Tampere University

MOHAMED NASURUDEEN MOHAMED BAHRUDEEN

Regulatory Mechanisms of Gene Transcription in *Escherichia coli*

Tampere University Dissertations 730

Tampere University Dissertations 730

MOHAMED NASURUDEEN MOHAMED BAHRUDEEN

Regulatory Mechanisms of Gene Transcription in *Escherichia coli*

ACADEMIC DISSERTATION
To be presented, with the permission of
the Faculty of Medicine and Health Technology
of Tampere University,
for public discussion in the auditorium A109
of the Arvo Building, Arvo Ylpön Katu 34, Tampere,
on 3 Feb 2023, at 12 o'clock.

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

Responsible supervisor and Custos	Professor Andre Sanches Ribeiro Tampere University Finland	
Pre-examiners	Professor Rahul V. Kulkarni University of Massachusetts Boston USA	Reader Baojun Wang University of Edinburgh UK
Opponent	Senior Lecturer Dr. Philipp Thomas Imperial College London UK	

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

Cover design: Roihu Inc.

ISBN 978-952-03-2716-3 (print)
ISBN 978-952-03-2717-0 (pdf)
ISSN 2489-9860 (print)
ISSN 2490-0028 (pdf)
http://urn.fi/URN:ISBN:978-952-03-2717-0

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

PunaMusta Oy – Yliopistopaino
Joensuu 2022

ACKNOWLEDGEMENTS

The work presented in this thesis was conducted from 2017 to 2022 in the Laboratory of Biosystem Dynamics (LBD) of the Faculty of Medicine and Health Technology of Tampere University.

First, I would like to thank Professor Andre S. Ribeiro for giving me this great opportunity to carry out this research work in LBD. He has provided me with active supervision and the opportunity to work on various projects collaborating with other lab members. This offered me a challenging atmosphere and helped me to learn many new things.

Second, I would like to thank all my LBD colleagues, who helped me a lot to complete my doctoral thesis. I would like to thank Samuel Oliveira, who co-supervised me during the early stage of my research career, and for teaching me the microscopy analysis tools, which greatly helped in my research work. Then, I would like to thank Sofia Startceva for helping me to learn the stochastic models. And I would like to thank Vinodh Kandavalli, Vatsala Chauhan, Nadia Goncalves, and Suchintak Dash, for helping me with the experiments. Finally, I would like to thank Bilena Almeida, Cristina Palma, and Ines Baptista for being good colleagues helping each other by solving problems and by having productive discussions.

I am so happy to have worked with the LBD team as they are wonderful people that always gave me a very good work atmosphere. They are so friendly, as well as caring, which was of great mental support to me during my work.

Mohamed Bahrudeen

Tampere, April 2022

ABSTRACT

Bacteria have always been exposed to a wide variety of environments, many of which are fluctuating. In order to survive in these environments, they have had to develop the ability to adapt to changing conditions, particularly hostile ones. The adaptiveness of these microorganisms primarily depends upon their gene regulatory mechanisms. Some of these are 'local', affecting only a few genes. For example, when a specific nutrient appears in the media, it activates a few genes. Other regulatory mechanisms are more complex, involving a large number of genes that need to be activated and/or repressed at specific time moments.

The survival of bacterial species, in some cases, depends on the existence of diversity in their genes' expression across the cell population, particularly since it is not always possible to predict the best action to take next. Understanding the mechanisms of bacteria that regulate the diversity in gene expression would help the bioindustries to benefit from them. Moreover, it would help finding ways to mitigate the harm caused by some species.

Bacterial genes are primarily regulated by their promoter strength in recruiting RNAP and specificity to a σ factor and, in some cases, by one or more global regulators. In addition, many genes are also regulated by specific transcription factors that can act as activators or as repressors, when present. Aside from these, other influential factors are the supercoiling in the DNA region occupied by the gene, whether there are other promoters closely spaced to the promoter of interest and, if so, their orientation, etc (Dash, et al., 2021).

This thesis focused on the study of some of the mechanisms that can affect genes' transcription kinetics. We focused on three mechanisms: i) Building up of positive supercoils, ii) Transcription interference between closely spaced promoters in tandem formation, and iii) Global regulation by input transcription factors.

First, we studied how the intrinsic and extrinsic sources of noise in gene expression could be regulated by tuning the relative duration of transcription initiation. The

study was done using stochastic models. It was found that the diversity in transcription kinetics across a cell population increases with the increase in the relative duration of the closed complex formation.

Second, a method was proposed to dissect the kinetics of transcription locking due to the effects of positive supercoiling buildup. Using RNA fluorescent protein tagging and microscopy, RNA transcripts were quantified in individual cells. It was found that increasing intracellular gyrase concentration decreases how often a promoter goes into the locked state, which in turn increases the gene's transcription rate. Using that information, it is possible to infer how long the promoter is locked.

Third, a method was proposed to quantify the RNA numbers in individual cells using information from flow cytometry. This method allows the quantification of RNA numbers in thousands of cells, and thus the mean and variability in those cells, with much less manual labour and in much lesser time than when using microscopy and image analysis.

Fourth, a method was proposed to dissect the rate-liming steps of gene transcription regulated by promoters in tandem orientation. Using protein fusion library and flow cytometry, the protein abundance was quantified. It was found that the gene's expression could be regulated by tuning the transcriptional interference by varying the promoters' strength and the distance between the transcription start site of the promoters.

Overall, the four studies above allow for, first, better extracting raw data from microscopy and flow cytometry, and from there, to either dissect the kinetics of rate-limiting steps during transcription initiation or, inversely, how they can be tuned to regulate the single-cell RNA and protein numbers.

Having studied two core mechanisms regulating transcription, the fifth and final study focus on a third mechanism, which is transcription factor (TF) regulation. For this, we used RNA-seq to study how RNAP and TFs affect the kinetics of gene cohorts from measurements after RNAP shifts. We found that the magnitude of genes' response is proportional to the asymmetry in the number of activators and repressors regulating them.

Overall, the works conducted in this thesis show that the gene expression and its products diversity in cell populations can be regulated by varying the rate-limiting

steps in transcription. These rate-limiting steps can be tuned by various mechanisms, such as the tuning of the accumulation of positive coils, tuning transcriptional interference of closely spaced promoters in tandem orientation and, tuning which and how many transcription factors act on each gene.

CONTENTS

ABBREVIATIONS

ara	L-Arabinose
aTc	anhydrotetracycline
bp	base pairs
CC	closed complex
CME	chemical master equation
CV^2	squared coefficient of variance
DNA	deoxyribonucleic acid
d_{TSS}	distance between the transcription start sites
FISH	fluorescent *in situ* hybridization
FITC	fluorescent isothiocyanate
[G]	gyrase concentration
GFP	green fluorescent protein
IPTG	isopropyl-β-D-1-Thiogalactopyranoside
LFC	$\log_2$ fold change
mRFP	monomeric red fluorescent protein
mRNA	messenger RNA
OC	open complex
ODE	ordinary differential equation
PCR	polymerase chain reaction
PSB	positive supercoiling buildup
qPCR	quantitative polymerase chain reaction
RBS	ribosome binding site
RNA	ribonucleic acid
[R]	RNAP concentration
RNAP	RNA polymerase
rRNA	ribosomal RNA
S	skewness
Sd	standard deviation
SSA	stochastic simulation algorithm
TF	transcription factor

TFN	transcription factor network
tRNA	transfer RNA
TSS	transcription start site
YFP	yellow fluorescent protein

ORIGINAL PUBLICATIONS

This thesis is a compilation of 5 studies. In the text, the first four studies are referred to as **Publication I, II, III, IV**, respectively, whereas the last study is currently under review and is referred to as **Study V**. The publications are reproduced with permission from the publishers.

I **M.N.M. Bahrudeen**, S. Startceva, A.S. Ribeiro. "Effects of extrinsic noise are promoter kinetics dependent", In Proceedings of the 9th International conference on Bioinformatics and Biomedical Technology (ICBBT), Lisbon, Portugal, 2017. DOI: 10.1145/3093293.3093295

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

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

IV V. Chauhan*, **M.N.M. Bahrudeen***, C.S.D. Palma, I.S.C Baptista, B.L.B Almeida, S. Dash, V. Kandavalli, A.S. Ribeiro. "Analytical kinetic model of native tandem promoters in E. coli", PLoS Computational Biology, 18, e1009824, 2021. *Equal contributions. DOI: 10.1371/journal.pcbi.1009824

Unpublished Manuscript

V B.L.B. Almeida, **M.N.M. Bahrudeen***, V. Chauhan*, S. Dash*, V. Kandavalli, A. Häkkinen, J. Lloyd-Price, C.S.D. Palma, I.S.C. Baptista, A. Gupta, J. Kesseli, E. Dufour, O.P. Smolander, M. Nykter, P. Auvinen, H.T. Jacobs, S.M.D. Oliveira and A. S. Ribeiro. "The transcription factor network of E. coli steers global responses

to shifts in RNAP concentration", bioRxiv, 2022. *Equal contributions.[1] DOI: 10.1101/2022.03.07.483226

[1] A revised version of this pre-print has later been published. The updated reference reads: B.L.B. Almeida, **M.N.M. Bahrudeen***, V. Chauhan*, S. Dash*, V. Kandavalli, A. Häkkinen, J. Lloyd-Price, C.S.D. Palma, I.S.C. Baptista, A. Gupta, J. Kesseli, E. Dufour, O.P. Smolander, M. Nykter, P. Auvinen, H.T. Jacobs, S.M.D. Oliveira, A.S. Ribeiro. "The transcription factor network of E. coli steers global responses to shifts in RNAP concentration", Nucleic Acids Research, 12. 6801–6819. *Equal contributions. DOI: 10.1093/nar/gkac540

AUTHOR CONTRIBUTIONS

In Publication I, the author conceived the study with A.S. Ribeiro. He also performed the data analysis. Further, he contributed to the design of the stochastic models and performed the simulations with S. Startceva. Finally, he contributed to the writing of the manuscript.

In Publication II, the author performed the derivations of the model, assisted by C.S.D. Palma. In addition, he also built the statistical tools used to extract the parameter values of the model. Further he contributed to carrying out some data analysis, such as the RNA quantification, and assisted in the writing of the methods that he performed.

In Publication III, the author conceived the study with V. Chauhan and A.S. Ribeiro. In this methods paper, the author developed the method and built the tools required for executing it. He also performed all the image and data analysis. In addition, he significantly contributed to the writing of the manuscript.

In Publication IV, the author conceived the study with V. Chauhan, C.S.D. Palma and A.S. Ribeiro. He designed the models and built most of the tools. He also performed most of the data and image analysis. In addition, he significantly contributed to the writing of the manuscript.

In Study V, the author did the RNA-seq data analysis along with B.L.B. Almeida. He also built some of the statistical tools for the study along with B. Almeida. In addition, he contributed to the writing of the manuscript.

For all the studies above, the author has not contributed to the wet lab experiments.

1 INTRODUCTION

Gene regulation is the fundamental process by which bacteria adapts to changing environments (Stoebel, et al., 2009). During gene regulation, decisions are made that force the cell to take specific actions that will determine its future success. During all the steps, from detecting external signals, to making a decision involving the activation and repression of specific genes, to performing specific actions, there are many variables that introduce uncertainty into the outcome of each step, and thus, to the final outcome of the cellular actions.

These uncertainties are the sources of phenotypic diversity in bacterial cell populations. Interestingly, this diversity can increase the survival probability of cell population in fluctuating environments (Kussell & Leibler, 2005). If the whole population did not make the same decision, it is more likely that some of them made the right decision. In case of bacteria, the most important source of phenotypic diversity is the stochasticity in the process of gene expression and subsequent decision-making nature of genetic networks after sensing the environmental conditions (Arkin, et al., 1998; Ribeiro, 2010; Razo-Mejia, et al., 2020).

Research has shown that most regulation of gene expression in bacteria occurs during the stage of transcription initiation (Browning & Busby, 2004; Browning & Busby, 2016; McLeod & Johnson, 2001; Ruff, et al., 2015). The transcription initiation includes multiple rate-limiting steps, whose dynamics depends on various factors, such as the promoter sequence, its σ factor specificity, the binding of different regulatory molecules to the promoter, deoxyribonucleic acid (DNA) supercoiling due to local topological constraints, promoters' orientation, etc. (Chong, et al., 2014; deHaseth, et al., 1998; Duchi, et al., 2016; Kandavalli, et al., 2016; Kærn, et al., 2005; Lutz, et al., 2001; McClure, 1985; Shearwin, et al., 2005).

Besides single-gene regulation, cells have another feature that makes their genes capable of performing complex programs. Namely, some genes produce transcription factors that are able to force other genes to change their activity. The

set of genes and transcription factor interactions of organisms are named transcription factor networks (TFN).

The TFN of *E. coli* is likely more studied than the TFN of any other organism. TFNs play an important role of regulating thousands of genes, for example, coordinating their response to environmental perturbations (Martínez-Antonio, et al., 2008). Using signals from the environment, TFNs process them and make decisions. These decisions coordinate cellular functioning following environmental stresses, such as varying nutrient availability, pH, temperature, and other factors.

The study of TFNs and gene expression has advanced much faster since the finding of fluorescent proteins along with live single-cell imaging methods (Kærn, et al., 2005; Elowitz, et al., 2002). Nowadays, microscopy imaging methods along with synthetic proteins allow the *in vivo* tracking of individual proteins (Yu, et al., 2006) and ribonucleic acids (RNAs) (Fusco, et al., 2003; Golding, et al., 2005; Femino, et al., 1998; Raj, et al., 2008).

With the use of experimental and modelling techniques, this thesis aims to study and quantify some of the regulatory mechanisms of transcription, namely DNA supercoiling, closely spaced promoters in tandem orientation, and transcription factors binding. We hope that this knowledge will be useful, in the near future, to synthetic biologists for engineering bacterial strains with predefined gene expression dynamics.

This thesis is organized as follows. Chapter 2 contains a review of the past studies that were considered most significant to our research. It focuses on the process of gene expression and the most relevant regulatory mechanisms. Then it introduces the methods for modelling and simulating gene expression. Chapter 3 describes the aims of this study. Chapter 4 presents all relevant materials and methods applied in our works. Meanwhile, Chapter 5 presents the models, derivations, and estimations of parameter values performed during our studies. Finally, Chapter 6 contains a summary of the results and Chapter 7 contains the discussion. In the end, we provide a list of the references used in this thesis.

2 LITERATURE REVIEW

In this chapter, we provide an overview of the processes by which bacterial gene expression takes place, particularly transcription and translation. We also provide a brief review of some gene regulatory mechanisms caused by DNA supercoiling, closely spaced promoters' arrangement, and input transcription factors. Next, we summarize the analytical approaches to model gene expression in the works below. Finally, we provide a short review of the stochastic modelling approaches and the simulation tools for running the stochastic models of gene expression dynamics.

2.1 Bacterial gene expression

Most living organisms, including bacteria, consist of hereditary information in the form of deoxyribonucleic acid (DNA). The DNA consists of two strands of nucleotide bases, which run in opposite directions. The nucleotide bases in one strand are complementary to the nucleotide bases in the other strand. There are four types of nucleotide bases in the DNA: Adenine (A), thymine (T), cytosine (C), and guanine (G). Adenine pairs with thymine, whereas cytosine pairs with guanine. The nucleotide bases of one strand bind to the complementary nucleotide bases in the other stand through hydrogen bonds, forming a double-strand DNA. During cell division, the DNA is replicated by an enzyme called DNA polymerase, allowing the DNA to be passed on to two daughter cells.

The cells employ the RNAP to transcribe the DNA codes into messenger RNA (mRNA), which is then translated into functional proteins by ribosomes (Crick, 1970). The proteins regulate the essential functions of the cell such as metabolism, growth, and reproduction.

During transcription, the nucleotide sequences in the DNA segment is coded into the RNA molecule. The RNAP molecule starts by recognizing the promoter region, located upstream of a gene in the DNA. Thus, one of the regulating factors of RNA production is the RNAP affinity to a promoter region.

In addition, transcription is regulated by certain proteins called transcription factors, which either enhance transcription (Activators) or repress transcription (Repressors) (Schlax, et al., 1995). At the transcription state site (TSS), the RNAP unwinds the DNA and synthesizes RNA nucleotide bases complementary to the nucleotide bases in one of the DNA strands. The RNAP continues to produce the RNA nucleotide bases one by one until it reaches the terminator site in the DNA. Then the RNAP dissociates from the DNA strand, and so does the completely synthesized RNA molecule.

Next, during translation, the information in the RNA will allow the production of proteins, which perform specific functions in the cell, such as catalysing the metabolic reactions, intra- or extracellular signalling, or forming subcellular structures.

A protein is composed of a chain of amino acids, and each amino acid is encoded by three nucleotide sequences (codon) in the RNA. The synthesis of proteins is performed by ribosomes, which take the mRNA as the template. The translation of mRNA starts after the subunits of ribosomes are assembled at the Ribosome binding site (RBS), following which the ribosome starts elongating the mRNA, synthesizing proteins. The protein synthesis continues until the ribosome reaches the stop codon and after that, it stops. The synthesized protein molecules then undergo folding and transforming into a 3-dimensional structure, to attain the functional property.

2.1.1 Transcription

The first step of transcription is transcription initiation, during which an RNAP recognizes a promoter region and binds to it (McClure, 1985). The second step is transcription elongation, during which the RNAP synthesizes the mRNA nucleotide-by-nucleotide based on the nucleotide sequence in the DNA template (Uptain, et al., 1997). The last step is the transcription termination, during which the complete mRNA is released and the RNAP is also dissociated from the DNA (Nudler & Gottesman, 2002).

In addition to RNAP, other regulatory molecules, called sigma factors, are required to recognize the promoters, and start transcription (see e.g., (Baptista, et al., 2022)). Each sigma factor (σ^{70}, σ^{54}, σ^{38}, σ^{32}, σ^{28}, σ^{19}, σ^{24}) is specific to a cohort of promoter

sequences. Most of the *E. coli* genes' promoters are specific to σ^{70} (Santos-Zavaleta, et al., 2019).

In detail, first, the RNAP core unit reaches the promoter primarily through 3-dimensional diffusion (Wang, et al., 2013). Next, the core unit binds with a sigma factor and forms an RNAP holoenzyme. Each sigma factor has a higher binding affinity towards a certain set of promoters (Burgess, et al., 1969). The holoenzyme binds to the promoter sequence at 10 and 35 nucleotides upstream of the TSS.

Most regulation of transcription occurs during the initiation phase. There are two main rate-limiting steps during this initiation phase: the closed and the open complex formations (Figure 1). In detail, initiation starts once the RNAP holoenzyme binds to the TSS of the promoter region, following which the RNAP forms an inactive transcription complex called the 'closed complex (CC)'. Next, the closed complex isomerizes to form an active transcription complex, called the 'open complex (OC)'. After this, the RNAP escapes the promoter, which becomes free for other RNAPs, and starts synthesizing the mRNA.

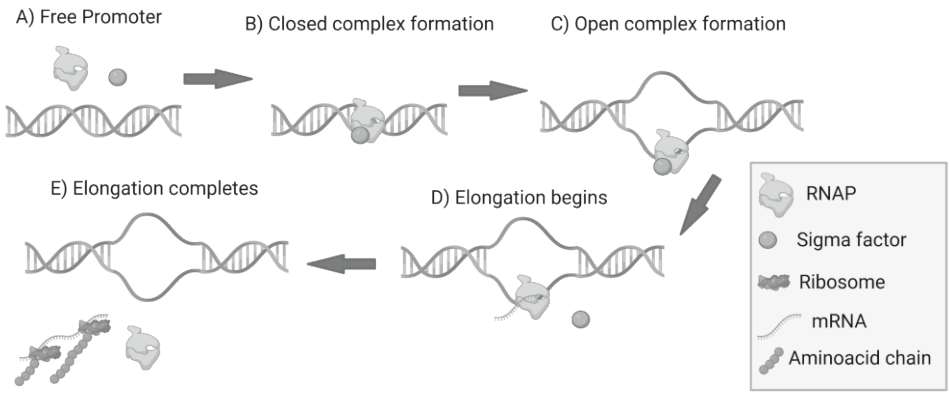


Figure 1. Steps during transcription. A) A promoter is free. B) One RNAP and one sigma factor bind to a promoter, forming a closed complex. C) The RNAP unwinds the DNA strand to form an open complex. D) The sigma factor dissociates, the RNAP starts elongation and forming an mRNA. E) The RNAP and the newly formed mRNA are released, following the completion of elongation. Created with BioRender.com.

2.1.2 Translation

Translation is a sequential process that includes initiation, elongation, and termination. In prokaryotes, this process can begin even before the RNA is fully formed. Translation is carried out by complex molecular machines called ribosomes, which are made up of ribosomal proteins and specialized RNA molecules (rRNAs). There are two main subunits in *E. coli* ribosomes: a small (30S) and a large (50S) subunit (Ramakrishnan, 2002; Metelev, et al., 2022). The large subunit consists of 5S and 23S RNA subunits, and 31 proteins, whereas the small subunit consists of 16S RNA and 21 proteins.

During translation initiation, first, the small ribosomal subunit binds to the RBS region of mRNA, which is located upstream of the start codon AUG. Then an initiator transfer RNA (tRNA) carrying the first amino acid, N-formylmethionine (fMet) recognizes the start codon of the mRNA AUG. Next, the fMet-tRNA forms a 30S-RNA complex after binding to the P-site of the ribosomal subunit (Ramakrishnan, 2002). After this, the large subunit of ribosome binds to the 30S-RNA complex to form the complete ribosome (70S) and starts elongation.

During the elongation phase, the amino acids are carried by tRNAs, and they bind to the codons of the mRNA in the A-site of the ribosome. The amino acids are linked one by one during each elongation event, allowing the polypeptide chain to grow until the stop codon is reached (Ramakrishnan, 2002). After this, the release factor recognizes the stop codon, binds to the ribosome, and releases the completed polypeptide, as well as the ribosome.

2.1.3 Noise in gene expression

Gene expression in *E. coli* is a sequence of complex biochemical reactions (Xie, et al., 2008). Mostly, in these random biochemical processes, a small number of molecules are involved. This results in significant fluctuations in the number of reacting species, which in turn makes gene expression a noisy process when compared to other processes in the cells (Kærn, et al., 2005). This noise can make cells phenotypically diverse within the population, despite the cells being genetically identical and exposed to the same environment (Elowitz, et al., 2002; Eldar & Elowitz, 2010; Bury-Moné & Sclavi, 2017). Noise in gene expression could be beneficial to the cell population because the phenotypic diversity is advantageous to

the cells to adapt to the fluctuating environment (Acar, et al., 2008; Raser & O'Shea, 2005).

The sources of diversity between cells on the products of gene expression (RNA and proteins) have been classified as being "intrinsic" and "extrinsic" (Lin & Amir, 2021). Intrinsic noise sources are those events inherent in the biochemical process of gene expression (e.g., the speed of protein folding which differs between events), whereas the extrinsic noise sources are other events that also cause diversity, such as differences in cellular components between cells that influence the gene expression kinetics, and diversity in the partitioning of those components in cell division (Elowitz, et al., 2002).

With the help of fluorescent protein tagging techniques along with high-resolution microscopy and flow cytometry, researchers have been able to quantify the variability in RNA and protein numbers in the cell population (Golding, et al., 2005; Yu, et al., 2006; So, et al., 2011; Jones, et al., 2014; Skinner, et al., 2013). Also, with the help of time-lapse microscopy imaging, the real-time RNA production events of synthetic $P_{lac/ara-1}$ promoter in *E. coli* cells were observed and quantified. It was found that transcription can occur in bursts (Golding, et al., 2005; Cao, et al., 2020; Engl, et al., 2020). In another study, protein production events were observed, and it was observed that translation events also occur in bursts (Yu, et al., 2006). These random bursts lead to diversity in RNA and protein numbers in the cell population (Sanchez & Golding, 2013).

In addition, studies have shown that variability in RNA and proteins in the cell population is also expected as a result of variations in ribosomes, RNAP, transcription factors (TFs), sigma factors, and other regulating factors, that take part in transcription and translation (Engl, 2019; Jones, et al., 2014; Yang, et al., 2014). Finally, there are some mechanisms although not directly interfere in the gene expression process, they still can contribute to noticeable variabilities in some processes, such as DNA replication, DNA supercoiling, and partitioning of cellular components during cell division (Peterson, et al., 2015; Chong, et al., 2014; Huh & Paulsson, 2011).

2.1.4 DNA supercoiling

When the RNAP transcribes a gene, it causes torsional stress in the DNA as it underwinds the DNA behind it and overwinds the DNA ahead of it. Therefore, in front of the RNAP forms positive supercoils, while behind the RNAP forms negative supercoils (Tripathi, et al., 2022). These do not cancel each other out due to the presence of local topological barriers in the DNA (Deng, et al., 2005; Ma, et al., 2013; Le, et al., 2013; Rovinskiy, et al., 2012) as shown in Figure 2.

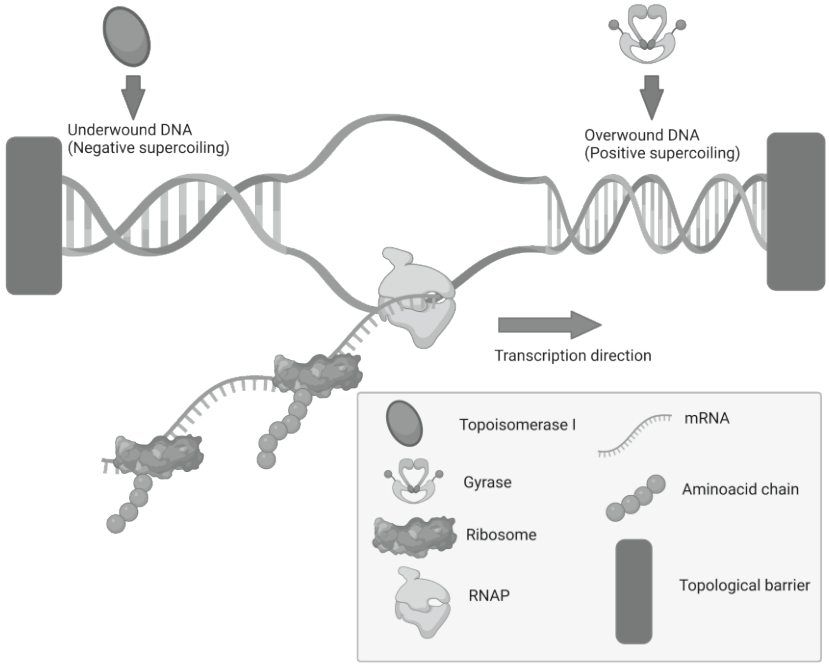


Figure 2. Schematic of DNA supercoils regulation. The transcribing RNAP causes underwinding of the DNA (negative supercoiling) behind it and overwinding of the DNA (positive supercoiling) ahead of it. Created with BioRender.com.

Altering the state of the torsional levels of the DNA affects the activity of the genes (Blot, et al., 2006). There are two specific proteins in *E. coli* whose function is to remove supercoiling buildup so as to maintain the torsional stress of the DNA relatively constant. Gyrase is the protein that removes the positive supercoils, whereas Topoisomerase I is the protein that removes the negative supercoils (Chong, et al., 2014). Positive supercoils, when in large numbers, start slowing down the transcription elongation. At very high levels, they may even halt transcription

initiation (Chong, et al., 2014). These halts can decrease the transcription rate and increase transcriptional noise (Chong, et al., 2014; Mitarai, et al., 2008).

2.1.5 Closely spaced promoters

Genes regulated by closely spaced promoters have been found in many organisms. The promoters can be in convergent, divergent, and tandem arrangements (Shearwin, et al., 2005; Crampton, et al., 2006; Ahmad, et al., 2021). In *E. coli*, nearly 15 percent of all promoters in the genome have another promoter closely spaced (Gama-Castro, et al., 2011).

When in convergent formation, the promoters are oriented in opposite directions (tail-to-tail), separated by some distance, and necessarily transcribe different genes (Figure 3A). Meanwhile, when in divergent formation, the promoters are oriented in opposite directions (head-to-head), separated by some distance, and also transcribe different genes (Figure 3C). Only when in tandem formation, are the promoters oriented in the same direction (head-to-tail), separated by a distance, and transcribe the same gene (Figure 3B) (Shearwin, et al., 2005; Häkkinen, et al., 2019).

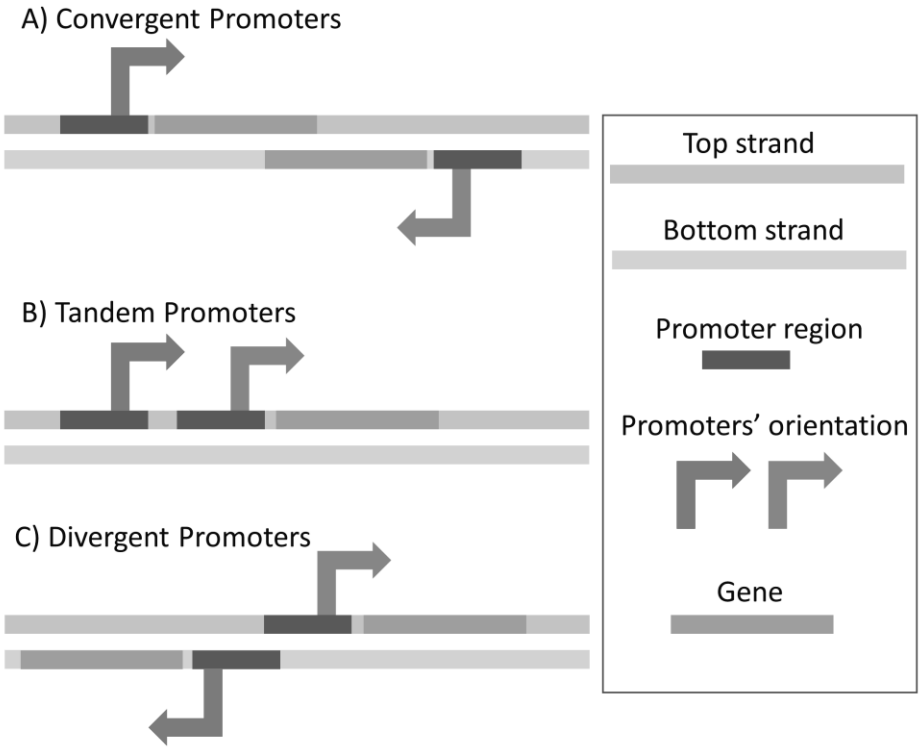


Figure 3. Closely spaced promoters. (A) Convergent promoters: the two promoters are oriented in opposite directions (facing each other), regulating different genes. (B) Tandem promoters: the two promoters are oriented in the same direction, regulating the same gene. (C) Divergent promoters: the two promoters are oriented in opposite directions (facing away from one another), regulating different genes.

Studies have shown that the RNAPs transcribing closely spaced promoters in different orientations undergo transcriptional interference (Sneppen, et al., 2005; Bendtsen, et al., 2011). Also, it has been possible to build models of traffic between RNAPs, while they transcribe the convergent promoters, that match with measurements (Sneppen, et al., 2005; Bendtsen, et al., 2011).

The interference between RNAPs transcribing closely spaced promoters can occur in 3 possible distinct events, shown in Figure 4. First, in the occlusion mechanism, an RNAP bound to a DNA region (being static in closed and open complex formations or being dynamic in elongation) hinders the binding of another RNAP to a promoter (Adhya & Gottesman, 1982). Second, in the sitting duck mechanism, an RNAP elongating on one of the DNA strands interferes with another RNAP, which is bound to a promoter (either in closed or open complex formations) and

dislodges the bound RNAP (Sneppen, et al., 2005; Callen, et al., 2004). Finally, in the collision mechanism, where the two RNAPs elongating in the opposite direction collide, leading to the dislodgement of one or both RNAPs (Ward & Murray, 1979; Prescott & Proudfoot, 2002; Sneppen, et al., 2005).

When an RNAP binds to the promoter, it undergoes multiple binding and unbinding events. Eventually, it successfully commits to the formation of the open complex at the promoter TSS, occupying 35 base pairs (bp) downstream of TSS (Krummel & Chamberlin, 1992; deHaseth, et al., 1998). If the TSS location of the other promoter in tandem orientation is closer than 35 bp, it will be not possible for RNAPs to occupy the TSS of both the promoters simultaneously. Because the RNAP occupying one of the promoters' TSS, occludes the TSS of the other promoter, and the RNAP cannot reach it (Sneppen, et al., 2005).

On the other hand, if the two tandem promoters are separated by more than 35 bp, the RNAP occupying one of the promoters' TSS will not occlude the TSS of the other promoter. Instead, the RNAP elongating from the upstream promoter may encounter an RNAP occupying the TSS (in open or closed complex formation) of the downstream promoter (Callen, et al., 2004) and this should lead one of the RNAPs to fall off.

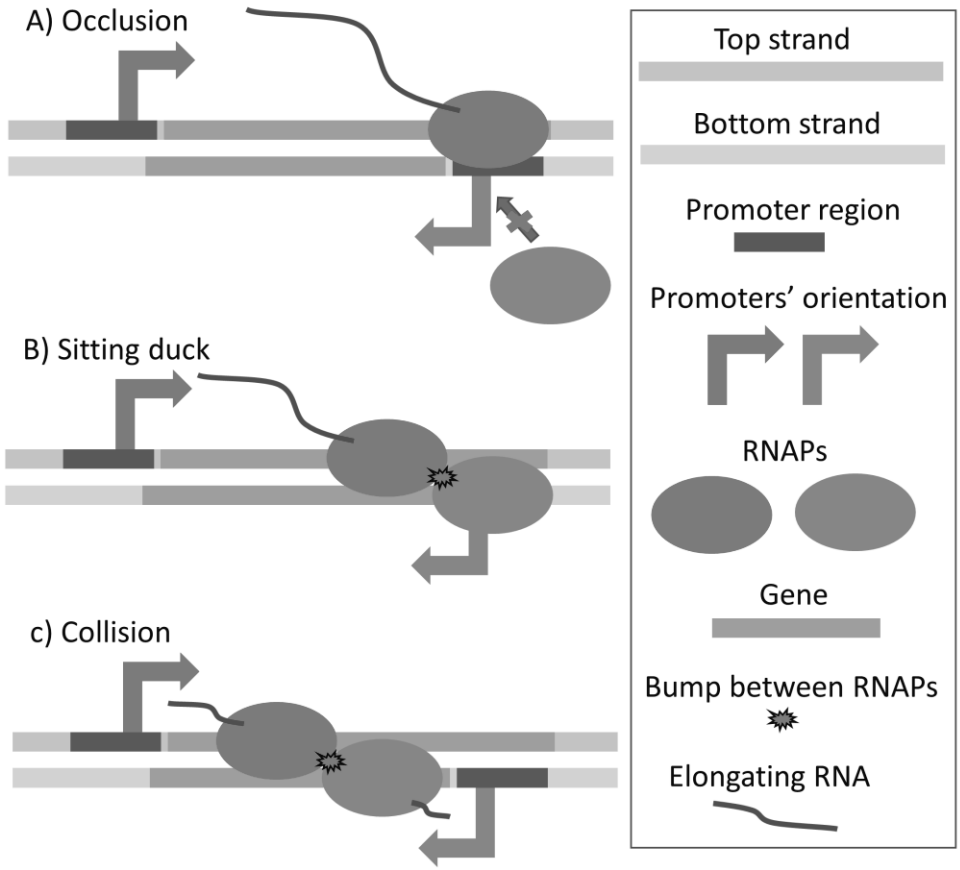


Figure 4. Transcription interference mechanisms. A) Occlusion: An elongating RNAP obstructs another RNAP's attempt to bind to the promoter. B) Sitting duck: An elongating RNAP progression is stopped by an RNAP bound to a promoter. C) Collision: Two elongating RNAPs undergo collision, potentially causing fall-offs.

Furthermore, studies suggest that the RNAPs elongating from the two tandem promoters will not undergo collision, as these occurrences are rare. This is due to less frequent occurrences of transcriptional pauses as well as to the simultaneous firing of RNAPs from both promoters. Even if the RNAPs are simultaneously fired from both promoters, this will unlikely result in RNAP fall-off, because the RNAPs are transcribing in the same direction at approximately the same speeds (Sneppen, et al., 2005; Martins, et al., 2012).

2.1.6 Transcription factor regulation

Transcription factors (TFs) are regulatory proteins that bind to specific operator sites on or near the promoter region, either directly or in cooperation with other regulatory molecules. A TF either upregulates (activation) or downregulates (repression) transcription (Babu & Teichmann, 2003; Browning & Busby, 2004; Pérez-Rueda & Collado-Vides, 2000). A lot of TFs regulate a small group of genes, however there are some TFs act as global TFs, which regulate a huge bulk of genes (Babu & Teichmann, 2003; Hochschild & Dove, 1998; Martínez-Antonio & Collado-Vides, 2003).

TF-mediated transcription repression can occur by various mechanisms. One of the common repression mechanisms is steric hindrance, where the binding of repressor molecule(s) to the operator site(s) in the promoter region physically blocks the RNAP from accessing the -10 or -35 elements (Garcia & Phillips, 2011; Oehler, et al., 1990). Another repression mechanism occurs with the binding of TFs to the operator sites located near the RNAP binding region, thereby bending the DNA into a loop. This can also prevent RNAP from binding to the promoter (Müller, et al., 1996; Oehler, et al., 1990; Swint-Kruse & Matthews, 2009).

TF-mediated transcription activation occurs at many promoters. There are three common mechanisms (Browning & Busby, 2004; Browning & Busby, 2016; Lee, et al., 2012). The first type of activation is when a TF is bound to an operator site upstream of the promoter's -35 element, which helps in RNAP recruitment by undergoing interactions with the α subunits (Browning & Busby, 2004; Ebright, 1993). The second type of activation is when a TF is bound to the DNA, overlapping with the -35 element. This, in general, leads to interaction with the domain 4 of the σ subunit of the RNAP holoenzyme (Browning & Busby, 2004; Dove, et al., 2003). The third type of activation is when a TF bound to the DNA causes a conformational change at the promoter whose spacing between -35 and -10 elements are otherwise non-optimal. This conformational change enables the promoter DNA to position the -35 and -10 elements better, thereby enhancing RNAP binding.

Furthermore, there are several promoters, regulated by multiple TFs simultaneously, through cooperative regulation (Browning & Busby, 2004). Here, the same TF can act both as an activator or a repressor depending upon the promoter it binds to (Pérez-Rueda & Collado-Vides, 2000). In addition, the TF itself is subjected to some regulatory mechanisms, for instance, the DNA binding affinity of a TF can be

changed by binding with certain ligands whose concentration in the cells varies with the environment (Browning & Busby, 2004).

2.2 Fluorescent proteins

The discovery of fluorescent proteins has proven to be a breakthrough in the field of molecular biology. The first fluorescent protein was isolated from a bioluminescent jellyfish named Aequorea Victoria and emits green light. This organism produces 2 proteins, aequorin and a green fluorescence protein (GFP), which are fused. The Aequorin is activated by calcium ions and emits blue light. Then the GFP absorbs the blue light by resonance energy transfer and emits green light (Morise, et al., 1974; Shimomura, et al., 1962; Ward, et al., 1980). Based on this, a variety of fluorescent proteins have been engineered, and have since been serving as the tools to *in vivo* visualization of the cellular components and gene expression activity (Day & Davidson, 2009; Shaner, et al., 2005; Tsien, 1998).

By employing synthetic engineering methods, proteins of interest (e.g., RNAP, Ribosome) can be fused with fluorescent proteins to study their expression dynamics. Likewise, the RNA of interest can be engineered to have certain binding sequences where the fluorescent proteins fused with the binding proteins can come and bind, helping to visualize the RNA dynamics (Golding & Cox, 2004; Yu, et al., 2006). To reliably visualize the fluorescence signal, the signal must be much higher than the autofluorescence of the cell.

However, the fluorescence proteins have some drawbacks, the most important being that different environmental conditions affect their detectability and robustness. So, the properties of the fluorescent proteins themselves, such as photostability and protein maturation time in different environments need to be taken into consideration, so that experiments can be planned and designed accordingly. Often the folding of these proteins is affected by temperature. The pH (acidity) of the environment can also affect the performance of many of the fluorescent proteins (Shaner, et al., 2005). This hampers, in many cases, the comparison of results in measurements taken in different conditions.

2.3 Analytical models of gene expression

Biological systems have been modelled using Analytical or Mathematical modelling techniques for several decades. Analytical models of gene expression have been done that account for the binding of TFs and RNAP to the DNA, the cooperative and inhibitory interactions between TFs, the translation of mRNA to proteins, the degradation of mRNA and proteins, cell division, among other factors. Of the various mathematical models applied to gene expression, here we briefly discuss the thermodynamic- and differential equation-based models (Ay & Arnosti, 2011).

Thermodynamic models have been used to model gene regulation from cis-regulatory regions and the binding of TFs to these regions (Morrison, et al., 2021). For a given promoter and a set of transcription factors, these models predict how the various combinations of bound TFs to those regulatory regions function together to drive gene expression. Usually, it is assumed that the gene expression rate is proportional to the number of bound activators, but it is inversely proportional to the number of bound repressors (Ay & Arnosti, 2011).

These models are implemented in two steps. In step 1, all the possible states that a combination of TFs bind to the DNA are listed. Then, based on the interactions between TF and DNA, each state is assigned a statistical weight which is calculated using the TFs' concentration and their binding affinities to the DNA binding sites. Also, the probability of each possible state is calculated by dividing the statistical weight of each state by the total sum of the statistical weight of all the possible states. In step 2, the gene expression is calculated for each possible state. The rate of gene expression is then set to be proportional to the probability of RNAP binding to the promoter region and to the weighted sum of the TFs binding to the promoter region (Bintu, et al., 2005; Segal, et al., 2008; Fakhouri, et al., 2010; He, et al., 2010).

Differential equation models have been used to model the dynamics of gene expression over time. Here, the interactions regulating gene expression are defined by differential equations, which control the concentration of molecular species such as mRNAs and proteins, based on explicit rules defined in terms of rate equations.

There are two types of such models (Ay & Arnosti, 2011): a) Ordinary Differential Equations (ODE) and Partial Differential Equations (PDE). ODE models have a single independent variable, whereas PDE models have multiple independent

variables. In general, solving these models analytically is hard. Instead, approximate solutions can be obtained using numerical methods. PDE models are more complex and computationally expensive due to having multiple variables in the equation. ODE models have been applied to model bacterial operons such as lac and tryptophan with the independent variable being the inducer molecule. Instead, PDE models could be applied when multiple factors, such as TFs and other regulatory molecules, affect expression rates.

2.4 Stochastic models of gene expression

Molecular species usually exist in relatively low copy numbers inside the cells (McAdams & Arkin, 1999). Also, from the perspective of gene expression, RNA and other regulatory molecules are generally not abundant. Consequently, even small changes in their numbers could significantly affect the behaviour of gene regulatory networks (Mileyko, et al., 2008). As such, they can be influenced by fluctuations in the numbers of the molecular species due to biochemical noise. Such events could not be accounted for in continuous and deterministic reaction-rate equations. Instead, to model these biochemical events more accurately, stochastic models have been proposed (Arkin, et al., 1998; Gillespie, 1977; McAdams & Arkin, 1997).

The stochasticity in the dynamics of gene regulatory networks can be intuitively understood using the chemical master equation (CME). However, solving the CME is not a trivial task using analytical methods. Instead, the CME can be approximated to stochastic differential equations under certain conditions. These stochastic differential equations consist of a deterministic ODE with a noise term (Gillespie, 2000).

Stochastic models can also be solved using an alternative approach, which consists of directly simulating the time evolution of the regulatory system without the need of solving the CME. This stochastic simulation approach was developed by (Gillespie, 1977). In short, this algorithm a) determines when and which reaction will occur the next time based on the system state at time t, and b) it revises the state of the system in accordance with the reaction result.

2.4.1 Stochastic simulation algorithm

The stochasticity of gene expression and its influence on the temporal evolution of a system can be studied using the stochastic simulation algorithm, SSA (Gillespie, 1977). The SSA samples the time evolution of a molecular species at present times. For a system with a number N of molecular species, let X be the state vector: $X = (x_1, \ldots, x_N)$, where x_i is the number of molecules of the species S_i at the present time instant t. Some of these molecular species may be able to react with one another via the M possible chemical reactions R_j (j = 1, …., M).

Next, one can establish a propensity function a_j such that, given the state vector X and an infinitesimal time interval dt with [t, t+dt], $a_j(x)dt$ represents the probability of reaction R_j to occur in the time interval dt (Gillespie, 2007).

Based on this, first, for a unimolecular reaction R_j of the form $S_a \rightarrow product$ with rate constant c_j, the probability that any of the x_a number of S_a molecules undergoes reaction in the next infinitesimal time dt equals $x_a \cdot c_j \cdot dt$. Thus, the propensity function for this reaction can be written as: $a_j(X) = c_j x_a$.

Similarly, for a bimolecular reaction R_j of the form with rate constant c_j, the probability that any of the x_a number of S_a molecules and any of the x_b number of S_b molecules undergoes reaction in the next infinitesimal time dt equals $x_a \cdot x_b \cdot c_j \cdot dt$. Thus, the propensity function for this reaction can be written as: $a_j(X) = c_j x_a x_b$.

However, for a bimolecular reaction R_j of the form $S_a + S_a \rightarrow product$ with rate constant c_j, the probability that any two of the x_a number of S_a molecules undergoes reaction in the next infinitesimal time dt equals $\frac{x_a(x_a-1)}{2} \cdot c_j \cdot dt$, where $\frac{x_a(x_a-1)}{2}$ is the number of distinct molecular pairs of S_a that can react. Thus, the propensity function for this reaction can be written as: $a_j(X) = c_j \frac{x_a(x_a-1)}{2}$.

The SSA is an iterative approach, where in each iteration, it has to answer two questions: i) which reaction (μ) occurs next, and ii) when it will occur (τ), where τ is an exponential random variable with mean $\frac{1}{a_0(X)}$ (Gillespie, 1977). To answer this, in each iteration, the values of μ and τ are generated using a Monte Carlo procedure, for example by direct methods, which use the standard inverse transform sampling (Gillespie, 2007). In this case, in each iteration, two random numbers, r_1 and r_2, are drawn (independently) from uniform distributions within the interval between 0 and 1. Having these, next, τ and μ are obtained from equations (4.1 - 4.3).

$$a_0(X) = \sum_{j=1}^{M} a_j(X) \tag{4.1}$$

$$\tau = \frac{-\log(r_1)}{a_0(X)} \tag{4.2}$$

The reaction, which satisfies the condition in equation (4.3), will be chosen as the next reaction (μ) to occur.

$$\sum_{j=1}^{\mu-1} a_j(X) \le r_2 a_0(X) < \sum_{j=1}^{\mu} a_j(X) \tag{4.3}$$

Once τ and μ are chosen for a given iteration, the state vector X is updated based on the state-change vector v_μ, which stores the changes in the number of molecules of each of the molecular species in the system, due to the outcome of reaction μ.

The stepwise procedure of SSA is summarized in "Algorithm 1", assuming a start time t_i, a stop time t_{end}, and an initial state X_i.

Algorithm 1: Stochastic simulation algorithm

1: Let $t = t_i$; $X = X_i$

2: Calculate all $a_j(X)$ values and $a_0(X)$ value using equation (4.1)

3: **while** $(t < t_{end})$ and $(a_0(X) > 0)$ **do:**

4: Generate two uniform random numbers r_1 and r_2

5: Calculate τ and μ using equations (4.2 - 4.3)

6: Update $t = t + \tau$; $X = X + v_\mu$

7: Calculate all $a_j(X)$ and $a_0(X)$ new values

8: **end while**

2.4.2 Delayed stochastic simulation algorithm

Gene expression comprises many complex events that occur in sequence and are not instantaneous. Thus, one modelling approach for this process is to model each time-consuming step as a separate reaction, as in (Mäkelä, et al., 2011). The negative side of this approach is that as additional reactions and species are added, this can significantly increase the duration of the simulations. To overcome this problem, a strategy was proposed, in which multi-sequential reactions are modelled in 1 step, but the reaction's products are only released after a specific time delay. These time delays account for the time-length of the intermediate steps they represent, without having to explicitly add them in the model (Bratsun, et al., 2005; Gibson & Bruck, 2000; Roussel & Zhu, 2006).

For implementing this strategy, a new algorithm, the delay SSA, was developed. In this approach, the reaction products are not released instantaneously following the occurrence of a reaction. Rather they are released into the system after a time delay. The duration of the delay can be pre-defined, or it can be randomly drawn from a pre-defined distribution. The delays can differ between the reactions and also the reaction products of the same reaction can have different delays. This approach facilitates the implementation of more realistic models (from the point of view of temporal evolution) while reducing the computational costs significantly (Roussel & Zhu, 2006).

For the implementation of the delay SSA, the delayed events are separated into a) reaction events, which instantaneously consume the reactants, and b) generating events, which release the delay reaction products. The next reaction and the time taken for it to happen are selected as when using the SSA. In detail, generating events having no delays are executed instantaneously when the corresponding reacting events are executed. Meanwhile, the generating events which are having non-zero-time delays are moved into a waiting list L, where the generating events are sorted based on their release times. Once the time it takes for the next reaction to occur is sampled, it is compared with the lowest time delay in the waiting list L. If the time

for the reaction to occur is smaller than the first release time, the reaction is executed. Otherwise, the generating event with the least release time on the waiting list is executed and its product will be released. After the occurrence of one of the events, a new time interval is drawn for the next reaction or product release has to occur (Roussel & Zhu, 2006; Zhu, et al., 2007).

The steps to implement the delay SSA are shown in Algorithm 2. Let t_{min} be the release time of the generation events, whose waiting time is the smallest on the waiting list and the state change of this generation event is represented as g_{min}. Also, let v_μ be the state change vector for non-delayed generation events, whereas g_μ is the state change vector for delayed generation events.

Comparing Algorithms 1 and 2, the delay SSA behaves exactly like the regular SSA, if all the delays are set to zero (Roussel & Zhu, 2006; Zhu, et al., 2007).

Algorithm 2: Delayed stochastic simulation algorithm

1: $t = t_i$; $X = X_i$; L = Empty waiting list
2: Calculate all $a_j(X)$ values and $a_0(X)$ value using equation 4.1
3: **while** $(t < t_{end})$ and $(a_0(X) > 0)$ **do**
4: Generate two uniform random numbers r_1 and r_2
5: Get τ and μ using equations 4.2 - 4.3
6: **if** L is empty **then**
7: $t_{min} = inf$
8: **else**
9: t_{min} = Release time of the earlier event in L
10: g_{min} = State change due to the occurrence of earliest event in L
10: **end if**
11: **if** $\tau < t_{min}$ **then**
12: $t = t + \tau$; $X = X + v_\mu$
13: If the reaction μ has g_μ, add g_μ and its delay time to L
14: **else**
15: $t = t + t_{min}$; $X = X + g_{min}$
16: Remove the earliest g_μ from L
17: **end if**
18: Calculate all $a_j(X)$ values and $a_0(X)$ value using equation 4.1

19: **end while**

2.4.3 Simulators

To run stochastic models of gene regulatory networks, including the effects of cell division, various simulators have been proposed recently (Gupta & Mendes, 2018). Most of these simulators use the Stochastic Simulation Algorithm (SSA) such as in (Hoops, et al., 2006; Lok & Brent, 2005) to solve the chemical master equation, Also, there are other simulators available that combine the SSA with ODEs to speed up the simulation process (Gupta & Mendes, 2018).

More complex models can be done if the simulators could implement compartmentalization. For instance, cell compartments could be dynamically created and sized, to model cell division and partitioning of cellular components. Considering this, presently, there are simulators available that have incorporated compartmentalization, where the compartments are created dynamically at runtime (Lok & Brent, 2005; Lloyd-Price , et al., 2012) to simulate growing cell populations.

The Stochastic gene network simulator, SGNS2 (Lloyd-Price , et al., 2012), uses the next reaction method for solving the SSA. It also has the option of using the delayed SSA (Roussel & Zhu, 2006), which enables solving multiple delayed reactions with hierarchical, interlinked compartments, that can be created, destroyed, and divided at runtime. Hence, it can be used to simulate complex models of gene expression such as transcription and translation at the level of nucleotides and codons, respectively. Further, it is also possible to simulate cell lineages and to partition the cellular components based on various partition schemes.

Stochpy (Maarleveld, et al., 2013) is a simulator integrated into Python. It uses the direct method for solving the SSA and other stochastic simulation algorithms, such as tau-leaping and the next reaction method. It also provides inbuilt statistical functions and plotting features.

GeneCircuits simulates gene networks of cells growing exponentially, while allowing time-delayed events and changing reaction rate constants over time (Luo, et al., 2013). This enables the users to model complex events, such as changes (at runtime) in cellular components, temperature, and other perturbations by changing rate

constants. The delay feature can be used to model protein folding events, based on realistic parameter values.

Finally, COPASI (Hoops, et al., 2006), is a simulator with a graphical interface. It can simulate models using diverse methods such as stochastic, deterministic, and hybrid deterministic-stochastic methods.

In this thesis, we preferentially used the SGNS2 simulator to simulate the stochastic models.

3 AIMS OF THE STUDY

This thesis results from the study of some of the regulatory mechanisms of the dynamics of transcription of the bacterium *Escherichia coli*. The study was conducted using a multidisciplinary approach that, first, includes empirical data collection using single-cell measurements by microscopy and by flow cytometry, as well as genome-wide measurements by RNA-seq. Second, it includes the use of signal and data processing methods ranging from image analysis, noise subtraction, and artifacts removal, up to RNA-seq data analysis. Third, it makes use of analytical and stochastic models of gene expression, from the single-promoter level, up to the genome-wide transcription factor network level.

Our approach was unique in that we analysed several orders of magnitude, ranging from the rate-limiting steps in transcription initiation of individual genes, up to genome-wide perturbations of the TF regulatory network. Also, the influence of biophysical parameters was considered.

Overall, to better understand how events occurring at the small scale (e.g., the kinetics of rate-limiting steps in transcription) affect the global dynamics of gene networks, we set these five main objectives.

Our first objective was to study how the rate-limiting steps in transcription initiation regulate the propagation of cell-to-cell variability in RNAP numbers into the cell-to-cell variability in RNA numbers. For this, we modelled cell populations with diverse single-cell concentrations of RNAPs, which we obtained empirically along with other parameter values of the models. We then addressed our questions using stochastic simulations of model genes with various transcription initiation kinetics. The results were presented in **Publication I**. Noteworthily, some of the methods and tools developed in **Publication I** were used for achieving our second and fourth objectives.

Our second objective was to dissect the rate-liming steps in transcription due to transcription locking caused by positive supercoiling buildup. For this, we used

synthetically engineered genetic constructs and microscopy measurements of fluorescently tagged RNAs, along with analytical and stochastic models. The results were presented in **Publication II**.

Our third objective was to establish a method to quantify the mean and variability of single-cell RNA numbers using MS2-GFP tagging, from flow cytometry data. This would give more confidence in these estimations, compared to when using microscopy data, because one can easily collect data from tens of thousands of cells. The results were presented in **Publication III**. The methods and tools developed were later used in **Publication IV**.

Our fourth objective was to develop an analytical kinetical model able to predict the transcriptional interference between natural closely spaced tandem promoters, based on parameters of the individual promoters and the nucleotide distance between them. To support the study, we used a protein fusion strain library to measure the single-cell protein expression levels of genes regulated by tandem promoters by flow cytometry. The results were presented in **Publication IV**.

Our fifth and final objective was to study the influences from global parameters of the transcription factor network logic and topology that regulate genome-wide transcriptional responses to global cellular perturbations. We mainly used RNA-seq measurements to carry out this study. The results were presented in **Study V**.

4 MATERIALS AND METHODS

In this chapter, we give an overview of the experimental and data analysis methods. We also briefly explain a plotting technique for dissecting the time length of the rate-limiting steps in transcription initiation.

4.1 Single-cell RNA measurements

There are various techniques available to quantify RNA at the single-cell level, such as single-cell RNA sequencing (Croucher & Thomson, 2010) and *in situ* hybridization techniques (So, et al., 2011; Taniguchi, et al., 2010). However, most methods, including the ones mentioned, require the breaking down of the cells for the RNA measurement. Thus, they cannot be used to study transcription over time. Meanwhile, the fluorescence tagging system can be utilized to observe the transcriptional events over time (Suzuki, et al., 2007).

The MS2-GFP RNA tagging, which is an RNA protein fluorescence tagging, was synthetically engineered by Golding and his team in *E. coli* (Golding, et al., 2005). It allows the *in vivo* tracking of RNA production events (Peabody, 1993). The MS2-GFP RNA tagging system has two components: a reporter and a target as shown in Figure 5. The target is a gene coding for the RNA, which needs to be quantified. For this, the natural RNA is either replaced or simply followed by many tandem repeats of MS2 capsid protein binding sites. Meanwhile, the reporter gene codes for a green fluorescent protein (GFP) that is fused with a bacteriophage MS2 coat protein. The reporter gene needs to be hosted by a high copy plasmid and be expressed under the regulation of a strong constitutive promoter, in order to ensure that the MS2-GFP proteins are always abundant in the cells. As the binding of the MS2-GFP proteins to the respective binding sites on the RNA is rapid and stable, the detection of individual RNA production events is possible using time-lapse microscopy imaging. It is also critical for the efficiency of this method that the MS2 viral protein has high specificity and binding affinity to the MS2 binding sites (Johansson, et al., 1998).

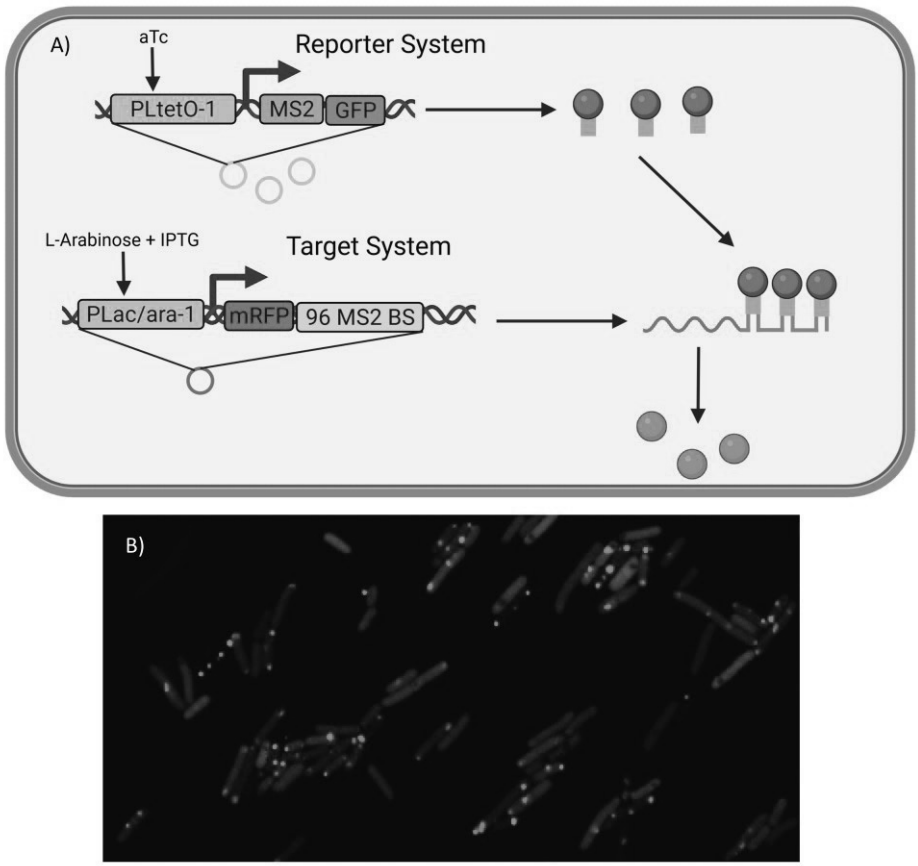


Figure 5. MS2-GFP RNA tagging system. A) Illustration of RNA being tagged with fused MS2-GFP proteins. The target system is produced by a single copy F-plasmid, wherein the promoter $P_{Lac/ara-1}$ regulates the production of the RNA sequences coding for monomeric red fluorescent protein (mRFP) and 96 MS2 binding sites. This target promoter is induced by L-Arabinose (ara) and isopropyl-β-D-1-Thiogalactopyranoside (IPTG). Meanwhile, the reporter system is implemented on a high copy plasmid, wherein the promoter $P_{LtetO-1}$ regulates the sequences coding for MS2-GFP. The reporter promoters are induced by anhydrotetracycline (aTc). When the target RNA is expressed, the MS2-GFP fused proteins quickly bind to its 96 MS2 binding sites. The target RNA also contains an mRFP coding sequence, which, after produced, is translated, and red fluorescent mRFP proteins are produced from it. B) Example confocal microscopy image of *E. coli* where both the target and reporter systems are fully induced. The formation of RNA-MS2-GFP complexes is responsible for the visible bright fluorescent spots, whereas the background fluorescence is due to free-floating unbound MS2-GFP proteins. Created with BioRender.com.

Unlike the reporter gene, the target gene coding for the MS2 binding sites is carried by a single-copy F-based vector, so as to observe transcription events at the resolution of a single promoter. In detail, when the gene is transcribed once, it produces one RNA with tandem repeats of MS2 binding sites. This allows the RNA to be tagged by the MS2-GFP proteins, which creates a bright glowing spot in the cell.

Furthermore, the binding of MS2-GFP to the target RNA MS2 binding sites was shown to prevent the degradation of the target RNA by the RNA-degrading enzymes, and this makes the target RNA nearly immortal during standard measurements (Golding & Cox, 2004; Muthukrishnan, et al., 2012). This critically facilitates RNA quantification, as it is not affected by the RNA degradation process. In detail, whenever a target RNA is produced in the cell, RNA fluorescent spot intensity should necessarily increase, which greatly facilitates RNA quantification.

4.2 Single-cell protein measurements

A chromosome yellow fluorescent protein (YFP) fusion library was engineered by researchers where, in each strain, a particular gene was fused with a YFP gene sequence as shown in Figure 6 (Taniguchi, et al., 2010). Importantly, YFP fluorescence can be detected with single-molecule sensitivity in live bacterial cells (Yu, et al., 2000).

Originally, to perform a high-throughput analysis of these YFP expressing genes, an automatic imaging platform using microfluidic technology was implemented (McDonald, et al., 2000). Since single-cell protein numbers are biased by cell size and gene copy number variation, the normalization by cell size of protein numbers was required (Taniguchi, et al., 2010).

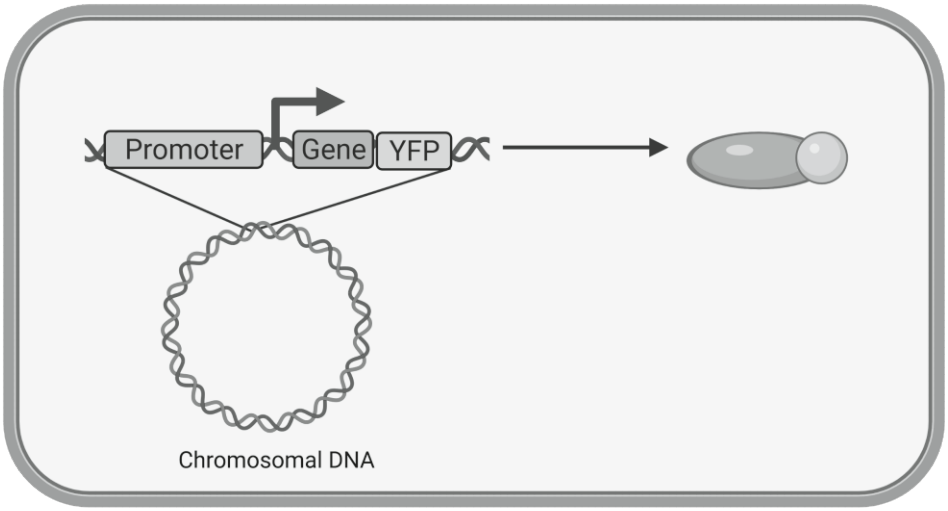


Figure 6. Schematic representation of a chromosomal-integrated gene-YFP fusion under the control of the natural promoter. When the gene in the DNA is expressed, a YFP is produced (yellow ball), and fused to the native protein (green ovoid). Created with BioRender.com.

4.3 Experimental methods and analysis

In this subsection, we listed the experimental methods that have been used to produce the data that was used in this thesis.

4.3.1 Bacterial strains

In **Publication II**, we used two strains from *E. coli* BW25993 (lacI^{q} hsdR514 ΔaraBAD$_{AH33}$ ΔrhaBAD$_{LD78}$) (Datsenko & Wanner, 2000). In both the strains the target gene is P$_{LacO3O1}$-mCherry-MS2-BS. In one of the stains, the target gene is instead integrated into the lac locus of the genome (Gene Bridges, Heidelberg, Germany). In the other strain, the target gene is integrated into the single-copy F-plasmid pBELOBAC11 (Hayakawa, et al., 1985; Mori, et al., 1986).

The promoter controlling the expression of the target gene, P$_{LacO3O1}$, is inducible by isopropyl-β-D-1-Thiogalactopyranoside (IPTG). It was engineered from a native lac promoter from which was removed the O2 repressor binding site, which originally was downstream from the TSS (Oehler, et al., 1994). The target gene expresses an

RNA that codes for an array of MS2 binding sites, where the MS2-GFP proteins expressed by the reporter genes can bind to.

In addition, to overexpress gyrase, a plasmid was constructed (pZe11 P_{rham} gyrAB-sfGFP), which includes the genes gyrA and gyrB. Their expression is under the regulation of a Rhamnose promoter, which can be expressed by adding Rhamnose to the media.

Meanwhile, in **Publication III**, we used the *E. coli* strain DH5α-PRO, which produces regulatory proteins for the tight regulation of the promoters of the genetic constructs used (AraC, LacI, and TetT). There were two genetic constructs in the strain: i) A single-copy target F-plasmid that expresses RNA coding for an array of 96 MS2 binding sites under the tight regulation of $P_{Lac/ara-1}$, which is inducible by IPTG and/or L-Arabinose (ara); and ii) a multicopy reporter plasmid that expresses MS2d-GFP (fusion of two MS2 proteins forming a dimer, MS2d) proteins under the tight regulation of the promoter $P_{LtetO-1}$, which is inducible by aTc.

Next, in **Publication IV**, we used YFP fusion library strains (where YFP is chromosomally integrated) to measure the protein expression of genes regulated by tandem promoters. In each strain, a particular gene was fused with the YFP gene sequence (Taniguchi, et al., 2010). Also, we used RL1314 stain where the RpoC gene is tagged with GFP (provided by Robert Landick) to measure the expression of one of the RNAP's subunits.

Finally, in **Study V**, we used wild-type MG1655 strain to carry out the genome-wide transcriptomic study. Next, we used RL1314 to measure the intracellular RNAP levels in different media conditions. Also, we used a strain that consists of rpoS::mcherry gene to measure the RpoS expression in the cells (Patange, et al., 2018). In addition, we also measured the expression of (p)ppGpp and some other genes using the YFP fusion library (Taniguchi, et al., 2010).

4.3.2 Microscopy, Image analysis and RNA Quantification

To capture the cell fluorescence, confocal images of the cells were taken using a C2+ (Nikon), a point scanning confocal microscope system. Meanwhile, to capture the cell morphology, a CCD camera (DS-Fi2, Nikon) was used to take the phase-contrast

images of the cells. Using the NIS-Elements software, all the captured microscopy images were extracted.

The image analysis was done using the software 'CellAging' (Häkkinen, et al., 2013). Under image analysis, first, cell segmentation was performed by applying a gradient path labelling algorithm (Mora, et al., 2011). Then classifiers were used for merging (to minimize over-segmentation) and discarding segments (unwanted artifacts such as air bubbles). The classifiers were built using a Classification and Regression Trees algorithm (Breiman, et al., 1984), and were manually trained by an expert using example images (Queimadelas, et al., 2012). Manual corrections were done in the case of necessity.

Next, CellAging aligns (semi-automatically) the confocal images with the corresponding phase-contrast images. This is done using thin-plate spline interpolation for the registration transform (by manual selection of 5-8 landmarks), in order to adjust the cell masks to overlap with the corresponding cells in the confocal image. Then, the fluorescent spots (MS2d-GFP tagged RNAs) inside the cells were automatically detected using the Gaussian surface fitting algorithm (Häkkinen & Ribeiro, 2015). An example of segmented RNA fluorescent spots inside the cells in a confocal image is shown in Figure 7. The fluorescence intensity of each pixel inside the segmented RNA spots is subtracted by the mean background fluorescence intensity, obtained from the mean fluorescence intensity of all pixels outside the segmented spot inside the cell.

Finally, to quantify RNA numbers, first, the total RNA spots fluorescence intensity on each cell was converted to integer-valued RNA numbers as in (Golding, et al., 2005). From the histogram of RNA spots intensity, the intensity that corresponds to the mode of the first peak (I_{peak}) should be approximately equal to the spot intensity corresponding to 1 RNA. Then, the total spots intensity of the cell is divided by I_{peak} and rounded to the nearest integer, to get the integer-valued RNA of the cell. In **Publication II**, we used this manual method to quantify RNA numbers from the distribution of RNA spots intensity per cell.

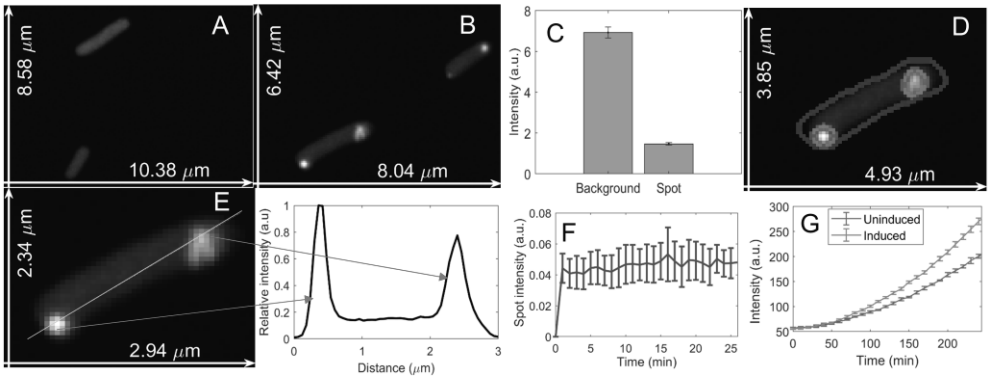


Figure 7. Microscopy and image analysis. A) An example confocal microscopy image of cells expressing only reporter proteins MS2d-GFP. The free-floating MS2d-GFP proteins fill the cells, making them fully visible. B) Example confocal microscopy image of cells expressing both the reporter proteins MS2d-GFP and the RNAs target for those proteins. The reporter proteins become bound to the multiple binding sites in the target RNA causing them to become bright glowing spots. C) The mean total fluorescence intensity of the background pixels of a cell and the mean total fluorescence intensity of pixels with the fluorescent spots inside the cell (units are arbitrary). The error in each bar is the standard error of its mean. D) Example of a cell segmented using the software 'CellAging' (Martins, et al., 2018; Häkkinen, et al., 2013). The segmented cell appears with blue borders and the segmented fluorescent spots inside the cell appear with red borders. E) An example of a fluorescent intensity profile along the major axis of a cell. Using Image J (Abramoff, et al., 2004), the fluorescent intensity profile along the yellow line was obtained and the values were normalized by the maximum value, which was then plotted as a function of distance from the left pole of the cell. (F) Mean fluorescence intensity (in arbitrary units) of an RNA spot (due to RNA tagged with MS2d-GFP) tracked over time after its appearance. We used timeseries microscopy data from 10 cells. After the RNA quantification, these cells were found to have only 1 RNA each. G) Total fluorescence intensity (in arbitrary units) of a cell population measured over time using spectrophotometry. The brown curve is for the cells with both the target and reporter genes induced. The blue curve is for the cells having only the reporter gene induced and not having the plasmid carrying the target gene. Adapted from (Bahrudeen, et al., 2019).

Meanwhile, in **Publication III**, we used an automated method to quantify RNA numbers from the distribution of total RNA fluorescence spot intensity per cell, (Häkkinen, et al., 2014). This method estimates the parameter values in maximum likelihood sense from the RNA spot fluorescence distribution, which is then used to build a maximum a posteriori classifier to estimate the total RNA numbers in each cell (Häkkinen, et al., 2014).

Finally, in **Publication IV** and **Study V**, we skipped the RNA spot segmentation and quantification steps as they were not needed to calculate the total protein fluorescence intensity of the cells.

4.3.3 Spectrophotometry

The fluorescence intensities of cells of populations expressing fluorescent proteins over time were measured using the BioTek Synergy HTX Multi-Mode Microplate Reader with the Gen5 software. From the overnight cultures, cells were taken and diluted to 1:1000 times in fresh LB medium. Then the cells were incubated at 37 °C in a shaking incubator until an OD_{600} of 0.3 was reached. Then, the cells were then aliquoted into 96 well microplates and they were allowed to grow while keeping the same temperature and shaking.

In **Publication III**, specifically, we measured the fluorescence intensity of: i) cells with both the reporter and target genes activated and ii) cells with only the reporter gene activated, and not carrying the target gene. The measurements were done using the excitation filter of wavelength 485/20 nm and the emission filter of wavelength 525/20 nm. The fluorescence of the cell population was recorded for 10 hours, at an interval of 10 min. For each condition, we have done 6 technical replicates.

4.3.4 Flow cytometry and data analysis

Flow cytometry experiments were done using an ACEA NovoCyte Flow cytometer. To detect the GFP fluorescence (from MS2d-GFP and from RNA-MS2d-GFP) or YFP fluorescence (from YFP proteins), we excited the cells with a blue laser (488 nm) and we used the fluorescein isothiocyanate (FITC) channel (530/30 nm) to capture the emitted fluorescence. Meanwhile, to detect the fluorescence of red fluorescent proteins (mRFP or mCherry), we excited the cells with a yellow laser (561 nm) and we used the PE-Texas Red fluorescence channel to capture the emitted fluorescence. Finally, to ignore the non-cell events due to particles smaller than bacteria, we have set the detection threshold to be 5000 in FSC-H. In the end, the FC data was collected using the ACEA NovoExpress software.

The raw FC data was cleaned by applying unsupervised gating (Razo-Mejia, et al., 2018). For this, $\log_{10}$ FSC vs $\log_{10}$ SSC was plotted and then a 2d Gaussian function

was fitted on top of it. We have set α to 0.99, where α is the area of the Gaussian fit, within which the fraction of single-cell events were chosen for further analysis. In general, we observed that the removed events outside this area could be due to non-cell events such as debris, doublets, fragments, cell clumps, and other undesired events. In addition to this, in **Publications III and IV**, we performed additional pre-processing of the FC data.

Specifically, in **Publication III**, we have removed possible biases by cell size (Cunningham, 1990) by normalizing FITC-H (height of the detected cell event in the FITC channel) by the pulse width, denoted as F/W. Likewise, we also normalized the PE-Texas Red-H (height of the detected cell event in the PE-Texas Red channel) by the pulse width, denoted by R/W. Also, from the distributions of F/W and R/W, we removed the < 0.01 % of the highest values.

Meanwhile, in **Publication IV**, we applied secondary gating to remove the outliers from the data. In detail, we first sorted the FITC-H data and then calculated the difference between consecutive sorted values. Next, we stored the indices of those whose difference between the consecutive values was greater than 10000 (which is approximately 10 times the mean of the background). Next, the minimum of those indices was selected to be the upper bound. Finally, the values above this index were considered outliers and were discarded.

4.3.5 RNA-seq experiments and data analysis

In **Publication IV**, the RNA-seq experiments were done using a single-index, 2×150 bp paired-end (PE) configuration on an Illumina HiSeq machine. RNA-seq samples were taken from the medium at time moments 0, 20, and 180 mins after the overnight cultures were diluted to OD_{600} of 0.05 in fresh LB medium.

In **Study V**, first, the RNA-seq experiments were performed for decreasing media richness (LB_{1x}, $LB_{0.75x}$, $LB_{0.5x}$, and $LB_{0.25x}$) at 180 min using 1×75 bp single-end (SE) configuration on an Illumina Miseq machine. We further did the RNA-seq experiments for decreasing media richness (LB_{1x}, and $LB_{0.5x}$) at 60 and 125 min using a single-index, 2×150 bp paired-end (PE) configuration on an Illumina Novaseq machine. Meanwhile, the RNA-seq experiments were performed for increasing media richness (LB_{1x}, $LB_{1.5x}$, $LB_{2.0x}$, and $LB_{2.5x}$) at 180 min using a single-index, 2×150 bp paired-end (PE) configuration on an Illumina HiSeq machine.

The sequenced raw data was obtained as fastq files. The pipeline to perform the RNA-seq data analysis was: i) RNA sequencing reads were trimmed using the Trimmomatic v.0.39 (Bolger, et al., 2014); ii) Trimmed reads were aligned to the reference genome, *E. coli* MG1655 (NC_000913.3), using STAR aligner v.2.5.2b (Dobin, et al., 2013) for **Publication IV** and Bowtie2 v.2.3.5.1 (Langmead & Salzberg, 2012) for **Study V**; iii) The aligned reads to the regions where the genes are located, were counted using 'featureCounts' in Rsubread package v.1.34.7 (Liao, et al., 2019); iv) These read counts were used to perform the differential expression analysis. For this, we performed pre-processing: Genes having less than 5 counts in more than 25 % of the samples were removed. Also, the genes having a mean count across the samples less than 10 were removed from further analysis. Finally, v) The differential expression analysis was performed using DEseq2 package v.1.24.0 in R (Love, et al., 2014) to calculate the gene-wise LFC ($\log_2$ fold change) of the target group of samples with respect to the control group of samples using Wald tests. P-values were adjusted using the BH procedure for multiple hypothesis testing (Benjamini & Hochberg, 1995).

4.3.6 Quantitative PCR and Western blot

Polymerase chain reaction (PCR) is a molecular biology technique that amplifies a specific DNA fragment. PCR in combination with reverse transcription is used to amplify RNAs, in order to quantify the relative RNA expression levels of a gene between conditions (Livak & Schmittgen, 2001; Schmittgen & Livak, 2008).

Quantitative PCR (qPCR) allows real-time gene expression measurements by using fluorescence labels. By measuring the fluorescence intensity, we can estimate the amount of the amplification product. Fluorescent labelling can be done either by the binding of fluorescent dyes to the DNA non-specifically or by the binding of fluorescent probes specific to the target sequence or by the binding of both dyes and the probes (Lind, et al., 2006). The qPCR experiments performed in this thesis work used a non-specific dye (SYBR Green I) (Livak & Schmittgen, 2001; Schmittgen & Livak, 2008).

In **Publication II**, qPCR experiments were done to estimate the relative RNA levels between the different experimental conditions, expressed under the regulation of the promoter $P_{LacO3O1}$ in both the chromosome and the plasmid. This data was used to make Lineweaver-Burk plots, which is explained in section 4.4.

Western blot is a molecular biology technique that has been commonly used to measure the relative abundance of a specific protein between conditions (Burnette, 1981; Mahmood & Yang, 2012). In **Publication II**, relative RNAP between the conditions was estimated by measuring RpoC proteins using western blot. Unlike qPCR, which quantifies the RNA transcripts, the western blot quantifies the proteins.

4.4 Lineweaver-Burk plots

Lineweaver-Burk Plot is a 2d graph to describe the relationship between the inverse of the rate of reaction and the inverse of the substrate concentration. It has been widely used to dissect the rate-limiting steps in transcription initiation using the linear relationship between the inverse of transcription rate and the inverse of concentration of RNAP, [R] (McClure, 1980; Lloyd-Price, et al., 2016; Kandavalli, et al., 2016). This method assumes that the rate at which the RNAP binds to a promoter is directly proportional to [R]. Namely, the average time that a promoter spends before committing to the open complex formation is proportional to $[R]^{-1}$.

Meanwhile, the average time taken for the subsequent rate-limiting steps after committing to the open complex formation should not depend upon [R], rather they should be constant in relation to [R]. Given this, the average transcription times should vary linearly with $[R]^{-1}$. As such, by measuring the average transcription times under different [R] in the cells, it should be possible to estimate the average time the promoter spends before committing to open complex formation ([R] dependent step) and after committing to open complex formation ([R] independent step).

In **Publication II**, a similar method was used to dissect the rate-limiting steps in transcription due to PSB. Here it is assumed instead that the rate at which the positive supercoils are removed is directly proportional to the concentration of gyrase, [G]. Thus, the average time that a promoter spends in the locked state is proportional to $[G]^{-1}$. As such, by measuring the average transcription times under different [G] in the cells, it should be possible to estimate the average time the promoter spends in the locked state.

5 MODEL DERIVATIONS AND ESTIMATIONS

In this chapter, we describe in detail the key model derivations made for the studies in this thesis. We also provide the derivations to estimate the model parameters.

5.1 Analytical kinetic models of gene expression

In this section, we explain the derivation of an analytical model, predictive of the inverse of the mean transcription rate of a gene subject to PSB. Next, we focus on an analytical model predictive of the mean and variability of the single-cell protein numbers of genes regulated by tandem promoters.

5.1.1 Modelling supercoiling effects on transcription

In **Publication II**, we modelled the expected transcription kinetics of a gene when subject to the effects of positive supercoiling as in equations 5.1-5.6. The model considers that the accumulation of positive supercoils is due to the RNAP activity on the gene of interest (Chong, et al., 2014) and on the neighbouring genes (Kouzine, et al., 2013; Naughton, et al., 2013; Teves & Henikoff, 2014; Lilley & Higgins, 1991).

The rate of accumulation of positive supercoils due to the neighbour genes' activity is assumed to be constant. However, this rate varies with the location of the neighbour gene (e.g., whether the neighbouring gene is in the same transcription unit as the gene of interest), with the distance between the gene of interest and the neighbour gene, and with their direction of transcription (Weinstein-Fischer, et al., 2000; Cheung, et al., 2003). It is to be noted that in the model, while the rates are constant, the accumulation process is stochastic.

The PSB effects considered in the model were: a) locking during transcription initiation, which causes long halts in transcription, and b) transient elongation arrests, which cause short pauses, slowing down elongation. To resolve transcription

initiation locking and elongation arrests, the intervention of gyrases is required (Chong, et al., 2014).

$$Pro \xrightarrow{k_1 \cdot [R]} Pro + RNA \tag{5.1}$$

$$Pro + SC^+ \xrightarrow{k_{lock}} Pro_{lock} + SC^+ \tag{5.2}$$

$$Pro_{lock} \xrightarrow{k_{unlock}} Pro \tag{5.3}$$

$$\xrightarrow{\lambda \cdot k_1 \cdot [R]} SC^+ \tag{5.4}$$

$$\xrightarrow{k_p \cdot [R]} SC^+ \tag{5.5}$$

$$G + SC^+ \xrightarrow{k_{rem}} G \tag{5.6}$$

In equation 5.1, $k_1 \cdot [R]$ is the mean active transcription rate (if there is no PSB), $k_{lock} \cdot SC^+$ is the mean rate at which the promoter goes to a locked state due to the accumulation of positive supercoiling, $k_{unlock}{}^{-1}$ is the mean time that the promoter takes to unlock when it goes to the locked state, λ is the mean number of SC^+ accumulated in 1 complete elongation event, $\lambda \cdot k_1 \cdot [R]$ is the mean rate at which the PSB accumulates because of the activity of the gene of interest, $k_p \cdot [R]$ is the mean rate at which the PSB accumulates because of the activity of neighboring genes in the same topological domain, and $k_{rem} \cdot [G]$ is the mean rate at which a positive supercoil is resolved by gyrase.

At steady state, the mean time taken between consecutive transcription events is derived as:

$$r^{-1} = \left(\frac{k_p \cdot [R]}{k_1 \cdot [R]} + \lambda \right) \cdot \frac{k_{lock}}{k_{unlock} \cdot k_{rem} \cdot [G]} + \frac{1}{k_1 \cdot [R]} \tag{5.7}$$

5.1.2 Modelling gene expression of promoters in tandem formation

In **Publication IV**, we modelled the transcription and translation kinetics of genes regulated by tandem promoters. To simplify the modelling of transcription initiation of upstream and downstream promoters in tandem formation as in reactions 5.8 and 5.9, we assumed that the overall time taken for the RNAP recruitment and binding to the promoter and the movement of RNAP towards the promoter's TSS is much higher than the time taken for closed complex formation after an RNAP reaches the TSS of the promoter.

In general, first, an RNAP binds to a free promoter (P_{free}) and moves towards TSS. This makes the promoter to go to the occupied state ($P_{occupied}$) with a rate constant $k_{bind} \cdot [R]$, where [R] is the concentration of RNAP. The bound RNAP is unstable and unbinds from the promoter with a rate constant k_{unbind} before it successfully forms the closed complex and it is represented as a reverse reaction, where $P_{occupied}$ becomes P_{free}.

The bound RNAP also unbinds if the TSS of the promoter is occluded by the RNAP occupying the TSS of the other promoter. This is represented as a reverse reaction, where $P_{occupied}$ becomes P_{free} at a rate constant, $k_{occlusion}$. In reaction 5.8, the superscript 'u/d' in $k_{occlusion}$ denotes the occlusion of the upstream promoter TSS due to the RNAP occupying the TSS of the downstream promoter (which differs with the affinity of the RNAP to the downstream promoter). So, the rate constants of these two reverse reactions can be summed up ($k_{occlusion} + k_{unbind}$) as in reactions 5.8 and 5.9. After the successful formation of a closed complex, the RNAP opens the DNA strand with a rate constant k_{after}, starts elongation (R_{elong}), and frees the promoter.

For the upstream promoter, transcription initiation is modelled as,

$$P^u_{free} \underset{k^{u/d}_{occlusion}+k^u_{unbind}}{\overset{k^u_{bind}\cdot[R]}{\rightleftarrows}} P^u_{occupied} \xrightarrow{k^u_{after}} P^u_{free} + R^u_{elong} \tag{5.8}$$

For the downstream promoter, transcription initiation is modelled as,

$$P^d_{free} \underset{k^{d/u}_{occlusion}+k^d_{unbind}}{\overset{k^d_{bind}\cdot[R]}{\rightleftarrows}} P^d_{occupied} \xrightarrow{k^d_{after}} P^d_{free} + R^d_{elong} \quad (5.9)$$

The phenomenon of occlusion is expected to occur only if the TSS of the promoters are close enough (at a distance smaller than the nucleotide length occupied by an RNAP).

Also, the rate constant of occlusion interference should increase with the increase in the RNAP occupancy (ω) of the other promoter. Also, it should decrease with the distance between the TSS of the promoters. In accordance, $I(d_{TSS})$ models the interference as a function of d_{TSS}, which varies between 0 (at higher d_{TSS}) and 1 (d_{TSS} ~0)

So, we defined $k_{occlusion}$ for the upstream promoter as in equation 5.10 and for the downstream promoter as in equation 5.11, respectively. $k^{\max}_{ocl}$ quantifies maximum occlusion possible, which is expected to occur when d_{TSS} ~0 and the TSS of the other promoter is always occupied (ω~1).

$$k^{u/d}_{occlusion} = k^{\max}_{ocl} \cdot I(d_{TSS}) \cdot \omega_d \quad (5.10)$$

$$k^{d/u}_{occlusion} = k^{\max}_{ocl} \cdot I(d_{TSS}) \cdot \omega_u \quad (5.11)$$

Even if the two promoters' TSS are separated by a length longer than what RNAP occupies, we expect RNAP interference to still occur. The RNAP elongating from the upstream promoter undergoes interference with the RNAP occupying the TSS of the downstream promoter, leading to the dislodgment of one of the RNAPS. The elongating RNAP (R_{elong}) from the upstream promoter either successfully produces RNA as in reaction 5.12, or it can be dislodged due to interference with the RNAP bound to the downstream promoter, as in reaction 5.13. The fraction of time that an elongating RNAP finds an RNAP occupying the downstream promoter is approximately equal to ω_d. Meanwhile, the parameter f defines the probability that an RNAP occupying downstream promoter gets dislodged and fall-off, due to the interference with RNAP elongating from the upstream promoter.

Likewise, the RNAP occupying the downstream promoter could be dislodged due to the encounter with the elongating RNAP from the upstream promoter, freeing the downstream promoter TSS as in reaction 5.14. This interference can be quantified by equation 5.15, where $k_{occupy}^{\max}$ is the maximum possible interference.

$$R_{elong}^{u} \xrightarrow{k_{elong}^{u} \cdot (1-\omega_d \cdot f)} RNA \tag{5.12}$$

$$R_{elong}^{u} \xrightarrow{k_{elong}^{u} \cdot \omega_d \cdot f} \varnothing \tag{5.13}$$

$$P_{occupied}^{d} \xrightarrow{k_{occupy}} P_{free}^{d} \tag{5.14}$$

$$k_{occupy} = \omega_u \cdot k_{after} \cdot k_{occupy}^{\max} \cdot (1-f) \tag{5.15}$$

Meanwhile, the elongating RNAP from the downstream promoter may not undergo any interference and successfully produce the RNA.

$$R_{elong}^{d} \xrightarrow{k_{elong}^{d}} RNA \tag{5.16}$$

The translation of RNA to proteins, degradation of RNA and proteins are modelled as, respectively:

$$RNA \xrightarrow{k_{pr}} Prot \tag{5.17}$$

$$RNA \xrightarrow{k_{rd}} \varnothing \tag{5.18}$$

$$Prot \xrightarrow{k_{pd}} \varnothing \tag{5.19}$$

Assuming that both the upstream as well as the downstream promoters in tandem formation have the same transcription initiation kinetics, the mean protein expression at steady state can be derived to be:

$$M_p = \frac{k_r \cdot k_{pr}}{k_{rd} \cdot k_{pd}} \tag{5.20}$$

$$k_r = \left(\frac{k_{bind} \cdot [R] \times k_{after} \cdot (1 - \omega_d \cdot f)}{k_{occlusion} + k_{bind} \cdot [R] + k_{unbind} + k_{after}} + \frac{k_{bind} \cdot [R] \times k_{after}}{k_{occlusion} + k_{occupy} + k_{bind} \cdot [R] + k_{unbind} + k_{after}} \right) \tag{5.21}$$

Since most single-cell protein distributions in optimal conditions can be well approximated by a negative binomial distribution (Taniguchi, et al., 2010), here we assumed that our protein distribution is also well approximated in the same way. For a negative binomial distribution with parameters r (number of failures) and p (probability of success per event), the mean (M_P), variance (Var_P) and the skewness (S_P) of the distribution should equal:

$$M_P = \frac{pr}{1-p} \tag{5.22}$$

$$Var_P = \frac{pr}{(1-p)^2} \tag{5.23}$$

$$S_P = \frac{1+p}{\sqrt{pr}} \tag{5.24}$$

Meanwhile, the relation between the squared coefficient of variance (CV_P^2) and M_P can be written as:

$$CV_P^2 = \frac{1}{M_P} \cdot \left(\frac{V_P}{M_P} \right) \tag{5.25}$$

Inserting equations (5.22) and (5.23) in equation (5.25):

$$CV_P^2 = \frac{\frac{1}{1-p}}{M_P} \tag{5.26}$$

The above equation can be rewritten by assuming a scaling factor C_1 as:

$$C_1 = \frac{1}{1-p} \tag{5.27}$$

$$CV_P^2 = \frac{C_1}{M_p} \tag{5.28}$$

From (Taniguchi, et al., 2010), C_1 can be approximated as:

$$C_1 = \frac{M_P}{M_{RNA}} \cdot \frac{\frac{1}{\tau_P}}{\frac{1}{\tau_P} + \frac{1}{\tau_{RNA}}} \tag{5.29}$$

where M_{RNA} is the mean RNA numbers of a gene in the cell population, while $\tau_P = \frac{1}{k_{pd}}$ and $\tau_{RNA} = \frac{1}{k_{rd}}$ are the lifetimes of proteins and RNAs, respectively. Hence the above equation can be rewritten as:

$$C_1 = \frac{k_{pr}}{k_{pd}} \cdot \frac{k_{pd}}{k_{pd} + k_{rd}} \tag{5.30}$$

$$C_1 = \frac{k_{pr}}{k_{pd} + k_{rd}} \tag{5.31}$$

Taking the $\log_{10}$ of both sides of equation (5.28), one obtains:

$$\log_{10}\left(CV_P^2\right) = \log_{10}\left(C_1\right) - \log_{10}\left(M_P\right), \text{ with } C_1 = \frac{k_{pr}}{k_{pd} + k_{rd}} \tag{5.32}$$

The relationship between the mean and skewness can be written from the equations (5.22 and 5.24) as:

$$S_P = \frac{\frac{1+p}{\sqrt{1-p}}}{\sqrt{M_P}} \tag{5.33}$$

The equation can be rewritten assuming constant C_2 as:

$$C_2 = \frac{1+p}{\sqrt{1-p}} \tag{5.34}$$

$$S_P = \frac{C_2}{\sqrt{M_P}} \tag{5.35}$$

The constants C_1 and C_2 are related to each other as follows. From equation (5.27):

$$p = 1 - \frac{1}{C_1} \tag{5.36}$$

From equations (5.34 and 5.36), it can be written as:

$$C_2 = \frac{2 - \frac{1}{C_1}}{\sqrt{\frac{1}{C_1}}} \tag{5.37}$$

$$C_2 = 2\sqrt{C_1} - \sqrt{\frac{1}{C_1}} \tag{5.38}$$

Taking $\log_{10}$ on both sides on equation (5.35), we get:

$$\log_{10}(S_P) = \log_{10}(C_2) - \frac{\log_{10}(M_p)}{2}, \text{ with } C_2 = 2\sqrt{C_1} - \frac{1}{\sqrt{C_1}} \tag{5.39}$$

5.2 Estimation of parameters of transcription with PSB

In **Publication II**, we estimated the parameters of the kinetic model of transcription due to the effects of PSB as shown in reactions 5.1-5.6. For condition x, the inverse of RNA production rate, r_x^{-1} equals:

$$r_x^{-1} = \left(\frac{k_{lock}}{k_{unlock} \cdot k_{rem}} \cdot \frac{k_p \cdot [R_x]}{k_1 \cdot [R_x]} \cdot \frac{1}{[G_x]} \right) + \left(\frac{k_{lock} \cdot \lambda}{k_{unlock} \cdot k_{rem}} \cdot \frac{1}{[G_x]} \right) + \frac{1}{k_1 \cdot [R_x]} \quad (5.40)$$

where $[R_x]$, $[G_x]$ refer to the RNAP and gyrase concentrations at condition x, respectively.

For the reference condition c, the equation 5.40 becomes:

$$r_c^{-1} = \left(\frac{k_{lock}}{k_{unlock} \cdot k_{rem}} \cdot \frac{k_p \cdot [R_c]}{k_1 \cdot [R_c]} \cdot \frac{1}{[G_c]} \right) + \left(\frac{k_{lock} \cdot \lambda}{k_{unlock} \cdot k_{rem}} \cdot \frac{1}{[G_c]} \right) + \frac{1}{k_1 \cdot [R_c]} \quad (5.41)$$

Next, we assumed: $w = \frac{[R_x]}{[R_c]}$ (5.42), $y = \frac{[G_x]}{[G_c]}$ (5.43), $z = \frac{r_x^{-1}}{r_c^{-1}}$ (5.44). Thus, dividing equation 5.40 by 5.41:

$$z = \left(\frac{\frac{k_{lock}}{k_{unlock} \cdot k_{rem}} \cdot \frac{k_p \cdot [R_c]}{k_1 \cdot [R_c]} \cdot \frac{1}{[G_c]}}{r_c^{-1}} + \frac{\frac{k_{lock} \cdot \lambda}{k_{unlock} \cdot k_{rem}} \cdot \frac{1}{[G_c]}}{r_c^{-1}} \right) \cdot \frac{1}{y} + \left(\frac{\frac{1}{k_1 \cdot [R_c]}}{r_c^{-1}} \right) \cdot \frac{1}{w} \quad (5.45)$$

Assuming: $\alpha = \frac{k_{lock} \cdot k_p \cdot [R_c]}{k_{unlock} \cdot k_{rem} \cdot [G_c]}$ (5.46), $\beta = \left(\frac{1}{\frac{k_1 \cdot [R_c]}{r_c^{-1}}} \right)$ (5.47), and $\eta = \frac{\frac{k_{lock} \cdot \lambda}{k_{unlock} \cdot k_{rem}} \cdot \frac{1}{[G_c]}}{r_c^{-1}}$ (5.48), equation 5.45 becomes:

$$z = (\alpha\beta + \eta) \cdot \frac{1}{y} + \beta \cdot \frac{1}{w} \tag{5.49}$$

In this regard, we used the empirical data of gene expression from two promoters ($LacO_3O_1$ and a native Lac promoter) to quantify the parameters of the model. Since these two promoters are present at the same location in the chromosome (in different strains), it is expected that the propensity of getting locked due to PSB because of factors other than their own transcription rate, is the same for both promoters. Finally, the removal of PSB is not expected to be different in the two promoters (mechanically or energetically). Hence:

$$z_{LacO3O1} = (\alpha\beta_1 + \eta) \cdot \frac{1}{y} + \beta_1 \cdot \frac{1}{w} \tag{5.50}$$

$$z_{Lac} = (\alpha\beta_2 + \eta \cdot X) \cdot \frac{1}{y} + \beta_2 \cdot \frac{1}{w} \tag{5.51}$$

where: $\beta_1 = \frac{\frac{1}{k_{1_{LacO3O1}} \cdot [R_c]}}{r_{c_{LacO3O1}}^{-1}}$ (5.52), $\beta_2 = \frac{\frac{1}{k_{1_{Lac}} \cdot [R_c]}}{r_{c_{Lac}}^{-1}}$ (5.53),

$\eta = \frac{\frac{k_{lock} \cdot \lambda}{k_{unlock} \cdot k_{rem}} \cdot \frac{1}{[G_c]}}{r_{c_{LacO3O1}}^{-1}}$ (5.54), and $X = \frac{r_{c_{LacO3O1}}^{-1}}{r_{c_{Lac}}^{-1}}$ (5.55).

The parameters α, β_1, β_2, *and* η should not be negative. They were estimated by finding the set of values for them that minimizes the mean squared error in equation 5.56:

$$MSE = \frac{\sum_{i=1}^{N_1}\left(Z^i_{LacO3O1} - \hat{Z}^i_{LacO3O1}(\alpha,\beta_1,\lambda)\right)^2 + \sum_{i=1}^{N_2}\left(Z^i_{Lac} - \hat{Z}^i_{Lac}(\alpha,\beta_2,\lambda,X)\right)^2}{N_1 + N_2} \quad (5.56)$$

5.3 Extraction of RNA production rates from microscopy data

We estimated the RNA production rate from empirical data on the mean RNA numbers per cell in cell populations obtained from two microscopy images taken at two different time moments (Häkkinen & Ribeiro, 2015; Zimmer, et al., 2016). We considered that the cells were dividing over time. This causes a "dilution" of RNAs by the daughter cells. The RNA 'dilution rate' due to cell division was considered in our estimation of RNA production rate.

The RNA dilution rate constant (k_{dil}) can be estimated from the number of cells at the start (t_i) and the end (t_f) of the measurements. Let M_i be the mean RNA per cell at the initial time moment t_i, while M_f is the mean RNA per cell at the time moment t_f. From this, the RNA production rate can be estimated as follows:

$$r = \frac{k_{dil}}{\log(2)} \cdot \frac{M_f - M_i \cdot e^{-k_{dil}\cdot(t_f - t_i)}}{1 - e^{-k_{dil}\cdot(t_f - t_i)}} \quad (5.57)$$

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

In **Publication III**, we derived a relationship between the statistics, i.e., the mean, variability, and asymmetries, of the single-cell distribution of cell fluorescence as

measured by flow cytometry (F/W) and the statistics of the single-cell RNA numbers distribution, as quantified by microscopy and image analysis.

Let N_B be the single-cell distribution of the number of free-floating MS2d-GFP molecules in the cells. Meanwhile, let N_R be the single-cell distribution of the number of mRNA molecules in the cell. Next, let I be the fluorescence intensity of 1 MS2d-GFP molecule, which is assumed to be constant, in agreement with GFP having approximately constant fluorescence intensity in a stable environment. Further, let BS be the number of MS2d-GFPs bound to each mRNA. BS is expected to be constant also, in that the MS2d-GFPs are known to quickly bind to all binding sites of the target RNA. In agreement, past studies (Tran, et al., 2015) produced empirical data suggesting that the tagged RNAs fluorescence is also approximately constant.

Given this, the single-cell distribution of total cell fluorescence intensities, F_T, is:

$$F_T = N_B \cdot I + N_R \cdot BS \cdot I \tag{5.58}$$

Next, we define $F_B = N_B \cdot I$, which is the single-cell distribution of background fluorescence intensity due to free-floating MS2d-GFPs (in the absence of MS2d-GFP tagged RNA spots). Further, we define $F_R = N_R \cdot BS \cdot I$, which is the single-cell distribution of MS2d-GFP fluorescence intensity due to the MS2d-GFP tagged RNAs (also called spots fluorescence intensity).

It is to be noted that we assumed that F_B and F_R are independent. This assumption should be approximately valid, provided that there are enough numbers of MS2d-GFP so that, when a target RNA is produced, the background fluorescence is not significantly decreased. For this to be (nearly) true, the MS2d-GFPs production must be much higher than the target RNA production, which is the case in our measurements.

From equation 5.58, one can write that the relationship between the mean of N_R, $M(N_R)$, and the mean of F_T (M(F_T)) is linear, as below:

$$M\left(N_R\right) = \frac{M\left(F_T\right)}{BS \cdot I} - \frac{M\left(N_B\right)}{BS} \tag{5.59}$$

Here, $\frac{1}{BS \cdot I}$ is the slope and $-\frac{M(N_B)}{BS}$ is the intercept.

Also, the relationship between the variance of N_R and the variance of F_T is linear, given the independence between them:

$$Var(N_R) = \frac{Var(F_T)}{BS^2 \cdot I^2} - \frac{Var(N_B)}{BS^2} \tag{5.60}$$

where, $\frac{1}{BS^2 \cdot I^2}$ is the slope and $-\frac{Var(N_B)}{BS^2}$ is the intercept.

From equation 5.60, the standard deviation (Sd) of N_R is:

$$Sd(N_R) = \sqrt{Var(N_R)} \tag{5.61}$$

Further, one also has that the relationship between the third moments of N_R and F_T is linear:

$$\mu_3(N_R) = \frac{\mu_3(F_T)}{BS^3 \cdot I^3} - \frac{\mu_3(N_B)}{BS^3} \tag{5.62}$$

where, $\frac{1}{BS^3 \cdot I^3}$ is the slope and $-\frac{\mu_3(N_B)}{BS^3}$ is the intercept. From equation 5.60 and 5.61, the skewness (S) of N_R is:

$$S(N_R) = \frac{\mu_3(N_R)}{Var(N_R)^{\frac{3}{2}}} \tag{5.63}$$

5.5 Uncertainty estimation in FC statistics using technical replicates of control cells

In **Publication III**, we estimated the uncertainty in the fluorescence of tagged RNAs (F/W) and in the fluorescence of the mRFP proteins that they code for (R/W).

Specifically, we estimated the uncertainty between the technical replicates in cells with the target gene, using the uncertainty in the fluorescence of the cells' background (measured using cells without the target gene) between the technical replicates.

First, the standard error of the mean of target cells' fluorescence was estimated as:

$$SE(M) = \sqrt{SE(M_T)^2 + SE(M_C)^2} \tag{5.64}$$

where $SE(M_T)$ is the standard error of the mean cell fluorescence in target cells, and $SE(M_C)$ is the standard deviation of the mean cell fluorescence in control cells from multiple conditions.

Next, the standard error of the Sd in target cells was estimated as:

$$SE(Sd) = \frac{1}{2}\sqrt{\frac{SE(Var_T)^2 + SE(Var_C)^2}{Var_T}} \tag{5.65}$$

where, $SE(Var_T)$ is the standard error of the variance in cell fluorescence in target cells, and $SE(Var_C)$ is the standard deviation of the variance in cell fluorescence in control cells from multiple conditions.

Finally, the standard error of the Skewness (S) in target cells was estimated as:

$$SE(S) \approx \frac{\sqrt{SE\left(\mu_3^T\right)^2 + SE\left(\mu_3^C\right)^2}}{\left(Sd_T + SE(Sd)\right)^3} \tag{5.66}$$

where $SE\left(\mu_3^T\right)$ is the standard error of the third moment of cell fluorescence in target cells, and $SE\left(\mu_3^C\right)$ is the standard deviation of the third moment of cell fluorescence in control cells from multiple conditions. In skewed distributions, it is expected that the variance and the third moment are correlated. To compensate for

this correlation, in equation 5.66, we summed the deviation to its error in the denominator.

6 SUMMARY OF THE RESULTS

In **Publication I,** we used stochastic simulations to investigate if the single-cell diversity in the RNA numbers expressed from a gene could be regulated by tuning the relative duration of the closed complex formation. In detail, our assumption was the following. Imagine two processes that take the same mean time, but one is a one-step process whose time interval between events follows an exponential distribution, while the other is a two-step process, with both steps following exponential distributions in time length. Interestingly, the second process will be less noisy than the first. Thus, it should be possible to regulate noise in RNA production, by regulating how long the closed and open complex formations take while maintaining the total time to produce an RNA to be constant.

In what concerns the resulting single-cell distributions of RNA numbers, the assumption above should be valid, regardless of whether the noise source is intrinsic or extrinsic. Here, we focused on the regulatability of the extrinsic noise by tuning the relative durations of the two rate-limiting steps of transcription. Let the relative duration of the closed complex formation be $t_{cc}/\Delta t$, where Δt is the mean RNA production time, while 't_{cc}' refers to the mean time length of a successful closed complex formation.

Meanwhile, to regulate extrinsic noise, we tuned the single-cell diversity in RNAP, and we studied the effects on the cell-to-cell variability in RNA numbers. We quantified the cell-to-cell variability in RNA and RNAP numbers using the squared coefficient of variance (CV^2).

We started by assuming the reversible model of transcription initiation proposed in (Lloyd-Price, et al., 2016) and we also have taken several parameter values from (Lloyd-Price, et al., 2016). Then, to introduce single-cell diversity in RNAP numbers in the model, we measured the CV^2 of RNAP fluorescence intensity per cell from microscopy data (we found that it equalled 0.03). Next, to obtain a realistic range of values for $t_{cc}/\Delta t$ (Table 1), we used the data published in (Kandavalli, et al., 2016).

Table 1. Empirical values of $t_{cc}/\Delta t$ of several promoters under several activation schemes

Promoter and induction	$t_{cc}/\Delta t$	Reference
PBAD (0.1% ara)	0.71	(Kandavalli, et al., 2016)
PBAD (0.01% ara)	0.55	(Kandavalli, et al., 2016)
PBAD (0.001% ara)	0.17	(Kandavalli, et al., 2016)
Plac-O1O3 (1 mM IPTG)	0.55	(Kandavalli, et al., 2016)
Plac-O1O3 (0.05 mM IPTG)	0.46	(Kandavalli, et al., 2016)
Pac-O1O3 (0.005 mM IPTG)	0.12	(Kandavalli, et al., 2016)
PtetA (no inducers)	0.07	(Kandavalli, et al., 2016)
Plac-O1 (1 mM IPTG)	0.33	(Kandavalli, et al., 2016)
Plac-ara1 (1 mM IPTG and 0.1% ara)	0.49	(Kandavalli, et al., 2016)

Finally, using the model, we explored the effects of different values for $t_{cc}/\Delta t$. For this, we varied k_1 and k_2 while keeping all the other rate constants and the mean transcription time unchanged. Centred on realistic values for $t_{cc}/\Delta t$, we then uniformly chose ten values between 0.05 and 0.95. For each $t_{cc}/\Delta t$, we then set 7 different values of CV^2 of RNAP between 0 and 0.09.

Having this, for each possible pair of values for $t_{cc}/\Delta t$ and $CV^2(RNAP)$, we simulated 1000 model cells, each for 10000 seconds, using SGNS2 software (Lloyd-Price , et al., 2012). From the results, we calculated the CV^2 of RNAs produced from model cells for each pair of values. Results are plotted in Figure 8. From Figure 8, one can conclude that the cell-to-cell variability in RNA numbers increases with increase in cell-to-cell variability in intracellular RNAP as well as with the relative duration of the closed complex formation, since the longer lasting is the latter, the more susceptible the system is to changes in the RNAP variability

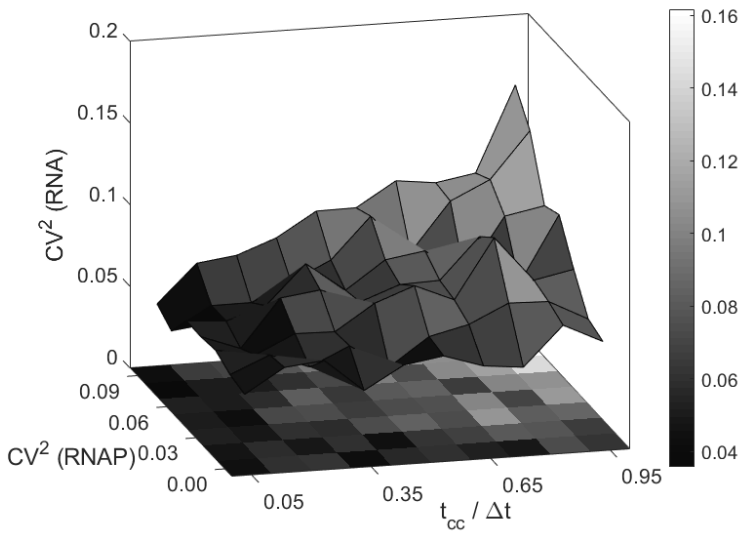


Figure 8. CV^2 of single-cell RNA numbers plotted as a function of the relative duration of the closed complex formation ($t_{cc}/\Delta t$) and the single-cell CV^2 of RNAP.

In **Publication II**, we proposed a methodology to dissect the rate-limiting steps in transcription, due to the effects of positive supercoiling buildup. Since gyrase proteins are the ones responsible for resolving the accumulated positive supercoils (Chong, et al., 2014), we first investigated how varying gyrase intracellular concentration would affect the kinetics of transcription. To vary this concentration, we introduced a plasmid in the cell that expresses the gyrase genes, gyrA, and gryB, under the control of a Rhamnose promoter.

Since the highly expressive genes can be supercoiling sensitive, this could affect the RNAP concentration. Thus, we measured RNAP levels at different gyrase concentrations and found that it changed by 12 % when the rhamnose concentration increased from 0 to 0.2 %.

To quantify the effects of changing gyrase concentration, we implemented a minimal model of transcription (where elongation is not considered, as it is not expected to affect the average transcription time) as in section 5.1.1. Then, using stochastic simulations, we showed that this minimal model was a good approximation of the more detailed single-nucleotide level model, in that the mean RNA production intervals of both models were almost identical.

Then, we derived an analytical solution of the mean RNA production time (r^{-1}) for the minimal model. It is a function of RNAP and gyrase concentration, with r^{-1} being the sum of the mean RNA production time of an unlocked/active promoter (τ_{active}) and the mean time a promoter is in a locked state (τ_{locked}). We expected that the average transcription interval decreases with an increase in gyrase concentration. This is because the average time that a promoter spends in the locked state should decrease since positive supercoils are being resolved faster.

In agreement with the model, measurements showed that increasing [G] decreased the average time spent in locked states per transcription event, $\tau_{locked,}$ of a chromosome-integrated gene under the control of the promoter $P_{LacO3O1.}$ For this, using microscopy, we first found that the time interval between RNA production events equalled 1476±145 s. Then, using qPCR, we estimated the mean intervals when gyrase was overexpressed. However, since the RNAP also varies with [G], we necessarily re-estimated the intervals to account for the changes in RNAP, so as to extract solely the effects of changing [G] (this step was referred to as 'RNAP correction').

The mean interval between RNA production events (r^{-1}) is expected to change linearly with $[G]^{-1}$, thus, it should be possible to get the time taken for transcription in the absence of positive supercoiling buildups (τ_{active}). For this, we extrapolated r^{-1} for when $[G]^{-1}=0$. Then, by subtracting τ_{active} from r^{-1}, it is possible to estimate τ_{locked} in the control condition. Hence, we then fitted a line to the data points of the graph relating the inverse of the gyrase concentration (relative to the control) and r^{-1}. From the point where the line intersects the Y-axis, we estimated τ_{active} to be 749±247 s. Based on this value, we then estimated the mean number of transcription events between consecutive locked states due to PSB to be ~1.8.

It is expected that r^{-1} of a plasmid-borne gene, under the regulation of the same promoter $P_{LacO3O1}$, should be approximately equal to τ_{active} of the chromosome-integrated gene. The reason is that the PSB effects are expected to be much weaker in plasmids, without topological constraints, that allow positive and negative supercoiling to move in opposite directions and annihilate each other (Chong, et al., 2014).

Using microscopy measurements of a promoter, $P_{LacO3O1}$, on a single-copy plasmid in control conditions, we estimated r^{-1} = 775±50 s. Then, under overexpressed

gyrase levels, we re-estimated r^{-1} after having corrected for the RNAP effects. We found that r^{-1} has not changed significantly due to the overexpression of gyrase, with τ_{active} not differing significantly from τ_{active} estimated from the chromosome construct.

Next, as a control experiment, we showed that only when inhibiting gyrase, there is the locking of transcription, confirming that the cause for the decreased transcription rate is the locking due to positive supercoiling buildup. Using time-lapse microscopy, we measured the mean integer-valued RNA numbers of the cell population, produced under the control of $P_{LacO3O1}$, every 15 min for 45 mins after the target gene induction and addition of Novobiocin, which is a gyrase inhibitor. When Novobiocin was added, RNA production almost stopped. Meanwhile, as a control experiment, the same was repeated without adding Novobiocin, and the production of RNAs did not cease.

We further showed that PSB effects differ with the basal transcription rate of the gene. We hypothesized that the gene's transcription will cause PSB. If it holds, then genes with longer RNA production intervals (i.e., slower transcription rate) should suffer less due to PSB, because gyrases have more time to resolve the PSB. To test this, we did the measurements on another promoter, P_{Lac} (with a transcription rate slower than $P_{LacO3O1}$), for which r^{-1} was estimated to be 2704±493 s. To assess whether the PSB effects differ with the promoter's transcription strength, we plotted r^{-1} versus the inverse of gyrase concentration relative to the control, for both chromosomal constructs (P_{Lac} and $P_{LacO3O1}$). The results showed that varying gyrase has lesser effects on the transcription rate of P_{Lac} as hypothesized.

In **Publication III**, we addressed a present problem in the study of transcription using single-cell biology techniques, which is that the constructs available could only be studied using microscopy. To address this, we proposed a method to estimate the statistics of single-cell RNA numbers, namely the mean, standard deviation, and skewness from flow cytometry data of total cell fluorescence, using cells capable of expressing RNA that can be tagged with MS2d-GFP, along with the free-floating MS2d-GFP. The estimation was shown to be possible by calibrating the fluorescence measured by flow cytometry with the single-cell integer-valued RNA number statistics (mean, standard deviation, and skewness) measured by microscopy.

For this study, we chose an *E. coli* bacterial strain carrying a single-copy plasmid with a promoter $P_{Lac/ara-1}$ (inducible by ara and IPTG) coding the RNA target for the

reporter MS2d-GFP, which is produced by a promoter $P_{LtetO-1}$ (inducible by aTc) placed on the multi-copy plasmid. For the control experiment, we used a similar bacterial strain, except that it lacks the target gene.

Using flow cytometry, we measured the total fluorescence (F/W) of cells with both the target and the reporter. We also measured cells with only the reporter. The target gene was induced by 0.1 % ara + different concentrations of IPTG (0 to 1000 μM), while the reporter gene was fully induced by 100 ng/μl aTc.

Then, we showed that the mean F/W of the cells having only the reporter gene remains unchanged irrespective of the IPTG concentration. This is expected since the inducers of the target gene should not affect the expression of the reporter gene. Meanwhile, the mean F/W of the cells having both the target and reporter genes increased with the IPTG concentration, due to the increased presence of RNAs tagged with MS2d-GFP. We also estimated the uncertainty in the mean, standard deviation, and skewness of F/W of these cells as described in section 5.5.

Next, using microscopy and image analysis, we calculated the mean, standard deviation, and skewness of the integer-valued RNA numbers of cells having both target and reporter systems. We found that the mean, standard deviation, and skewness of integer-valued RNA numbers are correlated with the respective statistics of F/W, measured by flow cytometry. From this, we hypothesized that it should be possible to quantify the single-cell RNA number statistics from the single-cell distribution of F/W of cells having both the target and reporter genes induced.

From section 5.4, it is expected that the first three moments of single-cell integer-valued RNA numbers from microscopy should be linearly correlated with F/W. In agreement, we plotted the first three moments of single-cell integer-valued RNA numbers against F/W and it was found to be highly correlated (p-value < 0.05 and $R^2 > 0.75$).

We further observed whether the increase in F/W is due to an increase in RNA numbers in the cells. For this, we considered that the increase in RNA expression should increase the protein expression since most of the gene regulation occurs in the transcription. Thus, to confirm whether the increase in F/W of the cells is due to the increased production of RNAs, we measured the fluorescence of mRFP proteins expressed by the target gene (R/W). We found that the statistics, mean, standard deviation, and skewness of F/W are highly correlated with R/W (p-value

< 0.05 and R^2 >0.85). Hence, it confirms that the F/W signal indeed captures the appearance of MS2d-GFP tagged RNA in the system.

Next, we calibrated the data on RNA fluorescence (F/W) obtained by flow cytometry with the integer-valued RNA numbers obtained by microscopy. The respective mean, second and third moments were calibrated independently (as explained in section 5.4). For the calibration, one needs microscopy data from at least two conditions (two data points) the same as obtained by flow cytometry. Choosing two conditions that differ significantly in IPTG induction achieves better calibration. Next, using the calibration, we estimated the M , Sd , and S of the single-cell RNA numbers from the cell fluorescence measurements (F/W) obtained from flow cytometry.

In **Publication IV**, we proposed an analytical model to explain the transcriptional interference that occurs in closely spaced promoters in tandem orientation. For this study, we obtained the list of known natural tandem promoters in *E. coli* from information available in RegulonDB, dated 30th January of 2020. Then, using the YFP protein fusion library, we measured the protein expression of these genes regulated by tandem promoters (only those genes which are available in YFP protein fusion library) by flow cytometry. We then fitted our analytical model to the measurements, in order to estimate the model parameters. Using these parameter values, we made predictions of the expected protein expression dynamics of tandem promoters, for various strengths and nucleotide distances between them.

In detail, first, we obtained the list of natural genes regulated by tandem promoters in the genome of *E. coli*. We used the RegulonDB database for this purpose. In total, we have found 102 pairs of tandem promoters, which do not have any genes in the region of the promoters or in between the promoters. Out of the 102, only 30 of them were tagged by YFP and made available in the YFP strain library (Taniguchi, et al., 2010). We measured their protein fluorescence by flow cytometry.

We then modelled the transcription and translation processes of tandem promoters while accounting for the transcription interference due to occlusion and downstream promoter occupancy, as described in section 5.1.2. Transcription interference due to occlusion was modelled by equations 5.10 and 5.11. This occlusion should increase with the RNAP occupancy of the other promoter and decrease with increasing d_{TSS}. Meanwhile, transcription interference due to downstream promoter occupancy was modelled by equation 5.15. The interference due to the occupancy of the

downstream promoter should increase with the occupancy of the downstream promoter.

Table 2. Potential models of transcriptional interference due to promoter occlusion as a function of d_{TSS}

Interference due to occlusion	$I(d_{TSS})$	$k_{occlusion}$
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")	$1 - \frac{1}{1 + e^{-m \cdot (d_{TSS} - L)}}$	$k_{ocl}^{\max} \left(1 - \frac{1}{1 + e^{-(d_{TSS} - L)}}\right) \cdot \omega$, for m = 1 bp^{-1}
Zero order ("ZeroO")	k	$k_{ocl}^{\max} \cdot \omega$

To account for transcriptional interference due to promoter occlusion as a function of d_{TSS} ($I(d_{TSS})$ in equations 5.10 and 5.11), we assumed three potential models (Exponential 1, Exponential 2 and Step), and a control, zero-order model (Table 2).

We then derived the analytical solution of the M, CV^2, and S of protein numbers at steady state as in section 5.1.2. We assumed that the promoters forming a pair in tandem orientation have similar strengths, since quantifying their exact strength in Nature is nontrivial. Also, if one promoter is much stronger than the other, we expect their behaviour to be similar to having only the strong promoter.

Having established the models and analytical solutions, we then measured the single-cell protein fluorescence distributions of 30 genes regulated by tandem promoters. From the results, we subtracted the background fluorescence and calculated the first three moments, using the method explained in section 5.4. In addition, we independently showed that the protein fluorescence of genes regulated by tandem promoters, measured using flow cytometry and microscopy were correlated (p-value<0.05).

Having established that the measurements are reliable, we converted the protein fluorescence of tandem promoter genes measured by flow cytometry to protein numbers. For this, we first plotted the protein fluorescence of these genes against

the corresponding protein numbers reported in (Taniguchi, et al., 2010) having the same experimental conditions. We then fitted the data with a line (without intercept). The best fit line had an $R^2 = 0.66$. Using it, we calculated the scaling factor to convert the protein fluorescence to protein numbers to be 0.09.

We further measured RNAP levels in the diluted media (0.5X and 0.25X) relative to the 1X M9 minimal media. We confirmed that the RNAP levels decreased slightly at 0.5X, while then remaining unchanged when the media was further diluted to 0.25X. We also measured the cells' doubling time in each media condition. These doubling times were almost identical in the 1X and 0.5X conditions (115±3 min). But for 0.25X, the growth rate slowed down and the doubling time increased. Since the growth changed significantly in 0.25X, we did not include data from that condition in our study.

We next quantified the relationship between the mean and the higher-order normalised moments, namely, the CV^2 and Skewness of single-cell protein numbers of genes regulated by tandem promoters. We plotted the $\log_{10}$ of the mean vs $\log_{10}$ of the CV^2 and of the Skewness of 1X and 0.5X data (in the same graph). We found that the relationship between the mean and the normalised moments (CV^2 and Skewness) did not differ between 1X and 0.5X. Thus, we fitted equation 5.22 to the data between mean and CV^2 to obtain parameter C_1. We also fitted equation 5.23 to the data between mean and Skewness to obtain parameter C_2.

Parameters C_1 and C_2 relate to each other as shown in 5.23. We observed that the values of C_1 slightly differ between the fittings when we fitted the equations 5.22 and 5.23 to the data independently. Then, we used the 'fminsearch' function in MATLAB (Lagarias, et al., 1998) to search for C_1 value that maximized the mean R^2 of both fittings. The search result was $C_1 = 72.71$ and $C_2 = 16.94$ with an R^2 of 0.80 for fitting the mean vs CV^2 and an R^2 of 0.61 for fitting the mean vs skewness.

We next quantified the regulatory effects of d_{TSS} and the promoter occupancy times on the protein expression dynamics of tandem promoters. For this, we plotted the $\log_{10}$ of mean, CV^2, and Skewness of protein numbers of tandem promoter genes as a function of d_{TSS}. We then fitted equations 5.20 and 5.21 to the scatter plot of d_{TSS} vs mean (using data from 1X M9 media). For this fitting, we assumed some of

the parameter values from the literature, and we estimated the other parameter values from the fitting.

Using these parameter values along with equations 5.20-5.23, we predicted the CV^2 and skewness in 1X M9 media as a function of d_{TSS}. We quantified the prediction accuracy using R^2 square. It equalled 0.43 for d_{TSS} vs CV^2 and 0.27 for d_{TSS} vs skewness when using the Step model. Also, we predicted the mean, CV^2 and skewness in 0.5X M9 media as a function of d_{TSS}. The mean R^2 equalled 0.15 for the Step model.

We next chose the best model of transcription occlusion. Of the four models in Table 2, the step model had a better overall R^2. From its fitting, we argue that if d_{TSS} is smaller than 35 bp (which equals the nucleotide distance occupied by an RNAP), the mean protein expression becomes significantly lower, in comparison to when d_{TSS} is larger than 35 bp.

We then investigated how the RNAP promoter occupancy of each of the tandem promoters and their d_{TSS} contribute to the mean expression of the upstream and downstream promoters. For this, we varied the promoter occupancy of each of these promoters by tuning k_{bind}, while keeping all other parameters unchanged. The promoter occupancy was varied between 0 and 1 for each promoter independently and we calculated the mean protein expression of the upstream and the downstream promoter, separately, for both when $d_{TSS} \leq 35$ bp and when $d_{TSS} > 35$ bp. We also calculated the total protein expression of the two tandem promoters for all possible combinations of upstream and downstream promoter occupancies.

We found that increasing the RNAP occupancy on the downstream promoter decreases the protein expression from the upstream promoter to a greater extent when $d_{TSS} \leq 35$ bp. Meanwhile, when $d_{TSS} > 35$, the protein expression of the upstream promoter did not decrease significantly when the occupancy of the downstream promoter was increased.

Likewise, we found that increasing the RNAP occupancy on the upstream promoter decreases the protein expression from the downstream promoter to a greater extent when $d_{TSS} \leq 35$ bp. Meanwhile, when $d_{TSS} > 35$, the protein expression of the

downstream promoter did not decrease significantly when the promoter occupancy of the upstream promoter was increased.

Finally, we have found that the total protein expression level from tandem promoters decreases when the occupancy of each of the promoters decreases, irrespective of d_{TSS}, as expected. Also, when the promoter occupancy of upstream and downstream promoters is high, the total expression from the tandem promoters decreases, when $d_{TSS} \leq 35$ bp. This could be due to increased interference between RNAPs occupying the upstream and downstream promoters. However, when $d_{TSS} > 35$ bp, the total expression from the tandem promoters did not decrease when both the promoters' occupancy is high.

Overall, having promoters in tandem configuration allows for the tuning of the overall kinetics in a manner that can alter the mean and variability in protein numbers differently than when having a single promoter. We found that the mean protein expression level of the genes with two closely spaced tandem promoters is less than twice the mean protein expression level of the genes with single promoters.

In **Publication V**, we studied how the transcription factor network (TFN) of *E. coli* regulates the global response of genes following the shift in RNAP concentration. In summary, first, we extracted the TFN of *E. coli* from the RegulonDB database. We then varied the intracellular RNAP concentration by diluting the media. Following the RNAP concentration shifts, we observed genome-wide changes in RNA numbers, which we measured using RNA-seq. We showed that the differences in the numbers of input TF activators and input TF repressors of a gene regulate its response.

In detail, we started by altering the intracellular concentration of RNAP by varying LB media richness (Lloyd-Price, et al., 2016). As expected, following the decreasing LB richness of the media (0.75X, 0.5X, and 0.25X), the RNAP concentration decreased inside the cells, while the growth rate did not vary significantly for at least 180 mins. Thus, we performed RNA-seq measurements at 180 mins, to quantify the mid-term response of genes across the genome, since this interval sufficed for genes to respond to the changes in RNAP and initial changes in TFs.

Next, we showed that the average magnitude of genes' response across the genome is correlated with the magnitude of RNAP shifts. Also, the number of differentially

expressed genes (DEGs) increased with the increase in the magnitude of the RNAP concentration shift.

We next investigated the potential role of global regulators other than RNAP in the global response of the genes. We found that the σ factors and global transcription factors/regulators were not influential in the mid-term genome-wide responses. Also, we found that the (p)ppGpp responsive genes responded only in the short-term but did not show unusual responses in the mid-term.

In addition, we found that noncoding sRNAs were not influential in the short and mid-term responses. Nevertheless, 14 out of 22 genes coding for rRNAs reported in the RegulonDB database showed unusual behaviours, for unknown reasons.

We next investigated the influence of the input TFs of the genes in their responses. First, we showed that the transcriptional response of genes having input TFs is greater than the transcriptional response of genes without having input TFs. Further, the magnitude of genes' response is correlated with the magnitude of its input TFs' changes.

We further showed that the responses of genes in the same operon or transcription units were correlated. So, the position of genes within each operon was not influential in their response to the TF changes, suggesting that there was no strong loss of "signal" when propagating the perturbation inside the operon.

We also showed that there is a correlation in RNA changes at 180 mins between the genes and transcription factors directly connected to them. However, there are less correlated responses between the genes and transcription factors that are separated by a pathlength of two or more. From this, we commented that the changes in the kinetics of direct input TFs are the ones that explain the genes mid-term responses.

Interestingly, we found that the magnitude of the transcriptional response of a gene cohort is correlated with the number of input TFs per gene. We also found that the correlation between each gene and its input TF decreases with the increasing number of input TFs.

Next, we investigated whether the increase in the magnitude of genes' response could be due to the increase in the asymmetry of the number of activators and repressors ($|b|$) of a gene. We have found that the average magnitude of the

response of a cohort of genes (grouped using an ensemble approach) increases with the mean asymmetry of the cohort. In addition, we have found that the mean asymmetry of a cohort (grouped based on the number of input TFs) increases with the increase in the number of TFs (k_{TF}) until $k_{TF}>5$. This explains how the increase in mean k_{TF} of the gene's cohort increases the average response of that cohort.

Finally, we confirmed that the genome-wide changes when subject to different LB media richness are mainly caused by the RNAP shift, as originally hypothesized. For this, we increased the media richness to 1.5X, 2.0X, and 2.5X respectively. Interestingly, we observed that the RNAP concentration of the media decreased despite the increase in media richness. We then again found a correlation between the responses of transcription factors and the genes which they directly control. Also, as expected, there was no correlation between the responses of the genes and their transcription factors separated by a path length of two or more.

7 DISCUSSION

In this thesis, we have studied some of the transcriptional regulatory mechanisms that affect the abundance of gene expression products, namely RNA and proteins, at the single-cell level. First, using an *insilico* model of gene transcription, we have described how the tuning of rate-limiting steps in transcription initiation regulates the abundance and diversity of RNAs in the cell population. We then focused on three different mechanisms by which it is possible to tune the kinetics of transcription initiation: i) Accumulation of positive coils due to local topological barriers, ii) Transcriptional interference between closely spaced promoters in tandem arrangement, and iii) Transcription factor binding. These studies resulted in five publications.

In **Publication I**, our aim was to understand how the rate-limiting steps in transcription initiation regulate the propagation of variability in RNAP numbers in cell population into the cell-to-cell variability in RNA numbers. We have answered this question by showing that varying the duration of the closed complex formation in transcription initiation, relative to the time between RNA production events, can alter the variability in transcripts abundance in a cell population. Due to this, it is possible to use this step kinetics as filter, regulator of the propagation of extrinsic noise (generated by the cell-to-cell variability in RNAP numbers) to the cell-to-cell variability in RNA numbers. Interestingly, promoters with relatively longer duration for closed complex formations are more prone to be affected by extrinsic noise, as such, over time, it is possible for a gene network to evolve which genes are much subject to this noise. Overall, evolving in this manner the diversity of RNA numbers at the genome-wide scale between sister cells may give organisms (or populations of single-celled organisms) an adaptive advantage in fluctuating environments.

In **Publication II**, we aimed to dissect the rate-liming steps in transcription due to transcription locking caused by positive supercoiling buildup. We have shown that by tuning the intracellular gyrase expression levels and measuring the resulting RNA numbers, it is possible to dissect these rate-limiting steps in transcription. The findings from this study could help synthetic biologists, when building novel genetic

constructs, to consider supercoiling as an influential factor in their dynamics. Meanwhile, recent studies have shown that genes that are quickly responsive to cold shock also have a higher probability of being supercoiling sensitive than the normal genes (Oliveira, et al., 2019). It would be interesting to apply this methodology to quantify the supercoiling sensitivity of cold-shock responsive genes.

In **Publication III**, we aimed to develop a method to quantify the mean and variability in single-cell RNA numbers using MS2-GFP tagging, from flow cytometry data. We proposed a method which estimates the statistics of single-cell RNA numbers, i.e., the mean, standard deviation, and skewness from the distribution of total cell fluorescence measured using flow cytometry. We expect that our proposed methodology would benefit the scientific community by enabling better quantification of single-cell RNA statistics, thereby adding more reliability to the conclusions. Although we used an MS2-GFP tagging method for this RNA quantification approach, we expect that the same quantification method can also be applied to other RNA tagging techniques, such as FISH. Meanwhile, one drawback of using MS2-GFP tagging is that the RNA quantification is less accurate when the cells express a smaller number of RNAs. This could be solved by adding more reporter binding sites to the target RNA, such that every newly produced RNA could produce strong enough fluorescence intensity so as to be easily differentiated from the background.

In **Publication IV**, we aimed to develop an analytical kinetical model able to predict the transcriptional interference between natural closely spaced tandem promoters. For this, we proposed a simplified model of tandem promoters based on chemical reactions. In the simplified model, transcription interference differs with the parameters of the individual promoters and the nucleotide distance between them. The knowledge acquired about the genes regulated by promoters in close proximity in tandem orientation can help to estimate gene expression. It can also guide synthetic biologists to build synthetic constructs, based on these promoters, with desired gene expression dynamics. Nevertheless, there is much potential for future improvements. For example, studies have shown that promoters in close proximity can be more sensitive to supercoiling (Jia, et al., 2017; Yeung, et al., 2017). In the future, the effects of supercoiling should be included in the models of closely spaced tandem promoters, so as to explore how to take advantage of this phenomenon. Another potential improvement in the models could be estimating some of its parameters using synthetic constructs. For example, synthetic tandem promoter

constructs of various promoter strengths and distances between the TSSs could be used to quantify with more precision the influence of these two parameters while removing the influence of unknown variables.

Finally, in **Study V**, we aimed to study the influences of the transcription factor network (logic and topology) on the genome-wide transcriptional responses to global cellular perturbations. From the experimental data, we have found that the genes' response following shifts in media richness (within a certain range), are mostly due to variations in RNAP and transcription factor numbers, and in accordance with the regulatory roles (activation or repression) of those transcription factors.

Meanwhile, we observed a strong linear correlation (slope, m=0.2 and y-intercept, c=0.1 with $R^2 = 0.5$ and $p < 0.05$) between LFC of RNA (LFC_{RNA}) and LFC of Proteins (LFC_{Prot}), when the media was diluted. The LFC_{Prot} is related to LFC_{RNA} as in the expression : $LFC_{Prot} = m \cdot LFC_{RNA} + \log_2\left(\frac{k_{pr}^{rich}}{k_{pr}^{dil}}\right)$, where $\frac{k_{pr}^{rich}}{k_{pr}^{dil}}$ is the ratio of post-transcription or translation rate constant between the rich and the diluted media. As $c = \log_2\left(\frac{k_{pr}^{rich}}{k_{pr}^{dil}}\right)$, $\frac{k_{pr}^{rich}}{k_{pr}^{dil}}$ can be approximated to 1.1. Hence, we claim that the regulatory effect due to post-transcription or translation did not change significantly between the moment the media was diluted and the moment the measurements were performed.

In summary, the models were an excellent guide for detecting which mechanisms are regulating the most gene expression in a desirable direction. Moreover, by modelling directly the regulatory steps of gene expression and by using realistic intervals of parameter values, it was possible to predict with some accuracy the values of those parameters beyond which significant changes in behaviour occurred.

As the genes are components of gene regulatory networks, it would be worthwhile to investigate if these findings can be applied to the networks. First, we could test on small network motifs, such as toggle switches, oscillators, and feed-forward

network motifs using stochastic simulations. The findings could then be tested on large networks.

In the future, large-scale TF network models may help us to explore transcriptional programs involving many genes over time following stresses, fluctuating environments, and antibiotics exposure. The potential findings could guide the development of better models of transcription regulation and of novel strategies for coping with antibiotic-resistant bacteria.

Finally, expanding these modelling techniques to the Mammalian genome would be interesting. However, in Mammalian genomes, the gene regulation process is more complicated than in *E. coli*, E.g., following transcription, the RNA has to undergo various post-transcriptional regulatory mechanisms such as interaction with RNA binding proteins (Gebauer, et al., 2021), alternative splicing (Wang, et al., 2015), RNA methylation (Zhou, et al., 2020), etc. These regulatory mechanisms, in turn affects the translation of RNA to proteins. Hence, incorporating these post-transcriptional regulatory mechanisms into the models would be a significant development towards quantifying the regulation, from those mechanisms.

8 BIBLIOGRAPHY

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

Acar, M., Mettetal, J. T. & van Oudenaarden, A., 2008. Stochastic switching as a survival strategy in fluctuating environments. *Nat Genet.,* 40(4), pp. 471-5.

Adhya, S. & Gottesman, M., 1982. Promoter occlusion: transcription through a promoter may inhibit its activity. *Cell,* 29(3), pp. 939-44.

Ahmad, S. S. et al., 2021. Bidirectional promoters: an enigmatic genome architecture and their roles in cancers. *Mol Biol Rep,* Volume 48, p. 6637–6644.

Arkin, A., Ross, J. & McAdams, H. H., 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics,* 149(4), pp. 1633-48.

Ay, A. & Arnosti, D. N., 2011. Mathematical modeling of gene expression: a guide for the perplexed biologist. *Crit Rev Biochem Mol Biol.,* 46(2), pp. 137-51.

Babu, M. M. & Teichmann, S., 2003. Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res.,* 31(4), pp. 1234-44.

Bahrudeen, M. N. et al., 2019. Estimating RNA numbers in single cells by RNA fluorescent tagging and flow cytometry. *J Microbiol Methods,* Volume 166, p. 105745.

Baptista, I. S. et al., 2022. Sequence-dependent model of genes with dual σ factor preference. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms,* 1865(3), p. 194812.

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

Benjamini, Y. & Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc.,* 57(1), p. 289–300.

Bintu, L. et al., 2005. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev.,* 15(2), pp. 116-24.

Blot, N. et al., 2006. Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome. *EMBO Rep.,* 7(7), p. 710–715.

Bolger, A. M., Lohse, M. & Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30(15), p. 2114–2120.

Bratsun, D., Volfson, D., Tsimring, L. S. & Hasty, J., 2005. Delay-induced stochastic oscillations in gene regulation. *Proc Natl Acad Sci U S A.,* 102(41), pp. 14593-8.

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

Browning, D. & Busby, S., 2004. The regulation of bacterial transcription initiation. *Nat Rev Microbiol,* Volume 2, p. 57–65.

Browning, D. F. & Busby, S. J., 2016. Local and global regulation of transcription initiation in bacteria. *Nat Rev Microbiol.,* 14(10), pp. 638-50.

Burgess, R. R., Travers, A. A., Dunn, J. J. & Bautz, E. K. F., 1969. Factor stimulating transcription by RNA polymerase. *Nature,* 221(5175), p. 43 – 46.

Burnette, W. N., 1981. "Western Blotting": Electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. *Anal Biochem.,* 112(2), pp. 195-203.

Bury-Moné, S. & Sclavi, B., 2017. Stochasticity of gene expression as a motor of epigenetics in bacteria: from individual to collective behaviors. *Res Microbiol.,* 168(6), p. 503–514.

Callen, B. P., Shearwin, K. E. & Egan, J. B., 2004. Transcriptional interference between convergent promoters caused by elongation over the promoter. *Mol Cell.,* 14(5), pp. 647-56.

Cao, Z., Filatova, T., Oyarzún, D. A. & Grima, R., 2020. A Stochastic Model of Gene Expression with Polymerase Recruitment and Pause Release. *Biophysical Journal,* 119(5), pp. 1002-1014.

Cheung, K. J. et al., 2003. A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of Escherichia coli. *Genome Res.,* 13(2), pp. 206-15.

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

Crampton, N. et al., 2006. Collision events between RNA polymerases in convergent transcription studied by atomic force microscopy. *Nucleic Acids Research,* 34(19), p. 5416–5425.

Crick, F., 1970. Central Dogma of Molecular Biology. *Nature,* Volume 227, p. 561–563.

Croucher, N. J. & Thomson, N. R., 2010. Studying bacterial transcriptomes using RNA-seq. *Curr Opin Microbiol.,* 13(5), pp. 619-24.

Cunningham, A., 1990. Fluorescence pulse shape as a morphological indicator in the analysis of colonial microalgae by flow cytometry. *J. Microbiol. Methods,* Volume 11, p. 27–36.

Dash, S. et al., 2021. Positive supercoiling buildup is a trigger of E. coli's short-term response to cold shock. *bioRxiv*.

Datsenko, K. A. & Wanner, B. L., 2000. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci U S A.,* 97(12), pp. 6640-5.

Day, R. N. & Davidson, M. W., 2009. The fluorescent protein palette: tools for cellular imaging. *Chem Soc Rev.,* 38(10), pp. 2887-921.

deHaseth, P. L., Zupancic, M. L. & Record, M. T., 1998. RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. *J Bacteriol.,* 180(12), pp. 3019-25.

Deng, S., Stein, R. A. & Higgins, N. P., 2005. Organization of supercoil domains and their reorganization by transcription. *Mol. Microbiol.,* 57(6), pp. 1511-1521.

Dobin, A. et al., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29(1), p. 15–21.

Dove, S. L., Darst, S. A. & Hochschild, A., 2003. Region 4 of sigma as a target for transcription regulation. *Mol Microbiol.,* 48(4), pp. 863-74.

Duchi, D. et al., 2016. RNA polymerase pausing during initial transcription. *Mol. Cell,* 63(6), p. 939–950.

Ebright, R. H., 1993. Transcription activation at Class I CAP-dependent promoters. *Mol Microbiol.,* 8(5), pp. 797-802.

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

Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S., 2002. Stochastic gene expression in a single cell. *Science,* 297(5584), p. 1183–1186.

Engl, C., 2019. Noise in bacterial gene expression. *Biochem Soc Trans,* 47(1), p. 209–217.

Engl, C. et al., 2020. The route to transcription initiation determines the mode of transcriptional bursting in E. coli. *Nature Communications,* Volume 11, p. 2422.

Fakhouri, W. D. et al., 2010. Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo. *Mol Syst Biol.,* Volume 6, p. 341.

Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H., 1998. Visualization of single RNA transcripts in situ. *Science,* 280(5363), pp. 585-90.

Fusco, D. et al., 2003. Single mRNA Molecules Demonstrate Probabilistic Movement in Living Mammalian Cells. *Curr. Biol.,* 13(2), p. 161–7.

Gama-Castro, S. et al., 2011. egulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Research,* Volume 39, p. D98–D105.

Garcia, H. G. & Phillips, R., 2011. Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci U S A.,* 108(29), pp. 12173-12178.

Gebauer, F., Schwarzl, T., Valcárcel, J. & Hentze, M. W., 2021. RNA-binding proteins in human genetic disease. *Nature Reviews Genetics,* Volume 22, p. 185–198.

Gibson, M. A. & Bruck, J., 2000. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *J. Phys. Chem. A,* 104(9), p. 1876–1889.

Gillespie, D. T., 1977. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry,* 81(25), p. 2340 – 2361.

Gillespie, D. T., 2000. The chemical Langevin equation. *J. Chem. Phys.,* 113(1), p. 297 – 306.

Gillespie, D. T., 2007. Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem.,* Volume 58, pp. 35-55.

Golding, I. & Cox, E. C., 2004. RNA dynamics in live Escherichia coli cells. *Proc Natl Acad Sci U S A.,* 101(31), pp. 11310-11315.

Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C., 2005. Real-time kinetics of gene activity in individual bacteria. *Cell,* 123(6), pp. 1025-36.

Gupta, A. & Mendes, P., 2018. An overview of network-based and -free approaches for stochastic simulation of biochemical systems. *Computation,* 6(1), p. 9.

Hayakawa, Y., Murotsu, T. & Matsubara, K., 1985. Mini-F protein that binds to a unique region for partition of mini-F plasmid DNA. *J. Bacteriol.,* 163(1), p. 349–354.

He, X., Samee, M. A., Blatti, C. & Sinha, S., 2010. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol.,* Volume 6, p. e1000935.

Hochschild, A. & Dove, S. L., 1998. Protein–Protein Contacts that Activate and Repress Prokaryotic Transcription. *Cell,* 92(5), pp. 597-600.

Hoops, S. et al., 2006. COPASI—a COmplex PAthway SImulator. *Bioinformatics,* 22(24), p. 3067–3074.

Huh, D. & Paulsson, J., 2011. Random partitioning of molecules at cell division. *Proc Natl Acad Sci U S A.,* 108(36), pp. 15004-9.

Häkkinen, A., Kandhavelu, M., Garasto, S. & Ribeiro, A. S., 2014. Estimation of fluorescence-tagged RNA numbers from spot intensities. *Bioinformatics,* 30(8), p. 1146–1153.

Häkkinen, A. et al., 2013. CellAging: a tool to study segregation and partitioning in division in cell lineages of Escherichia coli. *Bioinformatics,* 29(13), p. 1708–1709.

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. *J. R. Soc. Interface.,* 16(161).

Häkkinen, A. & Ribeiro, A. S., 2015. Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data. *Bioinformatics,* 31(1), p. 69–75.

Jia, J. et al., 2017. Three tandem promoters, together with IHF, regulate growth phase dependent expression of the Escherichia coli kps capsule gene cluster. *Sci Rep.,* Volume 7, p. 17924.

Johansson, H. E. et al., 1998. A thermodynamic analysis of the sequence-specific binding of RNA by bacteriophage MS2 coat protein. *Proc Natl Acad Sci U S A.,* 95(16), pp. 9244-9.

Jones, D. L., Brewster, R. C. & Phillips, R., 2014. Promoter architecture dictates cell-to-cell variability in gene expression. *Science,* 346(6216), pp. 1533-6.

Kærn, M., Elston, T., Blake, W. & Collins, J., 2005. Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet,* 6(6), p. 451–464.

Kandavalli, V. K., Tran, H. & Ribeiro, A. S., 2016. Effects of σ factor competition on the in vivo kinetics of transcription initiation in Escherichia coli. *BBA Gene Regulatory Mechanisms,* Volume 1859, p. 1281–1288.

Kouzine, F. et al., 2013. Transcription-dependent dynamic supercoiling is a short-range genomic force. *Nat Struct Mol Biol.,* 20(3), pp. 396-403.

Krummel, B. & Chamberlin, M. J., 1992. Structural analysis of ternary complexes of Escherichia coli RNA polymerase. Deoxyribonuclease I footprinting of defined complexes. *J Mol Biol.,* 225(2), pp. 239-50.

Kussell, E. & Leibler, S., 2005. Phenotypic diversity, population growth, and information in fluctuating environments. *Science,* 309(5743), pp. 2075-8.

Lagarias, J. C., Reeds, J. A., Wright, M. H. & Wright, P. E., 1998. Convergence Properties of the Nelder—Mead Simplex Method in Low Dimensions. *SIAM J Optim.,* 9(1), p. 112–147.

Langmead, B. & Salzberg, S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods.,* 9(4), p. 357–359.

Lee, D. J., Minchin, S. D. & Busby, S. J., 2012. Activating transcription in bacteria. *Annu Rev Microbiol.,* 66(1), pp. 125-52.

Le, T. B., Imakaev, M. V., Mirny, L. A. & Laub, M. T., 2013. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science,* 342(6159), p. 731–734.

Liao, Y., Smyth, G. K. & Shi, W., 2019. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research,* 47(8), p. e47.

Lilley, D. M. & Higgins, C. F., 1991. Local DNA topology and gene expression: the case of the leu-500 promoter. *Mol Microbiol.,* 5(4), pp. 779-83.

Lind, K., Ståhlberg, A., Zoric, N. & Kubista, M., 2006. Combining sequence-specific probes and DNA binding dyes in real-time PCR for specific nucleic acid quantification and melting curve analysis. *BioTechniques,* 40(3), p. 315 – 319.

Lin, J. & Amir, A., 2021. Disentangling Intrinsic and Extrinsic Gene Expression Noise in Growing Cells. *Phys. Rev. Lett.,* 126(7), p. 078101.

Livak, K. J. & Schmittgen, T. D., 2001. Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods,* 25(4), pp. 402-408.

Lloyd-Price , J., Gupta, A. & Ribeiro, A. S., 2012. SGNS2: a compartmentalized stochastic chemical kinetics simulator for dynamic cell populations. *Bioinformatics,* 28(22), pp. 3004-5.

Lloyd-Price, J. et al., 2016. Dissecting the stochastic transcription initiation process in live Escherichia coli. *DNA Research,* 23(3), pp. 203-214.

Lok, L. & Brent, R., 2005. Automatic generation of cellular reaction networks with Moleculizer 1.0. *Nat Biotechnol.,* 23(1), pp. 131-6.

Love, M. I., Huber, W. & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol,* 15(12), p. 550.

Luo, R., Ye, L., Tao, C. & Wang, K., 2013. Simulation of E. coli Gene Regulation including Overlapping Cell Cycles, Growth, Division, Time Delays and Noise. *PLoS ONE,* 8(4), p. e62380.

Lutz, R., 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 Res.,* 29(18), pp. 3873-3881.

Maarleveld, T. R., Olivier, B. G. & Bruggeman, F. J., 2013. StochPy: a comprehensive, user-friendly tool for simulating stochastic biological processes. *PLoS One,* 8(11), p. e79345.

Mahmood, T. & Yang, P. C., 2012. Western blot: technique, theory, and trouble shooting. *N Am J Med Sci.,* 4(9), pp. 429-34.

Ma, J., Bai, L. & Wang, M. D., 2013. Transcription under torsion. *Science,* 340(6140), p. 1580–1583.

Martínez-Antonio, A. & Collado-Vides, J., 2003. Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiolog,* 6(5), pp. 482-489.

Martínez-Antonio, A., Janga, S. C. & Thieffry, D., 2008. Functional organisation of Escherichia coli transcriptional regulatory network. *J Mol Biol.,* 381(1), pp. 238-47.

Martins, L. et al., 2012. Dynamics of transcription of closely spaced promoters in Escherichia coli, one event at a time. *J Theor Biol.,* Volume 301, pp. 83-94.

Martins, L. et al., 2018. SCIP: a single-cell image processor toolbox. *Bioinformatics,* 34(24), p. 4318–4320.

McAdams, H. H. & Arkin, A., 1997. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences of the United States of America,* 94(3), p. 814 – 819.

McAdams, H. H. & Arkin, A., 1999. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.,* 15(2), pp. 65-9.

McClure, W. R., 1980. Rate-limiting steps in RNA chain initiation. *Proc Natl Acad Sci U S A.,* 77(10), p. 5634 – 5638.

McClure, W. R., 1985. Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem,* Volume 54, p. 171–204.

McDonald, J. C. et al., 2000. Fabrication of microfluidic systems in poly(dimethylsiloxane). *Electrophoresis,* 21(1), pp. 27-40.

McLeod, S. M. & Johnson, R. C., 2001. Control of transcription by nucleoid proteins. *Curr Opin Microbiol.,* 4(2), pp. 152-9.

Metelev, M. et al., 2022. Direct measurements of mRNA translation kinetics in living cells. *Nature Communications,* Volume 13, p. 1852.

Mileyko, Y., Joh, R. I. & Weitz, J. S., 2008. Small-scale copy number variation and large-scale changes in gene expression. *Proc Natl Acad Sci U S A.,* 105(43), pp. 16659-64.

Mitarai, N., Dodd, I., Crooks, M. & Sneppen, K., 2008. The Generation of Promoter-Mediated Transcriptional Noise in Bacteria. *PLoS Comput. Biol.,* 4(7), p. e1000109.

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,* Volume 10, p. 59.

Mori, H. et al., 1986. Structure and function of the F plasmid genes essential for partitioning. *J. Mol. Biol.,* 192(1), p. 1–15.

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

Morrison, M., Razo-Mejia, M. & Phillips, R., 2021. Reconciling kinetic and thermodynamic models of bacterial transcription. *PLOS Computational Biology,* 17(1), p. e1008572.

Muthukrishnan, A. B. et al., 2012. Dynamics of transcription driven by the tetA promoter, one event at a time, in live Escherichia coli cells. *Nucleic Acids Research,* 40(17), p. 8472–8483.

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. *J. Mol.Biol.,* 257(1), pp. 21-29.

Mäkelä, J. et al., 2011. Automatic detection of changes in the dynamics of delayed stochastic gene networks and in vivo production of RNA molecules in Escherichia coli. *Bioinformatics,* 27(19), p. 2714–2720.

Naughton, C. et al., 2013. Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nat Struct Mol Biol.,* 20(3), pp. 387-95.

Nudler, E. & Gottesman, M. E., 2002. Transcription termination and anti-termination in E. coli. *Genes Cells,* 7(8), p. 755 – 768.

Oehler, S. et al., 1994. Quality and position of the three lac operators of E. coli define efficiency of repression. *EMBO J.,* 13(14), pp. 3348-55.

Oehler, S., Eismann, E., Krämer, H. & Müller-Hill, B., 1990. The three operators of the lac operon cooperate in repression. *The EMBO journal,* 9(4), pp. 973-979.

Oliveira, S. M. et al., 2019. Chromosome and plasmid-borne PLacO3O1 promoters differ in sensitivity to critically low temperatures. *Sci. Rep.,* 9(1), p. 4486.

Patange, O. et al., 2018. Escherichia coli can survive stress by noisy growth modulation. *Nat Commun,* Volume 9, p. 5333.

Peabody, D. S., 1993. The RNA binding site of bacteriophage MS2 coat protein. *EMBO J.,* 12(2), pp. 595-600.

Pérez-Rueda, E. & Collado-Vides, J., 2000. The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Research,* 28(8), p. 1838–1847.

Peterson, J. R. et al., 2015. Effects of DNA replication on mRNA noise. *Proc Natl Acad Sci U S A.,* 112(52), pp. 15886-91.

Prescott, E. M. & Proudfoot, N. J., 2002. Transcriptional collision between convergent genes in budding yeast. *Proc Natl Acad Sci U S A.,* 99(13), pp. 8796-801.

Queimadelas, C. et al., 2012. Segmentation and tracking of Escherichia coli expressing tsr-venus proteins from combined DIC/Fluorescence images. *In fifth International Conference on MEDSIP. Liverpool, UK.*

Raj, A. et al., 2008. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods,* Volume 5, p. 877–879.

Ramakrishnan, V., 2002. Ribosome structure and the mechanism of translation. *Cell,* 108(4), pp. 557-72.

Raser, J. M. & O'Shea, E. K., 2005. Noise in gene expression: origins, consequences, and control. *Science,* 309(5743), pp. 2010-3.

Razo-Mejia, M. et al., 2018. Tuning Transcriptional Regulation through Signaling: A Predictive Theory of Allosteric Induction. *Cell Syst.,* 6(4), pp. 456-469.

Razo-Mejia, M. et al., 2020. First-principles prediction of the information processing capacity of a simple genetic circuit. *Phys. Rev. E,* 102(2), p. 022404.

Ribeiro, A. S., 2010. Stochastic and delayed stochastic models of gene expression and regulation. *Math Biosci.,* 223(1), pp. 1-11.

Roussel, M. R. & Zhu, R., 2006. Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. *Phys Biol.,* 3(4), pp. 274-84.

Rovinskiy, N. et al., 2012. Rates of Gyrase Supercoiling and Transcription Elongation Control Supercoil Density in a Bacterial Chromosome. *PLOS Genetics,* 8(8), p. e1002845.

Ruff, E. F. et al., 2015. E. coli RNA Polymerase Determinants of Open Complex Lifetime and Structure. *J Mol Biol.,* 427(15), pp. 2435-2450.

Sanchez, A. & Golding, I., 2013. Genetic determinants and cellular constraints in noisy gene expression. *Science,* 342(6163), pp. 1188-93.

Santos-Zavaleta, A. 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(D1), pp. D212-D220.

Schlax, P. J., Capp, M. W. & Record, M. T., 1995. Inhibition of transcription initiation by lac repressor. *J Mol Biol.,* 245(4), pp. 331-50.

Schmittgen, T. D. & Livak, K. J., 2008. Analyzing real-time PCR data by the comparative CT method. *Nature Protocols,* 3(6), pp. 1101-1108.

Segal, E. et al., 2008. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature,* Volume 451, p. 535–540.

Shaner, N. C., Steinbach, P. A. & Tsien, R. Y., 2005. A guide to choosing fluorescent proteins. *Nature Methods,* 2(12), p. 905 – 909.

Shearwin, K. E., Callen, B. P. & Egan, J. B., 2005. Transcriptional interference – a crash course. *Trends in Genetics,* 21(6), pp. 339-345.

Shimomura, O., Johnson, F. H. & Saiga, Y., 1962. Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, Aequorea. *J Cell Comp Physiol.,* 59(3), pp. 223-39.

Skinner, S., Sepúlveda, L., Xu, H. & Golding, I., 2013. Measuring mRNA copy number in individual Escherichia coli cells using single-molecule fluorescent in situ hybridization. *Nat Protoc,* 8(6), p. 1100–1113.

Sneppen, K. et al., 2005. A Mathematical Model for Transcriptional Interference by RNA Polymerase Traffic in Escherichia coli. *J Mol Biol.,* 346(2), pp. 399-409.

So, L. H. et al., 2011. General properties of transcriptional time series in Escherichia coli. *Nat Genet.,* 43(6), pp. 554-60.

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 Genet.,* 5(10), p. e1000671.

Suzuki, T. et al., 2007. Recent advances in fluorescent labeling techniques for fluorescence microscopy. *Acta Histochem Cytochem.,* 40(5), pp. 131-7.

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

Taniguchi, Y. et al., 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science,* 329(5991), pp. 533-8.

Teves, S. S. & Henikoff, S., 2014. Transcription-generated torsional stress destabilizes nucleosomes. *Nat Struct Mol Biol.,* 21(1), pp. 88-94.

Tran, H., Oliveira, S. M., Goncalves, N. & Ribeiro, A. S., 2015. Kinetics of the cellular intake of a gene expression inducer at high concentrations. *Mol. BioSyst.,* Volume 11, p. 2579–2587.

Tripathi, S., Brahmachari, S., Onuchic, J. N. & Levine, H., 2022. DNA supercoiling-mediated collective behavior of co-transcribing RNA polymerases. *Nucleic Acids Research,* 50(3), p. 1269–1279.

Tsien, R. Y., 1998. The green fluorescent protein. *Annual Review of Biochemistry,* Volume 67, p. 509 – 544.

Uptain, S. M., Kane, C. M. & Chamberlin, M. J., 1997. Basic mechanisms of transcript elongation and its regulation. *Annu. Rev. Biochem.,* 66(1), p. 117 – 172.

Wang, F. et al., 2013. The promoter-search mechanism of Escherichia coli RNA polymerase is dominated by three-dimensional diffusion. *Nat. Struct Mol. Biol.,* 20(2), p. 174 – 181.

Wang, Y. et al., 2015. Mechanism of alternative splicing and its regulation. *Biomed Rep,* 3(2), pp. 152-158.

Ward, D. F. & Murray, N. E., 1979. Convergent transcription in bacteriophage lambda: interference with gene expression. *J Mol Biol.,* 133(2), pp. 249-66.

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 protein. *Photochemistry and Photobiology,* 31(6), p. 611 – 615.

Weinstein-Fischer, D., Elgrably-Weiss, M. & Altuvia, S., 2000. Escherichia coli response to hydrogen peroxide: a role for DNA supercoiling, topoisomerase I and Fis. *Mol Microbiol.,* 35(6), pp. 1413-20.

Xie, X. S. et al., 2008. Single-Molecule Approach to Molecular Biology in Living Bacterial Cells. *Annual Review of Biophysics,* 37(1), p. 417–444.

Yang, S. et al., 2014. Contribution of RNA polymerase concentration variation to protein expression noise. *Nat Commun,* Volume 5, p. 4761.

Yeung, E. et al., 2017. Biophysical Constraints Arising from Compositional Context in Synthetic Gene Networks. *Cell Syst.,* 5(1), pp. 11-24.e12.

Yu, D. et al., 2000. An efficient recombination system for chromosome engineering in Escherichia coli. *Proc Natl Acad Sci U S A.,* 97(11), pp. 5978-83.

Yu, J. et al., 2006. Probing gene expression in live cells, one protein molecule at a time. *Science,* 311(5767), pp. 1600-3.

Zhou, Y. et al., 2020. Principles of RNA methylation and their implications for biology and medicine. *Biomedicine & Pharmacotherapy,* Volume 131, p. 110731.

Zhu, R., Ribeiro, A. S., Salahub, D. & Kauffman, S. A., 2007. Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models. *J Theor Biol.,* 246(4), pp. 725-45.

Zimmer, C., Häkkinen, A. & Ribeiro, A. S., 2016. Estimation of kinetic parameters of transcription from temporal single-RNA measurements. *Math Biosci.,* Volume 271, pp. 146-53.

Publications

PUBLICATION I

Effects of extrinsic noise are promoter kinetics dependent.

M.N.M. Bahrudeen, S. Startceva, A.S. Ribeiro

In Proceedings of the 9th International conference on Bioinformatics and Biomedical Technology (ICBBT 2017).

https://doi.org/ 10.1145/3093293.3093295

Effects of Extrinsic Noise are Promoter Kinetics Dependent

Mohamed N. M. Bahrudeen
Laboratory of Biosystem Dynamics, Biomeditech, Tampere University of Technology, Finland.
P.O Box 553, 33101 Tampere, Finland.
mohamed.mohamedbahrudeen@tut.fi

Sofia Startceva
Laboratory of Biosystem Dynamics, Biomeditech, Tampere University of Technology, Finland.
P.O Box 553, 33101 Tampere, Finland.
sofia.startceva@tut.fi

Andre S. Ribeiro
Laboratory of Biosystem Dynamics, Biomeditech, Tampere University of Technology, Finland.
Office TC338, P.O Box 553, 33101 Tampere, Finland.
Phone: +358 408490736.
andre.ribeiro@tut.fi

ABSTRACT
Studies in *Escherichia coli* using *in vivo* single-RNA detection and time-lapse confocal microscopy showed that transcription is a multiple rate-limiting steps process, in agreement with previous *in vitro* measurements. Here, from simulations of a stochastic model of transcription validated empirically that accounts for cell-to-cell variability in RNA polymerase (RNAP) numbers, we investigate the hypothesis that the cell-to-cell variability in RNA numbers due to RNAP variability differs with the promoter rate-limiting steps dynamics. We find that increasing the cell-to-cell variability in RNAP numbers increases the cell-to-cell diversity in RNA numbers, but the degree with which it increases is promoter kinetics dependent. Namely, promoters whose open complex formation is relatively longer lasting dampen more efficiently this noise propagation phenomenon. We conclude that cell-to-cell variability in RNA numbers due to variability in RNAP numbers is promoter-sequence dependent and, thus, evolvable.

CCS Concepts
• Computing methodologies → Modeling and simulation → Model development and analysis

Keywords
Phenotypic diversity; gene expression; extrinsic noise; transcription initiation; rate-limiting steps; stochastic models.

1. INTRODUCTION
In *Escherichia coli*, major behavioral changes including metabolic, are driven by changes in the numbers of the molecules composing the transcriptional and translational machineries, such as RNA polymerase (RNAP) core enzymes and σ factors [1]. E.g., changes in σ factor numbers allow *E. coli* cells to quickly, and simultaneously, enhance and/or reduce the transcriptional activity of a large number of genes in a selective fashion [2].

ICBBT '17, May 14-16, 2017, Lisbon, Portugal

ACM ISBN 978-1-4503-4879-9/17/05…$15.00
http://dx.doi.org/10.1145/3093293.3093295

This is made possible by the limited number of RNAP core enzymes [3]. As the numbers of a specific σ factor increase, the RNAPs carrying that σ factor increase, and thus the activity of the promoters associated to that σ factor is expected to increase by direct positive regulation. Meanwhile, the activity of the promoters associated to other σ factors is expected to decrease by indirect negative regulation.

Interestingly, following changes in σ factor numbers, while these 'expectations' based on present models of transcription do, in general, take place on a global scale, some genes' activity is unaffected [3], and those that respond do so heterogeneously, i.e., they differ in degree of change, even in the case of genes associated to the same σ factor.

This 'behavioral' diversity in responses is due to differences in the promoters' selectivity for σ factors [4], the action of transcription factors [3] and, according to a recent study, in the case of indirect negative regulation, due to differing dynamics of the multiple steps in transcription initiation [5, 6], which were first observed by *in vitro* measurement techniques (for a review see [7]). In particular, it was shown that in promoters preferentially transcribed by σ^{70}, the smaller the time-scale of the closed complex formation relative to the open complex formation, the weaker is the promoter's responsiveness to changes in σ^{38} numbers. It was thus concluded that, in *E. coli*, a promoter's responsiveness to indirect regulation by σ factor competition is determined by the kinetics of the irate limiting steps in initiation.

Given this observation, validated by various measurement techniques of RNA production dynamics applied to several promoters [6], we here hypothesize that the dynamics of the rate-limiting steps in transcription initiation [7] influences also a gene's degree of responsiveness to extrinsic noise sources.

Here, we investigate this hypothesis by, first, establishing a stochastic model of transcription that accounts for cell-to-cell diversity in RNA polymerase numbers and whose parameter values are taken from state-of-the-art, single-cell RNAP levels and single-RNA microscopy measurements, and then performing stochastic simulations of model cells [8, 9] carrying the multi-step stochastic model of transcription [10] and whose RNAP numbers are, while constant in time, initially randomly drawn from a normal distribution. By tuning this model's parameter values, we assess to which extent variability in RNAP numbers, as function of transcription initiation kinetics, affects the cell-to-cell diversity in RNA numbers.

2. METHODS

2.1 Model of Transcription

We consider a dynamically broad model of transcription initiation that allows RNA production kinetics to range from sub-Poissonian to super-Poissonian, depending on the rate constant values. This model was derived from data from multiple studies, including genome-wide studies of RNA numbers variability [11, 2] and of the transcription dynamics of individual genes [13, 14].

The model includes the steps in transcription initiation in *E. coli* (e.g. open complex formation [15] and ON/OFF process [16]). Rate constant values were obtained by fitting the model to empirical data on RNA production kinetics at the molecule level of the Lac-Ara-1 promoter and from single-cell measurements of intracellular concentration of RNAPs reported in [5].

This model is applicable to common promoters in *E. coli*, differing between promoters in the rate constant values, and it consists of the following reactions:

$$P_{\mathrm{ON}} + R \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} RP_c \xrightarrow{k_2} RP_o \xrightarrow{k_3} R + P_{\mathrm{ON}} + \mathrm{RNA} \quad (1)$$

$$P_{\mathrm{ON}} \underset{k_{ON}}{\overset{k_{OFF}}{\rightleftarrows}} P_{\mathrm{OFF}} \quad (2)$$

Reactions (1) represent the multi-step transcription initiation of an active promoter, P_{ON} [17]. First, the closed complex (RPc) is formed as the RNAP (R) binds to a free promoter [18]. Next, intermediate steps occur to form the open complex (RPo) [17], [18]. Finally, elongation begins after promoter clearance [19], after which the promoter, the produced RNA, and the RNAP are released. In (1), k_1 is the rate at which RNAPs find and bind to the promoter successfully. k_2 is the open complex formation rate. Finally, k_3 is the promoter escape rate. k_{-1} is the rate of reversibility of the closed complex. In this model, a promoter occupied by an RNAP is unavailable to other initiation events.

Reactions (2) represent the intermittent transitioning of the promoter to an inactive state (P_{OFF}) due to e.g. binding/unbinding of repressors/activators [20], accumulation of positive DNA supercoiling [21], etc.

As the number of RNAPs differs between live cells (see measurements below), in each model cell the number of RNAPs is constant but initially randomly generated from a normal distribution, N(x,y), where x and y are obtained from empirical data [1]. This is the source of extrinsic noise of the model cell population here considered and is the main innovation of our model when compared to previous stochastic models [5, 10, 22].

2.2 Stochastic Simulations

Simulations are performed by SGNS [8], a simulator of chemical reaction systems whose dynamics is driven by the Stochastic Simulation Algorithm [9] that allows for multi-time-delayed reactions [10]. SGNS also allows hierarchical, interlinked compartments to be created, destroyed and divided at runtime, a feature used to generate dynamically independent model cells.

3. RESULTS AND CONCLUSIONS

Here, each model cell 'contains' one promoter and RNAP molecules, which interact via reactions (1) and (2). Parameter values of the 'control condition' are shown in Table 1 and Table 2.

The parameter values associated with RNAP numbers in individual cells (Table 1) are obtained from measurements of RNAP fluorescence intensities in individual *E. coli* RL1314 cells with fluorescently tagged β' subunits reported in [5]. From these, we have set the mean RNAP fluorescence in individual cells arbitrarily to 1 and then obtained the fraction of cells with a given relative fluorescence level. The resulting distribution of relative RNAP fluorescence levels is shown in Figure 1. Note that the 2.5% cells with lowest and highest total fluorescence intensity were discarded, as they were clear outliers.

Next, to obtain the CV^2 of these RNAP relative levels in individual cells, we fitted a normal distribution to the data using the MATLAB package *Statistics and Machine Learning Toolbox*™ [23] (Figure 1). The CV^2 of the fit is shown in Table 1.

Table 1. Parameter values of RNAP numbers in model cells in the control condition of the simulations

Parameter	Value	Reference
Mean RNAp available/cell	1*	[5]
CV^2 of RNAp available/cell	0.03	[5]

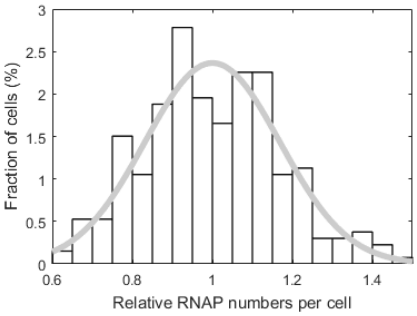


Figure 1. Fraction of cells with a given relative RNAP-fluorescence level as measured by microscopy in *E. coli* RL1314 cells with fluorescently tagged β′ subunits (bars) [1]. The mean absolute RNAP level in individual cells was set to 1. Also shown is the best fitting normal distribution (grey line).

To validate the fitting, we performed a Kolmogorov-Smirnov (KS) test and verified that the two distributions (empirical and best fit) cannot be distinguished in a statistical sense (p-value of 0.69). Thus, we use the best fit distribution to set random RNAP numbers in individual model cells. The mean and CV^2 of RNAP numbers of the model cell population are shown in Table 1.

Table 2. Parameter values of the transcription model (control)

Parameter	Value	Reference
k_{ON}	0.01 s^{-1}	[5]
k_{OFF}	281 s^{-1}	[5]
k_1	6469 s^{-1}	[5]
k_{-1}	1 s^{-1}	[5]
k_2	0.005 s^{-1}	[5]
k_3	∞	[5]

Table 2 shows the values of k_{ON}, k_{OFF}, k_1, k_{-1}, k_2, and k_3 of the transcription model, which were inferred from empirical distributions of time intervals between consecutive RNA productions in individual cells, under the control of the Lac-Ara-1 promoter in DH5α-PRO *E. coli* cells [5]. For this statistical inference, it was assumed the same model of transcription as here.

Next, we ran simulations of model cells (control condition) using SGNS2 [8]. Example time series of RNA production events in 5 individual model cells are shown in Figure 2. Visibly, most cells produced 2 RNAs during 2000 s, as expected [5].

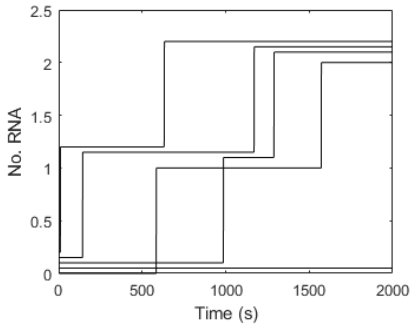


Figure 2. Example time series of the number of RNAs produced by 5 individual model cells in 2000 s. These numbers are offset of each other on the y axis to distinguish between the lines of different cells (only integer RNA numbers are possible).

Next, we study the overall cell to cell diversity in number of produced RNAs as a function of the cell-to-cell variability in RNAP numbers and as a function of the rate constants controlling the kinetics of closed (k_1) and open (k_2) complex formation.

According to the model, e.g., increasing k_1 results in shorter time-length closed complex. Meanwhile, increasing k_2 results in shorter time-length open complex. Here, we change the values of both k_1 and k_2 so that the mean RNA production rate remains unaltered. For that, we use the formula derived in [1], and reproduced here:

$$I(R) = \frac{(k_{ON} + k_{OFF})(k_{-1} + k_2)}{R k_1^* k_2 k_{ON}} + \frac{1}{k_2} + \frac{1}{k_3} \qquad (3)$$

In (3), I(R) is the mean interval between consecutive RNA productions in individual cells. This equation, derived from reactions (1) and (2), assumes infinite cell lifetime. Here, we vary k_1, and then, based on (3), vary k_2 so as to maintain *I(R)* constant.

Given this, we selected 10 values for k_1 and, consequently, k_2 [5]. From these, using (3), we calculate for each case the fraction of time between consecutive RNA production events (Δt) that is spend in closed complex formation ($\tau_{cc}/\Delta t$). The range of these values was set so as to be in agreement with recent measurements made for various promoters subject to various induction levels in cells whose RNAP numbers distribution is similar to that in Table 1. These empirical values are shown in Table 3.

Also, we selected 7 different values of CV^2 in RNAP numbers in individual cells, around the empirical value of 0.03 (Table 1). Based on these sets of parameter values, we produced 70 models of cells, combining in all possible ways the two parameter sets.

For each model, we simulated 1000 model cells for 10000 s each, and extracted the number of produced RNAs per cell. The mean number of RNAs produced per cell in the various models equaled ~10. In Figure 3, we show the values of the CV^2 of the number of produced RNAs in individual cells in all conditions.

Table 3. Empirical values of $\tau_{cc}/\Delta t$ of various promoters subject to different induction levels

Promoter and induction	$\tau_{cc}/\Delta t$	Reference
P_{BAD} with 0.1% arabinose	0.71	[6]
P_{BAD} with 0.01% arabinose	0.55	[6]
P_{BAD} with 0.001% arabinose	0.17	[6]
$P_{lac\text{-}O1O3}$ with 1 mM IPTG	0.55	[6]
$P_{lac\text{-}O1O3}$ with 0.05 mM IPTG	0.46	[6]
$P_{lac\text{-}O1O3}$ with 0.005 mM IPTG	0.12	[6]
P_{tetA} with no inducers	0.07	[6]
$P_{lac\text{-}O1}$ with 1 mM IPTG	0.05	[6]
$P_{lac\text{-}ara1}$ with 1 mM IPTG and 0.1% arabinose (full induction)	0.49	[6]

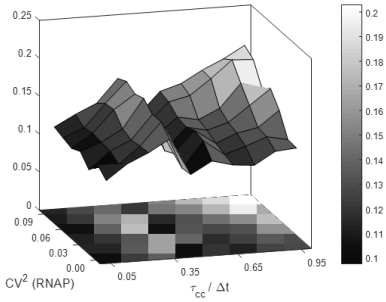


Figure 3. CV^2 of number of produced RNAs in model cells during their lifetime as a function of relative durations of closed and open complex formation and of the cell-to-cell variability in RNAP numbers.

From Figure 3, as $\tau_{cc}/\Delta t$ increases, so does the cell-to-cell variability in RNA numbers. Similarly, the higher the cell-to-cell variability in RNAP numbers, the higher the CV^2 in RNA numbers. Finally, increasing both these two parameters leads to a much higher increase in the cell-to-cell variability in RNA numbers, than if changing only one of these parameters.

The conclusion from these results is that, while as expected the cell-to-cell variability in RNAP numbers 'propagates' to the cell-to-cell diversity in RNA numbers, the degree with which it propagates is heavily promoter kinetics dependent.

Also, there is an unexpected decrease in CV^2 in RNA numbers at $\tau_{cc}/\Delta t$ ~ 0.35, that will require further research to explain.

4. DISCUSSION

We explored the effects of cell-to-cell variability in RNAP numbers in the cell-to-cell variability in RNA production rates, as a function of the kinetics of transcription initiation of a promoter. For this, we simulated the dynamics of RNA production in model cells, making use of a detailed stochastic model that combines

multiple steps in transcription initiation with cell-to-cell variability in RNAP numbers. All parameter values of the model were inferred from single-cell microscopy measurements.

We observed that as the cell-to-cell variability in RNAP numbers increases, so does the variability in RNA numbers. However, genes are not entirely void of 'filters' of this phenomenon. Namely, within the range of realistic parameter values, we observed that different promoter kinetics results in different degrees of variability in RNA numbers in individual cells. Specifically, RNAs whose production is controlled by promoters with relatively slow closed complex formation will exhibit much wider variability in numbers between cells.

As the initiation dynamics of promoters is both sequence-dependent and subject to regulation (e.g. inducers and repressors), we expect the level of cell-to-cell diversity in RNA (and protein) numbers of a gene due to the variability in RNAP numbers to be both evolvable as well as adaptable.

In addition, given the observed degree of changes in variability in RNA numbers as a function of the two parameter values changed in the course of the simulations, we expect this phenomenon to also be observable at the level of small genetic circuits. In the future, it would be of interest to investigate the extent to which this effect influences the behavior of such small circuits, such as genetic switches and clocks.

5. ACKNOWLEDGMENTS

Work supported by Academy of Finland (295027 to A.S.R.), Academy of Finland Key Project Funding (305342 to A.S.R.), Jane and Aatos Erkko Foundation (610536 to A.S.R.), and Tampere University of Technology President's Graduate Program (to S.S). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

6. REFERENCES

[1] Jishage, M., Iwata, A., Ueda, S., and Ishihama, A. 1996. Regulation of RNA polymerase sigma sub-unit synthesis in Escherichia coli: intracellular levels of four species of sigma subunit under various growth conditions. *J. Bacteriol.* 178, 5447–51.

[2] Rahman, M., Hasan, M.R., Oba, T., and Shimizu, K. 2006. Effect of rpoS gene knockout on the metabolism of Escherichia coli during exponential growth phase and early stationary phase based on gene expressions, enzyme activities and intracellular metabolite concentrations. *Biotechnol. Bioeng.* 94, 585–95.

[3] Farewell, A., Kvint, K., and Nyström, T. 1998. Negative regulation by RpoS: A case of sigma factor competition, *Mol. Microbiol.* 29, 1039–51.

[4] Hengge-Aronis, R. 2002. Recent insights into the general stress response regulatory network in Escherichia coli. *J. Mol. Microbiol. Biotechnol.* 4, 341–6.

[5] Lloyd-Price, J., Startceva, S., Kandavalli, V., Chandraseelan, J., Goncalves, N., Oliveira, S.M.D., Häkkinen, A. and Ribeiro, A.S. 2016. Dissecting the stochastic transcription initiation process in live Escherichia coli. *DNA Research* 23(3), 203-214.

[6] Kandavalli, V.K., Tran, H. and Ribeiro, A.S. 2016. Effects of σ factor competition on the *in vivo* kinetics of transcription initiation in *Escherichia coli*. *BBA Gene Regulatory Mechanisms* 1859, 1281–1288.

[7] McClure, W.R. 1980. Rate-limiting steps in RNA chain initiation. *Proc. Natl. Acad. Sci. U. S. A.* 77, 5634–5638.

[8] Lloyd-Price, J., Gupta, A, and Ribeiro, A.S. 2012. SGNS2: A Compartmentalized Stochastic Chemical Kinetics Simulator for Dynamic Cell Populations. *Bioinformatics* 28, 3004-5.

[9] Gillespie, D. T. 1977. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* 81(25), 2340–2361.

[10] Roussel, M.R. and Zhu, R. 2006. Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. *Phys Biol.* 3(4), 274-284.

[11] Taniguchi, Y., Choi, P.J., Li, G.-W., et al. 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329, 533–538.

[12] Bernstein, J.A., Khodursky, A.B., Pei-Hsun, L., 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. USA* 99, 9697–702.

[13] Muthukrishnan, A.-B., Kandhavelu, M., Lloyd-Price, J., et al. 2012, Dynamics of transcription driven by the tetA promoter, one event at a time, in live Escherichia coli cells. *Nucleic Acids Res.* 40, 8472–83.

[14] Kandhavelu, M., Lloyd-Price, J., Gupta, A., Muthukrishnan, A., Yli-Harja, O. and Ribeiro, A.S. 2012, Regulation of mean and noise of the in vivo kinetics of transcription under the control of the lac/ara-1 promoter. *FEBS Lett.* 586, 3870–5.

[15] McClure, W.R. 1980. Rate-limiting steps in RNA chain initiation. *Proc. Natl Acad. Sci. USA* 77, 5634–8.

[16] So, L.-H., Ghosh, A., Zong, C., Sepúlveda, L.A., Segev, R. and Golding, I. 2011. General properties of transcriptional time series in Escherichia coli, *Nat. Genet.* 43, 554–60.

[17] Saecker, R.M., Record, M.T. and Dehaseth, P.L. 2011. Mechanism of bacterial transcription initiation: RNA polymerase - promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. *J. Mol. Biol.* 412, 754–71.

[18] Chamberlin, M.J. 1974. The selectivity of transcription, *Annu. Rev. Biochem.* 43, 721–75.

[19] DeHaseth, P.L., Zupancic, M.L. and Record, M.T. 1998. RNA polymerasepromoter interactions: The comings and goings of RNA polymerase. *J. Bacteriol.* 180, 3019–25.

[20] 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, 3873–81.

[21] Chong, S., Chen, C., Ge, H. and Xie, X.S. 2014. Mechanism of transcriptional bursting in bacteria. *Cell* 158, 314–26.

[22] Ribeiro, A.S., Zhu, R. and Kauffman, S.A. 2006. A General Modeling Strategy for Gene Regulatory Networks with Stochastic Dynamics. *J. of Comput. Biol.* 13, 1630-1639.

[23] MATLAB and Statistics and Machine Learning Toolbox Release 2016b, The MathWorks, Inc., Natick, Massachusetts, United States.

PUBLICATION II

Dissecting the in vivo dynamics of transcription locking due to positive supercoiling buildup.

Cristina S.D. Palma, Vinodh Kandavalli, Mohamed N.M. Bahrudeen, Marco Minoia, Vatsala Chauhan, Suchintak Dash, Andre S. Ribeiro

Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 1863, 194515, 2020.

https://doi.org/10.1016/j.bbagrm.2020.194515

BBA - Gene Regulatory Mechanisms 1863 (2020) 194515

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

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

Dissecting the *in vivo* dynamics of transcription locking due to positive supercoiling buildup

Cristina S.D. Palma, Vinodh Kandavalli, Mohamed N.M. Bahrudeen, Marco Minoia, Vatsala Chauhan, Suchintak Dash, Andre S. Ribeiro*

Laboratory of Biosystem Dynamics, BioMediTech, Faculty of Medicine and Health Technology, Tampere University, 33101 Tampere, Finland

ARTICLE INFO

Keywords:
Single-RNA production dynamics
Positive supercoiling buildup
LineWeaver-Burk plots
Transcription locking kinetics

ABSTRACT

Positive supercoiling buildup (PSB) is a pervasive phenomenon in the transcriptional programs of *Escherichia coli*. After finding a range of Gyrase concentrations where the inverse of the transcription rate of a chromosome-integrated gene changes linearly with the inverse of Gyrase concentration, we apply a LineWeaver-Burk plot to dissect the expected *in vivo* transcription rate in absence of PSB. We validate the estimation by time-lapse microscopy of single-RNA production kinetics of the same gene when single-copy plasmid-borne, shown to be impervious to Gyrase inhibition. Next, we estimate the fraction of time in locked states and number of transcription events prior to locking, which we validate by measurements under Gyrase inhibition. Replacing the gene of interest by one with slower transcription rate decreases the fraction of time in locked states due to PSB. Finally, we combine data from both constructs to infer a range of possible transcription initiation locking kinetics in a chromosomal location, obtainable by tuning the transcription rate. We validate with measurements of transcription activity at different induction levels. This strategy for dissecting transcription initiation locking kinetics due to PSB can contribute to resolve the transcriptional programs of *E. coli* and in the engineering of synthetic genetic circuits.

1. Introduction

Transcription in *Escherichia coli* generates positive supercoiling ahead of the RNAP and negative supercoiling behind it ([11,46,51]; [87,95,99]). Discrete, topologically constrained segments along the DNA cause this process to generate local supercoiling buildup [31,33,41,70,77]. Evidence suggests that this torsional stress can affect gene activity [2,96].

E. coli has (at least) two proteins to resolve torsional stress. Namely, Gyrase removes positive supercoils [9,15,46] while Topoisomerase I removes negative supercoils [9,15,22,35,46,90]. Interestingly, in normal conditions, Topoisomerase I removes the negative supercoils at sufficient speed for R loops to not emerge, which is essential for cell survival [16]. This is made possible by the existence of a direct physical interaction between the RNAP and Topoisomerase I, allowing the latter to remove the negative supercoils, as soon as they form [7]. Contrarily to this, the removal of positive supercoils is not as efficient (being an ATP-dependent reaction likely contributes to this [77]), in the sense that positive supercoiling buildup (PSB) is commonly observed, particularly in highly active operons [17,28]. In support, measurements have shown that Topoisomerase I can relax plasmid DNA ~6 times faster than Topoisomerase IV [97], which has the same catalytic rate as Gyrase [84].

As positive supercoils accumulate, elongation slows down and, eventually, there are transient halts in transcription initiation [9,73]. These halts in initiation tangibly decrease RNA production rates and increase transcriptional noise [9,55,59]. Thus, dissection of the *in vivo* kinetics of transcription locking due to PSB is needed in order to dissect the transcriptional programs of *E. coli*.

A strategy was recently introduced for dissecting the *in vivo* kinetics of rate-limiting steps of active transcription initiation from *in vivo* measurements of individual RNA production events at different RNA polymerase (RNAP) concentrations [48]. It uses a Lineweaver–Burk plot [44] to infer the time-length of events *prior* and *after* commitment to open complex formation [58] from measurements of *in vivo*transcription rates at different RNAP concentrations ([RNAP]) [48]. This is possible due to the independence of the kinetics of the open complex formation from [RNAP], and because there is a range of values of [RNAP] for which the inverse of RNA production rate changes linearly with the inverse of [RNAP] [48].

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

https://doi.org/10.1016/j.bbagrm.2020.194515
Received 14 June 2019; Received in revised form 7 February 2020; Accepted 20 February 2020
Available online 27 February 2020

Similarly, chromosomal RNA production rates (particularly of genes in highly transcribed operons) are expected to differ with Gyrase concentration, due to the existence of discrete topological constraints [9,69,70]. Thus, it should be possible to, from *in vivo* RNA production rates at different Gyrase concentrations, infer the kinetics of *in vivo* transcription locking due to PSB. For this, it must hold true that there is a range of conditions for which the inverse of the RNA production rate changes linearly with the inverse of Gyrase concentration.

Here we verify this hypothesis and then use this strategy to dissect the contribution of transcription initiation locking due to PSB on the kinetics of RNA production of a chromosome-integrated gene. We validate the estimation by time-lapse microscopy of single-RNA production kinetics of the same gene when single-copy plasmid-borne, shown to be impervious to Gyrase inhibition. Based on this, we estimate the fraction of time in locked states and the number of transcription events prior to locking, which we validate by measurements of RNA production under the inhibition of Gyrase activity by the addition of Novobiocin (see Section 2.3). Replacing our gene by a gene with a slower transcription rate, we show that changes in the *basal* transcription rate (expected rate of RNA production in the absence of effects from PSB) affect the contribution of locking due to PSB on *effective* transcription rates (measured rate of RNA production). Finally, we infer a range of possible transcription initiation locking kinetics in a chromosomal location, obtainable by tuning the basal transcription rate, and validate this inference using measurements of transcription activity at different induction levels.

2. Materials and methods

2.1. Strains and plasmids

We engineered two strains from *E. coli* BW25993 (*lacI^q hsdR514 ΔaraBAD$_{AH33}$ ΔrhaBAD$_{LD78}$*) [10]. In one strain, the target gene $P_{LacO3O1}$-mCherry-MS2-BS is integrated into a single-copy F-plasmid (~11 kbp), pBELOBAC11 (target plasmid). This plasmid is not known to form long-lasting bounds to the cell membrane and is originally responsible for the expression of transient DNA-binding proteins [27,32,60]. In the other strain this plasmid is absent and the same target gene, $P_{LacO3O1}$-mCherry-MS2-BS, is integrated into the lac locus of the genome using Red/ET recombination (Gene Bridges, Heidelberg, Germany) (Supplementary Figs. S1A and S1B). We found no significant differences in the growth rates of the two strains and the original strain.

$P_{LacO3O1}$, inducible by IPTG, was engineered from the *E. coli* native P_{Lac} by removing the O2 repressor binding site downstream of the transcription start site [62]. Thus, strong topological barriers are not expected to form when fully induced [21]. Also, both strains were transformed with the medium-copy reporter plasmid pZA25-GFP [61] (kind gift from Orna Amster-Choder, Hebrew University of Jerusalem, Israel), coding for the reporter MS2-GFP under the control of the P_{BAD} promoter. The strain with the target gene in a single-copy F-plasmid also contains the native Lac promoter in the chromosome. Thus, it has 4 LacI binding sites. The strain with the chromosome-integrated target gene only has 2 LacI binding sites, as the original Lac promoter was replaced by the target promoter, $LacO_3O_1$. However, as both strains overexpress LacI [10], effects of this difference are expected to be negligible.

Both target genes (plasmid and chromosome constructs) code for an RNA with an array of bindings sites (BS) for the modified viral coat protein MS2-GFP [25,67,68]. Due to the multiple BS in the target RNA and the strong binding affinity of each site [25], MS2-GFP tagged RNAs appear as bright spots soon after produced (Supplementary Figs. S2B and S2D). Their maximum fluorescence is reached rapidly (< 1 min) and have long half-lives (Supplementary Section I).

For overexpressing Gyrase, we constructed a plasmid (pZe11 P_{rham} gyrAB-*sfGFP*, with ampicillin resistance) with the *gyrA* and *gyrB* genes under the control of a Rhamnose promoter. These genes were arranged in a polycistronic manner, using their (identical) ribosome-binding site to maintain the physiological stoichiometry of the two subunits. We amplified the sfGFP using the primers: Forward: 5′CATATGAGCAAAG GAGAAGAACTTTT 3′, Reverse: 5′ CGGCCGTTTGTAGAGCTCATCCA TGC 3′ with restriction enzymes and cloned it after the *gyrAB* genes by digestion followed by ligation (Supplementary Fig. S3). We also constructed a plasmid without sfGFP, by digesting with the restriction enzymes *Nde*I and *Nae*I, followed by ligation, which was transferred to *E. coli* BW25993 with the $P_{LacO3O1}$-mCherry-MS2-BS integrated into a single-copy F-plasmid [27] and to *E. coli* BW25993 with the $P_{LacO3O1}$-mCherry-MS2-BS integrated in the chromosome. Finally, in another strain, we replaced the chromosome-integrated $P_{LacO3O1}$ by the native Lac promoter, followed by the same array of binding sites for MS2-GFP.

To access the intracellular levels of Gyrase A proteins, we used a strain with a gyrA gene endogenously tagged with the YFP coding sequence [85]. From the glycerol stock (-80 °C), cells were streaked on the LB agar plates and incubated at 37 °C overnight. From the plate, a single colony was picked, inoculated in an LB medium supplemented with the antibiotics, and incubated at 30 °C overnight with shaking at 250 RPM. Next, cells were diluted into fresh LB medium to an OD of 0.03 (Optical Density, 600 nm; Ultrospec 10, Amersham biosciences, UK) and grown at 37 °C with 250 RPM until it reaches to the mid-exponential phase (OD ~0.4–0.5).

2.2. Nucleoid visualization by DAPI staining

DAPI (4′,6-diamidino-2-phenylindole) stains nucleoids specifically with little or no cytoplasmic labelling. Gyrase induced and un-induced cells were grown at 37 °C and fixed with 3.7% formaldehyde in phosphate buffered saline (PBS, pH 7.4) for 30 min at room temperature, followed by washing with PBS to remove excess formaldehyde. The pellets were suspended in PBS, and DAPI (2 μg/ml) was added to the suspension. After incubating for 20 min in the dark, cells were centrifuged and washed twice with PBS to remove excess DAPI. Cells were then re-suspended in PBS and 3 μl of these cells were placed on a 1% agarose gel pad for microscopy.

2.3. Growth conditions and induction of the reporter and target gene

From a -80 °C glycerol stock, cells were placed in LB medium agar plates with 34 μg/ml Chloramphenicol and 35 μg/ml Kanamycin (Sigma-Aldrich, USA) and incubated overnight at 37 °C (Innova® 40 incubator, New Brunswick Scientific, USA). Cells were cultured in LB medium from single colonies on LB agar plates with the appropriate concentration of antibiotics and incubated overnight at optimal temperature at 250 rpm with aeration. These cultures were diluted to an optical density (OD_{600}) of 0.05 in fresh M9 medium, with a culture volume of 20 ml supplemented with the appropriate antibiotics and 0.4% of Glycerol (Sigma-Aldrich, USA), and incubated for 3 h with a 250 rpm agitation until an OD_{600} of ~ 0.3. Next, to induce MS2-GFP expression, 0.4% of L-Arabinose (Sigma-Aldrich, USA) was added and cells were incubated for another 45 min for sufficient MS2-GFP to accumulate for detecting target RNAs [26]. Next, the target gene was induced by IPTG (Sigma-Aldrich, USA) and cells were incubated for 1 h, prior to image acquisition or RT-PCR. To obtain induction curves of target genes (under the control of $P_{LacO3O1}$ and P_{Lac}), 0, 50, 100, 250, 500 and 1000 μM IPTG was added (Supplementary Fig. S4A). Unless stated otherwise, the target genes are always fully induced by 1000 μM IPTG.

We also performed measurements when inactivating and when overexpressing Gyrase. To inactive gyrase, we follow the protocol above, but when MS2-GFP expression is induced, we further added Novobiocin (100 μg/ml) [22]. Since all strains used here contain the gene *acrA*, Novobiocin at this concentration is not expected to affect cell division rate [50]. We verified this by measuring growth rates by OD_{600} for varying Novobiocin concentration (0, 25, 50, 75, 100, 200,

400, 500 ng/μl). The measurements show that growth rates to not differ significantly at 100 μg/ml or lower (Supplementary Fig. S5A). We further verified that Novobiocin does not affect morphology at these concentrations (see Section 3.6).

To overexpress Gyrase, we use Rhamnose (see previous Section) [93]. We follow the protocol above but, when inducing MS2-GFP expression, we also added Rhamnose. When applicable, Gyrase and RNAP concentrations were measured 1 h after adding Rhamnose. Gyrase overexpression did not affect bacteria growth (Supplementary Fig. S5B) nor morphology (see Section 3.2).

2.4. RT-PCR

One hour after inducing the target gene, cells were fixed by RNAProtect bacteria reagent (Qiagen, Germany), followed by enzymatic lysis with Tris-EDTA Lysozyme (15 mg/ml) buffer (pH 8.3). From the lysates, the RNA content was isolated using RNeasy purification kit (Qiagen) as per the manufacturer instructions. The RNA was then separated by electrophoresis using 1% agarose gel stained with SYBR® Safe DNA Gel Stain (Thermo Scientific, USA). The RNA was intact, with clear bands for the 16S and 23S ribosomal RNA. The RNA yield (~2 μg/μl) and absorbance ratios A260/A280 nm and A260/A230 nm were measured by a Nanovue Plus Spectrophotometer (GE Healthcare Life Sciences, USA). The ratio (2.0–2.1) indicates highly purified RNA. To remove DNA contamination, samples were treated with DNaseI (Thermo Scientific, USA) as per the manufacturer instructions. The cDNA was synthesized from 1 μg of RNA using iScript Reverse Transcription Supermix (Biorad, USA) as per the manufacturer instructions. cDNA samples (10 ng/μl) were mixed with qPCR master mix with iQ SYBR Green supermix (Biorad, USA) with primers (200 nM) for target and reference genes. 16S rRNA was used as reference. Primers set for target mRNA (mCherry) and reference (16S rRNA) genes were: mCherry (Forward: 5' CACCTACAAGGCCAAGAAGC 3′, Reverse: 5′ TGGTGTAGTCCTCGTTGTGG 3′), 16S rRNA (Forward: 5' CGTCAGCTC GTGTTGTGAA 3′, Reverse: 5′ GGACCGCTGGCAACAAAG 3′). To determine fold changes in mRNA Gyrase, cells were grown in M9 media supplemented with different Rhamnose concentrations. For the Gyrase mRNA (GyrA) gene the primer set was: Forward: 5' GGATTATGCGAT GTCGGTTCAT 3′, Reverse: 5′ CTAGCACAGTATCTGGCGGCT 3′. For mRNA sfGFP the primer set used was: Forward: 5′ GGAAAACTACCTG TTCCGTGGC 3′, Reverse: 5' ACATAACCTTCGGGCATGGCAC 3′. Experiments were performed by a Biorad MiniOpticon Real Time PCR System (Biorad, USA). The thermal cycling protocol was 40 cycles of 95 °C for 10 s, 52 °C for 30 s, and 72 °C for 30 s, with the fluorescence being read after each cycle. For each condition, we performed 3 biological replicates. qPCR efficiencies of these reactions were > 95%. No-RT and no-template controls were used to crosscheck non-specific signals and contamination. Cq values from the CFX ManagerTM Software were used to calculate fold changes in the target gene (normalized to the reference gene) and standard error, using Livak's $2^{-\Delta\Delta CT}$ method (40). RT-qPCR results are presented in Table S1.

2.5. Flow cytometry

To measure single cell Gyrase-GFP expression levels, cells were grown as described in Section 2.1. Upon reaching mid exponential phase, cells were diluted 1:1000 into 1 ml PBS vortexed for 10 s and 50,000 cells were tested in each run. Data was collected by an ACEA NovoCyte Flow Cytometer (ACEA Biosciences Inc., San Diego USA) using a blue laser (488 nm) for excitation and the fluorescein isothiocyanate detection channel (FITC) (530/30 nm filter) for emission, at a flow rate of 14 μl/min and a core diameter of 7.7 μM. A PMT voltage of 417 was used for FITC. To avoid background signal from particles smaller than bacteria, the detection threshold was set to 5000 in FSC-H analyses. We set the fraction of the cells used in the analysis (α) to 0.55, to remove any undesired data points from debris, cell doublets etc. Reducing α further did not change the results.

2.6. Western blot

Cells were grown as above until reaching an OD_{600} of 0.6. Pelleted cells were lysed with B-PER bacterial protein extraction reagent (Thermo scientific) and proteins were extracted. Protein samples were diluted with 4× laemmli sample loading buffer and boiled for 5 mins at 95 °C. 30 μg of proteins were loaded on the 4–20% TGX stain free pre cast gel (Biorad) and separated by electrophoresis. Proteins were then transferred to PVDF membrane using Trans-Blot Turbo transfer system (Biorad). The membrane was blocked with 5% non-fat milk and incubated with primary RpoC antibody 1:2000 dilutions (Biolegend) overnight at 4 °C and followed by HRP-secondary antibodies 1:5000 dilutions (Sigma Aldrich) for 1 h at room temperature. For band detection, the membrane was treated with a chemiluminescence reagent (Biorad). Images were acquired by the Chemidoc XRS system (Biorad). Band quantification was done using Image lab software (version 5.2.1). For each condition, we performed 3 biological replicates.

2.7. Microscopy and image analysis

Cells were grown as above, and pelleted and re-suspended in ~100 μl of the remaining media. Prior to imaging, cells were placed on a 2% agarose gel pad of M9 medium and kept in between the microscope slide and a coverslip. Cells were visualized by a Nikon Eclipse (Ti-E) inverted microscope with a 100× Apo TIRF (1.49 NA, oil) objective. Confocal images were taken by a C2+ (Nikon) confocal laser-scanning system with a pinhole size of 1.2 AU. In confocal images, the size of a pixel corresponds to 0.062 μm using a scan area resolution of 2048 × 2048 pixels. MS2-GFP-RNA spots and GyrA-YFP regions were visualized by a 488 nm laser and a 514/30 emission filter, while DAPI-stained nucleoids were visualized by a 405 nm laser and a 447/60 emission filter.

Phase contrast images were taken by an external phase contrast system and DS-Fi2 CCD camera (Nikon). Image sizes were 2560 × 1920 pixels, each pixel corresponding to 0.048 μm. Phase contrast and confocal images were taken simultaneously by Nis-Elements software (Nikon).

From phase contrast images, we segmented cells with the software iCellFusion [78] (Supplementary Figs. S2A and S2C). Errors were manually corrected. Next, phase-contrast and corresponding fluorescence images were aligned by the software CellAging [29]. We used CellAging to detect RNA-MS2-GFP fluorescence spots (Supplementary Figs. S2B and S2D) and assess the intensity of each spot. From these, integer-valued RNA numbers were calculated for each spot (Supplementary Section I).

Nucleoid(s) segmentation was performed as in [63], using a 2D Gaussian approximation, followed by manual corrections. Cells whose size is smaller than 500 pixels were excluded from the analysis since, in general, they were not real cells (e.g. only half of the cell appeared in the image). Also removed were cells larger than 1000 pixels, as they were abnormally elongated. In general, this led to removing < 5% of the cell population.

The segmentation of the intracellular regions with significant GyrA-YFP was done using a tailored software, SCIP [56]. Errors were manually corrected. To remove measurement noise, we applied a 2D Gaussian filter to each region [94].

2.8. Models and simulations

We use stochastic models of gene expression to test if arrests during elongation, caused by PSB, disturb significantly the mean RNA production rate (within realistic intervals of parameter values).

These models are at single-cell, and single-molecule level. Specifically, two models are simulated. One is the 'Single-Nucleotide

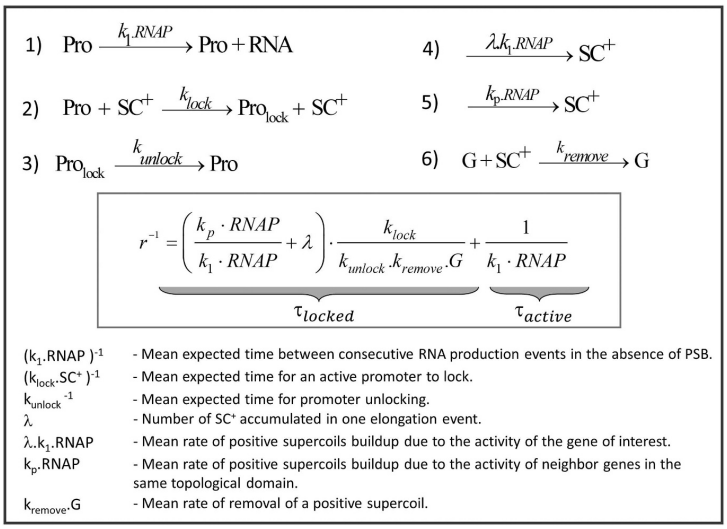


Fig. 1. Minimal (reduced) model of the dynamics of RNA production and transcription locking due to PSB of a chromosome-integrated gene in *E. coli*. The model includes the promoter when active (*Pro*) and when locked due to PSB (Pro_{Lock}), RNA molecules, RNA polymerases (*RNAP*), Gyrases (*G*), and positive supercoils (SC^+). Reaction 1 represents transcription at the rate k_1, which is the basal rate of RNA production of an active promoter assuming one RNAP in the cell. Reaction 2 models promoter locking due to PSB, with k_{lock} being the rate at which an active promoter is locked given the presence of one SC^+. Positive supercoils emerge during transcription of the gene of interest (reaction 4) and/or from the activity of genes in the same topological domain (reaction 5). λ corresponds to a tenth of the number of nucleotides of the gene of interest. Reaction 3 accounts for the unlocking of the promoter at the rate k_{unlock}. Finally, reaction (6) models the removal of positive supercoils by Gyrases. All parameter values are extracted or derived from empirical data , including measurements of a chromosome-integrated $LacO_3O_1$ promoter (k_1 = 0.0014 s^{-1}, k_{lock} = 0.0012 s^{-1} and k_{unlock} = 7 × 10^{-4} s^{-1}) (Supplementary Sections IV and V).

model' (SN Model), since elongation is modelled at the single nucleotide level (Supplementary section II). The other is the "Minimal model', as it is designed from the former, but lacks elongation at the nucleotide level (Fig. 1, Section 3.1).

The time length of each simulation is 10^5 s, found to be long enough for not underestimating the mean length of the time intervals between consecutive RNA production events (which would bias the data with right-censoring) [30]. The simulations have a reading time of 1 s^{-1}. The results shown in the Results section are obtained from 100 runs per condition, as this number suffices to obtain consistent results. Finally, the initials components at the start of simulations are 1 promoter (where the transcription start site is located), 1 Gyrase, and 28 RNAPs. In addition, the SN model has also 4058 nucleotides, in the state "unoccupied" (Supplementary Section II) along which elongation will occur.

The models are implemented in the simulator SGNSim [74] and their dynamics follow the Stochastic Simulation Algorithm [23,24]. In short, the stochastic nature of their dynamics arises from the generation of two random numbers at each step. As described in [24], one of these random numbers determines what is the next reaction (which differs with the propensity of each reaction at that moment), while the other random number determine when the next event will occur (which depends on the total propensity when considering all possible reactions combined). SGNSim makes use of 'Mersenne Twister' to produce these random variables at each step [57].

3. Results

3.1. Expected effects of changing gyrase concentration on the dynamics of transcription

We started by designing a stochastic model of transcription at the single nucleotide level (here named 'SN model'), described in detail in Supplementary Section II and shown in Supplementary Table S2. The model is based on a model proposed in [75] and later used in [53,72], to which we add positive supercoiling buildup/removal. The reactions composing the model should not be interpreted as elementary transitions. Instead, they represent the rates of the rate-limiting steps of the various events. The model dynamics and simulations are described in Materials and Methods, Section 2.8, while its assumption of homogeneous mixing of RNAP and Gyrases is validated in Supplementary Section III.3.

The model consists of the following events (Table S2): transcription starts when an RNAP finds the promoter (reaction S2.1) and unwinds the DNA for reading and escapes the promoter (reaction S2.5). After this, stepwise transcription elongation is initiated, accounting for realistic RNAP footprint in the DNA template, transcriptional pausing, arrests, editing, premature terminations, pyrophosphorolysis and collisions between RNA polymerases [53].

In addition, the model accounts for the phenomenon of production of positive supercoils during elongation [46,51,95], in reaction S2.6 in Supplementary Table S2. As these supercoils accumulate, they enhance the propensity for RNAP arrest (reaction S2.11) [20,51] and transcription initiation locking (reaction S2.2) [9]. The removal of positive supercoils by Gyrase [84] is also modelled explicitly (reaction S2.17). Finally, the model accounts for the potential accumulation of positive supercoils due to transcriptional activity of neighbor genes (reaction S2.4).

This model does not include RNA degradation, as we measured RNA numbers by MS2-GFP tagging, which prevents degradation for a few hours (Supplementary Section I and Figs. S1 and S2) [26,86], thus avoiding this source of noise. Further, for the purposes of this work, we are only interested in RNA production rates, which do not depend on degradation.

Results in Supplementary Sections III.1 and III.2 show that this SN model mimics the effects of PSB on the kinetics of stepwise transcription elongation. Namely, from Supplementary Fig. S6A, one finds that the mean elongation time increases as Gyrase numbers are decreased. Meanwhile, from Supplementary Fig. S6B, one finds that this slowdown of stepwise elongation does not affect the mean rate of RNA production, within the range of parameter values tested, which is expected.

Note that in this model, for the realistic range of parameter values considered, once the system reaches steady state (near constant number of RNAPs on the DNA strand), the mean RNA production rate depends only on the rate with which RNAPs initiate new elongation events (reaction S2.1 in Table S2) and on the rate of abortions of elongation (reaction S2.14 in Table S2), with the latter being near negligible (~4% per transcription initiation event). Only in the unlikely scenario of excessive accumulation of RNAPs in the DNA template that would jam the promoter region, would events in elongation affect the mean RNA production rate.

Therefore, for purposes of estimating the effects of changing Gyrase

concentration on the mean RNA production rate, we instead use a minimal model, where elongation is not explicitly represented. Fig. 1 shows the minimal model that, as the SN model, also has a stochastic dynamics in accordance to the SSA (Materials and Methods, Section 2.8). In detail, reaction 1 models RNA production by an active promoter, *Pro*, and its propensity differs with the basal transcription rate, k_1, and with RNA polymerase numbers. Meanwhile, reaction 2, which models transcription locking, is in all identical to reaction S2.2 of the SN model and, thus, its propensity differs with the number of positive supercoils (Supplementary Section II).

Positive supercoils can be generated via reactions 4 and 5 (also as in the SN model) [1,9,51,70,83,84,89]. The propensity of reaction 4 depends on the basal transcription rate, k_1, of the gene of interest, as described in [17] and in agreement with results from anchored plasmids [9,87] as well as with results reported here. The parameter λ in reaction 4 accounts for the length of the gene of interest (RNAP will take longer to transcribe a longer gene, during which time positive supercoils are produced). As for reaction 5, responsible for the accumulation of PSB due to the transcription activity in the topological domain of the gene of interest, its kinetics differs with the neighboring activity, which can be tuned by k_p and RNAP numbers (Fig. 1).

Once locked, a promoter can become unlocked via reaction 3. The unlocking kinetics can be tuned by the rate constant k_{unlock}. Because the propensity for locking changes linearly with the number of positive supercoils, the propensity for reaction 3 is kept independent from this number. Else, the overall time spent in locked states would change quadratically with the inverse of Gyrase numbers, and not linearly (Fig. 3 provides empirical support for the assumption that this relationship is linear within realistic ranges of parameter values). Finally, reaction 6 represents the removal of positive supercoils by Gyrases. As this takes place, the propensity for reaction 2 decreases, thereby accounting for the expected decrease in the effects of PSB with increasing Gyrase numbers [9].

To verify that the minimal model constitutes a valid approximation of the SN model, we performed simulations for various Gyrase numbers. Visibly, from Supplementary Fig. S6E, the minimal model matches the mean rate of RNA production of the SN model (and the empirical data) as a function of Gyrase numbers. This is expected since, as noted, all its parameter values are the same as in the SN model, except for reaction 4 in Fig. 1, since this reaction needs to account for the number of nucleotides of the gene of interest (which are modelled explicitly in the SN model). This adjustment is done by having the rate of SC^+ production of the gene of interest dependent on its nucleotide length (with λ equaling a tenth of its number of nucleotides, as this is approximately the expected number of SC^+ produced during one elongation event [84]).

We then derived an analytical solution of the minimal model, for the inverse of the mean rate of RNA production (r^{-1}) as a function of Gyrase (inset of Fig. 1). Here, τ_{active} is the mean time between consecutive RNA production events of an unlocked/active promoter, which equals the inverse of $k_1 \times$ RNAP (with RNAP being the number of RNA polymerases). Meanwhile, r is the inverse of the sum of τ_{active} and τ_{locked}, with the latter being the mean time spent in locked states (equation in the larger inset in Fig. 1). From this solution, we find that increasing [G] decreases τ_{locked} [9,46], which increases r. In detail, r^{-1} is expected to change linearly with $[G]^{-1}$ (large inset, Fig. 1). If this holds true, from measurements of r and [G], it should be possible to extrapolate τ_{active}, since τ_{active} should equal r^{-1} for infinite [G]. Further, from τ_{active} and r, it should be possible to estimate τ_{locked}. Finally, note that while k_1 does not affect the mean time for Gyrase to release the gene from a locked state, it does affect the rate of occurrence of locked states.

Interestingly, many plasmids only have weak, transient topological barriers (such as short-term protein-DNA complexes [42]). In particular, aside from when they are anchored to the membrane [3,11,49,71] or have many tandem copies of a DNA-binding site [42], no long-term PSB is expected, since positive and negative supercoils diffuse in opposite directions and annihilate one another [42] (unlike in the chromosome that has topological barriers). As such, it should be possible to simulate the dynamics of plasmid-borne genes using the model in Fig. 1, by setting k_{lock} to null, causing τ_{locked} to be null. Consequently, r^{-1} of a model plasmid-borne gene should equal τ_{active} of the same model gene, when chromosome-integrated. Further, if this holds true, then a plasmid-borne gene can be used as a proxy for the same gene when chromosome-integrated when unaffected by PSB.

3.2. Changing intracellular concentration of gyrases

Above, we hypothesized that r^{-1} should be linear with respect to $[G]^{-1}$ within a given range of Gyrase concentrations (see Fig. 1 and Supplementary Section VI). If true, one should observe a line on a Lineweaver–Burk plot [44] of r^{-1} against $[G]^{-1}$, from which one can extrapolate τ_{active}. From τ_{active} and r, one can then estimate τ_{locked}.

To test this hypothesis, it is necessary to measure r in cells differing in [G]. For this, we inserted a plasmid carrying a copy of the *gyraseA* and *gyraseB* genes under the control of the Rhamnose promoter (pZe11 P_{rham} gyrAB, Materials and Methods). We further added sfGFP, also under the control of the Rhamnose promoter (pZe11 P_{rham} gyrAB-*sfGFP*, Section 2.1 and Supplementary Fig. S4). The region coding for sfGFP allows measuring mRNA coding for Gyrase and the corresponding protein levels produced solely by the plasmid.

We subjected cells to different Rhamnose concentrations until finding a range for which the production rate of the mRNA coding for Gyrase increases linearly with Rhamnose concentration. For this, we performed qPCR using the region of the RNA from the plasmid that is absent in the native RNA coding for Gyrase (i.e. the region coding for sfGFP). In Fig. 2A we find a linear relationship between mRNA fold changes (measured by qPCR) and Rhamnose (0, 0.1, 0.2 and 0.4%). In particular, small deviations from linearity were rejected (p-value > 0.5, see Fig. 2 legend for details).

Next, we verified that cell growth rates were not disturbed in this

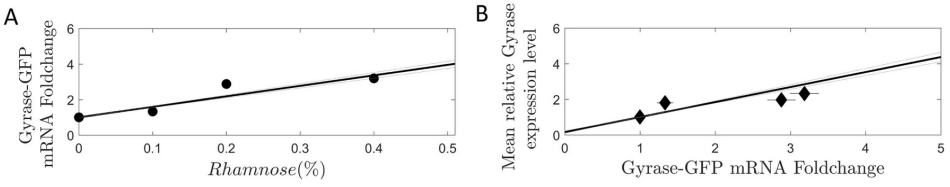


Fig. 2. Gyrase overexpression levels. (A) Fold change of mRNA Gyrase-GFP measured by qPCR, for different concentrations of Rhamnose (0, 0.1, 0.2 and 0.4%), standard error of the mean (vertical error bars) are not visible. Results are relative to the control condition (0% Rhamnose). Also shown is the best-fit line. (B) Calibration line between relative Gyrase-GFP mean expression levels (measured by flow cytometer with the FITC-H detection channel) and mRNA Gyrase-GFP fold change measured by qPCR for 0, 0.1, 0.2 and 0.4% Rhamnose. Gyrase numbers are relative to the 0% Rhamnose condition. Horizontal error bars represent the standard error of the mean. For both figures, we fitted a first order polynomial to the data points by WTLS by minimizing χ^2 [39]. To determine if small deviations from linearity are statistically significant, we performed a likelihood ratio test between the best linear fit and fits by higher order polynomials. In both cases, the test did not reject the linear model (p-values larger than 0.5 and 0.15, for Fig. 2A and B, respectively).

range (Supplementary Fig. S5B). Further, to test if morphology was affected, we measured cell areas in the control condition (165 cells analyzed) and when subject to 0.4% Rhamnose (182 cells analyzed). The cell area was obtained from phase contrast images, using the software iCellFusion (Section 2.7). We performed a 2-sample Kolmogorov-Smirnov test and found that, at the significance level of 0.05, the two distributions cannot be distinguished (p-value of 0.5).

Finally, we verified that, within this range of conditions, the mean relative Gyrase expression level changed linearly with the RNA production rate of the plasmid coding for Gyrase (Fig. 2B), as measured by Flow-cytometry (Section 2.5). In particular, small deviations from linearity were rejected (p-value of 0.15, see Fig. 2 legend for details). We thus conclude that the fold change in Gyrase-GFP protein levels corresponds to the fold change in the mRNA coding for Gyrase-GFP (Supplementary Section VII and Table S3 show the parameters of the calibration line and procedure).

We expect the quantitative relationship between mRNA and protein numbers of the plasmid-borne Gyrase to be the same as in the native Gyrase mRNA and proteins, since we used the native ribosome binding site in the plasmid construct. Thus, we measured by qPCR the fold change of the mRNA produced by both the native and the plasmid-borne Gyrase genes and used the line in Fig. 2B as a calibration line, to estimate the fold change with Rhamnose in Gyrase protein levels (Supplementary Section VII and Table S3).

Finally, we considered that Gyrase overexpression could change the proteome and, eventually, change cellular functioning (e.g. in 1–2 h). To mitigate effects from this eventuality (to avoid unknown changes in the processes represented in Fig. 1), subsequent measurements were conducted 1 h after inducing Gyrase overexpression (Materials and Methods). Given this and the above, we expect that, for 0.2% or lower Rhamnose concentrations (Fig. 2A), changes in RNA production rate in this time window are largely due to changes in concentrations of the components of the reactions in Fig. 1.

3.3. Transcription rate of a chromosome-integrated gene under the control of $P_{LacO3O1}$ in the absence of positive supercoiling buildup

Data in [85] indicates that the expression rate of (at least) three of the RNAP sub-units are, in normal conditions, approximately double the average expression rate of *E. coli* genes. Since several highly expressed genes are supercoiling sensitive [17], it is tangible that Gyrase overexpression may affect [RNAP], which according to the model (Fig. 1), could affect the transcription rate (r) of our gene of interest (Fig. 1). Thus, we first assessed for potential fast changes in [RNAP] when overexpressing Gyrase.

For this, we used the same plasmid as above, with the *gyrA* and *gyrB* genes controlled by a Rhamnose promoter, to overexpress Gyrase (but having removed sfGFP, so as to not affect RNA counting or Gyrase functioning, see Materials and Methods). Next, we measured [RNAP] at the different Rhamnose concentrations (0%, 0.1% and 0.2%) by measuring the RpoC protein by Western blot, 1 h after inducing Gyrase overexpression. From Fig. 3A and B, at the same OD_{600}, the [RNAP] differs by 12% between the two extreme conditions. This difference was found to be statistically significant by a 1-sample 2-tailed t-test, with the null hypothesis that the increase is 12% (p-value of 0.42). In addition, we performed a 2-sample, 2-tailed t-test with the null hypothesis that there is no difference between the conditions, which was rejected (p-value of 0.0008). This is expected to partially explain changes in r due to Gyrase overexpression and, thus, needs to be accounted for when quantifying the direct effects of changing [G] (Supplementary Section VIII).

In addition, it is tangible that overexpression of Gyrase could affect the negative supercoiling state of the chromosome, e.g. by introducing negative supercoils [6,47,80]. This, in turn, could affect DNA supercoiling density and its folding and compaction [36,92], which could alter transcription rates by affecting the time-lengths of open complex formations [52].

Unfortunately, we cannot measure directly the *in vivo* kinetics of open complex formation at a given Gyrase concentration, as this would require measuring the *in vivo* transcription in cells with differing RNAP concentrations [48,81], which would also affect the intracellular Gyrase concentration. Therefore, instead, we estimated indirectly if Gyrase overexpression (between 0% and 0.2% Rhamnose) suffices to alter significantly the chromosome folding and compaction. For this, we assessed if the nucleoid area (with area being a proxy for compaction strength) is altered by Gyrase overexpression, using DAPI staining and image analysis (Sections 2.2 and 2.7).

The mean and standard deviation of the nucleoid area, when and when not overexpressing Gyrase, are shown in Table S4. We performed a 2-sample student t-test for the null hypothesis that the two data sets of absolute nucleoid area come from the same distribution. The test did not reject the null-hypothesis (p-value > 0.01). We thus conclude that, in the range of Gyrase overexpression levels used here, the nucleoid size was not significantly affected. As such, we do not expect the indirect effects of Gyrase overexpression on DNA supercoiling density to significantly affect the kinetics of open complex formation.

Given this, we again used the plasmid with the *gyrA* and *gyrB* genes controlled by a Rhamnose promoter (without sfGFP) to study the effects of Gyrase overexpression on transcription initiation locking due to PSB of a chromosome-integrated gene under the control of $P_{LacO3O1}$. This promoter was used as its dynamics has been previously studied when plasmid-borne, including using single-RNA MS2-GFP tagging [34,54,64,66,81].

We first measured the absolute mean r^{-1} in the control condition (Materials and Methods, Section 2.3) by microscopy measurements of integer valued RNA numbers in single cells at different time moments (Supplementary Section IX). The absolute mean r^{-1} in the control condition was found to equal 1476 s, with a standard error of 145 s.

Using the value of r^{-1} in the control condition, we scaled the relative qPCR values to obtain the values of r^{-1} in conditions where Gyrase is overexpressed. Results from qPCR are shown in Supplementary Table S1. In these, the gradually increasing expression of Gyrase did not affect significantly the expression of the 16S rRNA gene. This is expected, since 16S rRNA is a stable component of ribosomes and, thus, should not change significantly between conditions when growth rates are not affected significantly [88] (Supplementary Fig. S5B). Namely, even if the small changes between conditions were considered significant, there is no monotonic change with increasing Rhamnose concentration. As such, 16S rRNA is used as the reference gene.

Next, from Supplementary Table S1 and the microscopy data in the control condition, we obtained absolute rates of RNA production in each condition (black circles in Fig. 3C). Finally, we fitted a line by weighted total least squares (WTLS) [39] (black line in Fig. 3C) to estimate τ_{active} (where the line intersects the Y-axis), when not accounting for changes in [RNAP]. We performed a likelihood ratio test between the best linear fit and fits of higher order polynomials which showed that the linear model best fits the data (p-value > 0.9).

Next, from the [RNAP] in each condition (Fig. 3B) and the model fitting (Supplementary Section IV), we estimated the effects of changes in [RNAP] (Supplementary Section VIII and Supplementary Table S5). Supplementary Figs. S7A and S7B show the Z surfaces of the best fitting models and empirical results. From the R^2 values (legend of Supplementary Fig. S7), one finds that the model well-fits the empirical data. We thus used this model to estimate the weight of the changes in [RNAP] (Fig. 3B) on r^{-1} and then quantified the changes in r^{-1} due to changes in [G] alone. Results are shown in the blue circles in Fig. 3C. Next, we fitted a line (blue line in Fig. 3C) to these data points using WTLS, from which we estimated r^{-1} for infinite [G] (i.e. τ_{active}) to be 749 ± 247 s.

Finally, we determined if the small deviations from linearity are statistically significant by performing a likelihood ratio test between

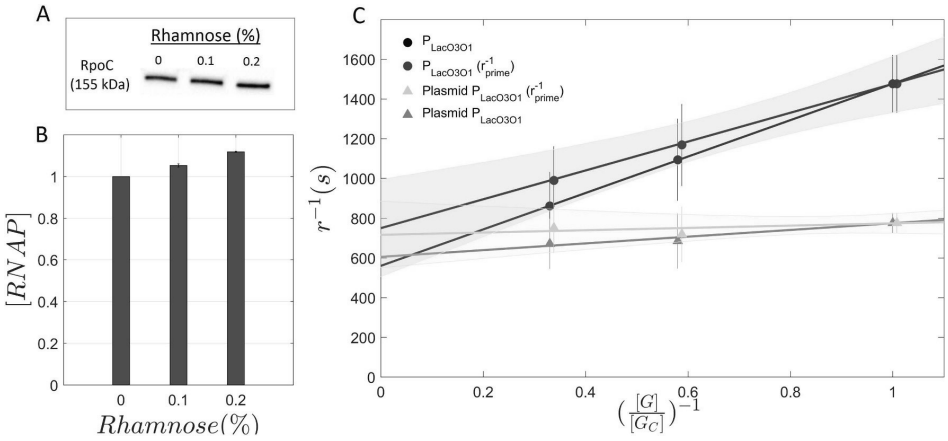


Fig. 3. Effects of Gyrase overexpression in the RNA production rate of $LacO_3O_1$ when chromosome-integrated and when plasmid-borne. (A) Replicate of Western Blot image of RpoC subunit for cells subjected to 0, 0.1% and 0.2% Rhamnose. (B) Bar chart of [RNAP] fold change with Gyrase overexpression, relative to the control condition (0% Rhamnose). In all conditions, OD_{600} was 0.6. (C) LineWeaver-Burk plot of the inverse of the RNA production rate (r^{-1}), for different Gyrase concentrations (black circles), relative to the control ($[G]/[G_C] = 1$) of the chromosome-integrated construct. Also shown is the standard error of the mean (vertical error bars), along with the best-fit line (black line). Further shown are the RNA production rates after correcting for the weight of the changes in [RNAP] (r^{-1}_{prime}), when overexpressing Gyrase (blue circles) and the correspondent best-fit line (blue solid line) and its standard error of the mean (light blue area) obtained by Monte Carlo simulations (5000 iterations). Blue circles are 0.008 units deviated to the right, for figure legibility. The equations of the black and blue lines are $r^{-1} = (917 \pm 329) \times \left(\frac{[G]}{[G_C]}\right)^{-1} + (559 \pm 246)$ and $r^{-1} = (726 \pm 329) \times \left(\frac{[G]}{[G_C]}\right)^{-1} + (749 \pm 247)$, respectively. Finally, the dark grey triangles are the values of r^{-1} for the plasmid-borne construct, when subject to the same levels of Gyrase overexpression while the light grey triangles correspond to r^{-1} after correcting for the weight of the changes in [RNAP] on r^{-1} (dark grey triangles). Light grey triangles are 0.008 units deviated to the right, for figure legibility. Also shown are the respective best-fit lines and its standard errors of the mean (light grey area) obtained by Monte Carlo simulations (5000 iterations). The equation of the dark grey line is $r^{-1} = (168 \pm 184) \times \left(\frac{[G]}{[G_C]}\right)^{-1} + (605 \pm 167)$. The equation of the light grey line is $r^{-1} = (58 \pm 184) \times \left(\frac{[G]}{[G_C]}\right)^{-1} + (715 \pm 167)$. Data from 368 cells (chromosome-integrated gene) and 476 cells (plasmid-borne gene). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the best linear fit and fits by higher order polynomials (by WTLS by minimizing χ^2) [39]. The test did not reject the linear model (p-value > 0.9), from which we conclude that r^{-1} decreases linearly with $[G]^{-1}$.

Several phenomena could have forced this plot to be non-linear. E.g., if the ratio between free and total Gyrase concentrations would increase as Gyrase is overexpressed, the plot would exhibit negative curvature (see Section VI in Supplementary). Meanwhile, if the resolution of supercoils in the control condition was near-saturation, overexpressing Gyrase would result in positive curvature. We therefore interpret the observed linearity as evidence that these changes in r^{-1} are largely due to changes in [RNAP] and [G] as assumed by the model in Fig. 1, rather than due to unknown factors.

3.4. Transcription kinetics of $P_{LacO3O1}$ when single-copy plasmid-borne

To validate the estimation of τ_{active}, we integrated the same gene under the control of $P_{LacO3O1}$ into a single-copy plasmid (Materials and Methods). We expect this to reduce the effects of PSB on the activity of the gene of interest to a minimum. I.e., the value of r^{-1} of the single copy plasmid-borne gene should approximate the estimated τ_{active} of the chromosome-integrated gene. If this holds true, adding Novobiocin, which inhibits Gyrase activity [9,15,22,91], should not disturb significantly its activity.

To test this, we performed time-lapse microscopy measurements of RNA numbers in cells subject to 100 μg/ml Novobiocin (Materials and Methods). Images were taken every 15 min, starting 30 min after introducing 1 mM IPTG in the media to ensure full induction of the target gene [86]. We also performed measurements where Novobiocin was not added.

From Supplementary Fig. S4B, the RNA production rate of the plasmid-borne gene is not affected by the addition of Novobiocin, as expected if PSB is absent. Meanwhile, in the absence of Novobiocin, we observe the same behavior but higher r, which is consistent with the cells subject to Novobiocin having lesser number of active RNAP and/or σ factors [8,13,18,76], etc. Further, both behaviors are significantly different from cells with the chromosome-integrated construct subject to Novobiocin, where a clear blocking of the RNA production is observed shortly after adding Novobiocin (Fig. 4, blue line). We conclude that the gene in the single-copy plasmid is not directly affected by Novobiocin, suggesting that it is impervious to the effects from PSB.

In support, according to the model (Fig. 1), for equal mean RNA production rate, the kinetics of RNA production from a gene unaffected by PSB (such as when on a single-copy plasmid) should be less noisy than otherwise (e.g. when chromosome-integrated) [64]. Lesser noise should reduce cell-to-cell variability in RNA numbers. To test this, we compared the squared coefficient of variation of RNA numbers in single cells, CV^2(RNA), in conditions where the two constructs exhibit the same mean RNA numbers per cell (50 μM IPTG for the plasmid-borne gene and 1000 μM IPTG for the chromosome-integrated gene, Supplementary Fig. S4A). The CV^2(RNA) in cells with the chromosome-integrated construct is found to be much higher than in cells with the single-copy plasmid-borne gene (3.18 and 1.58, respectively), in agreement with the model prediction, even though the plasmid-borne gene is being partially affected by LacI repression, which adds variability in RNA numbers [48].

Finally, we verified that the RNA production rate of the single-copy plasmid construct equals the inverse of τ_{active} of the chromosome construct. For this, we performed microscopy measurements of the integer valued RNA numbers in cells with the plasmid construct and estimated r^{-1} to be 775 ± 50 s (dark grey triangle in Fig. 3C, for the control condition). This result cannot be distinguished, in a statistical sense,

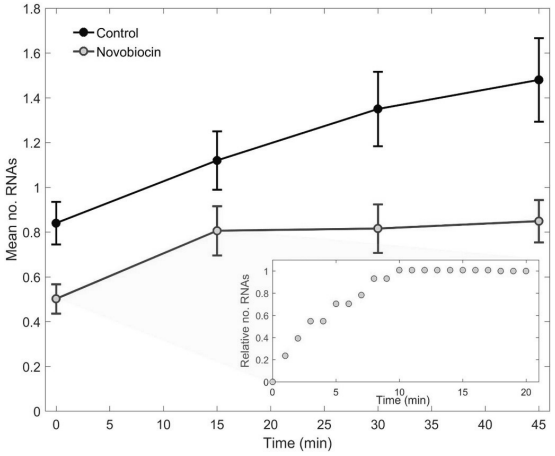


Fig. 4. RNA production over time. Mean integer-valued RNA numbers in individual cells with the chromosome-integrated $P_{LacO3O1}$ when subjected to 100 μg/ml Novobiocin (blue line) and in the control condition (black line, absent of Novobiocin). Measurements performed by microscopy, with single RNA tagging by MS2-GFP. For each time point, new cells were taken from the original culture. On average, 200 cells were used per condition. Error bars represent the standard error of the mean. Finally, the inset shows the number of RNA production events per cell relative to the total number of RNAs produced during the measurement time. Data collected at the single RNA level, from time-lapse microscopy measurements with images taken once per minute. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

from the estimate of τ_{active} for the chromosome construct assuming infinite [G] (749 ± 247 s) (in agreement with the model predictions, Fig. 1).

In this regard, in Supplementary Section X, we estimated the minimum PSB effects that would be detectable, provided the same degree of sensitivity in the measurements of r^{-1}. We found that there needs to exist a fold change between two conditions of, at least, 1.6. However, we estimate that tripling the number of data points collected allows a reduction of this number to 1.2.

Further, we also performed qPCR measurements of the plasmid construct when subject to the same levels of Gyrase overexpression. Results are shown in Fig. 3C. Next, we fitted a line (dark grey line in Fig. 3C) to the data points. From this, we find that the change in RNA production rate of the plasmid gene with gyrase overexpression is ~5 times weaker than in the chromosome-integrated construct. Next, using WTLS [39] we tested if the small deviations from linearity are statistically significant. The test did not reject the linear model (p-value > 0.8). Subsequently, as before, we corrected the data points to account for the changes in RNAP concentrations. Results in Fig. 3 (light grey line) show that the corrected line is nearly horizontal and, as expected, cannot be distinguished from a horizontal line, in a statistically sense, using the same test as above. We conclude that the RNA production kinetics from the plasmid construct is nearly non-responsive to Gyrase overexpression.

In this regard, note that estimation of τ_{active} when accounting for changes in [RNAP] (blue line in Fig. 3C) fits the measurements better (light grey line in Fig. 3C), then when not accounting for [RNAP] changes (black line in Fig. 3C). From comparing the blue and black circles in each condition, we also find that, e.g., for maximum Gyrase (0.2% Rhamnose), the increase in [RNAP] accounts for 31% of the decrease in r^{-1} relative to the control, with the remaining 69% being due to increased [G] (and/or other, unknown factors). Similarly, in the extrapolated condition of infinite [G], the increase in [RNAP], compared to the control condition, accounts for 41% of the decrease in r^{-1}, with the remaining 59% being due to increased [G].

3.5. Mean time spent in locked states and average number of transcription events between consecutive locking events

Since Fig. 3 shows that r^{-1} changes linearly with $[G]^{-1}$, we used the Lineweaver-Burk equation [44] to estimate the mean time spent in locked states, τ_{locked}, as follows:

$$\tau_{locked} = \frac{[G]_2(r_2^{-1} - r_1^{-1})}{([G]_1 - [G]_2)} \quad (1)$$

From (1), given the control and the condition where relative $[G]^{-1}$ is 0.33 (0% and 0.2% Rhamnose, respectively), we infer τ_{locked} to be 735 s, with a standard error of the mean (SEM) of 341 s (obtained by the Delta Method [5]). Using the other pair of conditions (0% and 0.1% Rhamnose) we obtain the same result, in a statistical sense. As the mean time interval between transcription events is 1476 s, we estimate transcription initiation locking due to PSB to account for ~50% of this interval.

Meanwhile, to estimate the mean number of transcription events between consecutive locked states, N, consider that, according to the model:

$$N = \frac{\tau_{escape}}{\tau_{locked}} \quad (2)$$

To solve for N, we used the value of τ_{locked} obtained above, and τ_{escape} obtained from measurements in [9], which reported that the average DNA binding time of Gyrase is ~333 s while the unbind time is $\sim 10^3$ s [9]. Since Gyrase is expected to resolve multiple positive supercoils during this time [1,84], we assumed that the sum of these times (~1333 s) is an upper bound of the time for a locked gene to escape PSB (i.e. τ_{escape}). Introducing the estimated values of τ_{locked} and τ_{escape} in eq. 2, we find that N equals ~1.8 ± 0.84.

3.6. Kinetics of transcription initiation locking in the presence of a gyrase inhibitor

To validate the above estimations, we performed time series measurements at the single-RNA level in cells carrying the chromosome-integrated $P_{LacO3O1}$ subject to Novobiocin, a Gyrase inhibitor [22]. Assuming that, when Novobiocin first enters the cytoplasm, $P_{LacO3O1}$ activity is not subject to PSB, then the mean number of RNAs produced until transcription ceases should correspond to the mean number of transcription events between consecutive locking events. As it is not likely that the gene of interest is absent of effects from PSB in all cells, the empirical result should correspond to a lower bound estimate. Interestingly, from the same experiment, it should also be possible to measure τ_{active} (Fig. 1) from the time for RNA production to cease in all cells.

First, we tested whether Novobiocin, at the concentrations used here, affects cell morphology. For this, as above, we measured cell areas in the control condition (165 cells analyzed) and when subject to 100 ng /ml Novobiocin (180 cells analyzed), and then performed a 2-sample Kolmogorov-Smirnov test. We found that, at the significance level of 0.05, the two distributions cannot be distinguished (p-value of 0.13).

Next, we measured integer-valued number of RNAs in individual cells over time, every 15 min, 45 min after inducing the target gene (with IPTG) and adding Novobiocin (Gyrase inhibitor), so as to account for the mean time taken by cells to intake IPTG [64,86] and because only at this moment did we observe any tangible reduction in transcription activity (inset in Fig. 4). RNAs were detected by MS2-GFP tagging, preventing RNA degradation (Materials and Methods). We also performed a control experiment, where Novobiocin was not introduced.

Results in Fig. 4 show that when and only when adding Novobiocin,

the RNA production ceases. In the presence of Novobiocin, on average we observed 0.8 ± 0.11 RNAs per cell after 15 min. Considering mean cell division times (Fig. S5), we estimated the mean number of RNAs produced per cell for 15 min to be ~1.04 ± 0.14. This agrees (statistically) with the above estimation of N (~1.8 ± 0.84). It also agrees with past estimations that, in live cells, transcription initiation locking can occur after less than 5 transcription events [9].

We also extracted the time for transcription events to cease after introducing Novobiocin. For this, we performed additional time-lapse microscopy (1 min interval between images). The number of RNAs produced in individual cells during the observation time were obtained as in (66) and verified by visual inspection. Results in the inset of Fig. 4 show that transcription activity started to be reduced at minute 1 and that no RNA was produced after 10 min, which can be used as a lower bound for τ_{active} (see above). This agrees with the previous estimation of τ_{active} (~12 ± 4 min) from Fig. 3.

3.7. Effects of PSB differ with the basal transcription rate

Previous works reported evidence that a gene's activity affects its own PSB when the gene is on a circular template tethered to a surface [9,87]. We hypothesized that the same occurs on a chromosome-integrated gene, due to discrete topological constraints. This follows from the reasoning that, if the expected time interval between consecutive transcription events becomes longer, while [G] is kept constant, there is more time for Gyrase to resolve transcription initiation locking due to PSB in between transcription events. The model in Fig. 1 accounts for this, as the responsiveness of r^{-1} to changes in [G] should decrease with k_1. To test this, we replaced $P_{LacO3O1}$ by a native Lac promoter (P_{Lac}). We chose this promoter because it has similar sequence and repression-activation mechanism (Methods), which could affect PSB, and because it exhibits slower RNA production when fully induced (Supplementary Fig. S4A). By being in the same location, we expect the contribution to PSB from the activity of neighboring genes to be the same.

First, we obtained an induction curve of P_{Lac} (Supplementary Fig. S4A). Visibly, under maximum induction, P_{Lac} has a slower transcription rate than $P_{LacO3O1}$ (less ~62% MS2-GFP tagged RNAs per cell). In detail, r^{-1} (P_{Lac}) equals 2704 ± 493 s (obtained as described in Supplementary Section IX).

Next, we measured by qPCR the transcription rate for various [G] (as in Fig. 3). Results were scaled by r^{-1} in the control condition (Fig. 5A, black diamonds). Afterwards, we fitted a line by WTLS (black line in Fig. 5A) and corrected its slope by accounting for changes in [RNAP] (Supplementary Section VIII). Finally, we fitted a (green) line by WTLS to the corrected data points (green diamonds in Fig. 5A). From the best fitting (green) line in Fig. 5A we find that, for maximum [G] (0.2% Rhamnose), the increase in [RNAP] explains 28% of the increase in r, with the remaining 72% being due to increased [G] and/or unknown factors.

To assess if the effects of PSB differ with the promoter strength, we plotted r^{-1} against $([G]/[G_C])^{-1}$ for both constructs (P_{Lac} and $P_{LacO3O1}$). Results in Fig. 5B show that r^{-1} decreases faster with [G] for $P_{LacO3O1}$ (in agreement with the model). We thus conclude that changing [G] has smaller effects in the effective transcription rate of the lesser active promoter (P_{Lac}).

3.8. Inference of the parameter values of the model that best fit the empirical data and prediction of τ_{locked} as a function of the basal transcription rate

We searched for parameter values for the model (Fig. 1) that best match the empirical data of both $P_{LacO3O1}$ and P_{Lac}, assuming that they differ only in the basal transcription rate (k_1). We found that the model fits the empirical data with a mean squared error of 0.0004 and R^2 values larger than 0.95 (Supplementary Figs. S7A and S7B).

From the fitting, we obtained the parameter values (α, β_1, β_2, and η, Supplementary Section V) and inferred the duty cycles of transcription initiation locking due to neighboring and 'self-produced' PSB, for each [G] (Table S6). From Table S6, the lower are [G] and τ_{active}, the longer will the gene remain locked and the higher is its OFF/ON duty cycle ratio.

In addition, we used the inferred values of α, β_1, β_2, and η, to extrapolate τ_{locked} relative to r^{-1} as a function of $\frac{[G]}{[G_C]}$ and $\frac{1}{k_1 \times [RNAP]}$. The inferred surface is shown in Fig. 6.

If the differences in the locking dynamics due to PSB of $P_{LacO3O1}$ and P_{Lac} are solely due to the difference in their values of k_1, as hypothesized, this surface should fit other empirical values of $\frac{1}{k_1 \times [RNAP]}$ and $\frac{\tau_{Locked}}{r^{-1}}$ obtained when changing k_1 (e.g. by tuning their induction strength).

We thus performed qPCR measurements when inducing $P_{LacO3O1}$ with 0 and 50 μM IPTG (from Supplementary Fig. S4A, note that, at these concentrations, the number of RNAs produced differs significantly from maximum induction). The results from qPCR measurements, added to Fig. 6, fit well the predicted surface, suggesting that combining data from a promoter(s) of differing basal transcription rates in the same location in the DNA one can predict a state space of possible kinetics of transcription initiation locking of genes differing in k_1 in a given chromosomal location.

4. Discussion and conclusions

Past studies have shown that DNA topology and gene expression mutually affect one another [4,14,37,38].

We found that, for a certain range of Gyrase concentrations, the inverse of the transcription rate of a chromosome-integrated gene controlled by $LacO_3O_1$ changes linearly with the inverse of Gyrase concentration, while not perturbing cell growth or morphology. Given this, we developed and validated a method that uses a LineWeaver-Burk plot to dissect, from single-cell, single-RNA data, key kinetic parameters of transcription initiation locking due to PSB. Namely, we dissected the rate of occurrence of these locks and their weight on the effective RNA production rate. Next, we compared with a promoter at the same chromosomal location and similar in structure and regulation but differing in strength. From this, we inferred a range of potential kinetics of transcription initiation locking in a given topological domain that can be achieved by tuning the basal transcriptional rate of the gene of interest. Relevantly, the method was sensitive to detect PSB effects causing a minimum of 1.6 fold changes in transcription rates. Further, we estimate that simple enhancements (e.g. increasing the number of data points used for the LineWeaver-Burk plot from 3 to 10) reduced this to 1.2 fold changes. Other improvements (e.g. higher precision in data collection) should further enhance the sensitivity, which should suffice to, e.g., dissect the effects of interference between closely spaced promoters (Supplementary Section X).

To an extent, the interpretation of the empirical data relies on the models and, thus, it is necessary to assess their reliability, i.e. the robustness of their predictions. In this regard, we observed that, first, the models accurately estimated how much of the change in r^{-1}, following Gyrase overexpression, is due to changes in RNAP numbers (Fig. 3). In detail, the dynamics of the chromosome integrated gene, when corrected for RNAP changes (with this correction relying on the model), only differs from the plasmid dynamics by 3.5% (not statistically significant) while, prior to considering the model, it differed by 28%. Second, the model predicted the mean time to lock the promoter due to PSB (τ_{active} ~12 min) from qPCR and population level microscopy data. This estimation was validated by direct measurements using time-lapse microscopy data at the single cell level (one image per minute). In detail, estimated and real data differed solely by ~15% that, when accounting for the measurement error, is also not statistically significant (Figs. 3 and 4). Further, the model accurately predicted (in a statistical sense) how much the basal transcription rate affects the

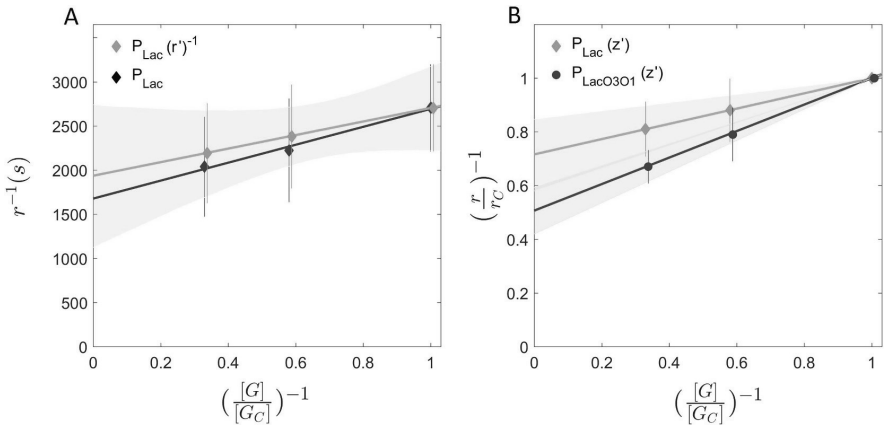


Fig. 5. LineWeaver-Burk plots for P_{Lac} and $P_{LacO3O1}$. (A) LineWeaver-Burk plot of the inverse of the RNA production rate of the chromosome-integrated Lac gene for different Gyrase concentrations (black diamonds), relative to the control (0% Rhamnose). Also shown are the standard error of the mean (vertical error bars), along with the best-fit line (black line). Further shown are the inverse of the RNA production rates corrected for the increased RNAP concentration when overexpressing Gyrase (r'^{-1}), and the correspondent best-fit line (green line) and its standard error of the mean (light green area) obtained by Monte Carlo simulations (10,000 iterations). The green diamonds are 0.008 units deviated to the right, for figure legibility. The line equations are $r^{-1} = (768 \pm 1096) \times \left(\frac{[G]}{[G_C]}\right)^{-1} + (1936 \pm 802)$ and $r^{-1} = (1016 \pm 1096) \times \left(\frac{[G]}{[G_C]}\right)^{-1} + (1677 \pm 802)$ for the green and black lines, respectively. RNA production rates were obtained by qPCR and microscopy. (B) LineWeaver-Burk plot of the inverse of the fold change in RNA production rate of the chromosome-integrated gene under the control of $LacO_3O_1$ (blue circles) and of the chromosome-integrated under the control of Lac gene (green diamonds) against the inverse of the Gyrase concentrations (0, 0.1 and 0.2% Rhamnose induction), measured by qPCR, relative to the control condition (0% Rhamnose). Vertical error bars represent the standard error of the mean. In addition, shown are the best-fit lines and their standard errors of the mean (green and light blue areas), obtained by Monte Carlo simulations (500 iterations). Both lines (blue and green) were corrected for the effects of the RNAP increase in the RNA production rate when overexpressing Gyrase. z' stands for the ratio $\left(\frac{r}{r_C}\right)^{-1}$ after the correction. The blue circles are 0.008 units deviated to the right, for legibility. The line equations are $\left(\frac{r}{r_C}\right)^{-1} = (0.28 \pm 0.13) \times \left(\frac{[G]}{[G_C]}\right)^{-1} + (0.72 \pm 0.13)$ and for the green and blue lines, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

fraction of time spent by the promoter in locked state (Fig. 6). Finally, the estimations of k_1 and k_{unlock} using the models agree with past estimations (respectively in [9,48]).

Overall, the results suggest that the weight of PSB on the effective RNA production rate of a gene depends not only on the mean activity of the DNA loop that the gene belongs to, but also on the basal transcription activity of the observed gene. This dependence was found to be sufficiently strong to require the introduction of this phenomenon in the model, if one is to predict the effects of changing Gyrase levels on the dynamics of transcription (reaction 4 in Fig. 1). This is because the fraction of time spent in locked states depends not only on the rate of accumulation of positive supercoils, but also on how much time Gyrases have to resolve enough supercoils (in between consecutive transcription events) to avoid reaching a supercoiling density that suffices for promoter locking.

Given that increasing the *basal* transcription rates enhances the influence of PSB on the *effective* transcription rate, we hypothesize that, at least in some genes, increasing the basal transcription rate may come at the cost of increased transcriptional noise due to PSB, even if lowering the noise from basal transcription dynamics. We thus expect that the relationship between basal transcription rate and PSB needs to be directly accounted for in models of prokaryotic gene expression. As such, when reducing the SN model (Supplementary Table S2) to a minimal model (Fig. 1), one of the critical components kept from the SN model was reaction 4 (Fig. 1), as it is responsible for the production positive supercoils at a rate that differs with the basal transcription rate of the gene interest.

This was required even though, similar to past models [4,9], there is also reaction 5, which introduces positive supercoils from 'external' sources, at a rate that differs with the average transcriptional activity of all genes in the same DNA loop (or topological domain) [17,43] and DNA replication [84]. Interestingly, the existence of this dependency suggests that it should be possible to, some extent, regulate the robustness of chromosome-integrated synthetic circuits to PSB, by tuning its own transcription rates, as well as placing it in a topological domain with desired mean activity.

In this regard, since increasing the basal transcription rate enhances the effects of PSB, is there an effective upper limit on the transcription rate? If so, this could potentially explain (at least partially) why some genes exist in multi-copy form. Such form would allow crossing this limit, while also supporting more stable expression levels.

Meanwhile, the combination of the results from two different constructs suggest that it may be possible to map a state space of transcription initiation locking of the topological domains of *E. coli*. However, since domain barriers are not likely to be at fixed sites [45,70,92], it may be necessary to set constructs in various regions of the DNA and measure not only the mean, but also the variability of the propensity for transcription locking as a function of DNA location. Using several constructs, differing in features (e.g. in regulatory mechanisms), should allow accounting for changes in parameters, other than the basal transcription rate. Namely, while here we mapped a 1-dimensional space by tuning the basal transcription rate, changing other variables would facilitate mapping a multi-dimensional state-space of transcription initiation locking kinetics. We expect such mapping to be of use in dissecting global transcription programs of *E. coli*, as well as for implementing chromosome-integrated synthetic circuits with predictable kinetics.

Our methodology may also assist in quantifying effects of environmental shifts (e.g. temperature) on the kinetics of transcription initiation locking. One could then explore whether *E. coli* uses this

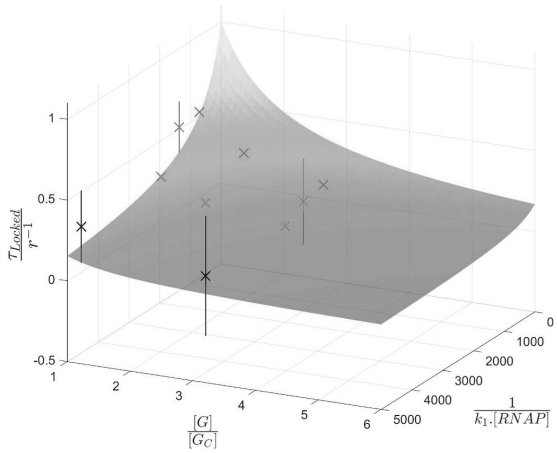


Fig. 6. Expected time in locked states relative to the expected interval between consecutive RNA production events as a function of Gyrase concentration and of the inverse of the basal transcription rate (k_1^{-1}), with $\tau_{active} = \frac{1}{k_1.RNAP}$. The surface is the model prediction of the relative τ_{Locked} as a function of τ_{active} and of the Gyrase concentration relative to the control. Red crosses are the empirical data for the $LacO_3O_1$ promoter under full induction (1000 μM IPTG), green crosses are the empirical data for the native Lac promoter under full induction (1000 μM IPTG), grey crosses are the empirical data for the $LacO_3O_1$ promoter under 50 μM IPTG induction, and black crosses are the empirical data for the $LacO_3O_1$ promoter uninduced (0 μM IPTG). The vertical bars are the standard error of the mean. All error bars intersect the surface. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

phenomenon to adapt to fluctuating environments. This hypothesis is supported by recent observations [64] that cold-shock genes have atypical supercoiling-sensitivity (for unknown reasons). I.e., genes with long-term responses to cold-shocks appear to be impervious to supercoiling, while genes with short-term responses have more-than-expected-by-chance sensitivity to supercoiling. Our methodology may assist in dissecting the responsible mechanisms, e.g. by measuring τ_{locked} and τ_{active} of these genes following mutations, etc.

We expect our methodology to be compatible with other techniques. E.g., it is potentially valuable to combine it with measurements of local DNA supercoiling density, such as trimethylpsoralen intercalation [40], to quantify the relationship between this density and the effects of PSB on transcription. Similarly, it may be valuable to combine it with the method in [48] to dissect the kinetics of rate limiting steps of active transcription initiation. For chromosome-integrated genes, we expect that only by using both methods will be possible to estimate the times spent prior to open complex formation, since models suggest that this state of activity will differ with the kinetics of promoter locking due to PSB [59], due to the expected competition between the formations of closed complexes and locked states.

Further, our methodology should be applicable using other techniques, such as RNA FISH (Fluorescence in situ hybridization) [79] and RNA aptamer-fluorogen systems [12,19,65,82,98].

Finally, our results derived from a first attempt at dissecting the *in vivo* dynamics of locking of transcription initiation using a Lineweaver-Burk plot. Many questions remain unanswered and require further study. It may turn out that fluctuations in Gyrase concentration have non-uniform effects at the genome-wide level, due to the dependency on basal transcription rates and mean rates of topological domains. Potentially, this could be used by cells as means to activate specific gene cohorts (e.g. of genes sharing the same topological domain), involved in responsive transcriptional programs. It could also be used to change the state of small genetic circuits responsible for triggering response programs to fluctuations in supercoiling density (e.g. fluctuations in supercoiling densities may alter the stable state of a, e.g., genetic switch with genes in different topological domains). If this holds true, the 'optimal' level of Gyrase may differ with the environment and/or internal cell state, depending on whether a given gene cohort (supercoiling density dependent) should be active or not.

In conclusion, the methods and results here presented are expected to support near-future research on the role of Gyrase on the global dynamics of gene regulatory networks.

Funding

This work was supported by the Finnish Academy of Science and Letters [to C.P.]; Pirkanmaa Regional Fund [to V.K.]; Tampere University Graduate Program (Finland) [to V.C. and M.B.]; EDUFI Fellowship [TM-19-11105 to S.D.]; 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.

Author's statement

C.P. and A.S.R. conceived the study. C.P. performed data analysis. C.P., M.B., and A.S.R. performed modelling. V.K.K., M.M., V.C., and S.D. performed measurements. C.P., V.K., and A.S.R. drafted the manuscript, which was revised by all authors. The authors declare no competing interests.

Transparency document

The Transparency document associated this article can be found, in online version.

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.bbagrm.2020.194515.

References

[1] R.E. Ashley, A. Dittmore, S.A. McPherson, C.L. Turnbough, K.C. Neuman, N. Osheroff, Activities of gyrase and topoisomerase IV on positively supercoiled DNA, Nucleic Acids Res. 45 (16) (2017) 9611–9624, https://doi.org/10.1093/nar/gkx649.
[2] N. Blot, R. Mavathur, M. Geertz, A. Travers, G. Muskhelishvili, Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome, EMBO Rep. 7 (7) (2006) 710–715, https://doi.org/10.1038/sj.embor.7400729.
[3] J.D. Boeke, P. Model, A prokaryotic membrane anchor sequence: carboxyl terminus of bacteriophage f1 gene III protein retains it in the membrane, Proc. Natl. Acad. Sci. U. S. A. 79 (17) (1982) 5200, https://doi.org/10.1073/pnas.79.17.5200.
[4] C.H. Bohrer, E. Roberts, A biophysical model of supercoiling dependent transcription predicts a structural aspect to gene regulation, BMC Biophys. 9 (1) (2015) 2, https://doi.org/10.1186/s13628-016-0027-0.
[5] G. Casella, R.L. Berger, Statistical Inference, Thomson Learning, 2002.
[6] J.J. Champoux, DNA topoisomerases: structure, function, and mechanism, Annu. Rev. Biochem. 70 (1) (2001) 369–413, https://doi.org/10.1146/annurev.biochem.70.1.369.
[7] B. Cheng, C.-X. Zhu, C. Ji, A. Ahumada, Y.-C. Tse-Dinh, Direct interaction between *Escherichia coli* RNA polymerase and the zinc ribbon domains of DNA topoisomerase I, J. Biol. Chem. 278 (33) (2003) 30705–30710, https://doi.org/10.1074/jbc.M303403200.
[8] B.-K. Cho, D. Kim, E.M. Knight, K. Zengler, B.O. Palsson, Genome-scale reconstruction of the sigma factor network in *Escherichia coli*: topology and functional states, BMC Biol. 12 (1) (2014) 4, https://doi.org/10.1186/1741-7007-12-4.
[9] S. Chong, C. Chen, H. Ge, X.S. Xie, Mechanism of transcriptional bursting in Bacteria, Cell 158 (2) (2014) 314–326, https://doi.org/10.1016/j.cell.2014.05.038.
[10] K.A. Datsenko, B.L. Wanner, One-step inactivation of chromosomal genes in

Escherichia coli K-12 using PCR products, Proc. Natl. Acad. Sci. 97 (12) (2000) 6640–6645, https://doi.org/10.1073/pnas.120163297.
[11] S. Deng, R.A. Stein, N.P. Higgins, Organization of supercoil domains and their reorganization by transcription, Mol. Microbiol. 57 (6) (2005) 1511–1521, https://doi.org/10.1111/j.1365-2958.2005.04796.x.
[12] E.V. Dolgosheina, S.C.Y. Jeng, S.S.S. Panchapakesan, R. Cojocaru, P.S.K. Chen, P.D. Wilson, N. Hawkins, P.A. Wiggins, P.J. Unrau, RNA mango aptamer-fluorophore: a bright, high-affinity complex for RNA labeling and tracking, ACS Chem. Biol. 9 (10) (2014) 2412–2420, https://doi.org/10.1021/cb500499x.
[13] T. Dong, H.E. Schellhorn, Global effect of RpoS on gene expression in pathogenic *Escherichia coli* O157:H7 strain EDL933, BMC Genomics 10 (1) (2009) 349, https://doi.org/10.1186/1471-2164-10-349.
[14] C.J. Dorman, M.J. Dorman, DNA supercoiling is a fundamental regulatory principle in the control of bacterial gene expression, Biophys. Rev. 8 (S1) (2016) 89–100, https://doi.org/10.1007/s12551-016-0238-2.
[15] K. Drlica, Control of bacterial DNA supercoiling, Mol. Microbiol. 6 (4) (1992) 425–433, https://doi.org/10.1111/j.1365-2958.1992.tb01486.x.
[16] M. Drolet, Growth inhibition mediated by excess negative supercoiling: the interplay between transcription elongation, R-loop formation and DNA topology, Mol. Microbiol. 59 (2006) 723–730, https://doi.org/10.1111/j.1365-2958.2005.05006.x.
[17] D. El Hanafi, L. Bossi, Activation and silencing of leu-500 promoter by transcription-induced DNA supercoiling in the Salmonella chromosome, Mol. Microbiol. 37 (3) (2000) 583–594, https://doi.org/10.1046/j.1365-2958.2000.02015.x.
[18] A. Farewell, K. Kvint, T. Nyström, Negative regulation by RpoS: a case of sigma factor competition, Mol. Microbiol. 29 (4) (1998) 1039–1051, https://doi.org/10.1046/j.1365-2958.1998.00990.x.
[19] G.S. Filonov, J.D. Moon, N. Svensen, S.R. Jaffrey, Broccoli: rapid selection of an RNA mimic of green fluorescent protein by fluorescence-based selection and directed evolution, J. Am. Chem. Soc. 136 (46) (2014) 16299–16308, https://doi.org/10.1021/ja508478x.
[20] K. Fujita, M. Iwaki, T. Yanagida, Transcriptional bursting is intrinsically caused by interplay between RNA polymerases on DNA, Nat. Commun. 7 (2016) 13788, , https://doi.org/10.1038/ncomms13788.
[21] G. Fulcrand, S. Dages, X. Zhi, P. Chapagain, B.S. Gerstman, D. Dunlap, F. Leng, DNA supercoiling, a critical signal regulating the basal expression of the lac operon in *Escherichia coli*, Sci. Rep. 6 (1) (2016) 19243, , https://doi.org/10.1038/srep19243.
[22] M. Gellert, M.H. O'Dea, T. Itoh, J. Tomizawa, Novobiocin and coumermycin inhibit DNA supercoiling catalyzed by DNA gyrase, Proc. Natl. Acad. Sci. U. S. A. 73 (12) (1976) 4474–4478, https://doi.org/10.1073/pnas.73.12.4474.
[23] D.T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, J. Comput. Phys. 22 (4) (1976) 403–434, https://doi.org/10.1016/0021-9991(76)90041-3.
[24] D.T. Gillespie, Exact stochastic simulation of coupled chemical reactions, J. Phys. Chem. 81 (1977) 2340–2361, https://doi.org/10.1021/j100540a008.
[25] I. Golding, E.C. Cox, RNA dynamics in live *Escherichia coli* cells, Proc. Natl. Acad. Sci. U. S. A. 101 (31) (2004) 11310–11315, https://doi.org/10.1073/pnas.0404443101.
[26] I. Golding, J. Paulsson, S.M. Zawilski, E.C. Cox, Real-time kinetics of gene activity in individual bacteria, Cell 123 (6) (2005) 1025–1036, https://doi.org/10.1016/j.cell.2005.09.031.
[27] N.S.M. Goncalves, S.M.D. Oliveira, V. Kandavalli, J.M. Fonseca, A.S. Ribeiro, Temperature dependence of leakiness of transcription repression mechanisms of *E. coli*, Lect. Notes Comput. Sci 9859 (2016) 341–342.
[28] P. Guptasarma, Cooperative relaxation of supercoils and periodic transcriptional initiation within polymerase batteries, BioEssays 18 (4) (1996) 325–332, https://doi.org/10.1002/bies.950180411.
[29] A. Hakkinen, A.-B. Muthukrishnan, A. Mora, J.M. Fonseca, A.S. Ribeiro, CellAging: a tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*, Bioinformatics 29 (13) (2013) 1708–1709, https://doi.org/10.1093/bioinformatics/btt194.
[30] A. Häkkinen, A.S. Ribeiro, Characterizing rate limiting steps in transcription from RNA production times in live cells, Bioinformatics 32 (9) (2016) 1346–1352, https://doi.org/10.1093/bioinformatics/btv744.
[31] C.D. Hardy, N.R. Cozzarelli, Alteration of *Escherichia coli* topoisomerase IV to novobiocin resistance. *Antimicrob*, Agents Chemother. 47 (3) (2003) 941–947, https://doi.org/10.1128/aac.47.3.941-947.2003.
[32] Y. Hayakawa, T. Murotsu, K. Matsubara, Mini-F protein that binds to a unique region for partition of mini-F plasmid DNA, J. Bacteriol. 163 (1) (1985) 349–354.
[33] N.P. Higgins, Species-specific supercoil dynamics of the bacterial nucleoid, Biophys. Rev. 8 (S1) (2016) 113–121, https://doi.org/10.1007/s12551-016-0207-9.
[34] V.K. Kandavalli, H. Tran, A.S. Ribeiro, Effects of σ factor competition are promoter initiation kinetics dependent, Biochim. Biophys. Acta - Gene Regul. Mech. 1859 (10) (2016) 1281–1288, https://doi.org/10.1016/j.bbagrm.2016.07.011.
[35] K. Kirkegaard, J.C. Wang, Bacterial DNA topoisomerase I can relax positively supercoiled DNA containing a single-stranded loop, J. Mol. Biol. 185 (3) (1985) 625–637, https://doi.org/10.1016/0022-2836(85)90075-0.
[36] N. Kleckner, J.K. Fisher, M. Stouf, M.A. White, D. Bates, G. Witz, The bacterial nucleoid: nature, dynamics and sister segregation, Curr. Opin. Microbiol. 22 (2014) 127–137, https://doi.org/10.1016/j.mib.2014.10.001.
[37] M.V. Kotlajich, D.R. Hron, B.A. Boudreau, Z. Sun, Y.L. Lyubchenko, R. Landick, Bridged filaments of histone-like nucleoid structuring protein pause RNA polymerase and aid termination in bacteria, Elife 4 (2015), https://doi.org/10.7554/eLife.04970.
[38] F. Kouzine, S. Sanford, Z. Elisha-Feil, D. Levens, The functional response of upstream DNA to dynamic supercoiling *in vivo*, Nat. Struct. Mol. Biol. 15 (2) (2008) 146–154, https://doi.org/10.1038/nsmb.1372.
[39] M. Krystek, M. Anton, A weighted total least-squares algorithm for fitting a straight line, Meas. Sci. Technol. 18 (11) (2007) 3438–3442, https://doi.org/10.1088/0957-0233/18/11/025.
[40] A. Lal, A. Dhar, A. Trostel, F. Kouzine, A.S.N. Seshasayee, S. Adhya, Genome scale patterns of supercoiling in a bacterial chromosome, Nat. Commun. 7 (1) (2016) 11055, , https://doi.org/10.1038/ncomms11055.
[41] T.B.K. Le, M.V. Imakaev, L.A. Mirny, M.T. Laub, High-resolution mapping of the spatial organization of a bacterial chromosome, Science 342 (6159) (2013) 731–734, https://doi.org/10.1126/science.1242059.
[42] F. Leng, B. Chen, D.D. Dunlap, Dividing a supercoiled DNA molecule into two independent topological domains, Proc. Natl. Acad. Sci. 108 (50) (2011) 19973–19978, https://doi.org/10.1073/pnas.1109854108.
[43] D.M. Lilley, C.F. Higgins, Local DNA topology and gene expression: the case of the leu-500 promoter, Mol. Microbiol. 5 (4) (1991) 779–783, https://doi.org/10.1111/j.1365-2958.1991.tb00749.x.
[44] H. Lineweaver, D. Burk, The determination of enzyme dissociation constants, J. Am. Chem. Soc. 56 (3) (1934) 658–666, https://doi.org/10.1021/ja01318a036.
[45] V.S. Lioy, A. Cournac, M. Marbouty, S. Duigou, J. Mozziconacci, O. Espéli, F. Boccard, R. Koszul, Multiscale structuring of the *E. coli* chromosome by nucleoid-associated and condensin proteins, Cell 172 (4) (2018) 771–783.e18, https://doi.org/10.1016/j.cell.2017.12.027.
[46] L.F. Liu, J.C. Wang, Supercoiling of the DNA template during transcription, Proc. Natl. Acad. Sci. U. S. A. 84 (20) (1987) 7024–7027, https://doi.org/10.1073/pnas.84.20.7024.
[47] Y. Liu, A.M. Berrido, Z.-C. Hua, Y.-C. Tse-Dinh, F. Leng, Biochemical and biophysical properties of positively supercoiled DNA, Biophys. Chem. 230 (2017) 68–73, https://doi.org/10.1016/j.bpc.2017.08.008.
[48] J. Lloyd-Price, S. Startceva, V. Kandavalli, J.G. Chandraseelan, N. Goncalves, S.M.D. Oliveira, A. Häkkinen, A.S. Ribeiro, Dissecting the stochastic transcription initiation process in live *Escherichia coli*, DNA Res. 23 (3) (2016) 203–214, https://doi.org/10.1093/dnares/dsw009.
[49] A.S. Lynch, J.C. Wang, Anchoring of DNA to the bacterial cytoplasmic membrane through cotranscriptional synthesis of polypeptides encoding membrane proteins or proteins for export: a mechanism of plasmid hypernegative supercoiling in mutants deficient in DNA topoisomerase I, J. Bacteriol. 175 (6) (1993) 1645, https://doi.org/10.1128/jb.175.6.1645-1655.1993.
[50] D. Ma, D.N. Cook, M. Alberti, N.G. Pon, H. Nikaido, J.E. Hearst, Genes acrA and acrB encode a stress-induced efflux system of *Escherichia coli*, Mol. Microbiol. 16 (1) (1995) 45–55, https://doi.org/10.1111/j.1365-2958.1995.tb02390.x.
[51] J. Ma, L. Bai, M.D. Wang, Transcription under torsion, Science 340 (6140) (2013) 1580–1583, https://doi.org/10.1126/science.1235441.
[52] J. Ma, M.D. Wang, DNA supercoiling during transcription, Biophys. Rev. 8 (Suppl. 1) (2016) 75–87, https://doi.org/10.1007/s12551-016-0215-9.
[53] J. Mäkelä, J. Lloyd-Price, O. Yli-Harja, A.S. Ribeiro, Stochastic sequence-level model of coupled transcription and translation in prokaryotes, BMC Bioinformatics 12 (1) (2011) 121, https://doi.org/10.1186/1471-2105-12-121.
[54] J. Mäkelä, V. Kandavalli, A.S. Ribeiro, Rate-limiting steps in transcription dictate sensitivity to variability in cellular components, Sci. Rep. 7 (1) (2017) 10588, , https://doi.org/10.1038/s41598-017-11257-2.
[55] H. Mannerstrom, O. Yli-Harja, A.S. Ribeiro, Inference of kinetic parameters of delayed stochastic models of gene expression using a markov chain approximation, EURASIP J. Bioinform. Syst. Biol. 2011 (1) (2011) 572876, , https://doi.org/10.1155/2011/572876.
[56] L. Martins, R. Neeli-Venkata, S.M.D. Oliveira, A. Häkkinen, A.S. Ribeiro, J.M. Fonseca, SCIP: a single-cell image processor toolbox, Bioinformatics (Oxford, England) 34 (24) (2018) 4318–4320, https://doi.org/10.1093/bioinformatics/bty505.
[57] M. Matsumoto, T. Nishimura, Mersenne Twister, ACM Trans. on Modeling and Comp. Simulation 8 (1998) 3–30, https://doi.org/10.1145/272991.272995.
[58] W.R. McClure, Mechanism and control of transcription initiation in prokaryotes, Annu. Rev. Biochem. 54 (1) (1985) 171–204, https://doi.org/10.1146/annurev.bi.54.070185.001131.
[59] N. Mitarai, I.B. Dodd, M.T. Crooks, K. Sneppen, The generation of promoter-mediated transcriptional noise in Bacteria, PLoS Comput. Biol. 4 (7) (2008) e1000109, , https://doi.org/10.1371/journal.pcbi.1000109.
[60] H. Mori, A. Kondo, A. Ohshima, T. Ogura, S. Hiraga, Structure and function of the F plasmid genes essential for partitioning, J. Mol. Biol. 192 (1) (1986) 1–15, https://doi.org/10.1016/0022-2836(86)90459-6.
[61] K. Nevo-Dinur, A. Nussbaum-Shochat, S. Ben-Yehuda, O. Amster-Choder, Translation-independent localization of mRNA in *E. coli*, Science 331 (6020) (2011) 1081–1084, https://doi.org/10.1126/science.1195691.
[62] S. Oehler, M. Amouyal, P. Kolkhof, B. von Wilcken-Bergmann, B. Müller-Hill, Quality and position of the three lac operators of *E. coli* define efficiency of repression, EMBO J. 13 (14) (1994) 3348–3355.
[63] S.M.D. Oliveira, R. Neeli-Venkata, N.S.M. Goncalves, J.A. Santinha, L. Martins, H. Tran, J. Mäkelä, A. Gupta, M. Barandas, A. Häkkinen, J. Lloyd-Price, J.M. Fonseca, A.S. Ribeiro, Increased cytoplasm viscosity hampers aggregate polar segregation in Escherichia coli, Mol. Microbiol. 99 (4) (2016) 686–699, https://doi.org/10.1111/mmi.13257.
[64] S.M.D. Oliveira, N.S.M. Goncalves, V.K. Kandavalli, L. Martins, R. Neeli-Venkata, J. Reyelt, J.M. Fonseca, J. Lloyd-Price, H. Kranz, A.S. Ribeiro, Chromosome and plasmid-borne PLacO3O1 promoters differ in sensitivity to critically low temperatures, Sci. Rep. 9 (1) (2019) 4486, https://doi.org/10.1038/s41598-019-39618-z.
[65] J.S. Paige, K.Y. Wu, S.R. Jaffrey, RNA mimics of green fluorescent protein, Science 333 (6042) (2011) 642–646, https://doi.org/10.1126/science.1207339.

[66] C.S.D. Palma, S. Startceva, R. Neeli-Venkata, M. Zare, N.S.M. Goncalves, J.M. Fonseca, S.M.D. Oliveira, A.S. Ribeiro, A strategy for dissecting the kinetics of transcription repression mechanisms, Proceedings of the European Medical and Biolgical Engineering Conference (EMBEC), June 11–15, Tampere, Finland, 65 Springer, Singapore, 2017, pp. 1097–1100, , https://doi.org/10.1007/978-981-10-5122-7_274 Also published in: IFMBE Proceedings.

[67] D.S. Peabody, The RNA binding site of bacteriophage MS2 coat protein, EMBO J. 12 (2) (1993) 595.

[68] D.S. Peabody, Role of the coat protein-RNA interaction in the life cycle of bacteriophage MS2, Mol. Gen. Genet. MGG 254 (4) (1997) 358–364, https://doi.org/10.1007/s004380050427.

[69] B.J. Peter, J. Arsuaga, A.M. Breier, A.B. Khodursky, P.O. Brown, N.R. Cozzarelli, Genomic transcriptional response to loss of chromosomal supercoiling in *Escherichia coli*, Genome Biol. 5 (11) (2004) R87, https://doi.org/10.1186/gb-2004-5-11-r87.

[70] L. Postow, C.D. Hardy, J. Arsuaga, N.R. Cozzarelli, Topological domain structure of the *Escherichia coli* chromosome, Genes Dev. 18 (14) (2004) 1766–1779, https://doi.org/10.1101/gad.1207504.

[71] G.J. Pruss, K. Drlica, Topoisomerase I mutants: the gene on pBR322 that encodes resistance to tetracycline affects plasmid DNA supercoiling, Proc. Natl. Acad. Sci. U. S. A. 83 (23) (1986) 8952–8956, https://doi.org/10.1073/pnas.83.23.8952.

[72] T. Rajala, A. Häkkinen, S. Healy, O. Yli-Harja, A.S. Ribeiro, Effects of transcriptional pausing on gene expression dynamics, PLoS Comput. Biol. 6 (3) (2010) 29–30, https://doi.org/10.1371/journal.pcbi.1000704.

[73] A. Revyakin, R.H. Ebright, T.R. Strick, Promoter unwinding and promoter clearance by RNA polymerase: detection by single-molecule DNA nanomanipulation, Proc. Natl. Acad. Sci. 101 (14) (2004) 4776–4780, https://doi.org/10.1073/pnas.0307241101.

[74] A.S. Ribeiro, J. Lloyd-Price, SGN Sim, a stochastic genetic networks simulator, Bioinformatics 23 (6) (2007) 777–779, https://doi.org/10.1093/bioinformatics/btm004.

[75] A.S. Ribeiro, O.-P. Smolander, T. Rajala, A. Häkkinen, O. Yli-Harja, Delayed stochastic model of transcription at the single nucleotide level, J. Comput. Biol. 16 (4) (2009) 539–553, https://doi.org/10.1089/cmb.2008.0153.

[76] P.E. Rouvière, A. De Las Peñas, J. Mecsas, C.Z. Lu, K.E. Rudd, C.A. Gross, rpoE, the gene encoding the second heat-shock sigma factor, sigma E, in *Escherichia coli*, EMBO J. 14 (5) (1995) 1032–1042.

[77] N. Rovinskiy, A.A. Agbleke, O. Chesnokova, Z. Pang, N.P. Higgins, Rates of gyrase supercoiling and transcription elongation control supercoil density in a bacterial chromosome, PLoS Genet. 8 (8) (2012) e1002845, , https://doi.org/10.1371/journal.pgen.1002845.

[78] J. Santinha, L. Martins, A. Häkkinen, J. Lloyd-Price, S.M.D. Oliveira, A. Gupta, T. Annila, A. Mora, A.S. Ribeiro, J.R. Fonseca, iCellFusion: Tool for Fusion and Analysis of Live-Cell Images from Time-Lapse Multimodal Microscopy, (2016), https://doi.org/10.4018/978-1-4666-8811-7.ch004.

[79] R.H. Singer, D.C. Ward, 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 (23) (1982) 7331–7335, https://doi.org/10.1073/pnas.79.23.7331.

[80] C. Sissi, M. Palumbo, In front of and behind the replication fork: bacterial type IIA topoisomerases, Cell. Mol. Life Sci. 67 (12) (2010) 2001–2024, https://doi.org/10.1007/s00018-010-0299-5.

[81] S. Startceva, V.K. Kandavalli, A. Visa, A.S. Ribeiro, Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression, Biochim. Biophys. Acta - Gene Regul. Mech. 1862 (2) (2019) 119–128, https://doi.org/10.1016/j.bbagrm.2018.12.005.

[82] R.L. Strack, M.D. Disney, S.R. Jaffrey, A superfolding Spinach2 reveals the dynamic nature of trinucleotide repeat–containing RNA, Nat. Methods 10 (12) (2013) 1219–1224, https://doi.org/10.1038/nmeth.2701.

[83] M. Stracy, C. Lesterlin, F. Garza de Leon, S. Uphoff, P. Zawadzki, A.N. Kapanidis, Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid, Proc. Natl. Acad. Sci. 112 (32) (2015) E4390–E4399, https://doi.org/10.1073/pnas.1507592112.

[84] M. Stracy, A.J.M. Wollman, E. Kaja, J. Gapinski, J.-E. Lee, V.A. Leek, S.J. McKie, L.A. Mitchenall, A. Maxwell, D.J. Sherratt, M.C. Leake, P. Zawadzki, Single-molecule imaging of DNA gyrase activity in living *Escherichia coli*, Nucleic Acids Res. 47 (1) (2019) 210–220, https://doi.org/10.1093/nar/gky1143.

[85] Y. Taniguchi, P.J. Choi, G.W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, X.S. Xie, Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells, Sci. (New York, NY) 329 (5991) (2010) 533–538, https://doi.org/10.1126/science.1188308.

[86] H. Tran, S.M.D. Oliveira, N. Goncalves, A.S. Ribeiro, Kinetics of the cellular intake of a gene expression inducer at high concentrations, Mol. BioSyst. 11 (9) (2015) 2579–2587, https://doi.org/10.1039/C5MB00244C.

[87] Y.P. Tsao, H.Y. Wu, L.F. Liu, Transcription-driven supercoiling of DNA: direct biochemical evidence from in vitro studies, Cell 56 (1) (1989) 111–118, https://doi.org/10.1016/0092-8674(89)90989-6.

[88] T. Větrovský, P. Baldrian, The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses, PLoS One 8 (2) (2013) e57923, , https://doi.org/10.1371/journal.pone.0057923.

[89] S.M. Vos, E.M. Tretter, B.H. Schmidt, J.M. Berger, All tangled up: how cells direct, manage and exploit topoisomerase function, Nat. Rev. Mol. Cell Biol. 12 (12) (2011) 827–841, https://doi.org/10.1038/nrm3228.

[90] J.C. Wang, DNA topoisomerases, Annu. Rev. Biochem. 54 (1) (1985) 665–697, https://doi.org/10.1146/annurev.bi.54.070185.003313.

[91] James C. Wang, DNA Topoisomerases, Annu. Rev. Biochem. 65 (1) (1996) 635–692, https://doi.org/10.1146/annurev.bi.65.070196.003223.

[92] X. Wang, P.M. Llopis, D.Z. Rudner, Organization and segregation of bacterial chromosomes, Nat. Rev. Genet. 14 (3) (2013) 191–203, https://doi.org/10.1038/nrg3375.

[93] A. Wegerer, T. Sun, J. Altenbuchner, Optimization of an *E. coli* L-rhamnose-inducible expression vector: test of various genetic module combinations, BMC Biotechnol. 8 (1) (2008) 2, https://doi.org/10.1186/1472-6750-8-2.

[94] A. Wheeler, Digital Microscopy, in: A. Wheeler, R. Henriques (Eds.), Standard and Super-Resolution Bioimaging Data Analysis, 2017, https://doi.org/10.1002/9781119096948.ch1.

[95] H.Y. Wu, S.H. Shyy, J.C. Wang, L.F. Liu, Transcription generates positively and negatively supercoiled domains in the template, Cell 53 (3) (1988) 433–440, https://doi.org/10.1016/0092-8674(88)90163-8.

[96] E. Yeung, A.J. Dy, K.B. Martin, A.H. Ng, D. Del Vecchio, J.L. Beck, J.J. Collins, R.M. Murray, Biophysical constraints arising from compositional context in synthetic gene networks, Cell Syst 5 (1) (2017) 11–24.e12, https://doi.org/10.1016/j.cels.2017.06.001.

[97] E.L. Zechiedrich, A.B. Khodursky, S. Bachellier, R. Schneider, D. Chen, D.M.J. Lilley, N.R. Cozzarelli, Roles of topoisomerases in maintaining steady-state DNA supercoiling in Escherichia coli, J. Biol. Chem. 275 (11) (2000) 8103–8113, https://doi.org/10.1074/jbc.275.11.8103.

[98] J. Zhang, J. Fei, B.J. Leslie, K.Y. Han, T.E. Kuhlman, T. Ha, Tandem spinach array for mRNA imaging in living bacterial cells, Sci. Rep. 5 (1) (2015) 17295, , https://doi.org/10.1038/srep17295.

[99] R. Samul, F. Leng, Transcription-coupled Hypernegative Supercoiling of Plasmid DNA by T7 RNA Polymerase in Escherichia coli Topoisomerase I-Deficient Strains, Journal of Molecular Biology 374 (4) (2007) 925–935, https://doi.org/10.1016/j.jmb.2007.10.011.

Supplementary Material

Dissecting the *in vivo* dynamics of transcription locking due to positive supercoiling buildup

Cristina S.D. Palma[1], Vinodh Kandavalli[1], Mohamed N.M. Bahrudeen[1], Marco Minoia[1], Vatsala Chauhan[1], Suchintak Dash[1], and Andre S. Ribeiro[1*]

[1] Laboratory of Biosystem Dynamics, BioMediTech, Faculty of Medicine and Health Technology, Tampere University, Tampere University, 33101 Tampere, Finland.

* To whom correspondence should be addressed. Email: andre.sanchesribeiro@tuni.fi

Present Address: Andre S. Ribeiro, BioMediTech Institute, Tampere University, Arvo Ylpön katu 34, P.O.Box 100, 33014 Tampere, Finland.

SUPPLEMENTARY METHODS

I. RNA quantification from fluorescent spots, spots lifetime, and spots full tagging time

Integer-valued number of MS2-GFP-tagged mRNA molecules in individual cells are obtained from microscopy images as in e.g. (Häkkinen et al., 2013; Oliveira et al., 2016). Shortly, MS2-GFP tagged RNA spots are segmented by Kernel Density Estimation (KDE). Example Figures S2A and S2E show cellular backgrounds generated by unbound MS2-GFP proteins in cells carrying the plasmid-borne and the chromosome integrated target genes, respectively. Meanwhile, Figures S2B and S2F show these cells when with tagged RNAs along with the results of the spot detection methods (Häkkinen and Ribeiro, 2015). These spots are visible to the Human eye and, as seen, detectable by image analysis (Santinha et al., 2016), since their fluorescence is much higher than in near-neighbour pixels (Figures S2D and S2H). In addition, the variability in fluorescence intensity of pixels without spots is much smaller than the mean difference in intensity between pixels with and without spots (Figures S2D and S2H), which lowers the risk of detecting 'false' spots and removing 'true' spots. Consequently, the background fluorescence intensity (average over all pixels not containing 'RNA-spots') can safely be subtracted from the intensity of each fluorescent RNA-spot.

From the resulting RNA-spot fluorescence intensities in arbitrary units (a.u.), we estimate the intensity of individual MS2-GFP tagged RNAs as in (Golding et al., 2005). From histograms of intensities of RNA-spots, we find the intensity of the first "peak" of the histogram (which should correspond to the intensity of one tagged RNA). Next, for each spot, we round its intensity value to the nearest integer, to obtain its integer-valued number of RNA molecules.

We also considered the possibility that some spots correspond to incomplete RNAS due to e.g. arrests during transcription elongation. This could lead to overestimation of RNA numbers. To determine whether this could occur, let us assume, as an example, that incomplete RNAs have, on average, 50% the total intensity of a completely tagged RNA. From the histogram of spots intensities for RNAs produced by the plasmid-borne gene, we estimated than only 1.6 % of the detected spots have less than 50% of the mean intensity of 1 RNA. Similar values were obtained for the

chromosome-integrated gene. This implies that small, forming RNAs introduce little to no error in RNA counting. Further, not all 'weak spots' will be forming RNAs (e.g. a few could be out-of-focus tagged RNAs). Thus, the value of 1.6 % could be considered to be an upper bound for such fraction of forming RNAs that are erroneously counted as fully formed RNAs. From this, we conclude that little to no error is added to the RNA numbers per cell due to this.

As in (Tran et al., 2015), we measured MS2-GFP tagged RNAs decrease in fluorescence intensity during time-lapse microscopy (Figure S9). We then estimated the mean half-life of tagged RNAs by fitting the intensity of each MS2-GFP tagged RNA over time with a decaying exponential function. We found the mean half-time of spots fluorescence to be longer than the measurement period (>150 min.), as in (Häkkinen and Ribeiro, 2015; Tran et al., 2015). In addition, 'bleaching' of MS2 tagged RNAs was not observed. Finally, we observed that target RNA molecules become fully tagged by MS2-GFP in <1 minute (Tran et al., 2015). As such, tagging times are not considered as influencing RNA counting.

To determine if MS2-GFP proteins formed clusters in the absence of target RNA, we analyzed cells with the reporter system (responsible for producing MS2-GFP proteins) but lacking the target system (coding for target RNAs). In these cells, the number of 'fake' spots detected by the image analysis algorithm was approximately 100 times smaller than in cells with a target system. As such, the influence of 'fake spots' or abnormal MS2-GFP clusters is considered to be negligible. Finally, visual inspection of the images showed that all fake spots were due to failures in the image analysis, rather than the presence of MS2-GFP clusters visible to the Human eye.

II. Single-nucleotide model of transcription subject to the effects of PSB

We use a stochastic model of transcription with stepwise elongation at the single nucleotide level based on past models (Rajala et al., 2010; Ribeiro et al., 2009; Mäkelä et al., 2011). All reactions are described in Supplementary Table S2. Parameter values are extracted or derived from empirical data. The main improvement of the model, compared to past similar models, is the introduction of a dynamic PSB phenomenon at the single-nucleotide level.

Positive supercoiling is generated by the RNAP activity on both the neighbour genes (Kouzine et al., 2013; Naughton et al., 2013; Teves et al., 2014; Lilley et al., 1991; Rhee et al., 1999) as well as on the gene of interest (Chong et al., 2014). We assume a constant (stochastic) rate of accumulation of positive supercoils due to the activity of neighbour genes. This rate is expected to differ with the neighbours' location (e.g. whether they are, or not, in the same transcriptional unit), distance from the gene of interest, direction of transcription activity, and external factors, such as environmental perturbations (Weinstein-Fisher et al., 2000; Cheung et al., 2003).

The effects of PSB in the model are: i) elongation arrests at the nucleotide level (resulting in short pauses, which slowdown elongation); and ii) transcription initiation locking, which causes longer transcription activity breakdowns whose resolution requires Gyrase intervention (Chong et al., 2014). The propensities of these events are dynamic, in that they differ with the global level of PSB in the region of the DNA where the gene of interest is located (Ma et al., 2014 and Chong et al., 2014).

To model this process at the single nucleotide level, we first introduce a reaction for transcription initiation, where a promoter is found by an RNAP (reaction S2.1), followed by promoter escape (reaction S2.5), which initiates stepwise elongation (reaction S2.6). As soon as the promoter becomes unoccupied, a new transcription initiation event can occur.

As the RNAP percolates the DNA, following each elongation step from one nucleotide to the next (reaction S2.6), an activation step (reaction S2.7) needs to occur for the RNAP to further progress to the subsequent nucleotide. However, the following events compete with activation: pausing (reaction S2.8), arrest (reactions S2.11), editing (reaction S2.12), premature termination (reaction S2.14) and pyrophosphorolysis (reaction S2.15). All these events, except premature termination, are modelled as reversible, due to the ability of the transcription machinery to resolve them. Finally, pyrophosphorolysis results in moving one step backwards, implying that it does not require resolution.

The model also allows for pauses and pause escapes to occur due to collisions between RNAPs (reactions S2.9 and S2.10, respectively). Further, misincorporation can occur at the end of the transcription process (reaction S2.13). Provided no misincorporation or premature termination, elongation is completed (reaction S2.16) and an RNA is produced and the RNAP is released (reaction S2.16).

In addition to all these events, we also model a dynamic process of accumulation of positive supercoils, which has a direct impact on RNA production (Travers et al., 2005; Lesne et al., 2018). Specifically, PSB causes short arrests to the moving RNAP (accounted for in reaction S2.11), which increase in frequency with increasing PSB (Ma et al., 2013; Fujita et al., 2016). Overall, this progressively decreases the rate of elongation. In addition, for high enough PSB, it can halt transcription initiation (accounted for in reaction S2.2) (Ma et al., 2013; Chong et al., 2014). Further, supercoils are not static, i.e., they can diffuse through the DNA. In some cases, they can reach regions located thousands of base pairs away from the point of origin. Evidence for this include, e.g., the observation of “topological promoter coupling” (Kouzine et al., 2013; Naughton et al., 2013; Teves et al., 2014; Lilley et al., 1991; Rhee et al., 1999), when supercoils produced in the activity of one gene reach the transcription start site of another gene.

Given the above, in the model, positive supercoils can accumulate from two sources: i) RNAP activity on neighbour genes (reaction S2.4) and, ii) RNAP activity on the gene of interest (reaction S2.6) via the production of positive supercoils (SC^+). The number of such SC^+ units allows quantifying the level of PSB in the region of the gene of interest at any given moment. In detail, the RNAP needs to percolate ~ 10 nucleotides for one positive supercoil to accumulate (Stracy et al., 2019; Rovinskiy et al., 2012). This is implemented in reaction S2.6, where the creation of a SC^+ requiring the percolation by the RNAP of 10 nucleotides (i.e. only in 1 out of 10 nucleotides will a supercoil be created). Finally, for simplicity, the model does not record the location of positive supercoils, only their total amount in the region of the gene of interest.

Also modelled is the process of SC^+ removal by the direct action of Gyrases (reaction S2.17) (Gellert et al., 1976; Chong et al., 2014; Stracy et al., 2019).

As a side note, it is physically possible for transcription initiation to halt because the RNAP becomes unable to bind to the promoter (Mitarai et al.,2008), as well as because the RNAP becomes

unable to unwind the promoter once bound (Revyakin et al., 2014). For simplicity, we model only the former phenomenon. Further, it is noted that *E. coli* has mechanisms to handle the effects of PSB in the kinetics of elongation other than the intervention of Gyrases. For example, GreB allows the RNAP to transcribe more efficiently through supercoiled regions of the DNA, by limiting backtracking (Ma et al., 2019). The detailed phenomena are not explicitly modelled here but, in most cases, their effects are indirectly accounted for in the rates of RNAP arresting, etc.

Finally, several phenomena may be more PSB-dependent than currently represented in the model (e.g. pausing, pyrophosphorolysis, etc. (Ma et al., 2019)). Due to the present lack of knowledge of the quantitative relationship between PBS levels and the rates of these events, we opted for modelling them as independent phenomena, using the currently available empirical parameter values in optimal growth conditions. Nevertheless, as noted, events in elongation (aside from misincorporation and pyrophosphorolysis) are not expected to affect the mean RNA production rate.

III. Dynamics of the single-nucleotide model of transcription when subject to the effects of PSB

III.1 Dynamics of stepwise transcription elongation

We first test whether the model in Supplementary Table S2 can mimic the effects of PSB in the kinetics of transcription elongation. For this, we performed simulations at various relative Gyrase concentrations ([G]/[G_C]), with [G_C] being the concentration of Gyrases in the control condition. From these simulations, we extracted the time-length of multiple stepwise transcription elongation events ($\Delta t_{elongation}$). Results in Figure S6A show that as [G] increases, both the mean and the standard error of the mean of $\Delta t_{elongation}$ decrease. This can be explained by the results in Figures S6C and S6D, which show that both the mean number of SC^+ in the DNA, as well as the mean rate of arrests during elongation decrease for increasing Gyrase.

Finally, also from Figure S6A, as expected, $\Delta t_{elongation}$ converges to a minimum value once the number of Gyrases approaches values that suffice to remove positive supercoils as fast as they appear, not allowing their accumulation.

III.2 Effects of elongation slowdown on the dynamics of RNA production

Next, we show that slowdown of stepwise elongation rates due to PSB does not affect the mean time interval between consecutive RNA production events (therefore not affecting the mean RNA production rate). For this, we simulated the model described in Table S2, but without reactions S2.2 to S2.4, so that one can change the number of Gyrases without affecting transcription initiation rates. This allows testing whether the effects of PSB on elongation (alone) alters the mean RNA production rate.

Using this model, we performed simulations for various values of [G] to obtain the mean rate of RNA production (r) as a function of [G]. Figure S6B shows that r is not significantly affected by [G], within realistic intervals of these parameters' values. This entails that the mean RNA production rate is independent from the effects of PSB on elongation. This is expected, provided that the events in

elongation do not affect the rate of transcription initiation and have negligible effects on the fraction of RNAPs that complete elongation, once initiated.

In this regard, only if the arrests due to PSB (reaction S2.11, Table S2) were long enough that the number of accumulated RNAPs in the DNA strand became so high that new transcription events would not be allowed to initiate due to promoter occupancy.

III.3 RNAP and Gyrase fluctuations within the nucleoid region

Both the SN model and the minimal model assume homogeneous mixing of free Gyrases and RNAP inside each cell (or, more precisely, inside the nucleoid region containing the DNA). There are two reactions in Supplementary Table S2 whose kinetics could be affected, in case this assumption does not hold true, specifically, reaction S2.1 (by which RNAPs bind to the promoter) and reaction S2.17 (by which Gyrases remove positive supercoils).

The assumption of homogenously distributed free RNAP in the nucleoid region is supported by live-cell super-resolution microscopy data (Stracy et al., 2015). Further, it has been showed that, within certain ranges, the total RNAP concentration can be used as a proxy for free RNAP concentration when estimating mean RNA production rates (Lloyd-Price et al, 2016), which would not be expected if the free RNAP had significant spatial fluctuations. It is also noteworthy that time intervals between consecutive transcription events are relatively long (700-1500 s, Figure 3) when compared to the diffusion rate of the RNAP (Bratton et al., 2011), supporting the assumption of a well-stirred system, which allows for the assumption of stochastic rate constants (Gillespie, 1977).

Meanwhile, to test the validity of the model assumption of 'homogeneous mixing of Gyrases in the nucleoid region', we measured their spatial heterogeneity by microscopy (example Figure S8A), using an *E. coli* strain where the gyrA gene is endogenously tagged with the YFP coding sequence (Taniguchi et al., 2010) (Materials and Methods, Section 2.1). From the data, we found that almost all Gyrases are located in the cell region(s) where the nucleoid(s) locate (example Figure S8A).

To estimate the spatial heterogeneity of Gyrases in those regions, we first applied a 2-dimensional Gaussian filter (Materials and Methods, Section 2.7), to remove measurement noise (Wheeler, 2017). Supplementary Figure S8B shows the 'raw' and 'filtered' distributions of pixel intensities (arbitrary units). Next, we multiplied each data point of the filtered distribution by a constant (equal to the ratio of the mean over the variance of the filtered distribution), which results in a Poisson distribution (named 'scaled distribution'). Figure S8C shows both this scaled distribution (inset) as well as its probability density function (pdf). The variance of the scaled distribution is expected to be a good proxy for the spatial (and thus temporal) variability in Gyrase numbers within the nucleoid region.

Subsequently, starting from the minimal model in Figure 1, we introduced the additional reactions in Table S7, so as to test if the *in silico* results are significantly affected by inserting in the model this degree of variability in Gyrase numbers over time. Since the reactions in Table S7 are first-order processes of production (reaction S7.1) and degradation (reaction S7.2) of Gyrases, they ensure that the number of Gyrases at any given time follows a Poisson distribution as the empirical data suggests (since the propensity of each event is constant and independent of the occurrence of the former

event). We then tuned the rate constants (*kp_G* and *kd_G*) so that the *in silico* distribution best fitted the empirical distribution (Figure S8C).

We then used the best fitted model, obtained an *in silico* Δt distribution, and compared to the same distribution, obtained prior to introducing the temporal variability in Gyrase's numbers. A 2-sample t-test did not reject the null-hypothesis that the two distributions cannot be statistical distinguished (p-value > 0.05), from which we conclude that the heterogeneity in Gyrases is not sufficiently high to affect the dynamics of the model.

Overall, we conclude that the additional process (Table S7), added to account for the measured heterogeneity in Gyrase numbers, does not change the RNA production kinetics sufficiently (in a statistical sense) for the single-cell distributions of RNA numbers to differ significantly. Given this and the above, for simplicity, we assume homogenous spatial distributions of RNAP and Gyrases in the region(s) occupied by the nucleoid(s).

IV. Model fitting

Assuming the model in Figure 1, we inferred the parameters that best fit the empirical data as follows. From the model, the inverse of the RNA production rate equals:

$$r_x^{-1} = \left[\frac{k_{lock}}{k_{unlock}\cdot k_{rem}}\cdot\frac{k_p\cdot[R_x]}{k_1\cdot[R_x]}\frac{1}{[G_x]}\right]+\left[\frac{k_{lock}\cdot\lambda}{k_{unlock}\cdot k_{rem}}\cdot\frac{1}{[G_x]}\right]+\frac{1}{k_1.[R_x]} \quad \text{(iv.1)}$$

The variables, r_x^{-1}, $[R_x]$ and $[G_x]$ refer to, respectively, the inverse of the RNA production rate, the concentration of RNAP and the concentration of Gyrases, for a condition 'x' of Gyrase overexpression. For the reference, control condition ('ref'), equation (iv.1) becomes:

$$r_{ref}^{-1} = \left[\frac{k_{lock}}{k_{unlock}\cdot k_{rem}}\cdot\frac{k_p\cdot[R_{ref}]}{k_1\cdot[R_{ref}]}\frac{1}{[G_{ref}]}\right]+\left[\frac{k_{lock}\cdot\lambda}{k_{unlock}\cdot k_{rem}}\cdot\frac{1}{[G_{ref}]}\right]+\frac{1}{k_1.[R_{ref}]} \quad \text{(iv.2)}$$

Next, we assume: $w=\frac{[R_x]}{[R_{ref}]}$ (iv.2.1), $y=\frac{[G_x]}{[G_{ref}]}$ (iv.2.2) and $z=\frac{r_x^{-1}}{r_{ref}^{-1}}$ (iv.2.3). Thus:

$$r_{ref}^{-1}\cdot z = \left[\frac{k_{lock}}{k_{unlock}\cdot k_{rem}}\cdot\frac{k_p\cdot[R_{ref}]}{k_1\cdot[R_{ref}]}\frac{1}{[G_{ref}]}\cdot\frac{1}{y}\right]+\left[\frac{k_{lock}\cdot\lambda}{k_{unlock}\cdot k_{rem}}\cdot\frac{1}{[G_{ref}]}\cdot\frac{1}{y}\right]+\left[\frac{1}{k_1.[R_{ref}]}\cdot\frac{1}{w}\right] \quad \text{(iv.3)}$$

From (iv.1), (iv.2.3), and (iv.3):

$$z = \left[\frac{\frac{k_{lock}}{k_{unlock}\cdot k_{rem}}\cdot\frac{k_p\cdot[R_{ref}]}{k_1\cdot[R_{ref}]}\frac{1}{[G_{ref}]}}{r_{ref}^{-1}}\right]\cdot\frac{1}{y}+\left[\frac{\frac{k_{lock}\cdot\lambda}{k_{unlock}\cdot k_{rem}}\cdot\frac{1}{[G_{ref}]}}{r_{ref}^{-1}}\right]\cdot\frac{1}{y}+\left[\frac{\frac{1}{k_1.[R_{ref}]}}{r_{ref}^{-1}}\right]\cdot\frac{1}{w} \quad \text{(iv.4)}$$

Assuming, $\alpha = \frac{k_{lock} \cdot k_p \cdot [R_{ref}]}{k_{unlock} \cdot k_{rem} \cdot [G_{ref}]}$ (iv.4.1), $\beta = \left[\frac{1}{\frac{k_1 \cdot [R_{ref}]}{r_{ref}^{-1}}} \right]$ (iv.4.2), $\eta = \frac{\frac{k_{lock} \cdot \lambda}{k_{unlock} \cdot k_{rem}} \cdot \frac{1}{[G_{ref}]}}{r_{ref}^{-1}}$ (iv.4.3),

equation (iv.4) becomes:

$$z = \alpha\beta \cdot \frac{1}{y} + \eta \cdot \frac{1}{y} + \beta \cdot \frac{1}{w} \quad \text{(iv.5)}$$

Since $LacO_3O_1$ and the native Lac promoter are located in the same position in the chromosome, we assume that they have the same propensity to become locked due to PSB due to variables other than their own transcription rate. Thus, it is imposed that k_{unlock}, k_{lock}, and k_p (Figure 1) do not differ between them. Positive supercoils' removal should also not differ. As such:

$$z_{LacO3O1} = \alpha.\beta_1.\frac{1}{y} + \beta_1.\frac{1}{w} + \eta.\frac{1}{y} \quad \text{(iv.6)}$$

$$z_{Lac} = \alpha.\beta_2.\frac{1}{y} + \beta_2.\frac{1}{w} + \eta.\mathrm{X}.\frac{1}{y} \quad \text{(iv.7)}$$

where $\beta_1 = \frac{\frac{1}{k_{1\,LacO3O1} \cdot [R_{ref}]}}{r_{ref\,LacO3O1}^{-1}}$ (iv.7.1) , $\beta_2 = \frac{\frac{1}{k_{1\,Lac} \cdot [R_{ref}]}}{r_{ref\,Lac}^{-1}}$ (iv.7.2), $\eta = \frac{\frac{k_{lock} \cdot \lambda}{k_{unlock} \cdot k_{rem}} \cdot \frac{1}{[G_{ref}]}}{r_{ref\,LacO3O1}^{-1}}$ (iv.7.3), and

$\mathrm{X} = \frac{r_{ref\,LacO3O1}^{-1}}{r_{ref\,Lac}^{-1}}$ (iv.7.4)

Given the empirical data in Table S5, from (iv.6) and (iv.7) we estimated α, β_1, β_2 and η by imposing the condition $\alpha, \beta_1, \beta_2, \eta \geq 0$. Namely, we searched for the set of solutions that minimizes the mean squared error (equation iv.8):

$$MSE = \frac{\sum_{i=1}^{N_1} \left(z_i^{LacO3O1} - \hat{z}_i^{LacO3O1}(\alpha, \beta_1, \lambda) \right)^2 + \sum_{i=1}^{N_2} \left(z_i^{Lac} - \hat{z}_i^{Lac}(\alpha, \beta_2, \lambda, X) \right)^2}{N_1 + N_2} \quad \text{(iv.8)}$$

The best fitting solution found was: $\alpha = 0.04$, and $\eta = 0.48$ for both promoters. Meanwhile, $\beta_1 = 0.5$ for $P_{LacO3O1}$, and $\beta_2 = 0.71$ for P_{Lac}.

Figures S7A and S7B show the resulting *z* surfaces for $P_{LacO3O1}$ and P_{Lac}. Both models best fit with a mean squared error of 0.004. To determine the goodness of fit of the surfaces we calculated R^2 values. Both surfaces had $R^2 > 0.95$, from which we conclude that the model well-fits the empirical data.

V. Inference of rate constants of $LacO_3O_1$

First, we infer the expected rate of transcription initiation events, in the absence of PSB. From the measurements (Figure 3, main manuscript) and model fitting (Section S.IV), we found that for $P_{LacO3O1}$ one has: r_{ref}^{-1} = 1476 s and β = 0.5. Given that, and the definition of β (equation iv.4.2), we find that $k_1.[R_{ref}]$ = 0.0014 s^{-1}. This is in agreement with the rate of initiation estimated in (Lloyd-Price 2016).

Next, we infer the expected rate of SC^+ production due to the transcription activity of neighboring genes. From equations (iv.4.1) and (iv.4.3), along with the values of k_{unlock} (Section 3.5, main manuscript), α, η, r_{ref}^{-1} (Supplementary Section IV) and λ (equaling a tenth of the number of nucleotides of the elongation region of $LacO_3O_1$) one has:

$$0.04 = \frac{k_{lock} \cdot k_p \cdot [R_{ref}]}{7\times10^{-4} \cdot k_{rem}.[G_{ref}]} \quad \text{(v.1)}$$

and

$$0.48 = \frac{\frac{k_{lock} \cdot 406}{7\times10^{-4} \cdot k_{rem}} \cdot \frac{1}{[G_{ref}]}}{1476} \quad \text{(v.2)}$$

Dividing equation (v.1) with equation (v.2) one has:

$$\frac{0.04}{0.48} = \frac{k_p.[R_{ref}].(1476)}{406} \quad \text{(v.3)}$$

From (v.3), one finds that $k_p \cdot [R_{ref}]$ equals 0.023. From this and equation (v.1), one obtains:

$$0.04 = \frac{0.023 \cdot k_{lock}}{7\times10^{-4} \cdot k_{rem}.[G_{ref}]} \quad \text{(v.4)}$$

From (v.4) one can write:

$$\frac{k_{lock}}{k_{rem}.[G_{ref}]} = 1.2\times10^{-3} \quad \text{(v.5)}$$

Next, we infer the expected rate of SC^+ removal. According to (Stracy et al, 2019) Gyrase dwell times are of 2 seconds to remove ~2 supercoils. Based on this, given that there is approximately one Gyrase molecule per DNA loop (Chong et al., 2014), we set: $k_{rem} \cdot [G_{ref}]$ = 1 s^{-1}. From this, along with equation (v.5), one estimates $k_{lock} = 1.2\times10^{-3}$ s^{-1}.

.

VI. Using the concentration of RNAP and of Gyrases as proxies for the concentrations of free RNAP and free Gyrases, respectively.

According to the model in Figure 1, one expects the transcription rate of a given gene to depend on the concentration of RNAP. However, at any given time, several RNA polymerases may not be available, if already committed to transcription. In detail, at any given time, a significant fraction of RNA polymerases are not available for new transcription events (estimations suggest that, at any given moment, ~48% of all RNAPs are bound to the DNA, interacting with promoter regions or involved transcription elongation (Stracy et al., 2015). Further, this task can take up to 75 s (Vogel and Jensen, 1994). As such, more accurately, and in accordance with the single nucleotide (SN) model, the transcription rate of a gene depends on the concentration of RNA polymerases that are *free* for transcription. Specifically, from reaction 1 in Figure 1, the inverse of the transcription rate of a gene should change linearly with the inverse of the free RNAP concentration.

Since one cannot easily measure the fraction of RNAP that is free for transcription at a given moment, for our estimations, we use the total concentration of RNA polymerases, [RNAP], as a proxy. This is possible because, within the range of conditions of the measurements, it was empirically verified in (Lloyd-Price et al., 2016) that the Lineweaver–Burk plot of the inverse of [RNAP] against the inverse of the transcription rate shows a straight line. This linear relationship was established by WTLS by minimizing χ^2 (Krystek and Anton, 2007), and then confirmed by showing that small deviations from linearity were not statistically significant, using likelihood ratio tests between the best linear fit and fits by higher order polynomials. In no case did the test reject the linear model (p-values above 0.1). This result was subsequently confirmed in (Kandavalli et al., 2016; Mäkelä et al., 2017; Oliveira et al., 2019; Startceva et al., 2019). Since this is strong evidence that the ratio between free and total RNAP concentrations is constant for the range of conditions tested here, we use the total RNAP concentration as a proxy for free RNAP concentration, relative to the control.

Similarly, from the equation in the large inset in Figure 1, one can also expect a straight line on a plot of the inverse of the free Gyrases concentration against the inverse of the transcription rate of the gene of interest, if [G] is a good proxy for the concentration of freely diffusing Gyrases [G^{free}]. In this regard, only ~49% of the Gyrases are expected to be free for resolving new transcription-generated positive supercoils, while the remaining ones are transiently maintaining steady state levels of negative supercoiling (~49%) and resolving replication-generated supercoiling (~2%) (Stracy et al., 2019). Further, resolving positive supercoils can take up to 1 s per supercoil (Stracy et al., 2019).

To test if [G] is a good proxy for [G^{free}], we measured RNA production rates in cells subject to various Gyrase concentrations (obtained by overexpressing Gyrase, see Materials and Methods in main manuscript). In Figure 3C (black line), we show the inverse of the transcription rate as a function of $[G]^{-1}$. Next, we fitted a line by WTLS and determined if small deviations from linearity are statistically significant by a likelihood ratio test between the best linear fit and fits by higher order polynomials (by WTLS by minimizing χ^2) (Krystek and Anton, 2007). The test did not reject the linear model (p-value > 0.99), from which we conclude that r^{-1} decreases linearly with $[G]^{-1}$. As such, in what pertains the extrapolation of τ_{active} and τ_{locked} from Lineweaver–Burk plots shown in the main manuscript, [G] is used as a proxy for [G^{free}] in the range of conditions considered (differing in the

concentration of Rhamnose, which is responsible for Gyrase, overexpression, see Materials and Methods in main manuscript).

Due to the above, in the main manuscript, for simplicity, we refer to the total concentrations of Gyrase and RNAP rather than to the concentrations of *free* Gyrase and RNAP molecules.

VII. Estimation of the quantitative relationship between the concentrations of Rhamnose and active gyrases

To obtain the fold change (F) in protein levels for a given mRNA fold change (due to adding Rhamnose), we use the calibration line ($Y_{Gyr.}$ = (0.85 +/- 0.06).X + (0.15+/- 0.07)) in Figure 2B**.** Let Y_{Gyr1} be the relative protein numbers corresponding to a Gyrase mRNA fold change of 1 and Y_{Gyr2} be the relative protein numbers for any given fold change. If b is the y-axes intersection (which equals 0.15, Figure 2B) one has:

$$F = \frac{Y_{Gyr_1} - b}{Y_{Gyr_2} - b} \qquad \text{(vii.1)}$$

Using this method, we found that fold changes in mRNA numbers resulted in the same fold changes in protein numbers (Table S3).

VIII. Dissection of the effects of RNAP overexpression in the RNA production rate of the target gene when overexpressing Gyrase.

To dissect the effects of RNAP overexpression from the direct effects of Gyrase overexpression on the RNA production rate of target gene, we estimated r^{-1} (equation iv.1) assuming that [G] has no effects on [RNAP] (i.e. considering that [R] = [R_{ref}]). Thus, we re-write equations iv.6 and iv.7 as:

$$z_x^{'} = \alpha.\beta_x.\frac{1}{y} + \beta_x + \eta.\frac{1}{y} \qquad \text{(viii.1)}$$

In (viii.1), $z_x^{'}$ represents z for promoter x assuming that [RNAP] is unaffected by Gyrase overexpression. The results for each condition are shown in Table S5.

IX. Extraction of RNA production rates from microscopy images

To estimate RNA production rates from the number of RNAs in individual cells obtained from microscopy images at two time points (Häkkinen and Ribeiro, 2015; Zimmer et al., 2016), we account for RNA dilution due to cell division, but not for RNA degradation, since the binding by multiple MS2-GFP molecules makes the tagged RNAs virtually immortal (Golding et al., 2005; Tran et al., 2015) and their fluorescence intensity constant for the duration of the measurements (Tran et al., 2015).

The rate of RNA dilution (k_d) due to cell division can be estimated from the numbers of cell division events between the start (t_0) and end (t_f) of the measurements, along with the mean RNA numbers at the start and end of the measurement period. In detail, let M_0 be the mean number of RNAs per cell at

moment t_0, and M be this number at moment t. It follows that, accounting for RNA dilution due to cell division, the rate of RNA production per cell (r) during that period of time equals:

$$r = \frac{k_d}{\ln 2} \cdot \frac{M - M_0 e^{-k_d \cdot t}}{1 - e^{-k_d \cdot t}} \qquad \text{(ix.1)}$$

From the values of t, M, M_0, and k_d extracted from empirical data, one can then obtain the mean of r^{-1}, and the standard error of the mean using the Delta Method (Casella et al., 2002).

X. Minimum supercoiling buildup effects that can be detected

It is possible to estimate the sensitivity of the method used in detecting effects of PSB on the transcription rate of the gene of interest. From the predicted bounds of the line fitting (obtained from the standard errors of the mean of each empirical data point), we estimated the minimum, detectable difference in transcription rates in the control condition assuming the same number of measurements as in Figure 3.

For this, using the blue line in Figure 3 as a starting point, we kept its y-intercept unchanged, and incrementally reduced its slope by 0.01 and calculated the resulting r^{-1} values for each of the 3 data points. The standard error of each data point was kept unchanged (which use it as an upper bound for their expected standard error of the mean). Next, we performed a 2-sample t-test with unequal variance to find if this slope differs from the slope of the grey line. We continued this procedure until this test could not find a difference.

Using this method, we found that the smallest slope below which the p-value is above the significance level (alpha=0.1) equals 512.5. This entails that the minimum fold change in transcription rate due to PSB that could be detected equals 1.6.

This value could be reduced by several procedures, such as collecting more data points per condition. For estimation, assume that the number of data points is increased from 3 to 10, selected within the range of 0.2 to 1 (inverse relative Gyrase concentration). From the best fitting line (blue line in Figure 3) and its predicted bounds, we estimate the mean and standard error for each data point (the mean is obtained from the blue line, with the error is set to be equal to the predicted bound, i.e. the width of the shadow area, in the same position in the x-axis). Next, using WTLS (Krystek and Anton, 2007) one can estimate a new best fit line using the estimated data points as above. From the resulting line, as above, we estimated the minimum fold change to equal 1.19. For improving this estimation, we repeated the process of collecting and processing 10 data points 1000 times, by randomly sampling 1000 data points from a normal distribution with the same mean and standard error for each data point. On average, the expected minimum fold change equalled 1.2. Other possible means of further reducing this minimum fold change include increasing the accuracy of the measurements of each data point (e.g. measuring more cells by microscopy, etc.).

Interestingly, this sensitivity is expected to suffice to detect effects of specific closely space promoter configurations. For example, using the same RNA detection technique, (Häkkinen et al., 2019) recently reported that the promoters tetA and $LacO_3O_1$, when in separate constructs, have a

mean RNA production rate of 1/800 RNA/s (~1/670 for tetA and ~1/1100 for $LacO_3O_1$), while in a tandem formation ($LacO_3O_1$ followed by TetA) the rate equals 1/700 RNA/s. Thus, the fold change between them equals 1.2, which is within the range of detectable fold changes.

SUPPLEMENTARY FIGURES

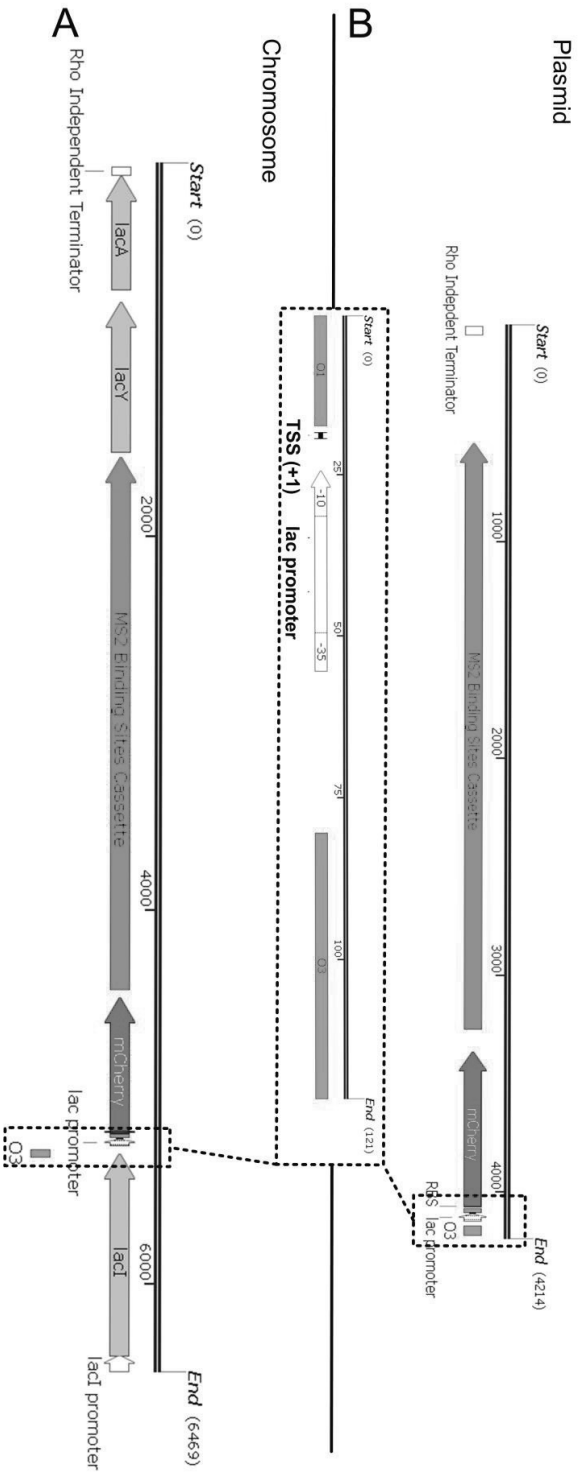


Figure S1. Genetic constructs. **(A)** Chromosome-integrated gene. mCherry-MS2-BS cassette under the control of $P_{LacO3O1}$ in the lac locus of *E. coli* strain BW25993 (ΔlacZ:mCherry-MS2-BSs), followed by the native lacY and lacA genes, and the native Rho-independent transcription termination site. **(B)** Single-copy plasmid-borne gene. mCherry-MS2-BS cassette under the control of $P_{LacO3O1}$ in a single-copy F-Plasmid in *E. coli* strain BW25993, followed by a Rho-independent transcription termination site. Constructs were confirmed by sequencing. Since the plasmid carrying the target gene does not code for lacY and lacA, and the cells carrying this plasmid also contain the original lacY and lacA genes in the chromosome, the two strains express lacY and lacA proteins similarly and, thus, do not differ significantly in the dynamics of IPTG intake. (Inset) The inset image in between the two constructs shows $P_{LacO3O1}$ with functional domains, which is identical in both constructs. It is in this region that the operator site O3 locates, followed by the RNA polymerase binding regions (positions -10 and -35), the transcription start site (TSS, position +1), and the operator site O1. Finally, the plasmid construct has a terminator upstream of the TSS that is 27 nucleotides long and is located 9 nucleotides downstream of the CmR gene (not represented in the figure). Thus, it is similar to the chromosome-integrated construct, where there is an upstream transcriptional terminator provided by the lacI gene. Related to Section 2.1, in main manuscript.

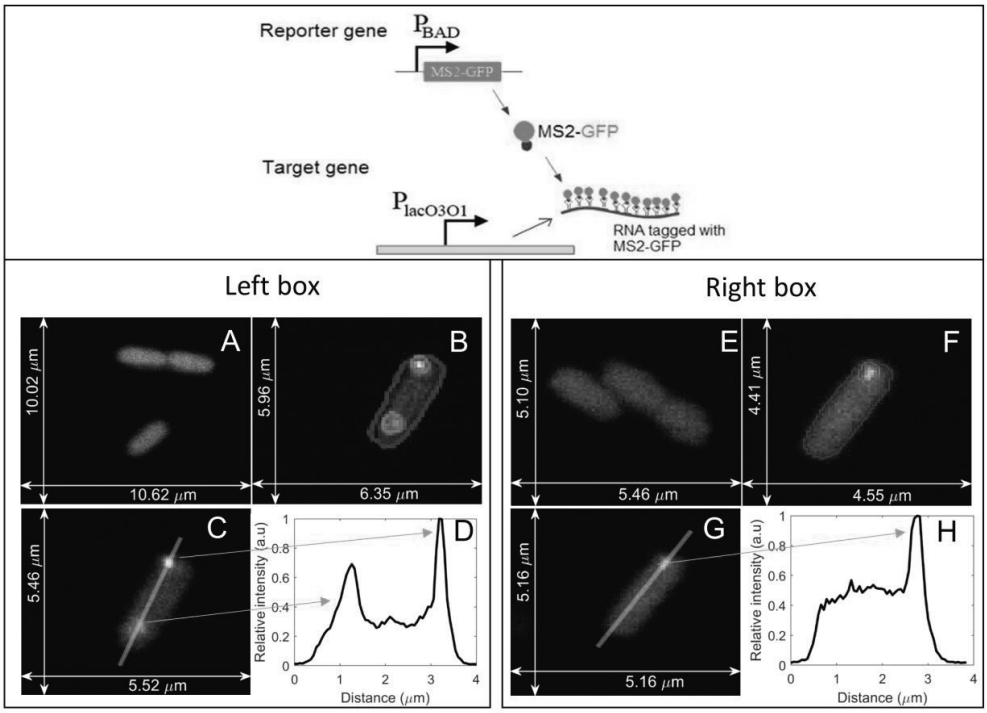


Figure S2. RNA quantification in individual cells by RNA tagging with MS2-GFP, due to which RNAs appear as fluorescent spots **(Top)** Schematic representation of the single-RNA MS2-GFP tagging detection system. Cells produce multiple MS2-GFP reporter proteins, under the control of P_{BAD}, while the production of RNAs target for MS2-GFP is under the control of $P_{LacO3O1}$. **(Left box)** Cells with a plasmid-borne promoter producing the target RNA. **(A)** Example microscopy image of cells carrying the reporter gene coding for MS2-GFP, prior to the production of target RNAs. The cells fluorescence is due to the large amount of MS2-GFP proteins. **(B)** Example microscopy image of cells carrying the reporter gene coding for MS2-GFP, after the production of target RNAs. The RNAs tagged with MS2-GFP are visible as bright spots. Blue line and red circles are the results of cell and RNA spot segmentation, respectively (Materials and Methods). **(C)** Example image of a cell along with a, manually introduced yellow line (using imageJ (Abramoff et al., 2004)) in order to obtain a fluorescence intensity profile along the major axis. **(D)** Pixel intensity (in arbitrary units) along the yellow line shown on (C). The peaks correspond to the regions where the two spots (tagged MS2-GFP RNAs) are located. **(Right box)** Cells with a chromosome-integrated promoter producing RNA target for MS2-GFP. Images from (E) to (H) have the same information as (A) to (D), respectively, but are obtained using cells with a chromosome-integrated promoter responsible for the production of the RNA target for MS2-GFP. Related to Supplementary Section I.

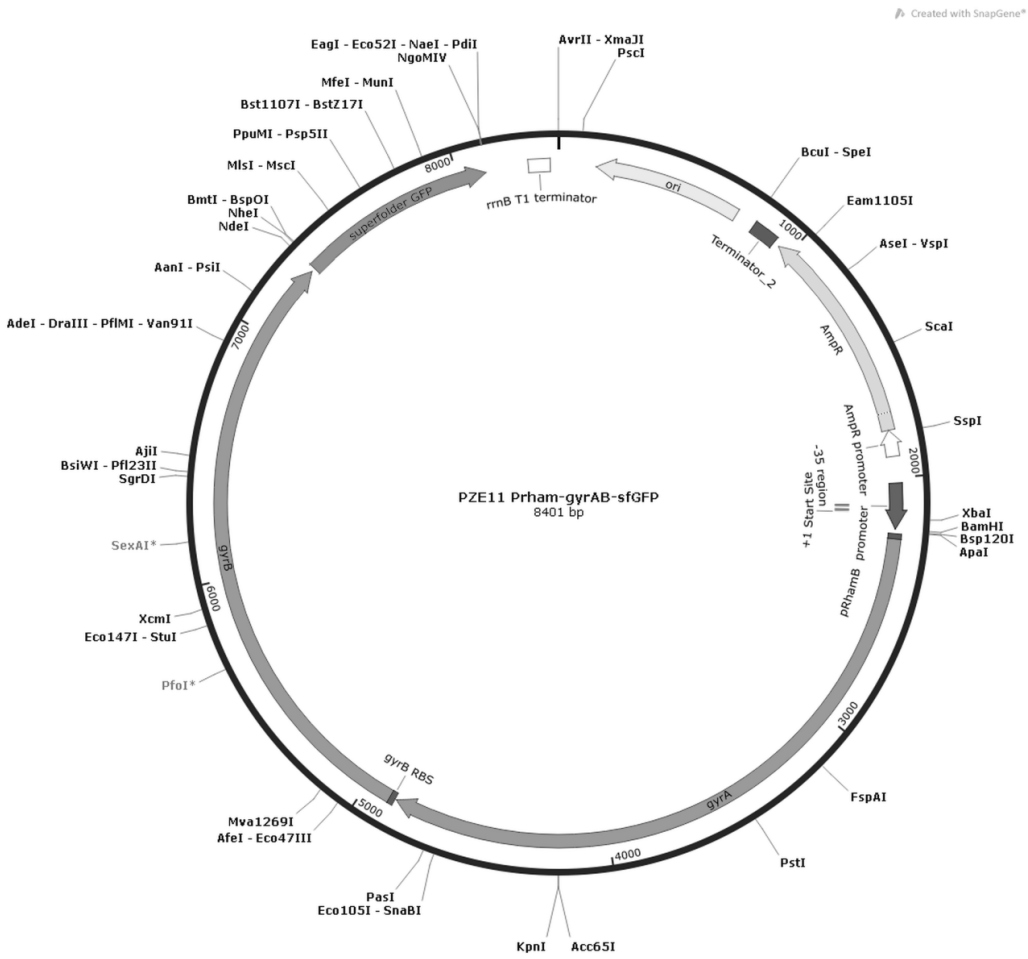


Figure S3. Schematic representation of the plasmid responsible for Gyrase overexpression. This plasmid was constructed by placing the gyraseA and gyraseB under the control of the PRhamB promoter, which is inducible by Rhamnose, and was transformed into BW25993 cells. Adapted from SnapGene® 1.5.2. Related to Section 2.1, in main manuscript.

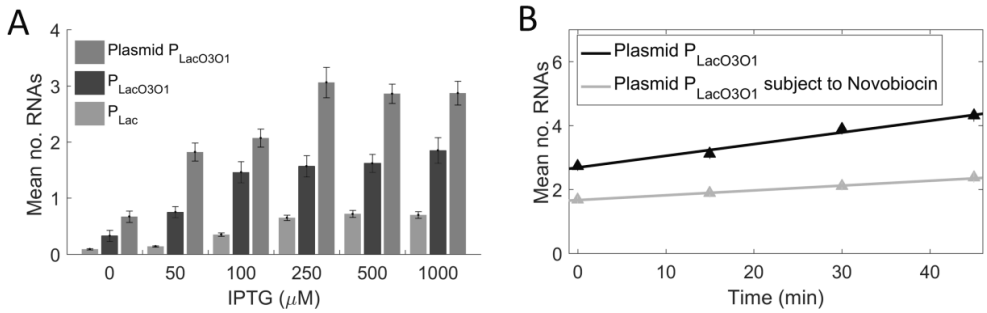


Figure S4. Mean number of RNAs produced by the target genes in individual cells, as measured by microscopy and RNA tagging by MS2-GFP. **(A)** Induction curves. Mean and Standard Error of the Mean (SEM) of the number of target RNA molecules in individual cells as a function of the induction strength. Results are shown for a chromosome-integrated gene controlled by the $P_{LacO3O1}$ promoter (blue bars), for a single-copy plasmid-borne gene controlled by the $P_{LacO3O1}$ promoter (grey bars) and for a chromosome-integrated gene controlled by the native P_{Lac} promoter (green bars), 1 hour after induction of the target gene. More than 100 cells were analyzed per condition. **(B)** Mean integer-valued RNA numbers produced over time in individual cells, each carrying the single-copy plasmid-borne gene under the control of $P_{LacO3O1}$, when subject to 100 μg/ml Novobiocin (grey triangles) and in the control condition (no Novobiocin, black triangles). For each time point, new cells were taken from the original culture. Best linear fits were calculated by WTLS (Krystek and Anton, 2007). The equation of the grey line is Y = (0.015 ± 0.004).X + (1.67 ± 0.11) and of the black line is Y = (0.037 ± 0.005).X + (2.69 ± 0.13). The errors are obtained by the standard error of the mean. Related to Figures 3 and 5. More than 100 cells were analyzed per condition.

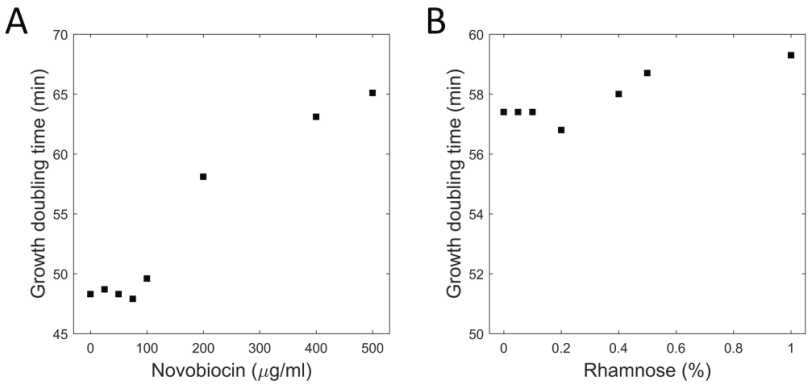


Figure S5**.** Cell growth doubling times. Doubling times versus the concentration of **(A)** Novobiocin and **(B)** Rhamnose. Doubling times were measured from the initial and final OD_{600} and, from the time interval in between (100 minutes). Related to Figures 3 and 4.

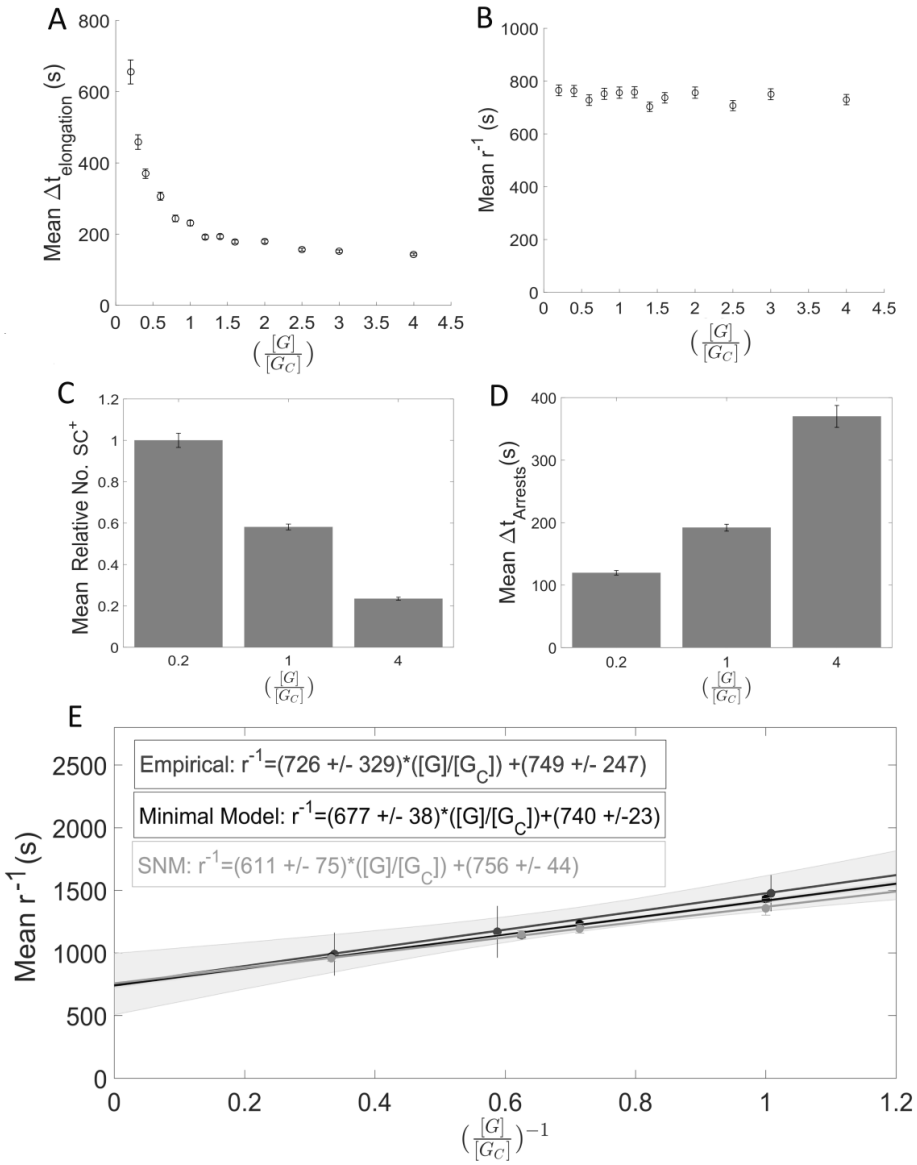


Figure S6. *In silico* results on the effects of Gyrase. **(A)** Mean $\Delta t_{elongation}$ (± SEM) as a function of Gyrase concentration (model in Table S2, with [G_c] being the concentration in the control condition). Relative Gyrase concentrations were set to [0.2, 0.3, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 2.0, 2.5, 3.0, 4.0]. **(B)** Mean Δt (± SEM) as a function of Gyrase (model in Table S2). **(C)** Mean relative SC^+ levels for different relative Gyrase concentrations [0.2, 1.0, 4.0]. **(D)** Mean time intervals (Δt) between consecutive arrests for different relative Gyrase concentrations [0.2, 1.0, 4.0]. As Gyrase concentration increases, the propensity of arrests during transcription elongation decreases. Related with reaction 6 in Table S2. **(E)** Effects of Gyrase overexpression on the dynamics of RNA production in the SN model (Table S2), the minimal model (Figure 1, main manuscript) and live cells (Figure 3,

main manuscript). In both the minimal and the SN model, the stochastic rate constants were set to k_1 = 0.0014 s^{-1}, k_{lock} = 0.0012 s^{-1}, k_{unlock} = 7x10^{-4} s^{-1}, k_{remove} = 1 s^{-1} and k_p=0.023. All rate constants were extracted or derived from empirical data (Supplementary Section VIII and Table S2). The reactions composing the SN model are shown in Table S2. The reactions of the minimal model are shown in Figure 1 and Section 3.1 of the main manuscript.

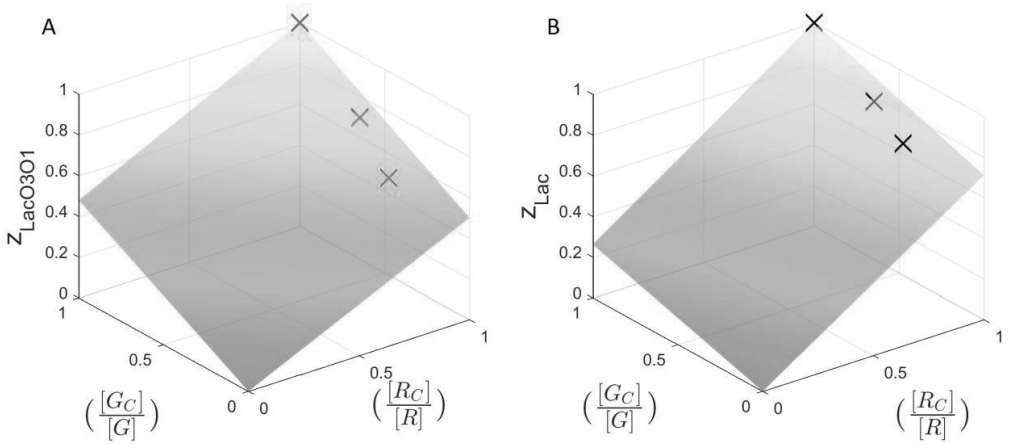


Figure S7. Z surfaces of the best fitting models. **(A)** Z surface for $P_{LacO3O1}$ (α = 0.04, β_1 = 0.5 and η = 0.48) as a function of $\frac{[R_C]}{[R]}$ and $\frac{[G_C]}{[G]}$. The red crosses mark the empirical data points for the $P_{LacO3O1}$ promoter. **(B)** Z surface for P_{Lac} (α = 0.04, β_2 = 0.71 and η = 0.48) as a function of $\frac{[R_C]}{[R]}$ and $\frac{[G_C]}{[G]}$. The black crosses mark the empirical data points for the P_{Lac} promoter. For both (A) and (B), the model fits the empirical data with a mean squared error of 0.0004. To estimate the goodness of fit of the surfaces, we calculated R^2 values. Both surfaces had R^2 > 0.95. Related to Figure 1 and Supplementary Section IV.

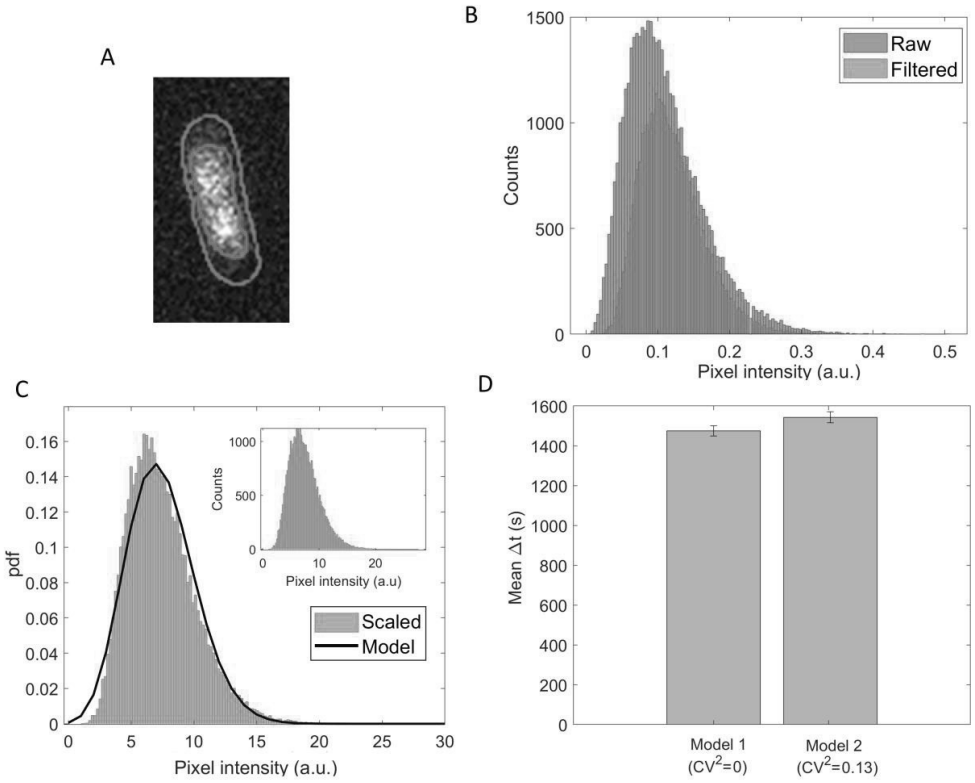


Figure S8. Microscopy measurements of pixel fluorescence intensities of *E. coli* cells expressing the gyrA gene endogenously tagged with a YFP coding sequence (data from 91 cells). **(A)** Example image of a cell with segmented borders (green line, based on phase contrast images) and segmented region with GyrA-YFP (red line). **(B)** Empirical pixel intensity distribution (in arbitrary units (a.u.)) prior ('Raw') and after applying a 2D Gaussian filtering ('Filtered'). The Raw distribution has a mean of 0.11 and a standard deviation of 0.054, while the Filtered distribution has a mean of 0.12 and a standard deviation of 0.044. **(C)** (inset) Empirical pixel intensity distribution multiplied by a constant (equal to the ratio between the mean and the variance of the filtered distribution in B). This distribution has a mean of 7.43 and a variance of 7.43 and can be well approximated by a Poisson (scaled) distribution. The scaled distribution is used to model the variability in Gyrase numbers over time (Table S7). Probability density function (PDF) of the scaled distribution and PDF of the *in silico* distribution of Gyrase numbers over time, fitted by a Poisson (scaled) distribution (black line). **(D)** Mean of the Δt distribution of intervals between RNA production events with (model 2) and without (model 1) variability in Gyrase numbers. Error bars correspond to the standard error of the mean. The imposed CV^2 is set (equal to) from the CV^2 of the scaled empirical distribution in (C). A two-sample t-test between the results from models 1 and 2 did not reject the null hypothesis that the two Δt distributions cannot be statistical distinguished (p-value > 0.05). Related to Supplementary Section III.3.

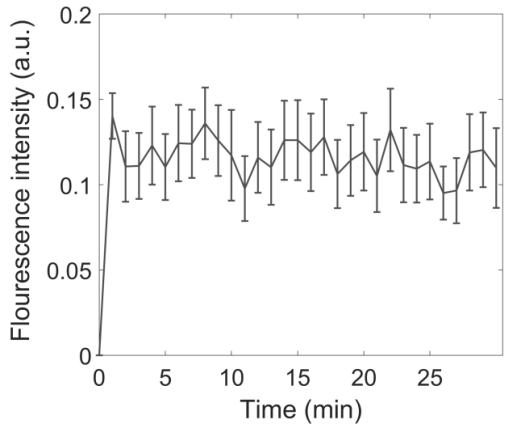


Figure S9. Mean fluorescence intensity of MS2-GFP tagged RNA molecules over time, since first appearing. 76 tagged RNAs were tracked, all from cells with only one tagged RNA. Also shown is the standard error of the mean (vertical bars). Related to Supplementary Section I.

SUPPLEMENTARY TABLES

Table S1. qPCR of the target genes under the control of $P_{LacO3O1}$ and P_{Lac}. Shown are the normalized average of the C_T value of the target gene by the respective mean C_T value of the respective reference condition ($\Delta C_{T(tested)}$), normalization of the $\Delta C_{T(tested)}$ value of the target gene to the $\Delta C_{T(tested)}$ value of the calibrator gene (0 % Rhamnose, $\Delta\Delta C_T$) and, finally, the fold change between the target and calibrator gene ($2^{-\Delta\Delta Ct}$). For each promoter, the samples were identically isolated, prepared and handled, and thus contain identical amounts of cDNA in each well. Related to Figures 3 and 5.

		mCherry			16SrRNA			$\Delta C_{T(tested)}$	$\Delta\Delta C_T$	$2^{-\Delta\Delta Ct}$
	Rhamnose (%)	1	2	3	1	2	3			
$LacO_3O_1$	**0.0**	26.63	26.49	26.26	13.52	13.33	13.28	13.08	0	1
	0.1	26.28	26.03	26.19	13.27	13.65	13.63	12.65	-0.43	0.74
	0.2	25.62	25.48	25.56	13.38	13.13	13.23	12.31	-0.78	0.58
Plasmid-Borne $LacO_3O_1$	**0.0**	23.17	23.18	23.04	7.79	7.81	7.36	15.48	0	1
	0.1	23.20	23.30	23.04	8.31	7.75	7.59	15.30	0.18	0.88
	0.2	23.18	23.18	23.28	8.34	7.72	7.78	15.27	-0.21	0.86
Lac	**0.0**	23.90	23.68	23.60	8.57	8.61	8.58	15.14	0	1
	0.1	22.71	22.76	23.27	8.15	8.09	7.94	14.86	-0.28	0.82
	0.2	23.39	23.23	22.87	8.60	8.34	8.36	14.73	-0.41	0.75

Table S2. Model of transcription at the single nucleotide (SN) level. Shown are the chemical reactions representing the various processes, and the stochastic rate constants (in s^{-1}) used to model transcription initiation, elongation and parallel and competing events at the nucleotide level (on example nucleotide *n*), and termination and RNA production. References from which rate constants were extracted are reported in column "Ref.". Pro stands for the promoter region, *RNAP* for the RNA polymerase, and *RNAP.Pro* for the promoter region when occupied by an *RNAP*. A_n, O_n and U_n stand for the n^{th} nucleotide when active, occupied, and unoccupied, respectively. Ranges of nucleotides are denoted as in *U*[start,end], which denotes a particular set of consecutive, unoccupied nucleotides from indexes start to end. $\mathrm{O}_{n_\mathrm{p}}$, $\mathrm{O}_{n_\mathrm{ar}}$ and $\mathrm{O}_{n_\mathrm{correcting}}$ are used to represent a paused, arrested, or error correcting *RNAP* at position n, respectively. On the template, each RNAP occupies (2Δ+1) nucleotides, where Δ = 12. These nucleotides cannot be occupied by any other *RNAP* at the same time.

Event	Reaction	Rate constant	Ref.
Reaction S2.1 Initiation	$\mathrm{Pro} + \mathrm{RNAP} \xrightarrow{k_1} \mathrm{RNAP.Pro}$	$k_1 = 0.0014$	Section V In accordance with (Lloyd-Price et al., 2016)
Reaction S2.2 Promoter locking	$\mathrm{Pro} + \mathrm{SC}^+ \xrightarrow{k_{lock}} \mathrm{Pro}_{\mathrm{lock}} + \mathrm{SC}^+$	$k_{lock} = 0.0012$	Section V Based on Empirical results and (Stracy et al., 2019)
Reaction S2.3 Promoter unlocking	$\mathrm{Pro}_{\mathrm{lock}} \xrightarrow{k_{unlock}} \mathrm{Pro}$	$k_{unlock} = 7\times10^{-4}$	(Chong et al., 2014)
Reaction S2.4 External Positive supercoils production	$\xrightarrow{k_p \cdot RNAP} \mathrm{SC}^+$	$k_p = 0.023$	Section V

Reaction S2.5 Promoter escape	$\mathrm{RNAP.Pro} + \mathrm{U}_{[1,(\Delta+1)]} \xrightarrow{k_m} \mathrm{O}_1 + \mathrm{Pro}$	k_m = 150	(Phroskin et al., 2010)
Reaction S2.6 Elongation	$\mathrm{A}_n + \mathrm{U}_{n+\Delta+1} \xrightarrow{k_m} \mathrm{O}_{n+1} + \mathrm{U}_{n-\Delta} + \left\{\mathrm{SC}^+, \text{if } n = 10 \times k, k \in \mathrm{N}\right\}$	k_m = 150	(Vogel and Jensen, 1994)
Reaction S2.7 Activation	$\mathrm{O}_{n+1} \xrightarrow{k_{act}} \mathrm{A}_{n+1}$	k_{act} = 150, n>10, k_{act} = 30, n<=10,	(Vogel and Jensen, 1994; Phroskin et al., 2010)
Reaction S2.8 Pausing	$O_n \underset{1/\tau_p}{\overset{k_{pause}}{\rightleftarrows}} O_{n_p}$	k_{pause} = 0.55 τ_p = 3	(Greive and von Hippel, 2005; Rajala et al., 2010; Landick, 2009)
Reaction S2.9 Pause release due to collision	$\mathrm{O}_{n_p} + \mathrm{A}_{n-2\Delta-1} \xrightarrow{k_{m1}} \mathrm{O}_n + \mathrm{A}_{n-2\Delta-1}$	k_{m1} = 120	(Epshtein and Nudler, 2003)
Reaction S2.10 Pause induced by collision	$\mathrm{O}_{n_p} + \mathrm{A}_{n-2\Delta-1} \xrightarrow{k_{m2}} \mathrm{O}_{n_p} + \mathrm{O}_{n-2\Delta-1_p}$	k_{m2} = 30	(Epshtein and Nudler, 2003)
Reaction S2.11 RNAP arrest due to supercoiling	$\mathrm{O}_n + \mathrm{SC}^+ \underset{1/d_{ar}}{\overset{k_{ar}}{\rightleftarrows}} \mathrm{O}_{n_{ar}} + \mathrm{SC}^+$	k_{ar} =0.03, d_{ar} = 100	(Fujita et al., 2016; Greive and von Hippel, 2005)

Reaction S2.12 Editing	$O_n \underset{1/d_{ed}}{\overset{k_{ed}}{\rightleftarrows}} O_{n_{correcting}}$	k_{ed} = 0.008, d_{ed} = 5	(Greive and von Hippel, 2005)
Reaction S2.13 Misincorporation	$A_{n_{last}} \xrightarrow{k_{mis}} RNA_{error} + RNAP + U_{n_{[last,last-\Delta]}}$	k_{mis}= 0.05	(Greive and von Hippel, 2005)
Reaction S2.14 Premature termination	$O_n \xrightarrow{k_{pre}} RNAP + U_{[(n-\Delta),(n+\Delta)]}$	k_{pre}= 0.00019	(Lewin, 2008)
Reaction S2.15 Pyrophosphorolysis	$O_n + U_{n-\Delta-1} \xrightarrow{k_{pyr}} O_{n-1} + U_{n+\Delta}$	k_{pyr} = 0.75	(Erie et al., 1993)
Reaction S2.16 Completion	$A_{n_{last}} \xrightarrow{k_f} R + RNAP + U_{n_{[last,last-\Delta]}}$	k_f = 2	(Greive et al., 2008)
Reaction S2.17 PSB removal	$Gyr + SC^{+} \xrightarrow{k_{remove}} Gyr$	k_{remove} = 1	(Stracy et al., 2019; Rovinskiy et al., 2012 ; Chong et al., 2014)

Table S3. Gyrase mRNA fold changes (measured by qPCR) for different concentrations of Rhamnose (0, 0.1 and, 0.2 %) and the corresponding Gyrase protein levels relative to the control condition (0 % Rhamnose), obtained from the calibration line (Figure 2B). Also shown are the fold changes in Gyrase protein levels (calculated using equation vii.1). Finally, it is shown the respective standard errors of the mean, calculated using the Delta method (11). Related to Figure 2.

Rhamnose (%)	**Gyrase mRNA fold change**	**Relative Gyrase Protein level (from calibration line Figure 2B)**	**Fold change in Protein levels (Equation vii.1)**
0.0 %	1.00	1± 0.09	1 .00 ± 0.15
0.1 %	1.71 ± 0.20	1.60 ± 0.21	1.71 ± 0.31
0.2 %	2.97 ± 0.32	2.67 ± 0.33	2.98 ± 0.53

Table S4. Average (μ) and standard deviation (σ) of the absolute nucleoid area (pixel) in the control condition (0 % Rhamnose) and when subject to Gyrase overexpression (0.2 % Rhamnose). A 2-sample student t-test for the null hypothesis that the two data sets are from the same distribution was not rejected (p-value > 0.01). The sample sizes are 224 and 237 cells for 0 % and 0.2 % Rhamnose conditions, respectively. Related to Section 3.3 in main manuscript.

	0 % Rhamnose	**0.2 % Rhamnose**
$\boldsymbol{\mu}$	421	449
$\boldsymbol{\sigma}$	105	130

Table S5. Dissection of the effects of RNAP overexpression on the transcription rate of $P_{LacO3O1}$ and P_{Lac} when overexpressing Gyrase. Shown are RNAP fold changes, ω, relative to the control (measured by Western blot) and the inverse of the fold change in RNA production rate, z (measured by qPCR). Also shown are the expected fold change in RNA production rate in the absence of change in [RNAP] (z', obtained from equation viii.1), and $(r')^{-1}$, the inverse of the RNA production rate of the target gene in the absence of indirect effects of Gyrase overexpression on the RNAP concentration. Related to Figures 3 and 5.

	$P_{LacO3O1}$		**P_{Lac}**	
	0.1 % Rhamnose	**0.2 % Rhamnose**	**0.1 % Rhamnose**	**0.2 % Rhamnose**
r_{ref}^{-1} (s)	1476	1476	2704	2704
r^{-1} (s)	1094	861	2223	2037
ω	1.05	1.12	1.05	1.12
y	1.71	2.98	1.71	2.98
z	0.74	0.58	0.82	0.75
z'	0.79	0.67	0.88	0.81
$(r')^{-1}$ (s)	1168	990	2380	2190

Table S6. Estimation of *OFF/ON* duty cycle ratios for $P_{LacO3O1}$ and for $P_{Lac,}$ based on model fitting, for different levels of Gyrase overexpression (0, 0.1 and, 0.2% Rhamnose). Shown are the total *OFF/ON* duty cycle ratio (OFF_{total} / ON), the *OFF/ON* duty cycle ratio due to neighbouring activity ($OFF_{neighboring} / ON$) and, the *OFF/ON* duty cycle ratio due to RNA polymerase activity (OFF_{self} / ON). Related to Figure 1 and Section 3.8 in main manuscript.

	$P_{LacO3O1}$			**P_{Lac}**		
Duty cycle	**0 %**	**0.1 %**	**0.2 %**	**0 %**	**0.1 %**	**0.2 %**
$OFF_{total} / ON = \frac{\left(\frac{k_p.RNAP}{k_1.RNAP} + \lambda\right).\frac{k_{lock}}{k_{unlock} \cdot k_{rem} \cdot G}}{\frac{1}{k_1.RNAP}}$	1.00	0.61	0.37	0.72	0.44	0.28
$OFF_{neighboring} / ON = \frac{k_{lock} \cdot k_p \cdot RNAP}{k_{unlock} \cdot k_{rem} \cdot G}$	0.04	0.02	0.01	0.04	0.02	0.01
$OFF_{self} / ON = \frac{\frac{k_{lock} \cdot \lambda}{k_{unlock} \cdot k_{rem} \cdot G}}{\frac{1}{k_1.RNAP}}$	0.96	0.59	0.36	0.68	0.42	0.27

Table S7. Reactions added to the minimal model to introduce variability in Gyrase numbers. Related with Figure S8. The empirical distribution in Figure S8C is modelled by reactions (S7.1) and (S7.2), provided accurate fitting of the rate constants k_{p_G} and k_{d_G} (their ratio should equal the mean of the empirical distribution). Meanwhile, k_{remove} in reaction (S7.3) was tuned so that 1 Gyrase resolves approximately 1 positive supercoil (SC^+) per second (Stracy et al., 2019) (related to Supplementary Section III.3).

Event	**Reaction**	**Rate constant (s^{-1})**
Gyrase production (S7.1)	$\xrightarrow{k_{p\_G}} Gyr$	k_{p_G} =7.43
Gyrase degradation (S7.2)	$Gyr \xrightarrow{k_{d\_G}}$	k_{d_G} =1
PSB removal (S7.3)	$Gyr + SC^+ \xrightarrow{k_{remove}} Gyr$	k_{remove} = 0.13

SUPPORTING REFERENCES

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

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

Casella, G., and Berger, R. L. (2002). *Statistical inference*. Thomson Learning.

Cheung, K. J., Badarinarayana, V., Selinger, D. W., Janse, D., & Church, G. M. (2003). A Microarray-Based Antibiotic Screen Identifies a Regulatory Role for Supercoiling in the Osmotic Stress Response of Escherichia coli. Genome Research, 13(2), 206–215. doi:10.1101/gr.401003

Chong, S., Chen, C., Ge, H., and Xie, X. S. (2014). Mechanism of Transcriptional Bursting in Bacteria. *Cell*, *158*(2), 314–326. doi:10.1016/j.cell.2014.05.038

Epshtein, V., Nudler, E. (2003) Cooperation between RNA polymerase molecules in transcription elongation. Science 300: 801-5. doi:10.1126/science.1083219

Erie, D.A., Hajiseyedjavadi, O., Young, M.C., von Hippel, P.H. (1993) Multiple RNA polymerase conformations and GreA: control of the fidelity of transcription. Science, 262: 867-873. doi: 10.1126/science.8235608

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

Gellert, M., O'Dea, M. H., Itoh, T., and Tomizawa, J. (1976). Novobiocin and coumermycin inhibit DNA supercoiling catalyzed by DNA gyrase. *Proc. Natl. Acad. Sci. U. S. A.*, *73*(12), 4474–4478. doi: 10.1073/pnas.73.12.4474

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4), 403–434. doi:10.1016/0021-9991(76)90041-3

Golding, I., Paulsson, J., Zawilski, S. M., and 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

Greive, S.J., von Hippel, P.H.(2005) Thinking quantitatively about transcriptional regulation. Nat Rev Mol Cell Biol, 6: 221-232. doi: 10.1038/nrm1588

Greive SJ, Weitzel SE, Goodarzi JP, Main LJ, Pasman Z, et al. (2008) Monitoring RNA transcription in

real time by using surface plasmon resonance. Proc. Natl. Acad. Sci. USA 105: 3315-20. doi: 10.1073/pnas.0712074105

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*(13), 1708–1709. doi:10.1093/bioinformatics/btt194

Häkkinen, A. and Ribeiro, A. S. (2015). Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data. *Bioinformatics*, *31*(1), 69–75. doi:10.1093/bioinformatics/btu592

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. BioRxiv, 842484. doi:10.1101/842484

Kandavalli, V. K., Tran, H., and Ribeiro, A. S. (2016). Effects of σ factor competition are promoter initiation kinetics dependent. *Biochim. Biophys. Acta - Gene Regul. Mech.*, *1859*(10), 1281–1288. doi:10.1016/j.bbagrm.2016.07.011

Kouzine, F., Gupta, A., Baranello, L., Wojtowicz, D., Ben-Aissa, K., Liu, J., Przytycka, T.M., Levens, D. (2013). Transcription-dependent dynamic supercoiling is a short-range genomic force. *Nat Struct Mol Biol* 20:396–403.doi:10.1038/nsmb.2517

Krystek, M. and Anton, M. (2007). A weighted total least-squares algorithm for fitting a straight line. *Meas. Sci. Technol.*, *18*(11), 3438–3442. doi:10.1088/0957-0233/18/11/025

Landick R. (2009) Transcriptional pausing without backtracking. *Proc Natl Acad Sci USA*, 106(22): 8797-8798. doi: 10.1073/pnas.0904373106

Lesne, A., Victor, J.-M., Bertrand, E., Basyuk, E., Barbi, M. (2018). The Role of Supercoiling in the Motor Activity of RNA Polymerases. In Methods in molecular biology (Clifton, N.J.) 1805, 215–232. doi:10.1007/978-1-4939-8556-2_11

Lewin B (2008) Genes IX, 256-299. Jones and Bartlett Publishers, USA

Lilley, D. M., & Higgins, C. F. (1991). Local DNA topology and gene expression: the case of the leu-500 promoter. Mol. Microbiol., 5(4), 779–783. doi: 10.1111/j.1365-2958.1991.tb00749.x

Lloyd-Price, J., Startceva, S., Kandavalli, V., Chandraseelan, J. G., Goncalves, N., Oliveira, S. M. D., A. Häkkinen, Ribeiro, A. S. (2016). Dissecting the stochastic transcription initiation process in live *Escherichia coli*. *DNA Res.*, *23*(3), 203–214. doi:10.1093/dnares/dsw009

Ma, J., Bai, L., and Wang, M. D. (2013). Transcription Under Torsion. Science., 340(6140), 1580–1583. doi:10.1126/science.1235441

Ma, J., & Wang, M. (2014). Interplay between DNA supercoiling and transcription elongation. Transcription, 5(3), e28636. doi:10.4161/trns.28636

Ma, J., Tan, C., Gao, X., Fulbright, R. M., Roberts, J. W., Wang, M. D., & Wang, M. D. (2019). Transcription factor regulation of RNA polymerase's torque generation capacity. *Proc.Nat. Aca.of Sci. U.S.A.*, 116(7), 2583–2588. doi:10.1073/pnas.1807031116

Mäkelä, J., Lloyd-Price, J., Yli-Harja, O., & Ribeiro, A. S. (2011). Stochastic sequence-level model of coupled transcription and translation in prokaryotes. BMC Bioinformatics, 12(1), 121. doi:10.1186/1471-2105-12-121

Mäkelä, J., Kandavalli, V., and Ribeiro, A. S. (2017). Rate-limiting steps in transcription dictate sensitivity to variability in cellular components. *Sci. Rep.*, *7*(1), 10588. doi:10.1038/s41598-017-11257-2

Mitarai, N., Dodd, I. B., Crooks, M. T., and Sneppen, K. (2008). The Generation of Promoter-Mediated Transcriptional Noise in Bacteria. *PLoS Comput. Biol.*, *4*(7), e1000109. doi:10.1371/journal.pcbi.1000109

Muthukrishnan, A.-B., Martikainen, A., Neeli-Venkata, R., and Ribeiro, A. S. (2014). *In vivo* transcription kinetics of a synthetic gene uninvolved in stress-response pathways in stressed *Escherichia coli* cells. *PLoS One*, *9*(9), e109005. doi:10.1371/journal.pone.0109005

Naughton, C., Avlonitis, N., Corless, S., Prendergast, J. G., Mati, I. K., Eijk, P. P., Cockroft,S., Bradley.M., Ylstra, B., Gilbert, N. (2013). Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. Nat Struct Mol Biol 20(3), 387–395. doi:10.1038/nsmb.2509

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. and Ribeiro,A.S. (2019). Chromosome and plasmid-borne PLacO3O1 promoters differ in sensitivity to critically low temperatures. *Sci. Rep.*, 9(1), 4486, doi:10.1038/s41598-019-39618-z

Oliveira, S.M.D., Neeli-Venkata, R., Goncalves, N.S.M., Santinha, J.A., Martins, L., Tran, H., Mäkelä, J., Gupta, A., Barandas, M., Häkkinen, A., Lloyd-Price, J., Fonseca, J.M. and Ribeiro, A.S. (2016). Increased cytoplasm viscosity hampers aggregate polar segregation in *Escherichia coli*. *Mol. Microbiol.*, 99(4), 686–699, doi:10.1111/mmi.13257

Phroskin S, Rachid Rahmouni A, Mironov A, Nudler E (2010) Cooperation between translating ribosomes and RNA polymerase in transcription elongation. Science, 328(5977): 504-508

Rajala, T., Häkkinen, A., Healy, S., Yli-Harja, O., & Ribeiro, A. S. (2010). Effects of transcriptional pausing on gene expression dynamics. PLoS Computational Biology, 6(3), e1000704. doi:10.1371/journal.pcbi.1000704

Revyakin, A., Ebright, R. H., & Strick, T. R. (2004). Promoter unwinding and promoter clearance by RNA polymerase: Detection by single-molecule DNA nanomanipulation. Proceedings of the National Academy of Sciences, 101(14), 4776–4780. doi:10.1073/pnas.0307241101

Ribeiro, A. S., Smolander, O.-P., Rajala, T., Häkkinen, A., & Yli-Harja, O. (2009). Delayed Stochastic Model of Transcription at the Single Nucleotide Level. Journal of Computational Biology, 16(4), 539–553. doi: 10.1089/cmb.2008.0153

Rhee, K. Y., Opel, M., Ito, E., Hung, S. p, Arfin, S. M., & Hatfield, G. W. (1999). 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, 96(25), 14294–14299. doi: 10.1073/pnas.96.25.14294

Rovinskiy, N., Agbleke, A. A., Chesnokova, O., Pang, Z., and Higgins, N. P. (2012). Rates of Gyrase Supercoiling and Transcription Elongation Control Supercoil Density in a Bacterial Chromosome. *PLoS Genet.*, *8*(8), e1002845. doi:10.1371/journal.pgen.1002845

Santinha, J., Martins, L., Häkkinen, A., Lloyd-Price, J., Oliveira, S. M. D., Gupta, A., Annila, T., Mora, A., Ribeiro, A.S. and Fonseca, J. R. (2016). *iCellFusion: Tool for Fusion and Analysis of Live-Cell Images from Time-Lapse Multimodal Microscopy*. doi:10.4018/978-1-4666-8811-7.ch004

Stracy, M., Lesterlin, C., Garza de Leon, F., Uphoff, S., Zawadzki, P., and Kapanidis, A. N. (2015). Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid. *Proc. Natl. Acad. Sci.*, *112*(32), E4390–E4399. doi:10.1073/pnas.1507592112

Stracy, M., Wollman, A. J. M., Kaja, E., Gapinski, J., Lee, J.-E., Leek, V. A., McKie, S.J., Mitchenall, L.A., Maxwell,A., Sherratt, D.J., Leake, M,C., and Zawadzki, P. (2019). Single-molecule imaging of DNA gyrase activity in living *Escherichia coli*. *Nucleic Acids Res.*, *47*(1), 210–220. doi:10.1093/nar/gky1143

Startceva, S., Kandavalli, V. K., Visa, A., and Ribeiro, A. S. (2019). Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression. *Biochim. Biophys. Acta - Gene Regul. Mech.*, *1862*(2), 119–128. doi:10.1016/j.bbagrm.2018.12.005

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

Teves, S. S., Henikoff, S. (2014). Transcription-generated torsional stress destabilizes nucleosomes.

Nat Struct Mol Biol 21(1), 88–94. doi:10.1038/nsmb.2723

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

Travers, A., & Muskhelishvili, G. (2005). DNA supercoiling — a global transcriptional regulator for enterobacterial growth? Nature Reviews Microbiology, 3(2), 157–169. doi:10.1038/nrmicro1088

Vogel, U. and Jensen, K. F. (1994). The RNA chain elongation rate in *Escherichia coli* depends on the growth rate. *J. Bacteriol.*, *176*(10), 2807–2813. doi:10.1128/jb.176.10.2807-2813.1994

Weinstein-Fischer, D., Elgrably-weiss, M., & Altuvia, S. (2002). *Escherichia coli* response to hydrogen peroxide: a role for DNA supercoiling, Topoisomerase I and Fis. Molecular Microbiology, 35(6), 1413–1420. doi:10.1046/j.1365-2958.2000.01805.x

Wheeler, A. (2017). Digital Microscopy. In Standard and Super-Resolution Bioimaging Data Analysis (eds A. Wheeler and R. Henriques). doi:10.1002/9781119096948.ch1

Zimmer, C., Häkkinen, A., and Ribeiro, A. S. (2016). Estimation of kinetic parameters of transcription from temporal single-RNA measurements. *Math. Biosci.*, *271*, 146–153. doi:10.1016/j.mbs.2015.10.001

PUBLICATION III

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, Andre S. Ribeiro

Journal of Microbiological Methods, 166, 105745, 2019. *Equal contributions

https://doi.org/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

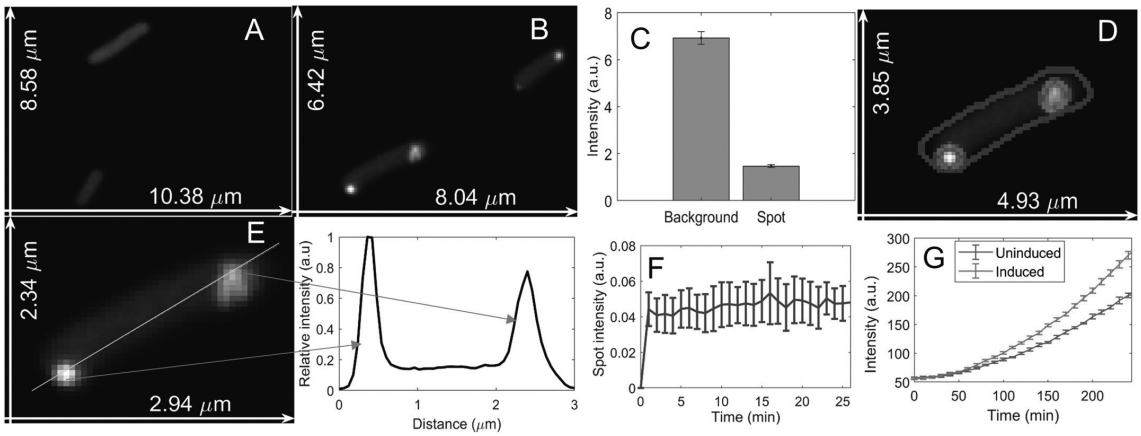


Fig. 1. A) Example microscopy image of cells carrying the reporter gene coding for MS2d-GFP, prior to the production of target RNAs. The cells are visible due to carrying a large amount of MS2d-GFP proteins; B) Example microscopy image of cells carrying the reporter gene coding for MS2d-GFP, after the 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:

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.

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	$\langle X \rangle$	$\langle (X - \langle X \rangle)^2 \rangle$	$\sqrt{\langle (X - \langle X \rangle)^2 \rangle}$	$\langle (X - \langle X \rangle)^3 \rangle$	$\frac{\langle (X - \langle X \rangle)^3 \rangle}{Sd^3}$

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 $P_{Lac/ara\text{-}1}$ (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.

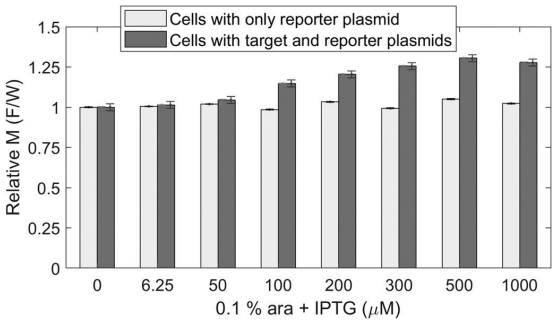


Fig. 2. (Light grey bars) Mean F/W values of the strain carrying only the multi-copy plasmid carrying the reporter gene, at various IPTG concentrations (x-axis), relative to its mean F/W value at the 0 μM IPTG condition. Its black error bars are the standard error of mean, estimated from the cells in 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 $P_{Lac/ara\text{-}1}$, 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.

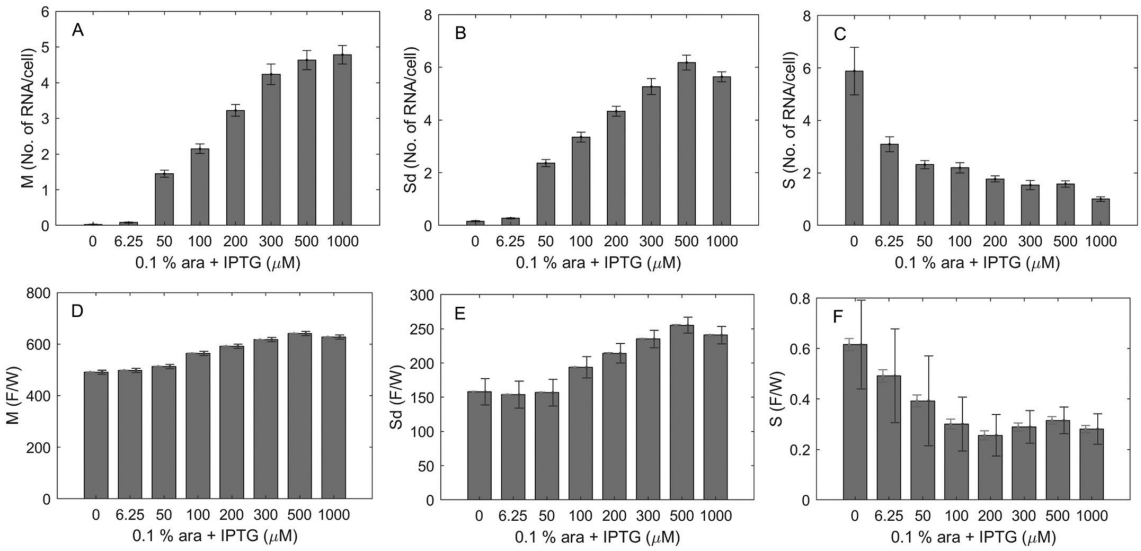


Fig. 3. (A) Mean, M, (B) Standard deviation, Sd, and (C) Skewness, S, of single-cell distributions of integer-valued RNA numbers obtained by microscopy, as a function of IPTG concentration (x-axis). The standard error of M, Sd and S of RNA numbers was estimated as described in Methods, section 2.7. (D) 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 $\sigma = 0$ 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

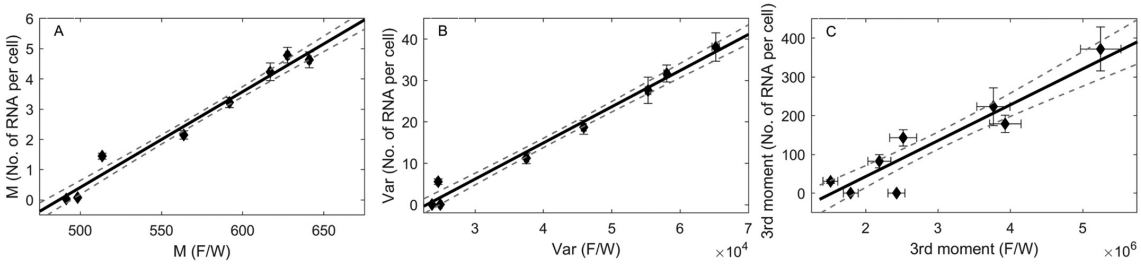


Fig. 4. Scatter plots between (A) Mean (M), (B) Variance (Var), (C) 3rd Moment of the single-cell distributions of F/W values obtained by flow cytometry against M, Var and 3rd Moment of the single-cell distributions of RNA numbers in individual cells obtained by Microscopy for various induction strengths (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.

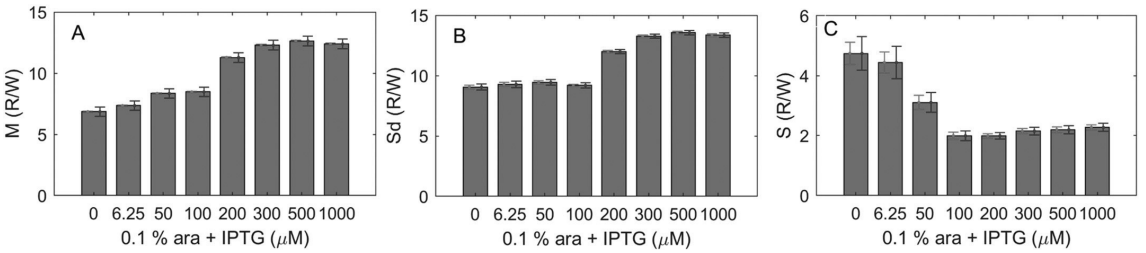


Fig. 5. (A) Mean, M, (B) Standard deviation, Sd, and (C) Skewness, S, of single-cell distributions of R/W values, when subject to various IPTG concentrations. The red error bars are standard errors (Methods, section 2.7). The blue error bars are the standard error of the statistics after adding empirical 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).

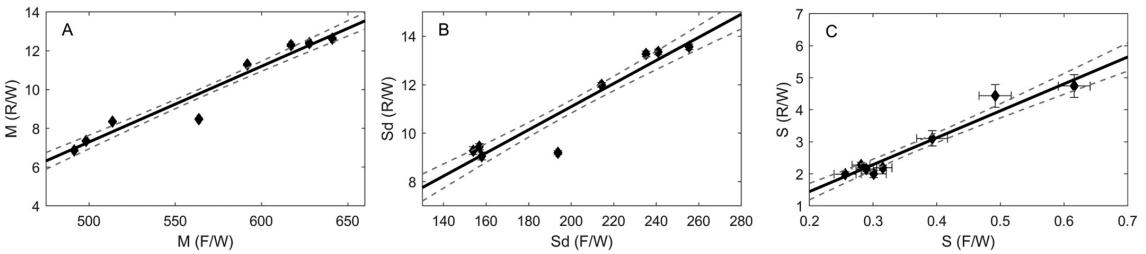


Fig. 6. Scatter plots between (A) Mean (M), (B) Standard deviation (Sd), (C) Skewness (S) of the single-cell distributions of F/W values against M, Sd and S of the single-cell distributions of R/W values for various induction strengths, differing in IPTG concentration (0, 6.25, 50, 100, 200, 300, 500, 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.)

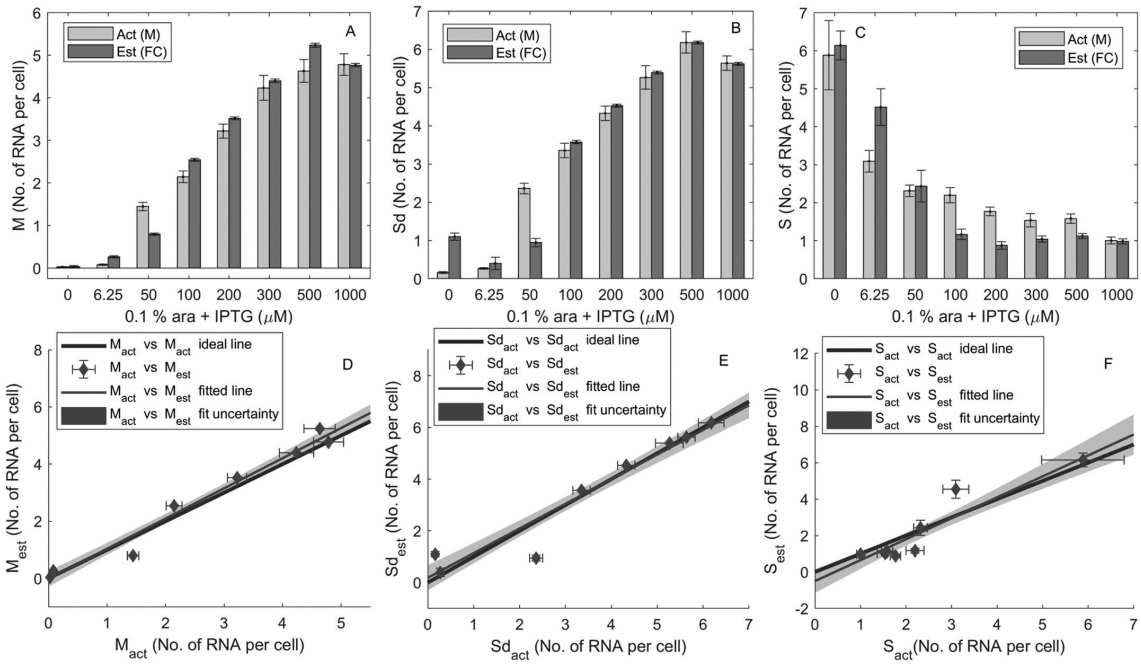


Fig. 7. (A) Mean single-cell RNA numbers estimated from Flow cytometry data, using microscopy data (Mean RNA numbers per cell) in the minimum (0 μM IPTG) and maximum induction (1000 μM IPTG) conditions for the calibration. (B) Standard deviation of single-cell RNA numbers estimated from Flow cytometry 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 $\sigma = 400$ 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:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} + \varepsilon \quad \text{(S1)}$$

Here, $\beta_0, \beta_1, \ldots, \beta_k$ 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:

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

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:

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

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

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

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:

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

From (S13):

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

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:

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

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

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:

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

1.5.3 Third moment

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

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

$$\mu_3(N_R) = \frac{\mu_3(F_T)}{BS^3 \cdot I^3} - \frac{\mu_3(N_B)}{BS^3} \quad \text{(S19)}$$

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 ($\mu_3(F_T)$) 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:

$$skew(N_R) = \frac{\mu_3(N_R)}{(Var(N_R))^{\frac{3}{2}}} \quad \text{(S20)}$$

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):

$$SE(M) = \sqrt{\left(SE\left(M_{target}\right)\right)^2 + \left(SE\left(M_{control}\right)\right)^2} \quad \text{(S21)}$$

where $SE\left(M_{target}\right)$ 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:

$$SE(Sd) = \frac{1}{2} \cdot \sqrt{\frac{\left(SE\left(Var_{target}\right)\right)^2 + \left(SE\left(Var_{control}\right)\right)^2}{Var_{target}}} \quad \text{(S22)}$$

Here, $SE\left(Var_{target}\right)$ 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_{target} 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:

$$SE(S) \approx \frac{\sqrt{\left(SE\left(\mu_3^{target}\right)\right)^2 + \left(SE\left(\mu_3^{control}\right)\right)^2}}{(Sd_{target} + SE(Sd))^3} \quad \text{(S23)}$$

where $SE\left(\mu_3^{target}\right)$ 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_{target} 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.

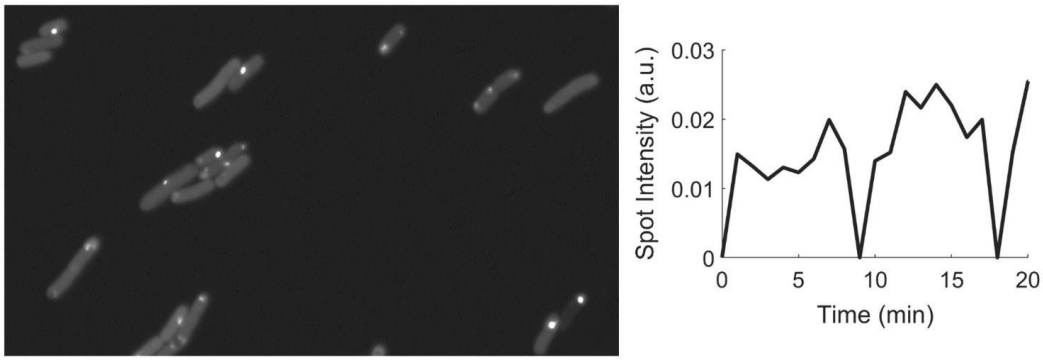
1.7 Estimation of single-cell RNA number statistics and standard error from F/W values from flow cytometry

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.

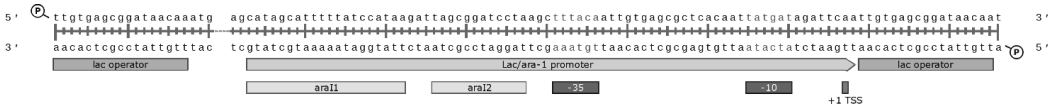
Similarly, to quantify the standard deviation (Sd) of RNA numbers per cell as measured by flow cytometry, we used the calibration line to map the variance of RNA numbers obtained by microscopy to the corresponding variance in F/W values obtained by flow cytometry. Next, we used bootstrapping to obtain 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.

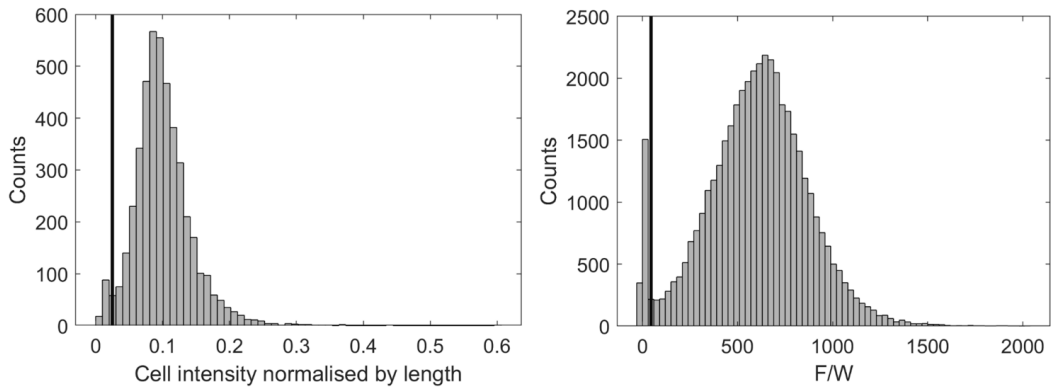
SUPPLEMENTARY FIGURES



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).

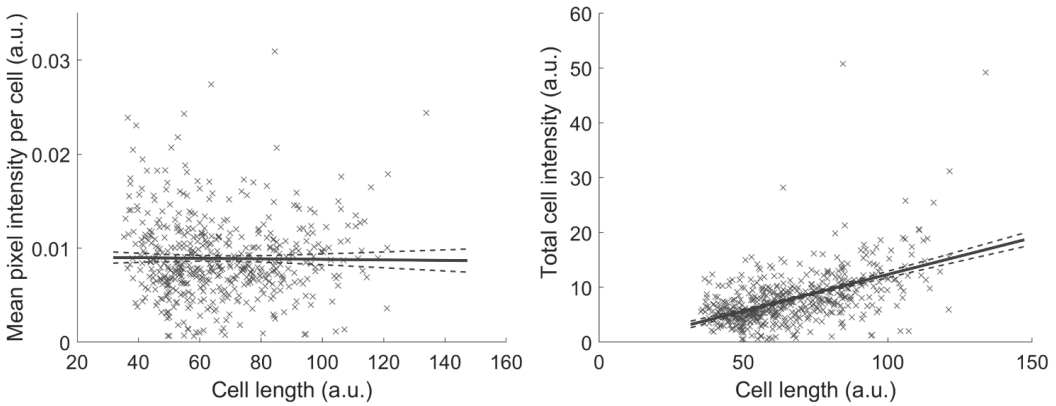


Supplementary Figure S2. Related to section 2.2 in main manuscript. Schematic representation of the target promoter sequence (marked in yellow) of the gene of interest. The -35 and -10 promoter elements are shown in light blue, while the transcription start site (+1 TSS) is marked by a small pink box.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

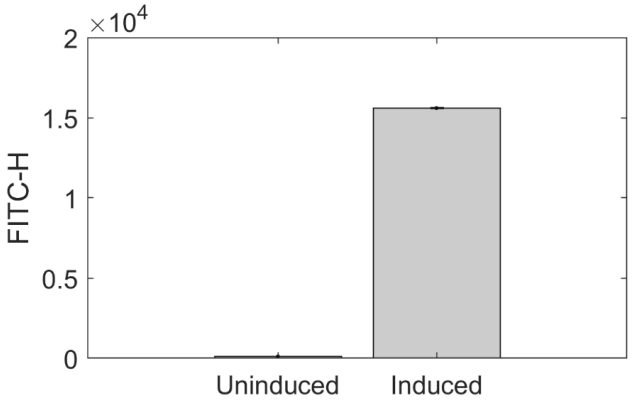


Supplementary Figure S3. Related to section 2.6 in main manuscript. (Left): Distribution of single-cell green fluorescence intensity normalized by cell length, as obtained by microscopy. We (manually) applied a minimum threshold (black vertical

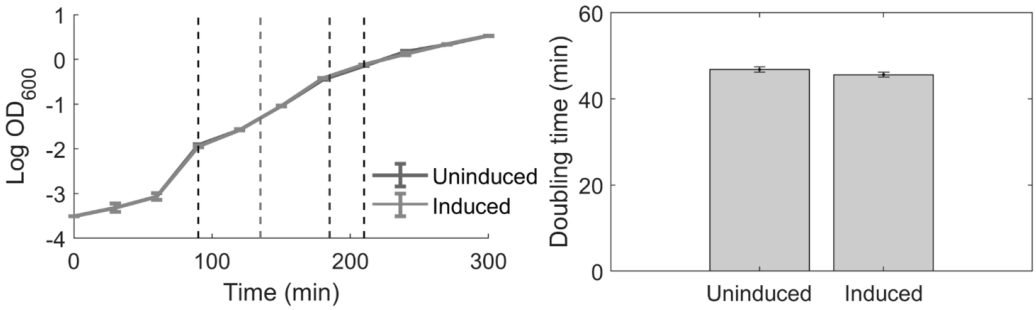
line, located at 0.025 in the x axis) to remove from the dataset the few cells not expressing sufficient MS2d-GFP for detecting target RNAs. Approximately 4500 cells were analyzed. (Right): Distribution of single-cell FITC-H values (F) divided by pulse Width (W), obtained by flow-cytometry (more than 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).



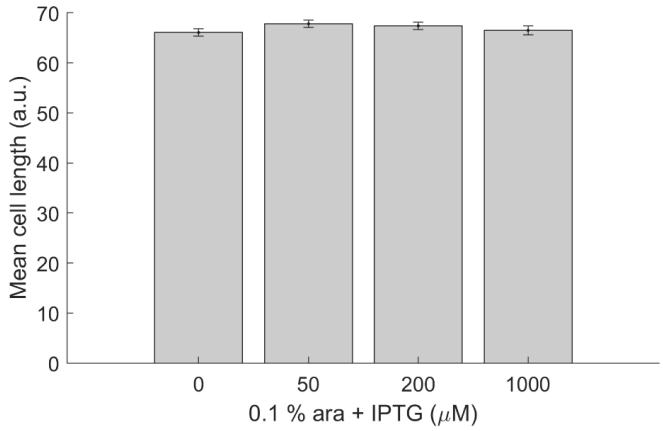
Supplementary Figure S4. Related to section 2.6 in main manuscript. (Left) Scatter plot of the single-cell fluorescence intensity (as estimated by the mean pixel intensity per cell) against the length of the major cell axis, as measured by microscopy. Approximately 500 cells were analyzed. The solid 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.



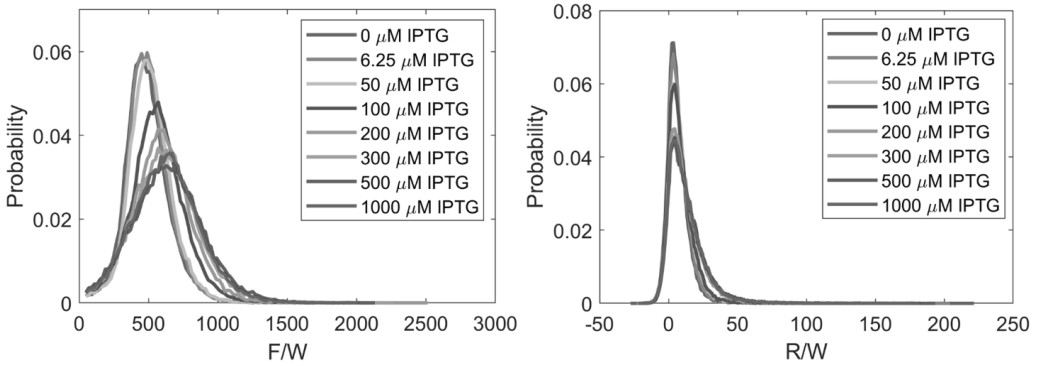
Supplementary Figure S5. Related to section 3.1 in main manuscript. Mean fluorescence intensity of cells, as measured by the FITC-H channel of the flow-cytometer. Measurements of the reporter gene expressing MS2d-GFP under the control of $P_{LtetO-1}$ were performed when induced (100 ng/µl aTc) and when 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.



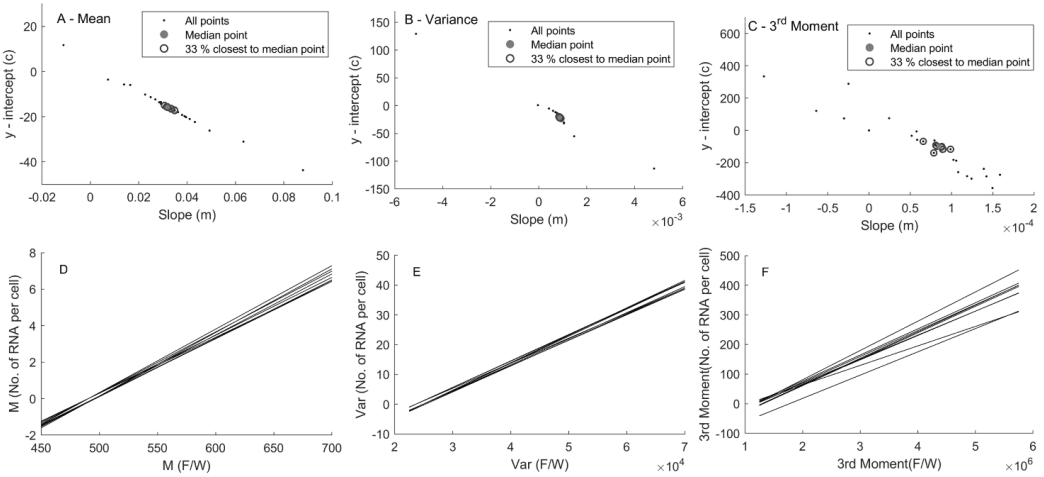
Supplementary Figure S6. Related to Figure 1 in main manuscript. (Left) Optical density (OD) curves of cell populations with the target plasmid in the presence ('induced', 1000 µM) and absence ('uninduced', 0 µM) of IPTG. Visibly, the two lines overlap. From a -80 °C glycerol stock, cells were streaked 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.

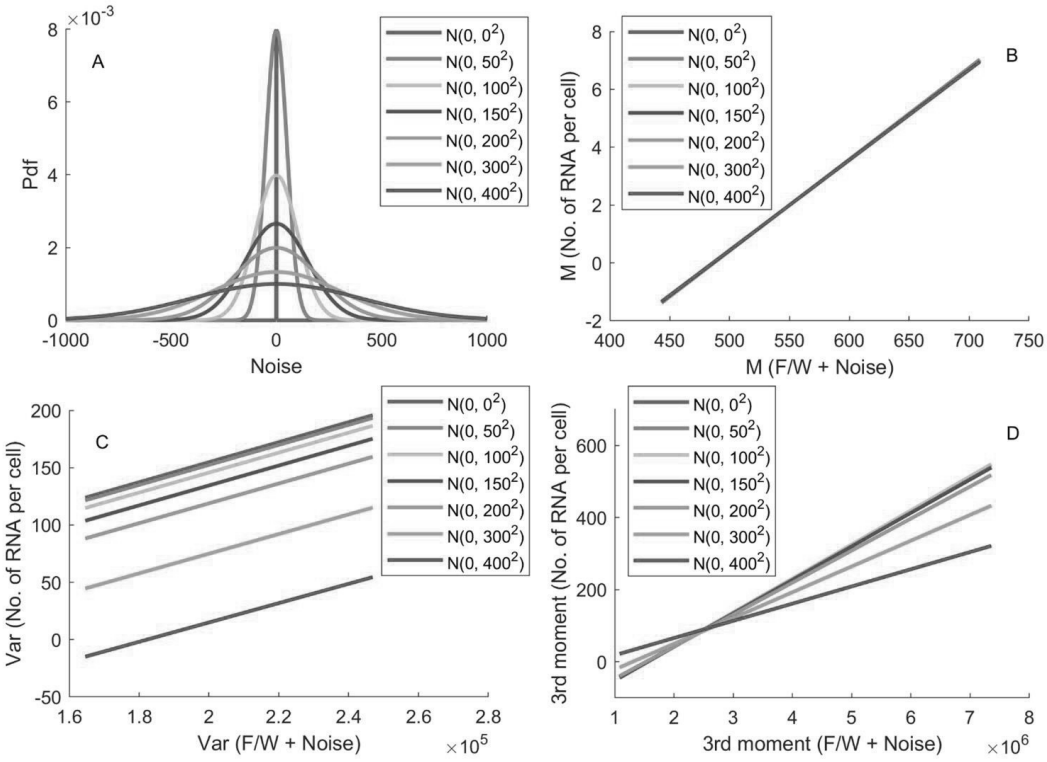


Supplementary Figure S8. Related to Figures 3D-F and 5 in main manuscript. (Left) Probability of single-cell F/W values (bin width = 20 units) for each condition differing in IPTG concentration, after gating (see Section 2.6 in main manuscript). In addition, we removed the 0.01% or less cells with highest 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.

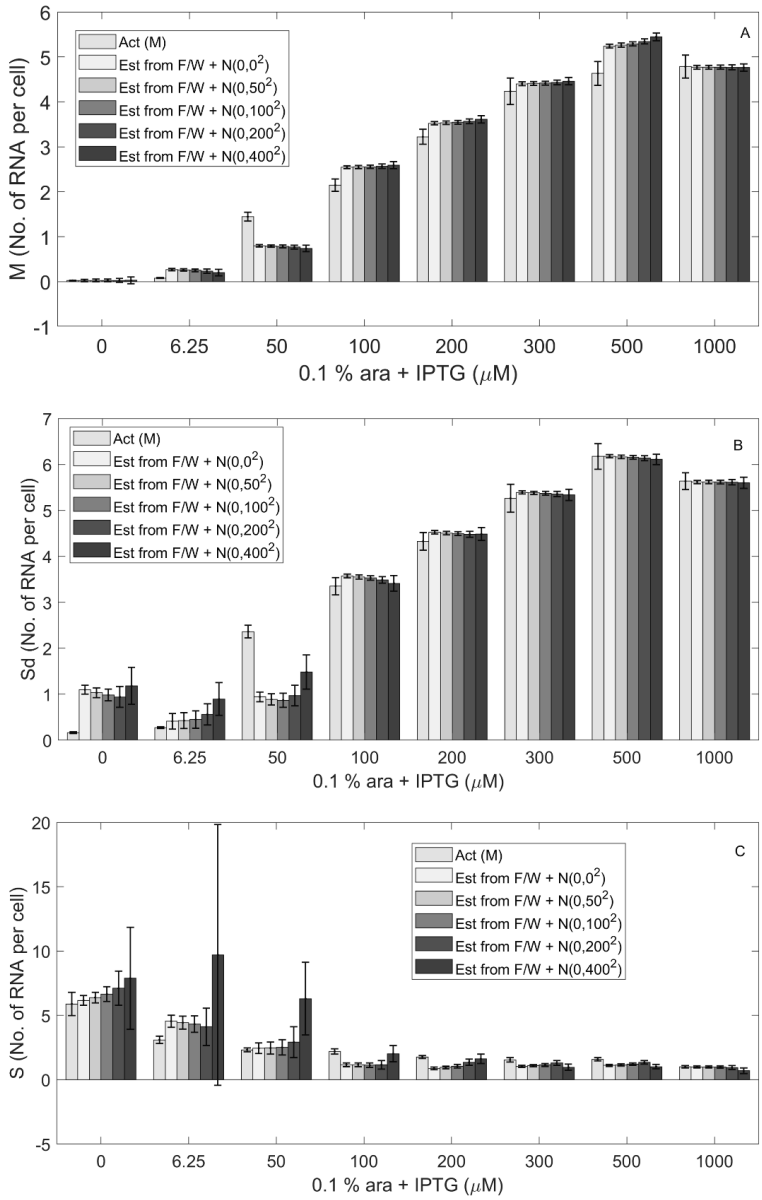


Supplementary Figure S9. Related to Figure 7 in the main manuscript. Slope vs intercept of the calibration lines between single-cell distributions of F/W and single cell distribution of RNA numbers for (A) Mean, (B) Variance and (C) 3rd moment. The small black dots correspond to each of all possible calibration lines (data from Figures 4A-C in main manuscript). The

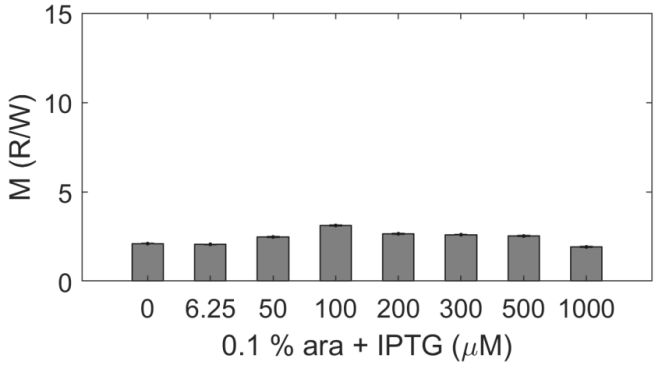
red points are the median point, whose coordinates are the median of the slopes versus the median of the intercepts of all possible calibration lines, respectively. The blue circles identify the black dots that are closer to the median point (the 33 % closest are encircled). Calibration lines of the 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 IV

Analytical kinetic model of native tandem promoters in E. coli.

Chauhan V*, Bahrudeen MNM*, Palma CSD, Baptista ISC, Almeida BLB, Dash S, Kandavalli V, Ribeiro AS

PLoS Computational Biology, 18, e1009824, 2022. *Equal contributions.

https://doi.org/ 10.1371/journal.pcbi.1009824

PLOS COMPUTATIONAL BIOLOGY

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

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'

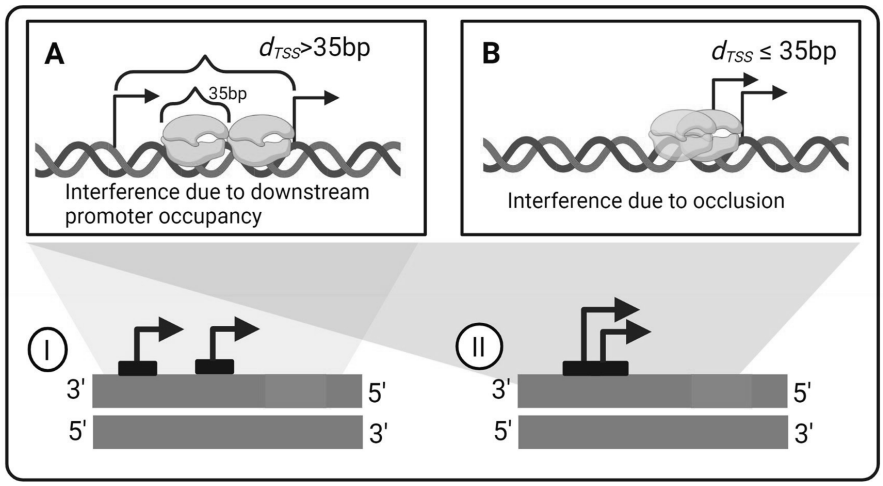


Fig 1. Interference between tandem promoters with different arrangements relative to each other. (A) Interference by an RNAP occupying the downstream promoter on the activity of the elongating RNAP from upstream promoter. The TSSs need to be at least 36 bp apart (the length occupied by an RNAP when in 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 $\leq$ 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

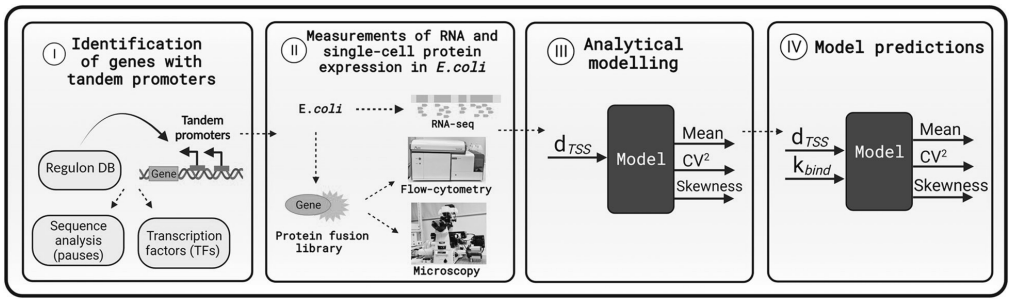


Fig 2. Workflow. (I) We identified genes controlled by tandem promoters in Regulon DB. (II) Next, we measured the single-cell protein levels of those genes with arrangements I and II that are tagged in the YFP strain library [28]. We also measured the mean RNA fold changes of these genes over time (S1 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

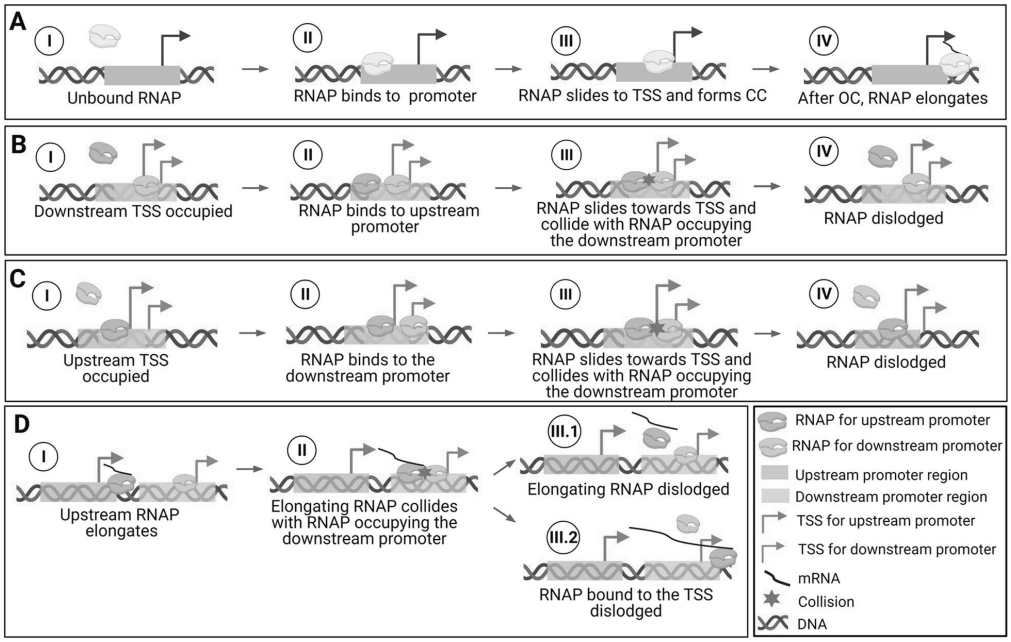


Fig 3. Events leading to transcriptional interference between tandem promoters. (**A**) Sequence of events in transcription in isolated promoters. A similar set of events occurs in tandem promoters, if only one RNAP interacts with them at any given time. (**B / C**) Interference due to the occlusion 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:

$$k_r = \left(\frac{k_{bind} \cdot [R] \times k_{after} \cdot (1 - \omega_d \cdot f)}{k_{occlusion} + k_{bind} \cdot [R] + k_{unbind} + k_{after}} + \frac{k_{bind} \cdot [R] \times k_{after}}{k_{occlusion} + k_{occupy} + k_{bind} \cdot [R] + k_{unbind} + k_{after}} \right) \tag{9}$$

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):

$$\log_{10}(CV_P^2) = \log_{10}(C_1) - \log_{10}(M_P), \qquad \text{with} \quad C_1 = \frac{k_p}{k_{pd} + k_{rd}} \tag{10}$$

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:

$$\log_{10}(S_P) = \log_{10}(C_2) - \frac{1}{2} \cdot \log_{10}(M_P), \qquad \text{with} \quad C_2 = 2\sqrt{C_1} - \frac{1}{\sqrt{C_1}} \tag{11}$$

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

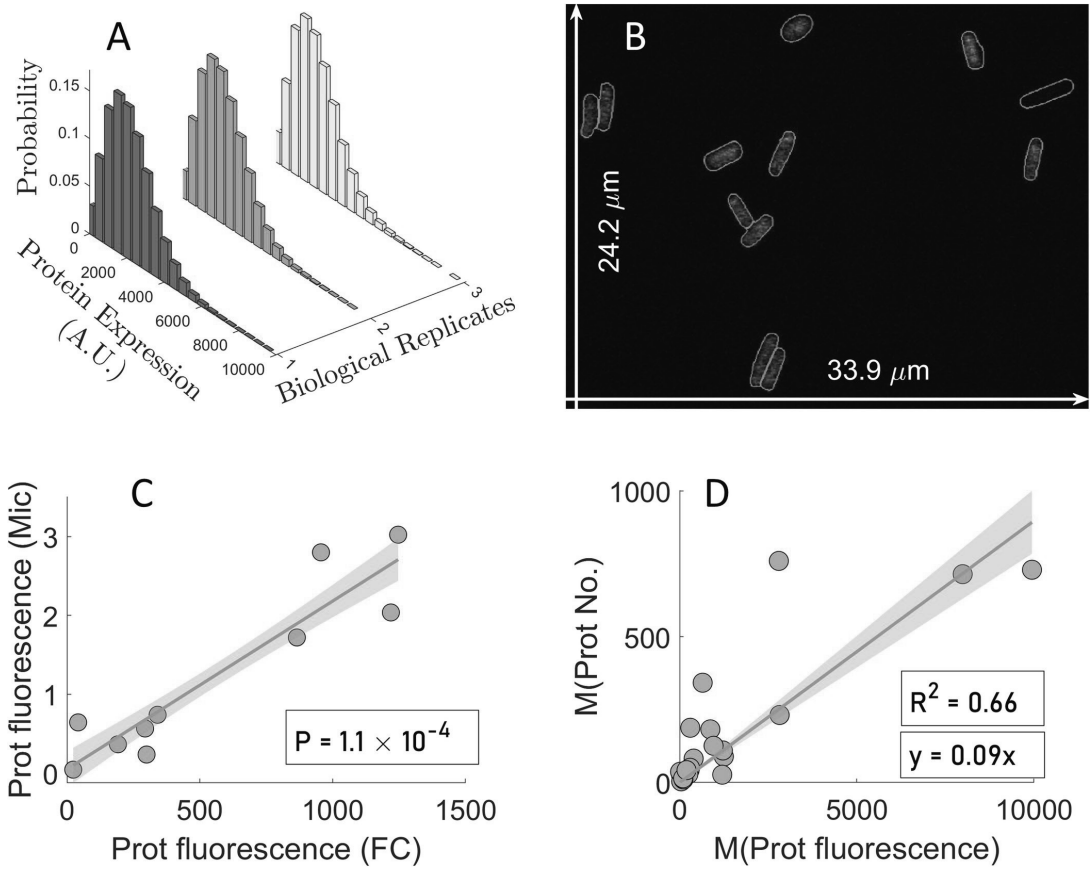


Fig 4. Single cell protein numbers by microscopy and flow-cytometry. (A) Example single-cell distributions (3 biological replicates) of fluorescence (in arbitrary units) of cells with a YFP tagged gene controlled by a pair of tandem promoters obtained by flow-cytometry, 'FC'. (B) Example confocal microscopy 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).

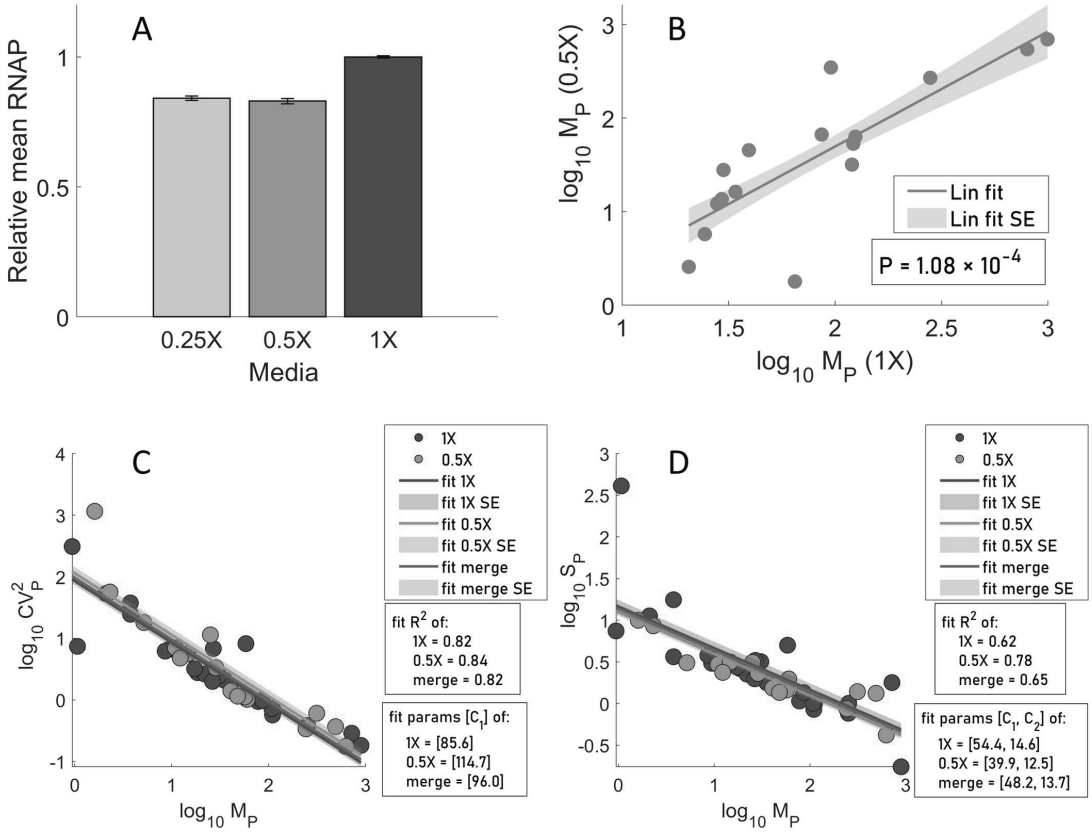


Fig 5. Relative RNAP concentrations along with the relationships between the moments of the single cell distributions of protein numbers. (A) Relative RNAP levels measured by flow-cytometry (Section 'flow-cytometry and data analysis' in the S1 Appendix) in three media. (B) Scatter plot between M_P in 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 $k^{\max} = k^{\max}_{occupy} = k^{\max}_{oci}$, 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_P^2 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_P^2 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_P^2, 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

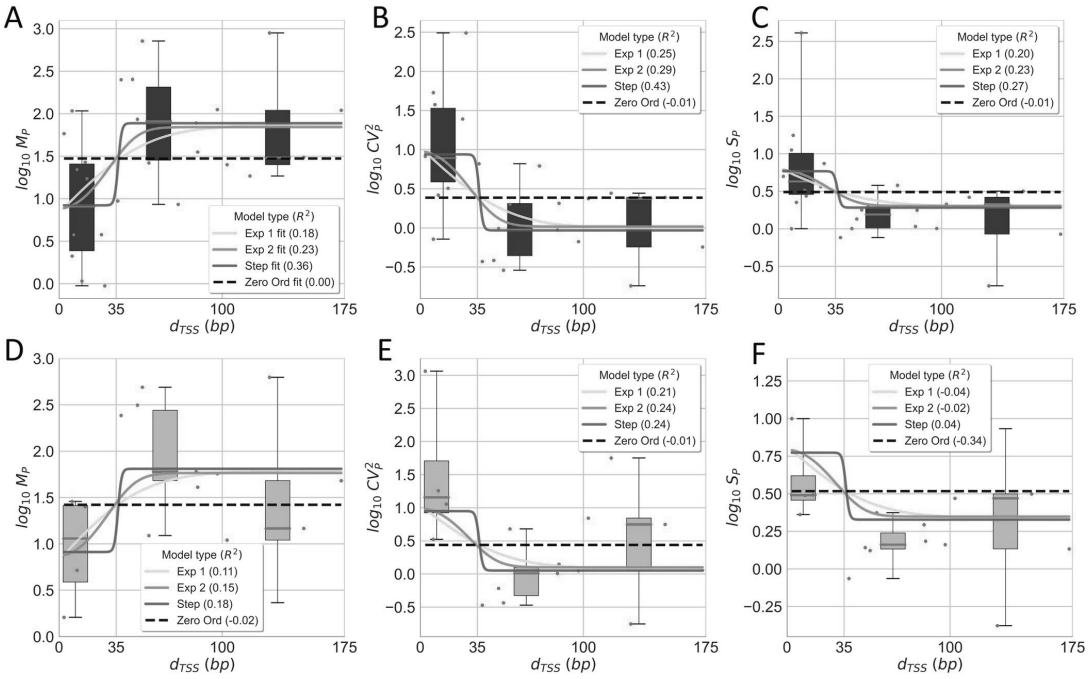


Fig 6. Empirical data and analytical model of how d_{TSS} influences the single-cell protein numbers of genes controlled by tandem promoters. (A) Mean, **(B)** CV^2, and **(C)** S of single protein numbers in the **1X** media as a function of d_{TSS}. **(D)**, **(E),** and **(F)** show the same for the **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

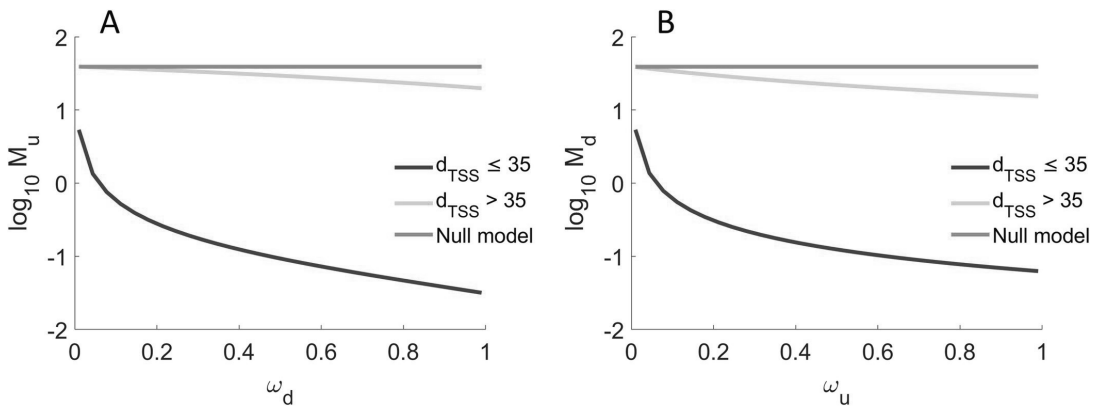


Fig 7. Mean protein numbers produced as a function of other promoter's occupancy. M_P of the single-cell distribution of the number of proteins produced (**A**) by the upstream promoter alone, and (**B**) by the downstream promoter alone. Results are shown as a function of the fraction of times that 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).

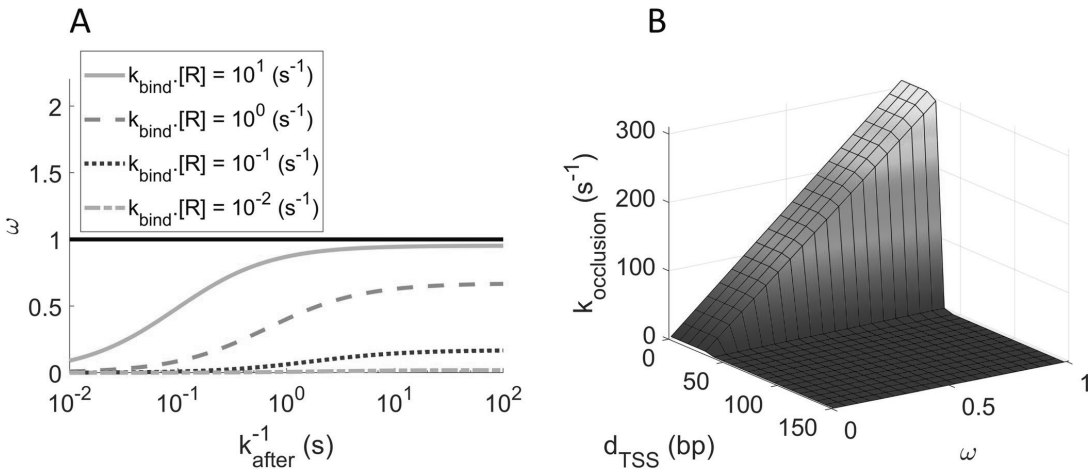


Fig 8. Promoter occupancy ω estimated for the step model. (A) ω as a function of the rate constant for a *free* RNAP to bind to the *unoccupied* promoter (k_{bind}·[R]) and of the time for that RNAP to start elongation after commitment to OC, k^{-1}_{after}. The horizontal black line at $\omega = 1$, 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^2_P, 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

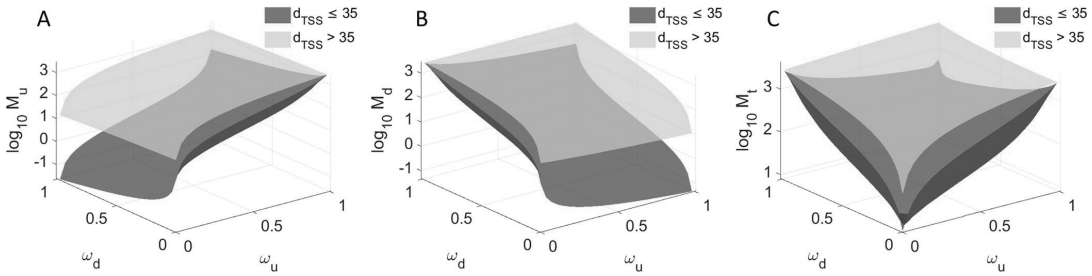


Fig 9. Mean protein expression as a function of both promoters' occupancy. Expected mean protein numbers due to the activity of: (**A**) the upstream promoter alone, (**B**) the downstream promoter alone, and (**C**) both promoters. M_P is shown as a function of the fraction of times that the upstream ($0 \leq \omega_u \leq 1$) 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:

$$CV^2(g) = \frac{Var(g)}{M(g)^2} \quad (3)$$

Finally, the third moment of protein fluorescence and the skewness after background subtraction are given by:

$$\mu_3(g) = \mu_3(T) - \mu_3(bg) \quad (4)$$

$$S(g) = \frac{\mu_3(g)}{Var(g)^{\frac{3}{2}}} \quad (5)$$

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:

$$P_{free} \underset{k_{unbind}}{\overset{k_{bind}\cdot[R]}{\rightleftarrows}} P_{occupied} \xrightarrow{k_{after}} P_{free} + R_{elong} \quad (6)$$

$$R_{elong} \xrightarrow{k_{elong}} RNA \quad (7)$$

At steady state $P_{occupied}$ is:

$$\frac{dP_{occupied}}{dt} = P_{free} \times k_{bind} \cdot [R] - P_{occupied} \times \left(k_{unbind} + k_{after}\right) = 0 \quad (8)$$

$$P_{free} = P_{occupied} \cdot \frac{\left(k_{unbind} + k_{after}\right)}{k_{bind} \cdot [R]} \quad (9)$$

Since necessarily:

$$P_{free} + P_{occupied} = 1 \quad (10)$$

From equations 9 and 10:

$$P_{occupied} \cdot \left(1 + \frac{k_{unbind} + k_{after}}{k_{bind} \cdot [R]}\right) = 1 \quad (11)$$

$$P_{occupied} = \frac{k_{bind} \cdot [R]}{k_{bind} \cdot [R] + k_{unbind} + k_{after}} \quad (12)$$

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:

$$\frac{dR_{elong}}{dt} = P_{occupied} \times k_{after} - R_{elong} \times k_{elong} = 0 \quad (13)$$

$$R_{elong} = \frac{P_{occupied} \times k_{after}}{k_{elong}} \quad (14)$$

From equations 12 and 14:

$$R_{elong} = \frac{k_{bind} \cdot [R]}{k_{bind} \cdot [R] + k_{unbind} + k_{after}} \times \frac{k_{after}}{k_{elong}} \quad (15)$$

At steady state, the mean *RNA* numbers, M_{RNA}, is:

$$\frac{dM_{RNA}}{dt} = R_{elong} \times k_{elong} - M_{RNA} \times k_{rd} = 0 \quad (16)$$

From equations 15 and 16s:

$$M_{RNA} = \frac{k_{bind} \cdot [R]}{k_{bind} \cdot [R] + k_{unbind} + k_{after}} \times \frac{k_{after}}{k_{elong}} \times \frac{k_{elong}}{k_{rd}} \quad \text{(S7)}$$

$$M_{RNA} = \frac{k_{bind} \cdot [R]}{k_{bind} \cdot [R] + k_{unbind} + k_{after}} \times \frac{k_{after}}{k_{rd}} \quad (18)$$

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:

$$\frac{d(RNA)}{dt} = R^u_{elong} \times k^u_{elong} \cdot (1 - \omega_d \cdot f) - RNA \times k_{rd} = 0 \quad (19)$$

From this and equation 6b in the main manuscript:

$$RNA = \frac{k^u_{bind} \cdot [R]}{k^{u/d}_{occlusion} + k^u_{bind} \cdot [R] + k^u_{unbind} + k^u_{after}} \times \frac{k^u_{after} \cdot (1 - \omega_d \cdot f)}{k_{rd}} \quad (20)$$

Meanwhile, for the downstream promoter, from reactions (2a1), (2a2), and (2a3) in the main manuscript, at steady state:

$$\frac{d(RNA)}{dt} = R^d_{elong} \times k^d_{elong} - RNA \times k_{rd} = 0 \quad (21)$$

$$RNA = \frac{k^d_{bind} \cdot [R]}{k^{d/u}_{occlusion} + k^d_{occupy} + k^d_{bind} \cdot [R] + k^d_{unbind} + k^d_{after}} \times \frac{k^d_{after}}{k_{rd}} \quad (22)$$

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:

$$M_{RNA}=\left(\frac{k_{bind}^{u}\cdot[R]\times k_{after}^{u}\cdot(1-\omega_d\cdot f)}{k_{occlusion}^{u/d}+k_{bind}^{u}\cdot[R]+k_{unbind}^{u}+k_{after}^{u}}+\frac{k_{bind}^{d}\cdot[R]\times k_{after}^{d}}{k_{occlusion}^{d/u}+k_{occupy}+k_{bind}^{d}\cdot[R]+k_{unbind}^{d}+k_{after}^{d}}\right)\cdot\frac{1}{k_{rd}} \quad (23)$$

Thus, the mean protein numbers is:

$$M_P=M_{RNA}\cdot\frac{k_p}{k_{pd}} \quad (24)$$

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:

$$M_P=\left(\frac{k_{bind}\cdot[R]\times k_{after}\cdot(1-\omega_d\cdot f)}{k_{occlusion}+k_{bind}\cdot[R]+k_{unbind}+k_{after}}+\frac{k_{bind}\cdot[R]\times k_{after}}{k_{occlusion}+k_{occupy}+k_{bind}\cdot[R]+k_{unbind}+k_{after}}\right)\cdot\frac{k_p}{k_{rd}\cdot k_{pd}} \quad (25)$$

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:

$$k_r=\left(\frac{k_{bind}\cdot[R]\times k_{after}\cdot(1-\omega_d\cdot f)}{k_{occlusion}+k_{bind}\cdot[R]+k_{unbind}+k_{after}}+\frac{k_{bind}\cdot[R]\times k_{after}}{k_{occlusion}+k_{occupy}+k_{bind}\cdot[R]+k_{unbind}+k_{after}}\right) \quad (26)$$

Thus, from equation 25 and 26:

$$M_P=\frac{k_r\cdot k_p}{k_{rd}\cdot k_{pd}} \quad (27)$$

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

$$C_1 = \frac{M_P}{M_{RNA}} \cdot \frac{\frac{1}{\tau_p}}{\frac{1}{\tau_p} + \frac{1}{\tau_{RNA}}} \tag{36}$$

$\tau_p = \frac{1}{k_{pd}}$ and $\tau_{RNA} = \frac{1}{k_{rd}}$ 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 $k_{bind} \cdot [R]$ 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

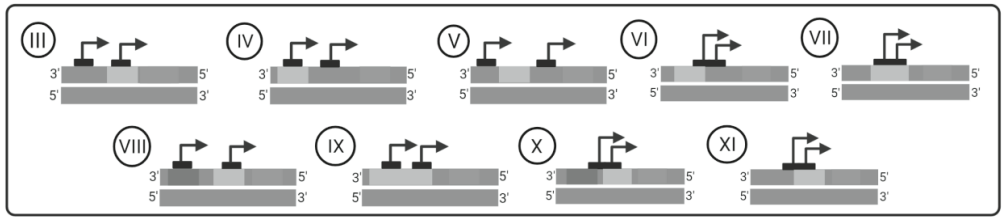


Fig A. Other arrangements of tandem promoters in *E. coli.* Unlike the arrangements I and II in Fig 1 in the main manuscript, the arrangements here (III-XI) allow for overlaps with or in between other gene(s). The red, green, and blue rectangles are DNA regions coding for RNA. These arrangements are not considered in this study. Figure created with BioRender.com.

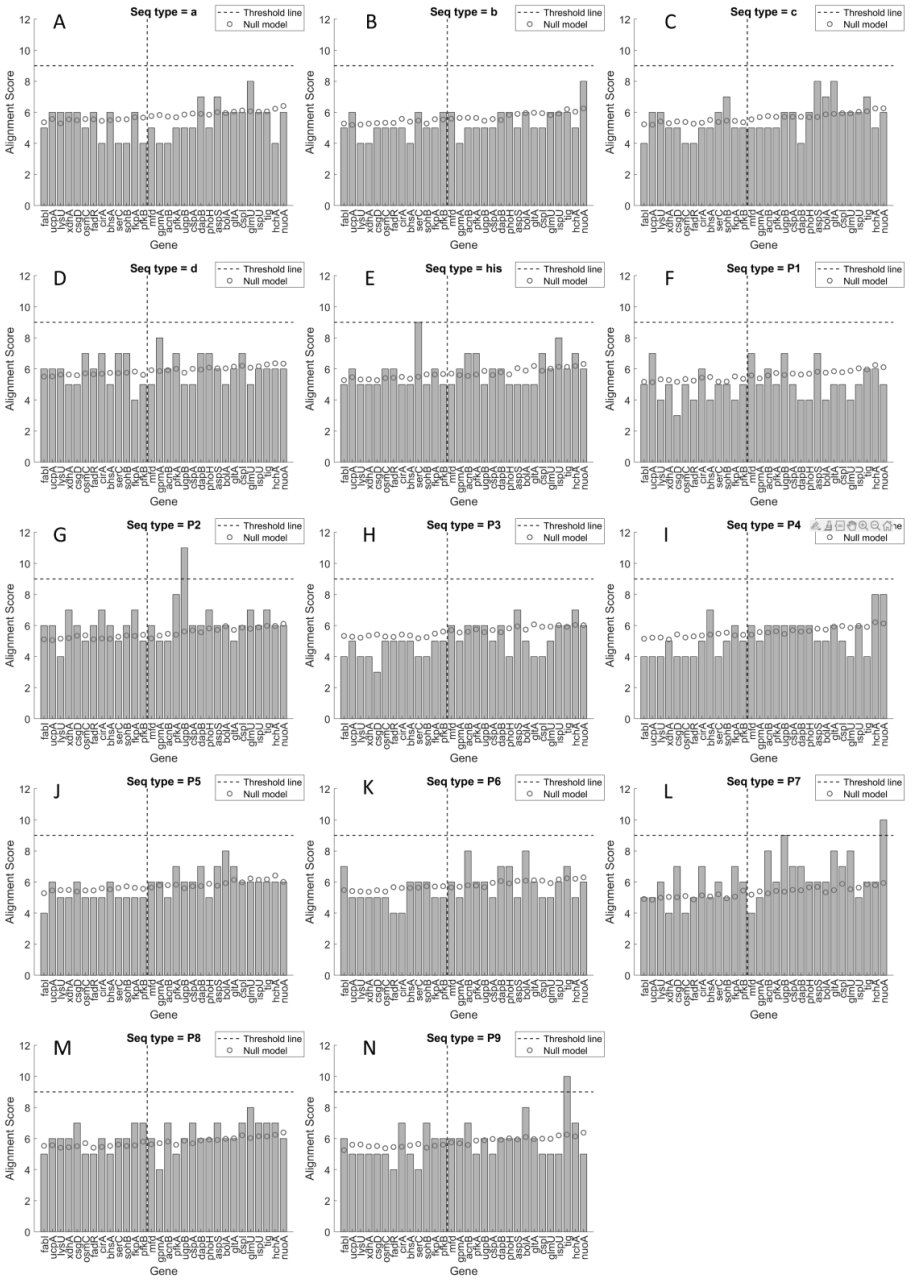


Fig B. Local alignment scores. Local alignment scores between known pause sequences and the sequences in between the tandem promoter regions (grey bars). Also shown by red circles are the alignment scores between each pause sequence and randomly generated sequences with the same d_{TSS} as the natural 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.

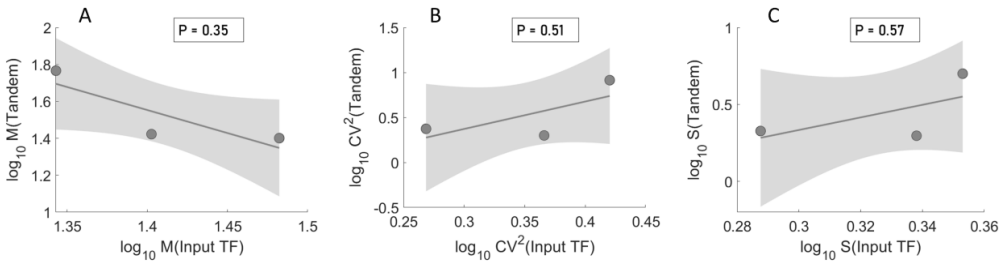


Fig C. Correlation of the moments of the single-cell protein numbers between genes and their input TFs. Scatter plots between the moments of the single-cell protein numbers (in $\log_{10}$ scale) of genes regulated by tandem promoters ('Tandem') and their input TFs. (A) Mean, (B) CV^2, and (C) Skewness. 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.

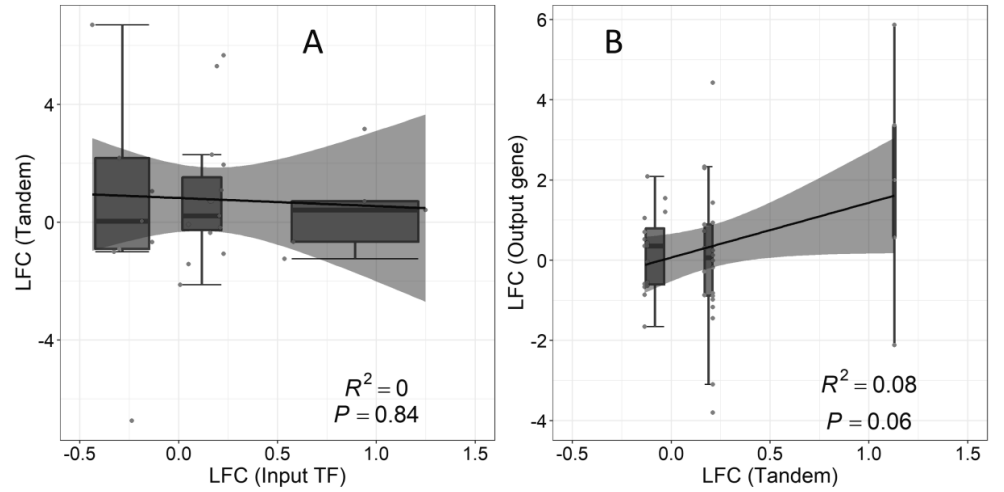


Fig D. Correlation of RNA fold changes of genes and their input TFs. Correlation plots between the LFCs of the RNA numbers of genes controlled by tandem promoters with their input and output genes. (A) LFCs (from 0 to 20 min) of 29 genes expressing input TFs plotted against the corresponding LFCs (from 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.

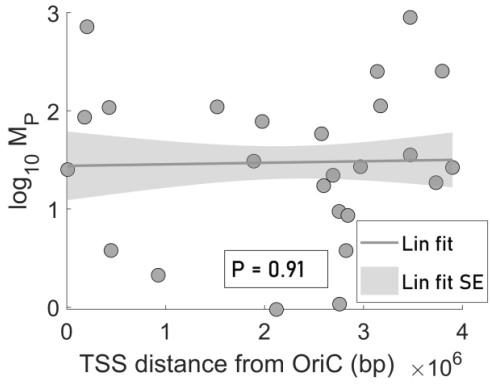


Fig E. Relationship between expression levels of the genes controlled by tandem promoters and the distance in nucleotides (bp) from the upstream promoter and the Oric region in the DNA. Data from 25 genes for the 1X condition. Also shown in a linear fit and the corresponding 1 standard error of the fit (shadow area). The p-value, *P*, is the probability that the slope of the line equals 0.

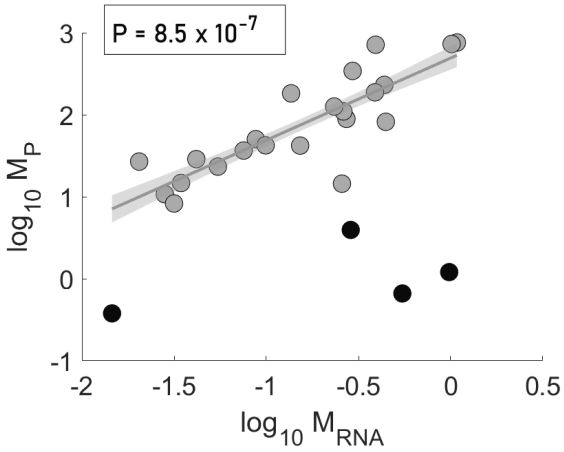


Fig F. Correlation plot between the mean single-cell RNA levels (M_{RNA}) and the mean single-cell protein numbers (M_P). Both data are obtained from Ref. [28] in main manuscript and are processed to include only genes controlled by tandem promoters (classes I and II, Table H in S3 Appendix). The line 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.

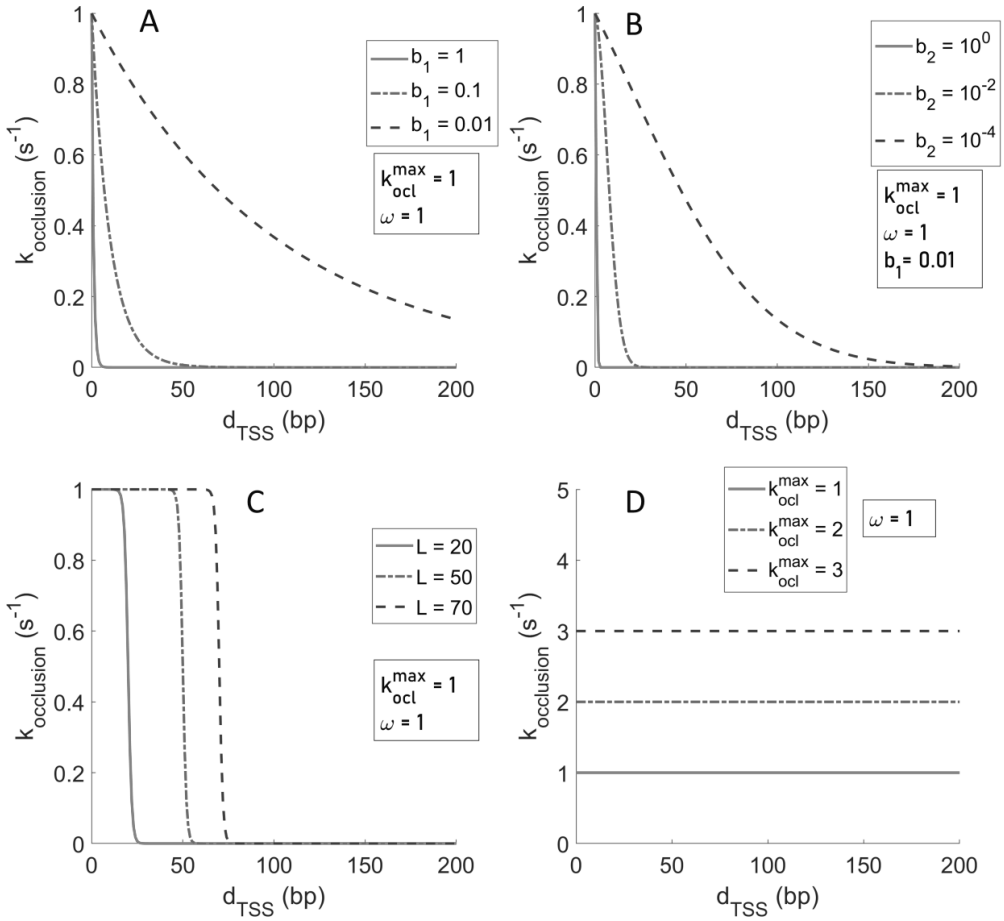


Fig G. Models of transcription interference. Models of transcription interference between RNAPs in tandem promoters as a function of the d_{TSS} between them. (A) 'Exponential 1' as a function for different values of 'b_1'. (B) 'Exponential 2' as a function at different values of 'b_2. (C) Continuous '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.

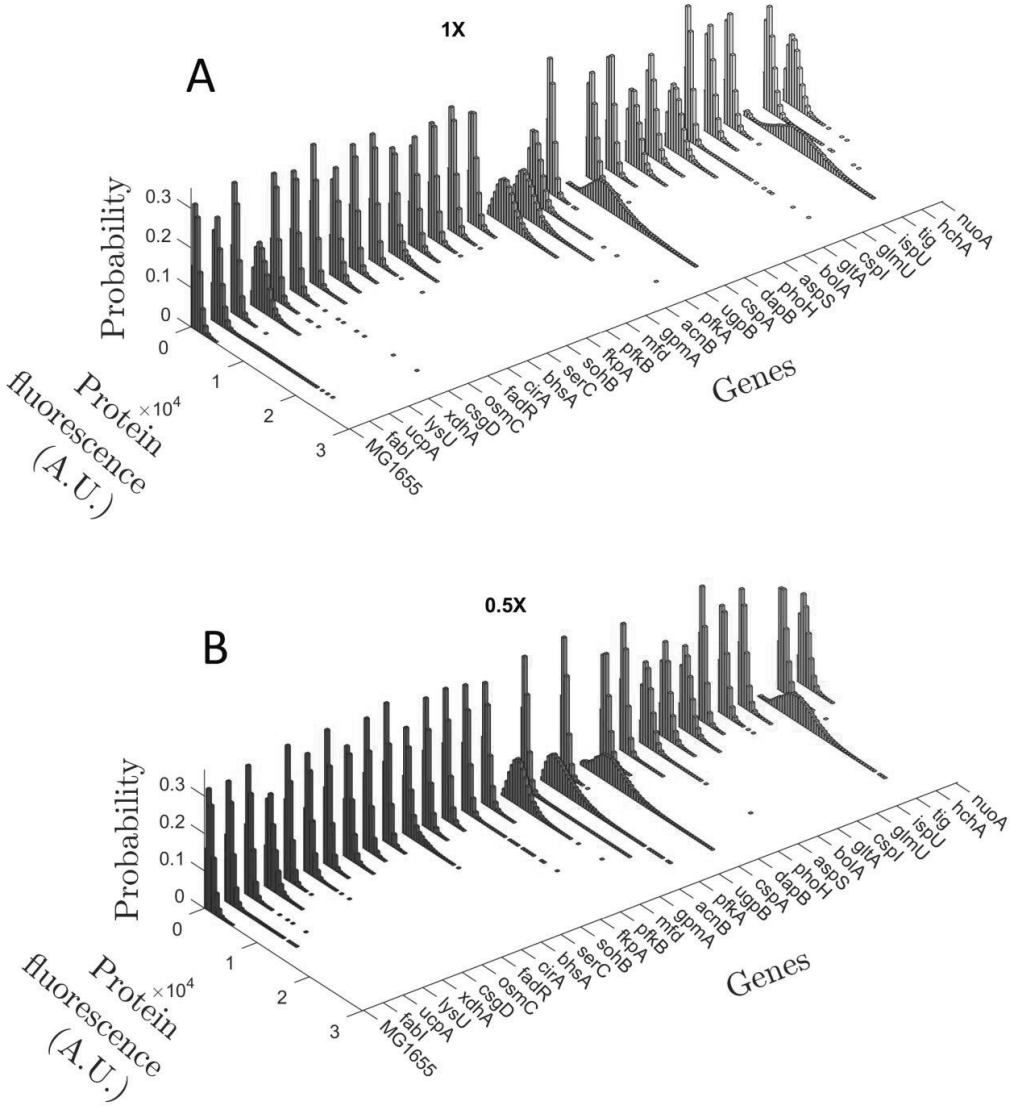


Fig H. Protein florescence distributions. Protein florescence distributions of genes controlled by tandem promoters measured by flow-cytometry**.** Each protein is tagged with a YFP (YFP strain library). Only 1 of 3 biological replicates is shown per gene. (A) M9 medium (1X). (B) Diluted M9 medium (0.5X). 'MG1655' are control cells, not carrying YFP. Protein fluorescence is shown in arbitrary units.

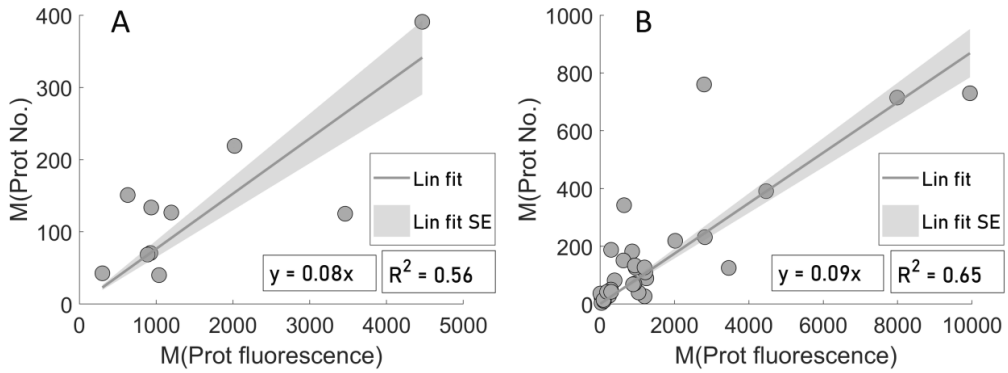


Fig I. Estimation of scaling factors using data from genes controlled by single promoters. A) Mean single-cell protein fluorescence (own measurements of genes controlled by single promoters) plotted against the corresponding mean single-cell protein numbers reported in [28]. From the equation of the 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.

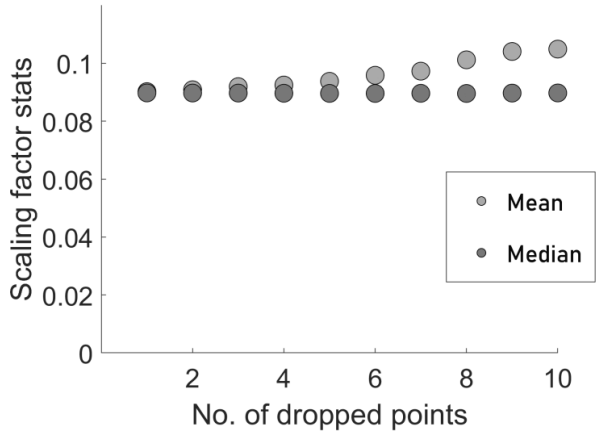


Fig J. Sensitivity test. Mean and median of scaling factor varies as a function of number of data points randomly dropped.

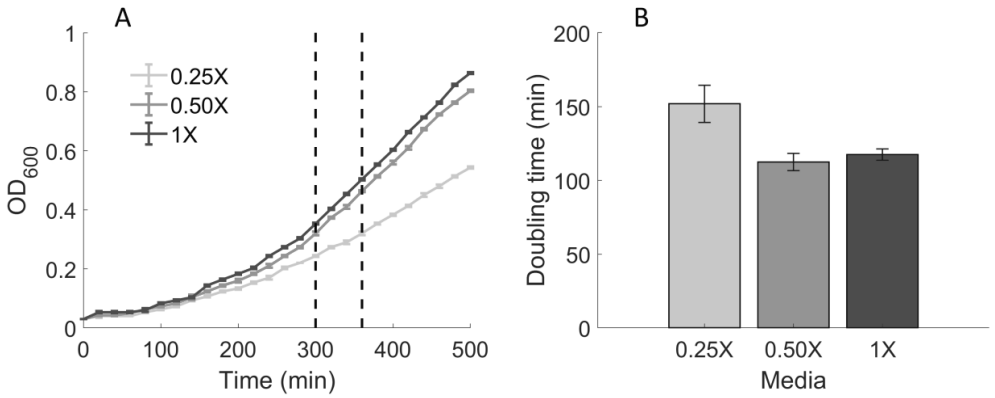


Fig K. Growth curves and doubling times. A. Optical density (OD_{600}) curves of *E. coli* MG1655 cells grown in 0.25X,0.5X and 1X media (section 'Media and Chemicals' in S1 Appendix). B. From these curves, the doubling time was estimated to be ~112 min in 0.5X and ~118 min in 1X. We used 115 min doubling time in the models. The estimation is made using the formula $D = \frac{\ln(2)}{\ln\left(\frac{OD(t_2)}{OD(t_1)}\right)} \times (t_2 - t_1)$, 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.

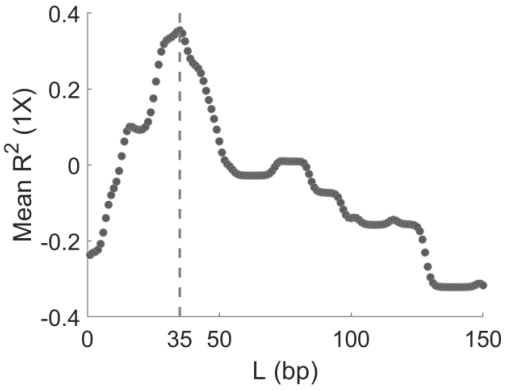


Fig L. Mean R^2 of the step interference model. Mean R^2 of the step interference model to the 1X data in Fig 6A, 6B, and 6C, as a function of L (d_{TSS} at which the step of the step function occurs). The Mean

R^2 is visibly maximized at L = 35, which marked by a grey dashed line. Relates to Fig 6 in the main manuscript.

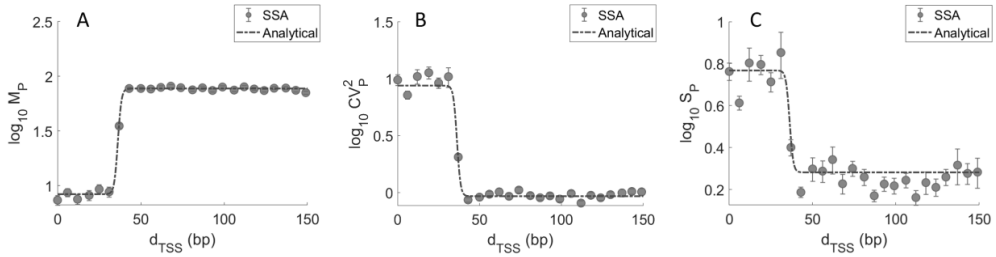
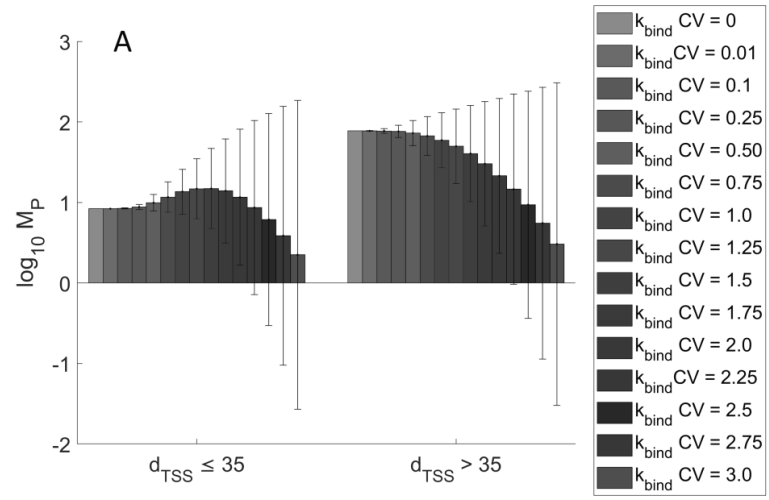
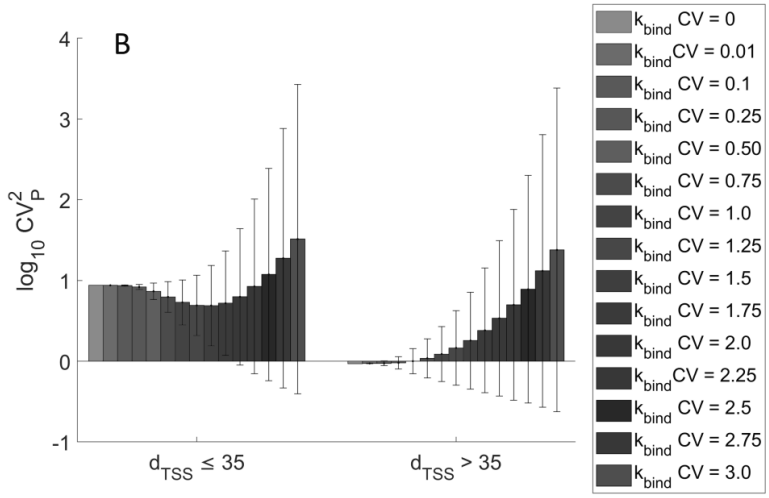


Fig M. Confronting the solutions of the analytical and stochastic model. (A) $\log_{10}$ of mean protein numbers, (B) $\log_{10}$ of CV^2 of protein numbers and (C) $\log_{10}$ of Skewness of protein numbers as a function of d_{TSS}. The blue line is the analytical solution of the step model. The blue 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
3
2
1
0
-1
-2
log10 MP
dTSS ≤ 35
dTSS > 35
kbind CV = 0
kbind CV = 0.01
kbind CV = 0.1
kbind CV = 0.25
kbind CV = 0.50
kbind CV = 0.75
kbind CV = 1.0
kbind CV = 1.25
kbind CV = 1.5
kbind CV = 1.75
kbind CV = 2.0
kbind CV = 2.25
kbind CV = 2.5
kbind CV = 2.75
kbind CV = 3.0


B
4
3
2
1
0
-1
log10 CV2P
dTSS ≤ 35
dTSS > 35
kbind CV = 0
kbind CV = 0.01
kbind CV = 0.1
kbind CV = 0.25
kbind CV = 0.50
kbind CV = 0.75
kbind CV = 1.0
kbind CV = 1.25
kbind CV = 1.5
kbind CV = 1.75
kbind CV = 2.0
kbind CV = 2.25
kbind CV = 2.5
kbind CV = 2.75
kbind CV = 3.0

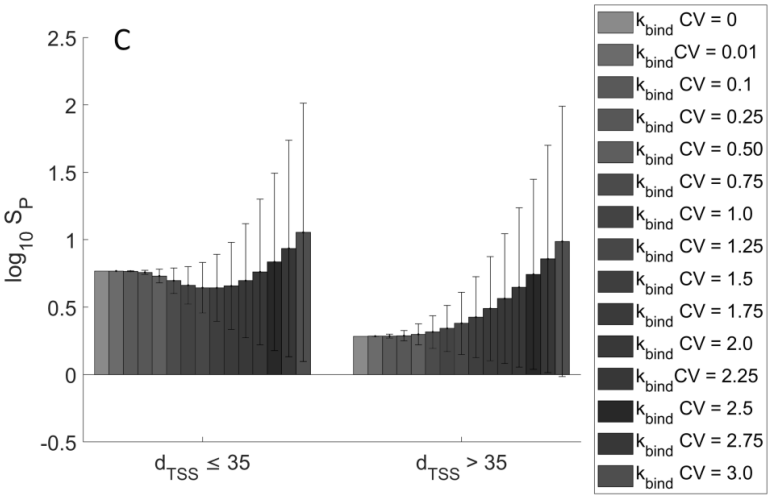

C
2.5
2
1.5
1
0.5
0
-0.5
log10 SP
dTSS ≤ 35
dTSS > 35
kbind CV = 0
kbind CV = 0.01
kbind CV = 0.1
kbind CV = 0.25
kbind CV = 0.50
kbind CV = 0.75
kbind CV = 1.0
kbind CV = 1.25
kbind CV = 1.5
kbind CV = 1.75
kbind CV = 2.0
kbind CV = 2.25
kbind CV = 2.5
kbind CV = 2.75
kbind CV = 3.0

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.

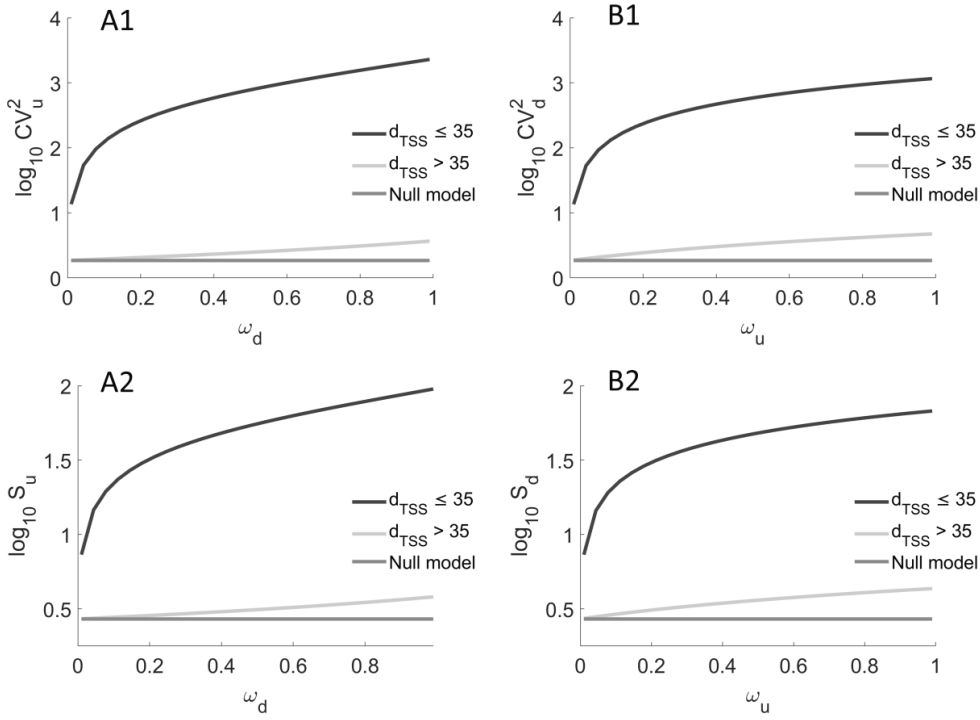


Fig O. Variability and skewness in single-cell protein numbers produced from an upstream and from a downstream promoter as a function of promoter occupancy of the other promoter. CV_P^2 and S_P of the single-cell distribution of the number of proteins produced (**A1 and A2**) by the upstream promoter 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.

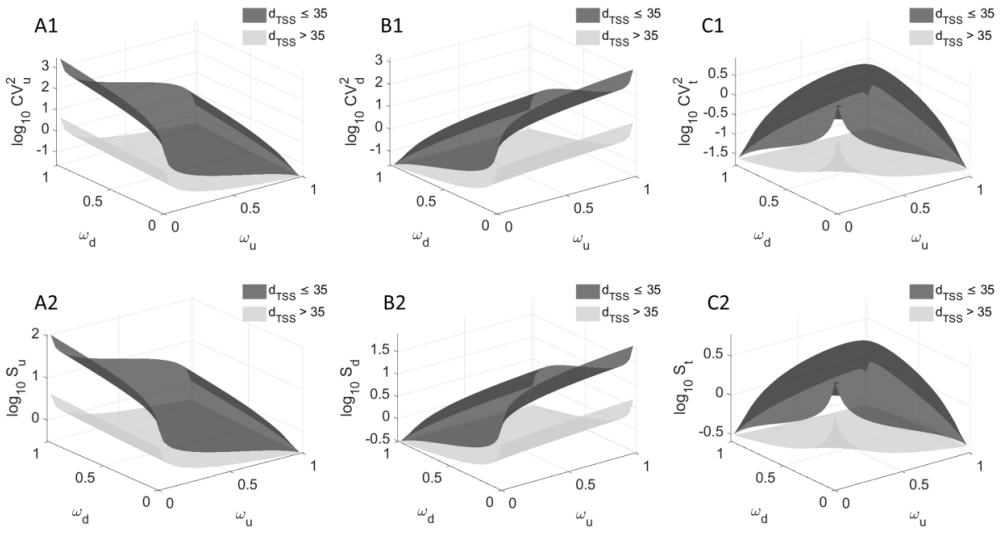


Fig P. Variability and skewness in single-cell protein numbers as a function of promoter occupancy. Expected variability (CV^2) and skewness (S) of the single cell distribution of protein numbers due to the activity of, respectively: (**A1** and **A2**) the upstream promoter alone, (**B1** and **B2**) 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.

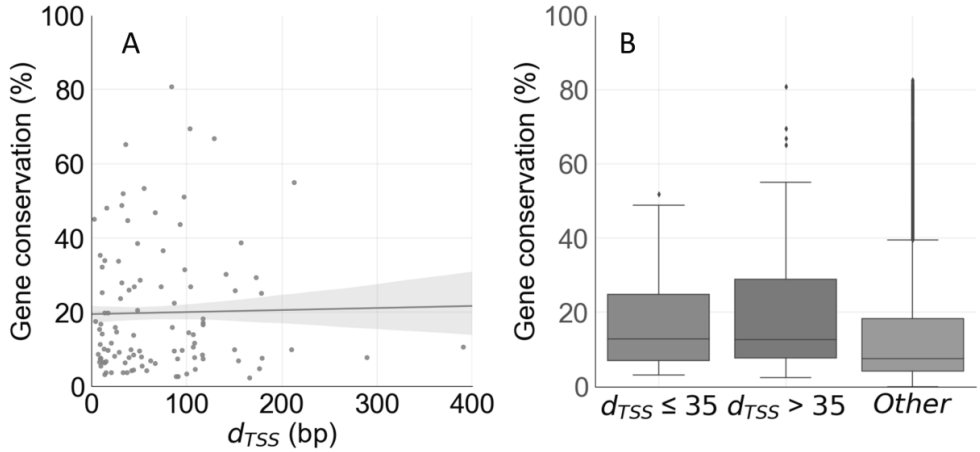


Fig Q. Gene conservation levels. (A) Correlation between d_{TSS} (bp) of the pairs of tandem promoters and the evolutionary conservation level of the gene that they express. The line shown is the best linear fit to the data, and its shadow is the standard error of the fit. (B) Box plot of the gene conservation 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 $d_{TSS} \leq 35$**	**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)	$0.11 \pm 4.34\times10^{-3}$
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 ($k_{bind} \cdot [R]$)	$Mean\left(abs\left(\begin{matrix} k^u_{bind} \cdot [R] - \\ k^d_{bind} \cdot [R] \end{matrix} \right) \right)$	$Mean\left(\frac{abs\left(\begin{matrix} k^u_{bind} \cdot [R] - \\ k^d_{bind} \cdot [R] \end{matrix} \right)}{k^u_{bind} \cdot [R]} \right) \times 100\%$
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 $k_{bind} \cdot [R]$ of each promoter in tandem configuration is sampled from. Also shown is the resulting expected mean absolute difference in $k_{bind} \cdot [R]$ between the upstream and downstream promoters. Furthermore, the last column shows how much larger (in percentage) is one of the $k_{bind} \cdot [R]$ 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

UNPUBLISHED MANUSCRIPT - STUDY V

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

B.L.B. Almeida, M.N.M. Bahrudeen*, V. Chauhan*, S. Dash*, V. Kandavalli, A. Häkkinen, J. Lloyd-Price, C.S.D. Palma, I.S.C. Baptista, A. Gupta, J. Kesseli, E. Dufour, O.P. Smolander, M. Nykter, P. Auvinen, H.T. Jacobs, S.M.D. Oliveira and A. S. Ribeiro.

bioRxiv. *Equal contributions.

https://doi.org/ 10.1101/2022.03.07.483226