFs correlates with the activity of their direct output genes (18,20–22,70,84). Because of these observations, it has been assumed that the topology and logic of TFNs should play a role. Here, we provided direct evidence of this, i.e. that the TFN topology and logic are major contributors to a global mid-term response. Our approach, relating genome-wide dynamics and global topological features, should now be applied to other genome-wide stresses, as well as to the genomes of other organisms, when data on their TFN becomes sufficient to permit such an analysis.

Expanding this research may inform on how to improve the robustness and plasticity of synthetic circuits. Further, as suggested in (20), bacteria subjected to stress, rather than under optimal conditions, may be a better proxy of their state when infecting a host. Thus, imposing stresses may be a valuable strategy to identify new target genes for antibiotics for disrupting bacterial adaptability to new conditions. The use of medium dilution as a genome-wide stress is a good proxy for nutrient imbalance, and we identified ∼ 900 responsive genes, even for moderate nutritional stress, of which only 58 are essential under optimal conditions. Plausibly, some of the responsive genes, particularly those responsive to all 3 medium dilutions, may be essential to adapt to poorer media, and thus are potential new drug targets. Conversely, it may be possible to tune these genes to assist in the performance of metabolic tasks, without disturbing the basic biology of the cells. As such, they are promising targets for modifications that could improve the yield and sustainability of bio-industrial processes.

Finally, our findings can be used to develop new large-scale, dynamic models of gene networks. These models should be able to predict short- and mid-term transcriptional responses of gene cohorts to genome-wide perturbations of transcription activities. The short-term responses should be mostly controlled by single-gene features (e.g. RNAP-promoter binding affinity). Meanwhile, the mid-term responses should be heavily influenced by the topology and logic of the TFN. The new dynamic models could be developed starting from the schematic and predictive models in Figures 1B and C, respectively, using empirical kinetic parameters of bacterial RNA and protein production and degradation.

Such dynamic models could then be used to predict how natural TFNs perform complex transcriptional programs, and how these programs can be modified to achieve desired goals. Interesting models involving many genes include models of programs responsive to environmental shifts, antibiotics, etc. These models could assist in identifying critical elements of the TFN during short-, mid-, and potentially long-term stress responses. These efforts should be facilitated by the ongoing information gathering on single-gene features (34,85–87), including on microorganisms other than *E. coli*.

DATA AVAILABILITY

RNA-seq *.fastq data (trimmed) and processed RNA-seq data are deposited in NCBI GEO with accession code GSE178281 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178281). The raw data of the control condition (180 min, $LB_{1.0x}$) was also used in (72) (GSE183139). In Dryad, we deposited a package with flow-cytometry, microscopy, spectrophotometry, and western blot data (DOI: 10.5061/dryad.wh70rxwnp). The package also has two *.xlsx files informing on the genes TFN topological and logical parameters, RNA-seq expression, global network topological features, gene-gene interactions, and the lists of pairs of genes separated by path lengths from 1 to 8.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author contributions: A.S.R., B.L.B.A. and S.M.D.O. conceived the study and A.S.R. and B.L.B.A. directed it. B.L.B.A. planned, directed and executed data analysis, to which C.S.D.P. and I.S.C.B. contributed to. M.N.M.B., A.S.R. and B.L.B.A. interpreted and integrated the data. V.C., S.D. and V.K. performed measurements. A.H., H.T.J. and E.D. assisted the data analysis. A.H., J.L.P., O.P.S., A.G., M.N., J.K. and P.A. assisted on the RNA-seq data analysis. B.L.B.A. and A.S.R., assisted by H.T.J., drafted the manuscript, supplementary information, figures and tables, which were revised by all co-authors. M.N.M.B., V.C. and S.D. contributed equally as second author and have the right to list their name first.

FUNDING

Jane and Aatos Erkko Foundation [10-10524-38 to A.S.R.]; Finnish Cultural Foundation [50201300 to S.D., and 00200193, 00212591 to I.S.C.B.]; EDUFI Fellowship [TM-19-11105 to S.D.]; Suomalainen Tiedeakatemia (to C.S.D.P.); Tampere University Graduate Program [to V.C., M.N.M.B., B.L.B.A.]; TTU development program 2016–2022 [2014-2020.4.01.16-0032 to O.P.S.]; Academy of Finland [307856 and 323576 to P.A., 322927 to A.H., and 272376 and 256615 to H.T.J.]. Funding for open access and for article processing charges provided by Tampere University Library [to B.L.B.A.]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Conflict of interest statement. None declared.

REFERENCES

1. Jacob,F. and Monod,J. (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, **3**, 318–356.
2. Mejía-Almonte,C., Busby,S.J.W., Wade,J.T., van Helden,J., Arkin,A.P., Stormo,G.D., Eilbeck,K., Palsson,B.O., Galagan,J.E. and Collado-Vides,J. (2020) Redefining fundamental concepts of transcription initiation in bacteria. *Nat. Rev. Genet.*, **21**, 699–714.

3. Sanchez-Vazquez,P., Dewey,C.N., Kitten,N., Ross,W. and Gourse,R.L. (2019) Genome-wide effects on escherichia coli transcription from ppGpp binding to its two sites on RNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 8310–8319.
4. Kao,K.C., Yang,Y.L., Boscolo,R., Sabatti,C., Roychowdhury,V. and Liao,J.C. (2004) Transcriptome-based determination of multiple transcription regulator activities in escherichia coli by using network component analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 641–646.
5. Jozefczuk,S., Klie,S., Catchpole,G., Szymanski,J., Cuadros-Inostroza,A., Steinhauser,D., Selbig,J. and Willmitzer,L. (2010) Metabolomic and transcriptomic stress response of escherichia coli. *Mol. Syst. Biol.*, **6**, 364.
6. Mitosch,K., Rieckh,G. and Bollenbach,T. (2019) Temporal order and precision of complex stress responses in individual bacteria. *Mol. Syst. Biol.*, **15**, e8470.
7. Phadtare,S. and Inouye,M. (2004) Genome-wide transcriptional analysis of the cold shock response in wild-type and cold-sensitive, quadruple-csp-deletion strains of escherichia coli. *J. Bacteriol.*, **186**, 7007–7014.
8. Dash,S., Palma,C.S.D., Baptista,I.S.C., Bahrudeen,M.N.M., Almeida,B.L.B., Chauhan,V., Jagadeesan,R. and Ribeiro,A.S. (2021) Positive supercoiling buildup is a trigger of *e. coli*'s short-term response to cold shock. bioRxiv doi: https://doi.org/10.1101/2021.12.22.473827, 23 December 2021, preprint: not peer reviewed.
9. Rau,M.H., Calero,P., Lennen,R.M., Long,K.S. and Nielsen,A.T. (2016) Genome-wide escherichia coli stress response and improved tolerance towards industrially relevant chemicals. *Microb. Cell Fact.*, **15**, 176.
10. Kochanowski,K., Gerosa,L., Brunner,S.F., Christodoulou,D., Nikolaev,Y.V. and Sauer,U. (2017) Few regulatory metabolites coordinate expression of central metabolic genes in escherichia coli. *Mol. Syst. Biol.*, **13**, 903.
11. Engl,C., Jovanovic,G., Brackston,R.D., Kotta-Loizou,I. and Buck,M. (2020) The route to transcription initiation determines the mode of transcriptional bursting in e. coli. *Nat. Commun.*, **11**, 2422.
12. Deter,H.S., Hossain,T. and Butzin,N.C. (2021) Antibiotic tolerance is associated with a broad and complex transcriptional response in e. coli. *Sci. Rep.*, **11**, 6112.
13. Lal,A., Dhar,A., Trostel,A., Kouzine,F., Seshasayee,A.S. and Adhya,S. (2016) Genome scale patterns of supercoiling in a bacterial chromosome. *Nat. Commun.*, **7**, 11055.
14. Storz,G., Vogel,J. and Wassarman,K.M. (2011) Regulation by small RNAs in bacteria: expanding frontiers. *Mol. Cell*, **43**, 880–891.
15. Keren,L., Zackay,O., Lotan-Pompan,M., Barenholz,U., Dekel,E., Sasson,V., Aidelberg,G., Bren,A., Zeevi,D., Weinberger,A. *et al.* (2013) Promoters maintain their relative activity levels under different growth conditions. *Mol. Syst. Biol.*, **9**, 701.
16. Gerosa,L., Kochanowski,K., Heinemann,M. and Sauer,U. (2013) Dissecting specific and global transcriptional regulation of bacterial gene expression. *Mol. Syst. Biol.*, **9**, 658.
17. Berthoumieux,S., de Jong,H., Baptist,G., Pinel,C., Ranquet,C., Ropers,D. and Geiselmann,J. (2013) Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Mol. Syst. Biol.*, **9**, 634.
18. Larsen,S.J., Röttger,R., Schmidt,H.H.H.W. and Baumbach,J. (2019) E. coli gene regulatory networks are inconsistent with gene expression data. *Nucleic Acids Res.*, **47**, 85–92.
19. Brooks,A.N., Reiss,D.J., Allard,A., Wu,W.J., Salvanha,D.M., Plaisier,C.L., Chandrasekaran,S., Pan,M., Kaur,A. and Baliga,N.S. (2014) A system-level model for the microbial regulatory genome. *Mol. Syst. Biol.*, **10**, 740.
20. Côté,J.P., French,S., Gehrke,S.S., MacNair,C.R., Mangat,C.S., Bharat,A. and Brown,E.D. (2016) The genome-wide interaction network of nutrient stress genes in escherichia coli. *Mbio*, **7**, e01714-16.
21. Fang,X., Sastry,A., Mih,N., Kim,D., Tan,J., Yurkovich,J.T., Lloyd,C.J., Gao,Y., Yang,L. and Palsson,B.O. (2017) Global transcriptional regulatory network for escherichia coli robustly connects gene expression to transcription factor activities. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 10286–10291.
22. Urchueguía,A., Galbusera,L., Chauvin,D., Bellement,G., Julou,T. and van Nimwegen,E. (2021) Genome-wide gene expression noise in escherichia coli is condition-dependent and determined by propagation of noise through the regulatory network. *PLoS Biol.*, **19**, e3001491.
23. Bremer,H.D.P.P. and Dennis,P.P. (1996) Modulation of chemical composition and other parameters of the cell by growth rate. In: Neidhardt,F.C. (ed). *Escherichia Coli and Salmonella: Cellular and Molecular Biology*. ASM Press, Washington, D.C., pp. 1553–1569.
24. McClure,W.R. (1985) Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.*, **54**, 171–204.
25. Taniguchi,Y., Choi,P.J., Li,G.W., Chen,H., Babu,M., Hearn,J., Emili,A. and Xie,X.S. (2010) Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, **329**, 533–538.
26. Lloyd-Price,J., Startceva,S., Kandavalli,V., Chandraseelan,J.G., Goncalves,N., Oliveira,S.M., Häkkinen,A. and Ribeiro,A.S. (2016) Dissecting the stochastic transcription initiation process in live escherichia coli. *DNA Res.*, **23**, 203–214.
27. Kandavalli,V.K., Tran,H. and Ribeiro,A.S. (2016) Effects of σ factor competition are promoter initiation kinetics dependent. *Biochim. Biophys. Acta*, **1859**, 1281–1288.
28. Bernstein,J.A., Khodursky,A.B., Lin,P.H., Lin-Chao,S. and Cohen,S.N. (2002) Global analysis of mRNA decay and abundance in escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 9697–9702.
29. Maurizi,M.R. (1992) Proteases and protein degradation in escherichia coli. *Experientia*, **48**, 178–201.
30. Hebisch,E., Knebel,J., Landsberg,J., Frey,E. and Leisner,M. (2013) High variation of fluorescence protein maturation times in closely related escherichia coli strains. *PLoS One*, **8**, e75991.
31. Balleza,E., Kim,J.M. and Cluzel,P. (2018) Systematic characterization of maturation time of fluorescent proteins in living cells. *Nat. Methods*, **15**, 47–51.
32. deHaseth,P.L., Zupancic,M.L. and Record,M.T. (1998) RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. *J. Bacteriol.*, **180**, 3019–3025.
33. Mazumder,A. and Kapanidis,A.N. (2019) Recent advances in understanding σ70-dependent transcription initiation mechanisms. *J. Mol. Biol.*, **431**, 3947–3959.
34. Santos-Zavaleta,A., Salgado,H., Gama-Castro,S., Sánchez-Pérez,M., Gómez-Romero,L., Ledezma-Tejeida,D., García-Sotelo,J.S., Alquicira-Hernández,K., Muñiz-Rascado,L.J., Peña-Loredo,P. *et al.* (2019) RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in e. coli K-12. *Nucleic Acids Res.*, **47**, D212–D220.
35. Patange,O., Schwall,C., Jones,M., Villava,C., Griffith,D.A., Phillips,A. and Locke,J.C.W. (2018) Escherichia coli can survive stress by noisy growth modulation. *Nat. Commun.*, **9**, 5333.
36. Zaslaver,A., Bren,A., Ronen,M., Itzkovitz,S., Kikoin,I., Shavit,S., Liebermeister,W., Surette,M.G. and Alon,U. (2006) A comprehensive library of fluorescent transcriptional reporters for escherichia coli. *Nat. Methods*, **3**, 623–628.
37. Chantzoura,E. and Kaji,K. (2017) Chapter 10 - flow cytometry. In: Jalali,M., Saldanha,F.Y. and Jalali,M. (eds). *Basic Science Methods for Clinical Researchers*. Academic Press, Boston, pp. 173–189.
38. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
39. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nat. Methods*, **9**, 357–359.
40. Liao,Y., Smyth,G.K. and Shi,W. (2019) The r package rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.*, **47**, e47.
41. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
42. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B (Methodological)*, **57**, 289–300.
43. Evans,C., Hardin,J. and Stoebel,D.M. (2018) Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform*, **19**, 776–792.
44. Spaniol,V., Wyder,S. and Aebi,C. (2013) RNA-Seq-based analysis of the physiologic cold shock-induced changes in moraxella catarrhalis gene expression. *PLoS One*, **8**, e68298.
45. Hazen,T.H., Daugherty,S.C., Shetty,A., Mahurkar,A.A., White,O., Kaper,J.B. and Rasko,D.A. (2015) RNA-Seq analysis of isolate- and

growth phase-specific differences in the global transcriptomes of enteropathogenic *Escherichia coli* prototype isolates. *Front Microbiol*, **6**, 569.
46. Yung,P.Y., Grasso,L.L., Mohidin,A.F., Acerbi,E., Hinks,J., Seviour,T., Marsili,E. and Lauro,F.M. (2016) Global transcriptomic responses of escherichia coli K-12 to volatile organic compounds. *Sci. Rep.*, **6**, 19899.
47. Yanofsky,C. (1981) Attenuation in the control of expression of bacterial operons. *Nature*, **289**, 751–758.
48. Proshkin,S., Rahmouni,A.R., Mironov,A. and Nudler,E. (2010) Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science*, **328**, 504–508.
49. Dahan,O., Gingold,H. and Pilpel,Y. (2011) Regulatory mechanisms and networks couple the different phases of gene expression. *Trends Genet.*, **27**, 316–322.
50. Albert,B., Johnson,A., Lewis,J., Raff,M., Roberts,K. and Walter,P. (2008) In: *Molecular Biology of the Cell*. Garland Science, NY.
51. Doncheva,N.T., Assenov,Y., Domingues,F.S. and Albrecht,M. (2012) Topological analysis and interactive visualization of biological networks and protein structures. *Nat. Protoc.*, **7**, 670–685.
52. Bahrudeen,M.N.M., Chauhan,V., Palma,C.S.D., Oliveira,S.M.D., Kandavalli,V.K. and Ribeiro,A.S. (2019) Estimating RNA numbers in single cells by RNA fluorescent tagging and flow cytometry. *J. Microbiol. Methods*, **166**, 105745.
53. McDonald,J.H. (2009) In: *Analysis of covariance. In: Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, pp. 232–238.
54. Farewell,A., Kvint,K. and Nyström,T. (1998) Negative regulation by rpos: a case of sigma factor competition. *Mol. Microbiol.*, **29**, 1039–1051.
55. Chang,D.E., Smalley,D.J. and Conway,T. (2002) Gene expression profiling of escherichia coli growth transitions: an expanded stringent response model. *Mol. Microbiol.*, **45**, 289–306.
56. Tani,T.H., Khodursky,A., Blumenthal,R.M., Brown,P.O. and Matthews,R.G. (2002) Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 13471–13476.
57. Baptista,I.S.C., Kandavalli,V., Chauhan,V., Bahrudeen,M.N.M., Almeida,B.L.B., Palma,C.S.D., Dash,S. and Ribeiro,A.S. (2022) Sequence-dependent model of genes with dual σ factor preference. *Biochim. Biophys. Acta Gene Regul. Mech.*, **1865**, 194812.
58. Klumpp,S., Zhang,Z. and Hwa,T. (2009) Growth rate-dependent global effects on gene expression in bacteria. *Cell*, **139**, 1366–1375.
59. Huh,D. and Paulsson,J. (2011) Non-genetic heterogeneity from stochastic partitioning at cell division. *Nat. Genet.*, **43**, 95–100.
60. Lange,R. and Hengge-Aronis,R. (1994) The cellular concentration of the sigma s subunit of RNA polymerase in escherichia coli is controlled at the levels of transcription, translation, and protein stability. *Genes Dev.*, **8**, 1600–1612.
61. Gruber,T.M. and Gross,C.A. (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.*, **57**, 441–466.
62. Pratt,L.A. and Silhavy,T.J. (1998) Crl stimulates RpoS activity during stationary phase. *Mol. Microbiol.*, **29**, 1225–1236.
63. Nyström,T. (2004) Growth versus maintenance: a trade-off dictated by RNA polymerase availability and sigma factor competition. *Mol. Microbiol.*, **54**, 855–862.
64. Buhler,J.M., Riva,M., Mann,C., Thuriaux,P., Memet,S., Micouin,J.Y., Treich,I., Mariotte,S., Sentenac,A., Reznekoff,W.S. *et al.* (1987) In: *RNA Polymerase and the Regulation of Transcription*. Elsevier Science Publishing Co., NY, pp. 25–36.
65. Nishiuchi,Y., Inui,T., Nishio,H., Bódi,J., Kimura,T., Tsuji,F.I. and Sakakibara,S. (1998) Chemical synthesis of the precursor molecule of the aequorea green fluorescent protein, subsequent folding, and development of fluorescence. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 13549–13554.
66. Severinov,K., Mooney,R., Darst,S.A. and Landick,R. (1997) Tethering of the large subunits of escherichia coli RNA polymerase. *J. Biol. Chem.*, **272**, 24137–24140.
67. 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. Res.*, **25**, 1203–1210.
68. Gardner,T.S., Cantor,C.R. and Collins,J.J. (2000) Construction of a genetic toggle switch in escherichia coli. *Nature*, **403**, 339–342.
69. Lutz,R., Lozinski,T., Ellinger,T. and Bujard,H. (2001) Dissecting the functional program of escherichia coli promoters: the combined mode of action of lac repressor and AraC activator. *Nucleic. Acids. Res.*, **29**, 3873–3881.
70. Martínez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.*, **6**, 482–489.
71. Balderas-Martínez,Y.I., Savageau,M., Salgado,H., Pérez-Rueda,E., Morett,E. and Collado-Vides,J. (2013) Transcription factors in escherichia coli prefer the holo conformation. *PLoS One*, **8**, e65723.
72. Chauhan,V., Bahrudeen,M.N.M., Palma,C.S.D., Baptista,I.S.C., Almeida,B.L.B., Dash,S., Kandavalli,V. and Ribeiro,A.S. (2022) Analytical kinetic model of native tandem promoters in e. coli. *PLoS Comput. Biol.*, **18**, e1009824.
73. Mushegian,A.R. and Koonin,E.V. (1996) Gene order is not conserved in bacterial evolution. *Trends Genet.*, **12**, 289–290.
74. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
75. Wells,J.N., Bergendahl,L.T. and Marsh,J.A. (2016) Operon gene order is optimized for ordered protein complex assembly. *Cell Rep.*, **14**, 679–685.
76. Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.
77. Haas,B.J., Chin,M., Nusbaum,C., Birren,B.W. and Livny,J. (2012) How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes. *BMC Genomics*, **13**, 734.
78. Elowitz,M.B., Levine,A.J., Siggia,E.D. and Swain,P.S. (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.
79. Engl,C. (2019) Noise in bacterial gene expression. *Biochem. Soc. Trans.*, **47**, 209–217.
80. Dobrin,R., Beg,Q.K., Barabási,A.L. and Oltvai,Z.N. (2004) Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. *BMC Bioinf.*, **5**, 10.
81. Martínez-Antonio,A. (2011) Escherichia coli transcriptional regulatory network. *Network Biology*, **1**, 21.
82. Ribeiro,A.S., Kauffman,S.A., Lloyd-Price,J., Samuelsson,B. and Socolar,J.E. (2008) Mutual information in random boolean models of regulatory networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **77**, 011901.
83. Samuelsson,B. and Socolar,J.E. (2006) Exhaustive percolation on random networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **74**, 036113.
84. Gutiérrez-Ríos,R.M., Rosenblueth,D.A., Loza,J.A., Huerta,A.M., Glasner,J.D., Blattner,F.R. and Collado-Vides,J. (2003) Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res.*, **13**, 2435–2443.
85. Reed,J.L., Vo,T.D., Schilling,C.H. and Palsson,B.O. (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.*, **4**, R54.
86. Ishihama,A., Shimada,T. and Yamazaki,Y. (2016) Transcription profile of *Escherichia coli*: genomic SELEX search for regulatory targets of transcription factors. *Nucleic. Acids. Res.*, **44**, 2058–2074.
87. Gao,Y., Lim,H.G., Verkler,H., Szubin,R., Quach,D., Rodionova,I., Chen,K., Yurkovich,J.T., Cho,B.K. and Palsson,B.O. (2021) Unraveling the functions of uncharacterized transcription factors in *Escherichia coli* using chip-exo. *Nucleic. Acids. Res.*, **49**, 9696–9710.

Supplementary Information for

The transcription factor network of *E. coli* steers global responses to shifts in RNAP concentration

Bilena L B Almeida*, Mohamed N M Bahrudeen†, Vatsala Chauhan†, Suchintak Dash†, Vinodh Kandavalli, Antti Häkkinen, Jason Lloyd-Price, Cristina S D Palma, Ines S C Baptista, Abhishekh Gupta, Juha Kesseli, Eric Dufour, Olli-Pekka Smolander, Matti Nykter, Petri Auvinen, Howard T Jacobs, Samuel M D Oliveira, and Andre S Ribeiro*

† Equal contributions.

* To whom correspondence should be addressed. E-mails: andre.sanchesribeiro@tuni.fi, bilena.limadebritoalmeida@tuni.fi

SUPPLEMENTARY RESULTS

Expected effects of shifting RNA polymerase concentration on a gene's transcription dynamics

Present knowledge of transcription in *E. coli* suggests that this is a multi-step, highly regulated process, which usually can be well approximated by a two-step model (Reactions S1).

The regulators are usually RNAP, σ factors, and, in many cases, gene-specific input TFs, including global regulators. The two rate constants in (S1) differ with the input TF's concentration. Depending on them, in some conditions, transcription will be blocked, while in others it will be enhanced, compared to a basal rate (1).

According to (S1), an RNAP can find the promoter (*Pro*) of gene α , which will be in state β at that moment due to a specific set of bound/unbound input TFs, that control the rates of both closed (CC) and subsequent open complex (OC) formations. Supplementary Figure S11A illustrates forms of TF-promoter interactions that affect the propensity of transcription.

Here, when binding to the transcription start site of the promoter region, the RNAP will attempt to form a CC with the DNA (2), via a reversible process. $k_{\alpha,\beta}^{cc,F}$ and $k_{\alpha,\beta}^{cc,B}$ are the forward and backward rate constants of CC formation. Once forming a CC, the RNAP can commit to OC at the rate $k_{\alpha,\beta}^{oc}$, after which initiation is nearly irreversible. It follows elongation (which frees the promoter) and RNA completion, which frees the RNAP (1).

$$Pro_{\alpha}^{\beta(TF_1,...)} + RNAP \underset{k_{\alpha,\beta(TF)}^{cc,B}}{\overset{k_{\alpha,\beta(TF)}^{cc,F}}{\rightleftarrows}} CC_{\alpha}^{\beta(TF_1,...)} \xrightarrow{k_{\alpha,\beta(TF)}^{oc}} Pro_{\alpha}^{\beta(TF_1,...)} + RNAP + RNA_{\alpha} \qquad \text{(S1)}$$

The effective rate of transcription, which defines the promoter strength, differs with both $k_{\alpha,\beta}^{cc}$ and $k_{\alpha,\beta}^{oc}$ which then, combined with a reaction for RNA degradation, controls the mean RNA levels. Given the fast degradation rates of RNA (~1-2 min (3)), changes in mean RNA levels are expected to be quick once the transcription rate is altered.

According to (S1), changing RNAP concentration will particularly affect the kinetics of genes with long lasting CC.

Finally, RNA degradation as well as RNA dilution with cell division (reaction S2) is usually modelled as a single-step process since evidence suggests that it can be well fitted by a single exponential function (4). This step is not expected to be subject to significant sequence-specific regulation (3).

$$RNA_{\alpha} \xrightarrow{k_{\alpha}^{\deg}+k_{\alpha}^{dil}} \varnothing \qquad \text{(S2)}$$

where k_{α}^{deg} is the RNA degradation rate and k_{α}^{dil} is the RNA dilution rate.

Given the above, shifting RNA polymerase (RNAP) concentration should quickly change the RNA abundances of many genes, particularly those whose step of closed complex formation has a significant time length. The shifts in RNA abundances then propagate to protein abundances, via translation.

Other sources of variability in single-gene response will then emerge from single-TF properties, such as different input TFs binding affinities, and folding and maturation times, etc.

This model is presented here to facilitate the interpretation of the results, and it does not intend to be a strick representation of *all* events occurring during the genome-wide stresses. As an example, this model would fail to capture the influence of ppGpp(p), non-coding RNAs, positive supercoiling buildup, events during transcription, and post-transcription and post-translation regulation, among other.

Estimation of $\mu_{K_{TF}}$ and $\mu_{|b|}$ from empirical data using an ensemble approach

The behavior of networks as a function of specific topological features can be studied using an ensemble approach. This approach consists of generating model networks imposing some features (e.g., total number of connected nodes) in order to study their relevance, while other features are randomly generated (5) to decrease the chances that the conclusions generalize poorly. This assists the study of responses to a global perturbation, since it is complex to filter abnormal responses, due to lack of knowledge about the range of possible behaviors. Also, it is not easy to establish if the ‘abnormality’ is in an arbitrary local topological feature of the genes of interest, or in the response due to an arbitrary internal parameter. Overall, we expect this methodology to decrease the effects of arbitrary features (but not necessarily all effects or for all features).

In our hypothesis, the perturbation strength due to shifting RNAP concentration ($\mu_{|LFC|}$) on a gene is a function of (and, thus, can be predicted from) $\mu_{|b|}$ of the input TFs of that gene. However, since other ‘local’ factors can affect the $\mu_{|LFC|}$ of a gene (including its original state, its sensitivity to supercoiling build up, etc., we only expect our hypothesis to be true at the level of gene cohorts.

Similarly, for large enough gene cohorts, we hypothesize that $\mu_{|LFC|}$ can be predicted from $\mu_{|b|}$ alone. Thus, we compared the behavior of cohorts of genes differing in $|b|$ using an ensemble approach. For this, we proceeded by comparing genes differing in $|b|$. Thus, we randomly selected genes, from the empirical data, to form cohorts with a given $\mu_{|b|}$. We then compared the average behavior (i.e., $\mu_{|LFC|}$) of these sets of cohorts with a given $\mu_{|b|}$.

This was done in two ways (Supplementary Figure S30 and Figure 5B respectively). In one case, we assembled cohorts of genes with a given $\mu_{K_{TF}}$ (average K_{TF}), which obligatorily results in a given $\mu_{|b|}$ (Figure 5C). In the other, we assembled cohorts of genes with a given $\mu_{|b|}$. We found no difference in the results using the two ensemble methods (note the similarities between Supplementary Figure S30 ($LB_{0.75x}$) and Figure 5B ($LB_{0.75x}$), between S30 ($LB_{0.5x}$) and 5B ($LB_{0.5x}$) and, between S30 ($LB_{0.25x}$) and 5B ($LB_{0.25x}$), respectively). In detail, to obtain each of the 46 data points in Supplementary Figure S30, we randomly assembled 100000 cohorts of 10 genes each with specific $\mu_{K_{TF}}$ (and, thus, $\mu_{|b|}$). This was done 100 times, from which we obtained average values and error bars for each of the 46 data points. The same was done for 5B, except that the cohorts, as noted, were assembled directly based on their $\mu_{|b|}$.

The pseudo-algorithm (Algorithm 1) created in MATLAB to generate, using the ensemble approach, the empirical data points in Supplementary Figure S30 was:

1. For each RNAP shift, let **LFC_data** be the **|LFC|** (absolute of log2 of fold change) of each gene and **KTF_data** be the corresponding number of input TFs, i.e., $\mathbf{K_{TF}}$, from 0 to 5.
2. Get the corresponding **S_data**, i.e., $S = |b|$, the absolute of the sum of the regulatory effects of the inputs of each gene.
3. Get the fraction of genes, $\mathbf{p_i}$, with a given $\mathbf{K_{TF}^i}$ ($\mathbf{K_{TF}}$ = i), with i=0:5.
4. Sample $\mathbf{b_i}$ from a Beta PDF (Probability Density Function), for each value of $\mathbf{p_i}$, using the MATLAB function 'BetaPDF' with parameter values: a = b = 0.75.
5. Let **r** (set to 10000) be the number of target genes, with a given $\mathbf{K_{TF}^i}$, to be randomly sampled with replacement.
6. For each $\mathbf{K_{TF}^i}$, define $\mathbf{n_i}$ = $\mathbf{b_i}$ x **r**, and create a set of $\mathbf{n_i}$ genes with $\mathbf{K_{TF}^i}$, randomly sampled with replacement from **KTF_data**. Let all these new sets of values be named **new_KTF_data**.
7. Let **TM** be numbers from 0.5 to 5, with an increment of 0.1, and a precision of 0. Let each **TM** value be a 'target mean of $\mathbf{K_{TF}}$'.
8. To find cohorts of genes with a given value of **TM**, set the number of iterations to 100. In each iteration, named **h,** do as follows:
    a. For each $\mathbf{K_{TF}^i}$**,** sample with replacement from **KTF_data** a vector **index_i** with $\mathbf{n_i}$ indexes of genes with $\mathbf{K_{TF}}$ **=** i.
    b. Generate **new_S_data** combining the empirical **S**, from **S_data**, of all genes under the indexes found in all sets of **index_i**. The data is added in ascending order of sorted $\mathbf{K_{TF}}$, in accordance with point 3.

c. Generate, for each RNAP shift, **new_LFC_data** combining the empirical **|LFC|**, from **LFC_data**, of all genes under the indexes found in all sets of **index_i**. The data is added by in ascending order of sorted $\mathbf{K_{TF}}$, in accordance with point 3.

d. Set **m**, the target number of genes of cohorts with a given value of **TM**, to 10.

e. Let the number of sampling iterations be 100000. In each iteration **j**, do as follows:

i. In **new_sampledKTF_data** save **m** values sampled with replacement from **new_KTF_data**. Save in '**I(:, j)'** the indexes of the **m** values sampled.

ii. In '**M_KTF (1, j)'** save the mean value of **new_sampledKTF_data**.

f. For each value **k** in **TM** located in position **t**:

i. Save in **idx_mean** the indexes in **M_KTF** of all mean values equal to **k** with a precision of 0.

ii. Get the indexes **I(:, idx_mean)** of the genes included in all sampling sets whose mean is **k**.

iii. For each RNAP shift, calculate the mean **M_LFC(t, h)** of **|LFC|** from the data in **new_LFC_data** of the genes in **I(:, idx_mean)**.

iv. Calculate the mean **M_S(t, h)** of **S** in **new_S_data** of the genes in **I(:, idx_mean)**.

9. For each RNAP shift, get the expected mean **|LFC| M_LFC_final(t)** and **SEM_LFC_final(t)** for each **k** value in position **t** in **TM**. **M_LFC_final(t) and SEM_LFC_final(t)** are calculated, respectively, as the mean and standard deviation of the **M_LFC(t, h)** values over the **h** runs.
10. Get the expected mean **S M_S_final(t)** and **SEM_S_final(t)** for each **k** in position **t** in **TM**. **M_S_final(t)** and **SEM_S_final(t)** are calculated, respectively, as the mean and standard deviation of the **M_S(t, h)** values over the **h** runs.

Meanwhile, for Figure 5B, the pseudo-algorithm (Algorithm 2) used was:

1. For each RNAP shift, let **LFC_data** be the **|LFC|** (absolute of log2 of fold change) of each gene and **KTF_data** be the corresponding number of input TFs, i.e., $\mathbf{K_{TF}}$, from 0 to 5.
2. Get the corresponding **S_data**, i.e., $S = |b|$, the absolute of the sum of the regulatory effects of the inputs of each gene.
3. Get the fraction of genes, $\mathbf{p_i}$, with a given $\mathbf{S_i}$ (**S** = i), for all unique values found **S_data** in ascending order.
4. Sample $\mathbf{b_i}$ from a Beta PDF, for each value of $\mathbf{p_i}$, using the MATLAB function 'BetaPDF' with parameter values: a = b = 1.5.

5. Let **r** (set to 10000) be the number of target genes, with a given **S$_i$**, to be randomly sampled with replacement.
6. For each **S$_i$**, define **n$_i$** = **b$_i$** x **r**, and create a set of **n$_i$** genes with **S$_i$**, randomly sampled with replacement from **S_data**. Let all these new sets of values be named **new_S_data**.
7. Let **TM** be numbers from 0 to 3 with an increment of 0.25, and a precision of ± 0. Let each **TM** value be a 'target mean of **S**'.
8. To find cohorts of genes with a given value of **TM**, set the number of iterations to 100. In each iteration, named **h,** do as follows:
    a. For each value of **S$_i$,** sample with replacement from **S_data** a vector **index_i** with **n$_i$** indexes of genes with **S =** i.
    b. Generate, for each RNAP shift, **new_LFC_data** combining the empirical **|LFC|**, from **LFC_data**, of all genes under the indexes found in all sets of **index_i**. The data is added in ascending order of sorted **S**, in accordance with point 3.
    c. Set **m**, the target number of genes of cohorts with a given value of **TM**, to 10.
    d. Let the number of sampling iterations be 100000. In each iteration **j**, do as follows:
        i. In **new_sampledS_data** save **m** values sampled with replacement from **new_S_data**. Save in **I(:, j)** the indexes of the **m** values sampled.
        ii. In **M_S (1, j)** save the mean value of **new_sampledS_data**.
    e. For each value **k** in **TM** located in position **t**:
        i. Save in **idx_mean** the indexes in **M_S** of all mean values equal to **k** with a precision of ± 0.1.
        ii. Get the indexes **I(:, idx_mean)** of the genes including in all the sampling sets whose mean is **k**.
        iii. For each RNAP shift, calculate the mean **M_LFC(t, h)** of **|LFC**| from **new_LFC_data** of the genes found in **I(:, idx_mean)**.
9. For each RNAP shift, get the expected mean **|LFC| M_LFC_final(t)** and **SEM_LFC_final(t)** for each **k** value in position **t** in **TM**. **M_LFC_final(t) and SEM_LFC_final(t)** are calculated, respectively, as the mean and standard deviation of the **M_LFC(t, h)** values over the **h** runs.

We also created tailored cohorts with imposed values of $\mu_{K_{TF}}$ and $\mu_{|b|}$, to evaluate how K$_{TF}$ and $|b|$ independently affect |LFC|. For instance, for Figure 5E, we obtained cohorts with a given $\mu_{K_{TF}}$ and different $\mu_{|b|}$. For this, we used the following pseudo-algorithm (Algorithm 3, obtained by "inserting" algorithm 2 into algorithm 1):

1. For each RNAP shift, let **LFC_data** be the **|LFC|** (absolute of log2 of fold change) of each gene and let **KTF_data** be the corresponding number of input TFs, i.e., $\mathbf{K_{TF}}$, from 0 to 5.
2. Get the corresponding **S_data**, i.e., $S = |b|$, the absolute of the sum of the regulatory effects of the inputs of each gene.
3. Get the fraction of genes, $\mathbf{p_i}$, with a given $\mathbf{K_{TF}^i}$ ($\mathbf{K_{TF}}$ = i), with i=0:5.
4. Sample $\mathbf{b_i}$ from a Beta probability density function, for each value of $\mathbf{p_i}$, using the MATLAB function 'BetaPDF' with parameter values: a = b = 0.75.
5. Let **r** (set to 10000) be the number of target genes, with a given $\mathbf{K_{TF}^i}$, to be randomly sampled, with replacement.
6. For each $\mathbf{K_{TF}^i}$, define $\mathbf{n_i}$ = $\mathbf{b_i}$ x **r**, and create a set of $\mathbf{n_i}$ genes with $\mathbf{K_{TF}^i}$, randomly sampled with replacement from **KTF_data**. Let these new sets of values be named **new_KTF_data**.
7. Let **TM** be numbers from 0.5 to 5, with an increment of 0.1, and a precision of 0. Let each **TM** value be a 'target mean of $\mathbf{K_{TF}}$'.
8. To find and analyze cohorts of genes with a given value of **TM**, set the number of iterations to 100. In each iteration, named **h,** do as follows:
   a. For each value $\mathbf{K_{TF}^i}$ **,** sample with replacement from **KTF_data** a vector **index_i** with $\mathbf{n_i}$ indexes of genes with $\mathbf{K_{TF}}$ **=** i.
   b. Generate **new_S_data{1,h}** combining the empirical **S**, from **S_data**, of all genes under the indexes found in all sets of **index_i**. The data is added in ascending order of sorted $\mathbf{K_{TF}}$, in accordance with point 3.
   c. Generate, for each RNAP shift, **new_LFC_data{1, h}** combining the empirical **|LFC|**, from **LFC_data**, of all genes under the indexes found in all sets of **index_i**. The data is added by in ascending order of sorted $\mathbf{K_{TF}}$, in accordance with point 3.
   d. Set **m**, the target number of genes of cohorts with a given value of **TM**, to 10.
   e. Let the number of sampling iterations be 100000. In each iteration **j**, do as follows:
      i. In **new_sampledKTF_data** save **m** values sampled with replacement from **new_KTF_data**. Save in **I(:, j)** the indexes of the **m** values sampled.
      ii. In **M_KTF (1, j)** save the mean value of **new_sampledKTF_data**.
   f. For each value **k** in **TM** located in position **t**:
      i. Save in **idx_mean** the indexes in **M_KTF** of all mean values equal to **k** with a precision of 0.
      ii. In **genes_sampled_KTF_k{1, h}** save the indexes **I(:, idx_mean)** of the genes included in all sampling sets whose mean is **k.**

iii. Save in **M_KTF_all_data{t, h}** all values from **new_KTF_data(genes_sampled_KTF_k{1, h})**, i.e., all the **K_{TF}** values contained in **new_KTF_data** of the genes under the indexes saved in **genes_sampled_KTF_k{1, h}**.

iv. In **M_S_all_data{t, h}** save all the **S** values contained in **new_S_data{1, h}(genes_sampled_KTF_k{1, h})**.

v. For each RNAP shift, save in **M_LFC_all_data{t, h}** all the **|LFC|** values from the respective **new_LFC_data{1, h}(genes_sampled_KTF_k{1, h})**.

vi. Let **Unique_S** be the unique values of **S** found in **M_S_all_data{t, h}**.

vii. In **S_data_S** save all the values in **M_S_all_data{t, h}** in ascending order of sorted **S**.

viii. For each RNAP shift and each value **Unique_S**^{i} (**Unique_S**=i), obtain **M_LFC_all_data{t,h}(M_S_all_data{t, h}=Unique_S**^{i}**)**. Let all these sets of **|LFC|** be named **LFC_data_S** for each shift.

ix. Get the corresponding **KTF_data_S** for each value **Unique_S**^{i} from **M_KTF_all_data{t, h}(M_S_all_data{t, h}=Unique_S**^{i}**)**.

x. Get the fraction of genes, **p**$_{i}$**_S**, with a given **Unique_S**^{i}, for all unique values found **S_data_S** in ascending order.

xi. Sample **b**$_{i}$**_S** from a Beta PDF (MATLAB function 'BetaPDF' with a = b = 0.75), for each value of **p**$_{i}$**_S**.

xii. Let **r_S** (set to 10000) be the number of target genes, with a given **Unique_S**^{i}, to be randomly sampled with replacement.

xiii. For each **Unique_S**^{i}, define **n**$_{i}$**_S** = **b**$_{i}$**_S** x **r_S**, and create a set of **n**$_{i}$**_S** genes with **Unique_S**^{i}, randomly sampled with replacement from **S_data_S**. Let all these new sets of values be named **new_S_data_S**.

xiv. Let **TM_S** be numbers from 1 to the maximum value of **Unique_S**, with a precision of ± 0. Let each **TM_S** value be a 'target mean of **S**' given the values in **Unique_S**.

xv. To find cohorts of genes with a given value of **TM_S**, set the number of iterations to 10. In each iteration, named **h_S,** do as follows:

1. For each value **Unique_S**^{i}, sample with replacement from **S_data_S** a vector **index_i_S** with **n**$_{i}$**_S** indexes of genes with **Unique_S**^{i}.
2. Generate, for each RNAP shift, **new_LFC_data_S** combining the empirical **|LFC|**, from **LFC_data_S**, of all genes under the indexes found in all sets of **index_i_S**. The data is added in

ascending order of sorted **Unique_S**, in accordance with point f.x.

3. Get the corresponding **new_KTF_data_S**, from **KTF_data_S**, of all genes under the indexes found in all sets of **index_i_S**. The data is added in ascending order of sorted **Unique_S**, in accordance with point f.x.
4. Set **m_S**, the target number of genes of cohorts with a given value of **TM_S**, to 10.
5. Let the number of sampling iterations be 100000 and, in each iteration **j_S**, do as follows:
   a. In **new_sampledS_data_S** save **m_S** values sampled with replacement from **new_S_data_S**. Save in **I_S(:, j_S)** the indexes of the **m_S** values sampled.

   b. In **M_UNIQUE_S(1, j_S)** save the mean value of **new_sampledS_data_S**.
6. For each value **k_S** in **TM_S** located in position **t_S**:
   a. Save in **idx_mean_S** the indexes in **M_UNIQUE_S** of all mean values equal to **k_S** with a precision of ± 0.1.

   b. Get the indexes **I_S(:, idx_mean_S)** of the genes including in all the sampling sets whose mean is **k_S**.

   c. In **M_S_S(t_S, h_S)** save the mean of **S** from **new_S_data_S** of the genes found in **I_S(:, idx_mean_S)**.

   d. For each RNAP shift, calculate the mean **M_LFC_S(t_S, h_S)** of **|LFC**| from **new_LFC_data_S** of the genes found in **I_S(:, idx_mean_S)**.

   e. Calculate the mean **M_KTF_S(t_S, h_S)** of $\mathbf{K}_{\mathbf{TF}}$ from **new_KTF_data_S** of the genes in **I_S(:, idx_mean_S)**.

xvi. For each RNAP shift, save under **M_LFC_final(t,t_S,h)** the expected mean and SEM of **|LFC|** for each **k_S** value in position **t_S** in **TM_S**. **M_LFC_final(t,t_S,h)** is calculated as the mean of the **M_LFC_S(t_S, h_S)** values over the **h_S** runs.

xvii. Save under **M_S_final(t,t_S,h)** the expected mean **S** for each **k_S** in position **t_S** in **TM_S**. **M_S_final(t,t_S,h)** is calculated as the mean of the **M_S_S(t_S, h_S)** values over the **h_S** runs.

xviii. Save under **M_KTF_final(t,t_S,h)** the expected mean **K_{TF}** for each **k_S** in position **t_S** in **TM_S**. **M_KTF_final(t,t_S,h)** iscalculated as the mean of the **M_KTF_S(t_S, h_S)** values over the **h_S** runs.

9. For each position **t_S** corresponding to a given **k_S** value, do as follows:
   a. For each RNAP shift, save under **M_LFC_ks** and **SEM_LFC_ks**, respectively, the expected mean and SEM of **|LFC|** for each **k_S** value in position **t_S**. **M_LFC_ks** and **SEM_LFC_ks** are calculated, respectively, as the mean and standard deviation of the **M_LFC_final(t,t_S,h)** values over the **h** runs.
   b. Save under **M_S_ks** and **SEM_S_ks**, respectively, the expected mean and SEM of **S** for each **k_S** in position **t_S**. **M_S_ks** and **SEM_S_ks** are calculated, respectively, as the mean and standard deviation of the **M_S_final(t,t_S,h)** values over the **h** runs.
   c. Save under **M_KTF_ks** and **SEM_KTF_ks**, respectively, the expected mean and SEM of **K_{TF}** for each **k_S** in position **t_S**. **M_KTF_ks** and **SEM_KTF_ks** are calculated, respectively, as the mean and standard deviation of the **M_KTF_final(t,t_S,h)** values over the **h** runs.

Finally, from the above, one can produce the results in Figure 5B. In detail, we assembled gene cohorts (Supplementary Results section *Estimation of the expected $\mu_{K_{TF}}$ and $\mu_{|b|}$ using an ensemble approach*) with a given $\mu_{|b|}$ (up to 24500 cohorts of 10 genes per $\mu_{|b|}$ value). In Figure 5B, each blue cross is the $\mu_{|LFC|}$ of a set of cohorts. Visibly, $\mu_{|LFC|}$ increases with $\mu_{|b|}$.

SUPPLEMENTARY FIGURES

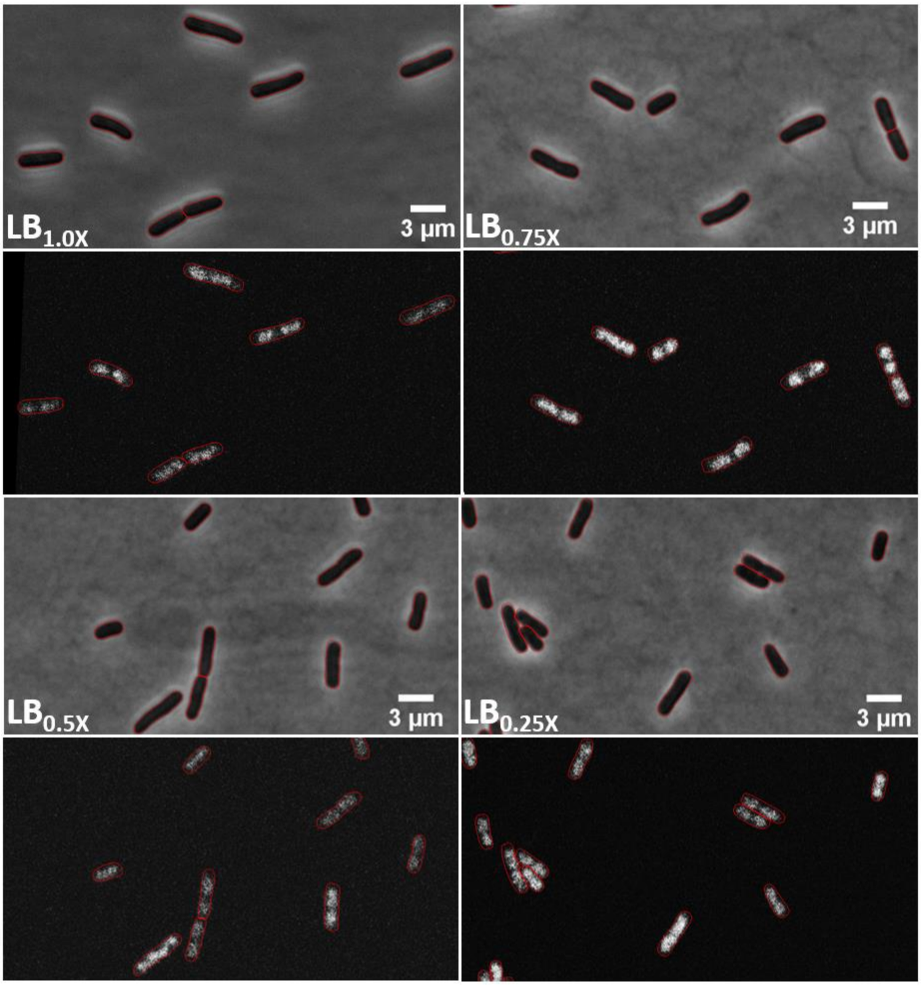


Figure S1. Example phase-contrast images (first and third rows) and corresponding confocal images of *E. coli* cells (second and forth rows, respectively). RL1314 cells grown in various media (Methods sections *Bacterial strains, media, growth conditions and curves* and *Microscopy*). Cells were segmented in phase-contrast images using CellAging (6).

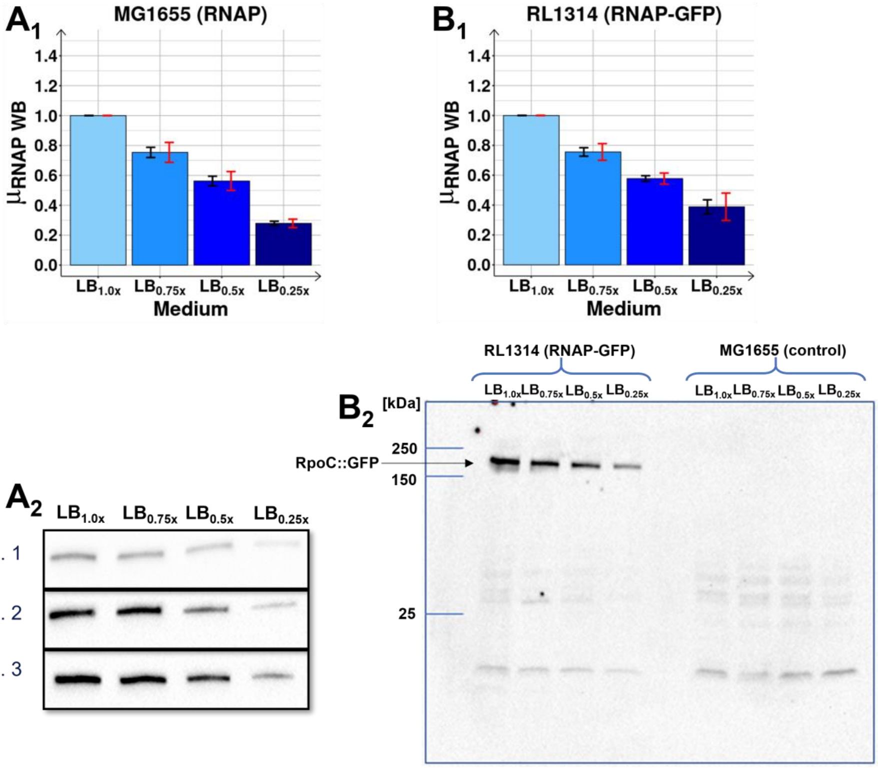


Figure S2. (Related to Figure 2E) RNAP and RNAP::GFP protein levels by Western Blot in LB$_{1.0x}$, LB$_{0.75x}$, LB$_{0.5x}$ and, LB$_{0.25x}$ media.

(A$_1$) RNAP abundance measured by Western Blot in LB$_{1.0x}$, LB$_{0.75x}$, LB$_{0.5x}$ and LB$_{0.25x}$ media using antibodies against β' subunit (Methods section *Protein isolation and western blotting*). Absolute values of 3 biological replicates (Supplementary Table S2) obtained from blot images by 'Image Lab' software (version 5.2.1), from which we calculated relative values to LB$_{1.0x}$. Finally, we obtained the mean, SEM (black error bars), and 95% confidence bounds (CB) of the SEM (red error bars) of the replicates.

(A$_2$) Blot images of 3 biological replicates (Rep) from which the RNAP abundances in A$_1$ were obtained. All bands are near 155 kDa, known to correspond to the β' subunit (7)(8).

(B$_1$) Same as (A$_1$), but using antibodies against GFP alone, which is tagged to RNAP (expressed by cells of the RL1314 strain). Absolute values in Supplementary Table S2.

(B$_2$) Blot image of RpoC::GFP measured in the RL1314 and MG1655 (control) strains. In RL1314 cells, the strongest bands are between 150 and 250 kDa, as expected, given the fusion of GFP (27 kDa (9)) with the β' unit (155 kDa). No clear bands were observed in that region in the control strain. Meanwhile, there is a weak band just above 25 kDa in RL1314 cells. This band might correspond to GFP that has been cleaved off from the chimeric protein, but it is so minor that one can conclude that its contribution to the total cell fluorescence is negligible.

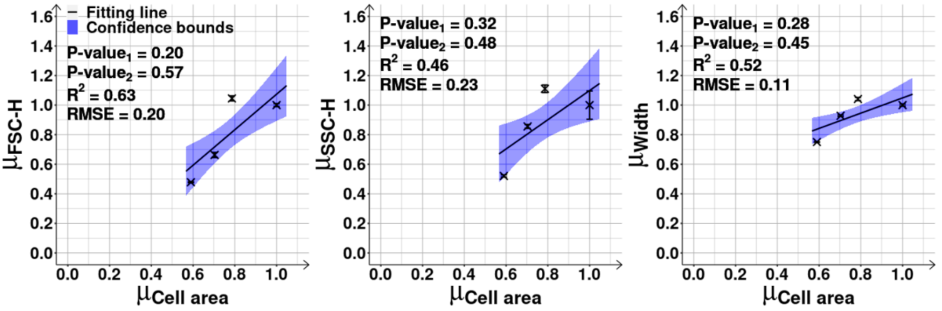


Figure S3. Related to Figures 2F and 2G) Flow-cytometry and microscopy data on cell size and composition. Correlation plot between cell areas ($\mu_{cell\ area}$, Methods section *Microscopy*) and each of the three flow-cytometry parameters positively correlated to cell size and composition (μ_{FSC-H} , μ_{SSC-H} , μ_{Width} from the FCS-H, SSC-H and Width parameters, respectively, Methods section *Flow-cytometry*), in each condition, 180 min after shifting the medium. All data points were obtained from 3 independent biological replicates and are shown relative to $LB_{1.0x}$ (control). Mean cell areas were obtained from phase-contrast images of MG1655 cells grown in the same conditions (Methods sections *Bacterial strains, media, growth conditions and curves* and *Microscopy*). The best fitting lines (solid black), their 68% confidence bounds (blue shadow areas) and statistics (coefficient of determination (R^2), root mean square error (RMSE), and P-values at 0.1 significance level) were obtained as described in Methods section *Statistical tests c*.

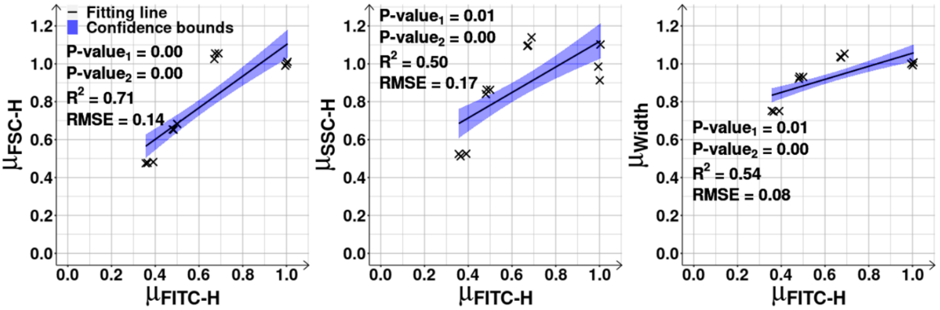


Figure S4. (Related to Figures 2G and 2D) Flow-cytometry data relating cell size and composition with the expression levels of the RpoC sub-unit of RNAP. Correlation plots between FITC-H (maximum peak 'Height', -H, of the FITC signal), which is linearly correlated to RNAP-GFP levels (μ_{FITC-H}), and each of the three flow-cytometry parameters positively correlated to cell size and composition (μ_{FSC-H}, μ_{SSC-H}, μ_{Width} from the FCS-H, SSC-H and Width parameters), in LB$_{1.0x}$, LB$_{0.75x}$, LB$_{0.5x}$ and LB$_{0.25x}$, at 180 min. All data is relative to LB$_{1.0x}$ (control). Mean background fluorescence levels were removed from the FITC-H signals (Methods section *Flow-cytometry*). The best fitting lines (solid black) along with their 68% confidence bounds (blue shadow areas) and statistics (coefficient of determination (R^2), root mean square error (RMSE), and P-values at 0.1 significance level) were obtained as described in Methods section *Statistical tests c*.

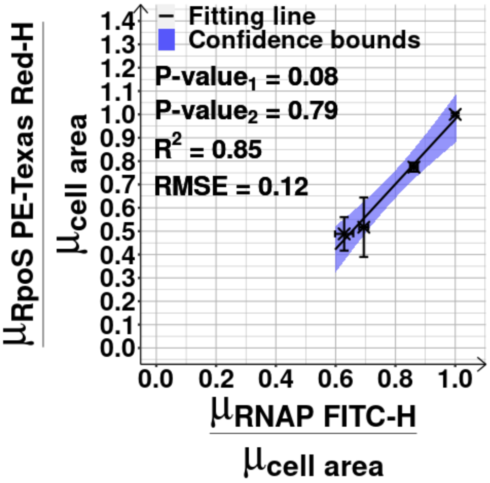


Figure S5. (Related to Figures 2I and 2D) Mean concentration of RpoS plotted against the mean concentration of RNAP. The mean concentration of of RpoS (σ^{38}) tagged with mCherry in LB$_{1.0x}$, LB$_{0.75x}$, LB$_{0.5x}$ and LB$_{0.25x}$ at 180 min, was measured by mean single-cell fluorescence by flow-cytometry (PE-Texas Red-H parameter), over mean cell area ($\mu_{cell\ area}$) (Methods section *Microscopy*). The mean RNAP concentration of RL1314 cells was obtained by mean single-cell fluorescence (FITC-H parameter, Methods section *Flow-cytometry*) over mean cell area ($\mu_{cell\ area}$). Mean background fluorescence levels were removed from both FITC-H and PE-Texas Red-H signals (Methods section *Flow-cytometry*). All data is relative to LB$_{1.0x}$ (control). The best fitting line (solid black) along with its 68% confidence bounds (blue shadow area) and statistics (coefficient of determination (R^2), root mean square error (RMSE), and P-values at 0.1 significance level) were obtained as described in Methods section *Statistical tests c*.

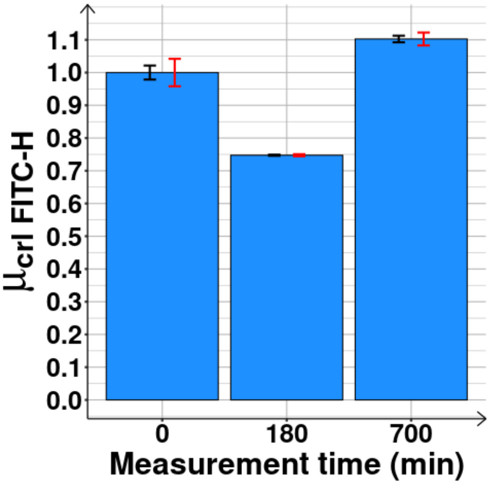


Figure S6. (Related to Figures 2I) crl gene expression levels in LB$_{0.5x}$. Mean protein expression levels ($\mu_{crl\ FITC-H}$) of the crl gene, first, at 0 and then at 180 (mid-log phase) and 700 (stationary growth phase) min in the LB$_{0.5x}$ medium, following dilution. The mean values are shown relative to 0 min.

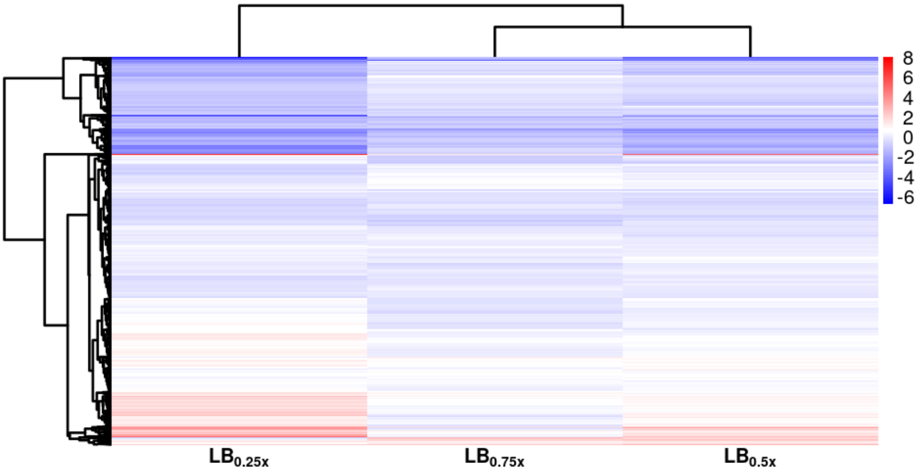


Figure S7. (Related to Figure 3B) Heatmap of shifts in RNA levels. Log2 of fold changes (LFC) from RNA-seq following medium dilutions from $LB_{1.0x}$ to $LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$. The compressed data includes all genes. The perturbation inflicted by $LB_{0.25x}$ was significantly higher than by $LB_{0.75x}$ and $LB_{0.5x,}$ regarding the number of genes perturbed and their response strengths.

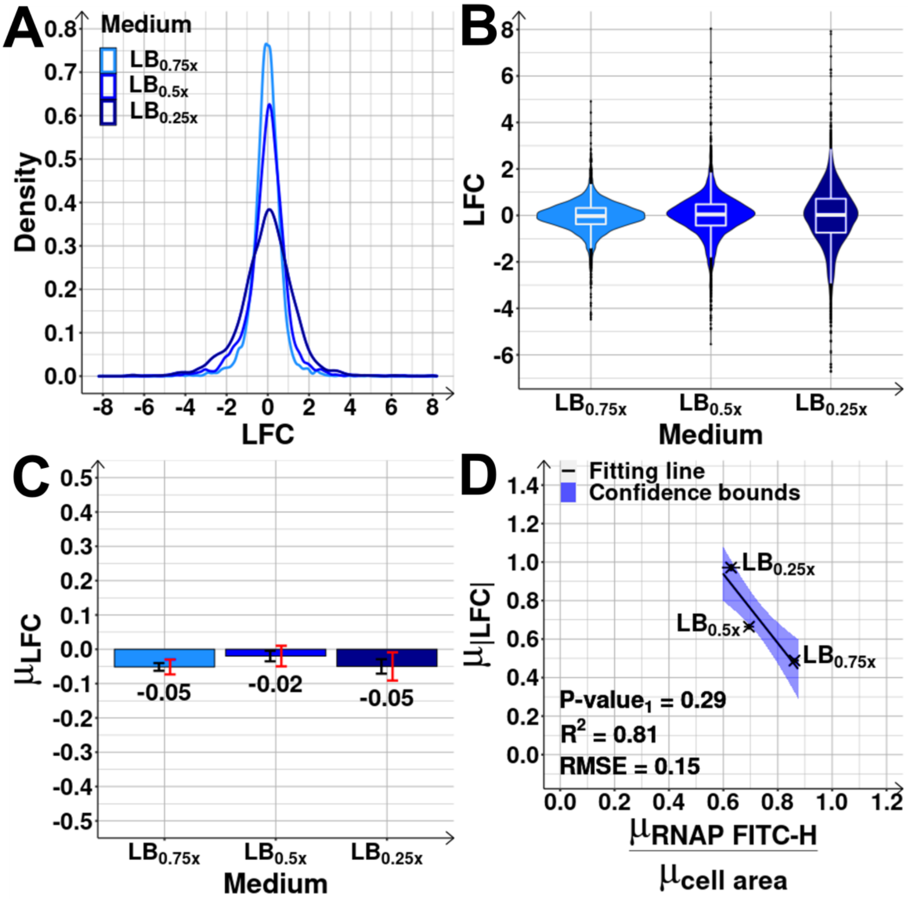


Figure S8. (Related to Figure 3B) Shifts in RNA levels.

(A) Kernel density estimates (Probability Distribution Function, PDF) of the distribution of genes with a given LFC (log2 of fold change) of RNA abundances, following the shifts from $LB_{1.0x}$ to $LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$, respectively. Non-parametric estimates of the PDF were obtained from data on all genes.

(B) The violin plot shows the maximum, minimum, median, interquartile ranges and probability density of the distributions in (A).

(C) Mean shifts in RNA levels, μ_{LFC} , following each shift. Black error bars represent the standard error of the mean (SEM), while red error bars represent the 95% confidence bounds of the SEM. From these, the mean cannot be distinguished between the conditions. We also performed 2-

sample T-tests of statistical significance between the distributions. Results in Supplementary Table S3 show that the $LB_{0.75x}$ and $LB_{0.5x}$ cannot be distinguished in a statistical sense.

(D) Correlation plot of $\mu_{|LFC|}$ and the respective RNAP concentration shift. All values are relative to the control condition. The best fitting line (solid black) along with its 68% confidence bounds (blue shadow area) and statistics (coefficient of determination (R^2), root mean square error (RMSE), and P-values at 0.1 significance level) were obtained as described in Methods section *Statistical tests c*.

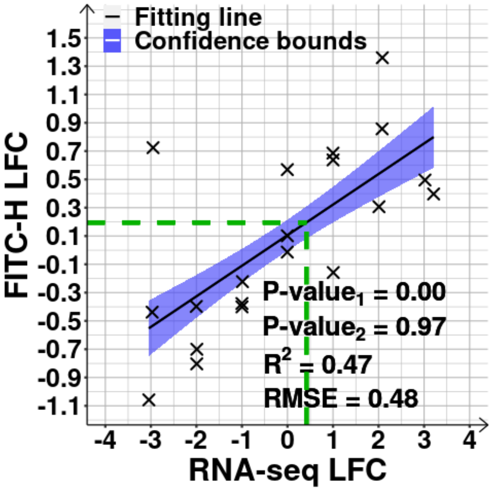


Figure S9. (Related to Figure 3B) Log$_2$ fold changes (LFCs) in RNA abundances of 20 randomly selected genes covering the whole spectrum of fold changes observed, plotted against their corresponding LFCs in protein abundances. Data from the shift from LB$_{1.0x}$ to LB$_{0.25x}$. We selected trios of genes whose LFC by RNA-seq was closest to -3, -2, -1, 0, +1, +2, and +3, respectively, to represent the entire spectrum (Methods section *RNA-seq e*). Events considered to be outliers by Tukey's fences (10) were removed. The LFCs in protein abundances were obtained by flow-cytometry (YFP strain library). Mean background fluorescence was removed (Methods section *Flow-cytometry*). The dashed green lines at positions (x, y) = (0.42, 0.20) indicate the average |LFC| of genes whose False Discovery Rate > 0.05, along with the corresponding estimated |LFC| by flow-cytometry (based on the fitting line). The best fitting line (solid black) along with its 68% confidence bounds (blue shadow area) and statistics (coefficient of determination (R^2), root mean square error (RMSE), and P-values at 0.1 significance level) were obtained as described in Methods section *Statistical tests c*. Supplementary Table S4 shows the list of strains measured by flow-cytometry.

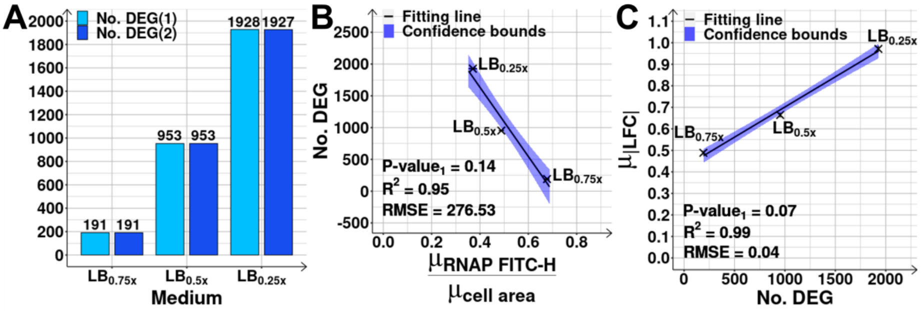


Figure S10. (Related to Figures 3B and 3C) Number of differentially expressed genes and its correlation with the shift in RNAP concentration.

(A) DEG(1) stands for Differentially Expressed Genes (4029 evaluated) (i.e., genes with False Discovery Rate (FDR) < 0.05), after diluting the medium to $LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$ (Methods section *RNA-seq c*). DEG(2) stands for genes whose FDR < 0.05 and absolute log2 of fold change (|LFC|) > 0.4248 ($LB_{0.75x}$), > 0.4085 ($LB_{0.5x}$) or > 0.4138 ($LB_{0.25x}$) (Methods section *RNA-seq d*). As the results using DEG(1) and DEG(2) slightly differ, from here onwards, DEG are selected using the more stringent criteria DEG(2).

(B) Number of DEG for each shift plotted against the respective RNAP concentration shift, estimated from the ratio between mean RNAP levels measured by FITC-H, $\mu_{RNAP\ FITC-H}$ using RL1314 cells (Methods section *Flow-cytometry*), and the mean cell area ($\mu_{cell\ area}$) estimated from phase-contrast images of MG1655 cells (Methods section *Microscopy*). All values are relative to the control condition ($LB_{1.0x}$).

(C) Correlation plot of the mean of |LFC| (absolute log2 fold change), $\mu_{|LFC|}$, and the respective number of DEG in each shift. The best fitting lines (solid black) along with their 68% confidence bounds (blue shadow area) and statistics (coefficient of determination (R^2), root mean square error (RMSE), and P-values at 0.1 significance level) were obtained as described in Methods section *Statistical tests c*.

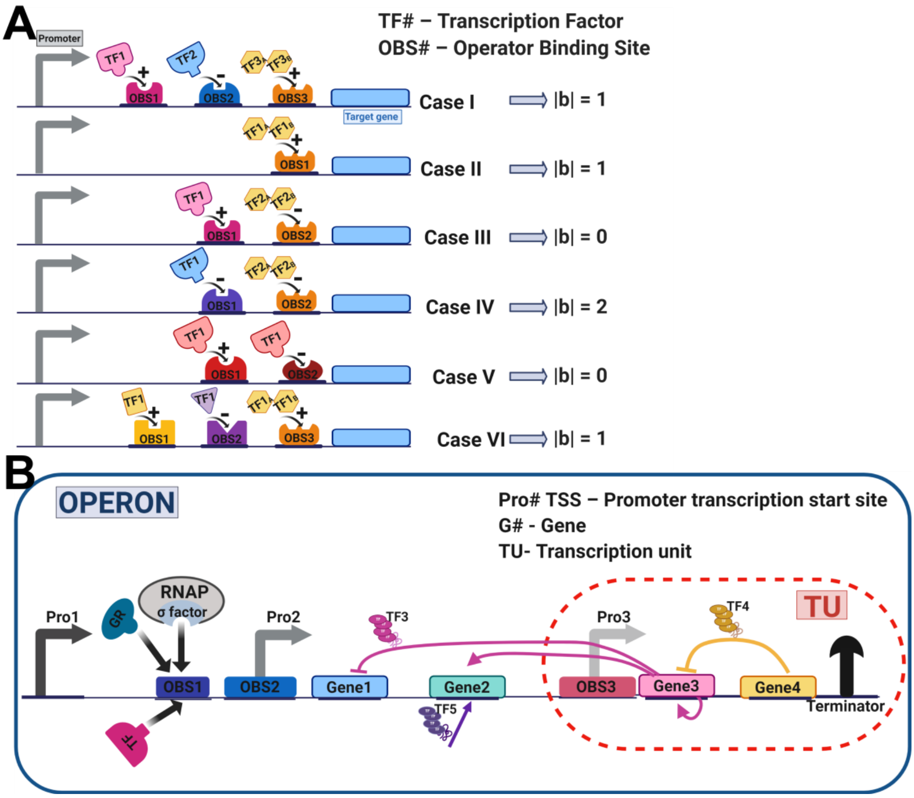


Figure S11. An example operon and examples of interactions of input transcription factors (TFs) with a total bias of 0 or 1.

(A) Illustrations of 6 forms of transcription regulation by sets of input TFs of an output protein. The input TFs regulation can be to activate (+1), repress (-1). The example also includes input TFs that are protein dimers. The right side of the figure shows the expected bias, i.e., overall regulatory effect (‘r’) of the set input TFs ($|b|$), set to equal the sum of the individual regulatory effect of each input TF, ‘r’ (shown by the black sign ‘+’ or ‘-‘, above each input TF on the left side). E.g., in case I, the pink OBS has a +1 effect, the blue has a -1 effect, and the orange has a +1 effect. Thus, $|b|$ = |+1-1+1| = +1.

(B) Illustration of an operon expressing 4 genes (G1, G2, G3 and G4) under the control of 2 promoters (Pro1 and Pro2). The colored arrows show the different regulations inside this operon. Also illustrated is a transcription unit (TU) inside the operon, controlled by one promoter, Pro3. The vertical black arrow at the end is a terminator while TF5 is produced by a gene not in the operon. This example assumes the definitions of operon and TU in RegulonDB. Figures (A) and (B) created with BioRender.com.

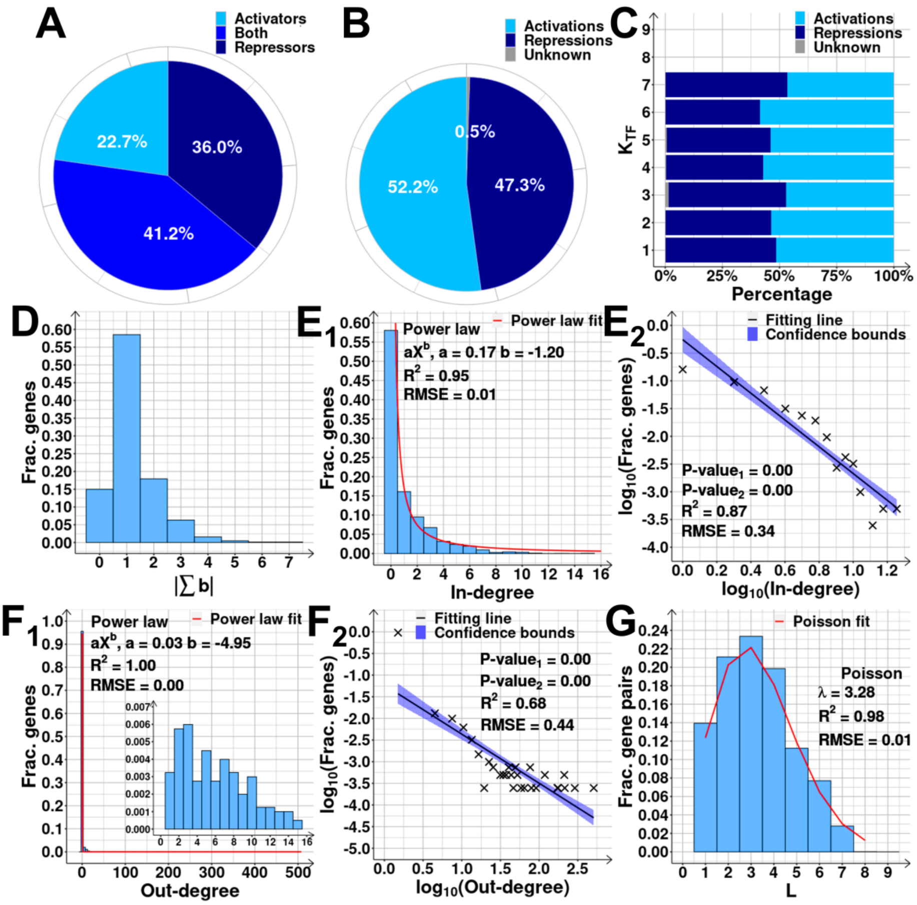


Figure S12. Topology and logic of the transcription factor (TF) network, TFN, of *E. coli*.

(A) Pie chart of the percentage of input TFs classified as ‘Activators’ (activate all genes they regulate) and ‘Repressors’ (repress all genes they regulate). ‘Both’ are the rare input TFs with dual effects (i.e., activate some genes and repress other, or have opposite strengths on the same promoter, depending on the conditions).

(B) Pie chart of the percentage of input TFs binding sites (interactions) classified as activations, repressions, or unknown.

(C) Same as (B), but for each cohort of genes defined by the genes’ K_{TF} (number of input TFs).

(D) Distribution of the fraction of genes with a given absolute sum of the regulatory effects, $|b|$, of the input TFs. Each input TF binding site is set to have a regulatory effect, ‘r’, equal to +1 (activation), -1 (repression), or 0 (unknown effect).

(E_1) In-degree (number of incoming edges) distribution.

(E_2) Scatter plot between the $\log_{10}$ of the In-degree and the $\log_{10}$ of the fraction of genes with corresponding In-degree.

(F_1) Out-degree (number of outgoing edges) distributions.

(F_2) Scatter plot between the $\log_{10}$ of the Out-degree and the $\log_{10}$ of the fraction of genes with corresponding out-degree.

(G) Path-length (L) distribution. The path length is the number of edges/input TFs needed to reach one node from another. A Poisson fitting (solid red) is shown with its fitting parameters, coefficient of variation (R^2) and root mean square error (RMSE).

For (E_1) and (F_1), we show power-law fittings (solid red lines) and their fitting parameters, R^2 and RMSE. In (E_2) and (F_2), the best fitting line (solid black), its 68% confidence bounds (blue shadow area), statistics (R^2, RMSE, and P-values at 0.1 significance level) were obtained as described in Methods section *Statistical tests c*.

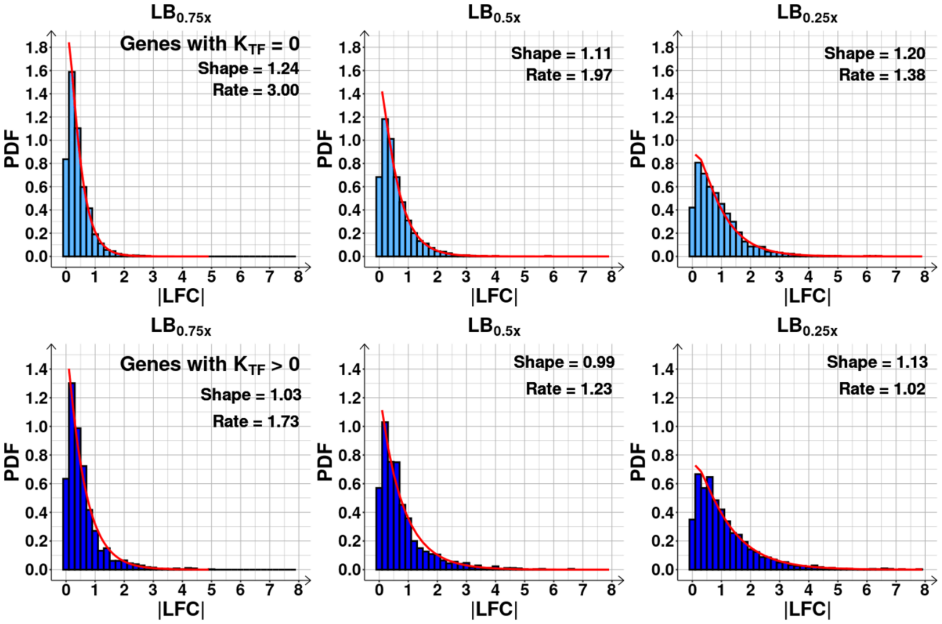
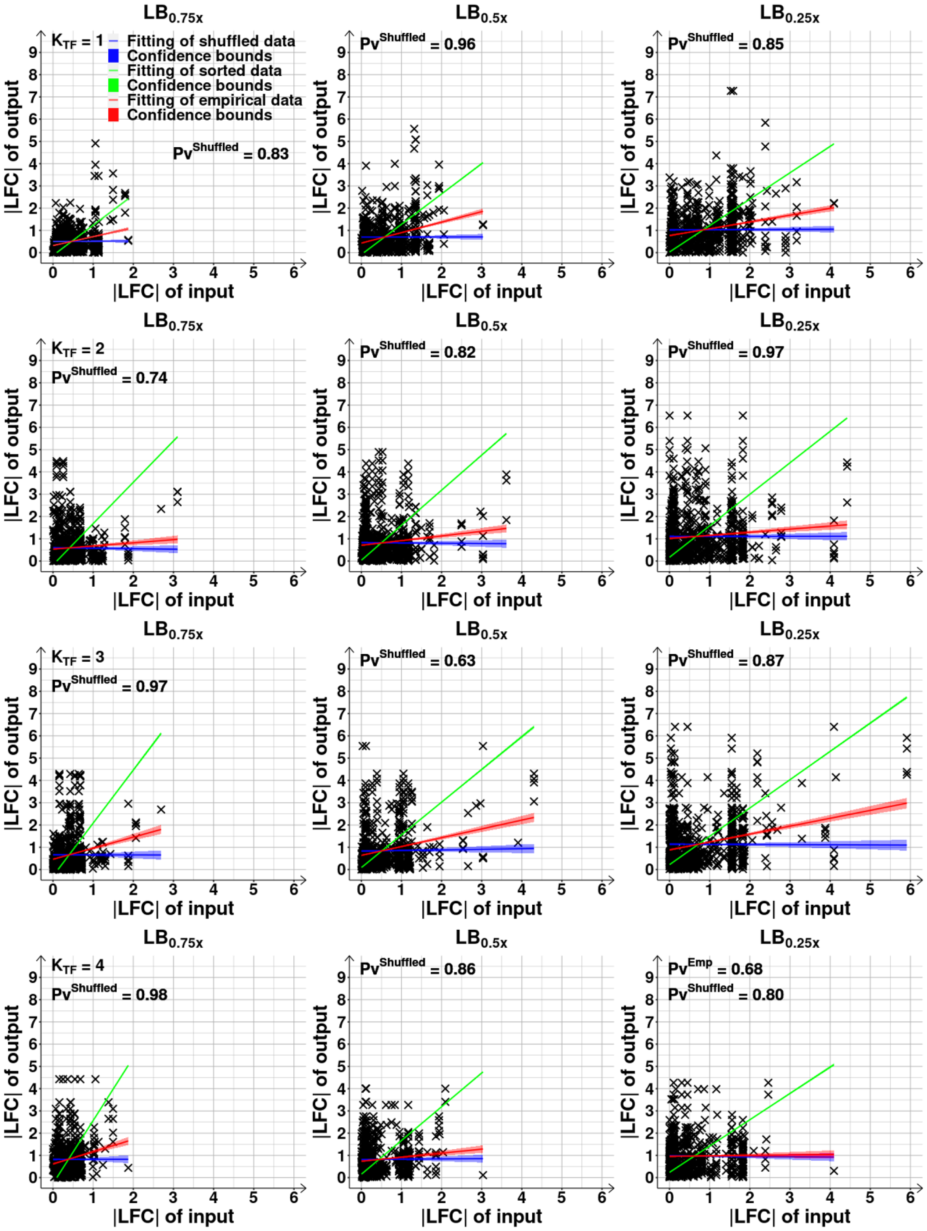


Figure S13. (Related to Figures 4A) Probability Density Function (PDF) of the number of genes with a given |LFC| (absolute log2 fold change) for each shift. PDFs obtained from genes without (light blue), and with input transcription factors (TFs) (medium blue). K_{TF} stands for number of input TFs. For each distribution, we fitted a gamma (Γ) function using *GAMFIT* of MATLAB which tunes the 'Shape' and 'Rate'. We also performed 2-sample T-tests and 2-sample K-tests of statistical significance between the distributions with and without input TFs. For both tests, all p-values were < 0.1.


LB$_{0.75x}$
LB$_{0.5x}$
LB$_{0.25x}$
K_{TF} = 1
Fitting of shuffled data
Confidence bounds
Fitting of sorted data
Confidence bounds
Fitting of empirical data
Confidence bounds
$Pv^{Shuffled}$ = 0.83
$Pv^{Shuffled}$ = 0.96
$Pv^{Shuffled}$ = 0.85
K_{TF} = 2
$Pv^{Shuffled}$ = 0.74
$Pv^{Shuffled}$ = 0.82
$Pv^{Shuffled}$ = 0.97
K_{TF} = 3
$Pv^{Shuffled}$ = 0.97
$Pv^{Shuffled}$ = 0.63
$Pv^{Shuffled}$ = 0.87
K_{TF} = 4
$Pv^{Shuffled}$ = 0.98
$Pv^{Shuffled}$ = 0.86
Pv^{Emp} = 0.68
$Pv^{Shuffled}$ = 0.80
|LFC| of output
|LFC| of input
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6

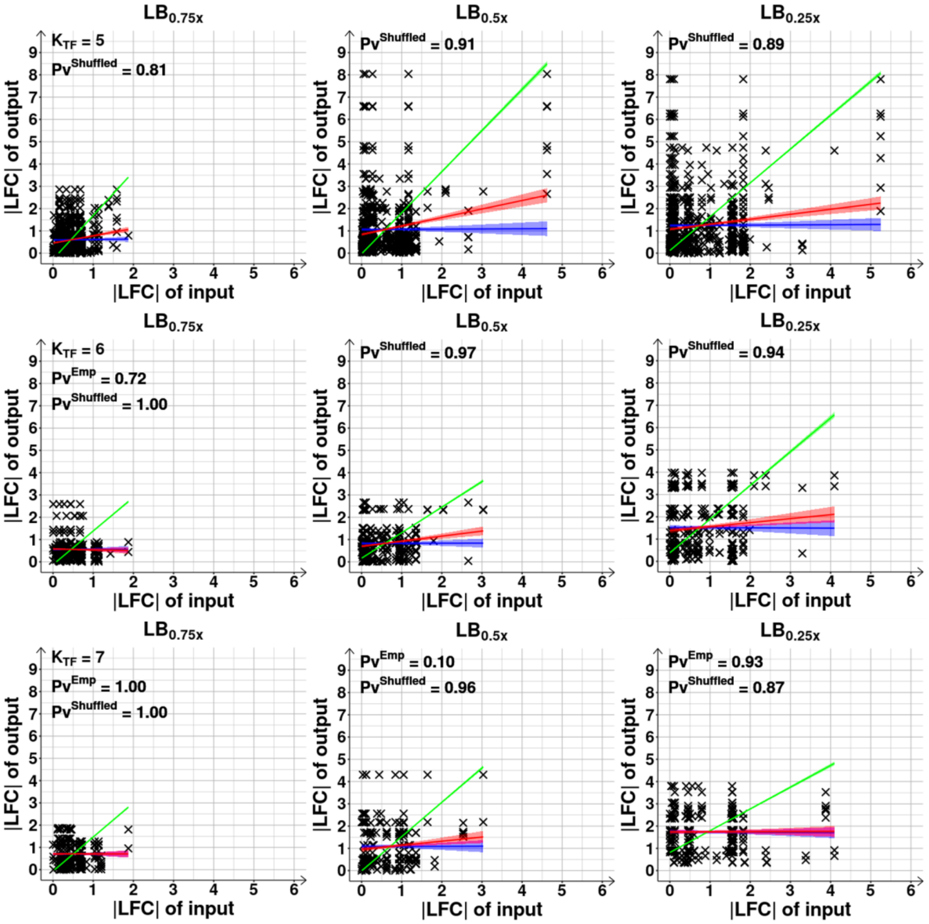
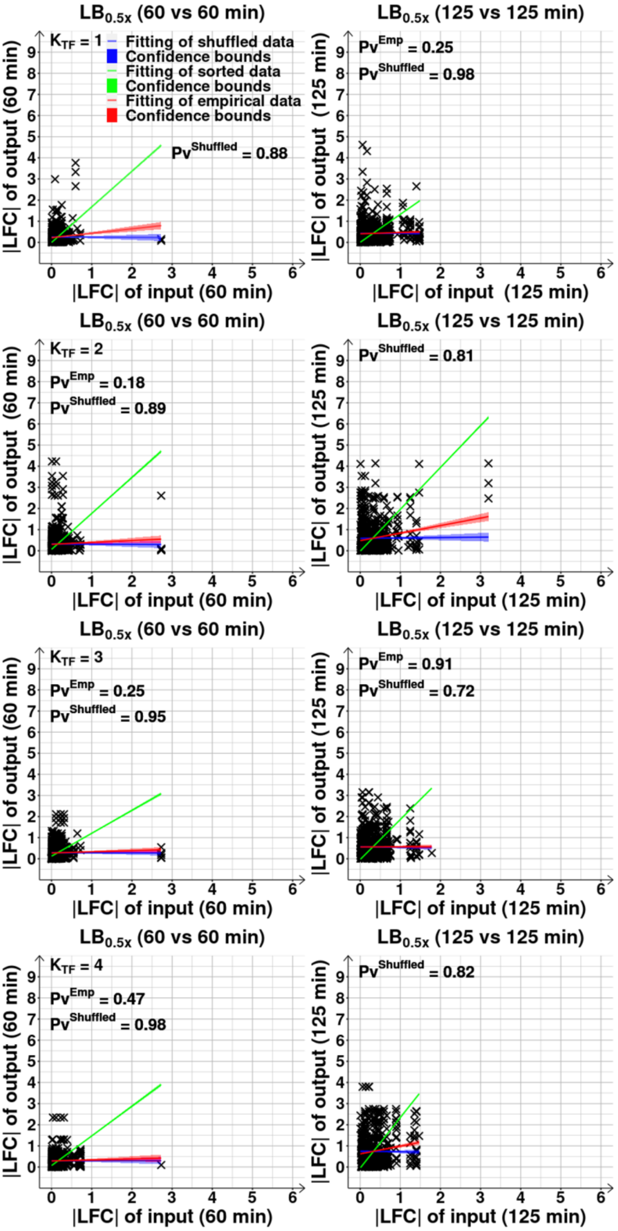


Figure S14. (Related to Figure 4B and Supplementary Figure S20B) Changes in RNA abundances of input genes plotted against those of the output genes, as a function of K_{TF}, the number of input transcription factors (TFs) of the output gene. Scatter plots of |LFC| (absolute log2 of fold change) of each output gene with the |LFC| of each gene expressing their direct input TFs (i.e., input genes), following the shifts from the control ($LB_{1.0x}$) to $LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$, for each class of genes defined by their K_{TF} (from 1 to 7). All genes are included, regardless of being differentially expressed. The red line is the best fitting one. The blue line is a null-model fitting line and was obtained as described in Methods section *Statistical tests c*. The green line is the best fitting one after sorting the values of the pairs input-output in ascending order, to estimate the maximum correlation possible between the two variables. Best fitting lines were obtained by linear least-squares regression fit using the MATLAB function *FITLM*. The colored shadow areas represent the 68% confidence bounds of the best fitted lines. We obtained the p-values of statistical significance for the red (Pv^{Emp}), blue ($Pv^{Shuffled}$) and green (Pv^{Sorted}) lines under the null hypothesis that the data is best fit by a horizontal line. We show the p-values when the null hypothesis was not rejected at 0.1 significance level. p-values and coefficients of determination (R^2) for the red ($R^2{}_{Emp}$) and green ($R^2{}_{Sorted}$) fitted lines are shown in Supplementary Table S7.


$LB_{0.5x}$ (60 vs 60 min)
$LB_{0.5x}$ (125 vs 125 min)
K_{TF} = 1
Fitting of shuffled data
Confidence bounds
Fitting of sorted data
Confidence bounds
Fitting of empirical data
Confidence bounds
$Pv^{Shuffled}$ = 0.88
Pv^{Emp} = 0.25
$Pv^{Shuffled}$ = 0.98
|LFC| of output (60 min)
|LFC| of input (60 min)
|LFC| of output (125 min)
|LFC| of input (125 min)
$LB_{0.5x}$ (60 vs 60 min)
$LB_{0.5x}$ (125 vs 125 min)
K_{TF} = 2
Pv^{Emp} = 0.18
$Pv^{Shuffled}$ = 0.89
$Pv^{Shuffled}$ = 0.81
$LB_{0.5x}$ (60 vs 60 min)
$LB_{0.5x}$ (125 vs 125 min)
K_{TF} = 3
Pv^{Emp} = 0.25
$Pv^{Shuffled}$ = 0.95
Pv^{Emp} = 0.91
$Pv^{Shuffled}$ = 0.72
$LB_{0.5x}$ (60 vs 60 min)
$LB_{0.5x}$ (125 vs 125 min)
K_{TF} = 4
Pv^{Emp} = 0.47
$Pv^{Shuffled}$ = 0.98
$Pv^{Shuffled}$ = 0.82
0 1 2 3 4 5 6
0 1 2 3 4 5 6 7 8 9

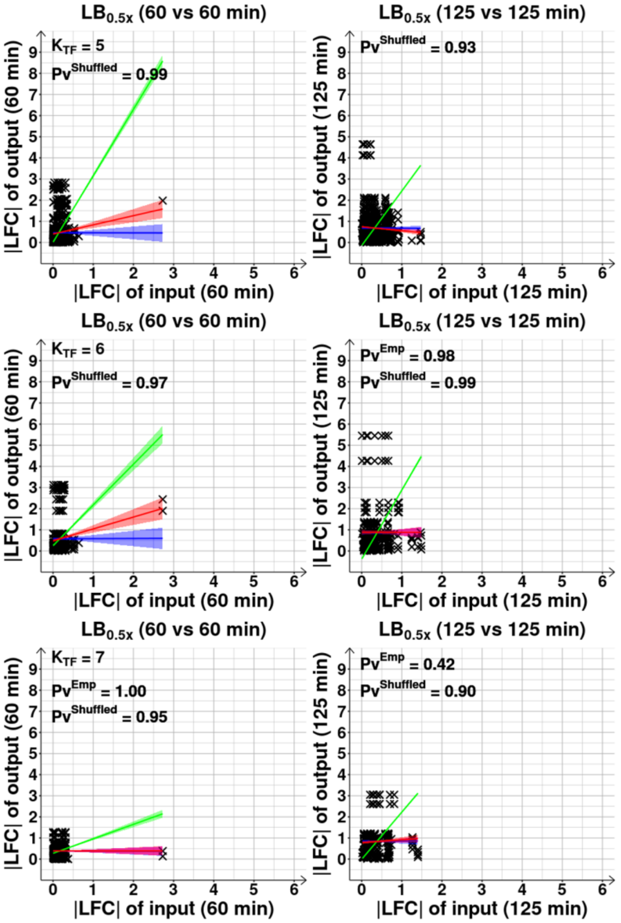


Figure S15. (Related to Figure 4B) Changes in RNA abundances (prior to the RNAP changes and during the short-term) of input genes plotted against their output genes, as a function of K_{TF}, the number of input transcription factors (TFs) of the output gene. Scatter plots of |LFC| (absolute log2 of fold change) of each output gene with the |LFC| of each gene expressing their direct input TFs (i.e., input genes), following the medium dilution to $LB_{0.5x}$, for each class of genes defined by their K_{TF} (from 1 to 7). The data regards the short-term responses (125 min) and the responses prior to the changes in RNAP concentration (60 min). All genes are included, regardless of being differentially expressed. The red line is the best fitting one. The blue line is the null-model fitting line and was obtained as described in Methods section *Statistical tests c*. The green line is the best fitting one after sorting the pairs input-output in ascending order, to estimate the maximum correlation possible between the two. Best fitting lines were obtained by linear least-squares regression fit using the MATLAB function *FITLM*. The colored shadow areas represent the 68% confidence bounds of the best fitted lines. We obtained the p-values of statistical significance for the red (Pv^{Emp}), blue ($Pv^{Shuffled}$) and green (Pv^{Sorted}) lines under the null hypothesis that the data is best fit by a horizontal line. We show the p-values when the null hypothesis was not rejected at 0.1 significance level. p-values and coefficients of determination (R^2) for the red (R^2_{Emp}) and green (R^2_{Sorted}) fitted lines are shown in Supplementary Table S8.

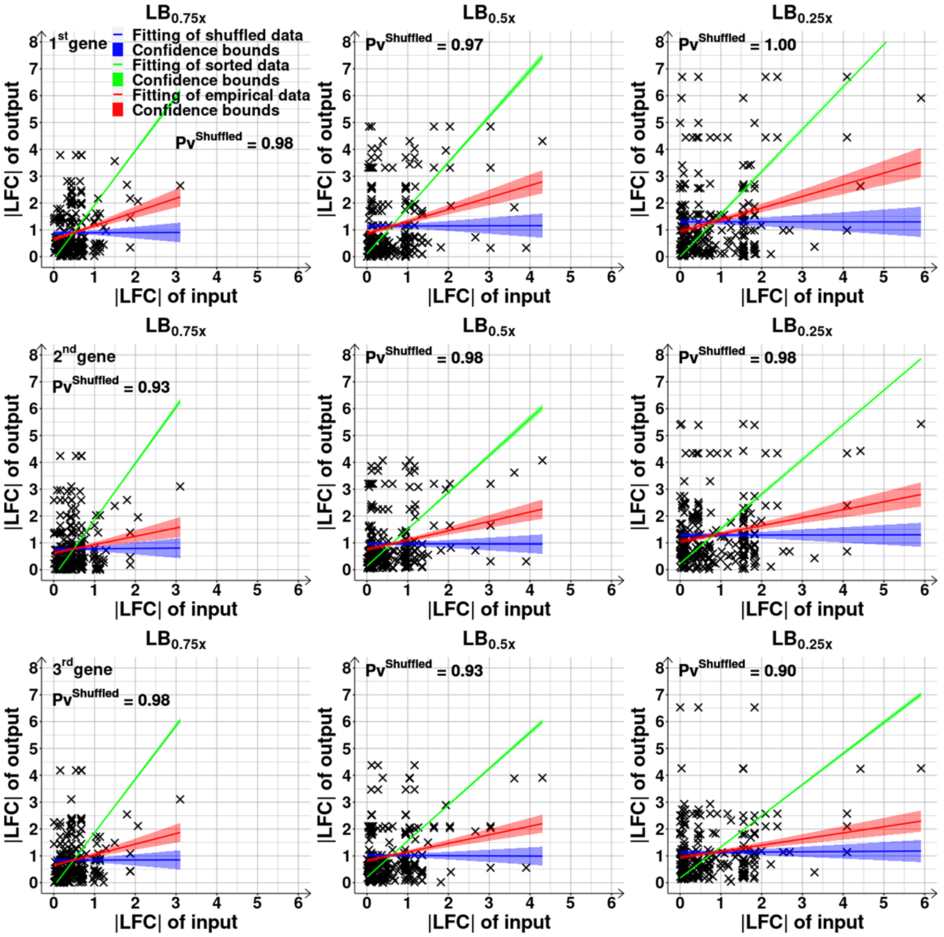


Figure S16. (Related to Figure 4B) Relationships between the changes in RNA abundances of input and output genes as a function of the position of the output gene in the operon. Scatter plots between the |LFC| (absolute log2 of fold change) of a gene expressing an input TF (i.e., input gene) and each of its direct output genes belonging to an operon of size 3 on the 1st, 2nd and 3rd positions in the operon following the transcription start site. The red line is the best fitting line between the two variables. The blue line is the null-model fitting line and was obtained as described in Methods section *Statistical tests c*. The green line is the best fitting line obtained after sorting the values of the pairs input-output in ascending order to obtain the maximum correlation possible. All best fitting lines were obtained by linear least-squares regression fit using the MATLAB function *FITLM*. The colored shadow areas are the 68% confidence bounds for the best fitted lines. We also obtained the p-values of statistical significance for the red (Pv^{Emp}), blue

($Pv^{Shuffled}$) and green (Pv^{Sorted}) lines under the null hypothesis that the data is best fit by a horizontal line. We only show the p-values when the null hypothesis was not rejected at 0.1 significance level. All p-values and the coefficients of determination (R^2) for the red (R^2_{Emp}) and green (R^2_{Sorted}) fitted lines are shown in Supplementary Table S9 (see also Supplementary Table S10).

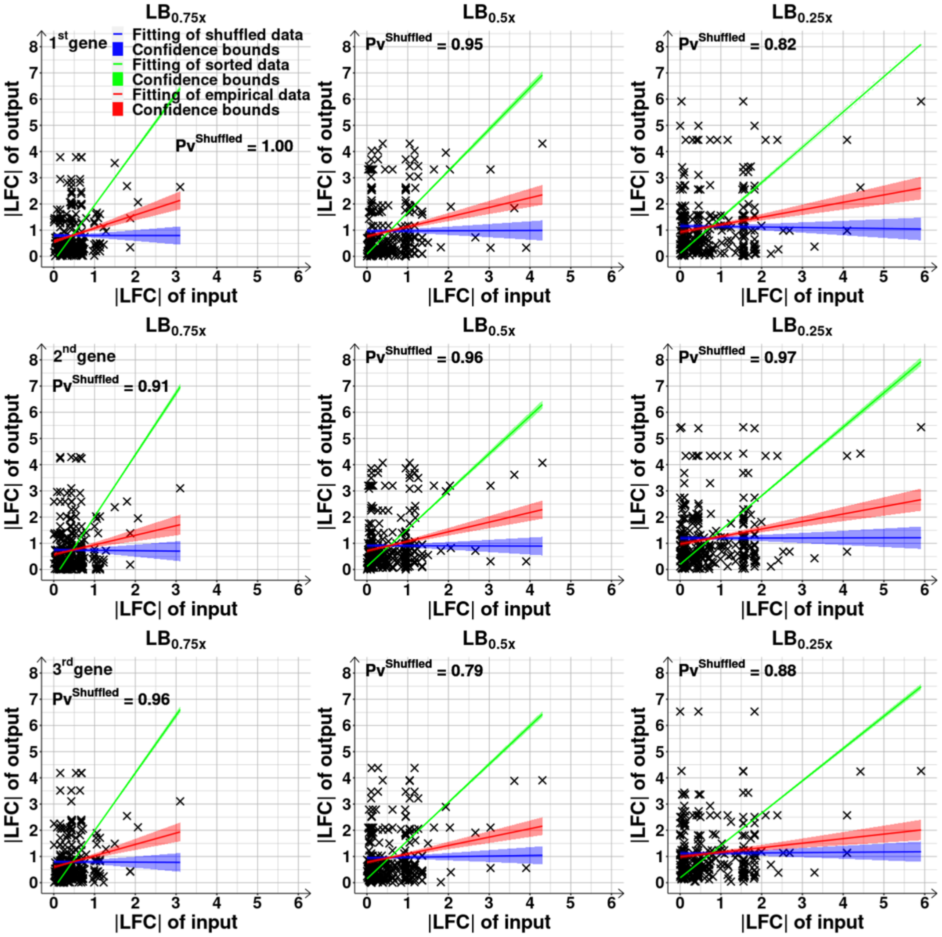
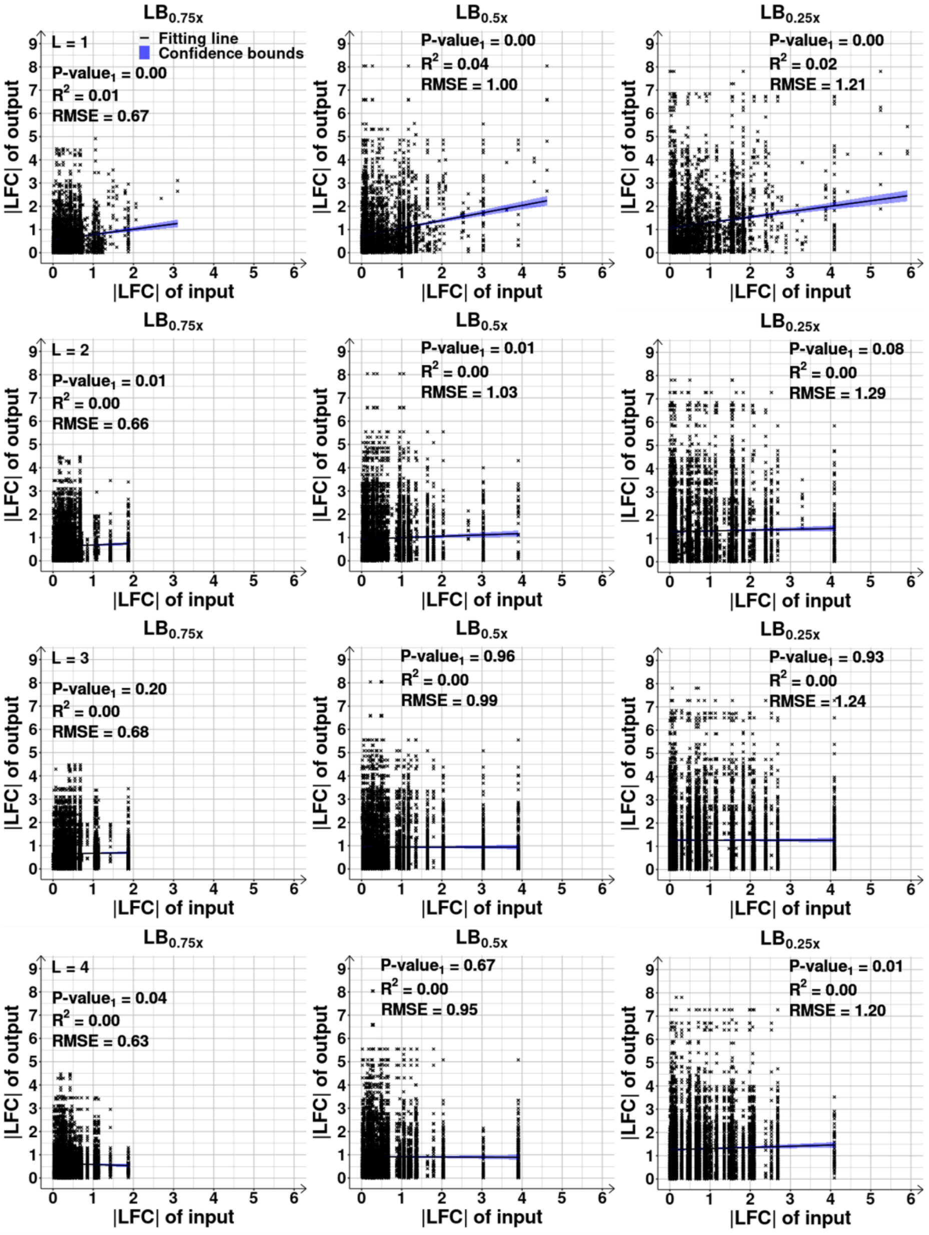


Figure S17. (Related to Figure 4B) Relationships between the changes in RNA abundances of input and output genes as a function of the position of the output gene in the Transcription Unit (TU). For each shift to $LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$, we plotted the absolute of log2 of fold change of each output gene (|LFC| of output) on the 1st, 2nd and 3rd positions in a TU following the transcription start site, against the |LFC| of each gene known to express a direct input transcription factor (|LFC| of input) of these 3 TU genes. Red lines are the best fitting lines. Blue lines are the null-model fitting lines and were obtained as described in Methods section *Statistical tests c*. Green lines are the best fitting after sorting the pairs input-output in ascending order to obtain the maximum correlation possible. All best fitting lines were obtained by linear least-squares regression fit using the MATLAB function *FITLM*. The colored shadow areas represent the 68% confidence bounds for the best fitted lines. We also obtained the p-values of statistical significance for the red (Pv^{Emp}), blue ($Pv^{Shuffled}$) and green (Pv^{Sorted}) lines under the null hypothesis that the data is best fit by a horizontal line. We only show the p-values when the null hypothesis was not rejected at 0.1 significance level. Supplementary Table S11 shows all p-values and coefficients of determination (R^2) of the red (R^2_{Emp}) and green (R^2_{Sorted}) fitted lines (see also Supplementary Table S12).


LB$_{0.75x}$
L = 1
Fitting line
Confidence bounds
P-value$_1$ = 0.00
R^2 = 0.01
RMSE = 0.67
LB$_{0.5x}$
P-value$_1$ = 0.00
R^2 = 0.04
RMSE = 1.00
LB$_{0.25x}$
P-value$_1$ = 0.00
R^2 = 0.02
RMSE = 1.21
LB$_{0.75x}$
L = 2
P-value$_1$ = 0.01
R^2 = 0.00
RMSE = 0.66
LB$_{0.5x}$
P-value$_1$ = 0.01
R^2 = 0.00
RMSE = 1.03
LB$_{0.25x}$
P-value$_1$ = 0.08
R^2 = 0.00
RMSE = 1.29
LB$_{0.75x}$
L = 3
P-value$_1$ = 0.20
R^2 = 0.00
RMSE = 0.68
LB$_{0.5x}$
P-value$_1$ = 0.96
R^2 = 0.00
RMSE = 0.99
LB$_{0.25x}$
P-value$_1$ = 0.93
R^2 = 0.00
RMSE = 1.24
LB$_{0.75x}$
L = 4
P-value$_1$ = 0.04
R^2 = 0.00
RMSE = 0.63
LB$_{0.5x}$
P-value$_1$ = 0.67
R^2 = 0.00
RMSE = 0.95
LB$_{0.25x}$
P-value$_1$ = 0.01
R^2 = 0.00
RMSE = 1.20
|LFC| of output
|LFC| of input
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6

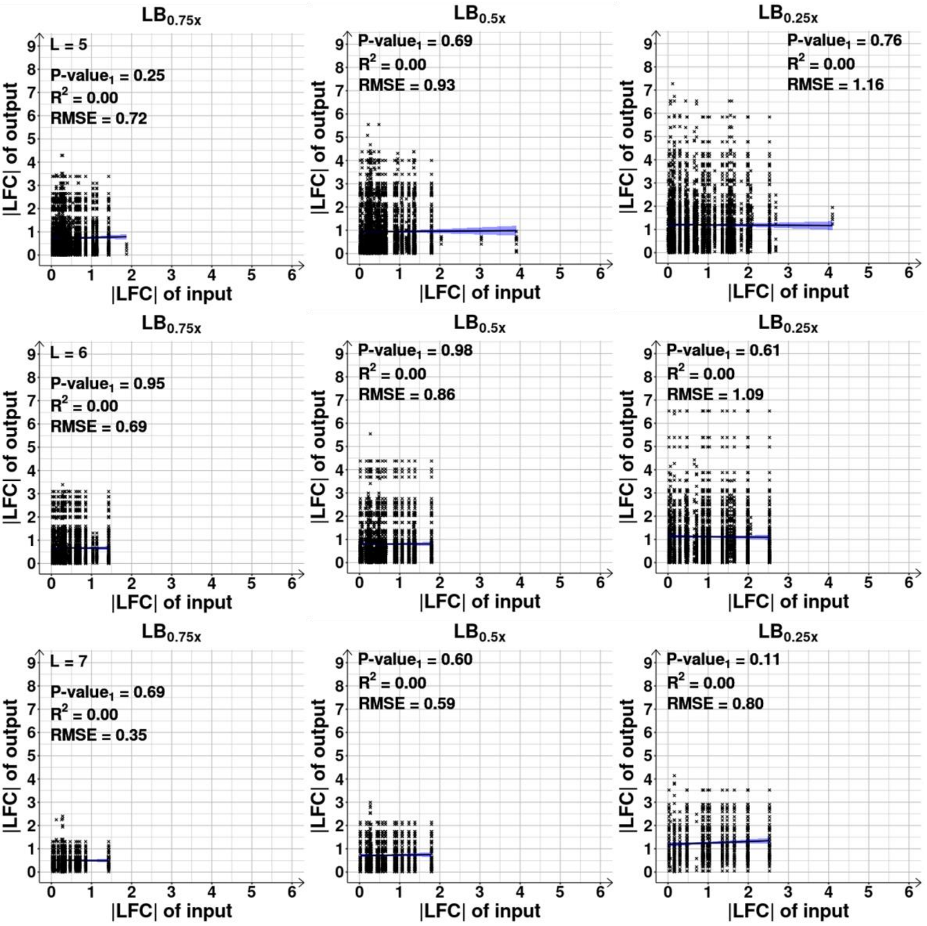

LB$_{0.75x}$
L = 5
P-value$_1$ = 0.25
R^2 = 0.00
RMSE = 0.72
LB$_{0.5x}$
P-value$_1$ = 0.69
R^2 = 0.00
RMSE = 0.93
LB$_{0.25x}$
P-value$_1$ = 0.76
R^2 = 0.00
RMSE = 1.16
LB$_{0.75x}$
L = 6
P-value$_1$ = 0.95
R^2 = 0.00
RMSE = 0.69
LB$_{0.5x}$
P-value$_1$ = 0.98
R^2 = 0.00
RMSE = 0.86
LB$_{0.25x}$
P-value$_1$ = 0.61
R^2 = 0.00
RMSE = 1.09
LB$_{0.75x}$
L = 7
P-value$_1$ = 0.69
R^2 = 0.00
RMSE = 0.35
LB$_{0.5x}$
P-value$_1$ = 0.60
R^2 = 0.00
RMSE = 0.59
LB$_{0.25x}$
P-value$_1$ = 0.11
R^2 = 0.00
RMSE = 0.80
|LFC| of output
|LFC| of input
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6

Figure S18. (Related to Figure 4C) Relationship between the changes in RNA abundances of output and input genes as a function of their distance in the Transcription Factor (TF) Network, TFN. Scatter plots between the |LFC| (absolute log2 of fold change) of pairs of genes as a function of their path length (L) (with L=1 to 7, L being the number of edges/input TFs in the TFN to go from one to the other), after shifting from the control condition. We included all gene pairs, regardless of being differentially expressed. The black lines are the best fitting ones (obtained by linear least-squares regression fit, MATLAB function *FITLM*) and the blue shadow areas are their 68% confidence bounds. Also shown are the coefficient of determination (R^2) and the root mean square error (RMSE) of the fitted lines, along with their p-values of statistical significance (P-value$_1$) (at 0.1 significance level, under the null hypothesis that the data is best fit by a horizontal line).

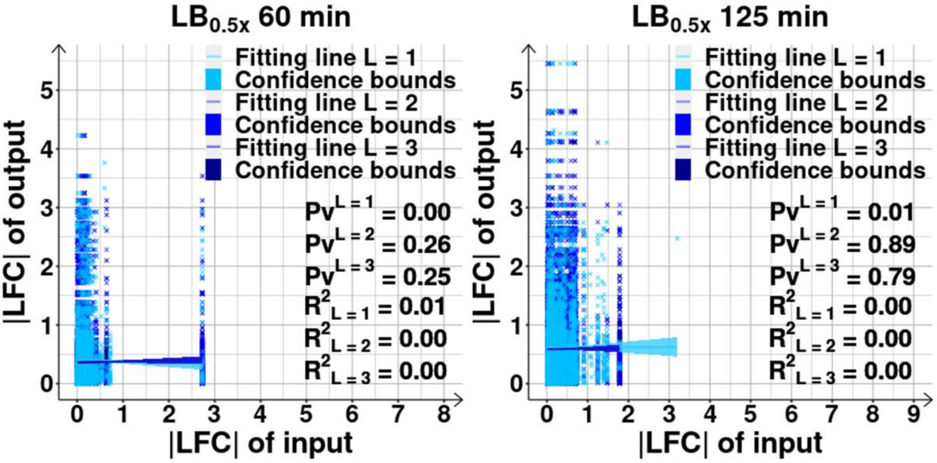


Figure S19. (Related to Figure 4C) Relationship between the changes in RNA abundances of output and input genes as a function of their path length prior to changes in RNAP and in the short-term following it. Scatter plots between |LFC| (absolute log2 of fold change) of output and input genes distanced by a minimum path length L of 1, 2, and 3 input TFs (edges) in the TFN, respectively (data from $LB_{0.5x}$). The data regards the short-term responses (125 min) and the responses prior to the changes in RNAP concentration (60 min). We included all gene pairs, regardless of being differentially expressed. The colored lines are the best fitting ones (obtained by linear least-squares regression fit, MATLAB function *FITLM*) and the corresponding blue shadow areas are their 68% confidence bounds. Also shown are the coefficient of determination (R^2) of the fitted lines, along with their p-values of statistical significance (Pv) (at 0.1 significance level, under the null hypothesis that the data is best fit by a horizontal line).

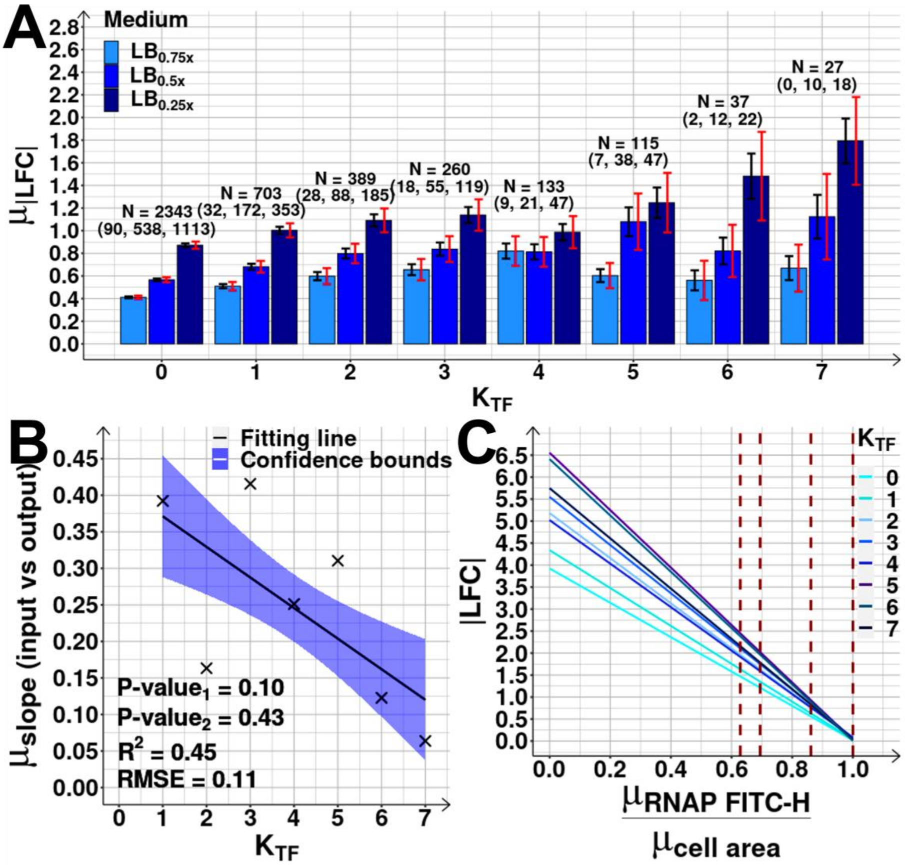


Figure S20. (Related to Figure 4D) Strengths of the shifts in RNA abundances of cohorts of genes with a given number of input transcription factors (TFs), K_{TF}.

(A) Mean of absolute log2 fold changes (|LFC|), $\mu_{|LFC|}$, of gene cohorts organized according to the K_{TF} of the component genes after shifting the medium. Black error bars represent the standard error of the mean (SEM), while red error bars represent the 95% confidence bounds of the SEM. Also shown is the number of genes of each class (*N*), and the numbers of differentially expressed genes (DEG) in each perturbation (Methods section *RNA-seq d*, assuming a False Discovery Rate < 0.05 and a |LFC| > 0.4248 ($LB_{0.75x}$), > 0.4085 ($LB_{0.5x}$) and > 0.4138 ($LB_{0.25x}$)).

(B) Scatter plot between K_{TF} and the average slope (over all shifts) of the fitting lines between |LFC| of the output gene and |LFC| of each of its gene expressing a direct input TF (i.e., input gene) (Supplementary Figure S14). The best fitting lines along with their 68% CI and statistics (coefficient of determination (R^2), root mean square error (RMSE), and P-values at 0.1

significance level) were obtained as described in Methods section *Statistical tests c*. A slope of 1 is expected if the input fully explains the output.

(C) Estimated rate of change of |LFC| with RNAP concentration for gene cohorts differing in K_{TF} (set to 0 in the control). RNAP concentration estimated from the ratio between mean RNAP levels measured by FITC-H, $\mu_{RNAP\ FITC-H}$ (Methods section *Flow-cytometry*), and the mean cell area ($\mu_{cell\ area}$) obtained from phase-contrast images. Only DEG are included. Vertical dashed red lines mark RNAP concentration levels at which RNA-seq was performed. The correlation between the two was obtained by linear least-squares regression fit using *FITLM* of MATLAB. Shown is the best fitting line for each K_{TF}. We also obtained p-values of statistical significance of the fitted regression line (at 0.1 significance level, under the null hypothesis that the data is best fit by a horizontal line). For all K_{TF}, the p-value < 0.1.

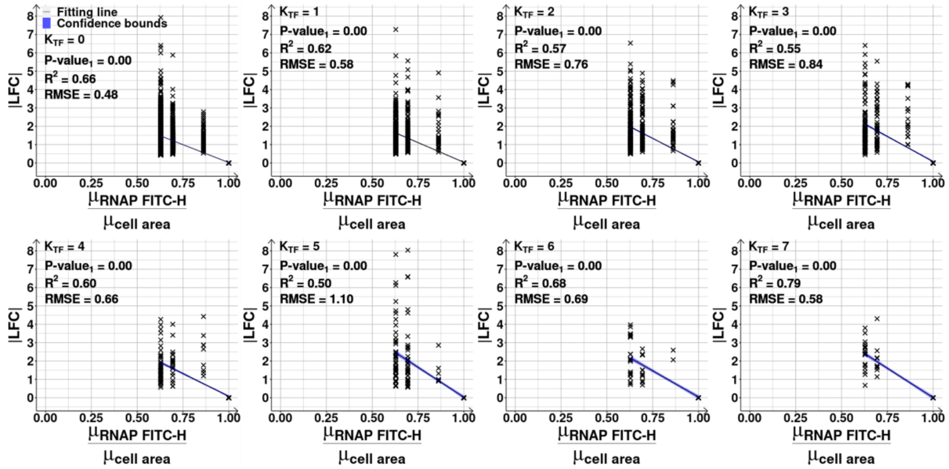


Figure S21. (Related to Supplementary Figure S20C) RNA changes as a function of the shift in RNAP concentration. Scatter plots between the |LFC| (absolute log2 of fold change) of each gene following each change in RNAP concentration. This concentration was estimated from the ratio between RNAP levels measured by FITC-H ($\mu_{RNAP\ FITC-H}$) using RL1314 cells (Methods section *Flow-cytometry*), and the mean cell area ($\mu_{cell\ area}$) obtained from phase-contrast images, relative to the control ($LB_{1.0x}$). Data for the cohorts of genes defined by K_{TF} (number of input transcription factors) from 0 to 7. Only differentially expressed genes are included (Methods section *RNA-seq d*, assuming a False Discovery Rate < 0.05 and |LFC| > 0.4248 ($LB_{0.75x}$), > 0.4085 ($LB_{0.5x}$) or > 0.4138 ($LB_{0.25x}$)). Best fitting lines obtained by linear least-squares regression fit using *FITLM* of MATLAB. Blue shadow areas are the 68% confidence bounds. Shown are the coefficient of determination (R^2) and the root mean square error (RMSE) of the fitted regression line, along with its p-value of statistical significance (P-value$_1$) (at 0.1 significance level, under the null hypothesis that the data is best fit by a horizontal line). Supplementary Table S13 shows tests of whether the lines differ statistically.

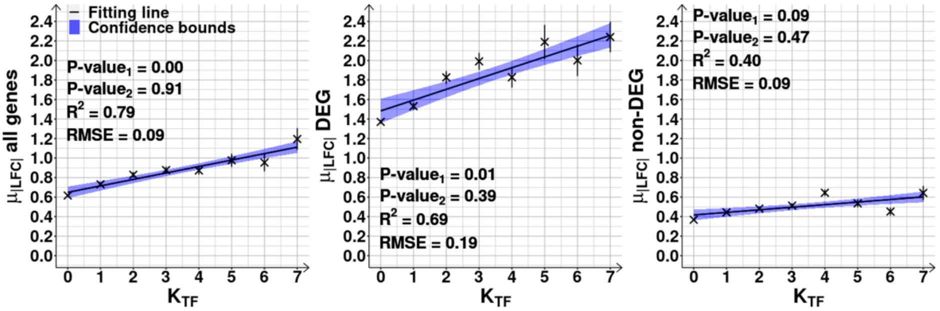


Figure S22. (Related to Figure 4D) RNA shifts of gene cohorts as a function of the number of input transcription factors (TFs), K_{TF}, of the component genes. Correlation plots between $\mu_{|LFC|}$, the mean of the absolute log2 fold changes (|LFC|), and K_{TF}. From left to right, results are shown for all genes, for differentially expressed genes (DEG) and, for non-differentially expressed genes 'non-DEG' (Methods section *RNA-seq d*), assuming a False Discovery Rate < 0.05 and |LFC| > 0.4248 ($LB_{0.75x}$), > 0.4085 ($LB_{0.5x}$) or > 0.4138 ($LB_{0.25x}$). $\mu_{|LFC|}$ obtained from data merged from three conditions ($LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$). The best fitting lines (solid black), their 68% confidence bounds (blue shadow area) and statistics (coefficient of determination (R^2), root mean square error (RMSE), and P-values at 0.1 significance level) were obtained as described in Methods section *Statistical tests c*. Both vertical and horizontal (not visible) error bars represent the standard errors of the mean.

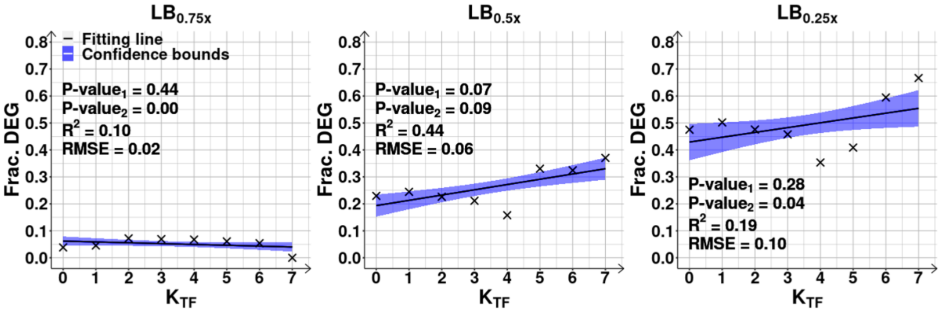


Figure S23. (Related to Figure 4D and Supplementary Figure S22) Fraction of differentially expressed genes (DEG) as a function of K_{TF}, the number of input transcription factors (TFs). DEG assessed assuming a False Discovery Rate < 0.05 and absolute log2 fold changes > 0.4248 for $LB_{0.75x}$, > 0.4085 for $LB_{0.5x}$ and > 0.4138 for $LB_{0.25x}$ (Methods section *RNA-seq d*). Data obtained for each shift from the control ($LB_{1.0x}$). Best fitting line (solid black) obtained by linear least-squares regression fit using the MATLAB function *FITLM*. The best fitting lines (solid black) along with their 68% confidence bounds (blue shadow areas) and statistics (coefficient of determination (R^2), root mean square error (RMSE), and P-values at 0.1 significance level) were obtained as described in Methods section *Statistical tests c*.

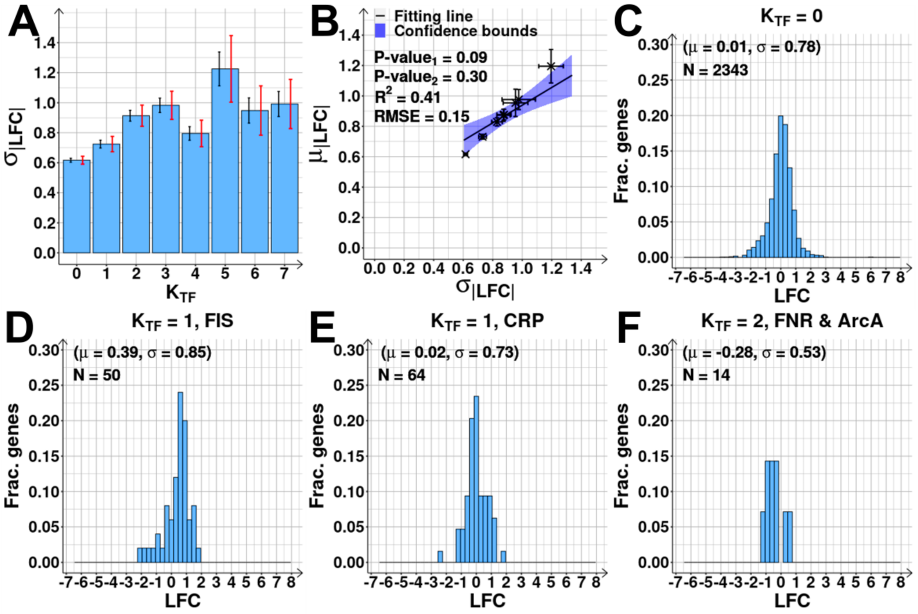


Figure S24. (Related to Figure 4D) Variability of the strengths of the shifts in RNA abundances of genes cohorts with a given number of input transcription factors (TFs), K_{TF}.

(A) Standard deviation of the absolute log2 of fold change (|LFC|), $\sigma_{|LFC|}$, of gene cohorts sharing the same K_{TF}, from 0 to 7. Results are from merged data from all shifts (LB$_{0.75x}$, LB$_{0.5x}$ and LB$_{0.25x}$). Black error bars are the standard error of the mean (SEM), while red error bars are the 95% confidence bounds (CB) of the SEM.

(B) Scattered plot of $\sigma_{|LFC|}$, following medium shifts, plotted against the respective mean of |LFC| ($\mu_{|LFC|}$). The best fitting line (solid black) along with its 68% CB (blue shadow area) and statistics (coefficient of determination (R^2), root mean square error (RMSE), and P-values at 0.1 significance level) were obtained as described in Methods section *Statistical tests c*.

Distribution of the fraction of genes with a given LFC, after shifting from the control (LB$_{1.0x}$) to LB$_{0.5x}$, for 4 gene cohorts: **(C)** Genes without input TFs; **(D)** Genes regulated only by FIS (K_{TF} = 1); and (E) Genes regulated only by CRP and, thus, K_{TF} = 1; **(F)** Genes regulated by FNR and ArcA (K_{TF} = 2).

Shown are the number of genes (N) and the mean (μ) and standard deviation (σ) of the LFC's of each distribution. Supplementary Table S14 shows statistical tests comparing the distributions.

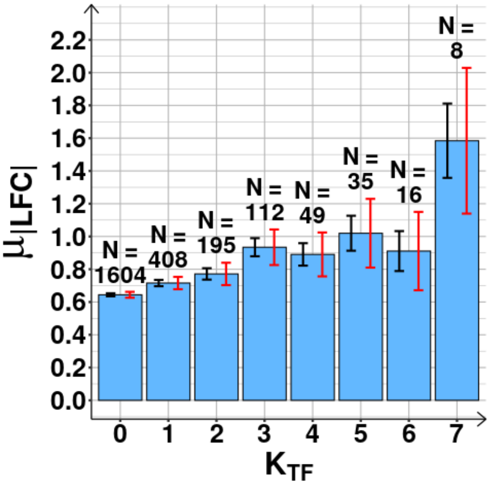


Figure S25. (Related to Figure 4D) Average RNA fold change of the first gene of each operon (including operons with only 1 gene) as a function of K_{TF}, the number of input transcription factors (TFs) of the output gene. Mean of absolute log2 fold change (|LFC|), $\mu_{|LFC|}$, of all genes in the first position of the operons, following the transcription start site. Results are from merged data from all shifts ($LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$). Black error bars are the standard error of the mean (SEM), while red error bars are the 95% confidence bounds of the SEM. *N* stands for the number of genes of each cohort.

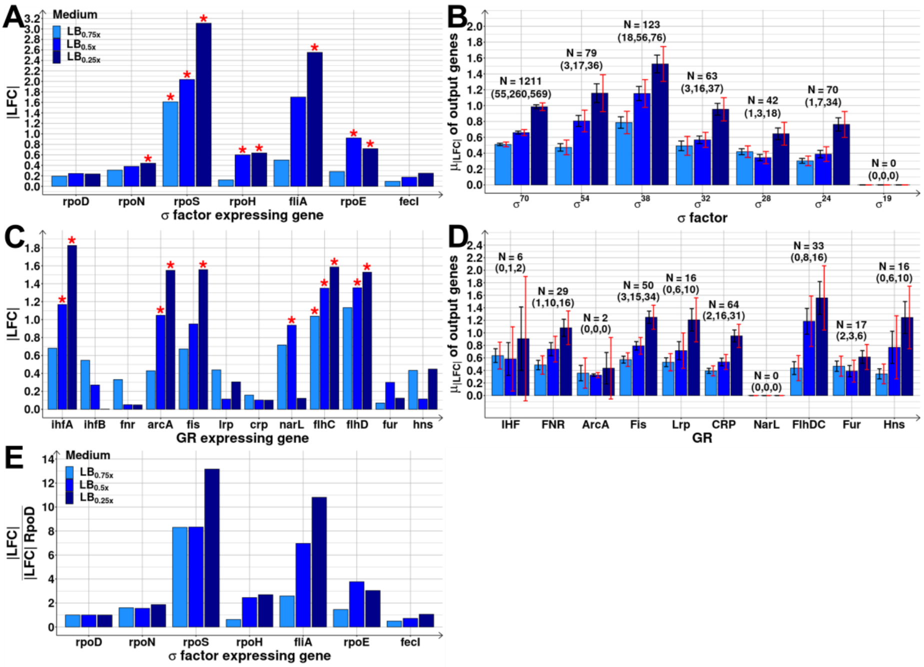


Figure S26. Shifts in the RNA abundances of global regulators and their output genes.

(A) Absolute of log2 fold change (|LFC|) of the genes expressing each σ factor (rpoD, rpoN, rpoS, rpoH, fliA, rpoE, fecI), following each dilution ($LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$).

(B) Mean |LFC|, $\mu_{|LFC|}$, of the *N* genes responsive only to σ^{70}, σ^{54}, σ^{38}, σ^{32}, σ^{28}, σ^{24} and σ^{19}, respectively.

(C) |LFC| of the genes expressing each global regulator (GR) (ihfA, ihfB, fnr, arcA, fis, lrp, crp, narL, flhC, flhD, fur, hns) (11,12).

(D) $\mu_{|LFC|}$ of the *N* genes that are responsive only to IHF, FNR, ArcA, Fis, Lrp, CRP, NarL, FlhDC, Fur and Hns, respectively.

In (A) and (C), the red asterisks denote changes that are classified as differentially expressed in Figure 3C, assuming a False Discovery Rate < 0.05 and a |LFC| > 0.4248 ($LB_{0.75x}$), > 0.4085 ($LB_{0.5x}$) or > 0.4138 ($LB_{0.25x}$) (Methods section *RNA-seq d*). In (B) and (D), the black error bars are the standard error of the mean (SEM), while red error bars are the 95% confidence bounds of the SEM.

(E) |LFC| of the genes expressing each σ factor (rpoN, rpoS, rpoH, fliA, rpoE, fecI), following each dilution ($LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$). Values are relative to the |LFC| of rpoD in the same conditions.

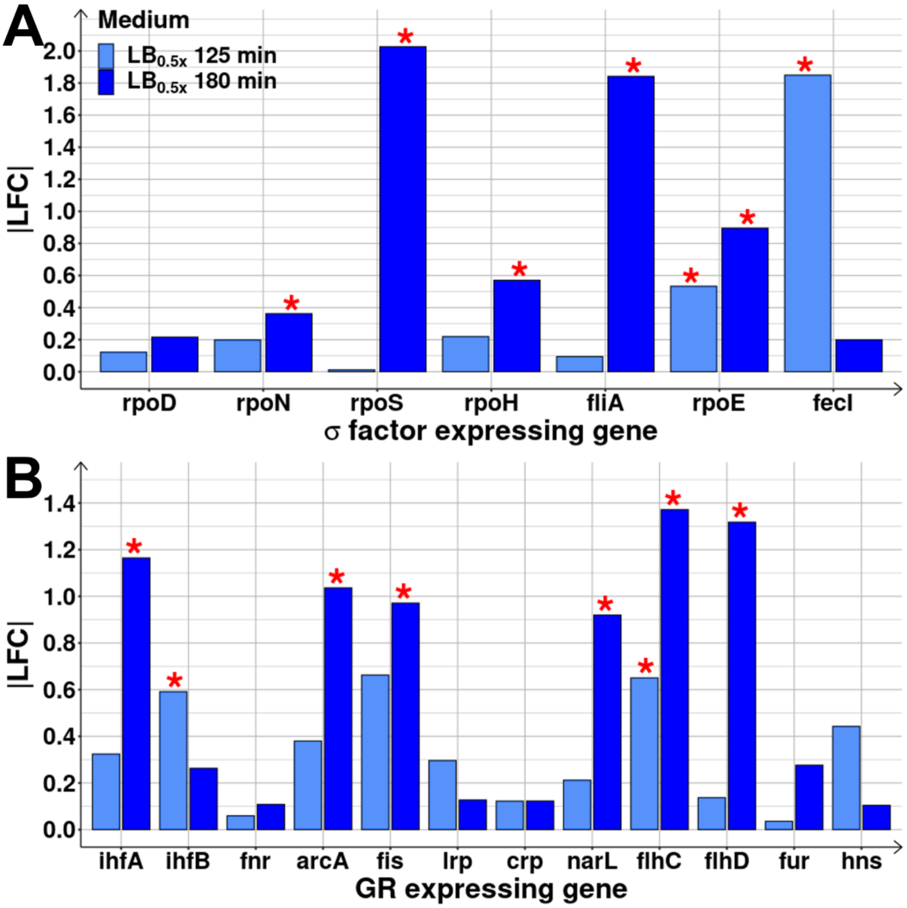


Figure S27. Shifts in the RNA abundances of global regulators and sigma factors at short- and mid-term response.

(A) Absolute of log2 fold change (|LFC|) of the genes expressing each σ factor (rpoD, rpoN, rpoS, rpoH, fliA, rpoE, fecI), when diluting the control medium ($LB_{1.0x}$) to the $LB_{0.5x}$ medium at 125 (short-term response) and 180 min (mid-term response). Here, the raw count matrices at 125 and 180 min were merged and only genes that passed the filtering were studied (Methods section *RNA-seq*).

(B) |LFC| of the genes expressing each global regulator (GR) (ihfA, ihfB, fnr, arcA, fis, lrp, crp, narL, flhC, flhD, fur, hns) under the same conditions as (A).

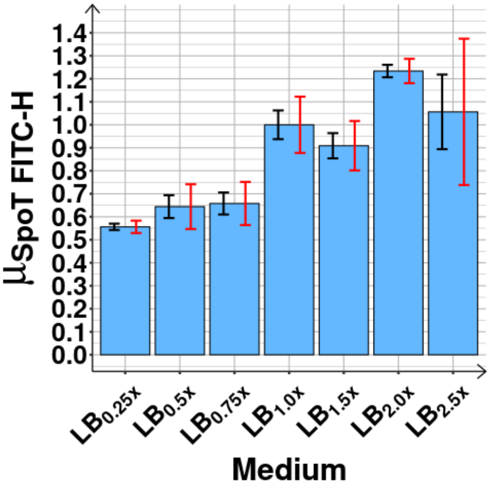


Figure S28. Expression levels of the spoT gene following medium shifts from LB$_{1.0x}$ to LB$_{0.75x}$, LB$_{0.5x}$, LB0$_{.25x}$, LB$_{1.5x}$, LB$_{2.0x}$ and LB$_{2.5x}$. Mean SpoT protein levels ($\mu_{SpoT\ FITC-H}$) at 180 min, from 3 biological replicates in each medium condition, measured by the mean single-cell fluorescence intensities (FITC channel, Methods section *Flow-cytometry*), after subtracting mean background fluorescence(s) and scaling to the control LB$_{1.0x}$ condition.

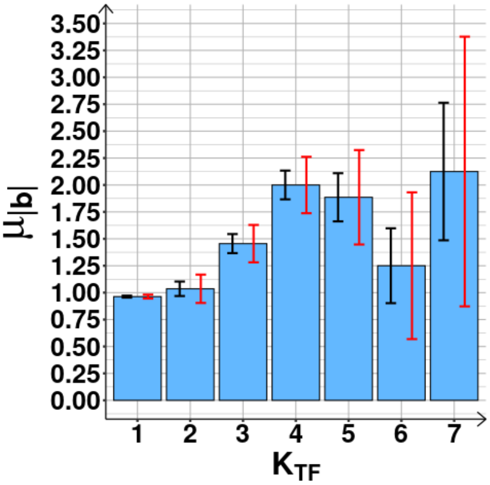


Figure S29. (Related to Figure 5A). **Mean absolute biases of the sets of input transcription factors (TFs) of the first gene of each operon, as a function of the number of input TFs, K_{TF}, of the output gene.** Mean of absolute of the sum of the regulatory effects 'r' of the inputs ($\mu_{|b|}$) obtained from RegulonDB for all genes assessed by RNA-seq. We removed the genes with positions higher than 1 following the transcription start site of the operon. Black error bars are the standard error of the mean (SEM), and red error bars are the 95% confidence bounds of the SEM.

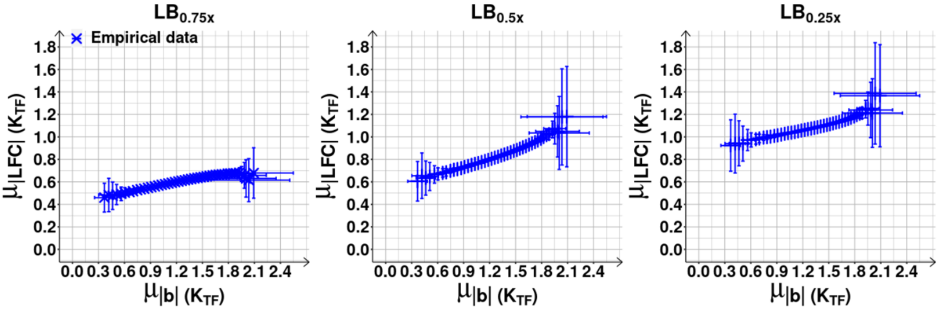


Figure S30. (Related to Figures 5B) Mean changes in RNA abundances of genes with an average K$_{TF}$, number of input transcription factors (TFs), as a function of $\mu_{|b|}$ (mean absolute bias in the regulatory effects of the input TFs) estimated using the ensemble approach. Mean of the absolute LFC (log2 of fold change), $\mu_{|LFC|}$, as a function of $\mu_{|b|}\left(K_{TF}\right)$ for each medium shift. Cohorts are assembled based on the K$_{TF}$ of the genes. Each blue cross is the average outcome from up to 7500 cohorts of 10 genes, Supplementary Results section *Estimation of the expected $\mu_{K_{TF}}$ and $\mu_{|b|}$ using an ensemble approach*, see also Supplementary Figure S31). Error bars (vertical and horizontal) are the standard error of the mean.

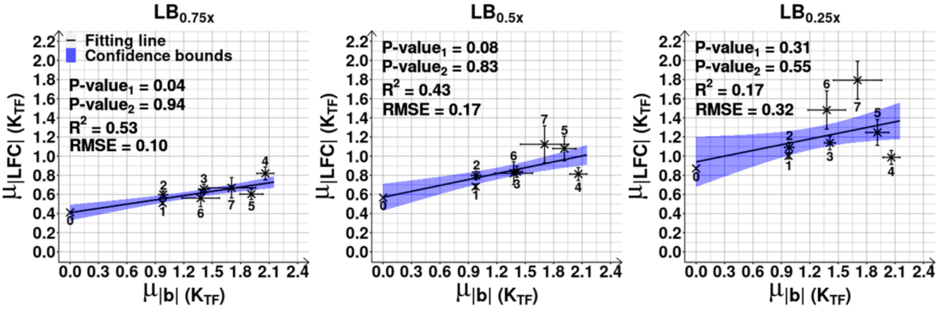
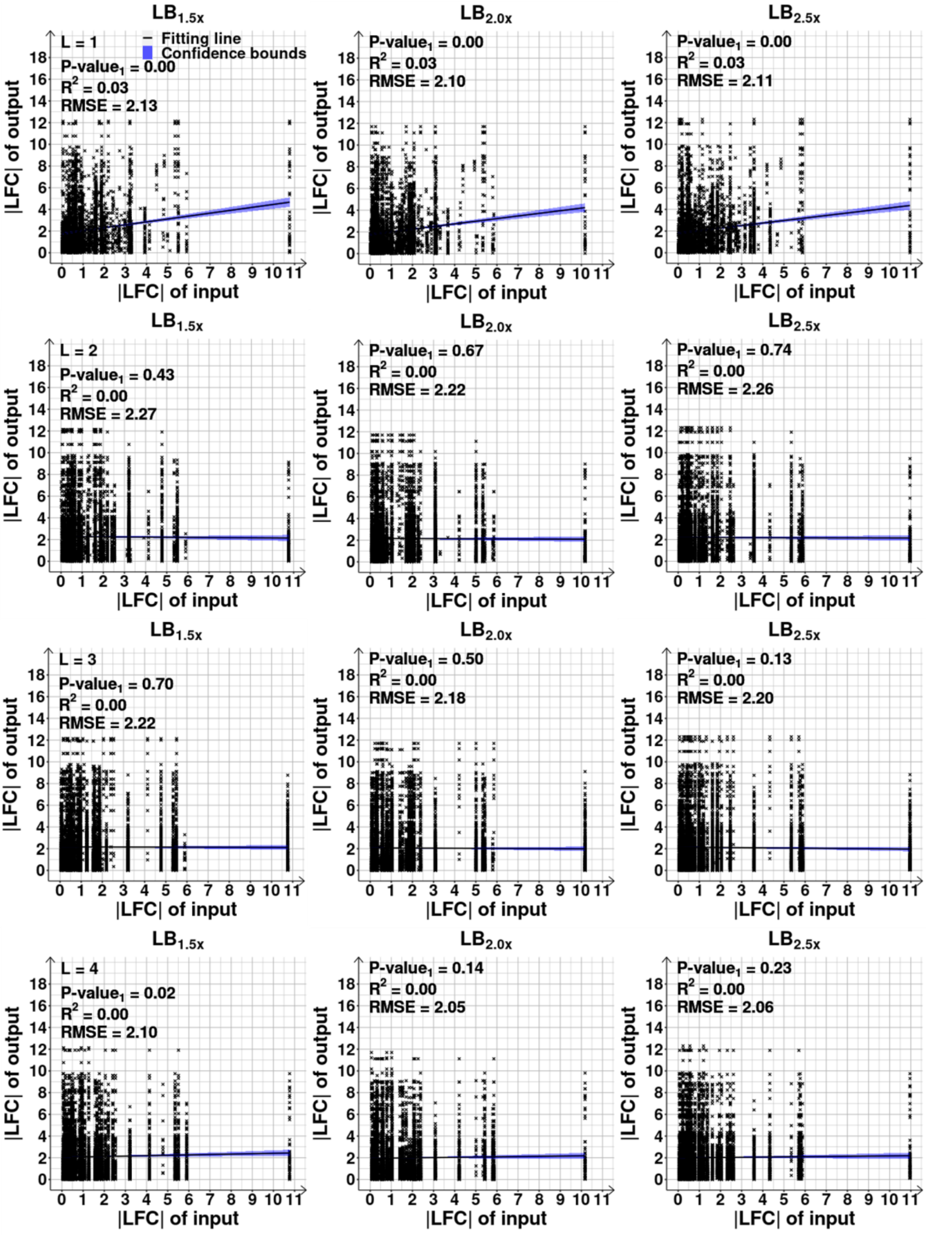


Figure S31. (Related to Supplementary Figure S30) Mean changes in RNA abundances of all genes with a specific K_{TF}, number of input transcription factors (TFs), as a function of the average $|b|$ (absolute of overall regulatory effect ('r') of the input TFs on the output gene). For each shift, we plot the mean of absolute LFC (log2 of fold change), $\mu_{|LFC|}$, against the mean of $|b|$, $\mu_{|b|}$, for each cohort of all genes with K_{TF} = 0 to 7. The error bars represent the standard error of the mean (SEM). We included all genes, regardless of being differentially expressed. The best fitting line (solid black) along with its 68% confidence bounds (blue shadow area) and statistics (coefficient of determination (R^2), root mean square error (RMSE), and P-values at 0.1 significance level) were obtained as described in Methods section *Statistical tests c*. Supplementary Table S18 shows the results of the statistical tests.


LB1.5x
L = 1
Fitting line
Confidence bounds
P-value1 = 0.00
R2 = 0.03
RMSE = 2.13
LB2.0x
P-value1 = 0.00
R2 = 0.03
RMSE = 2.10
LB2.5x
P-value1 = 0.00
R2 = 0.03
RMSE = 2.11
LB1.5x
L = 2
P-value1 = 0.43
R2 = 0.00
RMSE = 2.27
LB2.0x
P-value1 = 0.67
R2 = 0.00
RMSE = 2.22
LB2.5x
P-value1 = 0.74
R2 = 0.00
RMSE = 2.26
LB1.5x
L = 3
P-value1 = 0.70
R2 = 0.00
RMSE = 2.22
LB2.0x
P-value1 = 0.50
R2 = 0.00
RMSE = 2.18
LB2.5x
P-value1 = 0.13
R2 = 0.00
RMSE = 2.20
LB1.5x
L = 4
P-value1 = 0.02
R2 = 0.00
RMSE = 2.10
LB2.0x
P-value1 = 0.14
R2 = 0.00
RMSE = 2.05
LB2.5x
P-value1 = 0.23
R2 = 0.00
RMSE = 2.06
|LFC| of output
0 2 4 6 8 10 12 14 16 18
|LFC| of input
0 1 2 3 4 5 6 7 8 9 10 11

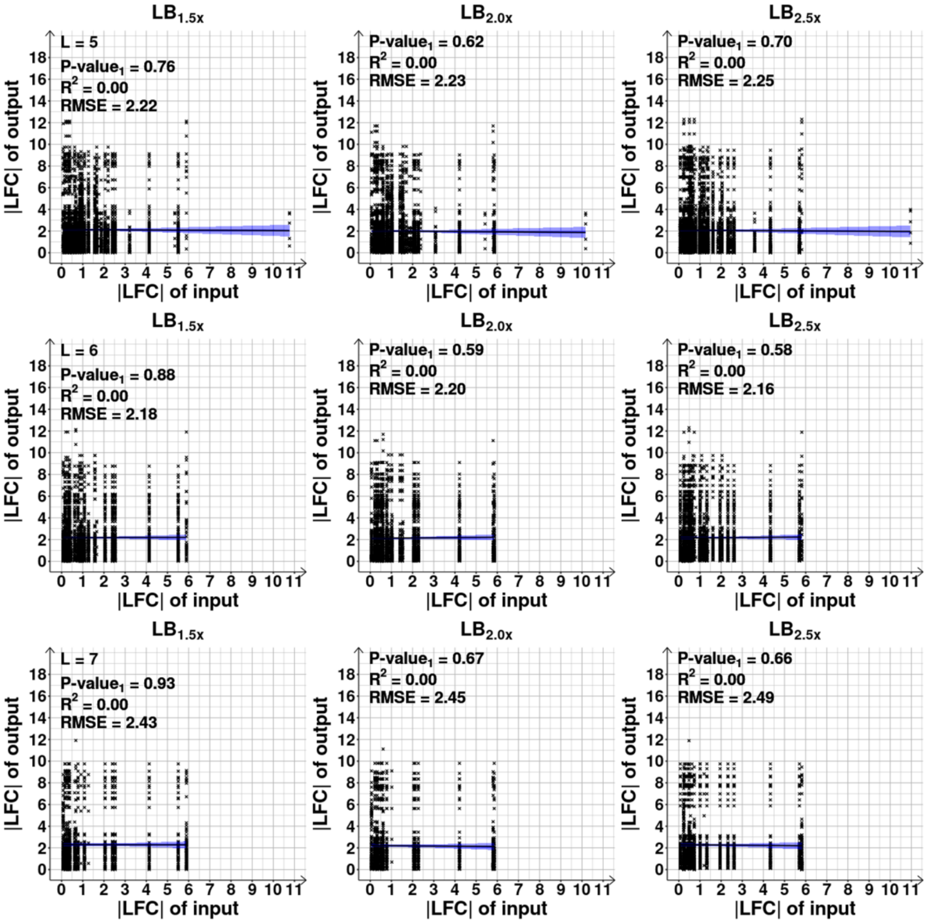


Figure S32. (Related to Figures 6D$_1$-6D$_3$) Changes in RNA abundances of output and input genes plotted as a function of their distance in the Transcription Factor (TF) Network, TFN. Scatter plots between the |LFC| (absolute log2 of fold change) of pairs of genes as a function of their path length (L) (with L=1 to 7, and L being the number of edges/input TFs in the TFN to go from one gene to the other), after shifting from the control condition. We included all gene pairs, regardless of being differentially expressed or not. Black lines are the best fitting ones (obtained by linear least-squares regression fit, MATLAB function *FITLM*) and blue shadow areas are their 68% confidence bounds. Also shown are the coefficient of determination (R^2), the root mean square error (RMSE) of the fitted lines, and their p-values of statistical significance (P-value$_1$) at 0.1 significance level, under the null hypothesis that the data is best fit by a horizontal line.

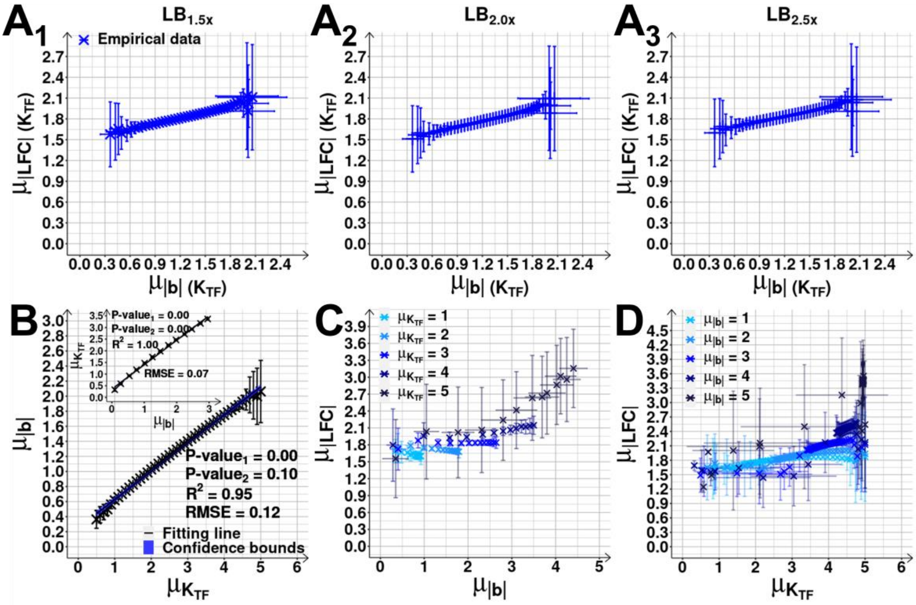


Figure S33. (Related to Figure 7D) Mean changes in RNA abundances of cohorts with an average K$_{TF}$ or $|b|$ (mean number of input TFs and mean absolute bias, respectively).

(A$_1$-A$_3$) Mean |LFC|, $\mu_{|LFC|}$, as a function of $\mu_{|b|}(K_{TF})$. Data obtained using the ensemble approach (Supplementary Results section *Estimation of the expected* $\mu_{K_{TF}}$ *and* $\mu_{|b|}$ *using an ensemble approach*). Each blue cross is the average outcome from up to 7.500 cohorts of 10 genes.

(B) Scatter plot of $\mu_{K_{TF}}$ against the corresponding $\mu_{|b|}$ of the cohorts in (A$_1$-A3). The inset shows the inverse, scatter plot between $\mu_{|b|}$ and $\mu_{K_{TF}}$, for the cohorts of Figure 7D, assembled based on $\mu_{|b|}$ (Supplementary Results section *Estimation of the expected* $\mu_{K_{TF}}$ *and* $\mu_{|b|}$ *using an ensemble approach*). Shown are best fitting lines and their 68% confidence bound (shadow areas, barely visible), coefficient of determination (R^2), root mean square error (RMSE), and P-value (Methods section *Statistical tests c*).

(C) Cohorts with increasing $\mu_{K_{TF}}$ but constant $\mu_{|b|}$ (from 1 to 5). $\mu_{K_{TF}}$ is plotted against the corresponding $\mu_{|LFC|}$, for each $\mu_{|b|}$.

(D) $\mu_{|b|}$ plotted against the corresponding $\mu_{|LFC|}$ for cohorts with constant $\mu_{K_{TF}}$ (from 1 to 5) and increasing $\mu_{|b|}$ (Supplementary Results section *Estimation of the expected* $\mu_{K_{TF}}$ *and* $\mu_{|b|}$ *using an ensemble approach*).

The error bars (vertical and horizontal) are the standard error of the mean. In (C) and (D), the data from the different conditions was merged and, since they slightly differ in mean (A_1-A_3), the SEM is larger than if in each condition separately. Also, comparatively, the SEM is much larger for $\mu_{K_{TF}}$ and $\mu_{|b|}$ equal to 5, due to which we did not extend the analysis further.

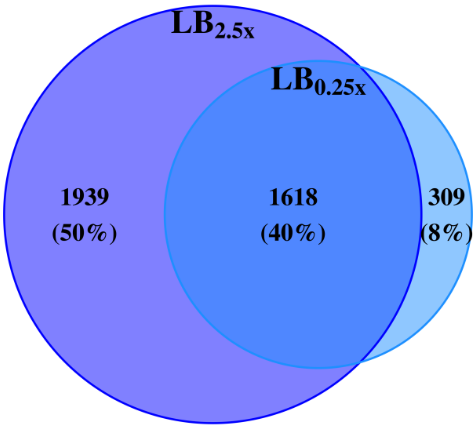


Figure S34. Venn diagram of the number and percentage of differentially expressed genes following RNAP changes. Data from 180 min following the shift to $LB_{0.25x}$ in the light blue circle and data from the shift to $LB_{2.5x}$ in the dark violet circle, when overlapping, become dark blue.

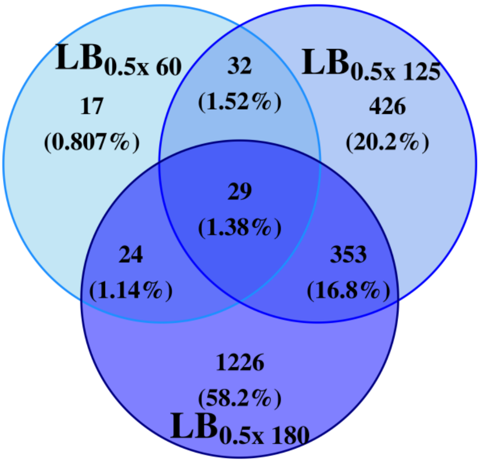


Figure S35. Venn diagram of the number and percentage of differentially expressed genes. Data from the shift to $LB_{0.5x}$ at 60 (prior to the changes in RNAP), 125 (short-term response) and 180 min (mid-term response).

SUPPLEMENTARY TABLES

Table S1. Variables. Short description of the main variables used.

Variable	Description
K_{TF}	Number of input transcription factors (TFs) of a gene.
$\mu_{K_{TF}}$	Mean of K_{TF} of a gene cohort.
r	Regulatory effect (+1, 0, -1)
$b = \left\|\sum r\right\|$	Bias. It equals the absolute of the sum of the regulatory effects, 'r' (each equaling +1 or -1) of the input TFs of a gene.
$\mu_{\|b\|}$	Mean absolute b of a gene cohort
$\mu_{\|LFC\|}$	Mean of |LFC| (absolute of log2 fold changes).
TFN	Transcription factor network
GR	Global regulator
L	Path length
CC	Closed complex formation
OC	Open complex formation
SEM	Standard error of the mean
CB	Confidence bounds

Table S2. (Related to Figure 2E and Supplementary Figure S2) Raw data of RNAP and GFP levels by western blotting. Shown are all the measurement values, for the three biological replicates, as reported by the software 'ImageLab' after analyzing the images obtained by Western Blot (Supplementary Figure S2).

RNAP Replicate 1

Sample	Channel	Band No.	Relative Front	Volume (Int)	Band %	Norm. Factor	Norm. Vol. (Int)
LB$_{1.0x}$	Chemi	1	0.344164	21470300	100	1.000000	21470300
LB$_{0.75x}$	Chemi	1	0.341757	16063000	100	0.916722	14725306
LB$_{0.5x}$	Chemi	1	0.327316	9521061	100	1.132143	10779204
LB$_{0.25x}$	Chemi	1	0.321300	3072728	100	1.798396	5525982

RNAP Replicate 2

Sample	Channel	Band No.	Relative Front	Volume (Int)	Band %	Norm. Factor	Norm. Vol. (Int)
LB$_{1.0x}$	Chemi	1	0.168000	19907978	100	1.000000	19907978
LB$_{0.75x}$	Chemi	1	0.162667	19135360	100	0.814673	15589056
LB$_{0.5x}$	Chemi	1	0.161333	8066324	100	1.511187	12189726
LB$_{0.25x}$	Chemi	1	0.152000	2059213	100	2.638734	5433715

RNAP Replicate 3

Sample	Channel	Band No.	Relative Front	Volume (Int)	Band %	Norm. Factor	Norm. Vol. (Int)
LB$_{1.0x}$	Chemi	1	0.213828	28023120	100	1.000000	28023120
LB$_{0.75x}$	Chemi	1	0.215109	22070880	100	1.006164	22206921
LB$_{0.5x}$	Chemi	1	0.212548	13065840	100	1.228384	16049866
LB$_{0.25x}$	Chemi	1	0.207426	4816080	100	1.782460	8584472

GFP Replicate 1

Sample	Channel	Band No.	Relative Front	Volume (Int)	Band %	Norm. Factor	Norm. Vol. (Int)
LB$_{1.0x}$	Chemi	1	0.142069	28705596	100	1.000000	28705596
LB$_{0.75x}$	Chemi	1	0.150345	17082868	100	1.355830	23161460
LB$_{0.5x}$	Chemi	1	0.155862	12007664	100	1.460927	17542319
LB$_{0.25x}$	Chemi	1	0.160000	4880930	100	1.847723	9018604

GFP Replicate 2

Sample	Channel	Band No.	Relative Front	Volume (Int)	Band %	Norm. Factor	Norm. Vol. (Int)
LB$_{1.0x}$	Chemi	1	0.236915	36092297	100	1.000000	36092297
LB$_{0.75x}$	Chemi	1	0.241047	26614700	100	0.961178	25581454
LB$_{0.5x}$	Chemi	1	0.245179	22808220	100	0.910526	20767477
LB$_{0.25x}$	Chemi	1	0.243802	14448294	100	1.187372	17155504

GFP Replicate 3

Sample	Channel	Band No.	Relative Front	Volume (Int)	Band %	Norm. Factor	Norm. Vol. (Int)
LB$_{1.0x}$	Chemi	1	0.201422	6419160	100	1.000000	6419160
LB$_{0.75x}$	Chemi	1	0.201422	3765849	100	1.281131	4824546
LB$_{0.5x}$	Chemi	1	0.202607	1445384	100	2.424255	3503979
LB$_{0.25x}$	Chemi	1	0.193128	1040612	100	2.320100	2414323

Table S3. (Related to Figure 3B and Supplementary Figure S8A-8C) Statistical test between the distributions in Supplementary Figure S8A-8C. P-values from the 2-sample T-test between the mean of the distributions (Methods section *Statistical tests a*).

P-value			
	$LB_{0.75x}$	**$LB_{0.5x}$**	**$LB_{0.25x}$**
$LB_{0.75x}$	1.00		
$LB_{0.5x}$	0.09	1.00	
$LB_{0.25x}$	0.95	0.24	1.00

Table S4. List of strains carrying an integrated YFP gene copy selected from a YFP strain library (Methods sections *Bacterial strains, media, growth conditions and curves* and *Flow-cytometry*). Cells used to test for correlations between the absolute log2 of fold change in protein levels and RNA abundances (Supplementary Figure S9).

	Strain [CGSC name]	**Genotype**	**Source**
1	cbpM [SX1494]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], cbpM791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13049)
2	tktB [SX1954]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], tktB792-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13509)
3	groS [SX1398]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1, groS791-YFP(::cat)	Yale CGSC (CGSC # 12953)
4	yafD [SX1626]	F-, yafD792-YFP(::cat), Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13181)
5	bolA [SX1087]	F-, Δ(argF-lac)169, bolA791-YFP(::cat), gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12642)
6	nudI [SX1271]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], nudI792-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12826)
7	cnu [SX1362]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], cnu-791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12917)
8	mobA [SX1354]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1, mobA791-YFP(::cat)	Yale CGSC (CGSC # 12909)
9	cpxR [SX1791]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1, cpxR791-YFP(::cat)	Yale CGSC (CGSC # 13346)
10	yciU [SX1384]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], yciU796-	Yale CGSC (CGSC # 12939)

		YFP(::cat), IN(rrnD-rrnE)1, rph-1	
11	yffL [SX1283]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], yffL791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12838)
12	recN [SX1220]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], recN796-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12775)
13	yceD [SX1638]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], yceD792-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13193)
14	rpsE [SX1340]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rpsE791-YFP(::cat), rph-1	Yale CGSC (CGSC # 12895)
15	mrcA [SX1938]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, mrcA791-YFP(::cat), rph-1	Yale CGSC (CGSC # 13493)
16	napD [SX1339]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], napD791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12894)
17	hyuA [SX1664]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], hyuA791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13219)
18	rbsB [SX1190]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1, rbsB791-YFP(::cat)	Yale CGSC (CGSC # 12745)
19	speC [SX1952]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], speC791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13507)
20	yjhP [SX1874]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1, yjhP794-YFP(::cat)	Yale CGSC (CGSC # 13429)

Table S5. (Related to Supplementary Figure S12) Features of the network topology. Network global topology parameters of the known TFN of *E. coli* and of (1000) randomly generated networks with the same number of nodes and edges but with 'Erdös Random' and with 'Scale-free' topology (using a Power-law exponent of -1.19) (13). Data from RegulonDB (14), from the genes in the RNA-seq data (Figure 3). For a description of the parameters, see Methods section *Transcription Factor Network of Escherichia coli*.

Parameters	***E. coli***	**Erdös random**	**Scale free**
No. nodes	4053	4053	4053
No. edges	4471	4471	4471
Clustering coefficient	0.06	$(1.79 \pm 1.51) \times 10^{-4}$	$(3.60 \pm 1.49) \times 10^{-4}$
No. connected components	2324	521.46 ± 16.98	563
Avg. path length (L)	3.28	24.06 ± 4.69	3.32
No. isolated nodes	2301	446.14 ± 17.25	485
No. self-loops	133	1.10 ± 1.02	0

Table S6. (Related to Figure 4A) Associations between having input transcription factors (TFs) and being a DEG (differentially expressed gene). P-values obtained by a Fisher test (Methods section *Statistical tests b*) to determine if there is an association between having one or more input TFs and being a DEG (Methods section *RNA-seq d*, assuming a False Discovery Rate < 0.05 and absolute log2 fold change (|LFC|) > 0.4248 for $LB_{0.75x}$, > 0.4085 for $LB_{0.5x}$ and > 0.4138 for $LB_{0.25x}$). For each condition, we use the specific values of $\mu_{|LFC|}$ (mean of |LFC|) and number of DEG. The null hypothesis is that there is random association between the two variables. The test rejects the null hypothesis at 0.1 significance level.

Medium	P-value
$LB_{0.75x}$	0.00
$LB_{0.5x}$	0.29
$LB_{0.25x}$	0.85

Table S7. (Related to Figure 4B, and Supplementary Figures S14 and S20B) Statistics of the linear fits in Supplementary Figure S14. R^2_{Emp}, R^2_{Sorted}, Pv^{Emp}, $Pv^{Shuffled}$ and Pv^{Sorted} values for gene cohorts differing in K_{TF} (number of known input transcription factors (TFs)) (from 1 to 7) and medium ($LB_{0.75x}$, $LB_{0.5x}$ and $LB_{0.25x}$), when testing in Supplementary Figure S14 the correlation between absolute of log2 fold change (|LFC|) of outputs genes and each gene known to express their input TFs.

	$LB_{0.75x}$				
	$Pv^{Shuffled}$	**Pv^{Emp}**	**Pv^{Sorted}**	**R^2_{Emp}**	**R^2_{Sorted}**
K_{TF} = 1	0.83	0.00	0.00	0.08	0.87
K_{TF} = 2	0.74	0.03	0.00	0.01	0.95
K_{TF} = 3	0.97	0.00	0.00	0.04	0.87
K_{TF} = 4	0.98	0.00	0.00	0.04	0.93
K_{TF} = 5	0.81	0.00	0.00	0.03	0.91
K_{TF} = 6	1.00	0.72	0.00	0.00	0.88
K_{TF} = 7	1.00	1.00	0.00	0.00	0.93
	$LB_{0.5x}$				
	$Pv^{Shuffled}$	**Pv^{Emp}**	**Pv^{Sorted}**	**R^2_{Emp}**	**R^2_{Sorted}**
K_{TF} = 1	0.96	0.00	0.00	0.11	0.88
K_{TF} = 2	0.82	0.00	0.00	0.02	0.92
K_{TF} = 3	0.63	0.00	0.00	0.06	0.88
K_{TF} = 4	0.86	0.01	0.00	0.01	0.96
K_{TF} = 5	0.91	0.00	0.00	0.04	0.85
K_{TF} = 6	0.97	0.00	0.00	0.03	0.88
K_{TF} = 7	0.96	0.10	0.00	0.01	0.93
	$LB_{0.25x}$				
	$Pv^{Shuffled}$	**Pv^{Emp}**	**Pv^{Sorted}**	**R^2_{Emp}**	**R^2_{Sorted}**
K_{TF} = 1	0.85	0.00	0.00	0.06	0.93
K_{TF} = 2	0.97	0.01	0.00	0.01	0.96
K_{TF} = 3	0.87	0.00	0.00	0.07	0.93
K_{TF} = 4	0.80	0.68	0.00	0.00	0.89
K_{TF} = 5	0.89	0.00	0.00	0.02	0.88
K_{TF} = 6	0.94	0.09	0.00	0.01	0.88
K_{TF} = 7	0.87	0.93	0.00	0.00	0.88

Table S8. (Related to Figure 4B, and Supplementary Figures S15) Statistics of the linear fits in Supplementary Figure S15. R^2_{Emp}, R^2_{Sorted}, Pv^{Emp}, $Pv^{Shuffled}$ and Pv^{Sorted} values for gene cohorts differing in K_{TF} (number of known input transcription factors (TFs)) (from 1 to 7) in $LB_{0.5x}$ medium (60 and 125 min), when testing in Supplementary Figure S15 the correlation between absolute of log2 fold change (|LFC|) of outputs genes and each gene known to express their input TFs.

	$LB_{0.5x}$ 60 mim					**$LB_{0.5x}$ 125 min**				
	$Pv^{Shuffled}$	**Pv^{Emp}**	**Pv^{Sorted}**	**R^2_{Emp}**	**R^2_{Sorted}**	**$Pv^{Shuffled}$**	**Pv^{Emp}**	**Pv^{Sorted}**	**R^2_{Emp}**	**R^2_{Sorted}**
K_{TF} = 1	0.88	0.00	0.00	0.01	0.84	0.98	0.25	0.00	0.00	0.86
K_{TF} = 2	0.89	0.18	0.00	0.00	0.78	0.81	0.00	0.00	0.03	0.91
K_{TF} = 3	0.95	0.25	0.00	0.00	0.64	0.72	0.91	0.00	0.00	0.96
K_{TF} = 4	0.98	0.47	0.00	0.00	0.76	0.82	0.00	0.00	0.02	0.95
K_{TF} = 5	0.99	0.01	0.00	0.01	0.56	0.93	0.09	0.00	0.00	0.90
K_{TF} = 6	0.97	0.01	0.00	0.03	0.38	0.99	0.98	0.00	0.00	0.88
K_{TF} = 7	0.95	1.00	0.00	0.00	0.32	0.90	0.42	0.00	0.00	0.90

Table S9. (Related to Figure 4B and Supplementary Figure S16) Statistics of the linear fits in Supplementary Figure S16. R^2_{Emp}, R^2_{Sorted}, Pv^{Emp}, $Pv^{Shuffled}$ and Pv^{Sorted} for results in Supplementary Figure S16, in various media. Data for the 1st, 2nd and 3rd genes (following the transcription start site) of operons of size 3, when confronting the correlation between absolute of LFC, log2 fold change of each of these output gene of the operon and each gene known to express an input transcription factor (TF) common to all 3 genes.

	$LB_{0.75x}$				
	$Pv^{Shuffled}$	**Pv^{Emp}**	**Pv^{Sorted}**	**R^2_{Emp}**	**R^2_{Sorted}**
1st gene	0.98	0.00	0.00	0.06	0.90
2nd gene	0.93	0.03	0.00	0.02	0.92
3rd gene	0.98	0.00	0.00	0.04	0.93
	$LB_{0.5x}$				
	$Pv^{Shuffled}$	**Pv^{Emp}**	**Pv^{Sorted}**	**R^2_{Emp}**	**R^2_{Sorted}**
1st gene	0.97	0.00	0.00	0.06	0.88
2nd gene	0.98	0.00	0.00	0.05	0.85
3rd gene	0.93	0.00	0.00	0.05	0.92
	$LB_{0.25x}$				
	$Pv^{Shuffled}$	**Pv^{Emp}**	**Pv^{Sorted}**	**R^2_{Emp}**	**R^2_{Sorted}**
1st gene	1.00	0.00	0.00	0.07	0.88
2nd gene	0.98	0.00	0.00	0.05	0.89
3rd gene	0.90	0.00	0.00	0.04	0.93

Table S10. (Related to Figure 4B and Supplementary Figure S16) Analysis of Covariance of results in Supplementary Figure S16. We performed an analysis of covariances, ANCOVA (Methods section *Statistical tests d*) test, to assess if the red fitting lines in Supplementary Figure S16 have the same intercept (I) and/or slope (S). The fitting lines were calculated for the shifts from the control, to assess the correlation between the absolute of log2 fold change of each output gene belonging to an operon of size 3 (i.e., 1st, 2nd and 3rd gene in the operon following the transcription start site), and each gene known to express an input TF, common to the 3 genes in the operon. We evaluated if the fitting lines differ between the 1st, 2nd, and 3rd gene, for each shift. Also, we evaluated if the fitting lines differ between the different shifts, for the 1st, 2nd, and 3rd gene, respectively. We consider 2 lines statistically different in I and/or S at 0.1 significance level.

P-value						
	$LB_{0.75x}$		**$LB_{0.5x}$**		**$LB_{0.25x}$**	
	I	**S**	**I**	**S**	**I**	**S**
1st gene vs 2nd gene	0.85	0.31	0.42	0.51	0.53	0.31
1st gene vs 3rd gene	0.96	0.54	0.77	0.38	1.00	0.12
2nd gene vs 3rd gene	0.81	0.67	0.54	0.82	0.43	0.56
	1st gene		**2nd gene**		**3rd gene**	
	I	**S**	**I**	**S**	**I**	**S**
$LB_{0.75x}$ vs $LB_{0.5x}$	0.12	0.75	0.30	0.83	0.14	0.66
$LB_{0.75x}$ vs $LB_{0.25x}$	0.06	0.72	0.00	0.94	0.03	0.33
$LB_{0.5x}$ vs $LB_{0.25x}$	0.63	0.94	0.03	0.69	0.37	0.45

Table S11. (Related to Figure 4B and Supplementary Figure S17) Statistics of the linear fits in Supplementary Figure S17. R^2_{Emp}, R^2_{Sorted}, Pv^{Emp}, $Pv^{Shuffled}$ and Pv^{Sorted} values for results in Supplementary Figure S17, in various shifts. Data for the 1st, 2nd and 3rd gene (following the transcription start site) belonging to a Transcription Unit (TU) of size 3, when confronting the correlation between the absolute LFC of each output gene of the TU and each gene expressing an input TF common to all 3 genes.

	$LB_{0.75x}$				
	$Pv^{Shuffled}$	**Pv^{Emp}**	**Pv^{Sorted}**	**R^2_{Emp}**	**R^2_{Sorted}**
1st gene	1.00	0.00	0.00	0.05	0.89
2nd gene	0.91	0.01	0.00	0.02	0.91
3rd gene	0.96	0.00	0.00	0.03	0.90
	$LB_{0.5x}$				
	$Pv^{Shuffled}$	**Pv^{Emp}**	**Pv^{Sorted}**	**R^2_{Emp}**	**R^2_{Sorted}**
1st gene	0.95	0.00	0.00	0.05	0.84
2nd gene	0.96	0.00	0.00	0.05	0.81
3rd gene	0.79	0.00	0.00	0.04	0.88
	$LB_{0.25x}$				
	$Pv^{Shuffled}$	**Pv^{Emp}**	**Pv^{Sorted}**	**R^2_{Emp}**	**R^2_{Sorted}**
1st gene	0.82	0.00	0.00	0.04	0.88
2nd gene	0.97	0.00	0.00	0.04	0.86
3rd gene	0.88	0.02	0.00	0.02	0.92

Table S12. (Related to Figure 4B and Supplementary Figure S17). Analysis of Covariance of results in Supplementary Figure S17. Analysis of covariances, ANCOVA (Methods section *Statistical tests d*) to test if the red fitting lines in Supplementary Figure S17 have the same intercept (I) and/or slope (S). Fitting lines calculated for the shifts from the control, to assess the correlation between the absolute of LFC, log2 fold change of each output gene belonging to a Transcription Unit (TU) of size 3 (i.e., 1st, 2nd and 3rd gene in the TU following the transcription start site), and the |LFC| of each gene known to express an input TF common to the 3 genes in the TU. We evaluate if the fitting lines differ between the 1st, 2nd, and 3rd gene, for each shift. Also, we evaluate if the fitting lines differ between the different shifts, for the 1st, 2nd, and 3rd gene, respectively. We consider 2 lines to be statistically different in I and/or S at 0.1 significance level.

P-value						
	$LB_{0.75x}$		**$LB_{0.5x}$**		**$LB_{0.25x}$**	
	I	**S**	**I**	**S**	**I**	**S**
1st gene vs 2nd gene	0.94	0.46	0.64	1.00	0.72	0.98
1st gene vs 3rd gene	0.79	0.69	0.88	0.72	0.61	0.32
2nd gene vs 3rd gene	0.86	0.73	0.52	0.72	0.89	0.30
	1st gene		**2nd gene**		**3rd gene**	
	I	**S**	**I**	**S**	**I**	**S**
$LB_{0.75x}$ vs $LB_{0.5x}$	0.09	0.42	0.23	1.00	0.09	0.50
$LB_{0.75x}$ vs $LB_{0.25x}$	0.00	0.20	0.00	0.66	0.00	0.12
$LB_{0.5x}$ vs $LB_{0.25x}$	0.18	0.52	0.03	0.52	0.06	0.21

Table S13. (Related to Supplementary Figures S20C and S21) Analysis of Covariance of results in Supplementary Figure S21. Analysis of covariances (ANCOVA, Methods section *Statistical tests d*) to assess if the fitting lines in Supplementary Figure S21 have the same intercept (I) and/or slope (S). We evaluate, as a function of K_{TF} (number of input transcription factors), the chance of correlation between |LFC| and the shift in mean RNA Polymerase (RNAP) concentration. We consider 2 lines to be statistically different in I and/or S if P-value < 0.1.

	P-value															
	K_{TF} = 0		**K_{TF} = 1**		**K_{TF} = 2**		**K_{TF} = 3**		**K_{TF} = 4**		**K_{TF} = 5**		**K_{TF} = 6**		**K_{TF} = 7**	
	I	**S**	**I**	**S**	**I**	**S**	**I**	**S**	**I**	**S**	**I**	**S**	**I**	**S**	**I**	**S**
K_{TF} = 0	1.00	1.00														
K_{TF} = 1	0.00	0.00	1.00	1.00												
K_{TF} = 2	0.00	0.00	0.00	0.00	1.00	1.00										
K_{TF} = 3	0.00	0.00	0.00	0.00	0.12	0.19	1.00	1.00								
K_{TF} = 4	0.00	0.00	0.00	0.02	0.65	0.64	0.14	0.18	1.00	1.00						
K_{TF} = 5	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	1.00	1.00				
K_{TF} = 6	0.00	0.00	0.00	0.00	0.20	0.26	0.70	0.69	0.11	0.15	0.24	0.31	1.00	1.00		
K_{TF} = 7	0.00	0.00	0.00	0.00	0.01	0.03	0.13	0.17	0.00	0.01	0.85	0.89	0.24	0.30	1.00	1.00

Table S14. (Related to Figure 4D and Supplementary Figures S24C-S24F) Statistical tests between the distributions in Supplementary Figures S24C-S24F. P-values from the 2-sample T-test and 2-sample KS-test confronting the distributions and the Z-test (Methods section *Statistical tests a*). For the z-test, the mean and standard deviation are from the distribution of genes without input transcription factors (TFs), K_{TF} = 0. The data is from the shift to $LB_{0.25x}$.

P-value, 2-sample T-test				
	K_{TF} = 0	**K_{TF} = 1, CRP**	**K_{TF} = 1, FIS**	**K_{TF} = 2, FNR & ArcA**
K_{TF} = 0	1.00			
K_{TF} = 1, CRP	0.88	1.00		
K_{TF} = 1, FIS	0.00	0.02	1.00	
K_{TF} = 2, FNR & ArcA	0.06	0.08	0.00	1.00
P-value, 2-sample KS-test				
	K_{TF} = 0	**K_{TF} = 1, CRP**	**K_{TF} = 1, FIS**	**K_{TF} = 2, FNR & ArcA**
K_{TF} = 0	1.00			
K_{TF} = 1, CRP	0.64	1.00		
K_{TF} = 1, FIS	0.00	0.00	1.00	
K_{TF} = 2, FNR & ArcA	0.10	0.50	0.00	1.00

P-value, Z-test	
K_{TF} = 0	1.00
K_{TF} = 1, CRP	0.89
K_{TF} = 1, FIS	0.00
K_{TF} = 2, FNR & ArcA	0.16

Table S15. (Related to Supplementary Figures S26A and S26C) Analysis of the potential role of each global regulatory in the genome-wide shifts in RNA abundances.

rpoD (σ^{70})	Likely not influential due to lack of changes in its concentration.
rpoN (σ^{54})	Likely not influential due to lack of changes in its concentration in 2 of 3 perturbations.
rpoS (σ^{38})	Likely not influential since it follows the RNAP changes (Supplementary Figure S5).
rpoH (σ^{34})	Likely not influential due to lack of changes in its concentration in the first perturbation and the inconsistency with changes in the LFCs of its outputs (Supplementary Figure S26B).
fliA (σ^{28})	Likely not influential due to lack of changes in its concentration in 2 of 3 perturbations.
rpoE (σ^{24})	Likely not influential since, while it responded to 2 of 3 perturbations, its response strength was inconsistent with the perturbation strengths. Also, its outputs responded inconsistently.
fecI (σ^{19})	Likely not influential since it did not respond to any perturbation.
ihfA	Likely not influential since its outputs did not respond consistently.
ihfB	Likely not influential due to lack of change in its concentration with the perturbations.
fnr	Likely not influential due to lack of change in its concentration with the perturbations.
arcA	Likely not influential since its outputs did not respond consistently to either arcA or RNAP.
fis	Likely not influential due to lack of change in its concentration in 2 of 3 perturbations.
lrp	Likely not influential due to lack of change in its concentration with the perturbations.
crp	Likely not influential due to lack of change in its concentration with the perturbations.
narL	Likely not influential due to lack of change in its concentration in 2 of 3 perturbations.
flhC	Likely not influential as it followed the RNAP changes and only controls 32 genes.
flhD	Likely not influential as it followed the RNAP changes and only controls 32 genes.
fur	Likely not influential due to lack of change in its concentration with the perturbations.
hns	Likely not influential due to lack of change in its concentration with the perturbations.

Table S16. (Related to Figures 5A and 5B and Supplementary Figure S12) Numbers of genes with input transcription factors (TFs) with a given overall bias. Number of genes with a given absolute of the sum of the regulatory effects, 'r' of inputs ($|b|$) from 0 to 5. Calculated from data from RegulonDB.

$\lvert b \rvert$	**0**	**1**	**2**	**3**	**4**	**5**
No. genes	2598	996	305	107	27	8

Table S17. (Related to Figure 5A) Number of genes with all input transcription factors (TFs) having the same regulatory effect (‘r’). Number of genes whose absolute sum of the r of inputs ($|b|$) equals their number of input TFs, K_{TF} (for K_{TF} = 0 to 7). As K_{TF} increases, the fraction of genes with $|b|$ = K_{TF} decreases.

K_{TF}	No. genes	No. genes with $\|b\|$ = K_{TF}
0	2343	2343
1	703	686
2	389	169
3	260	61
4	133	19
5	115	8
6	37	0
7	27	0

Table S18. (Related to Supplementary Figure S31) Analysis of Covariance of results in Supplementary Figure S31. We performed an analysis of covariances, ANCOVA (Methods section *Statistical tests d*) assessing if the fitting lines in Supplementary Figure S31 have the same intercept (I) and/or slope (S). The fitting lines were calculated for the shifts from the control ($LB_{1.0x}$) to other media, to assess the correlation between the mean of |LFC| (absolute of LFC, log2 fold change), $\mu_{|LFC|}$, and the mean of $|b|$ (absolute of the sum of the regulatory effects ('r') of inputs), for each cohort of genes with a specific K_{TF} (number of input transcription factors (TFs), from 0 to 5). We consider that 2 lines have statistically different I and/or S at a 0.1 significance level.

P-value						
	$LB_{0.75x}$		**$LB_{0.5x}$**		**$LB_{0.25x}$**	
	I	**S**	**I**	**S**	**I**	**S**
$LB_{0.75x}$	1.00	1.00				
$LB_{0.5x}$	0.32	0.56	1.00	1.00		
$LB_{0.25x}$	0.08	0.71	0.23	0.76	1.00	1.00

Table S19. Statistics of various features of the TFN topology of *E. coli*. Shown are the coefficients of determination (R^2s) and corresponding p-values of statistical significance of a linear fit between the |LFC| (absolute log2 fold change) of each gene and its corresponding topological feature (defined in (15), Methods section *Transcription Factor Network of Escherichia coli*). We reject the null hypothesis that the data is best fit by a horizontal line at 0.1 significance level.

	$LB_{0.75x}$		$LB_{0.5x}$		$LB_{0.25x}$	
Topological feature	**P-value**	**R^2**	**P-value**	**R^2**	**P-value**	**R^2**
Avg. short path length	0.72	0.00	0.42	0.00	0.50	0.00
Betweenness centrality	0.48	0.00	0.14	0.00	0.50	0.00
Stress centrality	0.41	0.00	0.13	0.00	0.15	0.00
Clustering coefficient	0.00	0.01	0.00	0.01	0.00	0.01
Eccentricity	0.54	0.00	0.65	0.00	0.79	0.00
Edge-count	0.26	0.00	0.27	0.00	0.73	0.00
Out-degree	0.55	0.00	0.39	0.00	0.19	0.00
Neighborhood connectivity	0.87	0.00	0.91	0.00	0.34	0.00

Table S20. Statistics of the mean behavior of the genes responsive to (p)ppGpp (16). Shown are the p-values from the 2-sample T-test (Methods section *Statistical tests a*) between the mean behavior of log2 fold change (LFC) of genes responsive to (p)ppGpp (16) and the LFC of all genes analyzed by RNA-seq.

	$LB_{0.25X}$	**$LB_{0.5X}$ 125 min**	**$LB_{0.5X}$ 180 min**	**$LB_{0.75X}$**	**$LB_{1.5X}$**	**$LB_{2.0X}$**	**$LB_{2.5X}$**
After 5 min	0.18	0.00	0.87	0.29	0.00	0.00	0.00
After 10 min	0.53	0.00	0.01	0.02	0.39	0.11	0.41

Table S21. Statistics of the behavior of sRNA. Shown are the p-values from a 2-sample T-test (Methods section *Statistical tests a*) between the mean log2 fold change (LFC) of sRNAs and the LFC of all genes analyzed by RNA-seq.

P-value, 2-sample T-test							
	LB$_{0.25X}$	**LB$_{0.5X}$ 125 min**	**LB$_{0.5X}$ 180 min**	**LB$_{0.75X}$**	**LB$_{1.5X}$**	**LB$_{2.0X}$**	**LB$_{2.5X}$**
sRNA (93 genes)	0.39	0.1	0.32	0.13	0.23	0.24	0.40
Number of DEG rRNAs							
sRNA	18 DEG	4 DEG	7 DEG	0	33 DEG	33 DEG	30 DEG

Table S22. Statistics of the behavior of rRNAs. Shown, for each shift, are the p-values from the 2-sample T-test (Methods section *Statistical tests a*) between the mean log2 fold change (LFC) of rRNAs and the LFC of all genes analyzed by RNA-seq. Also shown is the number of DEG genes.

<table>
<tr><th colspan="8">P-value, 2-sample T-test</th></tr>
<tr><th></th><th>$LB_{0.25X}$</th><th>$LB_{0.5X}$
125 min</th><th>$LB_{0.5X}$
180 min</th><th>$LB_{0.75X}$</th><th>$LB_{1.5X}$</th><th>$LB_{2.0X}$</th><th>$LB_{2.5X}$</th></tr>
<tr><td><b>rRNA (22 genes)</b></td><td>0.00</td><td>1.00</td><td>0.00</td><td>0.00</td><td>0.00</td><td>0.00</td><td>0.00</td></tr>
<tr><th colspan="8">Number of DEG rRNAs</th></tr>
<tr><td><b>rRNA</b></td><td>0</td><td>0</td><td>0</td><td>0</td><td>15 DEG</td><td>0</td><td>2 DEG</td></tr>
</table>

SI References

1. deHaseth,P.L., Zupancic,M.L. and Record,M.T. (1998) RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. *J Bacteriol,* 180, 3019-3025.
2. Taniguchi,Y., Choi,P.J., Li,G.W., Chen,H., Babu,M., Hearn,J., Emili,A. and Xie,X.S. (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science,* 329, 533-538.
3. Bernstein,J.A., Khodursky,A.B., Lin,P.H., Lin-Chao,S. and Cohen,S.N. (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A,* 99, 9697-9702.
4. Yu,J., Xiao,J., Ren,X., Lao,K. and Xie,X.S. (2006) Probing gene expression in live cells, one protein molecule at a time. *Science,* 311, 1600-1603.
5. Erdos,P., and Rényi,A. (1960) On the evolution of random graphs. *Publ Math Inst Hung Acad Sci,* 5, 17-60.
6. Häkkinen,A., Muthukrishnan,A.B., Mora,A., Fonseca,J.M. and Ribeiro,A.S. (2013) CellAging: a tool to study segregation and partitioning in division in cell lineages of Escherichia coli. *Bioinformatics,* 29, 1708-1709.
7. Severinov,K., Mooney,R., Darst,S.A. and Landick,R. (1997) Tethering of the large subunits of Escherichia coli RNA polymerase. *J Biol Chem,* 272, 24137-24140.
8. Buhler,J.M., Riva,M., Mann,C., Thuriaux,P., Memet,S., Micouin,J.Y., Treich,I., Mariotte,S., Sentenac,A., Reznekoff,W.S., Burgess,R.R., Dahlberg,J.E., Gross,C.A., Record Jr.,M.T., Wickens,M.P., (1987) RNA Polymerase and the Regulation of Transcription. Elsevier Science Publishing Co., New York, pp. 25-36.
9. Nishiuchi,Y., Inui,T., Nishio,H., Bódi,J., Kimura,T., Tsuji,F.I. and Sakakibara,S. (1998) Chemical synthesis of the precursor molecule of the Aequorea green fluorescent protein, subsequent folding, and development of fluorescence. *Proc Natl Acad Sci U S A*, 95, 13549-13554.
10. Tukey,J.W. (1977) Schematic summaries: Fences, and outside values. In *Exploratory data analysis*. Addison-Wesley Publishing, Reading, Massachusetts, pp. 43-44.
11. Martínez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol,* **6**, 482-489.
12. Balderas-Martínez,Y.I., Savageau,M., Salgado,H., Pérez-Rueda,E., Morett,E. and Collado-Vides,J. (2013) Transcription factors in Escherichia coli prefer the holo conformation. *PLoS One,* 8, e65723.
13. Airoldi,E.M. and Carley,K.M. (2005) Sampling algorithms for pure network topologies: a study on the stability and the separability of metric embeddings. *ACM SIGKDD Explorations Newsletter,* 7, 13-22.

14. Santos-Zavaleta,A., Salgado,H., Gama-Castro,S., Sánchez-Pérez,M., Gómez-Romero,L., Ledezma-Tejeida,D., García-Sotelo,J.S., Alquicira-Hernández,K., Muñiz-Rascado,L.J., Peña-Loredo,P.*, et al.* (2019) RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. *Nucleic Acids Res,* 47, D212-D220.
15. Doncheva,N.T., Assenov,Y., Domingues,F.S. and Albrecht,M. (2012) Topological analysis and interactive visualization of biological networks and protein structures. *Nat Protoc,* 7, 670-685.
16. Sanchez-Vazquez,P., Dewey,C.N., Kitten,N., Ross,W. and Gourse,R.L. (2019) Genome-wide effects on Escherichia coli transcription from ppGpp binding to its two sites on RNA polymerase. *Proc Natl Acad Sci U S A,* 116, 8310-8319.

PUBLICATION
III

Analytical kinetic model of native tandem promoters in *E. coli*

V. Chauhan*, M.N.M. Bahrudeen*, C.S.D. Palma, I. Baptista, B.L.B. Almeida, S. Dash, V. Kandavalli, and A.S. Ribeiro. *Equal contributions.

PLOS Computational Biology, 18(1), 2022
DOI: 10.1371/journal.pcbi.1009824

RESEARCH ARTICLE

Analytical kinetic model of native tandem promoters in *E. coli*

Vatsala Chauhan[1]☯, Mohamed N. M. Bahrudeen[1]☯, Cristina S. D. Palma[1], Ines S. C. Baptista[1], Bilena L. B. Almeida[1], Suchintak Dash[1], Vinodh Kandavalli[2], Andre S. Ribeiro[1]*

[1] Laboratory of Biosystem Dynamics, Faculty of Medicine and Health Technology, Tampere University, Finland, [2] Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden

☯ These authors contributed equally to this work.
* andre.sanchesribeiro@tuni.fi

OPEN ACCESS

Citation: Chauhan V, Bahrudeen MNM, Palma CSD, Baptista ISC, Almeida BLB, Dash S, et al. (2022) Analytical kinetic model of native tandem promoters in *E. coli*. PLoS Comput Biol 18(1): e1009824. https://doi.org/10.1371/journal.pcbi.1009824

Editor: Eli Zunder, University of Virginia, UNITED STATES

Received: August 13, 2021

Accepted: January 11, 2022

Published: January 31, 2022

Data Availability Statement: A data package was deposited in Dryad (Ref. [59] in main manuscript) under the DOI:10.5061/dryad.bnzs7h4bs. It contains the flow-cytometry and microscopy data, along with the MATLAB, R and Python codes used. The data is accessible through this link: https://datadryad.org/stash/share/CYS3FjMMhrs8aqPq4lLFsGunynao-Au0wTPvGYWS4oQ. Meanwhile, RNA-seq data is deposited in NCBI GEO with the accession code GSE183139 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183139).

Abstract

Closely spaced promoters in tandem formation are abundant in bacteria. We investigated the evolutionary conservation, biological functions, and the RNA and single-cell protein expression of genes regulated by tandem promoters in *E. coli*. We also studied the sequence (distance between transcription start sites 'd_{TSS}', pause sequences, and distances from oriC) and potential influence of the input transcription factors of these promoters. From this, we propose an analytical model of gene expression based on measured expression dynamics, where RNAP-promoter occupancy times and d_{TSS} are the key regulators of transcription interference due to TSS occlusion by RNAP at one of the promoters (when $d_{TSS} \leq 35$ bp) and RNAP occupancy of the downstream promoter (when $d_{TSS} > 35$ bp). Occlusion and downstream promoter occupancy are modeled as linear functions of occupancy time, while the influence of d_{TSS} is implemented by a continuous step function, fit to *in vivo* data on mean single-cell protein numbers of 30 natural genes controlled by tandem promoters. The best-fitting step is at 35 bp, matching the length of DNA occupied by RNAP in the open complex formation. This model accurately predicts the squared coefficient of variation and skewness of the natural single-cell protein numbers as a function of d_{TSS}. Additional predictions suggest that promoters in tandem formation can cover a wide range of transcription dynamics within realistic intervals of parameter values. By accurately capturing the dynamics of these promoters, this model can be helpful to predict the dynamics of new promoters and contribute to the expansion of the repertoire of expression dynamics available to synthetic genetic constructs.

Author summary

Tandem promoters are common in nature, but investigations on their dynamics have so far largely relied on synthetic constructs. Thus, their regulation and potentially unique dynamics remain unexplored. We first performed a comprehensive exploration of the conservation of genes regulated by these promoters in *E. coli* and the properties of their input transcription factors. We then measured protein and RNA levels expressed by 30

Funding: This work was supported by the Jane and Aatos Erkko Foundation (10-10524-38) [ASR]; Finnish Cultural Foundation (50201300 to [SD] and 00200193 to [ISCB]); Suomalainen Tiedeakatemia [CSDP]; Tampere University Graduate Program [VC, MNMB, BLBA]; and EDUFI Fellowship (TM-19-11105) [SD]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Escherichia coli tandem promoters, to establish an analytical model of the expression dynamics of genes controlled by such promoters. We show that start site occlusion and downstream RNAP occupancy can be realistically captured by a model with RNAP binding affinity, the time length of open complex formation, and the nucleotide distance between transcription start sites. This study contributes to a better understanding of the unique dynamics tandem promoters can bring to the dynamics of gene networks and will assist in their use in synthetic genetic circuits.

Introduction

Closely spaced promoters exist in all branches of life in convergent, divergent, and tandem formations [1–7]. Models of tandem promoters [8–10] have largely been based on measurements of synthetic constructs [11–13] and predict that such promoter arrangements result in unique transcription dynamics due to the interference between RNAPs transcribing the promoters [9,10,14–19].

When an RNAP is committed to form the open complex (OC), a process lasting up to hundreds of seconds [20–22], it occupies approximately 35 base pairs (bp), from the transcription start site (TSS, position 0) until position -35 [23–25]. If the TSS of a neighbouring promoter is closer than 35 bp it will not be possible for both promoters to be occupied simultaneously, since an RNAP occupying one of them will 'occlude' the other, preventing it from being reached [9]. However, if the promoters are more than 35 bp apart, this occlusion does not occur. Instead, interference will occur when RNAPs elongating from the upstream promoter collide with an RNAP occupying the downstream promoter [14] (in either closed or open complex formation), forcing one of the RNAPs to fall-off (both scenarios are likely possible, and we expect it to differ with, e.g., the binding affinity of the RNAP to the downstream promoter). Meanwhile, models based on empirical parameter values suggest that collisions between two elongating RNAPs are rare (because events such as pausing or simultaneous initiations from both promoters are rare). Also, even if and when such collisions occur, they are unlikely to result in fall-offs since the RNAPs are moving at similar speeds and in the same direction [9,10,26].

Models suggest that both forms of interference decrease the mean RNA production rate while increasing its noise based on the distance between promoters (d_{TSS}), their strengths [10], and the time spent between commitment of the RNAP to OC and escape from the promoter region [27]. These hypotheses have yet to be empirically validated in natural tandem promoters.

We studied how d_{TSS} and the time spent by RNAPs on the TSSs affect gene expression dynamics due to interference between the transcription processes of tandem promoters (Fig 1). We consider only the natural tandem promoters that neither overlap with nor have in between another gene (positionings I and II, which differ in if the promoter regions overlap or not) (see the other arrangements in Fig A in the S2 Appendix). The numbers of these arrangements in *E. coli* are shown in Table H in the S3 Appendix. From the measurements of these genes' protein levels, we then establish a model that we use to explore the state space of potential dynamics under the control of tandem promoters (Fig 2 illustrates our workflow).

Results

E. coli has 831 genes controlled by two or more promoters in tandem formation (RegulonDB and section 'Selection of natural genes controlled by tandem promoters'

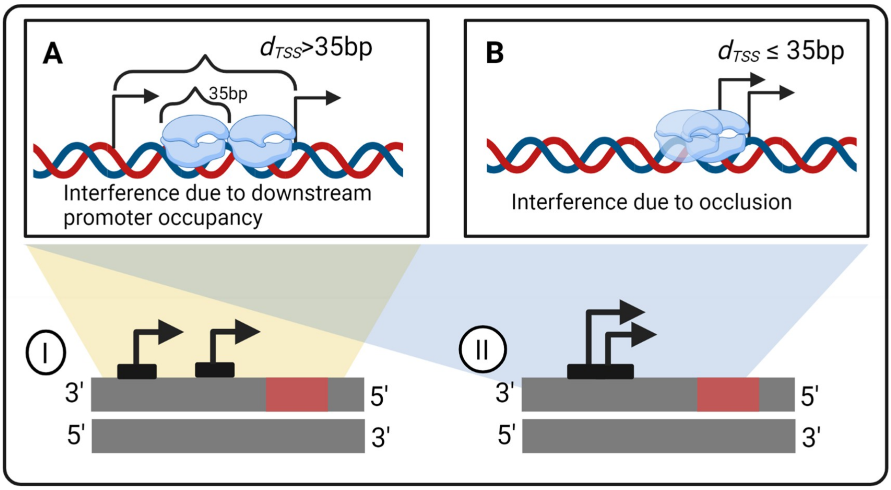

A
d_{TSS}>35bp
35bp
Interference due to downstream promoter occupancy
B
d_{TSS} ≤ 35bp
Interference due to occlusion
I
3'
5'
5'
3'
II
3'
5'
5'
3'

Fig 1. Interference between tandem promoters with different arrangements relative to each other. **(A)** Interference by an RNAP occupying the downstream promoter on the activity of the elongating RNAP from upstream promoter. The TSSs need to be at least 36 bp apart (the length occupied by an RNAP when in OC, [23,25]) **(B)** Interference by occlusion of one of the promoter's TSS by an RNAP on the TSS of the other promoter. The distance between the TSSs need to be ≲ 35 bp apart. Blue clouds are RNAPs. Black arrows sit on TSSs and point towards the direction of transcription elongation. Arrangements **(I-II)** of two promoters studied in the manuscript in tandem formation are represented. The red rectangles are the protein coding regions. We studied only the natural tandem promoters that neither overlap with nor have in between another gene (arrangements I and II, which differ based on whether the promoter regions overlap or not). Other arrangements (not considered in this study) are shown in Fig A in the S2 Appendix. Figure created with BioRender.com.

https://doi.org/10.1371/journal.pcbi.1009824.g001

in the S1 Appendix). However, to study the dynamics of genes controlled by tandem promoters, we focused on only 102 of them, because their activity is expected to be undisturbed by neighboring genes in the DNA (arrangements I and II in Fig 1), for reasons

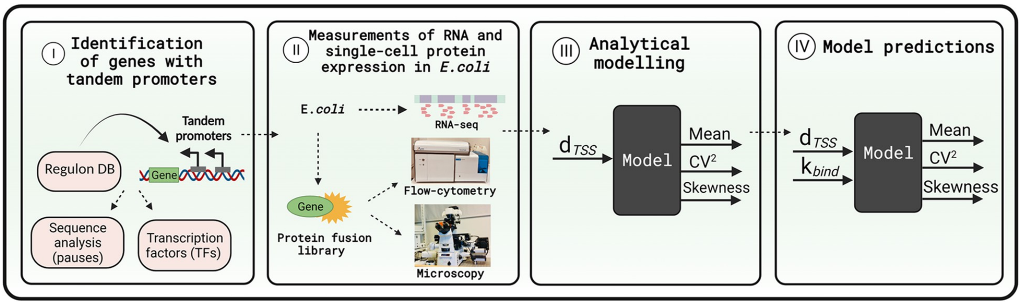


Fig 2. Workflow. (I) We identified genes controlled by tandem promoters in Regulon DB. (II) Next, we measured the single-cell protein levels of those genes with arrangements I and II that are tagged in the YFP strain library [28]. We also measured the mean RNA fold changes of these genes over time (S1 Appendix, section 'RNA-seq measurements and data analysis'). (III) We used the single-cell data to tune the model. (IV) Finally, we used the model to explore the state space of protein expression. Figure created with BioRender.com.

https://doi.org/10.1371/journal.pcbi.1009824.g002

described in section 'Selection of natural genes controlled by tandem promoters' in the S1 Appendix.

Further, these promoters do not have specific short nucleotide sequences capable of affecting RNAP elongation (section 'Pause sequences' in the S4 Appendix). Also, the 102 genes expressed by these promoters are not overrepresented in a particular biological process (section 'Over-representation test' in the S4 Appendix). From time-lapse RNA-seq data (S1 Appendix, section 'RNA-seq measurements and data analysis'), we also did not find evidence that their dynamics are affected by their input transcription factors (TFs) in our measurement conditions (section 'Input-output transcription factor relationships' in the S4 Appendix) nor by H-NS in a consistent manner (section 'Regulation by H-NS' in the S4 Appendix). Finally, they do not exhibit any particular TF network features (Table C in the S3 Appendix). As such, neither input TFs nor specific nucleotide sequences are considered in the model below. In addition to all of the above, we found no correlations between the shortest distance from the TSS of upstream promoters from the oriC region in the DNA and expression levels (section 'Relationship with the oriC region' in the S4 Appendix).

Model of gene expression controlled by tandem promoters

RNAPs bind, slide along, and unbind from a promoter several times until, eventually, one of them finds the TSS [29–30], commits to OC at the TSS, and initiates transcription elongation.

Reactions (1A1) are a 4-step (I-IV) model of transcription [20,31]. The forward reaction in step I in (1A1) models RNAP binding to a free promoter (P_{free}), which becomes no longer free albeit the RNAP might not yet have reached the TSS. This state, pre-finding of the TSS, is here named P_{bound} and its occurrence increases with RNAP concentration, *[R]*. Next, as it percolates the DNA, the RNAP should find and stop at the nearest TSS and form a closed complex (CC) with the DNA (step II, Reaction 1A1). CCs are unstable, i.e. reversible [22] (reaction 1A2) but, eventually, one of them will commit to OC irreversibly [32], via step III, Reaction 1A1 [21–22]. It follows RNAP escape from the TSS, freeing the promoter (step IV, Reaction 1A1) [33–37]. Then, the RNAP elongates (R_{elong}) until producing a complete RNA (reaction 1A3) and freeing itself.

These set of reactions usually model well stochastic transcription dynamics [20]. However, if two promoters are closely spaced in tandem formation, they can interfere [38]. Fig 3 shows sequences of events that can lead to interference between tandem promoters, not accounted for by the model above.

From Fig 3, if the TSSs are sufficiently close, the occupancy of one TSS by an RNAP will occlude the other TSS, blocking its kinetics [18]. This is accounted for by reaction 1A5, which competes with CC formation in reaction 1a1. Its rate constant, $k_{occlusion}$, is defined in the next section. In (1A5), 'u/d' stands for occlusion of the upstream promoter by an RNAP on the TSS of the downstream promoter.

Instead, if the TSSs are not sufficiently close, they will still interfere since the elongating RNAP (R_{elong}) starting from the upstream promoter can collide with RNAPs on the TSS of the downstream promoter. This can dislodge either RNAP via (reaction 1A4) or (reaction 2A3), depending on the sequence-dependent binding strength of the RNAP to the TSS [9].

Finally, once reaction 1A1 occurs, either reaction 1A3 or 1A4 occur. To tune their competition, we introduced the terms ω_d and (1- ω_d) in their rate constants, with ω_d being the fraction of times that an elongating RNAP from an upstream promoter finds an RNAP occupying the downstream promoter. Meanwhile, '*f* is the fraction of times that the RNAP occupying the downstream promoter falls-off due to the collision with an elongating RNAP, whereas '*1-f*' is

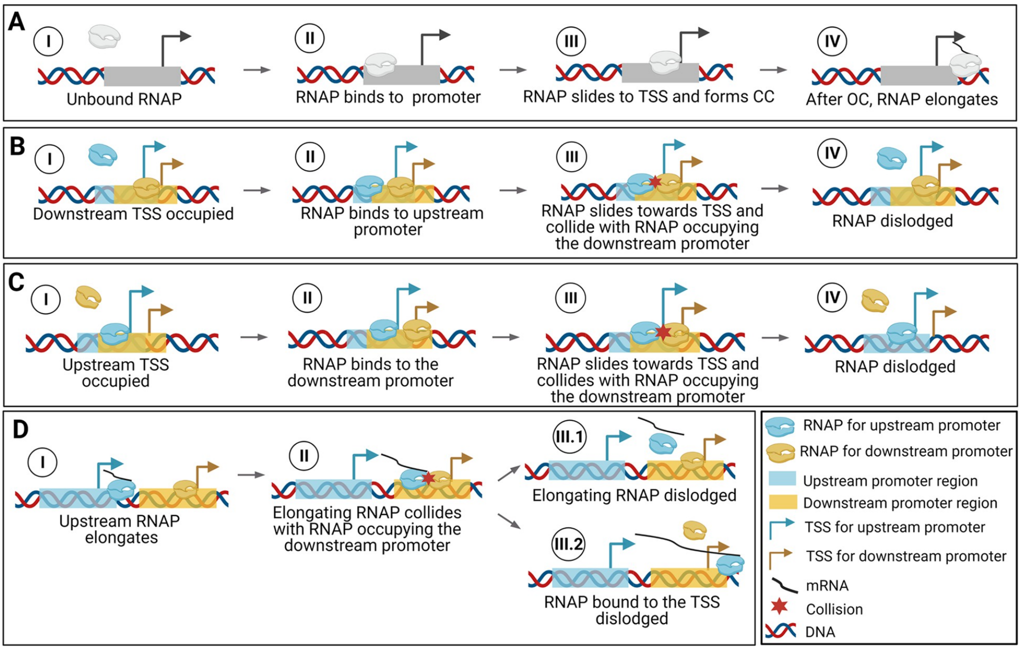

A
I
Unbound RNAP
II
RNAP binds to promoter
III
RNAP slides to TSS and forms CC
IV
After OC, RNAP elongates
B
I
Downstream TSS occupied
II
RNAP binds to upstream promoter
III
RNAP slides towards TSS and collide with RNAP occupying the downstream promoter
IV
RNAP dislodged
C
I
Upstream TSS occupied
II
RNAP binds to the downstream promoter
III
RNAP slides towards TSS and collides with RNAP occupying the downstream promoter
IV
RNAP dislodged
D
I
Upstream RNAP elongates
II
Elongating RNAP collides with RNAP occupying the downstream promoter
III.1
Elongating RNAP dislodged
III.2
RNAP bound to the TSS dislodged
RNAP for upstream promoter
RNAP for downstream promoter
Upstream promoter region
Downstream promoter region
TSS for upstream promoter
TSS for downstream promoter
mRNA
Collision
DNA

Fig 3. Events leading to transcriptional interference between tandem promoters. (**A**) Sequence of events in transcription in isolated promoters. A similar set of events occurs in tandem promoters, if only one RNAP interacts with them at any given time. (**B / C**) Interference due to the occlusion of the *downstream* / *upstream* promoter by a bound RNAP, which will impede the incoming RNAP from binding to the TSS. (**D**) Interference of the activity of the RNAP incoming from the upstream promoter by the RNAP occupying the downstream promoter. One of these RNAPs will be dislodged by the collision. Created with BioRender.com.

https://doi.org/10.1371/journal.pcbi.1009824.g003

the fraction of times that it is the elongating RNAP that falls-off.

$$P^u_{free} \xrightarrow[\text{I}]{k^u_{bind}\cdot[R]} P^u_{bound} \xrightarrow[\text{II}]{k^u_{cc}} P^u_{cc} \xrightarrow[\text{III}]{k^u_{oc}} P^u_{oc} \xrightarrow[\text{IV}]{k^u_{escape}} P^u_{free} + R^u_{elong} \tag{1A1}$$

$$P^u_{cc} \xrightarrow{k_{unbind}} P^u_{free} \tag{1A2}$$

$$R^u_{elong} \xrightarrow{k^u_{elong}\cdot(1-\omega_d\cdot f)} RNA \tag{1A3}$$

$$R^u_{elong} \xrightarrow{k^u_{elong}\cdot\omega_d\cdot f} \emptyset \tag{1A4}$$

$$P^u_{bound} \xrightarrow{k^{u/d}_{occlusion}} P^u_{free} \tag{1A5}$$

Next, we reduced the model and derived its analytical solution. First, since P_{cc} completion is expected to be faster than P_{bound} completion ([10] and references within) we merged them into a single state, $P_{occupied}$, which represents a promoter occupied by an RNAP prior to commitment to OC, whose time length is similar to P_{bound}.

Similarly, in standard growth conditions, the occurrence of multiple failures in escaping the promoter per OC completion should only occur in promoters with the highest binding affinity to RNAP. Thus, in general promoter escape should be faster than OC [20,32]. We thus merged OC and promoter escape into one step named 'events *after* commitment to OC', with a rate constant k_{after}. The simplified model is thus:

$$P^u_{free} \xrightarrow{k^u_{bind} \cdot [R]} P^u_{occupied} \xrightarrow{k^u_{after}} P^u_{free} + R^u_{elong} \tag{1B1}$$

These two steps are not merged since only the first differs with RNAP concentration [20,26,39]. Further, reports [40–41] indicate that *E. coli* has ~100–1000 RNAPs free for binding at any moment but ~4000 genes, suggesting that the number of free RNAPs is a limiting factor.

Finally, we merge (1A2), (1A5) and (1B1) in one multistep without affecting the model kinetics:

$$P^u_{free} \underset{k^{u/d}_{occlusion} + k^u_{unbind}}{\overset{k^u_{bind} \cdot [R]}{\rightleftharpoons}} P^u_{occupied} \xrightarrow{k^u_{after}} P^u_{free} + R^u_{elong} \tag{1C1}$$

Overall, this reduced model of transcription of upstream promoters has a multistep reaction of transcription initiation (1C1), a reaction of transcription elongation (1A3) and a reaction for failed elongation due to RNAPs occupying the downstream promoter (1A4).

Regarding RNA production from the downstream promoter, it should either be affected by occlusion if $d_{TSS} \leq 35$, or by RNAPs elongating from the upstream promoter if $d_{TSS} > 35$ (Fig 3). We thus use reactions (2A1), (2A2), and (2A3) to model these promoters' kinetics:

$$P^d_{free} \underset{k^{d/u}_{occlusion} + k^d_{unbind}}{\overset{k^d_{bind} \cdot [R]}{\rightleftharpoons}} P^d_{occupied} \xrightarrow{k^d_{after}} P^d_{free} + R^d_{elong} \tag{2A1}$$

$$R^d_{elong} \xrightarrow{k^d_{elong}} RNA \tag{2A2}$$

$$P^d_{free} \xrightarrow{k_{occupy}} P^d_{occupied} \tag{2A3}$$

Finally, one needs to include a reaction for translation (reaction 3), as a first order process since protein numbers follow RNA numbers linearly (Fig F in the S2 Appendix), and reactions for RNA and protein decay accounting for degradation and for dilution due to cell division (reactions 4A and 4B, respectively). TF regulation is not included as noted above (Fig C and panel A of Fig D in the S2 Appendix).

$$RNA \xrightarrow{k_p} Prot \tag{3}$$

$$RNA \xrightarrow{k_{rd}} \emptyset \tag{4A}$$

$$Prot \xrightarrow{k_{pd}} \emptyset \tag{4B}$$

Transcription interference by occlusion

In a pair of tandem promoters, the $k_{occlusion}$ of one of them should increase with the fraction of time that the other one is occupied. Further, it should decrease with increasing d_{TSS} between the two promoters' TSS. We thus define $k_{occlusion}$ for the upstream (Eq 5A) and downstream (Eq 5B) promoters, respectively as:

$$k^{u/d}_{occlusion} = k^{\max}_{ocl} \cdot I(d_{TSS}) \cdot \omega_d \tag{5A}$$

$$k^{d/u}_{occlusion} = k^{\max}_{ocl} \cdot I(d_{TSS}) \cdot \omega_u \tag{5B}$$

Here, $k^{\max}_{ocl}$ is the maximum occlusion possible. It occurs when the two TSSs completely overlap each other (*d_{TSS} = 0*) *and* the TSS of the 'other' promoter is always occupied. Meanwhile, *I(d_{TSS})* models distance-dependent interference.

We tested four models of interference: 'exponential 1', 'exponential 2', 'step', and 'zero order' (Table 1). The first two assume that the effects of occlusion decrease exponentially with d_{TSS} (first and second order dependency, respectively).

Meanwhile, the 'Step' model assumes that interference only occurs precisely in the region in the DNA occupied by the RNAP when in OC formation. For this, it uses a logistic equation to build a continuous step function, where *L* is the length of DNA (in bp) occupied by the RNAP in OC. As such, L tunes at what d_{TSS} the step occurs, while *m* is the steepness of that step (set to 1 bp^{-1}).

Finally, the 'Zero order' model assumes (unrealistically) that interference by occlusion, is independent of $d_{TSS.}$ Fig G in the S2 Appendix shows how $k_{occlusion}$ differs with d_{TSS} in each model, for various parameter values.

Finally, *ω* is the fraction of time that the 'other' promoter is occupied. It ranges from 0 (no occupancy) to 1 (always occupied). It is estimated for upstream and downstream promoters as:

$$\omega_u = \frac{k^u_{bind} \cdot [R]}{k^u_{unbind} + k^u_{bind} \cdot [R] + k^u_{after}} \tag{6A}$$

$$\omega_d = \frac{k^d_{bind} \cdot [R]}{k^d_{unbind} + k^d_{bind} \cdot [R] + k^d_{after}} \tag{6B}$$

Similarly, if $k^{\max}_{occupy}$ is the maximum possible interference due to RNAPs occupying the downstream promoter, k_{occupy} is defined as:

$$k_{occupy} = \omega_u \cdot k_{after} \cdot k^{\max}_{occupy} \cdot (1 - f) \tag{7}$$

Table 1. Potential models of transcriptional interference due to promoter occlusion considered.

Interference by occlusion	*I(d_{TSS})*	$k_{occlusion}$
Exponential 1 ("Exp1")	$e^{-(b_1 \cdot d_{TSS})}$	$k^{\max}_{ocl} \cdot e^{-(b_1 \cdot d_{TSS})} \cdot \omega$
Exponential 2 ("Exp2")	$e^{-(b_1 \cdot d_{TSS} + b_2 \cdot d^2_{TSS})}$	$k^{\max}_{ocl} \cdot e^{-(b_1 \cdot d_{TSS} + b_2 \cdot d^2_{TSS})} \cdot \omega$
Step ("Step")	$1 - \frac{1}{1+e^{-m \cdot (d_{TSS} - L)}}$	$k^{\max}_{ocl} \cdot \left(1 - \frac{1}{1+e^{-(d_{TSS} - L)}}\right) \cdot \omega$, for m = 1 bp^{-1}
Zero order ("ZeroO")	k	$k^{\max}_{ocl} \cdot \omega$

https://doi.org/10.1371/journal.pcbi.1009824.t001

Analytical solution of the moments of the single-cell protein numbers

Next, we derived an analytical solution of the expected mean single-cell protein numbers at steady state, M_P, which is later tuned to fit the empirical data. For any gene, regardless of the underlying kinetics of transcription, k_r is the *effective* rate of RNA production. Based on the reactions above, the mean protein numbers in steady state will be (see sections "Analytical model of mean RNA levels controlled by a single promoter in the absence of a closely spaced promoter" and "Derivation of mean protein numbers at steady state produced by a pair of tandem promoters" in the S1 Appendix):

$$M_P = \frac{k_r \cdot k_p}{k_{rd} \cdot k_{pd}} \tag{8}$$

This equation applies to a pair of tandem promoters as well. In that case, assuming that k_{bind} of the two tandem promoters is similar, we have:

$$k_r = \left(\begin{array}{c} \dfrac{k_{bind} \cdot [R] \times k_{after} \cdot (1 - \omega_d \cdot f)}{k_{occlusion} + k_{bind} \cdot [R] + k_{unbind} + k_{after}} + \\ \dfrac{k_{bind} \cdot [R] \times k_{after}}{k_{occlusion} + k_{occupy} + k_{bind} \cdot [R] + k_{unbind} + k_{after}} \end{array} \right) \tag{9}$$

To derive the other moments, we considered that empirical single-cell protein numbers in *E. coli* are well fit by negative binomials [28]. Consequently, M_p and the squared coefficient of variation CV_P^2, should be related as (Equations S28 to S38 in the S1 Appendix):

$$\log_{10}\left(CV_P^2\right) = \log_{10}(C_1) - \log_{10}(M_P), \qquad \text{with} \qquad C_1 = \frac{k_p}{k_{pd} + k_{rd}} \tag{10}$$

This relationship matches empirical data at the genome wide level, except for genes with high transcription rates [42]. Additionally, we further derived a relationship (Section 'CV^2 and Skewness of single-cell protein expression of a model tandem promoters' in the S1 Appendix) between M_P and the skewness, S_P, of the single-cell distribution of protein numbers:

$$\log_{10}(S_P) = \log_{10}(C_2) - \frac{1}{2} \cdot \log_{10}(M_P), \qquad \text{with} \qquad C_2 = 2\sqrt{C_1} - \frac{1}{\sqrt{C_1}} \tag{11}$$

Single-cell distributions of protein numbers

To validate the model, we measured by flow-cytometry the single-cell distributions of protein fluorescence of 30 out of the 102 genes known to be controlled by tandem promoters (with arrangements I and II). Measurements were made in 1X and 0.5X media (3 replicates per condition) using cells from the YFP strain library (section 'Strains and Growth Conditions' in the S1 Appendix). Data from past studies show that, in these 30 genes, RNA and protein numbers are well correlated (Fig F in the S2 Appendix) in standard growth conditions. Past studies also suggest that most of these genes are active during exponential growth (~95% of our 30 genes selected should be active, according to data in [43] using SEnd-seq technology).

Single-cell distributions of protein expression levels are shown in Fig 4A for one of these genes as an example. The raw data from all 30 genes (only one replicate) are shown in Fig H in the S2 Appendix. Finally, the mean, CV^2 and skewness for each gene, obtained from the triplicates, are shown in Excel sheets 1 and 2 in the S2 Table. In addition, we also show this mean, CV^2 and skewness after subtracting the first, second, and third moments of the single-cell distribution of the fluorescence of control cells, which do not express YFP (Sheets 3, 4 in the S2

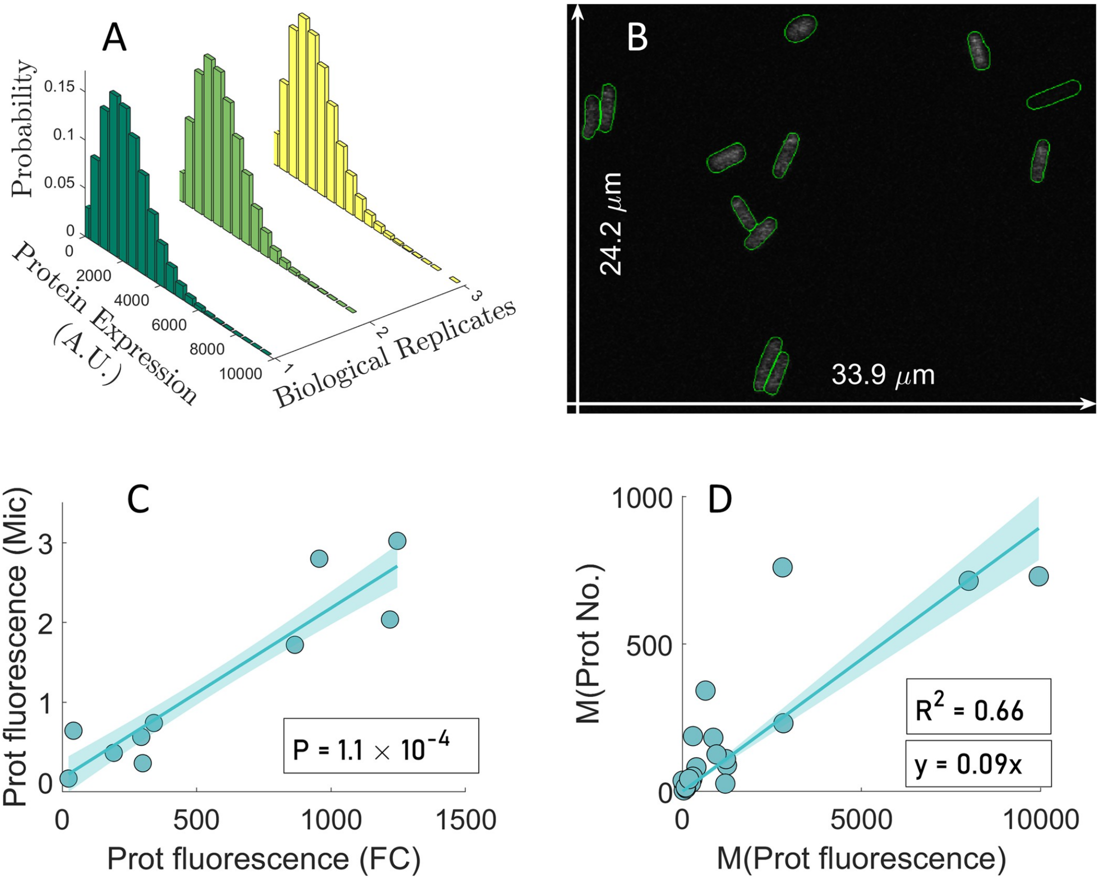


Fig 4. Single cell protein numbers by microscopy and flow-cytometry. (A) Example single-cell distributions (3 biological replicates) of fluorescence (in arbitrary units) of cells with a YFP tagged gene controlled by a pair of tandem promoters obtained by flow-cytometry, 'FC'. (B) Example confocal microscopy image of cells overlapped by the results of cell segmentation from the corresponding phase contrast image. The two white arrows show the dimensions of the image, for scaling purposes. (C) Mean single-cell protein fluorescence of 10 genes (Table G in the S3 Appendix) when obtained by FC plotted against when obtained by microscopy, 'Mic'. (D) Mean single-cell protein fluorescence (own measurements) plotted against the corresponding mean single-cell protein numbers reported in [28]. From the equation of the best fitting line without y-intercept (y-intercept = 0), we obtained a scaling factor, *sf*, equal to 0.09.

https://doi.org/10.1371/journal.pcbi.1009824.g004

Table) (Section 'Subtraction of background fluorescence from the total protein fluorescence' in flow-cytometry in the S1 Appendix).

Based on the analysis of the data of these 30 genes, we removed from subsequent analysis those genes (5 in 1X and 14 in 0.5X) whose mean, variance, or third moment of their protein fluorescence distributions are lower than in control cells (not expressing YFP), i.e., than cellular autofluorescence (Sheets 3, 4 in S2 Table). As such, only one gene studied here (in condition 1X alone) codes for a protein that is associated to membrane-related processes, which might affect its quantification (section 'Proteins with membrane-related positionings' in S4 Appendix). As such, we do not expect this phenomenon to influence our results significantly. The data from these genes removed from further analysis is shown in Fig F in S2 Appendix alone, for illustrative purposes.

We started by testing the accuracy of the background-subtracted flow-cytometry data by confronting it with microscopy data (also after background subtraction, see section 'Microscopy and Image Analysis' in the S1 Appendix). We collected microscopy data on 10 out of the 30 genes (Table G in the S3 Appendix). The microscopy measurements of the mean single-cell fluorescence expressed by these genes (example image in Fig 4B), were consistent, statistically, with the corresponding data obtained by flow-cytometry (Fig 4C).

Next, we converted the fluorescence distributions from flow-cytometry (25 genes in 1X and 16 genes in 0.5X) into protein number distributions. In Fig 4D we plotted our measurements of mean protein fluorescence in 1X against the protein numbers reported in [28] for the same genes, in order to obtain a scaling factor (sf = 0.09). Using sf, we estimated M_P, CV_P^2, and S_P of the distribution of protein numbers expressed by the tandem promoters in (Sheets 5, 6 in S2 Table) (Section 'Conversion of protein fluorescence to protein numbers' in S1 Appendix).

To test the robustness of the estimation of the scaling factor, we also estimated a scaling factor from 10 other genes present in the YFP strain library [28] (listed in Table B in S3 Appendix). These genes were selected as described in the section 'Selection of natural genes controlled by single promoters' in S1 Appendix. Using the data from this new gene cohort (Panel A of Fig I in S2 Appendix) reported in S3 Table, we estimated a scaling factor of 0.08, supporting the previous result. Meanwhile, since when merging the data from tandem and single promoters, the resulting scaling factor equals 0.09 (Panel B of Fig I in S2 Appendix), we opted for using 0.09 from here onwards.

We also tested how sensitive the estimated scaling factor is to the removal of data points. Specifically, for 1000 times, we discarded N randomly selected data points, and estimated the resulting scaling factor. We then compared, for each N, the mean and the median of the distribution of 1000 scaling factors (Fig J in S2 Appendix). Since the median is not sensitive to outliers, if mean and median are similar, one can conclude that the scaling factor is not biased by a few data points. Visibly, the mean and the median only start differing for N larger than 6, which corresponds to nearly 30% of the data.

Log-log relationship between the mean single-cell protein numbers of tandem promoters and the other moments

We plotted M_P against CV_P^2 and S_P in log-log plots, in search for the fitting parameters, 'C_1' and 'C_2', to estimate the rate of protein production per RNA (Eq 10). To increase the state space covered by our measurements, in addition to M9 media (named '1X'), we also used diluted M9 media (named '0.5X'), known to cause cells to have lower RNAP concentrations (Fig 5A) (Section 'Strains and growth conditions' in the S1 Appendix), without altering the division rate (Panels A and B of Fig K in the S2 Appendix). We note that 1X and 0.5X only refer to the degree of dilution of the original media and not to how much RNAP concentration and consequently, protein concentrations, were reduced by media dilution. From the same figures, we attempted stronger dilutions, but no further decreases in RNAP concentration were observed and the growth rate decreased.

Next, from Fig 5B, most genes (of those expressing tangibly in both media) suffered similar reductions (well fit by a line) in protein numbers with the media dilution, as expected by the model of gene expression (Eqs 8 and 9). This linear relationship could also be interpreted as evidence that the difference in expression of these genes between the two conditions is not affected by TFs in our measurement conditions. Namely, if TF influences existed, and TF numbers changed, they would likely be diversely affected by their output genes (weakly and strongly activated, repressed, etc.) and, thus, our proteins of interest would not have changed in such similar manners (linearly).

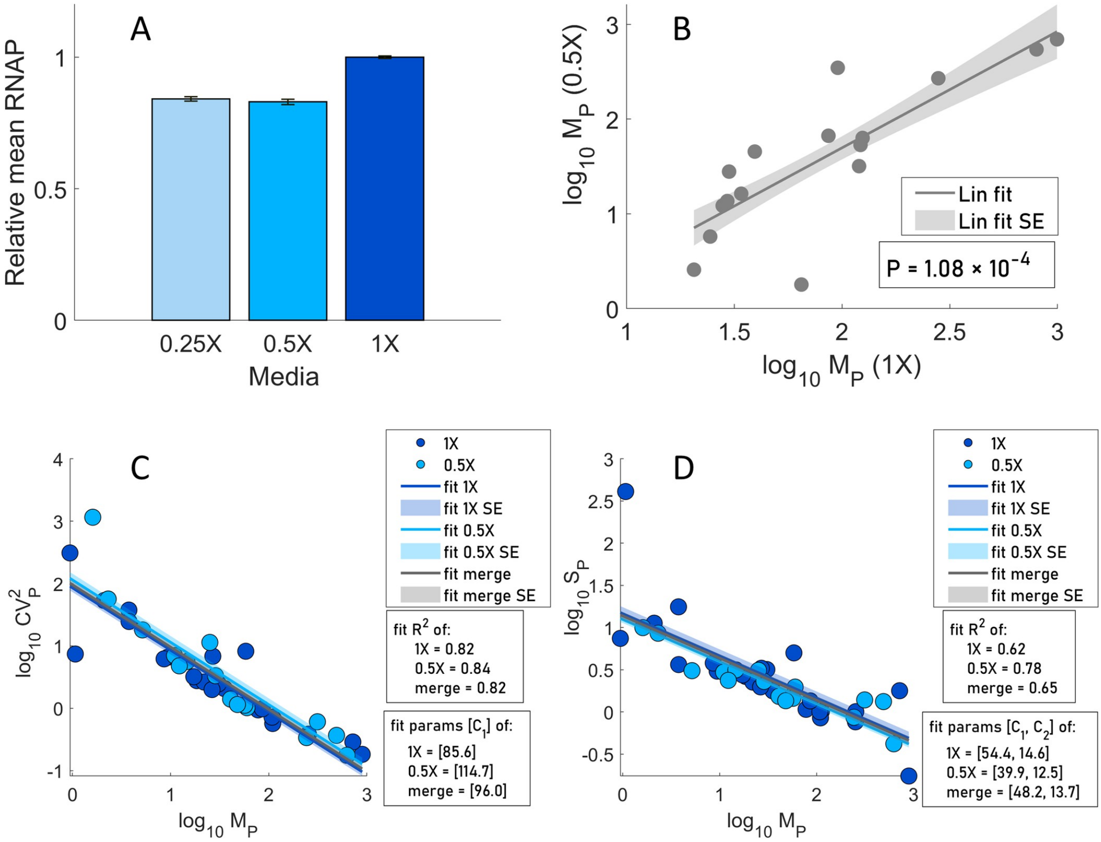


Fig 5. Relative RNAP concentrations along with the relationships between the moments of the single cell distributions of protein numbers. (A) Relative RNAP levels measured by flow-cytometry (Section 'flow-cytometry and data analysis' in the S1 Appendix) in three media. (B) Scatter plot between M_P in M9 (1X) and diluted M9 (0.5X) media. Also shown are the best fitting line and standard error and p-value for the null hypothesis that the slope is zero. (C) M_P vs CV_P^2 and (D) M_P vs S_P of single-cell protein numbers of genes with tandem promoters in M9 (1X) and M9 diluted (0.5X) media. The lines and their shades are the best fitting lines and standard errors, respectively. 'Merge' stands for data from both 0.5X and 1X conditions.

https://doi.org/10.1371/journal.pcbi.1009824.g005

Meanwhile, as in [42,44], CV_P^2 decreases linearly with M_P (log-log scale), irrespective of media ($R^2 > 0.8$ in all fitted lines), in agreement with the model (Fig 5C). Fitting Eq 10 to the data, we extracted C_1 in each condition. S_P also decreases linearly with M_P, irrespective of the media (Fig 5D). Similar to above, Eq 11 was fitted to each data set and C_1 and C_2 were obtained ($R^2 > 0.6$ for all lines).

Since C_1 from Fig 5C and 5D differed slightly (likely due to noise), we instead obtained C_1 and C_2 values that maximized the mean R^2 of both plots. Using 'fminsearch' function in MATLAB [45], we obtained $C_1 = 72.71$ and $C_2 = 16.94$ (R^2 of 0.80 and 0.61, respectively) for Fig 5C and Fig 5D, respectively.

Inference of parameter values and model predictions as a function of d_{TSS}

We next used the model, after fitting, to predict how d_{TSS} and the promoters' occupancy regulate the moments of the single-cell distribution of protein numbers (M_P, CV_P^2, and S_P) under

Table 2. Parameter values imposed identically on all models.

Parameter description	Parameter	Value	References
Inverse of the mean time to complete OC	k_{after}	0.005 s^{-1}	Differs between promoters. Since empirical data lacks, we used the data from *in vivo* single RNA measures for Lac-Ara-1 [20].
RNA and protein dilution due to division	$k_{dil} = \frac{\ln(2)}{D}$	1.005× 10^{-4} s^{-1}	Legend of Fig H in the S2 Appendix.
RNA degradation	k_{rdeg}	2.3 × 10^{-3} s^{-1}	[28]
RNA decay due to dilution from cell division and due to degradation	$k_{rd} = k_{rdeg} + k_{dil}$	2.4 × 10^{-3} s^{-1}	From row 2.
Protein degradation	k_{pdeg}	2.93 × 10^{-5} s^{-1}	[46], estimates it to be from ~6×10^{-5} to ~2×10^{-5}. We used the value in [47], in that interval.
Protein decay due to dilution by cell division and degradation	$k_{pd} = k_{pdeg} + k_{dil}$	1.3 × 10^{-4} s^{-1}	From rows 2 and 5.
Fall-off probability of the RNAP occupying the downstream promoter	f	50% (0.5)	Set here (likely sequence-dependent)
Protein production rate constant	$k_p = C_1 \times (k_{pd}+k_{rd})$	0.18 s^{-1}	C_1 is estimated here.
Free RNAP per cell	*[R]*	144/cell in 1X and 120/cell in 0.5X media	See main text.

https://doi.org/10.1371/journal.pcbi.1009824.t002

the control of tandem promoters. We started by assuming the parameter values from the literature listed in Table 2 and tuned the remaining parameters.

To set the RNAP numbers in Table 2, we considered that the RNAPs affecting transcription rates are the *free* RNAPs in the cell, and that, for doubling times of 30 min in rich medium, there are ~1000 free RNAPs per cell [41]. Meanwhile, for doubling times of 60 min in minimal medium, there are ~144 [40]. In both our media, we observed a doubling time of ~115 mins (Fig 5B). Thus, we expect the free RNAP in 1X to also be ~144/cell or lower. Meanwhile, in 0.5X, we measured the RNAP concentration to be 17% lower than in 1X (Fig 5A) and no morphological changes. Thus, we assume the free RNAP in 0.5X to equal ~120/cell.

Next, we fitted the Eqs (8) and (9) relating d_{TSS} with $\log_{10}(M_P)$ in all interference models (Table 1), using the data on M_P in 1X medium (Fig 6A) and the 'fit' function of MATLAB. For this, we set $k^{\max} = k^{\max}_{occupy} = k^{\max}_{ocl}$, for simplicity, as well as realistic bounds for each parameter to infer. To avoid local minima, we performed 200 searches, each starting from a random initial point, and selected the one that maximized R^2. Results are shown in Table 3.

Next, we inserted all parameter values (empirical and inferred) in Eqs (10) and (11) to predict CV^2_P and S_P in 1X medium (Fig 6B and 6C). Also, we inserted the same parameter values and the estimated RNAP numbers in 0.5X medium in Eqs (8–11) to obtain the analytical solutions for M_P, CV^2_P and S_P for 0.5X medium (Fig 6D,6E and 6F).

From Fig 6, the data is 'noisy', which suggests that it is not possible to establish if the models are significantly different. As such, here we only select the one that best explains the data, based on the R^2 values of the fittings. Table 3 shows the mean R^2 for M_P, CV^2_P, and S_P when confronting the model with the data. Overall, from the R^2 values, the step model is the one that best fits the data. Meanwhile, the 'ZeroO' model is the least accurate, which supports the existence of distinct kinetics when d_{TSS} is smaller or larger than 35 nucleotides, which is the length of the RNAP when committed to OC on the TSS [23–25].

In summary, the proposed model of expression of genes under the control of a pair of tandem promoters is based on a standard model of transcription of each promoter, which are subject to interference, either due to occlusion of the TSSs or by RNAP occupying the TSS of the downstream promoter. The influence of each occurrence of these events is well modeled by linear functions of TSS occupancy times, while their dependency on d_{TSS} is modeled by a

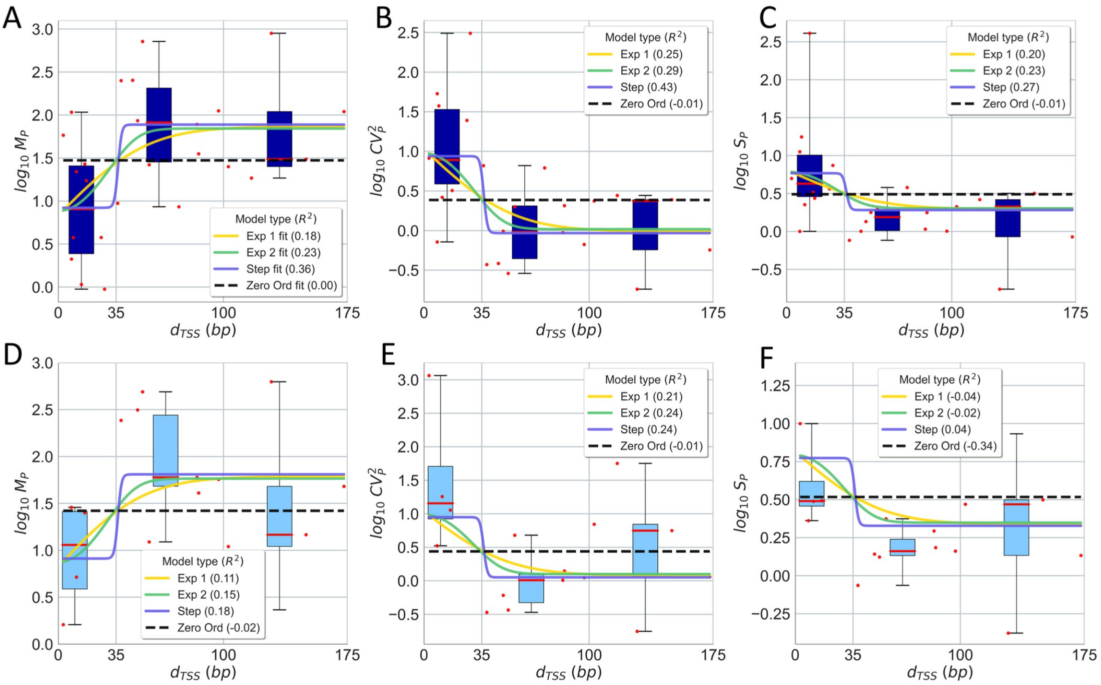


Fig 6. Empirical data and analytical model of how d_{TSS} influences the single-cell protein numbers of genes controlled by tandem promoters. **(A)** Mean, **(B)** CV^2, and **(C)** S of single protein numbers in the **1X** media as a function of d_{TSS}. **(D)**, **(E)**, and **(F)** show the same for the **0.5X** media, respectively. Each red dot is the mean from 3 biological repeats for a pair of promoters (S2 Table). The dots were also grouped in 3 'boxes' based on their d_{TSS}. In each box, the red line is the median and the top and bottom are the 3rd and 1st quartiles, respectively. The vertical black bars are the range between minimum and maximum of the red dots. In **A**, all lines are best fits. In **B**, **C**, **D**, **E**, and **F**, all lines are model predictions, based on the parameters used to best fit **A**. The insets show the R^2 for each model fit and prediction.

https://doi.org/10.1371/journal.pcbi.1009824.g006

continuous step function. If d_{TSS} is larger than 35 bp, effects from the RNAP occupying the downstream promoter can occur, else occlusion can occur.

We then confronted the analytical solutions of the step model with stochastic simulations (Section 'Stochastic simulations for the step inference model' in the S1 Appendix). We first assumed various d_{TSS}, but fixed k_{bind}, for simplicity. Visibly, M_P, CV_P^2, and S_P of the stochastic simulations are well-fitted by the analytical solution, supporting the initial assumption that CV_P^2, and S_P follow a negative binomial (Fig M in the S2 Appendix).

However, natural promoters are expected to differ in k_{bind} as they differ in sequence [48,49]. Thus, we introduced this variability and studied whether the analytical model holds. To change the variability, we obtained each k_{bind} from gamma distributions (means shown in Table 3 and CVs in Table I in the S3 Appendix). We chose a gamma distribution since its values are non-negative and non-integer (such as rate constants). Meanwhile, all parameters of the step model, aside from k_{bind}, are obtained from Tables 2 and 3. For $d_{TSS} \leq 35$ and $d_{TSS} > 35$, and each CV considered, we sampled 10000 pairs of values of $k_{bind}\cdot[R]$, and calculated M, CV^2 and S for each of them. Next, we estimated the average and standard deviation of each statistics. From Fig N in the S2 Appendix, if $CV(k_{bind})<1$, the analytical solution is robust. In that the standard error of the mean is smaller than $M_P/3$. Notably, for such CV, the strength of the

Table 3. Parameter values inferred for each model.

Interference model	Inferred parameter values	Average R^2 (M, CV^2, S) 1X medium	Average R^2 (M, CV^2, S) 0.5X medium
Exponential 1	$k_{bind}\cdot[R] = 1.09 \times 10^{-2}$ s^{-1} × (cell vol)$^{-1}$ $k_{bind} = 7.53 \times 10^{-5}$ s^{-1} $k_{unbind} = 0.84$ s^{-1} $k^{\max} = 677.7$ s^{-1} $b_1 = 5.08 \times 10^{-2}$ bp^{-1}	0.21 (Fig 6A–6C)	0.09 (Fig 6D–6F)
Exponential 2	$k_{bind}\cdot[R] = 9.71 \times 10^{-3}$ s^{-1} × (cell vol)$^{-1}$ $k_{bind} = 6.74 \times 10^{-5}$ s^{-1} $k_{unbind} = 0.80$ s^{-1} $k^{\max} = 554.8$ s^{-1} $b_1 = 7.92 \times 10^{-8}$ bp^{-1} $b_2 = 1.47 \times 10^{-3}$ bp^{-2}	0.25 (Fig 6A–6C)	0.12 (Fig 6D–6F)
Step	$k_{bind}\cdot[R] = 6.62 \times 10^{-3}$ s^{-1} × (cell vol)$^{-1}$ $k_{bind} = 4.60 \times 10^{-5}$ s^{-1} $k_{unbind} = 0.49$ s^{-1} $k^{\max} = 313.4$ s^{-1} $L = 35.11$ bp (by best fitting, which corresponds to 35 bp)	0.35 (Fig 6A–6C)	0.15 (Fig 6D–6F)
zero order	$k_{bind}\cdot[R] = 4.63 \times 10^{-3}$ s^{-1} × (cell vol)$^{-1}$ $k_{bind} = 3.22 \times 10^{-5}$ s^{-1} $k_{unbind} = 0.57$ s^{-1} $k^{\max} = 6.48$ s^{-1}	-0.007 (Fig 6A–6C)	-0.12 (Fig 6D–6F)

https://doi.org/10.1371/journal.pcbi.1009824.t003

two paired promoters would have to differ unrealistically by more than 2000%, on average (Table I in the S3 Appendix). Thus, we find the analytical solution to be reliable.

From our estimation of k_p, we further estimated a protein-to-RNA ratio, $\frac{M_P}{M_{RNA}} = \frac{k_p}{k_{pd}}$. From Eq 8 and Table 2, we find that $\frac{k_p}{k_{pd}} \sim 1418$ in both media, which agrees with previous estimations (~1832 in 27]).

Next, we used the fitted model to predict (using Eqs 8 to 11) the influence of promoter occupancy (ω) on the M_P, CV_P^2 and S_P of upstream and downstream promoters. We set d_{TSS} to 20 bp to represent promoters where ≤ 35, and to 100 bp to represent promoters with $d_{TSS} >$ 35. Then, for each cohort, we changed ω from 0.01 to 0.99 (i.e., nearly all possible values). In addition, we estimated these moments when $k_{occlusion}$, k_{occupy}, and ω are all set to zero (i.e., the two promoters do not interfere), for comparison.

From Fig 7, a pair of tandem promoters can produce less proteins than a single promoter with the same parameter values, if $d_{TSS} \leq 35$, which makes occlusion possible. Meanwhile, if $d_{TSS} > 35$, tandem promoters can only produce protein numbers in between the numbers produced by one isolated promoter and the numbers produced by two isolated promoters. In no case can two interfering tandem promoters produce more than two isolated promoters with equivalent parameter values. I.e., according to the model, the interference between tandem promoters cannot enhance production.

Meanwhile, the kinetics of the upstream (Fig 7A and panel A of Fig O in the S2 Appendix) and downstream promoters (Fig 7B and panel B of Fig O in the S2 Appendix) only differ in that the downstream promoter is more responsive to ω.

Finally, consider that the model predicts that transcription interference should occur in tandem promoters, either due to occlusion if $d_{TSS} \leq 35$ occupancy or due to occupancy of the downstream promoter if $d_{TSS} > 35$. Meanwhile, in single promoters, neither of these phenomena occurs. Thus, on average, two single promoters should produce more RNA and proteins than a pair of tandem promoters of similar strength. Using the genome wide data from [28] on

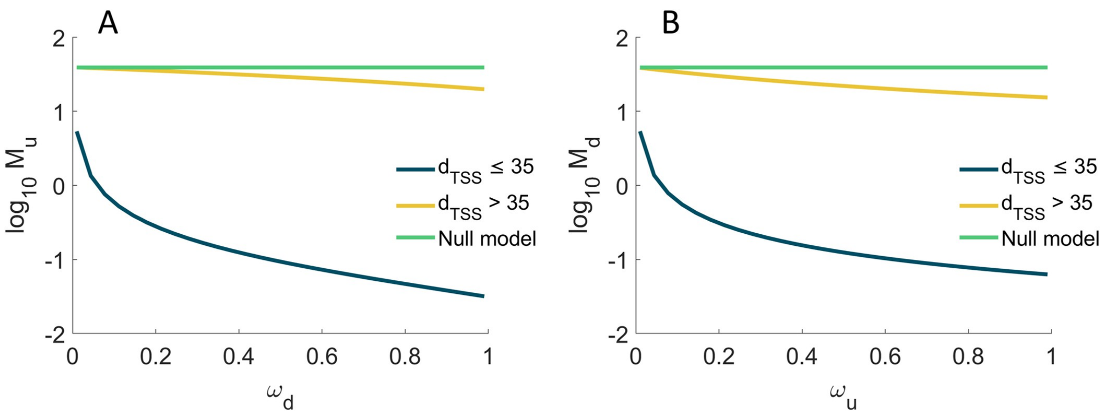


Fig 7. Mean protein numbers produced as a function of other promoter's occupancy. M_P of the single-cell distribution of the number of proteins produced (**A**) by the upstream promoter alone, and (**B**) by the downstream promoter alone. Results are shown as a function of the fraction of times that the upstream ($0.01 \leq \omega_u \leq 0.99$) and the downstream ($0.01 \leq \omega_d \leq 0.99$) promoter are occupied by RNAP. The null model is estimated by setting $k_{occlusion}$, k_{occupy}, and ω to zero.

https://doi.org/10.1371/journal.pcbi.1009824.g007

protein expression levels during exponential growth we estimated the double of the mean expression level (it equals 183.8) of genes controlled by single promoters (section 'Selection of natural genes controlled by single promoters' in the S1 Appendix). Meanwhile, also using data from [28], the mean expression level of genes controlled by tandem promoters equals 148 (estimated from the 26 that they have reported on), in agreement with the hypothesis. Nevertheless, this data is subject to external variables (e.g., TF interference). A definitive test would require the use of synthetic constructs, lesser affected by external influences.

Regulatory parameters of promoter occupancy and occlusion

Since the occupancy, ω, of each of the tandem promoters is responsible for transcriptional interference by occlusion and by RNAPs occupying the downstream promoter, we next explored the biophysical limits of ω. Eqs 6A and 6B define the occupancies of the upstream and downstream promoters, ω_u and ω_d, respectively. For simplicity, here we refer to both of them as ω. Fig 8A shows that ω increases with the rate of RNAP binding ($k_{bind}\cdot[R]$), but only within a certain range of (high) values of the time from binding to elongating (k_{after}^{-1}). I.e., RNAPs need to spend a significant time in OC, if they are to cause interference, which is expected. Similarly, ω changes with k_{after}^{-1}, but only for high values of $k_{bind}\cdot[R]$. I.e., if it's rare for RNAPs to bind, the occupancy will necessarily be weak.

In detail, from Fig 8A, ω can change significantly within $10^{-2} < k_{bind}\times[R] < 10$ s^{-1} and $10^{-2} < k_{after}^{-1} < 10^{2}$ s. For these ranges, we expect RNA production rates (k_r, Eqs 5A, 5B, 6B, 7 and 9) to vary from ~10^{-5} (if $d_{TSS} \leq 35$) and ~10^{-4} (if $d_{TSS} > 35$) until 10 s^{-1}. In agreement, in *E. coli*, promoters have RNA production rates from ~10^{-3} to 10^{-1} s^{-1} when induced [20–21,39,50–51] and ~10^{-4} to 10^{-6} s^{-1} when non-fully active [28]. Thus, ω can differ within realistic intervals of parameter values.

Next, we estimated $k_{occlusion}$, the rate at which a promoter occludes the other as a function of d_{TSS} and ω using Eqs 6A and 6B. k^{max} is shown in Table 3. To model $I(d_{TSS})$ we used the step function in Table 1. Overall, $k_{occlusion}$ changes linearly with ω, when and only when $d_{TSS} \leq$ 35 (Fig 8B).

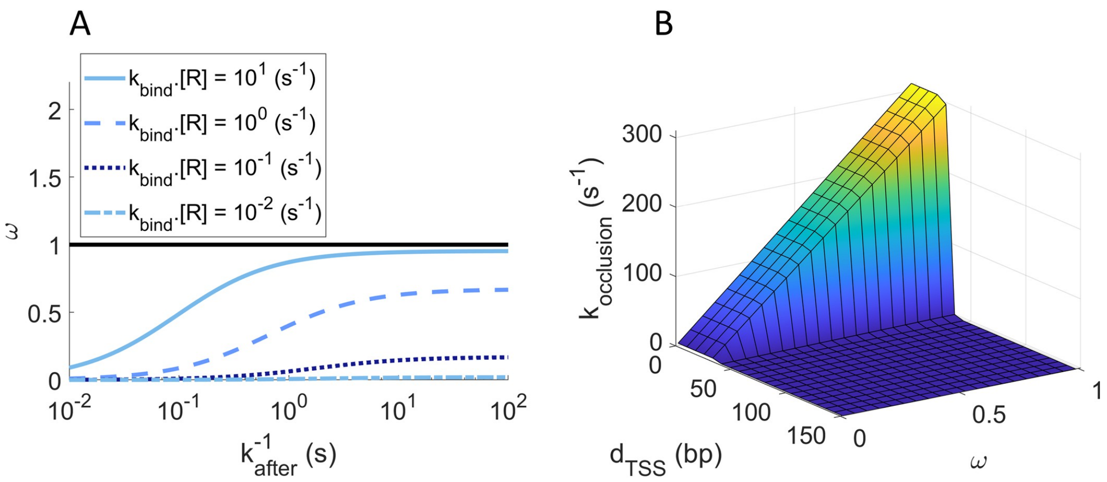


Fig 8. Promoter occupancy *ω* estimated for the step model. (A) ω as a function of the rate constant for a *free* RNAP to bind to the *unoccupied* promoter (k_{bind}·[R]) and of the time for that RNAP to start elongation after commitment to OC, k_{after}^{-1}. The horizontal black line at $\omega = 1$, is the maximum fraction of time that the promoter can be occupied (i.e., the maximum promoter occupancy). (B) $k_{occlusion}$ plotted as a function of ω and d_{TSS}. Since $k_{occlusion}$ increases with ω if and only if $d_{TSS} \leq 35$, it renders the simultaneous occupation of both TSS's impossible.

https://doi.org/10.1371/journal.pcbi.1009824.g008

State space of the single cell statistics of protein numbers of tandem promoters

We next studied how much the single-cell statistics of protein numbers (M_P, CV_P^2, and S_P) of the upstream, 'u', and downstream, 'd', promoters changes with ω_u, ω_d, and d_{TSS}. Here, ω_u and ω_d are increased from 0 to 1 by increasing the respective k_{bind} (Eqs 6A and 6B).

From Fig 9A, if $d_{TSS} \leq 35$ bp, reducing ω_d while also increasing ω_u is the most effective way to increase M_u, since this increases the number of RNAPs transcribing from the upstream promoter that are not hindered by RNAPs occupying the downstream promoter. If $d_{TSS} > 35$ bp, the occupancy the downstream promoter, ω_d, becomes ineffective.

Oppositely, from Fig 9B, if $d_{TSS} \leq 35$ bp, increasing ω_d while also decreasing ω_d, is the most effective way to increase M_d since this increases the number of RNAPs transcribing from the

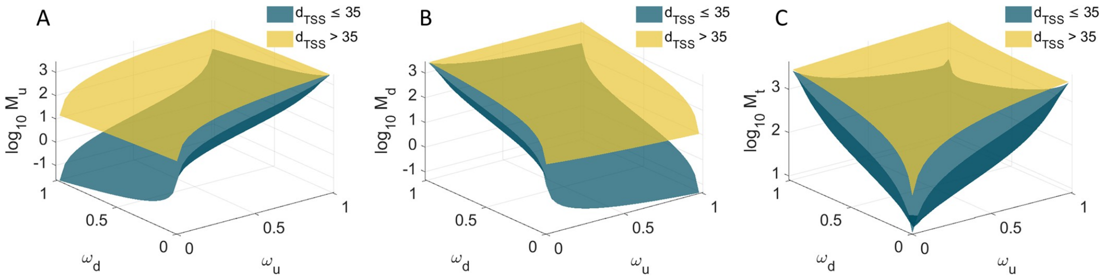


Fig 9. Mean protein expression as a function of both promoters' occupancy. Expected mean protein numbers due to the activity of: (**A**) the upstream promoter alone, (**B**) the downstream promoter alone, and (**C**) both promoters. M_P is shown as a function of the fraction of times that the upstream ($0 \leq \omega_u \leq 1$) and the downstream ($0 \leq \omega_d \leq 1$) promoters are occupied by RNAP, when $d_{TSS} > 35$ (yellow) and $d_{TSS} \leq 35$ (dark green) bp.

https://doi.org/10.1371/journal.pcbi.1009824.g009

downstream promoter does not interfere by RNAPs elongating from the upstream promoter. If $d_{TSS} > 35$ bp, the occupancy the upstream promoter, ω_u, becomes ineffective.

Finally, from Fig 9C, regardless of d_{TSS}, for small ω_d and ω_u, as the occupancies increase, M_t increases quickly and in a non-linear fashion. However, as both ω_d and ω_u reach high values, M_t decreases for further increases, if $d_{TSS} \leq 35$ bp. Instead, if $d_{TSS} > 35$ bp, Mt appears to saturate.

From Fig P in the S2 Appendix, CV_P^2 and S_P behave inversely to M_P.

Relevantly, in all cases, the range of predicted protein numbers (Fig 9C) are in line with the empirical values (~10^{-1} to 10^3 proteins per cell) (Fig 4D).

Discussion

E. coli genes controlled by tandem promoters have a relatively high mean conservation level (0.2, while the average gene has 0.15, with a p-value of 0.009), suggesting that they play particularly relevant biological roles (section 'Gene Conservation' in the S1 Appendix). From empirical data on single-cell protein numbers of 30 *E. coli* genes controlled by tandem promoters, we found evidence that their dynamics is subject to RNAP interference between the two promoters. This interference reduces the mean single-cell protein numbers, while increasing its CV^2 and skewness, and can be tuned by ω, the promoters' occupancy by RNAP, and by d_{TSS}. Since both of these parameters are sequence dependent [21,31] the interference should be evolvable. Further, since ω of at least some of these genes should be under the influence of their several input TFs, the interference has the potential to be adaptive.

We proposed models of the dynamics of these genes as a function of ω and d_{TSS}, using empirically validated parameter values. In our best fitting model, transcription interference is modelled by a step function of d_{TSS} (instead of gradually changing with d_{TSS}), since the only detectable differences in dynamics with changing d_{TSS} were between tandem promoters with $d_{TSS} \leq 35$ and $d_{TSS} > 35$ nucleotides (the latter cohort of genes having higher mean expression and lower variability). We expect that causes this difference tangible is the existence of the OC formation. In detail, the OC is a long-lasting DNA-RNAP formation that occupies that strict region of DNA at the promoter region [24,31]. As such, occlusion should share these physical features. Because of that, when $d_{TSS} \leq 35$, an RNAP bound to TSS always occludes the other TSS, significantly reducing RNA production. Meanwhile, if $d_{TSS} > 35$, interference occurs when an RNAP elongating from the upstream promoter is obstructed by an RNAP occupying the downstream promoter.

Meanwhile, contrary to d_{TSS}, if one considers realistic ranges of the other model parameters, it is possible to predict a very broad range of accessible dynamics for tandem promoter arrangements. This could explain the observed diversity of single-cell protein numbers as a function of d_{TSS} (Fig 6). At the evolutionary level, such potentially high range of dynamics may provide high evolutionary adaptability and thus, it may be one reason why genes controlled by these promoters are relatively more conserved.

One potentially confounding effect which was not accounted for in this model is the accumulation of supercoiling. Closely spaced promoters may be more sensitive to supercoiling buildup than single promoters [52–54]. If so, it will be useful to extend the model to include these effects [26]. Using such model and measurements of expression by tandem promoters when subject to, e.g. Novobiocin [55], may be of use to infer kinetic parameters of promoter locking due to positive supercoiling build-up.

Other potential improvements could be expanding the model to tandem arrangements other than I and II (Fig 1), to include a third form of interference (transcription elongation of a nearby gene).

One open question is whether placing promoters in tandem formation increases the robustness of downstream gene expression to perturbations (e.g., fluctuations in the concentrations of

RNAP or TF regulators). A tandem arrangement likely increases the robustness to perturbations which only influence one of the promoters. Another open question is why several of the 102 tandem promoters with arrangements I and II appeared to behave independently from their input TFs (according to the RNA-seq data), albeit having more input TFs (1.62 on average) than expected by chance (the average *E. coli* gene only has 0.95). As noted above, we hypothesize that these input TFs may become influential in conditions other than the ones studied here.

Here, we also did not consider any influence from the phenomenon of "RNAP cooperation" [56]. This is based on this being an occurrence in elongation, and we expect interactions between two *elongating* RNAPs to rarely affect the interference between tandem promoters [9]. However, potentially, it could be of relevance in the strongest tandem promoters.

Finally, a valuable future study on tandem promoters will require the use of synthetic tandem promoters (integrated in a specific chromosome location) that systematically differ in promoter strengths and nucleotide distances. This would allow extracting parameter values associated to promoter interference to create a more precise model than the one based on the natural promoters (which is influenced by TFs, etc). Similarly, measuring the strength of individual natural promoters would contribute to this effort.

Overall, our model, based on a significant number of natural tandem promoters whose genes have a wide range of expression levels, should be applicable to the natural tandem promoters not observed here (at least of arrangements I and II), including of other bacteria, and to be accurate in predicting the dynamics of synthetic promoters in these arrangements.

Currently, predicting how gene expression kinetics change with the promoter sequence remains challenging. Even single- or double-point mutants of known promoters behave unpredictably, likely because the individual sequence elements influence the OC and CC in a combinatorial fashion. Consequently, the present design of synthetic circuits is usually limited to the use of a few promoters whose dynamics have been extensively characterized (Lac, Tet, etc.). This severely limits present synthetic engineering.

We suggest that a promising methodology to create new synthetic genes with a wide range of predictable dynamics is to assemble well-characterized promoters in a tandem formation, and to tune their target dynamics using our model. Specifically, for a given dynamics, it is possible to invert the model and find a suitable pair of promoters with known occupancies and corresponding d_{TSS} (smaller or larger than 35), which achieve these dynamics. A similar strategy was recently proposed in order to achieve strong expression levels [57]. Our results agree and further expand on this by showing that the mean expression level can also be reduced and expression variability can further be fine-tuned.

Importantly, this can already be executed, e.g., using a library of individual genes whose expression can be measured [28]. From this library, we can select any two promoters of interest and arrange them as presented here, in order to obtain a kinetics of expression as close as possible to a given target. Note that these dynamics have a wide range, from weaker to stronger than that of either promoter (albeit no stronger than their sum, Fig 9C). Given the number of natural genes whose expression is already known and given the present accuracy in assembling specific nucleotide sequences, we expect this method to allow the rapid engineering of genes with desired dynamics with an enormous range of possible behaviours. As such, these constructs could represent a recipe book for the components of gene circuits with predictable complex kinetics.

Materials and methods

Using information from RegulonDB v10.5 as of 30th of January 2020 [58], we started by searching natural genes controlled by two promoters (Section 'Selection of natural genes controlled by tandem promoters' in the S1 Appendix). Next, we studied their evolutionary conservation and ontology (Sections 'Gene conservation' and 'Gene Ontology' in the S1 Appendix) and analysed their local topological features within the TFN of *E. coli* (Section 'Network topological properties' in the S1 Appendix).

RNA-seq measurements were conducted in two points in time (Section 'RNA-seq measurements and data analysis' in the S1 Appendix), to obtain fold changes in RNA numbers of genes controlled by tandem promoters with arrangements I and II, their input TFs, and their output genes (Fig 1). We used this data to search for relationships between input and output genes.

Next, a model of gene expression was proposed, and reduced to obtain an analytical solution of the single-cell protein expression statistics of tandem promoters (Sections 'Derivation of mean protein numbers at steady state produced by a pair of tandem promoters' and 'CV^2 and skewness of the distribution of single-cell protein numbers of model tandem promoters' in the S1 Appendix). This analytical solution was compared to stochastic simulations conducted using the simulator SGNS2. (Section 'Stochastic simulations for the step inference model' in the S1 Appendix).

We collected single-cell flow-cytometry measurements of 30 natural genes controlled by tandem promoters (Section 'Flow-cytometry and data analysis' in the S1 Appendix) to validate the model. For this, first, from the original data, we subtracted the cellular background fluorescence (Section 'Subtraction of background fluorescence from the total protein fluorescence' in the S1 Appendix). Then, we converted the fluorescence intensity into protein numbers (Section 'Conversion of protein fluorescence to protein numbers in the S1 Appendix). From this we obtained empirical data on M, CV^2*, and S* of the single-cell distributions of protein numbers in two media (Sections 'Media and chemicals' and 'Strains and growth conditions' in the S1 Appendix). Flow-cytometry measurements were also compared to microscopy data, supported by image analysis (Section 'Microscopy and Image analysis' in the S1 Appendix), for validation.

Comparing the data from RegulonDB (30.01.2020) used here, with the most recent (21.07.2021), we found that the numbers of genes controlled by tandem promoters of arrangements I and II differed by ~4% (from 102 to 98). Regarding those whose activity was measured by flow-cytometry, this difference is ~3% (30 to 31). Globally, 163 TF-gene interactions differed (~3.4%) while for the 98 genes controlled by tandem promoters of arrangements I and II, only 10 TF-gene interactions differ (~2.7%). Finally, globally the numbers of TUs differed by ~1%, promoters by ~0.6%, genes by ~1%, and terminators by ~15% (which did not affect the genes studied, as they changed by ~4% only). These small differences should not affect our conclusions.

Finally, a data package is provided in Dryad [59] with flow-cytometry and microscopy data and codes used. The RNAseq data has been deposited in NCBI's Gene Expression Omnibus [60] and are accessible through GEO Series accession number GSE183139 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183139).

Dryad DOI

10.5061/dryad.bnzs7h4bs.

Supporting information

S1 Appendix. Extended Materials and Methods.
(DOCX)

S2 Appendix. Supporting Figures.
(DOCX)

S3 Appendix. Supporting Tables.
(DOCX)

S4 Appendix. Supporting Results.
(DOCX)

S1 Table. Gene Ontology. Overrepresentation tests using the PANTHER Classification System. List of biological processes which are overrepresented using Fisher's exact tests are shown. (Excel)
(XLSX)

S2 Table. Protein statistics. Statistics of single-cell distributions of protein fluorescence of genes controlled by tandem promoters as measured by flow-cytometry in 1X and 0.5X diluted M9 media conditions. (Excel)
(XLSX)

S3 Table. Protein statistics. Statistics of single-cell distributions of protein fluorescence of genes controlled by single promoter as measured by flow-cytometry in 1X M9 media condition. (Excel)
(XLSX)

Acknowledgments

The authors thank Jason Lloyd-Price for proof-reading and editing the text.

Author Contributions

Conceptualization: Vatsala Chauhan, Mohamed N. M. Bahrudeen, Cristina S. D. Palma, Andre S. Ribeiro.

Formal analysis: Vatsala Chauhan, Mohamed N. M. Bahrudeen.

Funding acquisition: Andre S. Ribeiro.

Investigation: Vatsala Chauhan, Mohamed N. M. Bahrudeen, Cristina S. D. Palma, Ines S. C. Baptista, Bilena L. B. Almeida, Suchintak Dash, Vinodh Kandavalli, Andre S. Ribeiro.

Methodology: Vatsala Chauhan, Mohamed N. M. Bahrudeen, Ines S. C. Baptista.

Project administration: Andre S. Ribeiro.

Software: Mohamed N. M. Bahrudeen, Ines S. C. Baptista.

Supervision: Andre S. Ribeiro.

Writing – original draft: Vatsala Chauhan, Mohamed N. M. Bahrudeen, Ines S. C. Baptista, Andre S. Ribeiro.

Writing – review & editing: Vatsala Chauhan, Mohamed N. M. Bahrudeen, Cristina S. D. Palma, Ines S. C. Baptista, Andre S. Ribeiro.

References

1. Herbert M, Kolb A, Buc H. Overlapping promoters and their control in Escherichia coli: the gal case. Proc Natl Acad Sci U S A. 1986; 83: 2807–2811. https://doi.org/10.1073/pnas.83.9.2807 PMID: 3010319

2. Beck CF, Warren RA. Divergent promoters, a common form of gene organization. Microbiol Rev. 1988; 52: 318–326. https://doi.org/10.1128/mr.52.3.318-326.1988 PMID: 3054465

3. Adachi N, Lieber MR. Bidirectional gene organization: a common architectural feature of the human genome. Cell. 2002; 109: 807–809. https://doi.org/10.1016/s0092-8674(02)00758-4 PMID: 12110178

4. Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP, Myers RM. An abundance of bidirectional promoters in the human genome. Genome Res. 2004; 14: 62–66. https://doi.org/10.1101/gr.1982804 PMID: 14707170

5. Shearwin KE, Callen BP, Egan JB. Transcriptional interference—a crash course. Trends Genet. 2005; 21: 339–345. https://doi.org/10.1016/j.tig.2005.04.009 PMID: 15922833

6. Prescott EM, Proudfoot NJ. Transcriptional collision between convergent genes in budding yeast. Proc Natl Acad Sci U S A. 2002; 99: 8796–8801. https://doi.org/10.1073/pnas.132270899 PMID: 12077310

7. Korbel JO, Jensen LJ, von Mering C, Bork P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat Biotechnol. 2004; 22: 911–917. https://doi.org/10.1038/nbt988 PMID: 15229555

8. Wei W, Xiang H, Tan H. Two tandem promoters to increase gene expression in Lactococcus lactis. Biotechnol Lett. 2002; 24: 1669–1672. https://doi.org/10.1023/A:1020653417455

9. Sneppen K, Dodd IB, Shearwin KE, Palmer AC, Schubert RA, Callen BP, et al. A mathematical model for transcriptional interference by RNA polymerase traffic in Escherichia coli. J Mol Biol. 2005; 346: 399–409. https://doi.org/10.1016/j.jmb.2004.11.075 PMID: 15670592

10. Martins L, Mäkelä J, Häkkinen A, Kandhavelu M, Yli-Harja O, Fonseca JM, et al. Dynamics of transcription of closely spaced promoters in Escherichia coli, one event at a time. J Theor Biol. 2012; 301: 83–94. https://doi.org/10.1016/j.jtbi.2012.02.015 PMID: 22370562

11. Horowitz H, Platt T. Regulation of transcription from tandem and convergent promoters. Nucleic Acids Res. 1982; 10: 5447–5465. https://doi.org/10.1093/nar/10.18.5447 PMID: 6755394

12. Bordoy AE, Varanasi US, Courtney CM, Chatterjee A. Transcriptional Interference in Convergent Promoters as a Means for Tunable Gene Expression. ACS Synth Biol. 2016; 5: 1331–1341. https://doi.org/10.1021/acssynbio.5b00223 PMID: 27346626

13. Palmer AC, Ahlgren-Berg A, Egan JB, Dodd IB, Shearwin KE. Potent transcriptional interference by pausing of RNA polymerases over a downstream promoter. Mol Cell. 2009; 34: 545–555. https://doi.org/10.1016/j.molcel.2009.04.018 PMID: 19524535

14. Callen BP, Shearwin KE, Egan JB. Transcriptional Interference between Convergent Promoters Caused by Elongation over the Promoter. Mol Cell. 2004; 14: 647–656. https://doi.org/10.1016/j.molcel.2004.05.010 PMID: 15175159

15. Hoffmann SA, Hao N, Shearwin KE, Arndt KM. Characterizing Transcriptional Interference between Converging Genes in Bacteria. ACS Synth Biol. 2019; 8: 466–473. https://doi.org/10.1021/acssynbio.8b00477 PMID: 30717589

16. Masulis IS, Babaeva ZS, Chernyshov SV, Ozoline ON. Visualizing the activity of Escherichia coli divergent promoters and probing their dependence on superhelical density using dual-colour fluorescent reporter vector. Sci Rep. 2015; 5: 1–10. https://doi.org/10.1038/srep11449 PMID: 26081797

17. Vogl T, Kickenweiz T, Pitzer J, Sturmberger L, Weninger A, Biggs BW, et al. Engineered bidirectional promoters enable rapid multi-gene co-expression optimization. Nat Commun. 2018; 9: 1–13. https://doi.org/10.1038/s41467-017-02088-w PMID: 29317637

18. Adhya S, Gottesman M. Promoter occlusion: Transcription through a promoter may inhibit its activity. Cell. 1982; 29: 939–944. https://doi.org/10.1016/0092-8674(82)90456-1 PMID: 6217898

19. Eszterhas SK, Bouhassira EE, Martin DIK, Fiering S. Transcriptional interference by independently regulated genes occurs in any relative arrangement of the genes and is influenced by chromosomal integration position. Mol Cell Biol. 2002; 22: 469–479. https://doi.org/10.1128/MCB.22.2.469-479.2002 PMID: 11756543

20. Lloyd-Price J, Startceva S, Kandavalli V, Chandraseelan JG, Goncalves N, Oliveira SMD, et al. Dissecting the stochastic transcription initiation process in live Escherichia coli. DNA Res. 2016; 23: 203–214. https://doi.org/10.1093/dnares/dsw009 PMID: 27026687

21. Lutz R, Lozinski T, Ellinger T, Bujard H. Dissecting the functional program of Escherichia coli promoters: the combined mode of action of Lac repressor and AraC activator. Nucleic Acids Res. 2001; 29: 3873–3881. https://doi.org/10.1093/nar/29.18.3873 PMID: 11557820

22. McClure WR. Rate-limiting steps in RNA chain initiation. Proc Natl Acad Sci U S A. 1980; 77: 5634–5638. https://doi.org/10.1073/pnas.77.10.5634 PMID: 6160577

23. Krummel B, Chamberlin MJ. Structural analysis of ternary complexes of Escherichia coli RNA polymerase. Deoxyribonuclease I footprinting of defined complexes. J Mol Biol. 1992; 225: 239–250. https://doi.org/10.1016/0022-2836(92)90918-a PMID: 1593619

24. deHaseth Pieter L., Zupancic Margaret L., Record M. Thomas. RNA Polymerase-Promoter Interactions: the Comings and Goings of RNA Polymerase. J Bacteriol. 1998; 180: 3019–3025. https://doi.org/10.1128/JB.180.12.3019-3025.1998 PMID: 9620948

25. Greive SJ, von Hippel PH. Thinking quantitatively about transcriptional regulation. Nat Rev Mol Cell Biol. 2005; 6: 221–232. https://doi.org/10.1038/nrm1588 PMID: 15714199

26. Palma CSD, Kandavalli V, Bahrudeen MNM, Minoia M, Chauhan V, Dash S, et al. Dissecting the in vivo dynamics of transcription locking due to positive supercoiling buildup. Biochimica et Biophysica Acta (BBA)—Gene Regulatory Mechanisms. 2020; 1863: 194515. https://doi.org/10.1016/j.bbagrm.2020.194515 PMID: 32113983

27. Häkkinen A, Oliveira SMD, Neeli-Venkata R, Ribeiro AS. Transcription closed and open complex formation coordinate expression of genes with a shared promoter region. J R Soc Interface. 2019; 16: 20190507. https://doi.org/10.1098/rsif.2019.0507 PMID: 31822223

28. Taniguchi Y, Choi PJ, Li G-W, Chen H, Babu M, Hearn J, et al. Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. Science. 2010; 329: 533–538. https://doi.org/10.1126/science.1188308 PMID: 20671182

29. Friedman LJ, Mumm JP, Gelles J. RNA polymerase approaches its promoter without long-range sliding along DNA. Proc Natl Acad Sci U S A. 2013; 110: 9740–9745. https://doi.org/10.1073/pnas.1300221110 PMID: 23720315

30. Skinner GM, Baumann CG, Quinn DM, Molloy JE, Hoggett JG. Promoter Binding, Initiation, and Elongation by Bacteriophage T7 RNA Polymerase: A SINGLE-MOLECULE VIEW OF THE TRANSCRIPTION CYCLE*. J Biol Chem. 2004; 279: 3239–3244. https://doi.org/10.1074/jbc.M310471200 PMID: 14597619

31. McClure WR. Mechanism and control of transcription initiation in prokaryotes. Annu Rev Biochem. 1985; 54: 171–204. https://doi.org/10.1146/annurev.bi.54.070185.001131 PMID: 3896120

32. Saecker RM, Record MT Jr, Dehaseth PL. Mechanism of bacterial transcription initiation: RNA polymerase—promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. J Mol Biol. 2011; 412: 754–771. https://doi.org/10.1016/j.jmb.2011.01.018 PMID: 21371479

33. Mekler V, Kortkhonjia E, Mukhopadhyay J, Knight J, Revyakin A, Kapanidis AN, et al. Structural Organization of Bacterial RNA Polymerase Holoenzyme and the RNA Polymerase-Promoter Open Complex. Cell. 2002; 108: 599–614. https://doi.org/10.1016/s0092-8674(02)00667-0 PMID: 11893332

34. Margeat E, Kapanidis AN, Tinnefeld P, Wang Y, Mukhopadhyay J, Ebright RH, et al. Direct Observation of Abortive Initiation and Promoter Escape within Single Immobilized Transcription Complexes. Biophys J. 2006; 90: 1419–1431. https://doi.org/10.1529/biophysj.105.069252 PMID: 16299085

35. Hsu LM. Promoter clearance and escape in prokaryotes. Biochim Biophys Acta. 2002; 1577: 191–207. https://doi.org/10.1016/s0167-4781(02)00452-9 PMID: 12213652

36. Hsu LM. Promoter Escape by Escherichia coli RNA Polymerase. EcoSal Plus. 2008;3. https://doi.org/10.1128/ecosalplus.4.5.2.2 PMID: 26443745

37. Henderson KL, Felth LC, Molzahn CM, Shkel I, Wang S, Chhabra M, et al. Mechanism of transcription initiation and promoter escape by E. coli RNA polymerase. Proc Natl Acad Sci U S A. 2017; 114: E3032–E3040. https://doi.org/10.1073/pnas.1618675114 PMID: 28348246

38. Ponnambalam S, Busby S. RNA polymerase molecules initiating transcription at tandem promoters can collide and cause premature transcription termination. FEBS Lett. 1987; 212: 21–27. https://doi.org/10.1016/0014-5793(87)81549-1 PMID: 3542569

39. Kandavalli VK, Tran H, Ribeiro AS. Effects of σ factor competition are promoter initiation kinetics dependent. Biochim Biophys Acta. 2016; 1859: 1281–1288. https://doi.org/10.1016/j.bbagrm.2016.07.011 PMID: 27452766

40. Bremer H, Dennis P, Ehrenberg M. Free RNA polymerase and modeling global transcription in Escherichia coli. Biochimie. 2003; 85: 597–609. https://doi.org/10.1016/s0300-9084(03)00105-6 PMID: 12829377

41. Patrick M, Dennis PP, Ehrenberg M, Bremer H. Free RNA polymerase in Escherichia coli. Biochimie. 2015; 119: 80–91. https://doi.org/10.1016/j.biochi.2015.10.015 PMID: 26482806

42. Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, et al. Noise in protein expression scales with natural protein abundance. Nat Genet. 2006; 38: 636–643. https://doi.org/10.1038/ng1807 PMID: 16715097

43. Ju X, Li D, Liu S. Full-length RNA profiling reveals pervasive bidirectional transcription terminators in bacteria. Nat Microbiol. 2019; 4: 1907–1918. https://doi.org/10.1038/s41564-019-0500-z PMID: 31308523

44. Hausser J, Mayo A, Keren L, Alon U. Central dogma rates and the trade-off between precision and economy in gene expression. Nat Commun. 2019; 10: 1–15. https://doi.org/10.1038/s41467-018-07882-8 PMID: 30602773

45. Lagarias JC, Reeds JA, Wright MH, Wright PE. Convergence Properties of the Nelder—Mead Simplex Method in Low Dimensions. SIAM J Optim. 1998; 9: 112–147. https://doi.org/10.1137/S1052623496303470

46. Maurizi MR. Proteases and protein degradation in Escherichia coli. Experientia. 1992; 48: 178–201. https://doi.org/10.1007/BF01923511 PMID: 1740190

47. Koch AL, Levy HR. Protein turnover in growing cultures of Escherichia coli. J Biol Chem. 1955; 217: 947–957. Available: https://www.ncbi.nlm.nih.gov/pubmed/13271454 PMID: 13271454

48. Rydenfelt M, Garcia HG, Cox RS 3rd, Phillips R. The influence of promoter architectures and regulatory motifs on gene expression in Escherichia coli. PLoS One. 2014; 9: e114347. https://doi.org/10.1371/journal.pone.0114347 PMID: 25549361

49. Buchler NE, Gerland U, Hwa T. On schemes of combinatorial transcription logic. Proc Natl Acad Sci U S A. 2003; 100: 5136–5141. https://doi.org/10.1073/pnas.0930314100 PMID: 12702751

50. Golding I, Paulsson J, Zawilski SM, Cox EC. Real-Time Kinetics of Gene Activity in Individual Bacteria. Cell. 2005; 123: 1025–1036. https://doi.org/10.1016/j.cell.2005.09.031 PMID: 16360033

51. Startceva S, Kandavalli VK, Visa A, Ribeiro AS. Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression. Biochimica et Biophysica Acta (BBA)—Gene Regulatory Mechanisms. 2019; 1862: 119–128. https://doi.org/10.1016/j.bbagrm.2018.12.005 PMID: 30557610

52. Rhee KY, Opel M, Ito E, Hung S p., Arfin SM, Hatfield GW. Transcriptional coupling between the divergent promoters of a prototypic LysR-type regulatory system, the ilvYC operon of Escherichia coli. Proc Natl Acad Sci U S A. 1999; 96: 14294–14299. https://doi.org/10.1073/pnas.96.25.14294 PMID: 10588699

53. Jia J, King JE, Goldrick MC, Aldawood E, Roberts IS. Three tandem promoters, together with IHF, regulate growth phase dependent expression of the Escherichia coli kps capsule gene cluster. Sci Rep. 2017; 7: 1–11. https://doi.org/10.1038/s41598-016-0028-x PMID: 28127051

54. Yeung E, Dy AJ, Martin KB, Ng AH, Del Vecchio D, Beck JL, et al. Biophysical Constraints Arising from Compositional Context in Synthetic Gene Networks. Cell Syst. 2017; 5: 11–24.e12. https://doi.org/10.1016/j.cels.2017.06.001 PMID: 28734826

55. Chong S, Chen C, Ge H, Xie XS. Mechanism of transcriptional bursting in bacteria. Cell. 2014; 158: 314–326. https://doi.org/10.1016/j.cell.2014.05.038 PMID: 25036631

56. Epshtein V, Nudler E. Cooperation between RNA polymerase molecules in transcription elongation. Science. 2003; 300: 801–805. https://doi.org/10.1126/science.1083219 PMID: 12730602

57. Li M, Wang J, Geng Y, Li Y, Wang Q, Liang Q, et al. A strategy of gene overexpression based on tandem repetitive promoters in Escherichia coli. Microb Cell Fact. 2012; 11: 19. https://doi.org/10.1186/1475-2859-11-19 PMID: 22305426

58. Santos-Zavaleta A, Salgado H, Gama-Castro S, Sánchez-Pérez M, Gómez-Romero L, Ledezma-Tejeida D, et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. Nucleic Acids Res. 2019; 47: D212–D220. https://doi.org/10.1093/nar/gky1077 PMID: 30395280

59. Chauhan V, Bahrudeen MNM, Palma CSD, Ines SCB, Almeida BLB, Dash S, et al. Analytical kinetic model of native tandem promoters in E. coli, Dryad, Dataset. https://doi.org/10.5061/dryad.bnzs7h4b

60. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30: 207. https://doi.org/10.1093/nar/30.1.207 PMID: 11752295

S1 Appendix: Extended Materials and Methods

Selection of natural genes controlled by tandem promoters

We define a pair of tandem promoters as two promoters in a head-to-tail formation transcribing the same gene, as in [1]. In order to find them in the genome of *E. coli*, from RegulonDB, we obtained the lists of all known transcription units (TUs), promoters (defined as stretches of 60 upstream and 20 downstream nucleotide sequences from a TSS), gene sequences, TFs, and terminators [2].

From the list of TUs (3560), we extracted all genes (510) under the control of two and only two promoters in tandem formation with known TSS and DNA strand (information from the promoters' list). Then, we calculated the nucleotide distance between their pair of TSSs (d_{TSS}) and obtained the start and end positions of their sequence in the DNA. As a side note, we found additional 321 genes controlled by more than two promoters in tandem formation, which are not accounted for as they are not included in the model, for simplicity.

Next, we removed all genes with another gene or promoter sequence (associated to a TU) located in the opposing strand anywhere between the start of the upstream promoter and the end of the gene sequence (186 out of 510) since their dynamics may be subject to interference from convergent RNAPs [1,3,4]

Out of the remaining 324 genes, only 152 are in the first position of a TU or in a TU with only one gene. Since evidence suggests that the existence of multiple genes in a TU influences their transcription significantly, due to premature terminations, distance to the promoter etc. [5,6], we opted for keeping only those 152 genes. Subsequently, from the list of terminators, we obtained their start and end positions and DNA strand and filtered out (9 out of 152) genes with a terminator sequence in between the beginning of the upstream promoter and the end of the gene sequence, due to potential enhanced premature terminations. Finally, from these, we only considered promoter pairs (102 out of the 143 genes) such that no gene is coded in the regions containing them or the space in between them (Fig 1), so that elongation of other genes do not perturb their transcription.

Finally, of these 102 genes, we measured the expression levels at the single-cell level of 30 of them (Table A in S3 Appendix) using a YFP strain library [7]. These genes are of the categories 'I' (9 genes) and 'II' (21 genes) in Fig 1. Their d_{TSS} range from 84 to 173, and from 3 to 73 nucleotides, respectively.

Selection of natural genes controlled by single promoters

To select natural genes controlled by single promoters in the genome of *E. coli,* from RegulonDB, we obtained the lists of all known transcription units (TUs), promoters, gene sequences and terminators [2]. From the list of TUs (3560), we extracted all genes (1760) under the control of one and only one promoter with known TSS and DNA strand (information from the promoters' list). Next, we filtered out all genes with another gene or promoter sequence (associated to a TU) located in the opposing strand anywhere between the start of the promoter and the end of the gene sequence (446 out of 1760) since their dynamics may be subject to interference from convergent RNAPs [1,3,4] Out of the remaining 1314 genes, only 649 are in the first position of a TU or in a TU with only one gene and no other promoter sequence (associated to another TU) between the promoter and the end of the gene of interest. Since evidence suggests that the existence of multiple genes in a TU influences their transcription significantly, due to premature terminations, distance to the promoter etc. [5,6], we opted for keeping only those 649 genes. Subsequently, from the list of terminators, we obtained their start and end positions and DNA strand and filtered out (36 out of 649) genes with a terminator sequence in between the promoter and the end of the gene sequence, due to potential enhanced premature terminations. Finally, of these 613 genes, we obtained data on the expression levels of 126 genes from [7], which we used to compare expression levels of genes controlled by tandem promoters and genes controlled by single promoters.

Meanwhile, for purposes of validating the scaling factor between protein fluorescence and numbers, of these 613 genes, we measured the expression levels at the single-cell level of 10 of them, randomly selected (Table B in S3 Appendix) [7].

Gene Conservation

From a list of 5443 reference bacterial genomes [8], we used the Rentrez package [9] to obtain which genes are present in each genome. Next, we removed those genomes without gene entries (1310). Using the remaining genomes, we estimated the evolutionary conservation of each gene in the genome of MG1655 (GCF_000005845.2_ASM584v2), including those controlled by tandem promoters, by the ratio between the number of genomes where the gene is present, and the total number of genomes considered. Fig Q in S2 Appendix shows the conservation levels as a function of d_{TSS} of the tandem promoters controlling the genes' expression.

Gene Ontology (GO)

For gene ontology representations, we performed overrepresentation tests using the PANTHER Classification System [10], which finds statically significant overrepresentations using Fisher's exact tests. For p-values < α (here set to 0.05), the null hypothesis that there are no associations between the gene cohort and the corresponding GO of the biological process is rejected, which we interpret as the gene cohort being associated with corresponding GO of the biological process.

Network topological properties

By 'network topological property' we refer to some feature of a gene that is related to how that gene is integrated with the network formed by TFs linking genes. We used *Cytoscape* [11] to extract these features for the genes controlled by tandem promoters from the *known* transcription factor (TF) network of *E. coli*, using information from RegulonDB v10.5 on all known transcription factors (TFs) and their binding sites [2].

Next, for the two cohorts of genes with d_{TSS} larger or not than 35 bps, based on definitions in [12], we calculated (Table C in S3 Appendix) the mean and standard error of each cohort's average shortest path length (minimum number of edges between pairs of genes), clustering coefficient (fraction of input nodes to a node that are also linked), eccentricity (maximum non-infinite shortest path length between the node and another node in the network), edge count (number of edges/nodes that are connected to the node), indegree (number of incoming edges), neighbourhood connectivity (average connectivity of all nearest neighbours), and outdegree (number of outgoing edges).

For each feature, we also obtained a p-value, which is the probability that the genes of the cohort have a smaller mean than the mean from all genes of *E. coli*. This probability is estimated from 10^5 cohorts assembled from random samples from all genes with replacement, using a non-parametric bootstrap method. The sample size is equal to the size of the cohort being compared with.

Media and chemicals

Measurements were performed in Luria-Bertani (LB) and M9 media (standard and diluted). The chemicals, such as tryptone, sodium chloride, agarose, MEM amino acids (50X), MEM Vitamin solution (100X), Glucose and antibiotic chloramphenicol, etc. were purchased from Sigma Aldrich. Yeast extract was purchased from Lab M (Topley House, Bury, Lancashire, UK). The components of LB medium were 10 g tryptone, 10 g NaCl, and 5 g yeast extract in 1000 mL distilled water. For M9 medium, the components were 1x M9 Salts, 2 mM MgSO4, 0.1 mM CaCl2; 5x M9 Salts with 34 g/L Na2HPO4, 15 g/L KH2PO4, 2.5 g/L NaCl, 5 g/L NH4Cl supplemented with 100X vitamins, 0.2% Casamino acids and 0.4% glucose. We also used '0.5X' and '0.25X' media by diluting the M9 medium to 1:1 and to 1:3 respectively, using autoclaved distilled water [13-16].

Strains and growth conditions

To measure RNA polymerase (RNAP) levels at different medium, we used the RL1314 strain with RpoC endogenously tagged with GFP (generously provided by Robert Landick), which was engineered from the W3110 strain (used here to measure background fluorescence).

To measure single-cell protein levels of genes controlled by tandem promoters, we used genes endogenously tagged with the YFP coding sequence from the YFP fusion library [7]. These were purchased from the *E. coli* genetic stock center (CGSC) of Yale University, U.S.A. (Table B in S3 Appendix), which has wild type MG1655 cells as the reference genome (and thus was used to measure cellular background fluorescence). Measurements of protein levels using this library are expected to be precise for a wide range of expression levels, given evidence for strong correlation in single gene expression levels when measured by RNA-fish, RNA-seq, mass spectrometry and flow cytometry (taken using the YFP library) [7]. The lesser accurate estimations occur for the weakest expressing genes

[7][17], due to their values being near the level of cellular autofluorescence. For this reason as well, we do not consider all of the 30 genes in our analysis as described in the Results section.

From a glycerol stock (-80°C), cells were streaked on LB agar plates with the appropriate antibiotics and incubated at 37°C overnight. From the plates, a single colony was picked, inoculated in LB medium and supplemented with appropriate antibiotics and incubated at 30°C overnight with shaking at 250 rpm. Next, overnight cultures were diluted into freshly prepared tailored media (see 'Media and Chemicals'), with appropriate antibiotics with an O.D$_{600}$ of 0.03 (Optical Density, 600 nm; Ultrospec 10, Amersham biosciences, UK) and allowed to grow at 30°C with shaking at 250 rpm until reaching the mid-exponential phase (O.D$_{600}$ ~0.4-0.5). At this stage, measurements of protein levels were conducted using flow-cytometry and/or microscopy.

Growth curves

Growth curves were measured by O.D$_{600}$ using a spectrophotometer (Ultrospec 10; GE Healthcare). From the overnight culture, cells were diluted (1:10000) into the respective fresh media and allowed to grow while shaking (250 rpm). O.D.'s were recorded for 450 min. every 30 min. We performed 3 biological replicates for each condition. We found negligible variability between replicates. The results shown are the averages and standard error of the mean.

Microscopy and image analysis

When reaching the mid-exponential growth phase, cells were pelleted by centrifugation (10000 rpm for 1 min), and the supernatant was discarded. The pellet was re-suspended in 100 µL of the remaining medium Next, 3 µL of cells were placed in between 2% agarose gel pad and a coverslip and imaged using a confocal microscopy with a 100X objective. The fluorescence was measured with a 488 nm laser and a 514/30 nm emission filter. Phase-contrast images were simultaneously acquired for purposes of segmentation and to assess health, morphology, and physiology.

Using the software *CellAging* [18], from phase contrast images, we segmented cells semi-automatically, correcting errors manually. Next, phase-contrast and corresponding fluorescence images were aligned to extract single-cell fluorescence intensities (example image in Fig 4B). We then performed background subtraction, i.e., from each cell's total fluorescence we subtracted the mean fluorescence of control cells, not expressing YFP.

RNA-seq measurements and data analysis

We searched for correlations between the LFCs over time of genes controlled by tandem promoters ('Tg') and the LFCs over time of their output genes ('Og') as well as their input genes ('Ig').

Given known rates of RNA and protein production and degradation in *E. coli* [7, 19-22], we expect changes in RNA numbers to take at least 60 min. on average, to propagate to protein numbers. Thus, we performed RNA-seq of cells in exponential growth phase at moments ‘0 min’, and then 20 and 180 mins. later. We then calculated LFCs between 0 and 20 min, and between 0 and 180 min.

Specifically, to assess if LFCs in Ig propagate to Tg, we compared changes in Ig between moments 0 and 20, with changes in Tg between moments 0 and 180 min. Similarly, to assess LFCs in Tg propagate to Og, we compared changes in Tg between moments 0 and 20, with changes in Og between moments 0 and 180 min. Results are shown in Panels A and B of Fig D in S2 Appendix.

Sample preparation

For RNA-seq experiments, single colonies of K12 MG1655 cells were picked from LB Agar plates and inoculated into 5 ml of LB medium. Cultures were grown overnight with shaking at 250 rpm. Next, these cultures were diluted to O.D$_{600}$ of 0.05 in fresh LB medium and incubated, with a 250 rpm agitation. RNA-seq was performed over time (0, 20 and 180 min). Total RNA from 3 independent biological replicates in each medium was extracted using RNeasy kit (Qiagen). RNA was treated twice with DNase (Turbo DNA-free kit, Ambion) and quantified using Qubit 2.0 Fluorometer RNA assay (Invitrogen, Carlsbad, CA, USA). Total RNA amounts were determined by gel electrophoresis, using a 1% agarose gel stained with SYBR safe (Invitrogen). RNA was detected using UV with a Chemidoc XRS imager (Biorad).

Sequencing was performed by GENEWIZ, Inc. (Leipzig, Germany). The RNA integrity number (RIN) was obtained with the Agilent 4200 TapeStation (Agilent Technologies, Palo Alto, CA, USA). Ribosomal RNA depletion was performed using Ribo-Zero Gold Kit (Bacterial probe) (Illumina, San Diego, CA, USA). RNA-seq libraries were constructed using NEBNext Ultra RNA Library Prep Kit (NEB, Ipswich, MA, USA). Sequencing libraries were multiplexed and clustered on 1 lane of a flow-cell. Samples were sequenced using a single-index, 2x150 bp paired-end (PE) configuration on an Illumina HiSeq instrument. Image analysis and base calling were conducted with HiSeq Control Software (HCS). Raw sequence data (.bcl files) were converted into fastq files and de-multiplexed using Illumina bcl2fastq v.2.20. One mismatch was allowed for index sequence identification.

Data analysis

RNA-seq data analysis pipeline was: i) RNA sequencing reads were trimmed with Trimmomatic [23] v.0.39 to remove possible adapter sequences and nucleotides with poor quality. ii) Trimmed reads were mapped to the reference genome, *E. coli* MG1655 (NC_000913.3), using the using the STAR aligner v.2.5.2b, which outputs BAM files [24]. iii) Then, ‘*featureCounts*’ from the Rsubread R package v.1.34.7 was used to calculate unique gene hit counts [25]. iv) These counts were used for the differential expression analysis. Genes with less than 5 counts in more than 3 samples, and genes whose mean counts are less than 10 were removed from further analysis. We used the DESeq2 R package v.1.24.0 [26] to compare gene expression between groups of samples and calculate p-values and $\log_2$ of fold changes using Wald tests (function *nbinomWaldTest*). P-values were adjusted for multiple hypotheses testing (Benjamini–Hochberg, BH procedure, [27]).

Flow-cytometry and data analysis

We measured single-cell fluorescence using a ACEA NovoCyte Flow Cytometer (ACEA Biosciences Inc., San Diego, USA). Upon reaching the mid-exponential phase (OD~0.4-0.5), cells were diluted (1:10000) into 1 mL of phosphate buffer saline (PBS) solution and vortexed for 5 s. For a single run, 50000 events were collected at a flow rate of 14 µL/minute and a core diameter of 7.7 mm using the Novo Express software using a blue laser (488 nm) for excitation. We obtained the height of the fluorescein isothiocyanate channel (FITC-H) (530/30 nm filter). A PMT voltage of 600 volts was set for FITC. To avoid background signal from particles smaller than bacteria, the detection threshold was set to 5000 for FSC-H analyses. Three biological replicates were performed per condition.

We applied unsupervised gating [28] to the flow-cytometry data, setting the fraction of single-cell events used in the analysis, α, to 0.99. We proved to be enough to remove non-cell events due to debris, doublets, fragments, cell clumps, and other undesired events. Reducing α did not change the results qualitatively.

To remove outliers from the flow-cytometry distributions, we applied secondary gating. In detail, we sorted the data based on FITC-H values and calculated the difference between consecutive samples. Then, we obtained the indices of those differing by more than 10000 (approximately 10 times the mean fluorescent level observed). Next, we obtained the minimum of those indices to define the upper bound. Finally, values above this index were considered an outlier and discarded. In all measurements, never more than 10000 events were discarded, thus, more than 40000 were used for the analysis.

Subtraction of background fluorescence from total protein fluorescence in flow-cytometry

First, we collected mean background fluorescence from distributions of cells not carrying YFP. Then we measured the distributions of fluorescence of cells carrying the protein tagged with YFP. Having this, the protein fluorescence ‘g’ of a gene is obtained by subtracting mean background fluorescence ‘bg’ from the (total ‘T’) measured fluorescence. For the mean (M) protein fluorescence from a cell population, we write:

$$M\left(g\right)=M(T)-M(bg) \tag{1}$$

Similarly, the variance ‘Var’ is obtained by:

$$Var(g)=Var(T)-Var(bg) \tag{2}$$

The CV^2 of the distribution protein fluorescence of a gene after background subtraction is:

$$CV^2(g) = \frac{Var(g)}{M(g)^2} \tag{3}$$

Finally, the third moment of protein fluorescence and the skewness after background subtraction are given by:

$$\mu_3(g) = \mu_3(T) - \mu_3(bg) \tag{4}$$

$$S(g) = \frac{\mu_3(g)}{Var(g)^{\frac{3}{2}}} \tag{5}$$

After background subtraction, any genes with negative means, variance or third moment, will not be included in the data (except in Fig F in S2 Appendix for illustrative purposes).

Conversion of protein fluorescence into protein numbers

To convert protein fluorescence into protein numbers, we made a correlation plot between the mean protein fluorescence measured in our lab (after background subtraction) and the mean protein numbers reported in [7] for the same genes. We fitted a line to the data points by forcing the intercept with the Y axis to be at zero. The slope of the fitted line is used as a scaling factor (~0.09) with an R^2 value of 0.68 (Fig 4D). For protein fluorescence to protein numbers correction only the mean gets changed whereas the normalised moments CV^2 and S remain unchanged.

Analytical model of mean RNA levels controlled by a single promoter in the absence of a closely spaced promoter

From Reactions 1c1 and 1a4 in the main manuscript, for an isolated promoter, one would have:

$$P_{free} \underset{k_{unbind}}{\overset{k_{bind} \cdot [R]}{\rightleftarrows}} P_{occupied} \xrightarrow{k_{after}} P_{free} + R_{elong} \tag{6}$$

$$R_{elong} \xrightarrow{k_{elong}} RNA \tag{7}$$

At steady state $P_{occupied}$ is:

$$\frac{dP_{occupied}}{dt} = P_{free} \times k_{bind} \cdot [R] - P_{occupied} \times \left(k_{unbind} + k_{after}\right) = 0 \tag{8}$$

$$P_{free} = P_{occupied} \cdot \frac{\left(k_{unbind} + k_{after}\right)}{k_{bind} \cdot \left[R\right]} \qquad (9)$$

Since necessarily:

$$P_{free} + P_{occupied} = 1 \qquad (10)$$

From equations 9 and 10:

$$P_{occupied} \cdot \left(1 + \frac{k_{unbind} + k_{after}}{k_{bind} \cdot \left[R\right]}\right) = 1 \qquad (11)$$

$$P_{occupied} = \frac{k_{bind} \cdot \left[R\right]}{k_{bind} \cdot \left[R\right] + k_{unbind} + k_{after}} \qquad (12)$$

Note that, by definition (main manuscript, equations 6a and 6b), the fraction of time that an RNAP is bound to the promoter, *ω*, should equal $P_{occupied}$ in (12). Meanwhile, at steady state, R_{elong}becomes:

$$\frac{dR_{elong}}{dt} = P_{occupied} \times k_{after} - R_{elong} \times k_{elong} = 0 \qquad (13)$$

$$R_{elong} = \frac{P_{occupied} \times k_{after}}{k_{elong}} \qquad (14)$$

From equations 12 and 14:

$$R_{elong} = \frac{k_{bind} \cdot \left[R\right]}{k_{bind} \cdot \left[R\right] + k_{unbind} + k_{after}} \times \frac{k_{after}}{k_{elong}} \qquad (15)$$

At steady state, the mean *RNA* numbers, M_{RNA}, is:

$$\frac{dM_{RNA}}{dt} = R_{elong} \times k_{elong} - M_{RNA} \times k_{rd} = 0 \qquad (16)$$

From equations 15 and 16s:

$$M_{RNA} = \frac{k_{bind} \cdot [R]}{k_{bind} \cdot [R] + k_{unbind} + k_{after}} \times \frac{k_{after}}{k_{elong}} \times \frac{k_{elong}}{k_{rd}} \tag{S7}$$

$$M_{RNA} = \frac{k_{bind} \cdot [R]}{k_{bind} \cdot [R] + k_{unbind} + k_{after}} \times \frac{k_{after}}{k_{rd}} \tag{18}$$

From S18, the RNA numbers at steady state do not depend on k_{elong}.

Derivation of mean protein numbers at steady state produced by a pair of tandem promoters

For the upstream promoter, from (1c1), (1a3), and (1a4) in the main manuscript, at steady state:

$$\frac{d(RNA)}{dt} = R^{u}_{elong} \times k^{u}_{elong} \cdot (1 - \omega_d \cdot f) - RNA \times k_{rd} = 0 \tag{19}$$

From this and equation 6b in the main manuscript:

$$RNA = \frac{k^{u}_{bind} \cdot [R]}{k^{u/d}_{occlusion} + k^{u}_{bind} \cdot [R] + k^{u}_{unbind} + k^{u}_{after}} \times \frac{k^{u}_{after} \cdot (1 - \omega_d \cdot f)}{k_{rd}} \tag{20}$$

Meanwhile, for the downstream promoter, from reactions (2a1), (2a2), and (2a3) in the main manuscript, at steady state:

$$\frac{d(RNA)}{dt} = R^{d}_{elong} \times k^{d}_{elong} - RNA \times k_{rd} = 0 \tag{21}$$

$$RNA = \frac{k^{d}_{bind} \cdot [R]}{k^{d/u}_{occlusion} + k^{d}_{occupy} + k^{d}_{bind} \cdot [R] + k^{d}_{unbind} + k^{d}_{after}} \times \frac{k^{d}_{after}}{k_{rd}} \tag{22}$$

Having this, since at steady state the RNA numbers produced by a pair of tandem promoters should equal the sum of RNA numbers from the upstream (S20) and downstream (S22) promoters, we have:

$$M_{RNA} = \left(\begin{array}{c} \dfrac{k_{bind}^{u} \cdot [R] \times k_{after}^{u} \cdot (1 - \omega_d \cdot f)}{k_{occlusion}^{u/d} + k_{bind}^{u} \cdot [R] + k_{unbind}^{u} + k_{after}^{u}} + \\ \dfrac{k_{bind}^{d} \cdot [R] \times k_{after}^{d}}{k_{occlusion}^{d/u} + k_{occupy} + k_{bind}^{d} \cdot [R] + k_{unbind}^{d} + k_{after}^{d}} \end{array} \right) \cdot \frac{1}{k_{rd}} \tag{23}$$

Thus, the mean protein numbers is:

$$M_P = M_{RNA} \cdot \frac{k_p}{k_{pd}} \tag{24}$$

If the upstream and downstream promoters have similar strengths, i.e., if $k_{bind}^{d} \approx k_{bind}^{u}$, $k_{unbind}^{d} \approx k_{unbind}^{u}$, and $k_{after}^{d} \approx k_{after}^{u}$, we can expect that: $\omega_d = \omega_u$, $k_{occlusion}^{d/u} = k_{occlusion}^{u/d}$. If so, the equation above becomes:

$$M_P = \left(\begin{array}{c} \dfrac{k_{bind} \cdot [R] \times k_{after} \cdot (1 - \omega_d \cdot f)}{k_{occlusion} + k_{bind} \cdot [R] + k_{unbind} + k_{after}} + \\ \dfrac{k_{bind} \cdot [R] \times k_{after}}{k_{occlusion} + k_{occupy} + k_{bind} \cdot [R] + k_{unbind} + k_{after}} \end{array} \right) \cdot \frac{k_p}{k_{rd} \cdot k_{pd}} \tag{25}$$

Here, the symbols “u” and “d” are removed, as they no longer imply potentially different amounts. Having this, let k_r be the effective transcription rate constant of a pair of tandem proteins. It should equal:

$$k_r = \left(\begin{array}{c} \dfrac{k_{bind} \cdot [R] \times k_{after} \cdot (1 - \omega_d \cdot f)}{k_{occlusion} + k_{bind} \cdot [R] + k_{unbind} + k_{after}} + \\ \dfrac{k_{bind} \cdot [R] \times k_{after}}{k_{occlusion} + k_{occupy} + k_{bind} \cdot [R] + k_{unbind} + k_{after}} \end{array} \right) \tag{26}$$

Thus, from equation 25 and 26:

$$M_P = \frac{k_r \cdot k_p}{k_{rd} \cdot k_{pd}} \tag{27}$$

CV^2 and skewness of the distribution of single-cell protein numbers of model tandem promoters

The distributions of protein numbers in *E. coli* cells, can, in general, be well approximated by a Gamma or by a negative binomial distribution [7]. We assume here a negative binomial distribution. For a given number of events, if *r* is the number of failures, *p* is the probability of success per event, and an 'event' is an attempt to produce a protein, then the mean, variance, and skewness of the single-cell distribution of protein numbers should equal:

$$M_P = \frac{pr}{1-p} \tag{28}$$

$$Var_P = \frac{pr}{(1-p)^2} \tag{29}$$

$$S_P = \frac{1+p}{\sqrt{pr}} \tag{30}$$

The relationship between the mean, CV^2 could be written as:

$$CV_P^2 = \frac{1}{M_P}\cdot\left(\frac{Var_P}{M_P}\right) \tag{31}$$

Substituting (S28) and (S29) in (S31)

$$CV_P^2 = \frac{\left(\frac{1}{1-p}\right)}{M_P} \tag{32}$$

Rewriting the above equation by assuming a scaling factor C_1 as:

$$C_1 = \frac{1}{1-p} \tag{33}$$

$$CV_P^2 = \frac{C_1}{M_P} \tag{34}$$

Taking $\log_{10}$ on both sides

$$\log_{10}\left(CV_P^2\right) = \log_{10}(C_1) - \log_{10}\left(M_P\right) \tag{35}$$

From [17], C_1 is approximated as

$$C_1 = \frac{M_P}{M_{RNA}} \cdot \frac{\frac{1}{\tau_p}}{\frac{1}{\tau_p} + \frac{1}{\tau_{RNA}}} \tag{36}$$

$\tau_p = \dfrac{1}{k_{pd}}$ and $\tau_{RNA} = \dfrac{1}{k_{rd}}$ are the lifetimes of proteins and RNAs, respectively. The above equation is rewritten as:

$$C_1 = \frac{k_p}{k_{pd}} \cdot \frac{k_{pd}}{k_{pd} + k_{rd}} \tag{37}$$

$$C_1 = \frac{k_p}{k_{pd} + k_{rd}} \tag{38}$$

From (S28) and (S30), the relationship between the mean, skewness could be written as:

$$S_P = \frac{\frac{1+p}{\sqrt{1-p}}}{\sqrt{M_P}} \tag{39}$$

The equation can be rewritten assuming constant C_2 as:

$$C_2 = \frac{1+p}{\sqrt{1-p}} \tag{40}$$

$$S_P = \frac{C_2}{\sqrt{M_P}} \tag{41}$$

Taking $\log_{10}$ on both sides

$$\log_{10}\left(S_P\right) = \log_{10}\left(C_2\right) - \frac{1}{2} \cdot \log_{10}\left(M_P\right) \tag{42}$$

The constants C_1 and C_2 are related as follows. From equation 33:

$$p = 1 - \frac{1}{C_1} \tag{43}$$

Inserting S43 in S40:

$$C_2 = \frac{2 - \frac{1}{C_1}}{\sqrt{\frac{1}{C_1}}} \tag{44}$$

The equation can be rewritten as

$$C_2 = 2\sqrt{C_1} - \frac{1}{\sqrt{C_1}} \tag{45}$$

Stochastic simulations for the step inference model

Stochastic simulations of the models were done using the stochastic gene network simulator SGNS2 [29]. These stochastic models were compared to the analytical solutions to assess how much variability can there be in $k_{bind} \cdot \left[R\right]$ without the analytical solution deviating too much.

First, to compare analytical and stochastic solutions, we set d_{TSS} between 0 and 180 with an increment of 30. For each d_{TSS}, we calculated the occlusion rate constant ($k_{occlusion}$) for upstream and downstream promoters (Equations 5a and 5b in the main manuscript). The other parameters are listed in Tables 2 and 3 in the main manuscript. To obtain protein numbers at steady state, we have set the simulation time to 10^5 seconds and performed 1000 runs per condition. From these runs, for each condition, we calculated the mean, CV^2 and skewness, along with their standard errors using bootstrapping (10^4 resampling with replacement). Additional runs would slightly decrease the deviation between the two solutions.

References

1. Shearwin KE, Callen BP, Egan JB. Transcriptional interference--a crash course. Trends Genet. 2005;21: 339–345. doi: 10.1016/j.tig.2005.04.009
2. Santos-Zavaleta A, Salgado H, Gama-Castro S, Sánchez-Pérez M, Gómez-Romero L, Ledezma-Tejeida D, et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. Nucleic Acids Res. 2019;47: D212–D220. doi:10.1093/nar/gky1077

3. Crampton N, Bonass WA, Kirkham J, Rivetti C, Thomson NH. Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. Nucleic Acids Res. 2006;34: 5416–5425. doi:10.1093/nar/gkl668
4. Ward DF, Murray NE. Convergent transcription in bacteriophage λ: Interference with gene expression. J Mol Biol. 1979;133: 249–266. doi:10.1016/0022-2836(79)90533-3
5. Lewin B. Genes IX. 9th ed. Sudbury, Mass: Jones and Bartlett Publishers; 2008.
6. Turnbough CL Jr. Regulation of bacterial gene expression by transcription attenuation. Microbiol Mol Biol Rev. 2019;83. doi:10.1128/MMBR.00019-19
7. Taniguchi Y, Choi PJ, Li G-W, Chen H, Babu M, Hearn J, et al. Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. Science. 2010;329: 533–538. doi:10.1126/science.1188308
8. Xavier JC, Gerhards RE, Wimmer JLE, Brueckner J, Tria FDK, Martin WF. The metabolic network of the last bacterial common ancestor. Commun Biol. 2021;4: 413. doi:10.1038/s42003-021-01918-4
9. Winter D. rentrez: An R package for the NCBI eUtils API. R J. 2017;9: 520. doi:10.32614/rj-2017-058
10. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019;47: D419–D426. doi:10.1093/nar/gky1038
11. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13: 2498–2504. doi:10.1101/gr.1239303
12. Doncheva NT, Assenov Y, Domingues FS, Albrecht M. Topological analysis and interactive visualization of biological networks and protein structures. Nat Protoc. 2012;7: 670–685. doi:10.1038/nprot.2012.004
13. Lloyd-Price J, Startceva S, Kandavalli V, Chandraseelan JG, Goncalves N, Oliveira SMD, et al. Dissecting the stochastic transcription initiation process in live Escherichia coli. DNA Res. 2016;23: 203–214. doi:10.1093/dnares/dsw009
14. Kandavalli VK, Tran H, Ribeiro AS. Effects of σ factor competition are promoter initiation kinetics dependent. Biochim Biophys Acta. 2016;1859: 1281–1288. doi:10.1016/j.bbagrm.2016.07.011
15. Startceva S, Kandavalli VK, Visa A, Ribeiro AS. Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2019;1862: 119–128. doi:10.1016/j.bbagrm.2018.12.005
16. Oliveira SMD, Goncalves NSM, Kandavalli VK, Martins L, Neeli-Venkata R, Reyelt J, et al. Chromosome and plasmid-borne P LacO3O1 promoters differ in sensitivity to critically low temperatures. Sci Rep. 2019;9: 1–15. doi:10.1038/s41598-019-39618-z
17. Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, et al. Noise in protein expression scales with natural protein abundance. Nat Genet. 2006;38: 636–643. doi:10.1038/ng1807

18. Häkkinen A, Muthukrishnan A-B, Mora A, Fonseca JM, Ribeiro AS. CellAging: a tool to study segregation and partitioning in division in cell lineages of Escherichia coli. Bioinformatics. 2013;29: 1708–1709. doi:10.1093/bioinformatics/btt194
19. Bernstein JA, Khodursky AB, Lin P-H, Lin-Chao S, Cohen SN. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. Proc Natl Acad Sci U S A. 2002;99: 9697–9702. doi:10.1073/pnas.112318199
20. Balleza E, Kim JM, Cluzel P. Systematic characterization of maturation time of fluorescent proteins in living cells. Nat Methods. 2018;15: 47–51. doi:10.1038/nmeth.4509
21. Hebisch E, Knebel J, Landsberg J, Frey E, Leisner M. High variation of fluorescence protein maturation times in closely related Escherichia coli strains. PLoS One. 2013;8: e75991. doi:10.1371/journal.pone.0075991
22. Maurizi MR. Proteases and protein degradation in Escherichia coli. Experientia. 1992;48: 178–201. doi:10.1007/BF01923511
23. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170
24. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2012;29: 15–21. doi:10.1093/bioinformatics/bts635
25. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 2019;47: e47. doi:10.1093/nar/gkz114
26. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550. doi:10.1186/s13059-014-0550-8
27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc. 1995;57: 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x
28. Razo-Mejia M, Barnes SL, Belliveau NM, Chure G, Einav T, Lewis M, et al. Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction. Cell Syst. 2018;6: 456-469.e10. doi:10.1016/j.cels.2018.02.004
29. Lloyd-Price J, Gupta A, Ribeiro AS. SGNS2: a compartmentalized stochastic chemical kinetics simulator for dynamic cell populations. Bioinformatics. 2012;28: 3004–3005. doi:10.1093/bioinformatics/bts556

S2 Appendix: Supporting Figures

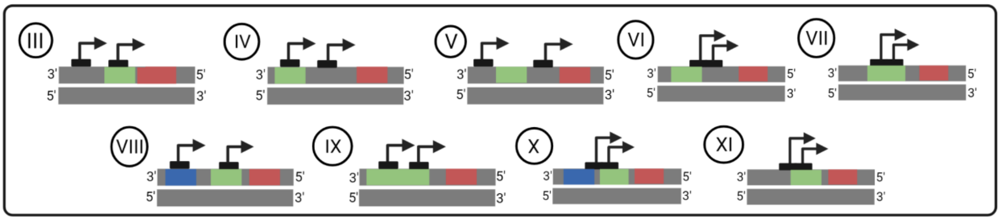


Fig A. Other arrangements of tandem promoters in *E. coli.* Unlike the arrangements I and II in Fig 1 in the main manuscript, the arrangements here (III-XI) allow for overlaps with or in between other gene(s). The red, green, and blue rectangles are DNA regions coding for RNA. These arrangements are not considered in this study. Figure created with BioRender.com.

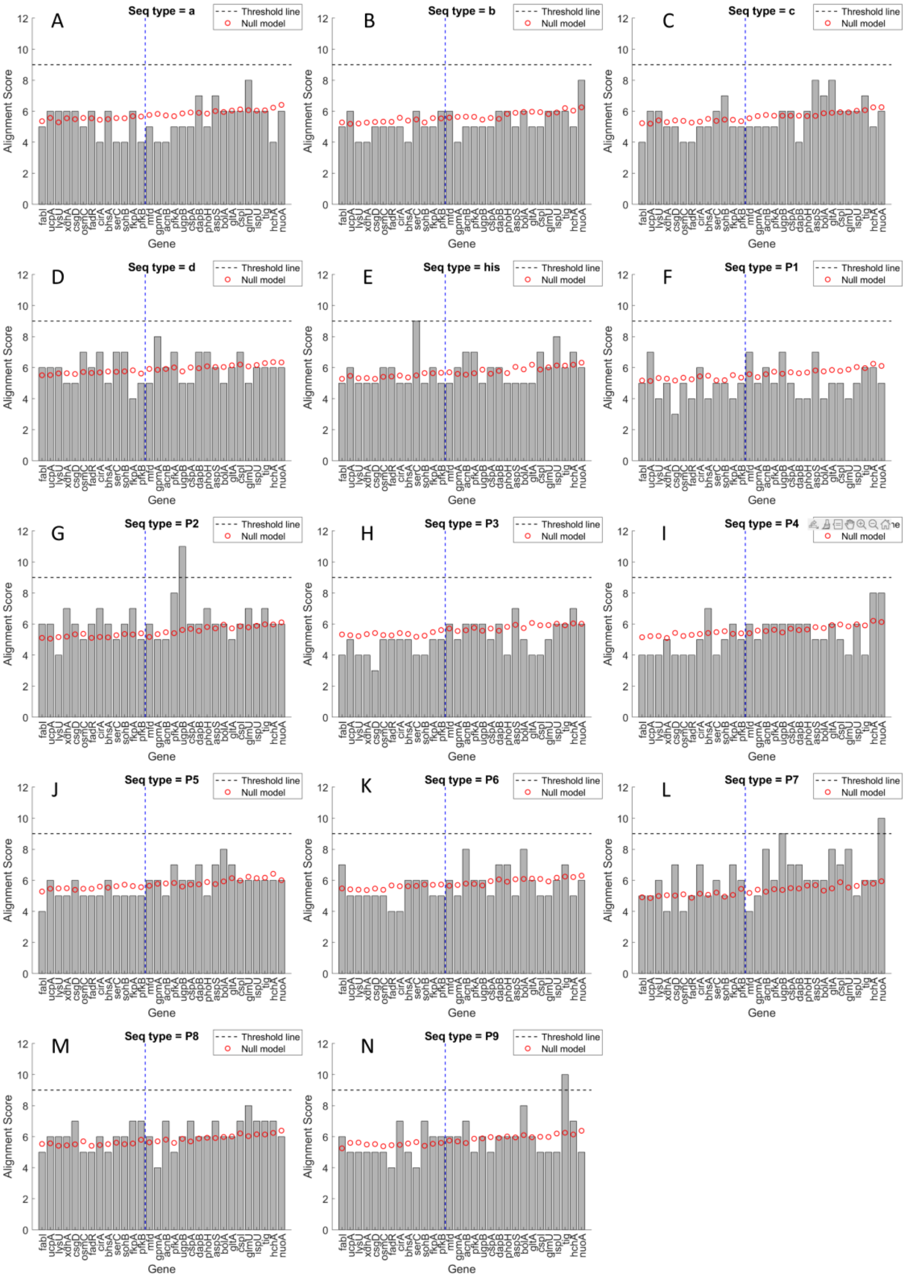


Fig B. Local alignment scores. Local alignment scores between known pause sequences and the sequences in between the tandem promoter regions (grey bars). Also shown by red circles are the alignment scores between each pause sequence and randomly generated sequences with the same d_{TSS} as the natural genes. The minimum alignment score to be considered significant is shown by a dashed black line. Finally, the blue vertical dashed line at d_{TSS} = 35 bp shows the separation between genes subject to occlusion or not.

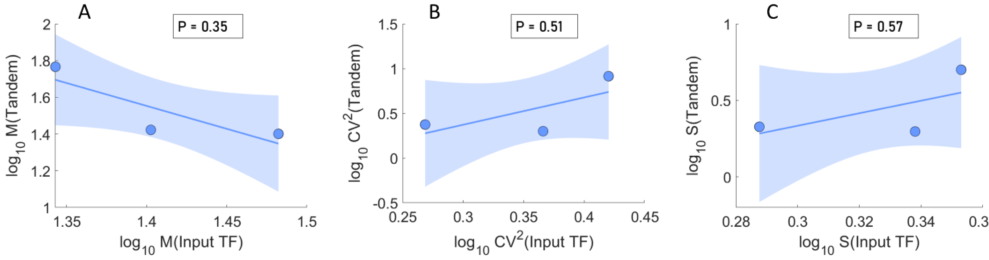


Fig C. Correlation of the moments of the single-cell protein numbers between genes and their input TFs. Scatter plots between the moments of the single-cell protein numbers (in $\log_{10}$ scale) of genes regulated by tandem promoters ('Tandem') and their input TFs. (A) Mean, (B) CV^2, and (C) Skewness. The blue line is the best linear fit, and its shadow is the standard error of the fit. The p-value, *P* is the probability that the slope of the line equals 0. If P < 0.05, there is a statistically significant correlation. The genes used in these results are listed in Table E in S3 Appendix. The axes differ widely in scales between the figures to facilitate visualization of the relationships.

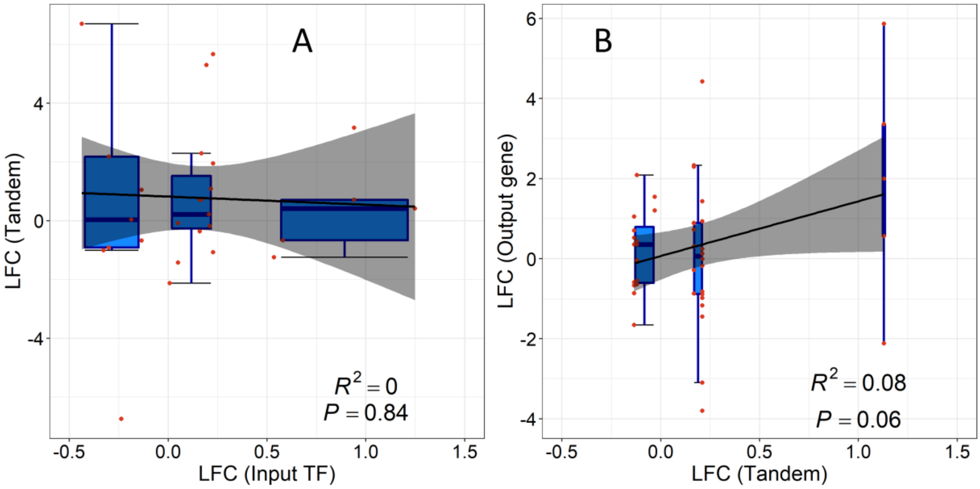


Fig D. Correlation of RNA fold changes of genes and their input TFs. Correlation plots between the LFCs of the RNA numbers of genes controlled by tandem promoters with their input and output genes. (A) LFCs (from 0 to 20 min) of 29 genes expressing input TFs plotted against the corresponding LFCs (from 0 to 180 min) of the genes controlled by tandem promoters. (B) LFCs (from 0 to 20 min) of genes controlled by tandem promoters plotted against the corresponding LFCs of their output genes (from 0 to 180 min). A total of 43 TF-gene interactions were analysed. RNA-seq measurements described in section "RNA-seq Measurements and Analysis in S1 Appendix". The black line is the best linear fit and the grey shadow area is the standard error of the fit. The blue horizontal lines inside the boxes are the median, the top of the boxes are the 3rd quartile (Q3) and the bottom of the boxes are the first quartile (Q1). The error bars at the top and bottom range from (Q3+1.5*IQR) to (Q1-1.5*IQR), with an interquartile range: IQR = Q3 – Q1. The three box plots correspond to the data points with LFCs < 0, LFC between 0 and 0.5, and LFC > 0.5. Related to Table E and F in S3 Appendix.

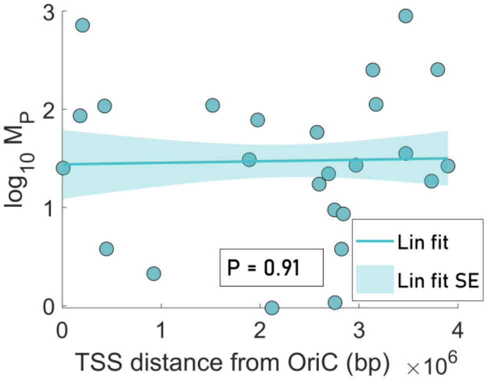


Fig E. Relationship between expression levels of the genes controlled by tandem promoters and the distance in nucleotides (bp) from the upstream promoter and the Oric region in the DNA. Data from 25 genes for the 1X condition. Also shown in a linear fit and the corresponding 1 standard error of the fit (shadow area). The p-value, *P*, is the probability that the slope of the line equals 0.

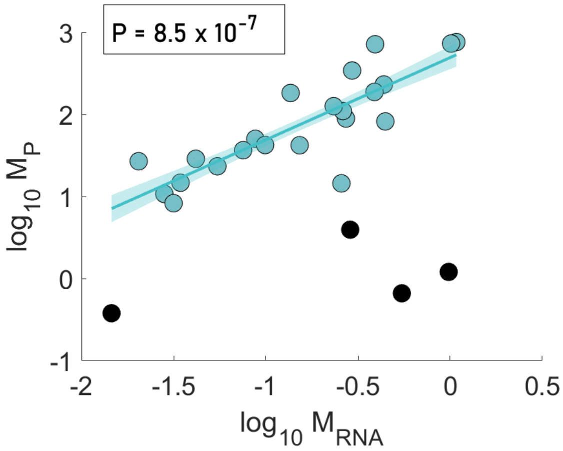


Fig F. Correlation plot between the mean single-cell RNA levels (*M_{RNA}*) and the mean single-cell protein numbers (*M_P*). Both data are obtained from Ref. [28] in main manuscript and are processed to include only genes controlled by tandem promoters (classes I and II, Table H in S3 Appendix). The line is the best linear fit to the data, and its shadow area is the standard error of the fit. The p-value, *P* is the probability that the slope of the line equals 0. Since P < 0.05, we conclude that *M_{RNA}* and *M_P* are significantly correlated. The black balls correspond to 4 genes that were not considered when fitting the line, due to being outliners. In our own data, cells carrying these same 4 genes exhibited a fluorescence that was equal or lower than the cellular background fluorescence in either 1X or 0.5X media.

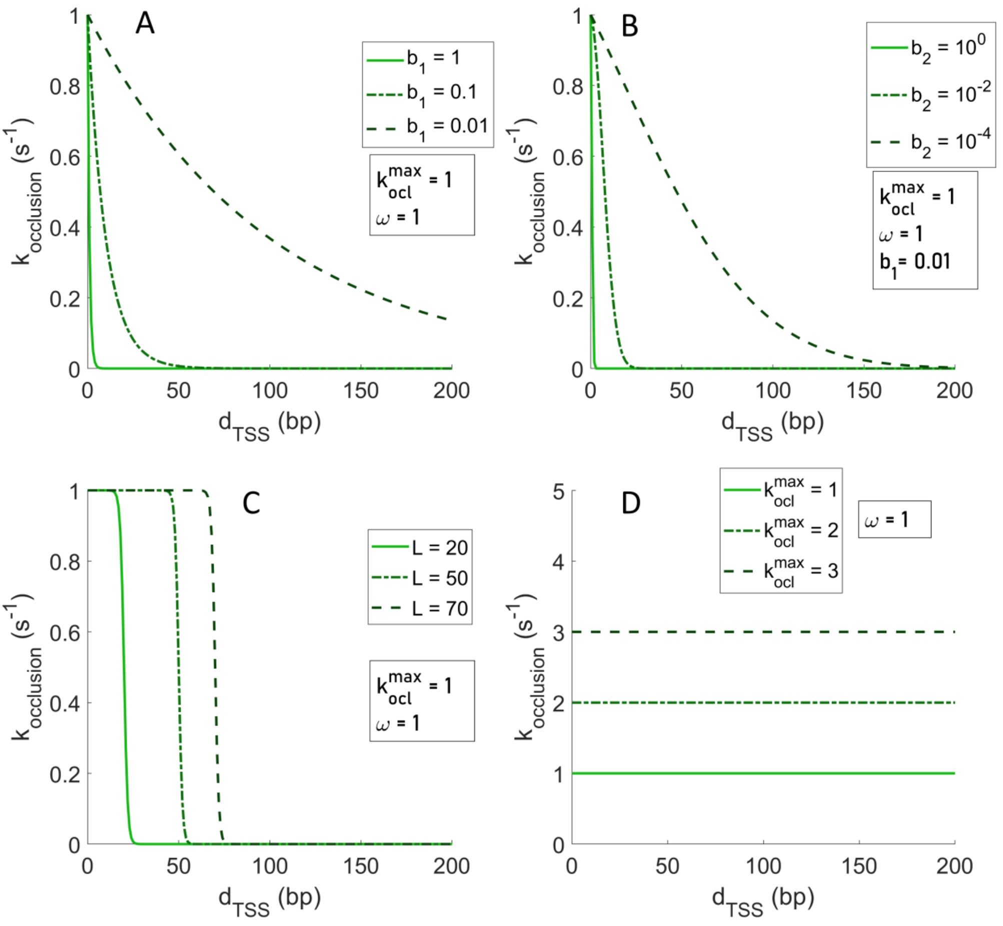


Fig G. Models of transcription interference. Models of transcription interference between RNAPs in tandem promoters as a function of the d_{TSS} between them. (A) ‘Exponential 1’ as a function for different values of ‘b_1’. (B) ‘Exponential 2’ as a function at different values of ‘b_2. (C) Continuous ‘step-like’ function for different values of ‘*L*’ (which is the d_{TSS} at which the step occurs). (D) Zero order polynomial for different values of $k_{ocl}^{\max}$. See Table 1 in the main manuscript for the definitions of these models and variables within.

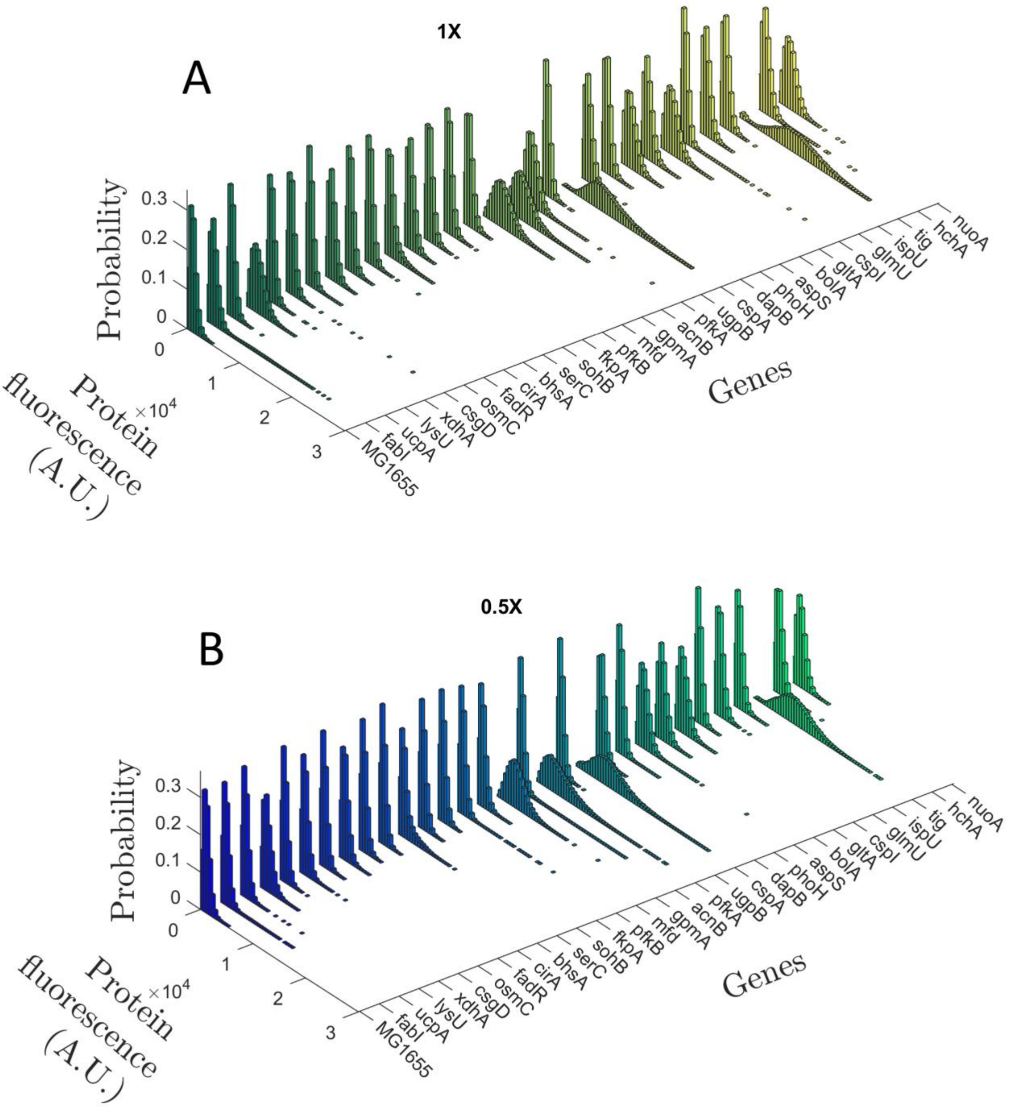


Fig H. Protein florescence distributions. Protein florescence distributions of genes controlled by tandem promoters measured by flow-cytometry**.** Each protein is tagged with a YFP (YFP strain library). Only 1 of 3 biological replicates is shown per gene. (A) M9 medium (1X). (B) Diluted M9 medium (0.5X). 'MG1655' are control cells, not carrying YFP. Protein fluorescence is shown in arbitrary units.

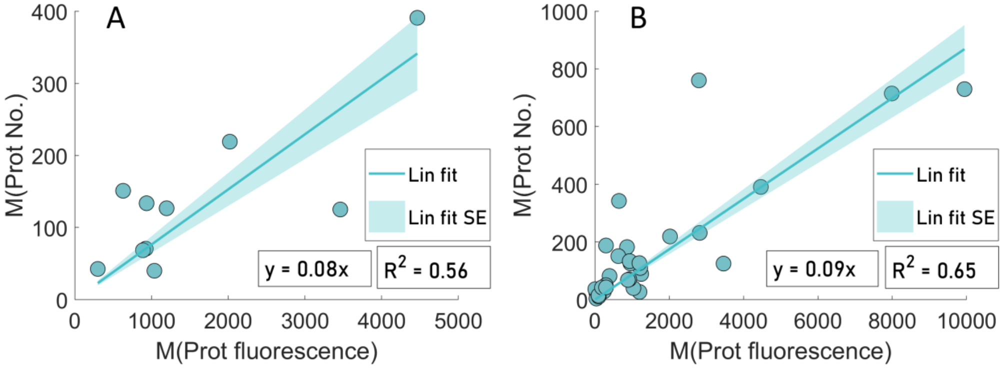


Fig I. Estimation of scaling factors using data from genes controlled by single promoters. A) Mean single-cell protein fluorescence (own measurements of genes controlled by single promoters) plotted against the corresponding mean single-cell protein numbers reported in [28]. From the equation of the best fitting line without y-intercept (y-intercept = 0), we obtained a scaling factor, s_f, equal to 0.08. B) Same as (A) but the own measurements are of both single promoters and tandem promoters, merged. From the equation of the best fitting line without y-intercept (y-intercept = 0), we obtained a scaling factor, s_f, equal to 0.09.

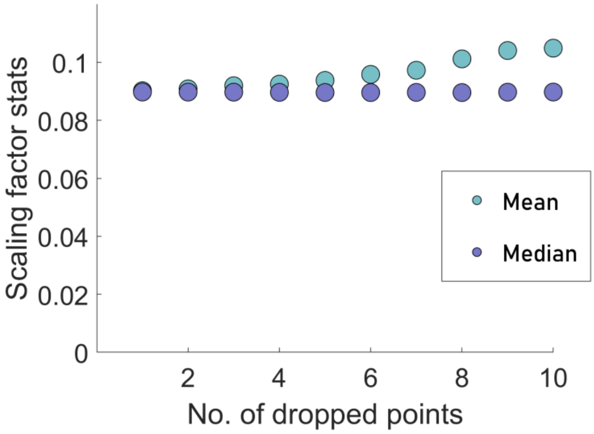


Fig J. Sensitivity test. Mean and median of scaling factor varies as a function of number of data points randomly dropped.

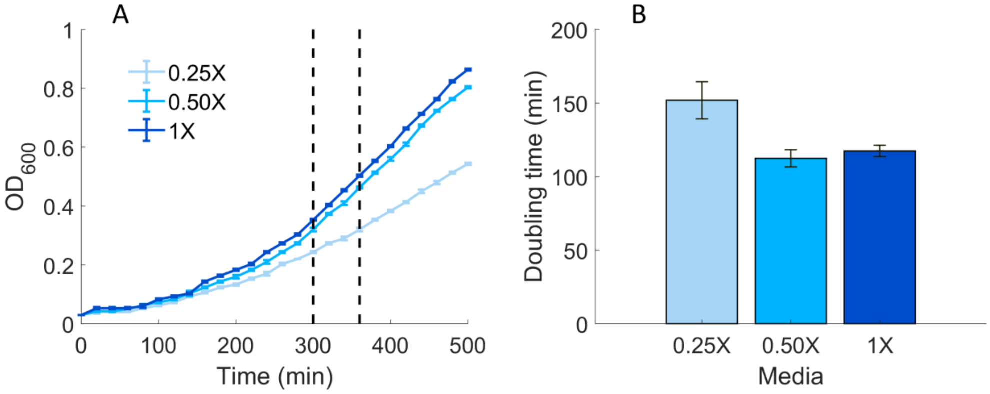


Fig K. Growth curves and doubling times. A. Optical density (OD_{600}) curves of *E. coli* MG1655 cells grown in 0.25X,0.5X and 1X media (section 'Media and Chemicals' in S1 Appendix). B. From these curves, the doubling time was estimated to be ~112 min in 0.5X and ~118 min in 1X. We used 115 min doubling time in the models. The estimation is made using the formula

$$D = \frac{\ln(2)}{\ln\left(\dfrac{OD(t_2)}{OD(t_1)}\right)} \times (t_2 - t_1)$$

, with t_2 and t_1 being the end and start times (in minutes), respectively. They are marked by two vertical dashed black lines. The error bars denote the standard error of the mean. Ref. [28] in main manuscript reported ~150 min using 96 well-plates in the same conditions. The fact that we used culture tubes may explain the difference.

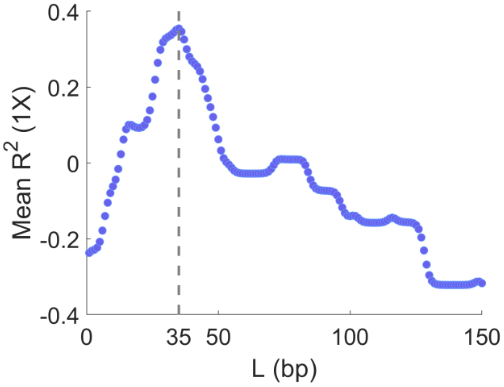


Fig L. Mean R^2 of the step interference model. Mean R^2 of the step interference model to the 1X data in Fig 6A, 6B, and 6C, as a function of L (d_{TSS} at which the step of the step function occurs). The Mean

R^2 is visibly maximized at L = 35, which marked by a grey dashed line. Relates to Fig 6 in the main manuscript.

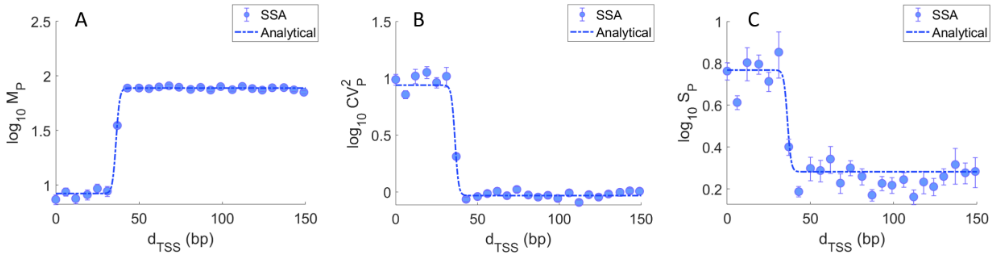


Fig M. Confronting the solutions of the analytical and stochastic model. (A) $\log_{10}$ of mean protein numbers, (B) $\log_{10}$ of CV^2 of protein numbers and (C) $\log_{10}$ of Skewness of protein numbers as a function of d_{TSS}. The blue line is the analytical solution of the step model. The blue dots are the mean results of stochastic simulations of the step model. The parameters used are shown in Tables 2 and 3 in the main manuscript. See Section 'Stochastic simulations for the step interference model' in S1 Appendix.

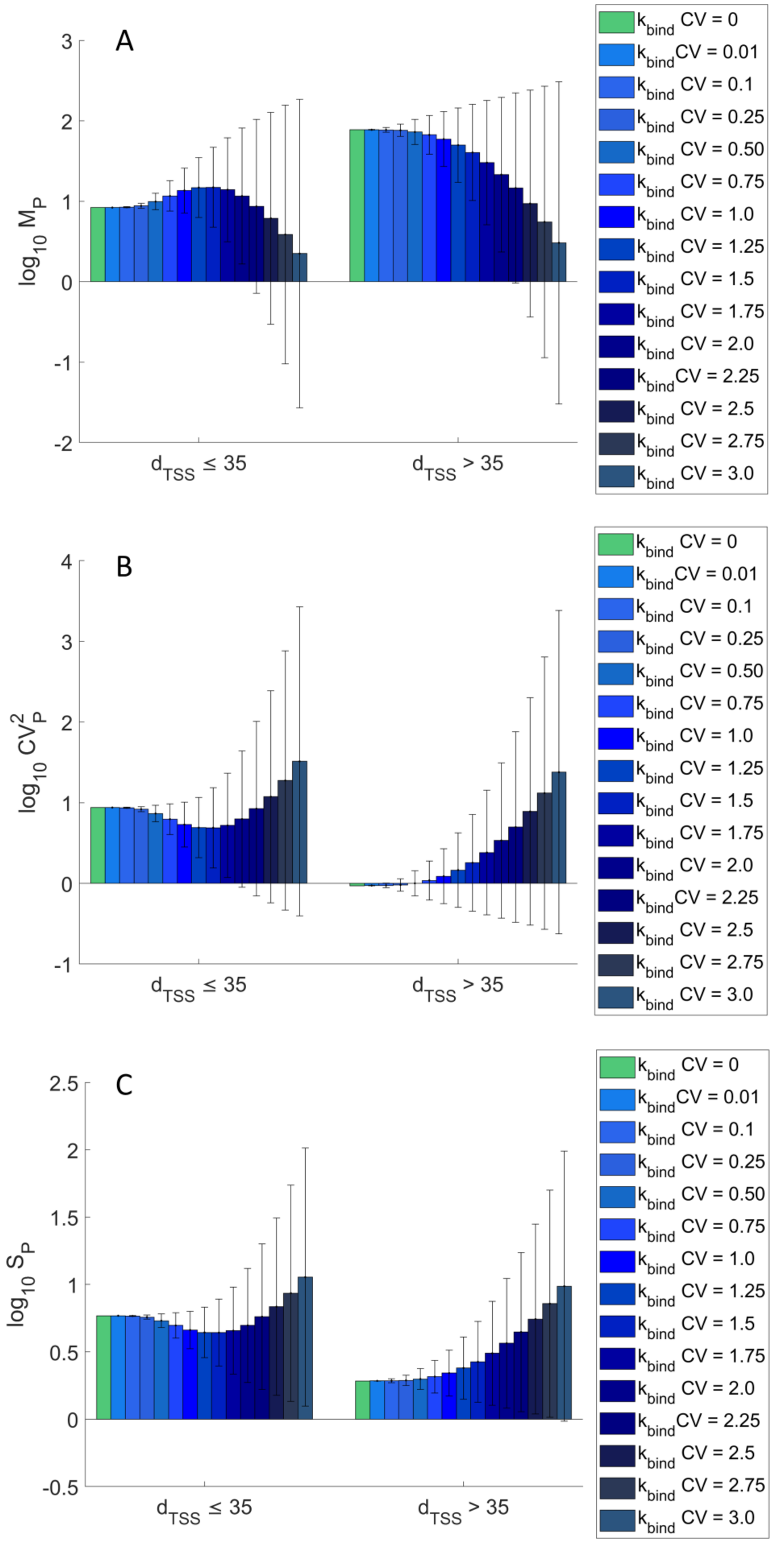

A
$\log_{10} M_P$
3
2
1
0
-1
-2
$d_{TSS} \leq 35$
$d_{TSS} > 35$
k_{bind} CV = 0
k_{bind}CV = 0.01
k_{bind} CV = 0.1
k_{bind} CV = 0.25
k_{bind} CV = 0.50
k_{bind} CV = 0.75
k_{bind} CV = 1.0
k_{bind} CV = 1.25
k_{bind} CV = 1.5
k_{bind} CV = 1.75
k_{bind} CV = 2.0
k_{bind}CV = 2.25
k_{bind} CV = 2.5
k_{bind} CV = 2.75
k_{bind} CV = 3.0
B
$\log_{10} CV^2_P$
4
3
2
1
0
-1
$d_{TSS} \leq 35$
$d_{TSS} > 35$
k_{bind} CV = 0
k_{bind}CV = 0.01
k_{bind} CV = 0.1
k_{bind} CV = 0.25
k_{bind} CV = 0.50
k_{bind} CV = 0.75
k_{bind} CV = 1.0
k_{bind} CV = 1.25
k_{bind} CV = 1.5
k_{bind} CV = 1.75
k_{bind} CV = 2.0
k_{bind}CV = 2.25
k_{bind} CV = 2.5
k_{bind} CV = 2.75
k_{bind} CV = 3.0
C
$\log_{10} S_P$
2.5
2
1.5
1
0.5
0
-0.5
$d_{TSS} \leq 35$
$d_{TSS} > 35$
k_{bind} CV = 0
k_{bind}CV = 0.01
k_{bind} CV = 0.1
k_{bind} CV = 0.25
k_{bind} CV = 0.50
k_{bind} CV = 0.75
k_{bind} CV = 1.0
k_{bind} CV = 1.25
k_{bind} CV = 1.5
k_{bind} CV = 1.75
k_{bind} CV = 2.0
k_{bind}CV = 2.25
k_{bind} CV = 2.5
k_{bind} CV = 2.75
k_{bind} CV = 3.0

Fig N. Solutions of the analytical model for different levels of variability of $k_{bind} \cdot [R]$. (Top) Mean, (Middle) CV^2 and (Bottom) S of single-cell protein numbers produced by tandem promoters when $d_{TSS} \leq 35$ (left) and $d_{TSS} > 35$ (right). The green bar is the analytical solution with $CV\left(k_{bind} \cdot [R]\right) = 0$. The other bars are from analytical solutions for various degrees of variability of $k_{bind} \cdot [R]$ of each promoter.

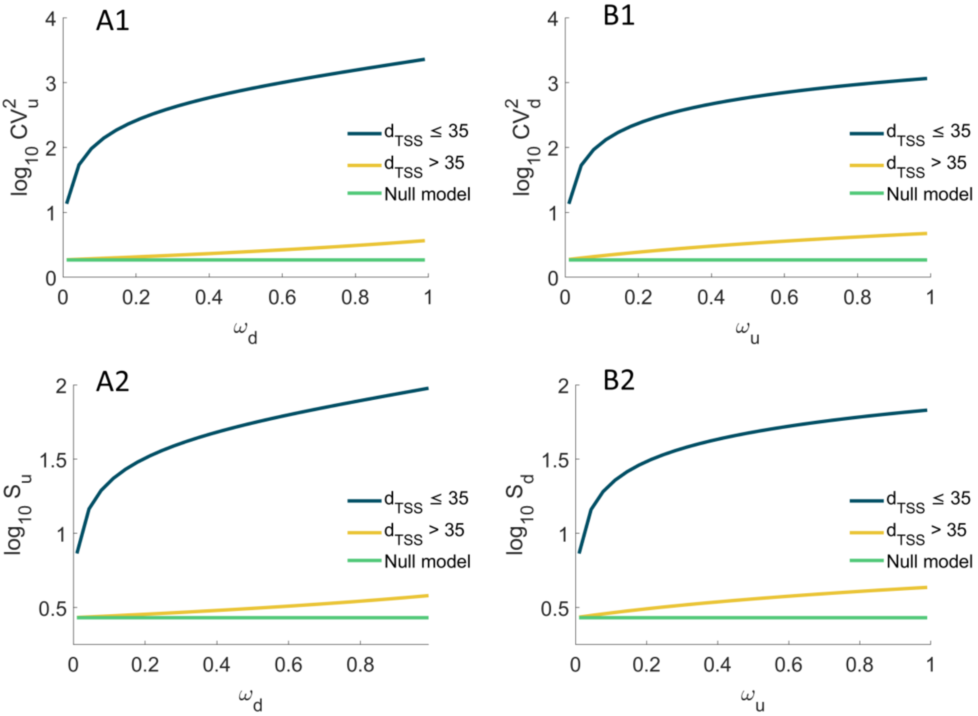


Fig O. Variability and skewness in single-cell protein numbers produced from an upstream and from a downstream promoter as a function of promoter occupancy of the other promoter. CV_P^2 and S_P of the single-cell distribution of the number of proteins produced (**A1 and A2**) by the upstream promoter alone, and (**B1 and B2**) by the downstream promoter alone. Results are shown as a function of the fraction of times that the upstream ($0.01 \leq \omega_u \leq 0.99$) and the downstream ($0.01 \leq \omega_d \leq 0.99$) promoter are occupied by RNAP. The null model is estimated by setting $k_{occlusion}$, $k_{sitting}$, and ω to zero.

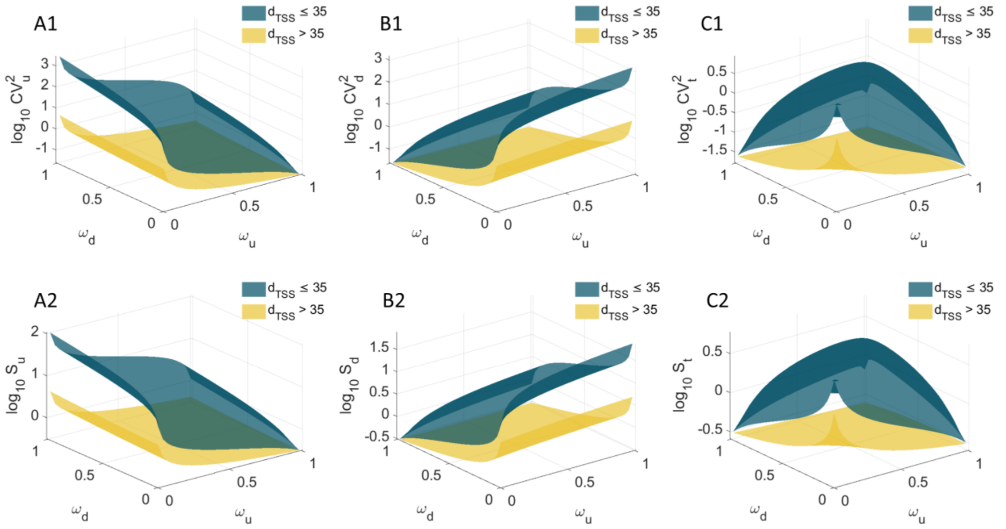


Fig P. Variability and skewness in single-cell protein numbers as a function of promoter occupancy. Expected variability (*CV*^{2}) and skewness (*S*) of the single cell distribution of protein numbers due to the activity of, respectively: (**A1** and **A2**) the upstream promoter alone, (**B1** and **B2**) the downstream promoter alone, and (**C1** and **C2**) both promoters. Shown is *CV*^{2}, *S* as a function of the fraction of times that the upstream ($0 \leq \omega_u \leq 1$) and the downstream ($0 \leq \omega_d \leq 1$) promoters are occupied by RNAP, when $d_{TSS} > 35$ (yellow) and $d_{TSS} \leq 35$ (dark green) bp.

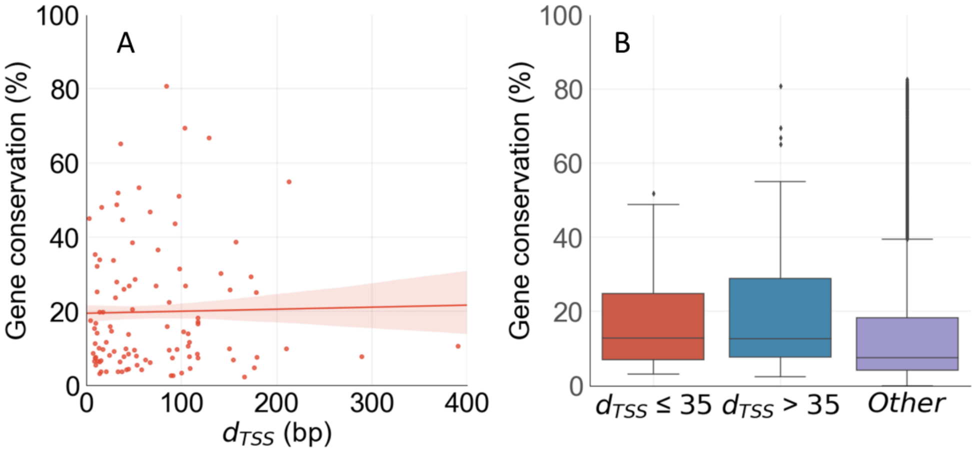


Fig Q. Gene conservation levels. (A) Correlation between d_{TSS} (bp) of the pairs of tandem promoters and the evolutionary conservation level of the gene that they express. The line shown is the best linear fit to the data, and its shadow is the standard error of the fit. (B) Box plot of the gene conservation levels of the cohorts of genes with $d_{TSS} > 35$ and with $d_{TSS} \leq 35$, along with genes other than those in tandem formation. The horizontal black line inside each box marks the median, the top of the box shows the 3rd quartile (Q3), and the bottom of the box shows the first quartile (Q1) of each gene cohort. The error bar above the box marks the range of values within (Q3+1.5*IQR), while the error bar below the bottom shows the range of values within (Q1-1.5*IQR). Here, IQR = Q3 – Q1.

S3 Appendix: Supporting Tables

Table A. List of genes controlled by tandem promoters.

S. No	Configuration (see Fig 1 main manuscript)	Gene	Promoters (upstream/ downstream)	Distance between TSS's (bp)
1	I	aspS	aspSp1/aspSp	84
2	I	bolA	bolAp2/bolAp1	85
3	I	cspI	cspIp/cspIp2	100
4	I	glmU	glmUp2/glmUp1	103
5	I	gltA	gltAp1/gltAp2	97
6	I	hchA	hchAp2/hchAp	150
7	I	ispU	ispUp1/ispUp2	117
8	I	tig	tigp1/tigp3	129
9	I	nuoA	nuoAp1/nuoAp2	173
10	II	acnB	acnBp/acnBp2	45
11	II	bhsA	bhsAp9/bhsAp	14
12	II	cirA	cirAp2/cirAp1	13
13	II	csgD	csgDp1/csgDp2	9
14	II	cspA	cspAp1/cspAp2	51
15	II	dapB	dapBp2/dapBp1	55
16	II	fabI	fabIp/fabIp1	3
17	II	fadR	fadRp/fadRp2	11
18	II	fkpA	fkpAp1/fkpAp2	26
19	II	gpmA	gpmAp2/gpmAp	38
20	II	lysU	lysUp1/lysUp2	8
21	II	mfd	mfdp1/mfdp2	36
22	II	osmC	osmCp1/osmCp2	10
23	II	pfkA	pfkAp2/pfkAp1	48
24	II	pfkB	pfkBp2/pfkBp1	28
25	II	phoH	phoHp1/phoHp2	73
26	II	serC	serCp2/serCp	16
27	II	sohB	sohBp1/sohBp2	17
28	II	ucpA	ucpAp2/ucpAp1	7
29	II	ugpB	ugpBp2/ugpBp1	48
30	II	xdhA	xdhAp/xdhAp2	8

List of genes controlled by tandem promoters whose single-cell protein numbers were measured by flow-cytometry using cells of the YFP strain library. Also shown are their promoters in tandem formation, their configuration, and the distance in base pairs (bp) between their TSSs.

Table B. List of strains of the YFP strain library observed by flow-cytometry.

S. No.	Strain name	Genotype	Source
1	acnB [SX1900]	F-, acnB791-YFP(::cat), Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13455)
2	argP [SX1436]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], argP794-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12991)
3	aspS [SX1044]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], aspS793-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12599)
4	bhsA [SX1979]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], bhsA791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13534)
5	bolA [SX1087]	F-, Δ(argF-lac)169, bolA791-YFP(::cat), gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12642)
6	cirA [SX1509]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], cirA791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13064)
7	csgD [SX1465]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], csgD791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13020)
8	cspA [SX1097]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, cspA791-YFP(::cat), rph-1	Yale CGSC (CGSC # 12652)
9	cspI [SX1106]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], cspI797-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12661)
10	dapB [SX1910]	F-, dapB792-YFP(::cat), Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13465)
11	fabD [SX2002]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], fabD793-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13557)
12	fabH [SX1474]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], fabH795-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13029)
13	fabI [SX1038]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], fabI796-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12593)
14	fadR [SX1521]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], fadR795-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13076)
15	fkpA [SX2015]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, fkpA791-YFP(::cat), rph-1	Yale CGSC (CGSC # 13570)
16	fur [SX1916]	F-, Δ(argF-lac)169, fur-791-YFP(::cat), gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13471)

17	glmU [SX1004]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1, glmU792-YFP(::cat)	Yale CGSC (CGSC # 12559)
18	gltA [SX1925]	F-, Δ(argF-lac)169, gltA791-YFP(::cat), gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13480)
19	gpmA [SX1553]	F-, Δ(argF-lac)169, gpmA791-YFP(::cat), gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13108)
20	hchA [SX1988]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], hchA791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13243)
21	ispU [SX1052]	F-, ispU796-YFP(::cat), Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12607)
22	lysU [SX1127]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1, lysU793-YFP(::cat)	Yale CGSC (CGSC # 12682)
23	mfd [SX1072]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], mfd-791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12627)
24	mreB [SX1466]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], mreB791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13021)
25	nagC [SX1561]	F-, Δ(argF-lac)169, nagC791-YFP(::cat), gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13116)
26	nlpA [SX1615]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1, nlpA791-YFP(::cat)	Yale CGSC (CGSC # 13170)
27	nuoA [SX1772]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], nuoA791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13327)
28	osmC [SX1758]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], osmC791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13313)
29	pepD [SX1530]	F-, pepD792-YFP(::cat), Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC #13085)
30	pfkA [SX1349]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1, pfkA791-YFP(::cat)	Yale CGSC (CGSC # 12904)
31	pfkB [SX1761]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], pfkB792-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13316)
32	phoH [SX1752]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], phoH791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13307)
33	serC [SX1390]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], serC791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12945)
34	sohB [SX1707]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], sohB791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13262)
35	tig [SX1140]	F-, Δ(argF-lac)169, tig-791-YFP(::cat), gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12695)

36	ucpA [SX1211]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], ucpA791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12766)
37	ugpB [SX1574]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, ugpB791-YFP(::cat), rph-1	Yale CGSC (CGSC # 13129)
38	wrbA [SX1718]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], wrbA791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13273)
39	xdhA [SX1671]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], xdhA792-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13226)
40	yccJ [SX1975]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], yccJ791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13530)
41	yccT [SX1368]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], yccT792-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 12923)
42	aldA [SX1901]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], aldA791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13456)
43	elaB [SX1695]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], elaB792-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13250)
44	feoA [SX1781]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, feoA791-YFP(::cat), rph-1	Yale CGSC (CGSC # 13336)
45	gcvT [SX1674]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], gcvT792-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13229)
46	glpD [SX1550]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, glpD792-YFP(::cat), rph-1	Yale CGSC (CGSC # 13105)
47	pepN [SX1519]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], pepN794-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13074)
48	wrbA [SX1718]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], wrbA791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13273)
49	ybeL [SX1822]	F-, Δ(argF-lac)169, ybeL794-YFP(::cat), gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13377)
50	ydfG [SX1986]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], ydfG791-YFP(::cat), IN(rrnD-rrnE)1, rph-1	Yale CGSC (CGSC # 13541)
51	yjbQ [SX1859]	F-, Δ(argF-lac)169, gal-490, Δ(modF-ybhJ)803, λ[cI857 Δ(cro-bioA)], IN(rrnD-rrnE)1, rph-1, yjbQ792-YFP(::cat)	Yale CGSC (CGSC # 13414)

Table C. Average ‘network’ properties of genes with 1 or more TFs.

Network properties	**Genes controlled by tandem promoters with d_{TSS} ≤ 35**	**Genes controlled by tandem promoters with d_{TSS} > 35**	**All promoters of genes with 1 or more TF interactions**

	Mean ± SEM	Random set from all genes Mean ±SEM (p-value)	Mean ± SEM	Random set from all genes Mean ±SEM (p-value)	Mean ± SEM
Average Shortest PathLength	0.31 ± 0.16	0.17 ± 0.11 (0.23)	0.13 ± 0.05	0.17±0.08 (0.60)	0.17 ±0.01
Clustering Coefficient	0.09 ± 0.03	0.11 ± 0.03 (0.68)	0.10 ± 0.03	0.11 ± 0.03 (0.62)	0.11 ± 4.34×10^{-3}
Eccentricity	0.56±0.31	0.25 ± 0.20 (0.22)	0.15 ± 0.06	0.26 ± 0.16 (0.73)	0.26 ± 0.03
Edge Count	5±1.64	4.64 ± 3.4 (0. 33)	3.3 ± 0.83	4.64 ± 2.73 (0.67)	4.63 ± 0.43
Indegree	2.33±0.48	2.32 ± 0.34 (0.52)	2.02 ± 0.17	2.31 ± 0.27(0.83)	2.32 ± 0.04
Neighborhood Connectivity	161.76 ± 29.09	131.95 ± 21.74 (0.20)	134.63 ± 15.1	131.87 ± 17.36 (0. 44)	131.91 ± 2.74
Outdegree	2.66 ± 1.34	2.33 ± 3.4 (0.30)	1.28 ± 0.83	2.31 ± 2.71 (0. 59)	2.32 ± 0.43

Shown are the network properties for genes controlled by tandem promoters at a distance $d_{TSS} \leq 35$ bp and at a distance $d_{TSS} > 35$ bp. For comparison, we show the same properties, when averaged from all genes of *E. coli*'s TF network. Genes without TF's are not considered. Note that all p-values are larger than 0.05.

Table D: Genes controlled by tandem promoters without input TFs.

S. No.	Gene	Availability in the YFP strain library
1	ampH	
2	ansP	
3	aroK	
4	aspS	✓
5	bepA	
6	cfa	
7	cobU	
8	crfC	
9	degQ	
10	fkpA	✓
11	ispU	✓
12	lpp	
13	mepS	
14	mfd	✓
15	narU	

16	opgG	
17	panD	
18	pfkB	✓
19	serW	
20	tig	✓
21	ucpA	✓
22	xapR	
23	ybgI	
24	ygiM	
25	yheO	
26	yobF	

Genes controlled by tandem promoters without input TFs. Those genes whose proteins are tagged with YFP in the YFP strain library are marked with the symbol '✓'.

Table E. Genes controlled by tandem promoters regulated by one and only one input TF.

	Tandem promoter's genes	**Availability in YFP strain library**	**Input TF**	**Availability in YFP strain library**
1	argR		argR	
2	cvpA		purR	
3	cysK		cysB	✓
4	dapB	✓	argP	✓
5	fabI	✓	fadR	✓
6	fadR	✓	fadR	✓
7	fliL		flhdC	
8	ftnB		cpxR	✓
9	glgS		crp	
10	glk		cra	
11	glmU	✓	nagC	✓
12	gpmA	✓	fur	✓
13	hchA	✓	h-ns	
14	ibaG		mlrA	✓
15	iraP		csgD	✓
16	leuL		leuO	
17	livK		lrp	
18	lysU	✓	lrp	
19	mqsR		mqsA	
20	ompA		crp	
21	ompX		fnr	
22	osmB		rcsB	✓
23	pfkA	✓	cra	

24	phoH	✓	phoB	
25	potF		ntrC	
26	slyB		phoP	
27	sohB	✓	crp	
28	wza		rcsaB	
29	xdhA	✓	fnr	
30	ydbK	✓	soxS	
31	yeaG		ntrc	
32	yhbT		csgD	✓
33	yqjA		cpxR	✓

When the proteins of these genes and of their input TFs can be measured using strains of the YFP strain library, they are flagged with the symbol '✓'.

Table F. Genes controlled by, and only by, a TF expressed by tandem promoters.

	Genes controlled by tandem promoters	**Availability in YFP strain library**	**Genes regulated by the protein expressed by the gene controlled by tandem promoters**	**Availability in YFP strain library**
1	argR		argA	✓
2	argR		argB	
3	argR		argC	
4	argR		argE	✓
5	argR		argF	
6	argR		argH	
7	argR		argI	
8	argR		argR	
9	argR		artI	
10	argR		artJ	
11	argR		artM	
12	argR		artP	✓
13	argR		artQ	
14	argR		lysO	
15	bolA	✓	ampC	
16	bolA	✓	dacC	
17	bolA	✓	mreB	✓
18	bolA	✓	mreC	
19	bolA	✓	mreD	
20	csgD	✓	dgcC	
21	csgD	✓	iraP	
22	csgD	✓	nlpA	✓
23	csgD	✓	pepD	✓
24	csgD	✓	wrbA	✓
25	csgD	✓	yccJ	✓
26	csgD	✓	yccT	✓

27	csgD	✓	yhbS	
28	csgD	✓	yhbT	
29	evgA		frc	
30	evgA		oxc	✓
31	evgA		yegR	✓
32	evgA		yegZ	
33	evgA		yfdE	
34	evgA		yfdV	
35	evgA		yfdX	
36	fadR	✓	accA	
37	fadR	✓	accD	
38	fadR	✓	fabD	✓
39	fadR	✓	fabG	
40	fadR	✓	fabH	✓
41	fadR	✓	fabI	✓
42	fadR	✓	fadM	
43	fadR	✓	fadR	✓
44	xapR		xapA	
45	xapR		xapB	

Table G. Protein levels and d_{TSS} of 10 genes as measured by Microscopy and Image Analysis.

Gene	TSS distance (d_{TSS})	Mean single-cell protein level (Microscopy)
xdhA	8	0.04
csgD	9	0.64
serC	16	0.24
sohB	17	0.37
pfkA	48	2.8
dapB	55	0.57
aspS	84	1.72
gltA	97	3.02
hchA	150	0.74
nuoA	173	2.04

Related to Fig 4C in the main manuscript.

Table H. Number of genes controlled by a pair of tandem promoters in each configuration.

Configuration	Number (in RegulonDB)	Present in the YFP strain library (measured here by flow-cytometry)
I	40	9(9)
II	62	21(21)
III	7	3

IV	4	1
V	6	2
VI	0	0
VII	3	1
VIII	2	2
IX	4	1
X	0	0
XI	9	2
Other	6	0

Related to Fig 1 in the main manuscript and Fig A in S2 Appendix.

Table I. Coefficient of variation, CV, of the gamma distribution.

CV **(**$k_{bind} \cdot \left[R\right]$**)**	$Mean\left(abs\begin{pmatrix}k_{bind}^{u} \cdot \left[R\right] - \\ k_{bind}^{d} \cdot \left[R\right]\end{pmatrix}\right)$	$Mean\left(\frac{abs\begin{pmatrix}k_{bind}^{u} \cdot \left[R\right] - \\ k_{bind}^{d} \cdot \left[R\right]\end{pmatrix}}{k_{bind}^{u} \cdot \left[R\right]}\right) \times 100\%$
0.01	7.52×10^{1}	1.14 %
0.1	7.64×10^{-4}	1.16×10^{1} %
0.25	1.86×10^{-3}	2.98×10^{1} %
0.5	3.63×10^{-3}	7.33×10^{1} %
0.75	5.27×10^{-3}	1.99×10^{2} %
1	6.62×10^{-3}	2.05×10^{3} %
1.25	7.81×10^{-3}	5.15×10^{4} %
1.5	8.66×10^{-3}	1.95×10^{7} %
1.75	9.41×10^{-3}	6.19×10^{12} %
2.0	9.89×10^{-3}	1.48×10^{15} %
2.25	1.04×10^{-2}	1.77×10^{17} %
2.5	1.10×10^{-2}	6.60×10^{18} %
2.75	1.12×10^{-2}	4.00×10^{24} %
3.0	1.20×10^{-2}	6.03×10^{30} %

Coefficient of variation, CV, of the gamma distribution from which $k_{bind} \cdot \left[R\right]$ of each promoter in tandem configuration is sampled from. Also shown is the resulting expected mean absolute difference in $k_{bind} \cdot \left[R\right]$ between the upstream and downstream promoters. Furthermore, the last column shows how much larger (in percentage) is one of the $k_{bind} \cdot \left[R\right]$ values compared to the other.

Table J. Location of the tandem promoters relative to the oriC.

Genes controlled by tandem promoters	Distance between the upstream TSS and the oriC
aspS	1975043
bolA	3471395
cspI	2286932
glmU	10418
gltA	3170977
hchA	1890114
ispU	3730960
nuoA	1520409
tig	3470751
acnB	3794225
bhsA	2756725
cirA	1678802
csgD	2822400
cspA	205855
dapB	3897456
fabI	2574623
fadR	2690839
fkpA	448219
gpmA	3138074
lysU	428830
mfd	2751716
osmC	2369148
pfkA	181499
pfkB	2119421
phoH	2840879
serC	2968165
sohB	2596460
ucpA	1381073
ugpB	333318
xdhA	925487

S4 Appendix: Supplementary Results

Pause sequences

We investigated if the nucleotide sequence of and in between the natural tandem promoters is coding for specific sequences known to perturb RNAP elongation. There are several events that compete with stepwise elongation. However, arrest, misincorporation and editing, pyrophosphorolysis, and premature termination are too rare in optimal growth conditions (rate constants listed in [1]) to be influential in several genes, and/or are not sequence dependent. Only sequences known to enhance transcriptional pausing [2] could fit both of these requirements. In *E. coli*, the mean rate of non-sequence specific pauses is 1 per 100 base pairs. These last 3 s on average [3-4]. However, a few sequences can enhance pausing frequency and/or duration (up to 15 or more seconds) [5] via various mechanically processes, which explains their variability in half-life and frequency of occurrence. For example, '*his*' pauses occur when the assembling RNA forms a hairpin-like loop, while '*ops*' pauses do not require it. Likely because of it, *his* pauses have longer half-life [6]. We searched in (and in between) the sequences of the 102 pairs of tandem promoters for the 14 sequences (each 12 nucleotides long) known to enhance pausing [7] (section 'Sequences prone to causes transcriptional pauses' in S1 Appendix) but found none. Thus, sequence-dependent transcriptional pausing should not be a common phenomenon in the tandem promoters of arrangements I and II. Even when allowing for 3 or less mistakes (sequence gaps, misalignments, duplicates, etc.), we only found 5 matches in the 30 of the 102 tandem promoter pairs studied with protein measurements below (Fig B in the S2 Appendix, note the 5 bars crossing the threshold).

Over-representation test

We performed an over-representation test to search for biological functions (as defined in [8,9] that are overrepresented by genes controlled by tandem promoters (using PANTHER 14 [10]). While based on a Fisher test, some biological processes appear to be overrepresented in our genes of interest (e.g., regulation of catabolic processes), none of them were significant to 'FDR correction' (FDR < 0.05, [10]. As such, we failed to identify a biological process significantly associated to genes controlled by tandem promoters (S1 Table).

Input-output transcription factor relationships

From time-lapse RNA-seq data, we assessed if the 102 genes controlled by tandem promoters (arrangements I and II, Fig 1) are affected by their input TFs. To facilitate this, we considered only those that have one and only input TF. I.e., we did not consider the 26 genes that do not have known input TFs (Table D in S3 Appendix), neither the 43 genes that have more than one input TF, making the detection of input-output relationships problematic. As such, of the 102, we considered only 33 genes (Table E in S3 Appendix). In these, we did not observe influences from input TFs (Fig C, Panel A in Fig

D in the S2 Appendix). Finally, and similarly, we observed genes whose only input TF is expressed by tandem promoters (Table F in S3 Appendix). Again, we found no correlation (Panel B in Fig D in the S2 Appendix). Note that, while we did not find influences from TF interactions in the conditions of our measurements, we expect these interactions to become active in other conditions (e.g., stress conditions).

Proteins with membrane-related positionings

From RegulonDB [11], of the 30 genes measured by flow-cytometry (Table A in S3 Appendix), only 3 are known to be related to membrane transportation and binding: bhsA, which is an outer membrane protein that is involved in copper permeability, stress resistance and biofilm formation, cirA, which is also an outer membrane transporter, and ugpB which is a periplasmic binding protein. Such membrane localizations could affect their quantification by YFP fusion, potentially by enhancing effects from avidity due to weakened diffusion.

However, none of these proteins significantly affect our results since, first, cirA and ugpB were removed from our analysis of the 1X condition, after preprocessing (gating, background subtraction and protein number conversion) (marked in red in S2 Table). Meanwhile, all three genes were removed from our analysis of the 0.5X condition after preprocessing (marked in red in S2 Table). Specifically, their removal was due to lack of expression above background autofluorescence.

Relationship with the OriC region

From EcoCyc [12], the OriC region has a length of 232 base pairs and is located in positions 3 925 744 and 3 925 975 in the DNA of *E. coli*. We calculated the shortest distance between the TSS of the upstream promoter and the Oric region. These positions in the DNA are shown in Table J in the S3 Appendix. Meanwhile, the corresponding protein expression levels of these genes in the 1X condition are shown in the S2 Table. Finally, we show a Fig E in the S2 Appendix of these distances from OriC plotted again $\log_{10} M_p$ which shows that the two quantities do not correlate statistically.

Regulation by H-NS

From RegulonDB [11], we investigated how many of the 102 genes controlled by tandem promoters (arrangements I and II) and how many of 30 of them observed by flow-cytometry are expected to be regulated by H-NS.

Of the 102 genes, 14 are regulated by H-NS (14%). Meanwhile, of the 30 genes, 5 are regulated by H-NS (17%). From this, we conclude that H-NS is not consistently a master regulator of these genes.

Nevertheless, of 4698 genes in E. coli, only 4 % are regulated by H-NS. This is significantly lower than in the case of the genes controlled by tandem promoters (p-value < 0.05 based on a Fisher test). As such, one could argue that H-NS regulation does occur higher than expected by chance. Future studies of the dynamics of those genes during environmental changes may thus be of interest.

References

1. Rajala T, Häkkinen A, Healy S, Yli-Harja O, Ribeiro AS. Effects of transcriptional pausing on gene expression dynamics. PLoS Comput Biol. 2010;6: e1000704. doi: 10.1371/journal.pcbi.1000704
2. Herbert KM, La Porta A, Wong BJ, Mooney RA, Neuman KC, Landick R, et al. Sequence-resolved detection of pausing by single RNA polymerase molecules. Cell. 2006;125: 1083–1094. doi: 10.1016/j.cell.2006.04.032
3. Greive SJ, von Hippel PH. Thinking quantitatively about transcriptional regulation. Nat Rev Mol Cell Biol. 2005;6: 221–232. doi:10.1038/nrm1588
4. Neuman KC, Abbondanzieri EA, Landick R, Gelles J, Block SM. Ubiquitous Transcriptional Pausing Is Independent of RNA Polymerase Backtracking. Cell. 2003;115: 437–447. doi:10.1016/S0092-8674(03)00845-6
5. Herbert KM, Greenleaf WJ, Block SM. Single-molecule studies of RNA polymerase: motoring along. Annu Rev Biochem. 2008;77: 149–176. doi: 10.1146/annurev.biochem.77.073106.100741
6. Artsimovitch I, Landick R. Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. Proc Natl Acad Sci U S A. 2000;97: 7090–7095. doi:10.1073/pnas.97.13.7090
7. Gabizon R, Lee A, Vahedian-Movahed H, Ebright RH, Bustamante CJ. Pause sequences facilitate entry into long-lived paused states by reducing RNA polymerase transcription rates. Nat Commun. 2018;9: 2930. doi:10.1038/s41467-018-05344-9
8. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Michael Cherry J, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25: 25–29. doi:10.1038/75556
9. The Gene Ontology Consortium, Carbon S, Douglass E, Good BM, Unni DR, Harris NL, et al. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2020;49: D325–D334. doi:10.1093/nar/gkaa1113
10. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019;47: D419–D426. doi:10.1093/nar/gky1038
11. Santos-Zavaleta A, Salgado H, Gama-Castro S, Sánchez-Pérez M, Gómez-Romero L, Ledezma-Tejeida D, et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12. Nucleic Acids Res. 2019;47: D212–D220. doi:10.1093/nar/gky1077
12. Karp PD, Weaver D, Paley S, Fulcher C, Kubo A, Kothari A, et al. The EcoCyc Database. EcoSal Plus. 2014;6. doi:10.1128/ecosalplus.ESP-0009-2013

PUBLICATION IV

Using synthetic tandem promoter formations to tailor genes with desired dynamics

V. Chauhan, I. Baptista, R. Jagadeesan, S. Dash, and A.S. Ribeiro.

Unpublished Submitted Manuscript