

VINODH KANDAVALLI

**Rate-limiting Steps in  
Transcription Initiation  
are Key Regulatory  
Mechanisms of  
*Escherichia coli* Gene  
Expression Dynamics**



VINODH KANDAVALLI

Rate-limiting Steps in Transcription Initiation  
are Key Regulatory Mechanisms  
of *Escherichia coli* Gene  
Expression Dynamics

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 F115  
of the Arvo Building, Arvo Ylpön Katu 34, Tampere,  
on 20 March 2020, 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*

Ph.D. Tineke Laura Lenstra  
The Netherlands Cancer Institute  
Netherlands

Associate Prof. Mukund Thattai  
National Centre for Biological  
Sciences TIFR  
India

*Opponent*

Ph.D. Libor Krasny  
Institute of Microbiology  
Czech Republic.

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

Copyright ©2020 author

Cover design: Roihu Inc.

ISBN 978-952-03-1465-1 (print)

ISBN 978-952-03-1466-8 (pdf)

ISSN 2489-9860 (print)

ISSN 2490-0028 (pdf)

<http://urn.fi/URN:ISBN:978-952-03-1466-8>

PunaMusta Oy – Yliopistopaino  
Tampere 2020

# ABSTRACT

In all living organisms, the “*blueprints of life*” are documented in the genetic material. This material is composed of genes, which are regions of DNA coding for proteins. To produce proteins, cells read the information on the DNA with the help of molecular machines, such as RNAP holoenzymes and  $\sigma$  factors.

Proteins carry out the cellular functions required for survival and, as such, cells deal with challenging environments by adjusting their gene expression pattern. For this, cells constantly perform decision-making processes of whether or not to actively express a protein, based on intracellular and environmental cues.

In *Escherichia coli*, gene expression is mostly regulated at the stage of transcription initiation. Although most of its regulatory molecules have been identified, the dynamics and regulation of this step remain elusive. Due to a limited number of specific regulatory molecules in the cells, the stochastic fluctuations of these molecular numbers can result in a sizeable temporal change in the numbers of transcription outputs (RNA and proteins) and have consequences on the phenotype of the cells. To understand the dynamics of this process, one should study the activity of the gene by tracking mRNA and protein production events at a detailed level.

Recent advancements in single-molecule detection techniques have been used to image and track individually labeled fluorescent macromolecules of living cells. This allows investigating the intermolecular dynamics under any given condition. In this thesis, by using *in vivo*, single-RNA time-lapse microscopy techniques along with stochastic modelling techniques, we studied the kinetics of multi-rate limiting steps in the transcription process of multiple promoters, in various conditions.

Specifically, first, we established a novel method of dissecting transcription in *Escherichia coli* that combines state-of-the-art microscopy measurements and model fitting techniques to construct detailed models of the rate-limiting steps governing the *in vivo* transcription initiation of a synthetic *Lac-ara-1* promoter. After that, we estimated the duration of the closed and open complex formation, accounting for the rate of reversibility of the first step. From this, we also estimated the duration of periods of promoter inactivity, from which we were able to determine the contribution from each step to the distribution of intervals between consecutive RNA productions in individual cells.

Second, using the above method, we studied the  $\sigma$  factor selective mechanisms for indirect regulation of promoters whose transcription is primarily initiated by RNAP holoenzymes carrying  $\sigma^{70}$ . From the analysis, we concluded that, in *E. coli*, a promoter’s responsiveness to indirect regulation by  $\sigma$  factor competition is determined by its sequence-dependent, dynamically regulated multi-step initiation kinetics.

Third, we investigated the effects of extrinsic noise, arising from cell-to-cell variability in cellular components, on the single-cell distribution of RNA numbers, in the context of cell lineages. For this, first, we used stochastic models to predict the variability in the numbers of molecules involved in upstream

processes. The models account for the intake of inducers from the environment, which acts as a transient source of variability in RNA production numbers, as well as for the variability in the numbers of molecular species controlling transcription of an active promoter, which acts as a constant source of variability in RNA numbers. From measurement analysis, we demonstrated the existence of lineage-to-lineage variability in gene activation times and mean transcription rates. Finally, we provided evidence that this can be explained by differences in the kinetics of the rate-limiting steps in transcription and of the induction scheme, from which it is possible to conclude that these variabilities differ between promoters and inducers used.

Finally, we studied how the multi-rate limiting steps in the transcription initiation are capable of tuning the asymmetry and tailedness of the distribution of time intervals between consecutive RNA production events in individual cells. For this, first, we considered a stochastic model of transcription initiation and predicted that the asymmetry and tailedness in the distribution of intervals between consecutive RNA production events can differ by tuning the rate-limiting steps in transcription. Second, we validated the model with measurements from single-molecule RNA microscopy of transcription kinetics of multiple promoters in multiple conditions. Finally, from our results, we concluded that the skewness and kurtosis in RNA and protein production kinetics are subject to regulation by the kinetics of the steps in transcription initiation and affect the single-cell distributions of RNAs and, thus, proteins. We further showed that this regulation can significantly affect the probability of RNA and protein numbers to cross specific thresholds.

Overall, the studies conducted in this thesis are expected to contribute to a better understanding of the dynamic process of bacterial gene expression. The advanced data and image analysis techniques and novel stochastic modeling approaches that we developed during the course of these studies, will allow studying in detail the *in vivo* regulation of multi-rate limiting steps of transcription initiation of any given promoter. In addition, by tuning the kinetics of the rate-limiting steps in the transcription initiation as executed here should allow engineering new promoters, with predefined RNA and, thus, protein production dynamics in *Escherichia coli*.

# PREFACE

This Ph.D. study is submitted to the Doctoral Program of Biomedical Sciences and Engineering and the research work was carried out in the Laboratory of Biosystem Dynamics (LBD) of the BioMediTech Institute of the Faculty of Medicine and Health Technology of Tampere University, under the supervision of Professor Andre Sanches Ribeiro.

First and foremost, I would like to deeply thank Prof. Ribeiro for accepting me as his student and providing me all that was necessary to execute this thesis work. He gave constant support, valuable suggestions, and encouragement throughout this work. Without his motivation and support, this thesis would not have been possible. His hardworking nature, extremely organizing skills, passion, and dedication to science has been contagious and reignited my passion for science.

I would also like to express my sincere gratitude to Prof. Mukund Thattai and Dr. Teneke Lenstra for the careful pre-examination of my thesis. Their comments and insightful suggestions have helped me improve the thesis.

Next, I would like to thank the Centre for International Mobility (CIMO) for funding the first year of my Ph.D. studies, as well as the Finnish Cultural Foundation (the Pirkanmaa Regional Fund) for funding the final year of the Ph.D. studies. Further, I would like to thank the short-term funds provided by BioMediTech of Tampere University of Technology to support my doctoral studies. Furthermore, I thank the Doctoral Education Network on Intelligent System (DENIS) for funding a participation in a conference.

I would like to thank all my co-workers and co-authors, especially Dr. Jason Llyod-Price, Dr. Huy Tran, Dr. Jarmo Mäkelä and Dr. Sofia Startceva for introducing me to the complex theoretical concepts that were developed in all the publications. I also thank all the past members of the Laboratory of Biosystem Dynamics (LBD), especially Dr. Jerome Chandraseelan, Dr. Samuel Oliveria, Dr. Ramakanth Neeli and Dr. Nadia Goncalves for their valuable scientific discussions on Microscopy and other experimental setup and for keeping me cherish and motivated during the lab hours. Further, I would like to thank the current LBD members and budding scientists, especially Cristina Palma, Mohamed Bahrudeen, Bilena Almeida and Ines Baptista, for beneficial discussions during the thesis writing. In my routine lab work, I am grateful to my lab mates and upcoming scientists, especially Vatsala Chauhan and Suchintak Dash, for creating a nice and supportive environment and for discussions on the scientific part and useless stuff. I offer my best regards to all others who supported me in any respect during the completion of the Ph.D. studies.

I am grateful to all the non-scientific staff of the University for their on-time help during my Ph.D. studies. On a lighter note, I would like to thank the cricket communities (Tampere Cricket Club and Indian Cricket Club) for making me fit and active all the time (especially winter time), which has been very supportive

during the stay in Finland. I am greatly thankful to all my friends in Tampere and India who engaged with me during the weekends and not to feel homesick.

I would like to thank countless time to my parents Prasada Rao and Sarala Devi for their endless love, support, and encouragement since my birth. Special thanks to my siblings Madhuri, Manoj and Bhanu for their love and support. Finally, I would like to thank once again Vatsala Chauhan for understanding and being next to me in all the situations, from the last two years both at on and off the work and make things possible with her love, care, and support. Hopefully looking forward to seeing life long journey.

Tampere, January 29, 2019.

Vinodh Kandavalli.



# CONTENTS

ABSTRACT

PREFACE

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

LIST OF PUBLICATIONS

1	INTRODUCTION.....	1
	1.1 Background and Motivation .....	1
	1.2 Aims of the Study.....	2
	1.3 Thesis Outline .....	4
2	REVIEW OF LITERATURE .....	5
	2.1 The Central Paradigm of Molecular Biology.....	5
	2.2 <i>Escherichia coli</i> as a model organism.....	7
	2.3 Gene expression in <i>Escherichia coli</i> .....	8
	2.4 Bacterial Transcription.....	10
	2.4.1 Transcription mechanism .....	10
	2.4.2 Rate-Limiting steps in transcription initiation.....	13
	2.4.3 Transcription elongation.....	16
	2.4.4 Transcription termination.....	16
	2.5 Gene Regulation at the Transcription Level .....	17
	2.5.1 Promoter region .....	17

2.5.2	Regulation by transcription factors .....	18
2.5.3	Regulation by $\sigma$ factors.....	20
2.5.4	Other regulatory factors.....	21
2.6	Transcription Noise.....	22
3	MATERIALS AND METHODS .....	24
3.1	Basics of Microscopy and Fluorescent proteins .....	24
3.2	Single-molecule methods for quantifying transcription dynamics.....	27
3.2.1	MS2-GFP tagging system .....	28
3.2.2	Engineering of Synthetic Genetic Constructs .....	30
3.2.3	Time-lapse microscopy .....	31
3.3	Validation methods.....	32
3.3.1	Quantitative polymerase chain reaction. ....	32
3.3.2	Western blotting.....	34
3.3.3	Flow Cytometry.....	35
3.4	Stochastic Simulation Algorithm and Stochastic Simulators .....	36
3.5	Models of Transcription.....	37
3.6	Tau ( $\tau$ ) plots.....	41
4	IMAGE AND DATA ANALYSIS.....	43
4.1	Cell Segmentation and Lineage Construction.....	43
4.2	Spot Detection .....	44
4.3	Extraction of Time intervals from Total Spot Fluorescence over time.....	45
4.4	Asymmetry and Tailedness of the Distribution of Transcription Intervals .....	48
5	RESULTS: SUMMARY AND CONCLUSIONS.....	50
6	DISCUSSION.....	54

7	REFERENCES .....	56
---	------------------	----

# LIST OF FIGURES

<b>Figure 2.1:</b> The central paradigm of molecular biology .....	6
<b>Figure 2.2:</b> Phase contrast image of <i>E. coli</i> cells .....	7
<b>Figure 2.3:</b> Schematic representation of typical <i>E. coli</i> operon .....	8
<b>Figure 2.4:</b> The translation process in <i>E. coli</i> .....	9
<b>Figure 2.5:</b> Structure of RNA polymerase and its interaction with a promoter region .....	11
<b>Figure 2.6:</b> Transcription cycle in <i>E. coli</i> .....	12
<b>Figure 2.7:</b> Depiction of the transcription initiation mechanisms in <i>E. coli</i> .....	14
<b>Figure 2.8:</b> Schematic representation of repression (a) and activation (b) mechanisms of promoter activity using transcription factors .....	19
<b>Figure 2.9:</b> Schematic representation of how various sources of noise affect the stochastic gene expression in a clonal population of bacteria .....	22
<b>Figure 3.1:</b> Purified fluorescent proteins .....	25
<b>Figure 3.2:</b> Lightpath illustration of wide-field epi-illumination and confocal microscopy .....	26
<b>Figure 3.3:</b> Schematic overview of the MS2-GFP component system .....	29
<b>Figure 3.4:</b> Overview of the single-step reaction of the Gibson Assembly® method .....	31
<b>Figure 3.5:</b> An example of qPCR results .....	33
<b>Figure 3.6:</b> RNAP subunits quantification by the western blot method .....	34
<b>Figure 3.7:</b> Schematics of a flow cytometer .....	35
<b>Figure 3.8:</b> Absolute $\tau$ -plot .....	42
<b>Figure 4.1:</b> Phase-contrast and confocal images of <i>E. coli</i> cells .....	44
<b>Figure 4.2:</b> Single-cell distribution of RNA spot and cell intensities .....	45
<b>Figure 4.3:</b> Quantification of integer-value RNA molecules in individual cells .....	46

**Figure 4.4:** Cartoon of RNAs spot appearance in the cells.....47

# LIST OF TABLES

**Table 1:** List of *E. coli*  $\sigma$  factors.....20

## LIST OF ABBREVIATIONS

<b>ATP</b>	Adenosine triphosphate
<b>BP</b>	Base pairs
<b>BS</b>	Binding sites
<b>cAMP</b>	Cyclic Adenosine Monophosphate
<b>CAP</b>	Catabolite activator protein
<b>CC</b>	Closed complex
<b>CME</b>	Chemical master equation
<b>CRP</b>	cAMP receptor protein
<b>CV</b>	Coefficient of Variation
<b>DNA</b>	Deoxyribonucleic Acid
<b>EC</b>	Elongation complex
<b>FISH</b>	fluorescence in situ hybridization
<b>FP</b>	Fluorescent probe
<b>GRN</b>	Gene Regulatory Network
<b>GFP</b>	Green fluorescent protein
<b>HILO</b>	Highly inclined and laminated optical sheet
<b>HSPs</b>	Heat-shock proteins
<b>ITC</b>	initial transcribing complex
<b>IPTG</b>	Isopropyl- $\beta$ -D-1-Thiogalactopyranoside
<b>kDa</b>	Kilo daltons
<b>KDE</b>	Kernel density estimation
<b>mRNA</b>	messenger RNA
<b>NAPs</b>	Nucleoid-associated proteins
<b>OC</b>	Open complex
<b>ORF</b>	Open reading frame
<b>PALM</b>	Photo-activation localization microscopy
<b>PCA</b>	Principal Component Analysis
<b>PCR</b>	Polymerase chain reaction
<b>qPCR</b>	Quantitative PCR
<b>RBS</b>	Ribosome binding site
<b>RNA</b>	Ribonucleic acid
<b>rRNA</b>	Ribosomal RNA
<b>RNAP</b>	RNA Polymerase
<b>RT-PCR</b>	Reverse transcriptase PCR
<b>SSA</b>	Stochastic Simulation Algorithm
<b>STORM</b>	Stochastic Optical Reconstruction Microscopy
<b>TFs</b>	Transcription Factors
<b>TIRF</b>	Total Internal Reflection Fluorescence

# LIST OF PUBLICATIONS

This thesis is a compilation of five studies. In the text, these are referred to as **Publication I, II, III, IV, and V**. The Publications are reproduced with permission from the publishers.

- I. J. Lloyd-Price, S. Startceva, **V. Kandavalli**, J.G. Chandraseelan, N. Goncalves, S.M.D Oliveira, A. Häkkinen and A.S. Ribeiro. “Dissecting the stochastic transcription initiation process in live *Escherichia coli*”, *DNA Res.*, 23 (3): 203-214, 2016.
- II. **V. Kandavalli**, H. Tran and A.S. Ribeiro. “Effects of  $\sigma$  factor competition are promoter initiation kinetics dependent”. *Biochimica et Biophysica Acta: Gene Regulatory Mechanisms*, 1859, 1281–1288, 2016.
- III. J. Mäkelä, **V. Kandavalli** and A.S. Ribeiro. “Rate-limiting steps in transcription dictate sensitivity to variability in cellular components”. *Scientific Reports*, 7:10588., 2(10), 2017.
- IV. **V. Kandavalli**, S. Startceva, and A.S. Ribeiro. “Transcription initiation controls skewness of the distribution of intervals between RNA productions”, *In Proceedings of the 31<sup>st</sup> European Simulation and Modelling (ESM)*, Portugal. EUROISIS, pp. 418-421. ISBN: 978-9492859-00-6. Paulo J.S (Ed.), 2017.
- V. S. Startceva, **V. Kandavalli**, A. Visa, and A.S. Ribeiro. “Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression”, *Biochimica et Biophysica Acta: Gene Regulatory Mechanisms*, 1862 (2), 119-128, 2019.

The author of this thesis contributed to the Publications as follows:

In **Publication I**, the author performed the microscopy and qPCR measurements and assisted J.G. Chandraseelan with other experiments. The author also assisted in image analysis. The author participated in analyzing the results with J. Lloyd-Price, S. Startceva, and A. S. Ribeiro, and contributed to the writing of the manuscript.

In **Publication II**, the author conceived the study with A.S. Ribeiro. The author designed and conducted all microscopy, qPCR, WB, and other experiments. The author also participated in the image analysis and analyzed the results with H. Tran and A.S. Ribeiro. The theoretical and mathematical model of this study was developed by H.



Tran, assisted by the author. Finally, the author participated in writing the manuscript with the corresponding author.

In **Publication III**, the author performed all qPCR, WB, and plate reader experiments. In addition, the author also performed some microscopy experiments, assisted in the image analysis and in the analysis of the results. Finally, the author contributed to the writing of the manuscript.

In **Publication IV**, the author conceived the study with A.S. Ribeiro. The author designed and conducted all experiments. The author assisted in the image analysis. Finally, the author analyzed the results, participated actively in all discussions during the execution of the project, and performed most of the writing of the manuscript.

In **Publication V**, the author conceived the study with S. Startceva and A.S. Ribeiro. The author is responsible for designing and conducting all experiments. The author assisted in the image analysis assisted by S. Startceva. Finally, the author analyzed the results with the other authors and participated actively in the discussions and in the writing of the manuscript.

**Publications I and V** will be included by Sofia Startceva in her Ph.D. thesis.



# 1 INTRODUCTION

## 1.1 Background and Motivation

During the course of evolution, all living organisms, from lower to higher-order, have developed highly sophisticated mechanisms that allow them to survive in a wide range of environmental conditions (Yamanaka 1999; Sleator & Hill 2001; López-Maury et al. 2008). This robustness is essential for microorganisms, as they constantly face fluctuating environments (Ramos et al. 2001). Studying how organisms achieve this robustness contributes to a better understanding of biological systems.

Several studies have shown that regulation of gene expression is the core process of adaptability to changing environments (Stoebel et al. 2009). Gene expression, a fundamental process of all living organisms, is the process by which the information on the genetic material (DNA) is, first, transcribed into RNA and, consequently, is translated from RNA into functional proteins (Crick 1970).

Dynamically, gene expression is highly complex, even in prokaryotes, since it is multi-stepped, sequence-dependent, and subject to regulation (Saecker et al. 2011; Muthukrishnan et al. 2012; Albert et al. 2014; McClure 1985). In *Escherichia coli* (*E. coli*), most regulation of gene expression dynamics occurs at the stage of transcription initiation (Shih & Gussin 1983; Golding & Cox 2004; Jones et al. 2014; Browning & Busby 2016; Mäkelä et al. 2011; Gonçalves et al. 2016). Transcription initiation starts once an RNA polymerase holoenzyme finds and binds to a promoter region, so as to form a closed complex. Next, the RNA polymerase unwinds the double helix DNA to form an open complex. For this, the RNA polymerase undergoes several isomerization steps. These will eventually lead to the initiation of RNA synthesis, once promoter clearance is achieved.

Evidence suggests that these two steps i.e. closed and open complex steps are rate-limiting, implying that they are amongst the slowest events in transcription, and thus, are the ones most affecting the rate of RNA production. Consequently, by regulating the kinetics of these steps, cells can fine-tune the production rate of transcripts. The RNA production is noisy due to intrinsic factors, such as the kinetics of the process (for example, the closed complex formation differs in time length between events) and due to extrinsic factors (for example, different cells might not have the same number of transcription factors).

Most of our knowledge on transcription initiation derives from *in vitro* studies using techniques such as the abortive initiation method, DNA foot-printing, and gel-based assays, among others (McClure 1980;

Bertrand-Burggraf et al. 1984; McClure 1985; deHaseth et al. 1998). However, recent advancement in single-molecule imaging techniques has made possible to study the dynamics of gene expression in living cells with high details (Peabody 1993; Fusco et al. 2003; Yu et al. 2006; Golding et al. 2005). It was these new techniques that exposed the stochastic nature of transcription.

One popular method for conducting *in vivo* studies of RNA production kinetics is the MS2-GFP tagging system since it allows monitoring the RNA molecules in real-time in individual cells with high spatial and temporal resolution (Fusco et al. 2003; Xie et al. 2008; Golding et al. 2005). Among others, these measurements made possible the quantification of the *in vivo* kinetics of the intermediate steps in transcription initiation, such as the closed and open complex (Muthukrishnan et al. 2012; Lloyd-Price et al. 2016; Kandavalli et al. 2016).

Other methods, such as fluorescence in situ hybridization (FISH) have focused on the quantification of RNA numbers with single-molecule sensitivity in cell populations. This technique has also been widely used (Jones et al. 2014; Sanchez et al. 2013; Sanchez & Golding 2013), for showing, e.g., that promoter sequences can control not only the mean but also the variability of the kinetics of constitutive expression in *E. coli* (Jones et al. 2014).

By observing the dynamics of transcription in various environmental conditions, much knowledge was obtained on the regulatory mechanisms used by cells to obtain single-gene, media-dependent expression patterns (Dong & Schellhorn 2009; Mäkelä et al. 2013). For example, when cells are in starvation, the expression of stress sigma factors is activated, changing the competition by sigma factors for the core RNA polymerase. As a result, not only some genes are upregulated, but there are also adverse expression regulations in genes recognized by other sigma factors (Farewell et al. 1998).

Overall, present, state-of-the-art, time-lapse, *in vivo* single-cell measurements, when combined with tailored image and signal analysis tools, now allow dissecting the dynamics of gene expression, at the RNA and rate-limiting step levels (Golding & Cox 2004; Golding et al. 2005; Mäkelä et al. 2013). Meanwhile, stochastic models of gene expression can be used to predict/mimic these dynamics and test novel hypotheses of underlying mechanisms that can explain the experimental data (Arkin et al. 1998; Roussel & Zhu 2006; Ribeiro 2010; Ribeiro & Lloyd-Price 2007).

This research is expected to assist in a better understanding of how cells regulate the kinetics of transcription and, consequently, the distributions of RNA numbers in cell populations. This knowledge is expected to assist in the engineering of robust and sensitive synthetic circuits.

## 1.2 Aims of the Study

This thesis aims to characterize the rate-limiting steps in transcription initiation and how their kinetics controls the RNA production kinetics of individual genes in *E. coli*, including its adaptability to various environmental conditions.

For this, we established five main objectives.

Our first objective was to dissect the kinetics of the rate-limiting steps in transcription initiation of live *E. coli* cells. In vitro studies using abortive initiation assays have shown that these steps kinetics can be dissected by measurements differing in the free RNA polymerase concentration (McClure 1980). After showing that there is a range of media conditions for which the concentration of RNA polymerases differs between conditions, while the fraction of RNA polymerases free for transcription remains approximately constant, we employed the same strategy, but to dissect the in vivo kinetics.

In particular, we quantified the duration of the steps prior (closed complex formation) and after commitment to the open complex formation of *E. coli* promoters. For this, first, we produced media with different richness, such that the intracellular RNA polymerase concentration differed, without affecting the growth rates. Next, we perform measurements of the time intervals between RNA productions, at the single-molecule level, in different intracellular RNA polymerase concentrations. From the data, we applied a standard model-fitting procedure to fully characterize the in vivo kinetics of the rate-limiting steps in transcription initiation of the *E. coli* promoters. The results were presented in **Publication I**.

Second, since *E. coli* expresses  $\sigma$  factors following stresses, which results in the direct activation of specific genes, because there is a limited pool of RNA polymerase core enzymes (Grigorova et al. 2006; Maeda et al. 2000), we hypothesized that this ought to cause indirect downregulation of genes expressed by other  $\sigma$  factors (Farewell et al. 1998). First, we showed by mathematical analysis and stochastic modelling that this hypothesis ought to be true, in accordance with standard stochastic models of transcription. In particular, the model showed that, by changing the concentration of a specific holoenzyme, one should affect the rate of transcription of promoters not directly affected by that holoenzyme. Next, to validate the model predictions, we performed measurements of RNA production at single-cell, single-molecule level, and qPCR in live *E. coli* cells in various growth phases. These showed that, when the levels of  $\sigma^{38}$  increased, the transcription rate of  $\sigma^{70}$ -dependent promoters decreased as predicted. Further, we showed that the degree of change by negative regulation differs with the kinetics of transcription initiation of the gene of interest, in that it's higher the longer is the fraction of time spent prior to commitment to open complex formation, also as predicted by the model. These results are published in **Publication II**.

Third, our objective was to demonstrate that cell-to-cell variability in transcription activation times (such as due to the intake of an activator for the media) introduces cell-to-cell variability in RNA numbers which propagate over time, resulting in lineage-to-lineage variability in gene expression products. In addition, we show that the amplitude of this phenomenon differs with the kinetics of transcription initiation of the promoter of interest. For this, we performed single-cell, single RNA

time-lapse microscopy, activating promoters in cells already under observation. Further, we tracked cells while replicating for several generations and quantified the inheritance of single MS2-GFP tagged RNAs. The empirical results, in agreement with our stochastic models, validate the hypotheses that the effects of extrinsic noise are promoter initiation kinetics dependent and thus are evolvable and adaptable. These results were reported in **Publication III**.

Finally, we focused on the study of to what level of detail gene expression in *E. coli* can be tuned by regulating the rate-limiting steps in transcription. In particular, having been established that mean and variability in RNA production can be tuned, we hypothesized that should also be possible to tune the probability of crossing thresholds by tuning the asymmetry (skewness) and tailedness (kurtosis) of the distribution of intervals between consecutive RNA production events. To test this, we performed live, single-cell, single-RNA microscopy measurements of the asymmetries of time intervals between consecutive RNA production events in individual cells in various conditions, including multiple promoters, single-point mutant promoters, stresses, induction schemes, and growth phases. These showed that it is possible to change skewness and kurtosis by regulation and by sequence. Next, using stochastic modelling and measurements, we showed that this tuning causes significant differences in the threshold crossing probabilities in protein numbers. The results were presented in **Publications IV and V**.

### 1.3 Thesis Outline

This thesis is organized as follows: Chapter 2 provides the biological background, which includes an overview of the gene expression process, followed by a more detailed description of the bacterial transcription process and regulatory mechanisms. Also, the concepts of noise in the transcription process are introduced. Chapter 3 presents the experimental and theoretical methods used in the Publications composing the thesis that are necessary to quantify transcription dynamics from measurements of single RNAs in live cells and models and then infer the rate-limiting steps of the models of transcription initiation. Chapter 4 introduces the computational tools used to analyze the microscopy images, including cell segmentation, lineage construction, tracking of fluorescence spots, and quantification of single RNA molecules. Finally, the results summary, conclusions, and discussion are presented in Chapter 5 and 6.

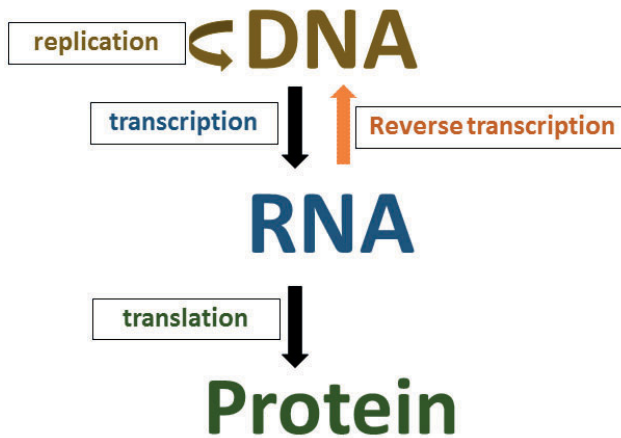
## 2 REVIEW OF LITERATURE

This chapter provides an overview of the biological process on which this work focuses on. Briefly, it describes the central paradigm of molecular biology and introduces the model organism used in this study, *E. coli*. This chapter also introduces the transcription and its regulatory mechanisms. Finally, the concept of noise in transcription is described.

### 2.1 The Central Paradigm of Molecular Biology

One of the greatest achievements of Molecular Biology is the discovery of the DNA (deoxyribonucleic acid) structure by Watson and Crick (Crick 1970). Several subsequent experiments proved that DNA is hereditary molecules that propagate the information from one generation to the next. Cells store hereditary information in the form of DNA, which is essential for controlling metabolism, growth, and reproduction. DNA alone cannot perform these tasks; it requires other biopolymers to decode its information (genetic material) and produce functional proteins. The central paradigm of molecular biology is a fundamental process that occurs in all living organisms to transfer of sequence information between the DNA, RNA, and Proteins (Crick 1970). In most cases, the flow of genetic information in cells is from DNA to RNA, to proteins (Figure 2.1). In rare cases, such as some viruses and in some specific laboratory conditions, reverse transcription (information flow from RNA to DNA), RNA replication, and direct translation of DNA into proteins have been observed.

The DNA molecule consists of two long polypeptide chains that coil around each other to form a double helix-like structure. Each chain is made up of nucleotide subunits, composed of a deoxyribose sugar, a nucleobase, and a phosphate group. There are four types of nucleobases in the DNA molecule: adenine (A), thymine (T), cytosine (C) and guanine (G). Adenine always pairs with thymine while cytosine pairs with guanine. The fraction of A's equals the fraction of T's in the DNA. Likewise, the fractions of C and G are identical, as these nucleotides are joined to one another in a chain by a covalent bond between the sugar and phosphate groups, forming a backbone structure. The nucleobases of each strand are bound together by a hydrogen bond to form the double-strand DNA. The linear sequences of A, T, G, and C, on the DNA, encode all of the hereditary information of living cells. A cell has one or more copies of its DNA. DNA is replicated with the help of an enzyme, called DNA-dependent DNA polymerase, and passed on to progeny cells during the division process of a cell.



**Figure 2.1:** The central paradigm of molecular biology. This includes duplication of DNA to many copies by the process of Replication. Next, the information stored in DNA is transcribed into RNA by the process of transcription, and finally, the RNA is translated into proteins by the process of Translation. In rare cases, the information from the DNA directly flows to proteins and, the information on the RNA can be transferred to the DNA by a process called reverse transcription, which is represented by the orange arrow.

The short-lived intermediate product of gene expression, the RNA (Ribonucleic acid) (Alberts et al. 2002) produced by the RNA polymerase, unlike the DNA, quickly degrades once formed (Bernstein et al. 2002). The RNA molecule is similar to a DNA molecule, except in structure and one nucleobase. In particular, unlike DNA, the RNA is a single strand, and it has an additional nucleobase, uracil (U), instead of thymine. RNA exists in three forms: mRNA (messenger RNA) which carries the information from DNA to ribosomes, tRNA (transfer RNA) that transfers the specific amino acid into a growing polypeptide chain and, rRNA (ribosomal RNA), which is a catalytic component of ribosomes. Another form of RNA that was discovered more recently is siRNA (small interfering RNA), which plays a role in the regulatory pathway, e.g. in gene silencing, unlike other forms of RNA, that do not have an active role in gene expression regulation.

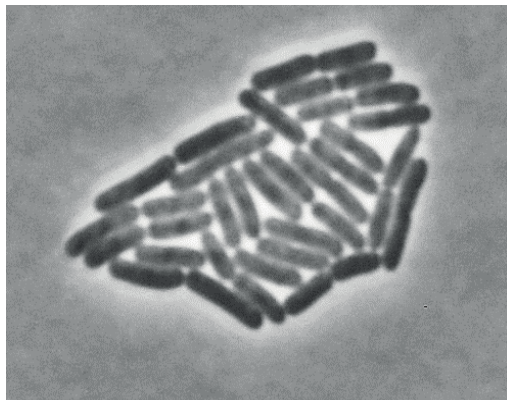
The final product of gene expression, Proteins, constitutes most of the cell's dry mass. They are not only the building blocks of the cells but also perform all cellular functions required for the maintenance of life (Alberts et al. 2002). The wide range of functionalities, which includes performing biological reactions, controlling gene expression, and responding to cellular signals. Proteins are synthesized from RNA by the process of translation, which is performed by protein-RNA complexes, named ribosomes. From a structural point of view, proteins are polypeptides made up of a linear chain of amino acids. Each amino acid is made up of  $\alpha$ -carbon, an amino group, a carboxyl group, and a side chain. The two ends of polypeptide chains are chemically different: the end carrying the free amino group is the amino terminus or N-terminus, and the end carrying the free carboxyl group is named carboxy-terminal, or C-terminus. Proteins are brought together by the covalent interaction between two peptides, forming a three-dimensional structure, which is necessary in order to become functional. In prokaryotes, such as *E. coli*, transcription and translation occur in parallel, allowing synthesis of proteins soon after the RNA



appearance (Conn et al. 2019; Mcgary & Nudler 2013). Meanwhile, in eukaryotes, there are additional steps, such as post-transcription regulation, including chaperon, intron, and exons modifications, which occur prior to protein synthesis.

## 2.2 *Escherichia coli* as a model organism.

*E. coli* is the most common bacterium found in the human gut and other warm-blooded organisms (Alberts et al. 2002). *E. coli* is a gram-negative, rod-shaped bacterium (Figure 2.2), that ranges between 2-4  $\mu\text{m}$  long and 0.5 to 0.8  $\mu\text{m}$  in diameter (Neidhardt 1987; Volkmer & Heinemann 2011) and has a cell volume of 0.6-0.7  $\mu\text{m}^3$  (Kubitschek 1990; Murray et al. 2009). It can reproduce very rapidly, both in the presence and in the absence of oxygen: a single cell can divide into millions of cells that form a colony in half a day. Many researchers have long been extensively using *E. coli* in laboratory conditions, making it an important research organism for Molecular Biology for over a century. Due to that, it is arguably one of the best-known prokaryotic organisms (Hufnagel et al. 2015).



**Figure 2.2:** Phase contrast image of *E. coli* cells. This rod-shaped bacterium has served as a model organism as it is widely used by biologists for research purposes.

Much is known about the physiology, genetics, biochemistry and molecular biology of *E. coli* (Alberts et al. 2002; Sambrook J & D W Russell 2001). The standard laboratory *E. coli* strain K-12 has a genome of 4.6 million base pairs of nucleotides, which is packed and condensed in a supercoiled single circular chromosomal double-stranded DNA. It contains approximately 2600 clusters of genes called operons and 4288 protein-coding sequences, (Blattner et al. 1997). Unlike in eukaryotes, *E. coli* has no nuclear envelope surrounding the bacterial chromosome. While, in general, the genes required for basic survival and reproduction are found in a single chromosome, *E. coli* cells can also contain plasmids, which are smaller DNA molecules that usually carry genes for specialized functions, such as resistance to a specific drug (Russo &

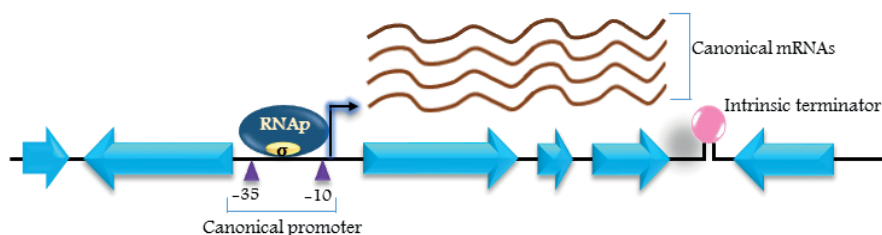
Johnson 2003). Standard experimental methods for manipulating *E. coli* extensively used by researchers include the modification of chromosomal DNA and biochemical analyses (Miller 1992; Sambrook J & D W Russell 2001). Due to its fast-growth and easy manipulation, it is an ideal experimental platform for quantitative, single-cell studies of gene expression (Elowitz & Leibler 2000; Golding et al. 2005; Yu et al. 2006; Alon 2007; Muthukrishnan et al. 2012).

## 2.3 Gene expression in *Escherichia coli*

Genes are the fundamental units of Biology. A gene is a storage of information in the form of DNA of how to code a protein. Gene expression regulation is made possible by a wide range of evolved mechanisms so that the cell can produce the specific amount of RNA and then proteins. The regulations also determine when the protein should be produced. Both how much and when are essential variables in determining the adaptability of organisms to a given environment. For example, *E. coli* might use different food sources at different times, implying that the cell will require different proteins in these conditions (F.C. Neidhardt et al. 1991).

Structurally, each gene consists of three molecular elements: Promoter, Operator and structural genes. A promoter is an upstream part of the DNA sequence of a gene. It is a region containing a specific site (the consensus region) to which the RNA polymerase first binds to, thereby initiating transcription (Figure 2.3). Genes also have regulatory regions, called operators, which are upstream or downstream of the open reading frame (ORF) that alter the expression of the gene. For example, Activator molecules enhance transcription activity by recruiting the RNA polymerase to the promoter. Conversely, repressor molecules can make promoter regions less available for RNA polymerases (Alberts et al. 2002; Herna et al. 2009).

Many prokaryotic structural genes are clustered and organized into operons, which are sets of genes under the control of a single promoter (Osborn & Field 2009; Eugene V. Koonin 2009). The genes in an operon are transcribed as continuous mRNAs (Polycistronic mRNA) which encodes for two or more proteins (Jacob & Monod 1960).

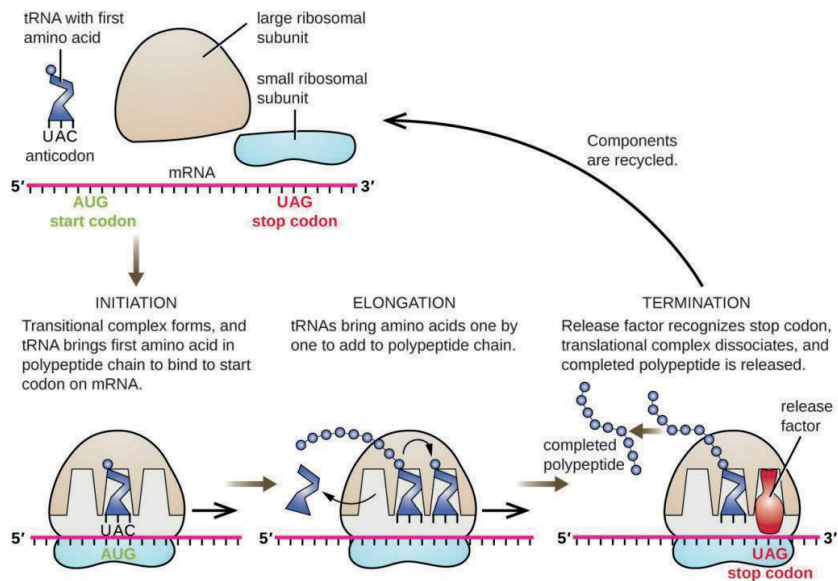


**Figure 2.3:** Schematic representation of typical *E. coli* operon. The mRNA synthesis begins with RNAP holoenzyme binding to the promoter region, followed by the formation of the elongation complex, produces the mRNA in a stem-loop structure called intrinsic termination. This picture is adopted from (Wade & Grainger 2014) and reprinted with permission from Macmillan Publishers Ltd: (Nature Reviews Microbiology).

Translation, the second step of the gene expression process, involves the synthesis of amino acids from the information encoded in the mRNA sequence and is performed by ribosomes. In *E. coli*, transcription and translation can occur simultaneously, since proteins can be assembled from RNAs that have not yet been completely assembled. This is accomplished by the ribosomes, which consist of ribosomal RNA (rRNA) molecules and many proteins components that are assembled by a tightly regulated process (Kaczanowska & Ryde 2007).

*E. coli* ribosomes sediment of 70S particles that consist approximately two-thirds RNA and one-third protein (Schuwirth et al. 2005) and are formed by two unequal subunits, 50S, and 30S. The larger subunit (the 50S) is composed of two rRNA (23S and 5S) and 33 ribosomal proteins involved in the catalysis of a peptide bond. The small subunit (the 30S) is made up of one rRNA (16S) and 21 ribosomal proteins, involved in decoding the mRNA sequence (Noller 2012; Horan & Noller 2007).

Each subunit has three binding sites for transfer RNA (tRNA). Namely, A (aminoacyl) site, P (peptidyl) site and E (exit) site. The A site is involved in the binding of incoming aminoacylated tRNA, the P site holds the tRNA with the nascent peptide chain, and the E site holds the deacylated tRNA before it leaves the ribosome. Both 50S and 30S subunits are involved in translocation, in which the tRNA and mRNA move through the ribosome, one codon at a time.



**Figure 2.4:** The translation process in *E. coli*. Initiation begins with the formation of the initiation complex, which includes the initiator factor, an mRNA sequence, a small ribosomal subunit, and N – formyl-methionine, which is a special initiator.

In *E. coli*, the translation mechanism includes initiation, elongation, and termination (Figure 2.4). Initiation starts with the assembly of the initiation complex, which includes a small ribosomal subunit, an mRNA molecule, initiation factors, a special initiator tRNA carrying N – formyl-methionine (fMet-tRNA<sup>fMet</sup>) and a guanine triphosphate (GTP). This is followed by the interaction between the Shine-Dalgarno (SD) sequence (6 to 9 nucleotides sequence in the mRNA, upstream transcription initiation codon, AUG) and the anti-SD sequence, at the end of the 16S rRNA. Next, the 50S ribosomal subunit attaches to the initiation complex to form a fully assembled ribosome (70S), which leads to translation elongation. During elongation, peptide bonds form between A-site tRNA and the P-site tRNA. The amino acid bound to the P-site tRNA link to the growing polypeptide chain. As the ribosome moves along the mRNA, the former P-site tRNA enters the E site, detaches from the amino acid, and is expelled. Several of the steps during elongation, including binding of a charged aminoacyl tRNA to the A site and translocation, require energy derived from GTP hydrolysis, which is catalyzed by specific elongation factors. Termination occurs when a stop codon (UAA, UAG, or UGA) is encountered and translocated into the A site. For this, termination factors bind to the ribosome, releasing both the ribosome and a new polypeptide chain. In *E. coli*, approximately 12-17 amino acids are translated per second in optimal conditions.

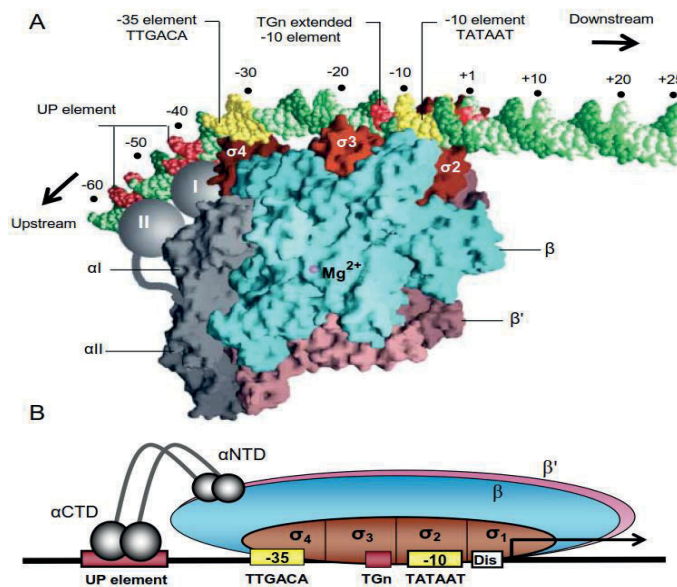
## 2.4 Bacterial Transcription

Bacteria can adapt to a wide range of different environmental conditions by changing their gene expression patterns (López-Maury et al. 2008), and, consequently, their proteome, making use of its molecular machinery for gene expression. However, since biosynthesis is expensive, the cell needs to be careful into produce the right amount (and solely this amount) of molecular species required to cope with the external conditions. Even though regulations in gene expression can take place at any stage, transcription is the primary step when cellular decisions are implemented by *E. coli* cells. In particular, most known gene expression regulatory molecules of *E. coli* interact with the promoter so as to modulate the transcription initiation process, rather than at other stages. Further, while different RNA molecules exhibit different degradation rates (Bernstein et al. 2002; Chen et al. 2015), these do not seem to be correlated with increases or decreases of RNA numbers, neither do they change significantly with other factors such as transcript length, operon length, codon composition, and G/C content.

### 2.4.1 Transcription mechanism

The enzyme that executes bacterial transcription is the DNA-dependent RNA polymerase (RNAP). This multisubunit enzyme complex has a core and a holoenzyme. The core enzyme consists of two  $\alpha$  subunits,  $\beta$  and  $\beta'$  subunit, and a  $\omega$  subunit, with a molecular mass of  $\sim 400$  kDa (Murakami & Darst 2003). Each subunit has the following size: the  $\beta'$  subunit is  $\sim 155$  kDa; the  $\beta$  subunit is  $\sim 151$  kDa; the  $\alpha$  subunits are  $\sim 37$  kDa and the  $\omega$  subunit is  $\sim 6$  kDa. Structural studies of the RNAP have revealed that it resembles a “crab- claw” (Tagami et al. 2011; Duchi et al. 2018; Browning et al. 2004) with an internal channel of 27 Å in diameter. The “two pincers of the claw” are made up of the  $\beta$  subunit and  $\beta'$  subunit and resemble a cleft. Between them, there is an active site located on the base of the channel where the  $Mg^{+2}$  ion is bound

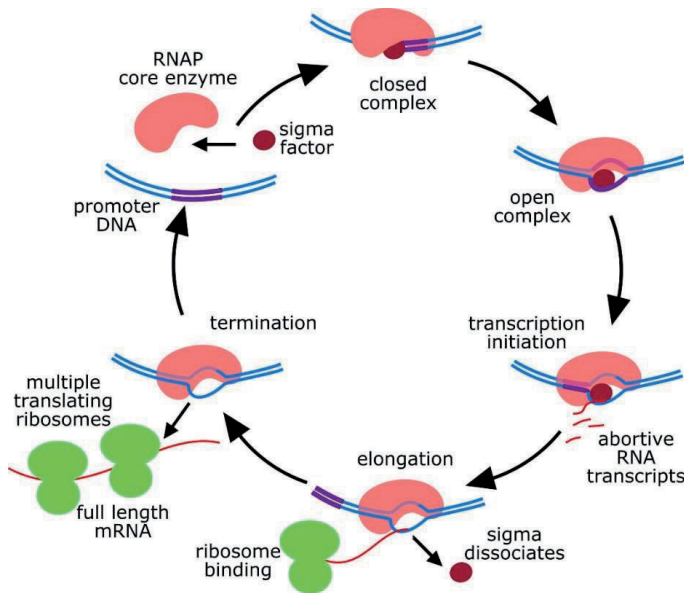
(Figure 2.5). The two  $\alpha$  subunits are located distal to the cleft and contain two domains: the C-terminal domain ( $\alpha$ CTD) and the N-terminal domain ( $\alpha$ NTD), that play distinct roles in transcription. These two domains are connected through a long flexible linker. The two  $\alpha$  subunits interact with two large subunits, triggering the formation of the core enzyme. The main function of the C-terminal domain is to interact with the upstream promoter element (Figure 2.5) and help in promoter recognition and binding, while the main function of the N-terminal domain is to gather the two large subunits together (Murakami et al. 1997). Meanwhile, the small subunit ( $\omega$ ) has no direct role in transcription, but it helps maintaining the correct formation of  $\beta'$  subunit and assists in promoting RNAP assembly (Ghosh et al. 2001; Chatterji et al. 2007; Weiss et al. 2017; Browning et al. 2004).



**Figure 2.5:** Structure of RNA polymerase and its interaction with a promoter region. A. The RNA polymerase holoenzyme complex consisting the  $\beta$  and  $\beta'$  subunits represented in blue and pink,  $\alpha$  subunits are in grey and different domains of  $\sigma$  factors are represented in red. The grey balls labeled I and II, represent the domains of  $\alpha$ CTD that interact with the promoter region. The DNA is represented in green and the -35 and -10 regions are represented in yellow. The active site of RNA polymerase is denoted by  $Mg^{2+}$ . B. Schematic representation of the RNAp holoenzyme interaction with a promoter region shown in A:  $\alpha$ CTDs interaction with UP element; the -35 element and the -10 element recognized by  $\sigma^{70}$  subdomains 4.2 and 2.4. The extended region of -10 elements is recognized by the  $\sigma^{70}$  domain 3.0. (Adapted from (Browning et al. 2004)).

Although the different subunits form a stable core enzyme, which is capable of carrying out transcript elongation, it cannot recognize the promoter sequence to initiate the transcription process. This can be achieved by binding the core enzyme with a specific factor, known as a sigma ( $\sigma$ ) factor, to form a holoenzyme form. The  $\sigma$  factors not only recognize the promoter sequence but also ensure that the binding

of the RNAP holoenzyme to the promoter is at a specific site. In *E. coli*, the most common  $\sigma$  factor that docks with the core enzyme and transcribes genes during the exponential growth phase is  $\sigma^{70}$  and belongs to the house-keeping  $\sigma^{70}$  family. The  $\sigma^{70}$  family have four different conserved domains ( $\sigma_1$ ,  $\sigma_2$ ,  $\sigma_3$  and  $\sigma_4$ ). Each domain has subdomains that interact with  $\beta'$  subunit of core RNAP and recognize the -35 region and -10 regions of the promoter sequence (Figure 2.5). The resulting holoenzyme complex binds to the different promoters with different higher affinities and regulates transcription.



**Figure 2.6:** Transcription cycle in *E. coli*. This cycle includes mainly three steps. First, the sigma factor (pink) binds to the core RNAP (red) to form a holoenzyme and finds the promoter region (blue). Next, after forming the closed complex, it forms an open complex and enters into elongation, where the sigma factor is released. Finally, when the elongation complex reaches the termination sequence, the new RNA, and the RNAP detaches from the DNA. The RNAP then binds to a sigma factor to start a new transcription event. Picture adapted from (Stracy & Kapanidis 2017).

In *E. coli*, the transcription cycle is a three steps process: Initiation, elongation, and termination (Figure 2.6). During initiation, the RNAP holoenzyme recognizes and binds to the promoter region, unwinds the double-strand DNA to form the transcription bubble (Browning & Busby 2016). Once the RNAP holoenzyme successfully escapes from the promoter region the elongation process starts, where the sigma subunit detaches from the core RNAP (Stracy & Kapanidis 2017). In elongation, the core RNAP moves along the DNA in the 3' to 5' direction, to read out the information on the DNA and to synthesize the mRNA until it reaches the termination sequence. In termination, the elongation complex disassociates into the new mRNA molecule, the DNA template and the core RNAP, which can then bind to a free sigma factor to form a holoenzyme, so as to once again start a transcription process (Nudler & Gottesman 2002).

## 2.4.2 Rate-Limiting steps in transcription initiation

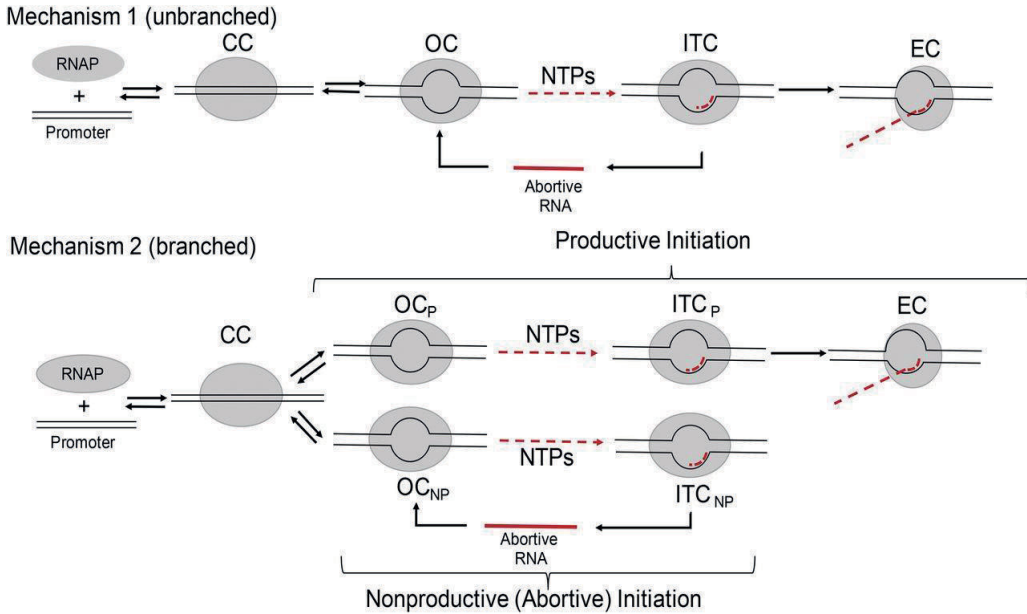
Several *in vitro* and *in vivo* studies on *E. coli* suggest that the transcription initiation is a complex, multi-step process (deHaseth et al. 1998; McClure 1985; Saecker et al. 2011; Bertrand-Burggraf et al. 1984; Browning & Busby 2016; Lloyd-Price et al. 2016; Duchi et al. 2018)(Chamberlin 1974). The multi-step nature of this process is represented in reaction 2.1. It contains three steps: promoter recognition and binding, isomerization and promoter clearance.



Reactions (2.1), include the binding and unbinding of an RNA polymerase holoenzyme complex (R) to a promoter (P) with an equilibrium rate constant ( $K_a$ ). The forward step forms a closed complex ( $RP_{CC}$ ). Due to the reversibility, these events can occur several times before the process successfully reaches the next step. The subsequent step is the formation of a stable open complex ( $RP_{OC}$ ), which is isomerized from the closed complex with a rate constant ( $K_b$ ). This step is irreversible, thus ensuring the stability of the transcription event from here onwards. Once the open complex is completed, it follows the promoter escape. In detail, the RNAP attempts to synthesize short nucleotide portions of the DNA template strand by a scrunching mechanism (that pulls off the downstream DNA into the active site, while on the promoter region (Revyakin et al. 2006; Kapanidis et al. 2006). After an initial RNA synthesis (6-8 nucleotides), the RNA polymerase holoenzyme moves forward and attempts to break contact within the promoter region and enter the elongation phase. It can do so via the *unbranched* or the *branched* mechanism (Henderson et al. 2017) (see Figure 2.7).

Most *in vitro* studies of transcription initiation focused on the *unbranched* mechanism (Hsu 2002; Gralla et al. 1980; Carpousis & Gralla 1985). In this mechanism, during the formation of the elongation complex (EC), the open complexes (OCs) undergo several cycles of synthesis of short RNA (abortive RNA) in the initial transcribing complex (ITCs) region, until a productive initiation is achieved (Straneyt & Crothers 1987; Goldman et al. 2009). This creates a stress in the initiation of transcription, which leads to release of the RNAP holoenzyme from the promoter region and begins the elongation process. The rate formation of this elongation process is expected to be much higher than the rates of other steps (CCs and OCs). Thus, in reactions 2.1, it is assumed as “negligible” for the rate of the whole process. As such, in reactions 2.1, as the time-length is much smaller than the CCs and OCs, the rate constant of this process is set to near-infinite.

Alternatively, in the *branched* mechanism, based on the transcription initiation kinetics of two promoters namely T7A1 and  $\lambda P_R$  (Susa et al. 2006), the initiation has been divided into two pathways: Productive initiation and non-productive initiation (Figure 2.7). In the productive initiation pathway, the RNAP escapes from the promoter region and enters into the elongation complex for the synthesis of long mRNA molecule, without releasing the abortive RNA. Meanwhile, in the other pathway, non-productive initiation complexes cannot escape from the promoter region and undergo several abortive cycles for the synthesis of only short RNAs (abortive RNA).



**Figure 2.7:** Depiction of the transcription initiation mechanisms in *E. coli*. In mechanism 1 (unbranched), the RNAP holoenzyme binds with the promoter region and forms a closed complex, followed by the formation of the open complex. During this step, a transcription bubble is formed, which is exposed to a short sequence of DNA. Next, a release of a short RNA synthesis (abortive RNA) occurs on the pathway to RNAP promoter escape, during the formation of the initial transcribing complex (ITC). Following an abortive initiation cycle, the RNAP enters into the transcription elongation complex (EC) to synthesize the mRNA. Meanwhile, in the branched mechanism (mechanisms 2), two classes of initiation complexes are proposed: productive complex and nonproductive complexes. Productive complexes (OC<sub>P</sub>) that are those that escape from the promoter region without releasing any short RNA sequence (abortive RNA). Nonproductive complexes cannot escape the promoter and only synthesize and release a short RNA. This picture is adapted from (Henderson et al. 2017) and reprinted with permission from PNAS.

Regulation of the rate-limiting steps of the transcription initiation process has been studied with both *in vivo* and *in vitro* methods (McClure 1980; McClure 1985; Lloyd-Price et al. 2016; Kandavalli et al. 2016)(McClure et al. 1978). For instance, from *in vitro* studies, it has been proposed that the rate of the open complex formation is derived from the time taken to reach the steady-state production of the abortive initiation products (McClure et al. 1978). Another study proposed that the rate of the closed complex formation is based on the different concentration of RNA polymerases (McClure 1985). This dependence of the closed complex formation on the concentration of RNA polymerase allows it to be distinguished, from a dynamical point of view, from the open complex formation (McClure 1980; McClure 1985). Specifically, from the direct relationship between the lag times of RNA productions and the reciprocal of RNA polymerase concentrations, it is possible to draw a Lineweaver-Burk plot (Lineweaver & Burk 1934), named as ‘tau ( $\tau$ )-plots’ (McClure 1980; Patrick et al. 2015).



From  $\tau$ -plots, the mean duration of the closed complex formation is obtained from its slope and the mean duration of the open complex formation is obtained from its intercept with the y-axis. This is possible because the duration of these steps is much longer than the time required for elementary steps to catalyze in the enzymatic reaction (McClure 1980; McClure 1985), and thus, they can be considered to be rate-limiting steps in the transcription initiation of *E. coli* genes (McClure 1980; Bertrand-Burggraf et al. 1984; Lutz et al. 2001; Buc & McClure 1985). In addition, these steps kinetics are sequence-dependent as they differ between promoters (Saecker et al. 2011).

Recently, the underlying concept of *in vitro*  $\tau$ -plots has been applied to be *in vivo* measurements to characterize the rate-limiting steps in the transcription initiation (Lloyd-Price et al. 2016; Kandavalli et al. 2016; Mäkelä et al. 2017; Startceva et al. 2019). It is worthwhile to mention that, compared with *in vitro* methods, performing these measurements in the live cells is more complex, because of limitations in the *in vivo* RNAP concentrations, as it affects cell functionality.

This method is based on the extracting the time intervals between the two consecutive RNA production events in individual cells using the *MS2-GFP* tagging system and then performing the statistical analysis of these distributions to infer the duration of rate-limiting steps in the transcription initiation. This approach has been used in several recent studies to characterize the underlying steps of transcription initiation of various promoters in different environmental conditions (Lloyd-Price et al. 2016; Kandavalli et al. 2016; Mäkelä et al. 2017; Startceva et al. 2019)(Oliveira et al. 2016). For example, when studies are conducted at an optimal temperature, the model that best fits the empirical data contains two main rate-limiting steps (associated with the closed and open complex formation), while lowering the temperature to 24<sup>0</sup>C, it has been reported the emergence of a third rate-limiting step. It has been hypothesized that it due to increased duration of an isomerization step that occurs before the completion of the open complex (Muthukrishnan et al. 2012). The results of this study suggest that the dynamics of transcription initiation could be explained by its multi-rate limiting steps, in agreement with the *in vitro* studies (Buc & McClure 1985).

Another *in vivo* technique, fluorescence in situ hybridization (FISH), has also been used to measure quantitatively the kinetics of transcription. Studies using this technique reported that the mean and variability of mRNA numbers in cell populations is dictated by transcription initiation (Jones et al. 2014; So et al. 2011).

Overall, the above-mentioned studies suggest that the mean rate and variability in transcription are promoter sequence-dependent and, thus are evolvable and that the regulatory molecules in the promoter region can accelerate or hinder the durations of underlying steps and, thus, they are adaptive. Furthermore, the kinetics of these steps are influenced by DNA supercoiling and other environmental factors, such as temperature. In all Publications, to characterize the rate-limiting steps in the transcription initiation process, we conducted measurements of time intervals between the two consecutive RNA production events in individual cells and applied best-fitting stochastic models.

### 2.4.3 Transcription elongation

Transcription elongation is the second step in the transcription process. This phase starts as soon as the RNAP clears the promoter region. During this phase, the  $\sigma$  factor detaches from the RNAP holoenzyme. The core RNAP, the template DNA strand, and the nascent mRNA forms the elongation complexes. This Elongation complex has no specific affinity towards the DNA template strand and advances on the template strand in a slide-like movement (Gusarov & Nudler 1999). Studies suggest that, on average, genes producing mRNA have a transcription elongation rate of 30 to 50 nucleotides per second (Murakawa et al. 1991; Vogel & Jensen 1994; Greive et al. 2005; Proshkin et al. 2010; Larson et al. 2011). This rate is higher (approximately 80 nucleotides per second) in genes producing ribosomal RNA (rRNA) (Dennis et al. 2009).

The movement of the elongation complex along the DNA template is not a continuous process, instead, it exhibits transcription pauses (Gabizon et al. 2018; Kireeva & Kashlev 2009) or arrest or backward diffusion on the DNA (known as backtracking) (Greive & Hippel 2005). It was shown that the pausing can significantly affect transcription elongation rates (Gabizon et al. 2018) as pauses can last from seconds to minutes (Herbert et al. 2010; Landick 2009). These pauses can be categorized as short and long-lived pauses, that can be further stabilized by RNAP backtracking (Komissarova & Kashlev 1997; Artsimovitch & Landick 2000) or by the formation of a nascent RNA hairpin structure (Wilson 1995; Artsimovitch & Landick 2000)(Landick 2006). In addition, transcription elongation factors and other DNA sequences are known to bound the DNA and obstruct the movement of RNAP, as such affect the dynamics of pausing (Uptain & Kane 1997). For example, NusA and NusG are transcription elongation factors that can increase (Yakhnin et al. 2016) and decrease (Herbert et al. 2010; Burmann et al. 2010) the transcription elongation rate and pause states.

Overall, in transcription pausing, the RNAP halts synthesis of RNA transcripts, but not release them neither it aborts synthesis of RNA. Aside from pauses, other pathways such as pyrophosphorolysis, editing, and premature termination can also occur in transcription elongation (Arndt & Chamberlin 1988; Erie et al. 1993; Kane et al. 1991).

### 2.4.4 Transcription termination

Transcription termination is the final step in the transcription process. Their location in the sequence demarcates gene boundaries and can be targets for regulation (Santangelo & Artsimovitch 2011). At the end of this process, newly formed mRNAs disassociate from the template DNA and the RNAP detaches from the DNA. In *E. coli*, transcription termination is carried out by one of the two mechanisms: intrinsic termination or Rho-dependent termination (Santangelo & Artsimovitch 2011; John P Richardson 1991).

Intrinsic termination occurs when the emerging RNA forms a hairpin loop stimulated by signals encoded within the nascent RNA. These signals are generated in the guanine – cytosine-rich region followed by approximately eight uridines (U stretch) at the 3'terminus (Gusarov & Nudler 1999). When the RNAP reaches the U stretch, it halts transcription and the nascent RNA folds and forms a stem-loop structure.

The formation 'RNA-DNA duplex' in the U stretch region is not stable as the bond between the uracil and adenine is weak. The weak adenine-uracil bond lowers the DNA-RNA stabilization energy and allows it to unwind and detach the nascent RNA and RNAP from the transcription elongation complex (Martin & Tinoco 1980; Arndt & Chamberlin 1988).

In Rho-dependent termination, the Rho protein, which belongs to the helicase family, unwinds the DNA-RNA duplex at the 5' end of the nascent RNA strand (Koslover1 et al. 2012; Hollands et al. 2014). The Rho termination factor has two main domains: RNA-binding domain and ATP binding domain. The Rho factor employed by a part of nascent RNA, which is rich in cytidine residues, moves along the nascent RNA in 5' to 3' direction following RNAP. This energy movement process involves ATP hydrolysis of the ATP binding domains of the Rho factor. Once the RNAP reaches the terminator, the Rho factor binds to it and unwinds the DNA-RNA duplex, followed by the release of the RNAP, nascent RNA and the Rho factor from the DNA template strand (Richardson 2002).

## 2.5 Gene Regulation at the Transcription Level

Bacterial cells constantly face challenging environmental conditions such as stress, temperature shifts, etc. To survive, they regulate gene expression to produce a specific amount of essential and functional proteins, at specific moments during their lifetime (López-Maury et al. 2008).

In *E. coli*, transcription is the step where more control is exerted (McClure 1985; Chamberlin 1974; Browning & Busby 2016; Browning et al. 2004; Rosenberg & Court 1979), while the degradation of RNA and proteins are kept nearly constant rates (Bernstein et al. 2002; Chen et al. 2015; Goldberg 1972). *E. coli* evolved several mechanisms to control the steps in the transcription initiation (William S Reznikoff et al. 1985; Browning & Busby 2016).

### 2.5.1 Promoter region

In *E. coli*, a promoter region is defined by a highly conserved consensus sequence (-10 and -35 position) which is upstream of the transcription start site (Harley & Reynolds 1987). This sequence is recognized by an RNAP holoenzyme, which binds to it and starts transcription initiation (Hippe et al. 1984; William S Reznikoff et al. 1985). In this region, transcription factors such as activators, repressors, etc., can bind and up-regulate or downregulate transcription either by interacting with the RNAP or by binding to the DNA and change DNA conformation. Also, the affinity of RNAP binding is affected by the promoter sequence itself (Brewster et al. 2012), thus affecting the rate of formation of the closed complex. Since the DNA conformation is also sequence-dependent, the promoter sequence also affects the rate of open complex formation. Thus, the promoter region plays a crucial role in the regulation of transcription initiation (William S Reznikoff et al. 1985).

Although such sequence-specific regulation is relevant for transcription, it only provides static regulation, as it cannot be tuned according to, among other, environmental conditions. Dynamic regulation, e.g. based on the environmental condition, requires, e.g. small ligands (e.g. ppGpp),  $\sigma$  factors (Kandavalli et al. 2016; Mauri & Klumpp 2014), the intracellular concentration of RNAP, all of which interacting with the promoter region as well, providing additional adaptability to *E. coli*.

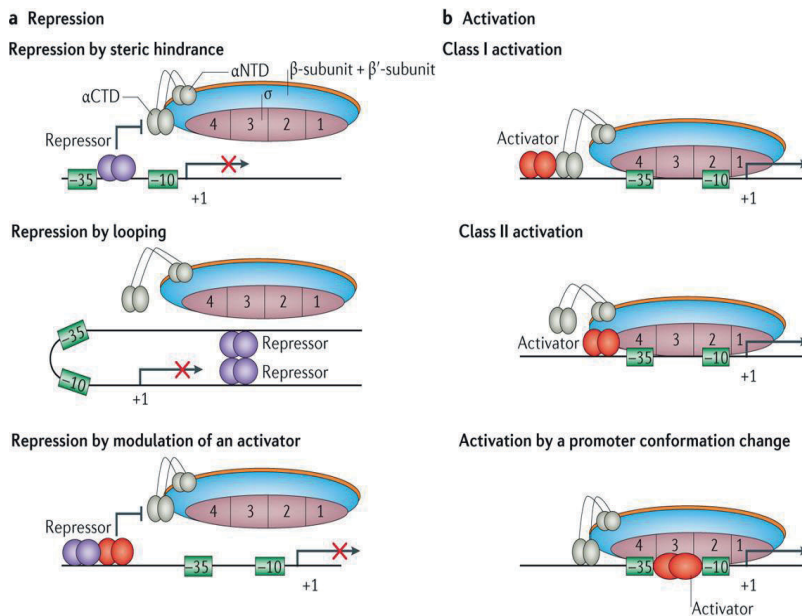
## 2.5.2 Regulation by transcription factors

The concept of regulating transcription initiation by transcription factors (TFs) was originally proposed in 1961 with the introduction of the Jacob and Monod operon model. An operon consists of a promoter, an operator and, structural genes. In addition, there are associated regulatory genes, located at some distance from the operon. These regulatory genes encode for the activator or repressor proteins that bind to an operator and serve as on-off switches of the gene activity. Usually, transcription factors are expressed in response to changes in, e.g., environmental conditions. It has been reported that, in *E. coli*, there are more than 300 genes encoding for transcription factors (Pérez-rueda & Collado-vides 2000). Most for transcription factors target specific sequences located in the promoter-operator region. However, a few TFs act as global regulators, by recognizing and interacting with many promoters. In *E. coli*, there are at least seven TFs that act as global regulators: CRP, FNR, IHF, Fis, ArcA, NarI, and Lrp. Combined, they affect the activity of more than 50 % of all promoters (Agustino Martínez-Antonio & Collado-vides 2003).

In general, TFs have two domains. One receives an internal/external signal, while the other directly interacts with the DNA (Babu & Teichmann 2003), leading to a modification of a gene's expression rate. This modification consists of accelerating or decreasing RNAP affinity with the promoter region. This effect depends on the promoter's architecture, i.e. the location of its binding sites and their affinities. In some cases, the TF action can change from activator to repressor, or the opposite, following a change in architecture (Pérez-rueda & Collado-vides 2000). For example, in the *gal* operon, CRP acts as an activator of the *gal* P1 promoter and as a repressor of the *gal* P2 promoter (E.Mussoa et al. 1977; Lewis & Adhya 2015).

In *E. coli*, the most common means of control of promoter activity is repression (Garcia et al. 2010). Several repression mechanisms have been found (Figure 2.8 a). In some promoters, repression occurs by steric hindrance, as the operator region overlaps with the consensus sequence (-35 and -10 element region) which the RNAP recognizes and binds to. Repressor binding to that region prevents RNAP recruitment. A classic example of this mechanism is the binding of the lacI repressor to the lac promoter, as it blocks the interaction of RNAP with the promoter region (Muller-Hill 1998). In some cases, the operator region is located upstream and/or downstream of the promoter region. It acts upon binding of the repressor to the operator site, by formation of a DNA loop (known as DNA looping mechanism), which blocks access of the RNAP to that region (Schleif 2010; Choy et al. 1995; Browning & Busby 2016). In some promoters, repressor proteins (acting as anti-activators) act by preventing the binding of activators (Browning & Busby 2016).

Activators act as positive regulators of transcription by actively recruiting RNAP to the promoter region, enhancing the transcription process. Similar to repression, there are many mechanisms of activation of transcription initiation (Lee et al. 2012; Browning & Busby 2016; Browning et al. 2004). They are divided into 3 categories: Class I activation, Class II activation and, activation by a promoter conformational change.



**Figure 2.8:** Schematic representation of repression (a) and activation (b) mechanisms of promoter activity using transcription factors. This image is adapted from (Browning and Busby, 2016) and reprinted with permission from Macmillan Publishers Ltd: [Nature Reviews Microbiology]

In class I activation, the activator binds to an operator site located upstream of the -35 element region of the promoter and then recruits the RNAP by interacting with the RNAP subunits. For example, in the *E. coli* lac operon, the cyclic adenosine monophosphate (cAMP) receptor protein (CRP) act as an activator, as it recruits RNAP by direct interaction with the C-terminal domain of the  $\alpha$ -subunit (Figure 2.8 b) (Ebright 1993)(Browning et al. 2004). In class II activation, the operator site overlaps with the -35 element region of the promoter. Once the activator binds to the target site, it recruits the RNAP by interacting with the domain four of the RNAP  $\sigma$  subunit (Ebright & Busby 1995; Lee et al. 2012), such as MarA and SoxS (Martin et al. 2002). Finally, some activators assist transcription without directly interacting with the RNAP. These activators bind to the DNA and alter promoter conformation, which increases the binding affinity of RNAP, thus increasing the rate at which the transcription process initiates (Sheridan et al. 1998).

The functionality of the TFs can also be altered by specific molecules, known as the inducers. These alter the expression rate of inducible genes by binding to its repressor, rendering them inactive. For example, in lac operon, the action of LacI as a repressor can be rendered ineffective by Lactose, as its binding reduces the binding affinity of LacI to the operator site, thus indirectly enhancing recruitment of RNAP to the promoter region (Lewis 2005). Studies show that, instead of lactose, Isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG), a molecular mimic of lactose, can be used instead. This molecule is preferred in laboratory studies as it is a synthetic substrate and cannot be metabolized. Due to this, its concentration remains constant during the course of the experiments (Lutz & Bujard 1997; Marbach & Bettenbrock 2012).

### 2.5.3 Regulation by $\sigma$ factors

$\sigma$  factors play a crucial role in the kinetics of transcription initiation (Mauri & Klumpp 2014; Kandavalli et al. 2016).  $\sigma$  factors can recognize specific promoter sequences and, once doing so, start transcription, e.g. following environmental shifts (Hengge-aronis 2002b; Hengge-aronis 2002a). In *E. coli* there are seven  $\sigma$  factors, namely  $\sigma^{70}$ ,  $\sigma^{38}$ ,  $\sigma^{54}$ ,  $\sigma^{24}$ ,  $\sigma^{32}$ ,  $\sigma^{19}$ , and  $\sigma^{28}$  named in accordance with the molecular weight (kDa) of the proteins. The primary  $\sigma$  factor is  $\sigma^{70}$ , as it control most active genes during the exponential phase (Tripathi et al. 2014). The expression of other  $\sigma$  factors occurs in unfavorable environmental conditions (Hengge-aronis 2002b). For example, upon heat shock,  $\sigma^{32}$  is expressed to control a specific set of otherwise largely inactive genes (Hengge-aronis 2002a; Arsene et al. 2000). Another  $\sigma$  factor,  $\sigma^{38}$ , also known as RpoS, is a master up regulator of stress response genes (Battesti et al. 2011; Ishihama 2000). The main functions of each  $\sigma$  factor in *E. coli* are listed in Table 1.

**Table 1:** List of *E. coli*  $\sigma$  factors, genes responsible to produce it, and function in the cell.

$\sigma$ factor	Gene producing the $\sigma$ factor	Functions or regulation
$\sigma^{70}$	rpoD	Housekeeping genes
$\sigma^{38}$	rpoS	Stress response and stationary phase genes
$\sigma^{54}$	rpoN	Nitrogen response genes
$\sigma^{32}$	rpoH	Heat shock response genes
$\sigma^{19}$	fecI	Ferric citrate uptake
$\sigma^{24}$	rpoE	Extracytoplasmic function
$\sigma^{28}$	fliA	Flagellar genes

Recent studies suggest that  $\sigma$  factors compete for a limited pool of RNAP core enzymes. This causes variability in RNAP holoenzyme distributions and, thus, modulates genome-wide transcription kinetics (Mauri & Klumpp 2014; Grigorova et al. 2006; Kandavalli et al. 2016). Specially, the number of core RNAP enzymes is limited, as it ranges from 3000 to 13000, depending on the growth conditions (Grigorova et al. 2006; Klumpp & Hwa 2008). Similarly, the intracellular numbers of the housekeeping  $\sigma^{70}$  ranges from 5000 to 17000 molecules in an exponential growth phase, while alternative  $\sigma$  factors vary in their number, based on the type of stress it responds to (see Table 1). In general, the number of core RNAP is always

smaller than the number of  $\sigma$  factors in a cell. When there are two or more species of  $\sigma$  factors present, they need to compete for binding with core RNAP to form an RNAP holoenzyme complex.

Aside from the numbers of  $\sigma$  factors, another variable that determines the holoenzyme formation is binding affinities of  $\sigma$  factors to core enzymes. Each  $\sigma$  factor has a specific binding affinity to the core RNAP and each holoenzyme has specific recognition of promoter sequences. For example, evidence from transcription assays suggests that the promoter with consensus sequences (TATAAT) on the -10 element region are recognized by holoenzymes carrying  $\sigma^{70}$  or  $\sigma^{38}$  (Gaal et al. 2001; Tanaka et al. 1995). Likewise, holoenzymes carrying  $\sigma^{54}$  can only recognize the promoter consensus sequence at -12 and -24 element region ((Zhao et al. 2010). Due to this diversity of promoter recognition by  $\sigma$  factors, any changes in the number of holoenzymes will lead to genome-wide modulation in gene expression. For instance, while most genes in *E. coli* are transcribed by the RNAP holoenzyme carrying  $\sigma^{70}$ , in the stationary phase the expression of these genes is downregulated, due to the appearance of other  $\sigma$  factors (Farewell et al. 1998; Dong & Schellhorn 2009). In **Publication II**, we identified mechanisms explaining how these genes are downregulated in this fashion, by studying the dynamics of the transcription initiation as a function of  $\sigma$  factor numbers.

Aside from the role of  $\sigma$  factors in transcription initiation, studies have shown that  $\sigma$  factors can also influence the rate of elongation by interacting with the elongation complex (Kapanidis et al. 2005; Harden et al. 2016).

## 2.5.4 Other regulatory factors

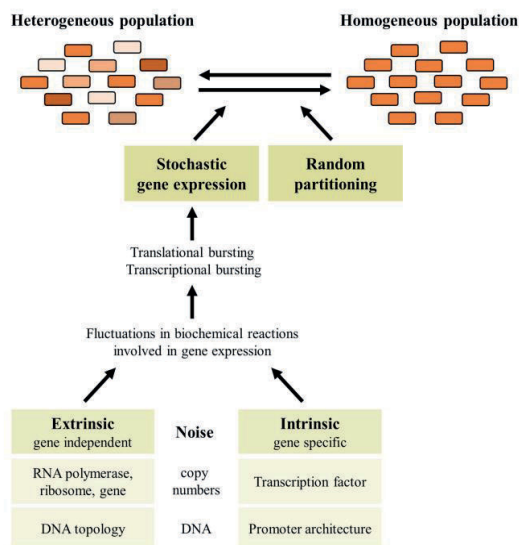
In *E. coli*, aside from the TFs and  $\sigma$  factors, there are many more factors regulating the activity of genes in the chromosome, such as ligands, nucleoid-associated proteins (NAPs), DNA supercoiling, etc (Holmes & Cozzarelli 2000; Browning et al. 2004). For instance, small ligands such as guanosine 3', 5'-diphosphate (ppGpp), an alarmone to stringent response, interacts with the RNAP and down-regulates the expression of genes responsible for stress (Ross et al. 2013). Also, structural analysis on a ppGpp-RNAP complex shows that ppGpp binds close to the active site of the RNAP and inhibits transcription initiation (Ross et al. 2013; Artsimovitch et al. 2004). Aside from inhibition, ppGpp also enhances the expression of genes responsible for proteins are required for amino acid biosynthesis and transport (Ross et al. 2013; Paul et al. 2005).

Meanwhile, NAPs including H-NS, Fis, and HU proteins, are responsible for DNA compaction in *E. coli* (Dillion and Dorman 2010). The folding of DNA by NAPs affects the distribution of RNAP on promoters. These proteins can also act as the regulators in transcription. For example, H-NS acts as a global repressor by binding to AT-rich region of DNA sequence, irrespective of the sequence, and silencing transcription activity (Navarre et al. 2006; Browning et al. 2009). In other cases, NAPs also act as activators, e.g., Fis proteins (Dillon & Dorman 2010). Further, DNA compaction is also affected by topological constraints in the structure of the chromosome (Rovinskiy et al. 2012).

## 2.6 Transcription Noise

In *E. coli*, the transcription process is a series of biochemical reactions, where reactant molecules exist in very few copies in a cell (Xie et al. 2008). Thus, fluctuations in their numbers due to random biochemical processes cause noise that can significantly impact transcripts numbers (Kærn et al. 2005). This acts as a source of phenotypic variability, even in genetically identical cells with the same histories of environmental exposure (Elowitz et al. 2002; Eldar & Elowitz 2010; Bury-Mone & Sclavi 2017). Noise in gene expression can be advantageous in fluctuating environmental conditions (Acar et al. 2008; Raser & O’Shea 2005).

To identify sources of noise-generating variability in gene expression products, studies have classified noise as Intrinsic or Extrinsic noise (Elowitz & Leibler 2000; Elowitz et al. 2002). In one study, to measure the intrinsic and extrinsic noise, the authors constructed a dual reporter system (YFP and CFP) under the control of identical promoters (Elowitz & Leibler 2000). The measured relative difference in fluorescence intensity between cells is termed as intrinsic noise and the measured correlation between the two fluorescence intensities in individual cells is termed as extrinsic noise (Elowitz & Leibler 2000).



**Figure 2.9:** Schematic representation of how various sources of noise affect the stochastic gene expression in a clonal population of bacteria. This image is obtained from (Engl 2018) and reprinted with permission from by Portland Press Limited: [Biochemical Society Transactions].



In general, intrinsic noise is gene-specific and arises from the inherent randomness of chemical processes between small numbers of molecules. Extrinsic noise affects gene expression product numbers in a non-specific manner and arises from cell to cell variability in RNAP, ribosomes, and gene copy numbers, etc., (Figure 2.9).

Using single-molecule sensitivity techniques, researchers have quantified cell-to-cell variability in mRNA and protein numbers (Golding et al. 2005; Yu et al. 2006; So et al. 2011; Jones et al. 2014; Skinner et al. 2013). One study measured the real-time mRNA kinetics of a synthetic  $P_{lac/ara-1}$  promoter in individual *E. coli* cells and proposed that transcription occurs in bursts, even in fully active genes (Golding et al. 2005). Other studies also observed the dynamics of protein production and suggested that it also occurs in bursts of different sizes (Yu et al. 2006). These bursts contribute to noise in gene expression (Sanchez & Golding 2013).

Additionally, studies proposed that the diversity in mRNAs and proteins is also due to fluctuations in molecular species, such as ribosomes, RNAP, TFs,  $\sigma$  factors, among others, involved in transcription and translation (Engl 2018; Jones et al. 2014; Brewster et al. 2014; Yang et al. 2014). Other studies suggested that other mechanisms not directly involved in gene expression can also contribute to observable variabilities, such as from DNA replication, DNA supercoiling, and condensation, partitioning cell division, etc. (Peterson et al. 2015; Chong et al. 2014; Huh & Paulsson 2011).

In **Publication III**, we investigated how the variability in intrinsic and extrinsic sources affect the kinetics of transcripts production in individual *E. coli* cells and cell lineages.

## 3 MATERIALS AND METHODS

This chapter provides an overview of the experimental and theoretical approaches for studying transcription initiation dynamics, with emphasis on the ones used in this thesis. These include the fluorescent proteins and microscopy, single-molecule techniques for RNA detection, and alternative methods of validation. Finally, we present theoretical approaches used to model and simulate the biological systems, with emphasis on stochastic simulation methods.

### 3.1 Basics of Microscopy and Fluorescent proteins

The discovery of a green fluorescent protein (GFP), which was isolated from the jellyfish *Aequorea victoria* (Shimomura 1962), along with its gene fusion (C. Prasher et al. 1992; Tsien 1998) has revolutionized the field of cell biology. Due to their versatility, specificity and quantitative capabilities for the live-cell imaging, fluorescent proteins fused to their target proteins have become crucial tools in cell biology experiments and have allowed the emergence of single-cell studies of gene expression.

Molecular cloning methods of fusing the fluorophore moiety to a protein of interest (e.g. an enzyme or an RNA polymerase subunit), have helped researchers to monitor cellular processes in living systems. Due to rapid evolution, fluorescent proteins can now cover the visible spectral wavelengths. Also enhanced were properties such as maturation and degradation times, folding, oligomerization, brightness, and photostability. Also, these improvements have helped to perform multicolor imaging of any protein (Shaner et al. 2004) and allow us to study the subcellular architecture at the sub-nano second-time resolution (Tsien 1998). Overall, fluorescent proteins have become essential tools in areas ranging from studies of the complex behavior of single-molecules, of the internal dynamics of the molecular process, quantitative studies of gene expression dynamics, etc. (Yu et al. 2006; Stracy & Kapanidis 2017; Golding et al. 2005).

During the last two decades, this field also has produced advanced fluorescent probes with unique characteristics, such as photoactivation and photoconversion (Daya & Davidson 2009; Wu et al. 2011). These modified fluorescent proteins can be switched on and off or, be converted from the non-fluorescent state to the stable fluorescent state in response to light at an appropriate wavelength. These works have motivated many advances in microscopy imaging techniques, e.g., super-resolution microscopy, that enables the monitoring of inner components of cells in great detail.

Some factors have to be carefully considered when conducting imaging experiments. For example, the brightness of fluorescent proteins should be above the cellular background level, their photostability needs to be robust, and there should be minimum cross-talk between the emission and excitation spectra. Also, when fused with the target protein, the effect of the fluorescent protein on the native protein functionality should be as weak as possible (Shaner et al. 2004).

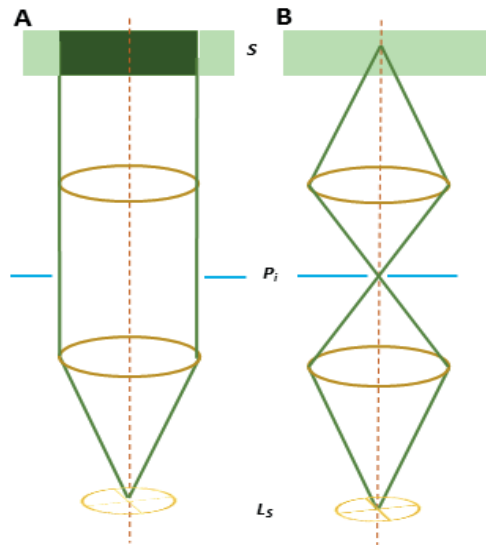
Although fluorescent proteins have several advantages, drawbacks of using them include fluctuations in the fluorescent intensity (Ha & Tinnefeld 2012) due to changes in the environmental conditions and photostability. For example, some wild-type GFP is sensitive to temperature, while Yellow fluorescent proteins (YFP) are sensitive to pH and chloride (Wachter & Remington 1999). To overcome this situation, several altered or mutated versions of fluorescent proteins (Shaner et al. 2004) ( see Figure 3.1) has been developed, in order to improve stability, folding and sensitivity to environmental conditions.



**Figure 3.1:** Purified fluorescent proteins shown in visible light. These fluorescent proteins from left to right (mHoneydew, mBanana, mOrange, tdTomato, mTangerine, mStrawberry, and mCherry) were derived from the *Discosoma* sp. red fluorescent protein. This image is extracted from (Shanner et al. 2004) with permission from the Nature publishing group.

In addition, when performing live-cell imaging, one common goal is to monitor cellular dynamics, which requires fast image acquisition. This demands short exposure times, meaning that fluorescent proteins that absorb and emit light have to be significantly bright than the cellular background (Ha & Tinnefeld 2012). This requests a highly sensitive detector and a bright light source. Further, the numerical aperture needs to pay attention when choosing the objective.

The most common optic system used in the illumination of fluorescence microscopy is a wide-field epi-illumination. In this optic system, the entire area of the sample is exposed to light either from above (in a standard upright microscope) or from below (in an inverted configuration). It excites the incident lamp excitation light of an area of  $\sim 10 \times 10 \mu\text{m}^2$  (Webb & Brown 2012). Thus, the volume illuminated is quite large, causing out-of-focus fluorescent molecules to contribute to the background fluorescence signal. Wide-field epi-illumination has a wide range of applications in bacterial studies, such as monitoring the dynamics of fluorescently tagged RNA molecules (Golding et al. 2005) and protein molecules (Yu et al. 2006). On the other hand, there are several limitations such as low resolution, excess of out of focus fluorescent signal and, photobleaching of the sample.



**Figure 3.2:** Lightpath illustration of wide-field epi-illumination and confocal microscopy. In confocal microscopy (B), light from the light source ( $L_s$ ) is focused through a pinhole for illumination ( $P_i$ ) and subsequently passes through a sample (S), resulting in a small focal volume. In wide-field epifluorescence microscopy (A), the entire sample volume is exposed to light.

To avoid such limitation, several other methods have been developed, including Confocal microscopy (Pawley 2006), Total Internal Reflection (TIRF) microscopy (Fish 2015), and Highly Inclined and Laminated Optical (HILO) sheet microscopy (Tokunaga et al. 2008). The primary goal of this microscopy is to eliminate excess of out of focus fluorescent light during the imaging process by restricting the illumination volume of the sample.

In confocal laser scanning microscopy, the light source for exciting the fluorescence molecule comes from the laser unit, and it is targeted to a region of interest. It can be used to obtain optical sections through a sample to exclude out of focus and background fluorescence (Figure 3.2). This is achieved with a pinhole aperture where excited light passes through a focal volume of a sample (Pawley 2006). The drawback of this optical system is the slowness of point scanning image acquisition, which restricts the area of the image. This speed can be increased by using spinning disc confocal microscopy that illuminates multiple regions of the sample and minimizes photobleaching or phototoxicity (Frigault et al. 2009; Nakano 2002).

When comparing TIRF microscopy with confocal microscopy, a better optical section of the sample is illuminated. TIRF uses an evanescent wave, which is generated when the incident light is reflected at the interface of two transparent media with different refractive indices. TIRF allows to selectively illuminate and excite the fluorophores in a restricted region of the sample. As the energy of the evanescent wave field decreases exponentially with distance from the interface, the only fluorophore at a certain distance from coverslip is excited, which allows creating images with an outstanding signal-to-noise ratio. Also, TIRF can illuminate the region of the sample with an outstandingly high axial resolution, below 100nm. As a

result, it can only probe molecules close to coverglass surfaces, e.g., membrane-associated molecules. To illuminate the region deeper than the TIRF range, HILO microscopy was developed (Tokunaga et al. 2008). HILO is generated by intense laser illumination, angled through a high numerical aperture objective to a sample, resulting in lower out-of-focus light signal (Tokunaga et al. 2008).

The methodologies above allow monitoring of fluorescent molecules *in-vivo* and *in-vitro*. Some of our studies also require the visualization of high contrast images of transparent live cells. Such images were acquired by phase-contrast microscopy. It employs an optical mechanism that converts minute differences in a phase into corresponding variations in amplitude, which can be seen as a difference in image contrast (Zernike F 1942). Phase-contrast microscopy enables to examine live cells, without exposing them to laser or staining dyes. It is one of the few methods available to quantify cell structure, shape, and size.

### 3.2 Single-molecule methods for quantifying transcription dynamics

Most knowledge of transcription was gathered from biochemical and biophysical studies conducted using *in vitro* techniques (McClure 1985; deHaseth et al. 1998). One classic example is the study to identify the binding region of RNAP to the sequence of DNA (Ishihama 2000). Another is the study that identified DNA-protein binding interactions, both studies used foot printing, a method based on gel electrophoresis.

However, during the last decade, the development of *in vivo* techniques with real-time observation has allowed to characterize and dissect transcription in the context of a living cell. In that sense, single-molecule studies have remarkably provided more advanced biological information. Also, they allow monitoring the spatial localization of macromolecules and other cellular components in live cells. These includes RNA (Golding & Cox 2004; So et al. 2011; Muthukrishnan et al. 2012; Pitchiaya et al. 2014; Lenstra et al. 2016), proteins (Yu et al. 2006; Taniguchi et al. 2010), RNAP (Bakshi et al. 2012; Stracy et al. 2015), ribosomal subunits (Sanamrad et al. 2014), transcription factors (Leon et al. 2017), plasmids (Reyes-lamothe et al. 2014), etc. This is made possible by the usage of photoactivatable or photoconvertible fluorophores fused to the target molecule. To detect accurately the fluorescent molecule, the target molecule must express at low concentration. Various strategies have been employed to lower the target molecules (Pitchiaya et al. 2014; Yu et al. 2006; Santangelo et al. 2009; Huang et al. 2009). This involves the use of super-resolution microscopy techniques, such as Photo-activation localization microscopy (PALM) and Stochastic optical reconstruction microscopy (STORM) which can achieve up to 20 nm spatial resolution (Xu & Liu 2018).

In recent years, several methods have been developed to probe RNA molecules with fluorescent proteins. RNA labeling can be done in two ways: direct and indirect. Direct RNA labeling involves the usage of a chemically reactive functional group or structural motifs present in RNA and RNA modifying proteins, for fluorophore conjugation (Pitchiaya et al. 2014). Conversely, indirect labeling methods involve sequence-based complementary hybridization of RNA labels carrying a fluorescent protein with multiple specific RNA motifs (Pitchiaya et al. 2014; Raj & Oudenaarden 2008)(Levsky & Singer 2003). Indirect

RNA labeling is more popular due to the ability to tag and detect different endogenous RNAs, as well as exogenous RNA (Raj & Oudenaarden 2008).

One method using indirect labeling of RNA is fluorescence in situ hybridization (FISH). This method uses fluorescent probes that specifically bind to parts of nucleic acid with a high degree of complementary sequence (Raj & Oudenaarden 2008). Advantages of FISH allows to detect multiple RNAs at the same time and measure their spatial localization. In addition, it can quantify cell-to-cell variability in endogenous RNAs (Raj & Oudenaarden 2008; Llopis et al. 2010), which is not possible with methods such as qPCR and RNA-Seq.

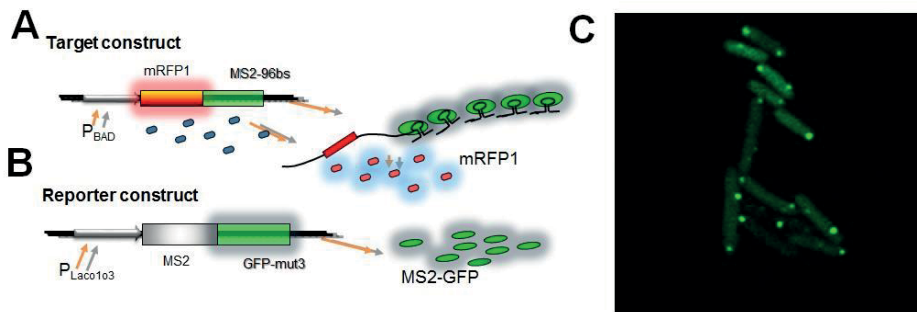
Although this method provides single-molecule sensitivity in individual cells, it lacks information on spatial and temporal resolution. For example, as it involves the fixation of cells, probe hybridization to the target RNA sequence, permeabilization of the cell membrane and extensive washing of cells to remove the unbound probes (Gasnier et al. 2013), it cannot monitor individual transcription events (Huber et al. 2018), which assists in the study of the *in vivo* dynamics of transcription, in real-time.

### 3.2.1 MS2-GFP tagging system

The MS2-GFP system is one of the most sensitive real-time single-molecule methods that allow studying the *in vivo* dynamics of the transcription process at the single-cell level. This method was initially developed by Robert Singer and co-workers to visualize RNA in higher eukaryotic cells (Bertrand et al. 1998). Later modifications allowed its use in bacteria (Golding et al. 2005; Golding & Cox 2004). This method allows tracking RNA molecules inside live cells, as soon as they appear.

The MS2-GFP system involves the expression of two components: (i) fusion of the RNA bacteriophage MS2 coat protein to a fluorescent protein, which allows it to bind specifically and (ii) a target RNA containing tandem repeats of the MS2 stem-loop sequences. These components can be genetically engineered either into a plasmid and transformed into cells or, it can be integrated into the genome of host cells. The two components are illustrated in Figure 3.3.

In detail, the MS2 coat protein is derived from the native bacteriophage, which binds with high specificity to 19 to 21 nucleotides of RNA stem-loop structure containing the initiation codon of the phage replicase gene (Bernardi & Spahr 1972). Upon binding to a unique site in the RNA genome of the phage, the coat protein represses translation of the RNA replicase gene and guides packaging into phage particles (Peabody 1993; Querido & Chartrand 2008). Over the years, the MS2 coat protein has been engineered to fluorescent fusion proteins and bind to any RNA that has specific stem-loop sequences or motifs. Such RNAs can be used for the study of various cellular process in different organisms (Golding et al. 2005; Lenstra et al. 2016a; Fusco et al. 2003).



**Figure 3.3:** Schematic overview of the MS2-GFP component system. (A) The target construct carrying mRFP1 fluorescent proteins followed by the 96 binding sites for the detection of the RNA by MS2-GFP proteins. The target construct is under the control of the PBAD promoter, which is inducible by Arabinose. (B) A reporter construct is responsible for the expression of MS2 GFP molecules (green balls), which is under the control of promoter  $P_{Lac03}$ . Once the target constructs producing the RNAs, the MS2-GFP molecules bind to it, allowing it to visualize as a cluster of GFPs. (C) Example confocal image of *E. coli* cells expressing both target RNAs and reporter MS2-GFP molecules. Individual RNA molecules appear as bright spots when visualized by confocal microscopy. The background of the cells is due to the unbound distribution of MS2-GFP molecules in the cells' cytoplasm.

The use of the MS2-GFP system in live *E. coli* cells allows to monitor the patterns of RNA localization and to study the transcription events inside the cells with single-molecule sensitivity (Golding & Cox 2004; Muthukrishnan et al. 2012; Mäkelä et al. 2013). To determine the transcription dynamics of a target gene, the reporter constructs containing the MS2 coat protein fused with GFP has to be highly expressed before the target RNA is produced. The high intracellular concentration of MS2-GFP protein guarantees that enough will bind to the target RNA containing the binding motifs, as soon as they produced. The specific binding of multiple MS2-GFP proteins to the same target RNA, create brighter fluorescent than the unbound MS2-GFP, freely diffusing inside the cell.

When visualizing the cells containing the MS2-GFP system under the confocal microscope, the target RNA bound by multiple MS2-GFP fused proteins appear as a bright spot (See Figure 3.3 C), that moves slowly inside the cells (Golding & Cox 2004; Muthukrishnan et al. 2012; Mäkelä et al. 2013). Since the target RNA is coated by MS2-GFP proteins, it is protected from natural degradation (Fusco et al. 2003). Also, the intensity of the fluorescent spots does not decrease during the measurement time (Tran et al. 2015; Kandavalli et al. 2016; Muthukrishnan et al. 2012).

Apart from the MS2-GFP system, other viral proteins have been used to tag and detect the target RNA, such as PP7 proteins, derived from PP7 bacteriophage (Larson et al. 2011; Lenstra et al. 2016b), and the  $\lambda_N$  peptide, derived from the  $\lambda$  bacteriophage (Daigle & Ellenberg 2007). All the above-mentioned systems are orthogonal to each other, meaning that the MS2 coat proteins do not bind to PP7 binding site or vice versa (Lim & Peabody 2002). This orthogonal functionality aids in detecting the three different RNA at the same time or three different regions of single RNA (Hocine et al. 2013).

The advantage of using the MS2-GFP system over the expression of fluorescent-tagged endogenous RNA binding proteins is that the MS2-GFP system is highly specific to RNA containing the MS2 stem-loop structure, while the other may bind to several mRNAs and reflect the behavior of all of them. Therefore this system provides two benefits: detection of specific RNA molecules under the microscope and study the dynamics of tagged RNA in live cells.

In this thesis, we made use of the MS2-GFP tagging system to study intervals between consecutive RNA production events of multiple *E. coli* promoters, under different inductions, and conditions in live cells. In all Publications, target genes are inserted in the single copy F<sup>+</sup> plasmid.

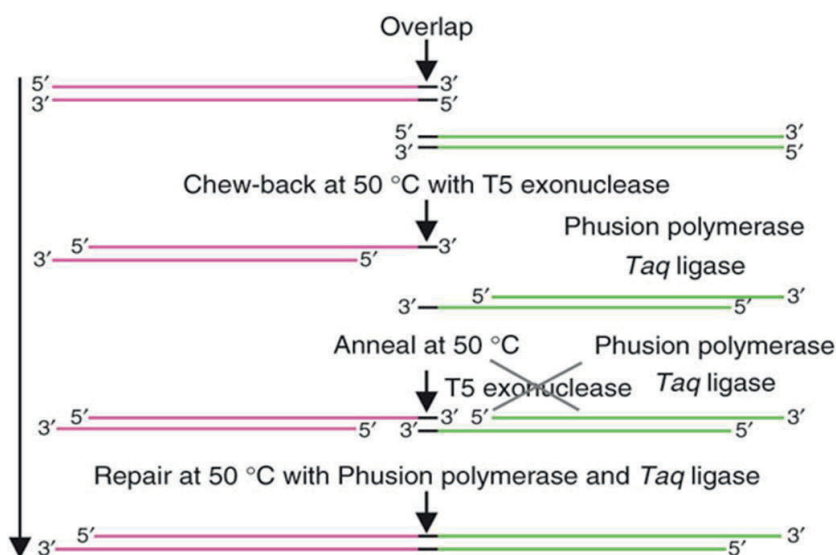
### 3.2.2 Engineering of Synthetic Genetic Constructs

A decade ago it was reported that using advanced molecular techniques like DNA assembly and *de novo* synthesis, it is possible to construct functional synthetic genomes (Gibson et al. 2009). This is achieved by synthesizing multiple small DNA fragments separately and assembling them to a larger piece of DNA, which is transferred into the genome-free host cell. These advancements have contributed to creating synthetic organisms with engineered genomes, with pre-defined specifications and functions. DNA assembly methods have become essential tools in synthetic biology to engineer complex systems with standardized and specific genetic parts. Also, it is reliable, cheap and fast. Thus, most researchers use these methods, instead of traditional methodologies such as molecular cloning using restriction enzymes, etc.

Several techniques made possible to construct genetic parts; one such technology is the Gibson Assembly<sup>®</sup> method. This method has proven its value by synthesizing the complete genome and transfer it into genome-free host cells (Gibson et al. 2010). It is a cloning method that allows assembling the multiple overlapping regions of DNA fragments in a single reaction mixture. In detailed, the Gibson assembly<sup>®</sup> master mix consists of three components: (i) T5 exonuclease enzyme, that cleaves the 5' ends of double-stranded DNA generating single-stranded complementary DNA overhangs, (ii) a DNA polymerase enzyme, which fills in the gaps of the annealed sequence, and (iii) a Taq ligase enzyme that joins the ends of the two DNA strand nicks (See Figure 3.4).

When performing the Gibson Assembly<sup>®</sup> method, the following steps have to be considered for successful construct. When designing primers for the DNA fragments, one must consider adding the overlapping sequences, such that when amplifying the DNA fragments, it contains at least 40 bp overlapping regions with the adjacent DNA fragment. Considering that these DNA fragments are assembled with a vector to form a circular product, this vector should also have the overlapping region at the terminal ends with the DNA fragments, which they will ligate with. When combining all fragments and the vector, the concentration must be in the ratio of 3:1. Gibson Assembly<sup>®</sup> method can be performed by a single isothermal reaction, by adding DNA fragments, Vector and Gibson Assembly<sup>®</sup> master mix.





**Figure 3.4:** Overview of the single-step reaction of the Gibson Assembly® method. The reaction mixture consists of multiple DNA fragments with overlapping regions, DNA polymerase, T5 exonuclease and ligase enzymes that are needed to ligate these fragments. In this picture, the two DNA fragments (green and pink coloured) are treated with T5 exonuclease at 50°C. Next, the products are treated with Phusion polymerase and Taq ligase to fill the gap of the final ligated DNA products. This picture is adopted and reprinted with permission from Macmillan Publishers Ltd: [Nature Methods] (Gibson et al. 2009), copyright (2009).

In this thesis, for studying the *in vivo* production of RNA molecules using the MS2 GFP system, we have constructed single-copy plasmids using the Gibson Assembly® method. Particularly in **Publication II and V**, using a computer-based simulator and the Gibson Assembly® method, we have built a DNA fragment containing multiple repeats of MS2-GFP binding sites for individual RNA detection and integrated them into single copy F-plasmids.

### 3.2.3 Time-lapse microscopy

As mentioned in section 3.2.1, MS2 GFP tagging allows monitoring individual RNAs by time-lapse fluorescence microscopy at the single-molecule level (Muthukrishnan et al. 2012; Mäkelä et al. 2013; Lloyd-Price et al. 2016; Startceva et al. 2019). To perform such experiments, cells containing the target and reporter system must be placed on the 2.5% agarose gel pad, which is sandwiched between the microscopic slide and the glass cover-slip. The agarose gel pad consists of necessary nutrients requires for cell growth, the inducers to activate the reporter and target systems, and the respective antibiotics (Golding et al. 2005; Muthukrishnan et al. 2012; Mäkelä et al. 2013; Lloyd-Price et al. 2016; Startceva et al. 2019). Also, there is a continuous supply of nutrient medium to the cells, done with the help of a peristaltic pump. This allows steady-state growth for many hours under the microscope. In addition, a temperature-controlled chamber was used to ensure that a specific temperature is maintained during the measurements.

The time-lapse images were acquired by a Nikon Eclipse inverted microscope (Ti-E, Nikon, Japan) with confocal laser scanning with a 100x Apo TIRF (1.49 NA, oil) objective. Confocal images were taken at specific time intervals with a Nikon C2 camera and respective phase-contrast images were captured with a DS-Fi2 CCD camera. GFP fluorescence measured by using a 488 nm argon ion laser (Melles-Griot) and a 515/30 nm emission detection filter. For image acquisition, we used the NIS-Elements software (Nikon).

In this thesis, we used this experimental approach for studying the *in vivo* transcription activation kinetics and subsequent RNA production of *E. coli* genes. Particularly in **Publication II**, to investigate the transcription dynamics of multiple genes in different growth phase conditions, cells (when under the microscope) were continuously supplied with respective growth phase medium and all the inducers by a peristaltic pump and maintained at a specific temperature during the measurement times. In **Publication III**, to study how the variability in gene activation times and RNA production intervals contribute to variability in RNA numbers between cell lineages, cells were placed under the microscope with a constant supply of fresh medium containing the inducers for the target and reporter genes. Here, the target gene was activated under the microscope to determine the time taken to produce the first RNA. Next, we captured fluorescence images every 2 min for 2 hours, and phase-contrast images every 5 min. During the two hours, we maintained a specific temperature using the temperature chamber (Bioptechs, FCS2). The tools used to analyze images are described in chapter 4.

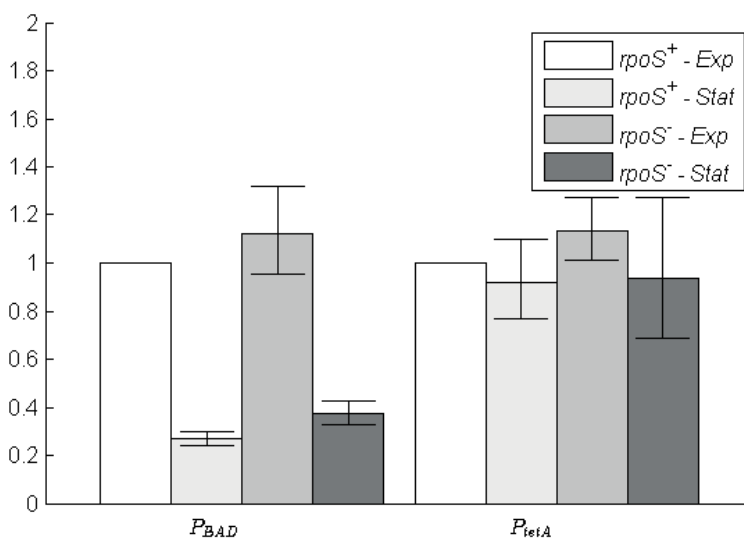
### 3.3 Validation methods

#### 3.3.1 Quantitative polymerase chain reaction.

The MS2 GFP system is more sensitive and informative than any other method to extract time intervals between RNA production events. Currently, there is no independent validation method with this level of precision and sensitivity. However, other techniques can partially validate these measurements. One such technique is quantitative polymerase chain reaction (qPCR). It allows measuring of RNA production rates (Schmittgen & Livak 2008).

In the traditional PCR method, the critical result is the end product generated at the final cycle of the PCR reaction. In qPCR, the detection of the sample is monitored at each cycle of PCR reaction. It quantifies the relative or absolute amount of the amplified product between the samples. In detail, to perform qPCR, a total RNA has to be extracted from the cells, then converted into complementary DNA (cDNA) using the reverse transcriptase enzyme. Next, using the modified version of the standard PCR reaction cycle, specific primers are used to amplify the coding region of interest that have an amplicon size of 150-200 bp. The amplified product can be detected in real-time in each cycle by fluorescent probes or fluorescent DNA binding dyes. These probes or dyes should be sequence-specific to an amplified region of the target. Considering when the DNA amplifies exponentially, the fluorescence amount should increase gradually above the background level. It is important to account for this background signal to generate meaningful information about the target, which is addressed by threshold cycle value ( $C_T$ ).

The threshold cycle value ( $C_T$ ), is the point at which the fluorescence signal is above the background level. This value tells the number of cycles it took to detect the signal from the sample.  $C_T$  values are inverse to the amount of target product in the sample and correlate the number of target copies in the sample. Lower  $C_T$  values indicate higher expression of the target product or vice versa (Schmittgen & Livak 2008). It is possible to determine the relative or absolute fold change in mRNA level between samples from  $C_T$  values. One such method to determine the mRNA fold change is “Delta–Delta  $C_T$ ” or the Livak’s method (Livak & Schmittgen 2001). In this, reference (housekeeping) genes such as 16s *rRNA* or 23s *rRNA*, were used as internal controls for the relative quantification of the target mRNA levels and production rates.

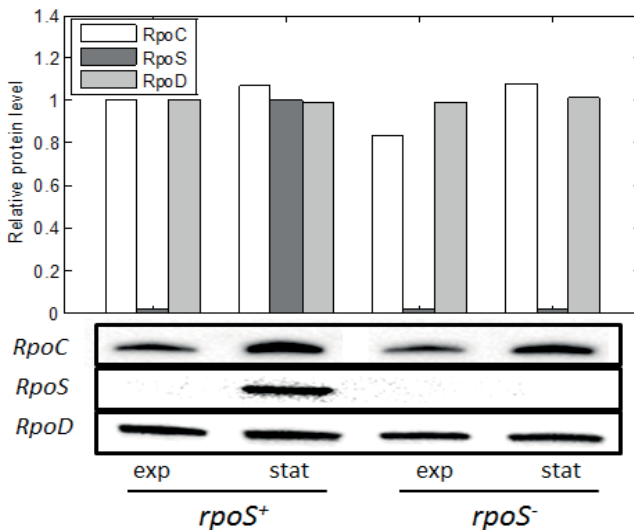


**Figure 3.5:** An example of qPCR results. Relative mRNA levels under the control of promoters  $P_{BAD}$  and  $P_{tetA}$  measured and compared in the exponential growth phase (Exp) and stationary growth phase (Stat) in different *E. coli* strains ( $rpoS^+$  and  $rpoS^-$ ). The error bars are based on three technical replicates from each sample. This image is used in **Publication II** and reprinted with permission from Elsevier.

In this thesis, particularly in **Publication II**, using this method, we quantified the relative fold change in the mRNA expression of two promoters ( $P_{BAD}$  and  $P_{tetA}$ ), under different growth phases (exponential and stationary), between two *E. coli* strains (Figure 3.5). The genes not affected by the experimental condition as used as internal reference genes (16s *rRNA*) for relative quantification of target mRNA expression levels (Livak & Schmittgen 2001). Apart from that, we also use this method to measure the induction curve of multiple promoters. In Publications **III, IV and V**, we also used qPCR to measure the mean transcription rates of various promoters in different media richness (discussed in detail in subsequent sections).

### 3.3.2 Western blotting

Western blotting is one of the traditional analytical techniques in molecular biology, used to separate and detect the specific protein from a total mixture of proteins extracted from the cells. The following steps achieve this: Extraction of total proteins from the cell lysate, followed by proteins separation based on molecular weight using gel electrophoresis. Then, the separated protein is transferred on to a nitrocellulose or PVDF membrane to produce a band of each protein. Next, the proteins are blocked on the membrane, using the blocking buffer of 5% BSA or nonfat dried milk. Next, they are treated with a primary antibody specific to a target protein. Before the unbound antibody is washed off, a secondary antibody treatment is done to recognize and bind to the primary antibody. The bound antibody is then chemically treated to detect them as a single band under the chemiluminescence doc, which implies that the antibodies bind to a specific target of interest. The thickness of the band determines the quantity of target protein present.

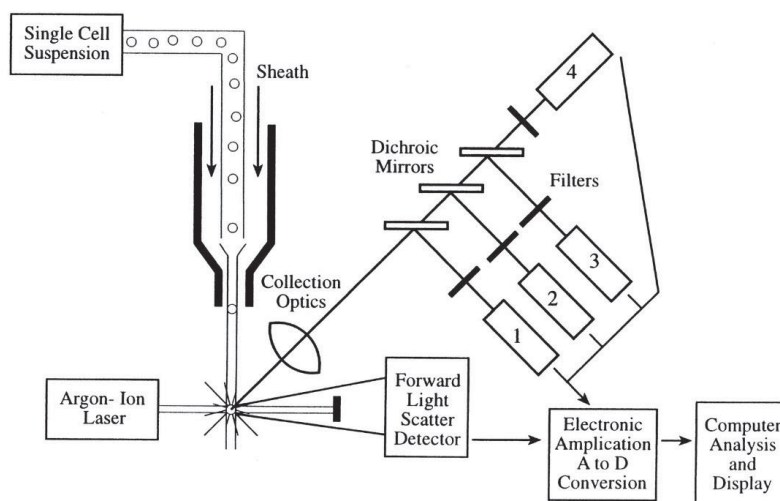


**Figure 3.6:** RNAP subunits quantification by the western blot method. Relative protein levels of RpoC ( a  $\beta'$  prime subunit of RNAP) and two sigma subunits namely,  $\sigma^{70}$  (RpoD) and  $\sigma^{38}$  (RpoS), measured in the exponential growth phase (exp) and stationary growth phase (stat) in *E. coli* wild type (*rpoS*<sup>+</sup>) and deletion mutation (*rpoS*<sup>-</sup>) strains. This image is used in Publication II and reprinted with permission from Elsevier.

In this thesis, in Publications **II**, **III** and **V**, to alter RNAP concentrations in live *E. coli* cells, we modified the Lauri broth medium ingredients with various concentrations. We then grew the cells in these tailored media so that the RNAP levels differed. By using the western blot technique, we measured the relative RpoC levels (one of the subunits of an RNAP enzyme complex) in different strains, for varying media richness. Additionally, in **Publication II** we also quantified the amount of the  $\sigma^{70}$  (RpoD) and  $\sigma^{38}$  (RpoS) concentrations in exponential and stationary growth phases (Figure 3.6).

### 3.3.3 Flow Cytometry

Flow cytometry is an optical technique, used to detect and characterize the properties of an individual cell in heterogeneous populations. In real-time, it is able to analyze thousands of particles per second and perform multiple quantitative measurements with specified optical properties at a similar rate. It has three main components: Fluidics system, optical system, and workstation.



**Figure 3.7:** Schematics of a flow cytometer. A single-cell suspension is focused on the light source (lasers). The signals are collected by the Forward light scatter detector and amplified, to convert them to digital form for further analysis. This picture is adopted from (Brown & Wittwer 2000).

The fluidic system consists of sheath fluid, where cells are suspended and passed through the center of the laser beam, to measure their optical properties (Figure 3.7). The optics systems consist of lasers, beam optics, light-collecting optics, dichroic mirrors/filters, and detectors. The laser is used to excite molecules of cells, where the beam optics shape the laser beam into an elliptical spot at the center of the flow cell. The light collecting optics collects the light information (including forward scatter light, side scatter light, and fluorescence) from the cells. The collected light is separated by the dichroic mirrors and filtered by bandpass filters. The optical detectors are used to convert the light into electronic signals. The information from the cells that are processed by the optical system is further converted into digital signals with the help of the workstation. The process of collecting data from the cells using the flow cytometer is called acquisition.

In this thesis, particularly in **Publication V**, we make use of the flow cytometer to study the single-cell distribution of protein expression levels in *E. coli*. Briefly, cells carrying plasmids having the target gene followed by mCherry fluorescence protein, under the control of the inducible promoter, were grown and induced with various concentrations of inducers. Using the flow cytometer, we measured the mCherry protein expression levels at various inductions. From these measurements, we obtained the single-cell distributions and extracted their mean, coefficient of variation, skewness, and kurtosis.

### 3.4 Stochastic Simulation Algorithm and Stochastic Simulators

Cellular functioning involves biochemical reactions between molecular species. As some of these molecular species are present in very low number, e.g. DNA, RNA and regulatory proteins in a cell, the simulation of the dynamics of such interactions needs a discrete model. In theory, using the Chemical masters equation (CME) it is possible, from knowing the current state of a chemical system, predict all possible future states of that system. However, for complex systems, one cannot solve this equation. Instead one can opt to simulate the dynamics of the model by sampling trajectories from the distribution described by the CME. This is the approach of the stochastic simulation algorithm (Gillespie 1976; Gillespie 1977; Gillespie 1992; Gillespie 2007).

The Stochastic Simulation Algorithm (SSA) of the chemical master equation, is an exact procedure for numerically simulating the time evolution of a well-stirred reacting system.

The propensity function ( $a_{\mu}$ ) is defined as follows:

$a_{\mu}(\mathbf{x})dt$  = the probability of the one reaction  $R_{\mu}$  that occurs in volume  $V$  in the next infinite time intervals  $(t, t+dt)$ .

The propensity of the reaction indicates the system's state and temporal evolution. The propensity function depends on the type of reacting species. For example, in unimolecular reactions, the constant  $c_{\mu}$  is the probability of particular molecules  $X$  will spontaneously react via reaction  $R_{\mu}$  in the next infinitesimal time interval  $dt$ . The propensity function of this reaction is:

$$a_{\mu}(\mathbf{x}) = c_{\mu} X$$

Similarly, for bimolecular reactions between two molecular species  $X_1$  and  $X_2$ , the constant  $c_{\mu}$  is the probability that a random pair of a molecule from  $X_1$  and  $X_2$  react with reaction  $R_{\mu}$  in the next infinitesimal time interval  $dt$ . For this, the propensity function will be:

$$a_{\mu}(\mathbf{x}) = c_{\mu} X_1 X_2$$

Instead, if the reaction is between two identical molecules, e.g. of species X, and  $c_\mu$  is the probability they will react in reaction  $R_\mu$  after a  $dt$  infinitesimal time moment, the propensity will instead be:

$$a_\mu(\mathbf{x}) = c_\mu X(X-1)/2$$

In order to simulate the dynamics of a chemical system, at each moment, one needs to generate two random variables: i) the time that will take for the next reaction to occur ( $\tau$ ) and what reaction will occur ( $\mu$ ). These are given by (Gillespie 1976; Gillespie 1977):

$$p(\tau, \mu | \mathbf{x}, t) = a_\mu(\mathbf{x}) \exp(-a_0(\mathbf{x})\tau),$$

$$\text{where, } a_0(\mathbf{x}) = \sum_{j=1}^M a_j(\mathbf{x})$$

One implementation of the stochastic simulation algorithm (SSA) is described below:

Step 1: Initialize the step by setting  $t = 0$  and  $\mathbf{x} = \mathbf{x}_0$ . Where  $\mathbf{x}$  is the vector state that consists of all molecular species in a system at a given time  $t$  and  $\mathbf{x}_0$  is the state vector consisting of the initial concentration of molecular species.

Step 2: Evaluate all propensity functions  $a_\mu(\mathbf{x})$  and their sum  $a_0(\mathbf{x})$ .

Step 3: Using a sampling procedure calculate the time taken for the next reaction to occur ( $\tau$ ) and the index of this reaction ( $\mu$ ).

Step 4: If  $t + \tau \geq t_{\text{stop}}$  terminate the simulation.

Step 5: Set  $t = t + \tau$  and  $\mathbf{x} = \mathbf{x} + \mathbf{v}_\mu$

Step 6: Go to step 2 or else end the simulation.

### 3.5 Models of Transcription

We produced several models of gene expression, with the aim of being able to make use of computer simulations in order to predict, mimic, or better visualize the cellular process. Once validated with empirical data, the model serves as a framework to test new hypotheses. The models used here were essential in generating new hypotheses and find the key underlying mechanism responsible for the empirical observations.

In general, the models built using the stochastic formulation are described as a set of chemical reactions. A basic process is represented in reactions 3.1.



In reaction 3.1, A and B are two reactant molecules that, when reacting, form a product called C. This reaction has a stochastic rate constant  $k$ , which, along with the amount of A and B, define the propensity of the reaction.

When modeling the transcription process, one can simplify it by considering only the most “rate-limiting” steps affecting the kinetics of the process (i.e the dynamics of RNA production) (Ribeiro et al. 2006). For example, the process of transcription can be modeled as a single-step reaction.



Here, RNAP represents the RNA polymerase holoenzyme, Pro is the promoter and RNA is the end product of this reaction. As neither RNAP nor Pro are consumed, they must also be products of the reaction. Meanwhile,  $k$  is the stochastic rate constant of this reaction. Note that, it is assumed that there are no regulatory proteins (e.g. repressors) that could block transcription. As such, the gene would be continuously producing the RNA in a constitutive fashion.

The model above is a much-simplified version of the transcription process, it does not consider any reversible steps or other rate-limiting steps in the transcription process, such as the open complex formation. For this, the model needs to be more complex.

In *E. coli*, transcription initiation is a complex multi rate-limiting step process (McClure 1985; Saecker et al. 2011; Browning & Busby 2016), and it can be represented as follows:



Reaction 3.3 involves the binding of the RNA polymerase holoenzyme ( $RNAP$ ), to the free Promoter (Pro) with a rate constant ( $K_s$ ) to form a closed complex ( $RP_{CC}$ ). Following this, the closed complex undergoes isomerization to form an open complex step ( $RP_{OC}$ ) with rate constant ( $K_f$ ). Next,  $RNAP$  enters into elongation steps via scrunching, promoter escape, reaches termination point, and release RNA. The rate of formation this step is much faster than the other rate-limiting steps, and so it is represented as negligible or infinitely-fast the reaction. Note that the first step is reversible. As such,  $K_s$  depends on the forward and backward reactions. This accounts for the chemical instability of the closed complex. This model representation was first proposed by Walter, Zillig, and colleagues (Walter et al. 1967).



Studies reported that in *E. coli*, at a given condition, transcription of highly expressed operons occurs in bursts, because of positive supercoiling buildup (PSB) (Chong et al. 2014). The PSB arises due to the presence of segments of topological constraints in the chromosomal DNA (Hardy & Cozzarelli 2005; Rovinskiy et al. 2012). Meanwhile, plasmids DNA only has transient topological constraints (Leng et al. 2011), except plasmids encoding for membrane-associated proteins, carrying tandem copies of multiple binding sites or when expressed in *E. coli* strains lacking topA gene. In these, the segments of topological constraints are more efficient than the transient ones which lead to PSB (Deng et al. 2006). Recently, in vitro measurements showed that in the plasmids having only transient topological constraints, when buildup arise, they freely diffusive in the opposite direction leading to their annihilation (Chong et al. 2014).

In this thesis, all promoters considered were inserted in a single-copy plasmid that lacks such segments of topological constraints. Due to this, the model for transcription initiation does not account PSB that could generate transcription bursts.

In the **Publication I**, based on the empirical data and model assumed, we characterized the rate-limiting steps (particularly duration of closed and open complex formation) in transcription initiation of Promoter  $P_{lac/ara-1}$  in live *E. coli* cells.

The transcription initiation model that best fitted the empirical data is



Reaction 3.4, represents the multi-step process of transcription initiation of an active promoter. It begins with the formation of the closed complex ( $RP_c$ ), i.e. the binding of the RNA polymerase holoenzyme (R) to an active promoter ( $P_{ON}$ ). Once polymerase reaches the start site, it opens the DNA double helix, leading to the formation of the open complex ( $RP_o$ ). Next, the polymerase enters in elongation, clearing the promoter region. Here  $k_1$  represents the rate at which polymerases find and bind to the promoter region.  $k_2$  and  $k_{-1}$  are interpreted as the product of the rates of the elementary reactions.

Reaction 3.5, represents the transitions between active ( $P_{ON}$ ) and inactive states ( $P_{OFF}$ ) of the promoter. In the context of our constructs, this is due to the binding and unbinding of regulatory molecules such as activator or repressor to the promoter region (Lutz et al. 2001), not an accumulation of positive supercoiling in the DNA (Chong et al. 2014).

From the empirical data and model, we concluded that the time spent by the promoter in OFF state is  $\sim 87$ s, while the time in closed complex formation is  $\sim 788$  s and the time in open complex formation is  $\sim 193$  s.

In **Publication II**, we assumed a transcription model proposed in (Ribeiro et al. 2006). In addition to this model, we also considered the  $\sigma$  factors competition towards the core RNA polymerase, as these are a limited number in *E. coli*. We studied genes that are primarily transcribed by  $\sigma^{70}$  thus, binding of  $\sigma^{70}$  to core RNAP is represented as in reaction (3.6). *E. coli* has other  $\sigma$  factors and binding of core RNAP, as represented in reaction 3.7.



In the reactions 3.6 and 3.7, RNAP stands for the core RNA polymerase,  $\sigma^{70}$  is the primary sigma factor, and  $\sigma^i$  stands for the other sigma factors.  $K_{70}$  and  $K_i$  are the rate constants of association and disassociation of sigma factor to core RNA polymerase to form RNAP holoenzyme ( $RNAP.\sigma^{70}$  or  $RNAP.\sigma^i$ ). From these reactions, one can estimate the approximate number of RNA holoenzyme carrying  $\sigma^{70}$ . At equilibrium, the approximate number of  $RNAP.\sigma^{70}$  is given by

$$[RNAP.\sigma^{70}] \sim [RNAP] \frac{[\sigma^{70}]K_{70}}{[\sigma^{70}]K_{70} + [\sigma^i]K_i} \quad (3.8)$$

In this equation 3.8,  $[RNAP]$  stands for the number of free-floating RNAP core enzymes,  $[\sigma^{70}]$  and  $[\sigma^i]$  are the number of  $\sigma^{70}$  and other  $\sigma$  factors, that are freely floating or in the holoenzyme form.

Based on the transcription initiation model (including sigma factors interaction with RNAP, and multi rate-limiting steps) and empirical data, we measured the time intervals between the consecutive RNA productions along with the mean duration of the open complex formation of several genes.

In **Publication III**, along with the transcription initiation model as described in reaction 3.3, we also model the gene activation times as follows,



Here,  $I_1$  is an uninduced state,  $I_2$  is an intermediate state, and  $S_0$  is the induced state, in which the promoter is available for transcription. The reactions occur at rates  $k_1$  and  $k_2$ , and are catalyzed by an uptake protein (Upt). The number of uptake proteins is expected to affect the rates of both steps, as the shape of the distribution did not change with inducer concentration. The process of activating transcription with an external inducer contains molecular events, such as the inducer being imported into the cytoplasm, binding to a transcription factor, releasing DNA repression loop, etc. (Megerle et al. 2008; Fritz et al. 2014; Choi et al. 2008; Schleif 2000). Also, these molecular level details differ between induction systems.

## 3.6 Tau ( $\tau$ ) plots

The kinetics of rate-limiting steps of transcription initiation have been measured making use of *in vitro* transcription assays and abortive initiation techniques (Buc & McClure 1985; McClure 1985).

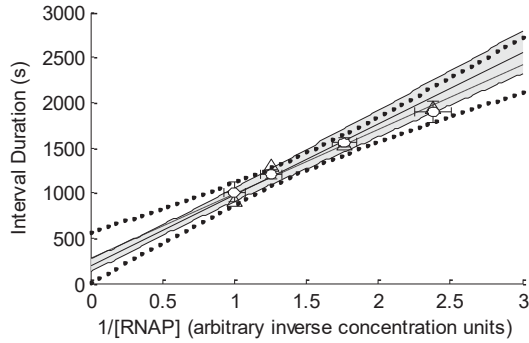
It was shown that there is a lag time before reaching a steady-state rate of abortive initiation products. This is the time taken by RNAP to find and bind to the promoter region to form the closed complex, and therefore it depends on the concentration of RNAP. Meanwhile, the next step, the open complex formation, does not depend on this concentration (McClure 1985). Therefore, measuring the times for RNA production, at increasing high RNAP concentrations, it should make the first step ever faster, while not altering the time length of the second step.

From this direct relationship between the lag times and the inverse of RNA polymerase concentrations, it is possible to draw a Lineweaver-Burk plot (Lineweaver & Burk 1934), named ‘tau ( $\tau$ )-plots’ (McClure 1980). In these, the point where the line intercepts with the y-axis should correspond to the mean duration of the open complex formation, since that height should correspond to the lag time if the closed complex was infinitely fast. Meanwhile, the difference in height between this point, and the height of the original inverse of the rate of RNA production in the control condition should correspond to the closed complex formation, as it is the remainder between open complex formation and total time to produce one RNA.

These measurements are relatively easy to perform *in vitro* using a transcription abortive assay because the concentrations of the components can be precisely controlled, no unknown components exist, and wide changes in concentrations are possible. Meanwhile, it is challenging to achieve this *in vivo*. For example, it is complex to determine how many RNAP molecules are, at a given moment, free for transcription, as there is a large amount of these molecules committed to transcription of various genes, at any given time. Further, changing RNA polymerases concentration in live cells is expected to disturb significantly their functionality (Gummesson et al. 2009).

To address this, in **Publication I**, we established a method to change the RNA polymerase concentration in live *E. coli* cells (by changing media richness) that overcomes two major impediments. First, it is shown that the changes in media richness did not tangibly affect cell growth rates. Also, the RNA production rates were shown to change linearly with RNAP concentration, which is evidence that the fraction of RNA polymerases free for transcription was kept approximately constant within this range of conditions. This implies that it is possible to recreate, *in vivo*, the conditions met by the *in vitro* measurements.

Given this, by altering media richness with the method proposed in **Publication I**, we dissected the mean duration of the steps prior and after commitment to open complex formation from measurements of the RNA production rate at different RNAP concentrations. Finally, assuming current models of transcription, we interpreted these times as estimates of the closed complex and open complex formations.



**Figure 3.8:** Absolute  $\tau$ -plot. Estimating the duration of open complex formation of Plac/ara -1, by plotting the mean intervals between the transcription events on X-axis versus the inverse of RNAP concentration on Y-axis. This picture is adopted from a **Publication I**.

In this thesis,  $\tau$ -plots (absolute and relative) have been used in all Publications. For absolute  $\tau$ -plots, the mean time-scale of intervals between two consecutive RNA production events is obtained from time-lapse microscopy data of time intervals between consecutive RNA production events in individual cells. These intervals duration is then plotted against the inverse of the concentration of RNAP on the x-axis (Figure 3.8). In relative  $\tau$ -plots, the rate of RNA production in each condition is obtained from qPCR data, which is also plotted against the inverse of the RNAP concentration. With this data, one can only measure the rate of RNA production relative to the control. Thus, one can also only estimate the relative duration of open complex formation.

## 4 IMAGE AND DATA ANALYSIS

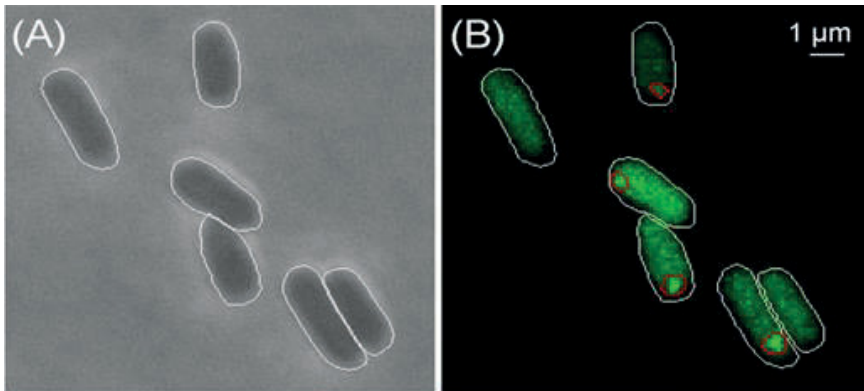
This chapter presents computational tools implemented for image analysis in the works composing this thesis. These tools are used in cell segmentation and RNA spot detection and quantification from time-lapse microscopy images. The quantification of the time intervals is discussed in the final section of this chapter.

### 4.1 Cell Segmentation and Lineage Construction

Many genes in *E. coli* are expressed in a rare and stochastic fashion, meaning that few transcripts are produced during the cells lifetime (So et al. 2011). Therefore, to characterize the kinetics of this process, one needs to observe many cells, sometimes for several generations. Performing such experiments requires multi-modal microscopy (e.g. fluorescence and phase contrast) and single-molecule fluorescent probes.

Manual extraction of the data from the microscope images is inappropriate due to being laborious, but also because it may lead to the introduction of errors (usually biased and differing between people collecting the data). Thus, robust image analysis and signal processing tools provide much support for performing accurate and unbiased extraction and quantification of the desired measure/variable of the study.

The first step in microscopy image analysis is cell segmentation, where cells are detected and automatically segmented from the image. The location, orientation, and size of the cells are measured using principal component analysis (PCA). The segmentation methods require the cells to be spatially sparse. In case of cell clusters, more sophisticated methods of cell segmentation are required, e.g. employing multi-scale morphological edge detection with denoising filters to segment the clusters into an initial set of candidate segments (Hakkinen et al. 2013). The level of accuracy obtained by this tool is therefore very high. Next, to analyze time-series microscopy images, consecutive images are aligned using cross-correlation. This alignment removes possible drifts occurred during image acquisition. The sources of drift include movement of the stage, temperature change, media inflow, etc. which hamper the tracking of cells over generations. Once individual cells are segmented in one image, it is possible to construct cell lineages, provided that there is no drift larger than, e.g. a cell's size.



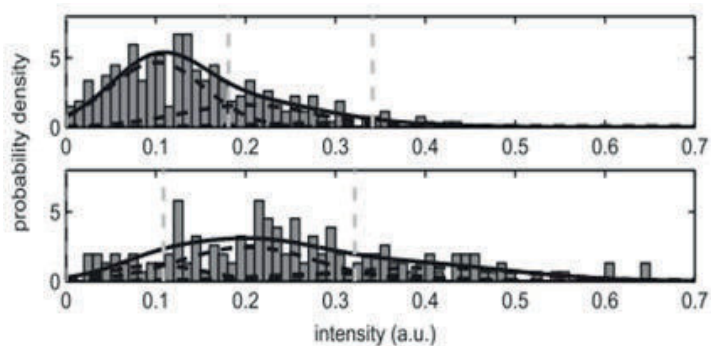
**Figure 4.1:** Phase-contrast and confocal images of *E. coli* cells. In the phase-contrast image (A) the cells were segmented and corresponding confocal image (B), cells with MS2-GFP-RNA spots, which were detected and segmented as well.

To perform multi-modal cell segmentation in acquired microscopic images, we made use of software “MAMLE” (multi-resolution analysis and maximum likelihood estimation) (Chowdhury et al. 2013). It performs automatic segmentation of the cells but allows the results to be manually corrected, if necessary. An example image of how the cells are segmented is shown in Figure 4.1a. To establish the relationships between the cells in sequential frames, we also use the software “CellAging” (Hakkinen et al. 2013). It is established as follows: a segment overlapping is most associated with each segment in the next frame. If the association is one to one, it is assumed that it is the same cell (no cell division occurred). Else, if the association is more than one, it is interpreted as a cell division. For one to zero or zero to one association, no relation between the segments is established.

In all **Publications I-V**, we made use of this software to perform cell and spots segmentation and tracking.

## 4.2 Spot Detection

Once the segmentation is concluded, the next step is to extract the information on the confocal images, i.e. to quantify the RNA molecules in each cell. For that, the alignment of the segmented mask cells of phase-contrast images to corresponding confocal images is required (Figure 4.1b). The confocal images contain the cells with RNA spots trapped inside the MS2 GFP complex (Figure 4.1b) (Golding et al. 2005; Golding & Cox 2004). In order to estimate the spot intensity inside the cells, it has to be segmented as in Figure 4.1b and the fluorescence intensity distribution of a spot is detected by Kernel density estimation (KDE) method using a Gaussian Kernel (Ruusuvaori et al. 2010) and Otsu’s Threshold (Otsu 1979).



**Figure 4.2:** Single-cell distribution of RNA spot and cell intensities. Spot intensities (Top), cell intensities (bottom). The solid black lines show the overall estimated distributions, the dashed black lines their components and the dashed gray lines the decision boundaries. Adapted and modified with permission from (Hakkinen et al. 2014).

Once the spot is detected, more features of the spot, such as position, total fluorescence intensities, and area are extracted. Then the spot intensity is corrected to background fluorescence by multiplying the area of the spot with the average intensity outside the spot and then subtracting that from the total intensity of each spot. From the histogram of all spots intensity, the number of RNA molecules of the cells can be extracted by normalizing it with the intensity of a single tagged RNA molecule (Golding et al. 2005). The peaks correspond to the integer-valued number of RNAs. An example image of RNA number detection is shown in Figure 4.2.

In general, the best estimations of the intensity of a single RNA are obtained by using the first peak of the distribution of spots intensity. This can be further improved by, e.g., measuring spots intensity in cells with very weak induction, with most cells not having a spot, thus implying that the existing spots are very likely to be a single RNA.

In this thesis, cell segmentation, spot detection, and RNA counting methods were used in all Publications.

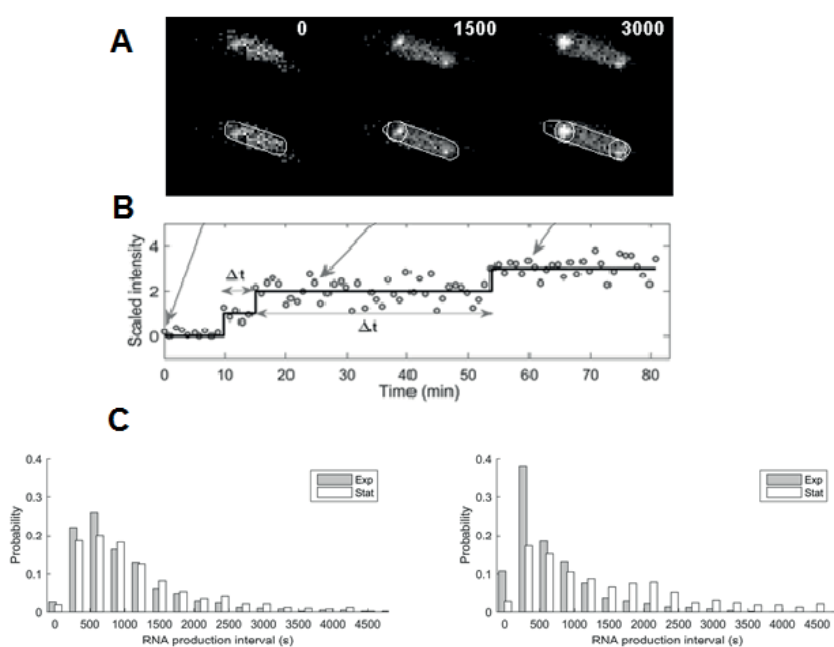
### 4.3 Extraction of Time intervals from Total Spot Fluorescence over time

In all **Publications I -V**, we used time-lapse images to monitor the dynamics of transcription events in live cells. The information from these images can be used to build detailed models of transcription at a single-cell level.

Time-series images have significantly more information on RNA production dynamics than single time point images. E.g. many production dynamics could generate the same RNA numbers in a given population. Time-lapse images allow identifying with precision the production dynamics.

Since the lifetime of RNA molecules trapped inside the MS2-GFP system is much longer than the cell division time (Peabody 1993; Golding & Cox 2004), the total spot intensity in the cell tends to increase over time, as new RNA molecules are produced (Tran et al. 2015). Only cell divisions are able to decrease RNA numbers in the cells.

Due to the high fluorescence of each spot, provided that the cells and cells are clearly visible and that detailed segmentation is performed, the moments when novel target RNAs appear are visible as a discrete “jump” in the total spot fluorescence intensity of the cell over time. In general, each jump corresponds to the production of a single RNA molecule. Figure 4.3b depicts an example of the usage of the jump detection method when applied to a time-lapse graph of the scaled intensity of spots in one.



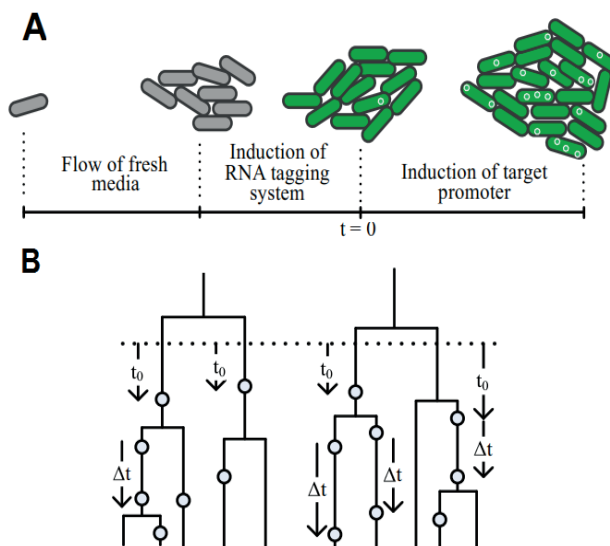
**Figure 4.3:** Quantification of integer-value RNA molecules in individual cells. A) Top: An example confocal image over time, with a cell expressing MS2 GFP along with two target RNA spots. Bottom: segmentation of the example cell and the RNA spots within, white circles. B) Scaled spot intensities over time along with the best-fitting monotonic piecewise-constant curve (black line) from which  $\Delta t$  intervals are estimated by measuring the times between jumps. C) Distributions of time intervals between consecutive RNA productions in individual cells under the control of promoter PtetA (Left) and PBAD (Right). The dark gray bars show the distribution of RNA intervals from cells measured in the exponential phase, while light gray is data from cells in the stationary phase. Images were obtained and modified from **Publication II**.

This method is used to extract the time intervals between two consecutive RNA production events in each cell (Figure 4.3c). From these distributions of time intervals, a stochastic model of transcription was built with a specific number of rate-limiting steps using the maximum likelihood ratio test, which evaluates the



best-fitted models to the data (Hakkinen & Ribeiro 2015). In particular, assuming the model in reactions 3.4 and 3.5, the data allows estimating the values of the rate constants.

In all Publications, we used this method to obtain time intervals between RNA production events from many promoters, under various growth conditions and induction schemes. Particularly, in **Publication II**, we extracted the transcription time intervals from wild type (WT) and deletion mutant *E. coli* cells, when growing in an exponential growth phase and stationary growth phase. Later, based on these distributions, we estimated a two-step model of transcription initiation which best-fits the data.



**Figure 4.4:** Cartoon of RNAs spot appearance in the cells. A) Cells containing the reporter (MS2-GFP system) and target plasmids (MS2-96BS) placed under the microscope and continuously supplied with media and inducers. At  $t = 0$ , cells are induced with the target plasmid inducers. B) Illustration of RNA production events in individual cells and their respective lineages (shown in circles). The dotted line represents the time when inducers were added.  $t_0$  represents the waiting time to produce the first RNA, and  $\Delta t$  represents the time intervals between the consecutive RNA production events in the cells. This image was obtained from **Publication III** and modified.

It is worth noting that, the “jump” detection method can also be used to determine the appearance of the first RNA production event inside the cells, following induction. In **Publication III**, we measured the time taken for the first RNA to appear in the cells, when under the control of *Lac/ara-1* in individual cells and their subsequent lineages. This promoter is regulated by two transcription factors; AraC is the activator molecule, inducible by Arabinose and the repressor of LacI, i.e., isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG) (Lutz & Bujard 1997; Mäkelä et al. 2017). To set up such experiments, we used a peristaltic pump, which allows supplying the media with the respective inducers to the cells, while monitoring under the microscope. From the time series images, we estimated the time for the appearance of the first RNA in

each cell after induction, which is denoted as  $t_0$ . Subsequent production events allowed measuring RNA production intervals, which are denoted as  $\Delta t$ . An example image is depicted in Figure 4.4.

## 4.4 Asymmetry and Tailedness of the Distribution of Transcription Intervals

Most of the knowledge of *E. coli* gene expression is limited to mean and noise level. It is worth noting that, from the distributions of time intervals between the RNA productions, it is possible to measure the asymmetry and tailedness in the dynamics of RNA production, which ought to contribute to phenotypic changes in the cells, particularly in the occurrence of threshold crossing in protein and/or RNA numbers.

In **Publication IV and V**, we measured the asymmetry and tailedness of distributions of time intervals between consecutive RNA production ( $\Delta t$ ) in individual cells. For this, we obtained the Skewness ( $S$ ) and Kurtosis ( $K$ ) of those distributions, as a means to quantify skewness and kurtosis, respectively.

Skewness equals the third moment of the distribution to the cube of standard deviation:

$$S = \frac{\langle (\Delta t - \langle \Delta t \rangle)^3 \rangle}{\sigma_{\Delta t}^3} \quad (1)$$

Whereas kurtosis is the fourth moment of the distribution to the power of four to the standard deviation:

$$K = \frac{\langle (\Delta t - \langle \Delta t \rangle)^4 \rangle}{\sigma_{\Delta t}^4} \quad (2)$$

To estimate the sample skewness ( $S_s$ ) and kurtosis ( $K_s$ ) of experimentally measured and simulated data, we apply a correction to increase the precision of the estimates for samples from asymmetric distributions in the above equations (D. N. Joanes & C. A. Gill 1998). This correction does not affect significantly distributions with a large sample size ( $>100$ ) (D. N. Joanes & C. A. Gill 1998).

$$S_s = \frac{\sqrt{n(n-1)}}{n-2} \cdot S \quad (3)$$

$$K_s = \frac{(n-1)}{(n-2)(n-3)} ((n+1)K - 3(n-1)) + 3 \quad (4)$$

To obtain confidence boundaries on sample skewness and kurtosis, we performed non-parametric bootstrap as in (DiCiccio & Efron 1996; Carpenter & Bithell 2000). Namely, for each data set, we resampled the data randomly with replacement (using the original amount of samples)  $10^5$  times, and calculated the bootstrap sample skewness ( $S_{sb}$ ), and kurtosis ( $K_{sb}$ ). As the obtained  $S_{sb}$  and  $K_{sb}$  distributions

were well-approximated by a normal distribution, we estimated the standard uncertainty as to the 68% percentile confidence interval of the obtained  $S_{sb}$  and  $K_{sb}$  distribution.

## 5 RESULTS: SUMMARY AND CONCLUSIONS

One of the key achievements of Molecular Biology is the unraveling of the mechanisms used by cells to perform transcription, including the complex process of initiation and its regulation apparatus. Most of these achievements were accomplished by making use of *in vitro* techniques (McClure 1985) that were able to dissect the apparatus and, from there, measure the kinetics and identify the parts composing the core machinery. From the results, it was possible to design detailed models that capture the rate-limiting, multi-step nature of transcription initiation (deHaseth et al. 1998).

With the advent of microscopy imaging of fluorescent molecules (Shimomura 1962; Tsien 1998), two new, main advances were made possible. First, it was observed that gene expression dynamics is stochastic in nature (McAdams & Arkin 1997; Arkin et al. 1998; Elowitz & Leibler 2000). Interestingly, previous evidence was presented in, e.g. (Neunauer & Calef 1970). Second, *in vivo* measurements of RNA numbers in individual cells identify an ON/OFF mechanism, preceding transcription initiation (Golding et al. 2005; Yu et al. 2006; Muthukrishnan et al. 2012), later explained to be the result of positive supercoiling buildup (Chong et al. 2014).

These, and several other findings, e.g. on the mechanics of promoter escape (Margeat et al. 2006), have allowed a better understanding of how core mechanics and regulatory mechanisms allow fine-tuning of gene expression in *Escherichia coli*.

Our work has focused on one such mechanism, namely, on the two-step nature of active transcription initiation and *E. coli*'s ability to regulate these steps independently. In particular, the five Publications composing the thesis contribute by, first, proposing a method to dissect and measure the *in vivo* dynamics of the rate-limiting steps of transcription initiation using time-lapse microscopy and stochastic modelling (**Publication I**). Next, we quantified the effects of selectivity and sensitivity of indirect regulation by global changes in  $\sigma$  factor numbers on promoters primarily transcribed by RNAP. $\sigma^{70}$  holoenzymes (**Publication II**). Afterward, we identified and quantified, at the RNA level, the effects of intake time of inducers on single-cell phenotypic variability. One interesting aspect of the effects of this phenomena is that, while long-lasting (several generations), they are, nevertheless, transient, unlike noise in transcription (**Publication III**). The results of the final work contribute to better profiling of the effects of the stochastic nature of gene expression, by providing evidence that there is regulation and evolvability in the asymmetry and

tailedness of single-cell distributions of RNA and protein numbers. Further, we identified one of the regulatory mechanisms of these noise components to be the relative durations of the steps prior and after commitment to open complex formation (**Publications IV and V**).

In detail, in **Publication I**, following similar principles as *in vitro* techniques, we proposed a method to dissect the kinetics of the rate-limiting steps in transcription, along with possible ON/OFF processes, from *in vivo* measurements of distributions of time intervals between consecutive transcription events in individual cells. Then, by applying a model-fitting procedure to the empirical data, we quantified the contribution of each rate-limiting step to those intervals.

We found that, in live, individual cells, under full *IPTG* and *Arabinose* induction, the closed complex formation of this promoter lasts  $\sim 788$  s, while subsequent steps last  $\sim 192$  s, on average. We also found evidence that the closed complex formation usually occurs multiple times prior to each successful commitment to open complex formation. Further, the promoter intermittently switches to an inactive state that, on average, lasts  $\sim 87$  s, consistent with the effects of intermittent repression of the promoter by LacI, even in the presence of inducers.

We expect the methodology proposed to be applicable to any gene whose changes in transcripts production rates are linear with changing RNAP concentrations. One interesting application of this methodology would be in quantifying the effects of changing environments on the dynamics of transcription, at the rate-limiting step level.

In **Publication II**, we studied the selectivity and sensitivity to indirect regulation by changes in  $\sigma$  factor numbers, in promoters primarily transcribed by RNAP. $\sigma^{70}$  holoenzymes.

First, from the mathematical analysis of a 2-step model of transcription initiation, we argued that the sensitivity of a promoter preferentially transcribed by RNAP. $\sigma^{70}$  holoenzymes to changes in  $\sigma$  factors numbers (other than  $\sigma^{70}$ ) should mainly depend on two factors. The first is the ratio between the duration of the closed and the open complex formation, which is sequence-dependent and subject to regulation. The second is the degree of change in  $\sigma$  factors numbers. To validate the model-based predictions, we used qPCR to compare RNA production rates of several promoters, under various induction schemes, and various growth phases, prior and after increases in  $\sigma^{38}$  numbers.

The measurements were found to be in agreement with the model predictions in a statistical sense, confirming that the response of a promoter to changes in  $\sigma^{38}$  numbers increases with the ratio between the duration of the closed and open complex formations.

For validation, we performed single-RNA *in vivo* microscopy in *rpoS*<sup>+</sup> and *rpoS*<sup>-</sup> cells of the transcriptional activity of the promoters used, with the highest and the lowest ratio between the duration of the closed and open complex formation. As predicted, only in the former is the transcriptional activity affected. By showing that this does not occur in the deletion mutant cells for  $\sigma^{38}$ , we concluded that the cause for the changes in transcription activity was the increase in  $\sigma^{38}$  numbers.

We conclude that, in *E. coli*, a promoter's responsiveness to indirect regulation by  $\sigma$  factor competition is determined by its sequence-dependent, dynamically regulated ratio between the duration of the closed and the open-complex formation.

In **Publication III**, we investigated the effects of noise in inducer intake times on cell-to-cell variability in RNA numbers of the target gene. Further, we investigated whether the effects of this extrinsic noise are promoter initiation kinetics dependent. Based on the multi-step nature of transcription, we hypothesized that, if the source of extrinsic noise (the inducer) only affects one of the rate-limiting steps (e.g. increases the rate of the open complex formation alone), then it would affect more strongly genes whose initiation kinetics is mostly rate-limited by that step. Further, we considered that there is significant variability between cell lineages, which should also be accounted for.

To study this, we followed by time-lapse microscopy independent cell lineages generated from individual cells for several generations. We then measured single-cell activation times and transcription intervals for different promoters induced by IPTG ( $P_{lac/ara-1}$  and  $P_{lac}$ ), and for different inducers on the same promoter ( $P_{lac/ara-1}$  induced by IPTG and by Arabinose).

Our results indicate that extrinsic noise from upstream processes, such as the intake time of external inducers, has a significant, but transient influence. As the mean and variability of these times differ with the inducer, lineage-to-lineage variability in RNA numbers also differs with the inducer. Meanwhile, we observed that lineage-to-lineage variability in RNA numbers also differ with the promoter.

In search of the mechanisms that could explain the latter, we considered that, depending on the source of extrinsic noise, different steps of transcription are expected to be affected (e.g. different transcription factors act at different steps, and thus the variability in their numbers will affect mostly the variability in the kinetics of those steps alone). For that, we considered the effects on RNA production should depend on the relative duration of the step affected (relative to the overall duration of the multistep transcription initiation process). I.e. if the steps affected are relative short time-lengthed, the effects ought to be weak, and vice versa. This explanation was found to be in agreement with the empirical data. We thus concluded that a promoter's susceptibility to external noise in activation times is both sequence-dependent and subject to regulation.

Finally, from the literature, the probability with which RNA and/or protein numbers of a gene cross a threshold is quantified from the mean and variance of protein numbers (Eldar & Elowitz 2010; Leibler & Kussell 2010; Raj & Oudenaarden 2008; Thattai & van Oudenaarden 2001; Thattai & Oudenaarden 2004). However, this probability would differ if the single-cell distribution of these numbers could change its asymmetry and tailedness. In **Publication IV and V**, we investigated whether there are asymmetry and tailedness in the distributions of time intervals between two consecutive RNA productions in individual cells and whether (and by which degree) the rate-limiting steps of transcription initiation can control these parameters.

To study this, first, we considered the stochastic model of transcription initiation and investigated how the asymmetry in the distribution of RNA and protein numbers differ within realistic ranges of parameter

values. Second, we performed live, single-cell, single-RNA microscopy measurements (time series and cell populations) and qPCR in live individual cells, in various conditions, including multiple promoters and induction schemes.

From the data, we measured the asymmetry (assessed by skewness) and tailedness (assessed by kurtosis) and showed that it differs between the conditions. In addition, we showed that it is independent of the mean and that they are sequence-dependent (by comparing the kinetics of multiple promoters) and subject to regulation (by comparing the kinetics of genes subject to various induction schemes and regulatory molecules). Next, we dissected how skewness and kurtosis in RNA and protein number distribution can be regulated by tuning the kinetics of the rate-limiting steps in transcription initiation. In particular, we showed that by tuning the skewness of the transcription initiation kinetics, within realistic parameter values, one observes modifications in protein numbers dynamics strong enough to, likely, affect the behavior of small genetic circuits.

Overall, we concluded that skewness and kurtosis are tunable via the regulation of rate-limiting which is both evolvable and adaptable. As such, the study should be of interest to a wide audience of researchers using experimental and/or theoretical methods to study gene regulatory mechanisms and genetic circuits, as it introduces another parameter of relevance in the regulation of threshold crossing dynamics.

## 6 DISCUSSION

Over the last few decades, the regulation of gene expression at the transcription level has been a major goal of Molecular Biology. The work presented in this thesis adds to this effort with new findings on how to dissect the different steps in transcription initiation and how to make use of the kinetics of these steps to better fine-regulate regulate gene expression in *E. coli*.

So far, our methodology to dissect transcription initiation as only be applied to synthetic promoters, whose expression is under control of an inducible promoter. It would be of interest in the future to study how this methodology works when applied to natural genes, e.g. genes associated to cold-shock response in *E. coli*. For example, when cells are subjected to such temperatures, what is different in the kinetics of initiation of cold-shock genes compared to general responses? This can be answered by using the strategy proposed in this thesis. Answers to this question will inform on whether fine-tuning of the kinetics of rate-limiting steps in transcription initiation is used as a means to provide robustness to cold-shock genes, or a means to ensure that they have a fast response. One interesting question in this regard is how have such genes evolved so that some cold-shock genes have a short-term, while others have a long-term response? Do they have opposing patterns of transcription initiation dynamics (e.g. the former have fast while the latter have slow close complex formation?).

Similarly, we expect that rate-limiting steps of closely space promoters will be significantly different, when compared to isolated promoters, e.g., to reduce transcription interference. Also, do these cells implement different kinetics of initiation in genes with one copy number and genes with multiple copies? Or, are these means of regulation to control cases where the gene becomes multi-copy?

Further, having observed that changes in gene expression in response to other  $\sigma$  factors populations differ from gene to gene, due to their distinctive kinetics of transcription initiations combined with the  $\sigma$  factors selectivity process. To what extent does this affect small genetic networks e.g. toggle switch, oscillator, genetic clocks, etc.? One can argue that, if the components genes in the network respond in opposite manner to sigma factor fluctuations, depending on their initiation kinetics, these networks may have a wider range of dynamic responses than previously suggested. It also opens new possibility in the field of synthetic engineering of circuits. We hypothesize that cells may have evolved the initiation kinetics of some component genes as a means to govern reactivity to stress conditions, as a means to enhance robustness.



Overall, we expect our studies to contribute to the development of more sophisticated synthetic genes and circuits with tailored RNA and protein production dynamics, by making use of fine-tuning of the dynamics of the rate-limiting steps in transcription initiation. We expect this to be a key strategy that will assist biomedicine and biotechnology in the effort of regulating cellular behavior for pharmaceutical purposes, industrial output enhancement, etc.

## 7 REFERENCES

- Acar, M., Mettetal, J.T. & Van Oudenaarden, A., 2008. Stochastic switching as a survival strategy in fluctuating environments. *Nature Genetics*, 40(4), pp.471–475.
- Agustino Martinez-Antonio & Collado-vides, J., 2003. Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology*, (6), pp.482–489.
- Albert, F.W. et al., 2014. Genetic Influences on Translation in Yeast. *PLoS Genetics*, 10(10).
- Alberts, B. et al., 2002. *Molecular Biology of the Cell*, New York: Garland Science.
- Alon, U., 2007. Network motifs: theory and experimental approaches. *Nature reviews. Genetics*, 8(6), pp.450–61.
- Arkin, A., Ross, J. & Mcadams, H.H., 1998. Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in. *Genetic Society of America*, 149(4), pp.1633–48.
- Arndt, K.M. & Chamberlin, M.J., 1988. Transcription Termination in *Escherichia coli* Measurement of the Rate of Enzyme Release From Rho-independent Terminators. *Journal of Molecular Biology*, 202, pp.271–285.
- Arsene, F., Tomoyasu, T. & Bukau, B., 2000. The heat shock response of *Escherichia coli*. *International Journal of Food Microbiology*, 55, pp.3–9.
- Artsimovitch, I. et al., 2004. Structural Basis for Transcription Regulation by Alarmone ppGpp. *Cell*, 117, pp.299–310.
- Artsimovitch, I. & Landick, R., 2000. Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals. *Proceedings of the National Academy of Sciences*, 97(13), pp.7090–7095.
- Babu, M.M. & Teichmann, S.A., 2003. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Research*, 31(4), pp.1234–1244.
- Bakshi, S. et al., 2012. Superresolution imaging of ribosomes and RNA polymerase in live *Escherichia coli* cells. *Molecular Microbiology*, 85, pp.21–38.
- Battesti, A., Majdalani, N. & Gottesman, S., 2011. The RpoS-Mediated General Stress Response in *Escherichia coli*. *Annual Review of Microbiology*, 65, pp.189–213.
- Bernardi, A. & Spahr, P.-F., 1972. Nucleotide Sequence at the Binding Site for Coat Protein on RNA of Bacteriophage R17. *Proceedings of the National Academy of Sciences*, 69(10), pp.3033–3037.
- Bernstein, J.A. et al., 2002. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proceedings of the National Academy of Sciences*, 99(15), pp.9697–9702.
- Bertrand-Burggraf, E. et al., 1984. Effect of superhelicity on the transcription from the tet promoter of pBR322. Abortive initiation and unwinding experiments. *Nucleic Acids Research*, 12(20), pp.7741–7752.
- Bertrand, E. et al., 1998. Localization of ASH1 mRNA Particles in Living Yeast. *Molecular Cell*, 2, pp.437–445.
- Blattner, F.R. et al., 1997. The complete genome sequence of *Escherichia coli* K-12. *Science*, 277(5331), pp.1453–1462.
- Brewster, R.C. et al., 2014. The Transcription Factor Titration Effect Dictates Level of Gene Expression. *Cell*, pp.1–12.
- Brewster, R.C., Jones, D.L. & Phillips, R., 2012. Tuning Promoter Strength through RNA Polymerase Binding Site Design in *Escherichia coli*. *PLOS Computational Biology*, 8(12).
- Brown, M. & Wittwer, C., 2000. Flow Cytometry : Principles and Clinical Applications in Hematology. *Clinical Chemistry*, 46(8), pp.1221–1229.
- Browning, D. et al., 2009. Assays for Transcription Factor Activity. In *Methods in Molecular Biology*. pp. 369–387.
- Browning, D.F. & Busby, S.J.W., 2016. Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*, 14(10), pp.638–650.

- Browning, D.F., Busby, S.J.W. & Correspondence, S.J.W.B., 2004. The Regulation Of Bacterial Transcription Initiation. *Nature Reviews Microbiology*, 2, pp.1–9.
- Buc, H. & McClure, W.R., 1985. Kinetics of open complex formation between *Escherichia coli* RNA polymerase and the lac UV5 promoter . Evidence for a sequential mechanism involving three steps Kinetics of Open Complex Formation between *Escherichia coli* RNA Polymerase and the lac UV5 Pr. *Biochemistry*, 24(11), pp.2712–2723.
- Burmann, B.M. et al., 2010. A NusE:NusG Complex Links Transcription and Translation. *Science*, 328, pp.501–503.
- Bury-Mone, S. & Sclavi, B., 2017. Stochasticity of gene expression as a motor of epigenetics in bacteria : from individual to collective behaviors St. *Research in Microbiology*, 168, pp.503–514.
- C.Prashera, D. et al., 1992. Primary structure of the Aequorea victoria green-fluorescent protein. *Gene*, 111(2), pp.229–233.
- Carpenter, J. & Bithell, J., 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19, pp.1141–1164.
- Carpousis, A.J. & Gralla, J.D., 1985. Carpousis1985 Interaction of RNA Polymerase with lacUV5 Promoter DNA during mRNA initiation and elongation. *Journal of Molecular Biology*, 183, pp.165–177.
- Chamberlin, M.J., 1974. The selectivity of transcription. *Annual review of biochemistry*, pp.727–725.
- Chatterji, D. et al., 2007. The role of the omega subunit of RNA polymerase in expression of the relA gene in *Escherichia coli*. *FEMS Microbiology Letters*, (267), pp.51–55.
- Chen, H. et al., 2015. Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*. *Molecular Systems Biology*, 11(781), pp.1–10.
- Choi, P. et al., 2008. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science*, 322, pp.442–446.
- Chong, S. et al., 2014. Mechanism of Transcriptional Bursting in Bacteria. *Cell*, 158(2), pp.314–326.
- Chowdhury, S. et al., 2013. Cell segmentation by multi-resolution analysis and maximum likelihood estimation (MAMLE). *BMC Bioinformatics*, 14(Suppl 10), p.S8.
- Choy, H.E. et al., 1995. Repression and activation of transcription by Gal and Lac repressors : involvement of alpha subunit of RNA polymerase. *EMBO Journal*, 14(18), pp.4523–4529.
- Conn, A.B. et al., 2019. Two Old Dogs , One New Trick : A Review of RNA Polymerase and Ribosome Interactions during Transcription-Translation Coupling. *international journal of Molecular sciences*, 20(2595), pp.1–14.
- Crick, F., 1970. Central Dogma of Molecular Biology. *Nature*, 227, pp.561–563.
- D. N. Joanes & C. A. Gill, 1998. Comparing Measures of Sample Skewness and Kurtosis. *Royal Statistical Society*, 47(1), pp.183–189.
- Daigle, N. & Ellenberg, J., 2007. 1N -GFP : an RNA reporter system for live-cell imaging. *Nature Methods*, 4(8), pp.633–636.
- Daya, R.N. & Davidson, M.W., 2009. The fluorescent protein palette: tools for cellular imaging. *Chem Soc Rev*, 38(10), pp.2887–2921.
- deHaseth, P.L., Zupancic, M.L. & Record, M.T., 1998. MINIREVIEW RNA Polymerase-Promoter Interactions : the Comings and Goings of RNA Polymerase. *Journal of bacteriology*, 180(12), pp.3019–3025.
- Deng, S., Stein, R.A. & Higgins, N.P., 2006. Organization of supercoil domains and their reorganization by transcription. *Molecular Microbiology*, 57(6), pp.53–66.
- Dennis, P.P. et al., 2009. Varying Rate of RNA Chain Elongation during rrn Transcription in *Escherichia coli* . *Journal of Bacteriology*, 191(11), pp.3740–3746.
- DiCiccio, T.J. & Efron, B., 1996. Bootstrap confidence intervals. *Statistical Science*, 11(3), pp.189–228.
- Dillon, S.C. & Dorman, C.J., 2010. Bacterial nucleoid-associated proteins , nucleoid structure and gene expression. *Nature Reviews Microbiology*, 8(3), pp.185–195.
- Dong, T. & Schellhorn, H.E., 2009. Global effect of RpoS on gene expression in pathogenic *Escherichia coli* O157:H7 strain EDL933. *BMC Genomics*, 10, pp.1–17.
- Duchi, D. et al., 2018. The RNA polymerase clamp interconverts dynamically among three states and is stabilized in a partly closed state by ppGpp. *Nucleic Acids Research*, 46(14), pp.7284–7295.

- E.Mussoa, R. et al., 1977. Dual control for transcription of the galactose operon by cyclic AMP and its receptor protein at two interspersed promoters. *Cell*, 12(3), pp.847–854.
- Ebright, R.H., 1993. Transcription activation at Class I CAP-dependent promoters. *Molecular Microbiology*, 8, pp.797–802.
- Ebright, R.H. & Busby, S., 1995. The Escherichia coli RNA polymerase subunit : structure and function. *Current Opinion in Genetics and Development*, (5), pp.197–203.
- Eldar, A. & Elowitz, M.B., 2010. Functional roles for noise in genetic circuits. *Nature*, 467(7312), pp.167–173.
- Elowitz, M.B. et al., 2002. Stochastic gene expression in a single cell. *Science*, 297(5584), pp.1183–1186.
- Elowitz, M.B. & Leibler, S., 2000. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403, pp.335–338.
- Engl, C., 2018. Noise in bacterial gene expression. *Biochemical Society Transactions*, pp.1–9.
- Erie, D.A. et al., 1993. Multiple RNA Polymerase Conformations and GreA : Control of the Fidelity of Transcription. *Science*, 262, pp.867–873.
- Eugene V. Koonin, 2009. Evolution of Genome Architecture. *Int J Biochem Cell Biol*, 41(2), pp.298–306.
- F.C. Neidhardt, Ingraham, J.L. & Schaechter, M., 1991. Physiology of the Bacterial Cell - A Molecular Approach. *Trends in Genetics*, 7(10), p.341.
- Farewell, A., Kvint, K. & Nyström, T., 1998. Negative regulation by RpoS: A case of sigma factor competition. *Molecular Microbiology*, 29(4), pp.1039–1051.
- Fish, K.N., 2015. Total Internal Reflection Fluorescence (TIRF) Microscopy Kenneth. *Curr Protoc Cytom.*, (2), pp.1–21.
- Frigault, M.M. et al., 2009. Live-cell microscopy – tips and tools. *Journal of cell science*, 122(6).
- Fritz, G. et al., 2014. Single cell kinetics of phenotypic switching in the arabinose utilization system of *E. coli*. *PLoS ONE*, 9(2), p.e89532.
- Fusco, D. et al., 2003. Single mRNA Molecules Demonstrate Probabilistic Movement in Living Mammalian Cells. *Current biology*, 23(13), pp.2512–2518.
- Gaal, T. et al., 2001. Promoter recognition and discrimination by EsS RNA polymerase. *Molecular Microbiology*, 42(4), pp.939–954.
- Gabizon, R. et al., 2018. Pause sequences facilitate entry into long-lived paused states by reducing RNA polymerase transcription rates. *Nature Communications*, 9(2930), pp.1–10.
- Garcia, H.G. et al., 2010. Transcription by the numbers redux : experiments and calculations that surprise. *Trends in Cell Biology*, 20(12), pp.723–733.
- Gasnier, M. et al., 2013. Fluorescent mRNA labeling through cytoplasmic FISH. *Nature Protocols*, 8(12), pp.4–13.
- Ghosh, P., Ishihama, A. & Chatterji, D., 2001. *Escherichia coli* RNA polymerase subunit  $\nu$  and its N-terminal domain bind full-length  $\beta$  H to facilitate incorporation into the  $\alpha 2 \beta$  subassembly. *Eur. j. Biochem*, pp.4621–4627.
- Gibson, D.G. et al., 2010. Chemical synthesis of the mouse mitochondrial genome. *Nature Methods*, 7(11), pp.11–15.
- Gibson, D.G. et al., 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, 6(5), pp.12–16.
- Gillespie, D.T., 1992. A rigorous derivation of the chemical master equation. *Physica A*, 188, pp.404–425.
- Gillespie, D.T., 1976. Gillespie1976 A general method for numerically simulating coupled chemical reactions.pdf. *Journal of Computational Physics*, 22, pp.403–434.
- Gillespie, D.T., 2007. Stochastic Simulation of Chemical Kinetics. *Annu. Rev. Phys. Chem*, 58, pp.35–55.
- Gillespie, D.T., 1977. Exact Stochastic Simulation of Coupled Chemical Reactions Danlel. *Journal of Physical Chemistry*, 81(25), pp.2340–2361.
- Goldberg, A.L., 1972. Degradation of Abnormal Proteins in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 69(2), pp.422–426.
- Golding, I. et al., 2005. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6), pp.1025–1036.
- Golding, I. & Cox, E.C., 2004. RNA dynamics in live *Escherichia coli* cells. *Proceedings of the National Academy of Sciences*, 101(31), pp.11310–11315.

- Goldman, S.R., Ebright, R.H. & Nickels, B.E., 2009. Direct Detection of Abortive RNA Transcripts in Vivo. *Science*, 324, pp.927–928.
- Gonçalves, N. et al., 2016. Temperature Dependence of Leakiness of Transcription Repression Mechanisms of *Escherichia coli*. In *Computational Methods in Systems Biology - 14th International Conference, CMSB*. Springer Verlag, pp. 341–342.
- Gralla, J.D., Carpousis, A.J. & Stefano, J.E., 1980. Productive and Abortive Initiation of Transcription in Vitro at the. *Biochemistry*, 19(25), pp.5864–5869.
- Greive, S.J. & Hippel, P.H. Von, 2005. Thinking Quantitatively About Transcriptional Regulation. *Nature Reviews Molecular cell biology*, 6, pp.221–232.
- Greive, S.J., Lins, A.F. & Hippel, P.H. Von, 2005. Assembly of an RNA-Protein complex binding of NUSb and NUSe (s10) proteins to b ox a rna nucleates the formation of the antitermination complex involved in controlling rRNA transcription in *Escherichia coli*\*. *Journal of biological chemistry*, 280(43), pp.36397–36408.
- Grigorova, I.L. et al., 2006. Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA. *Proceedings of the National Academy of Sciences*, 103(14), pp.5332–5337.
- Gummesson, B. et al., 2009. Increased RNA polymerase availability directs resources towards growth at the expense of maintenance. *EMBO Journal*, 28(15), pp.2209–2219.
- Gusarov, I. & Nudler, E., 1999. The Mechanism of Intrinsic Transcription Termination site (HBS), and the single-stranded RNA-binding site (RBS). DBS was defined as a region of strong nonionic interactions with 9 bp of DNA just downstream of the. *Molecular Cell*, 3, pp.495–504.
- Ha, T. & Tinnefeld, P., 2012. Photophysics of Fluorescent Probes for Single-Molecule Biophysics and Super-Resolution Imaging. *Annu. Rev. Phys. Chem*, 63, pp.595–617.
- Hakkinen, A. et al., 2013. CellAging : A tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*. *Bioinformatics*, 29(13), pp.1708–1709.
- Hakkinen, A. et al., 2014. Estimation of fluorescence-tagged RNA numbers from spot intensities. *Bioinformatics*, pp.1–8.
- Hakkinen, A. & Ribeiro, A.S., 2015. Gene expression Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data. *Bioinformatics*, 31(1), pp.69–75.
- Harden, T.T. et al., 2016. Bacterial RNA polymerase can retain  $\sigma 70$  throughout transcription. *Proceedings of the National Academy of Sciences*, 113(3), pp.602–607.
- Hardy, C.D. & Cozzarelli, N.R., 2005. A genetic selection for supercoiling mutants of *Escherichia coli* reveals proteins implicated in chromosome structure. *Molecular Microbiology*, 57, pp.1636–1652.
- Harley, C.B. & Reynolds, R.P., 1987. Analysis of *E. coli* promoter sequences. *Nucleic Acids Research*, 15(5), pp.2343–2361.
- Henderson, K.L. et al., 2017. Mechanism of transcription initiation and promoter escape by *E. coli* RNA polymerase . *Proceedings of the National Academy of Sciences*, 114(15), pp.E3032–E3040.
- Hengge-aronis, R., 2002a. Recent Insights into the General Stress Response Regulatory Network in *Escherichia coli*. *Journal of Microbiol Biotechnology*, 4(3), pp.341–346.
- Hengge-aronis, R., 2002b. Stationary phase gene regulation : what makes an *Escherichia coli* promoter  $\sigma S$  - selective ? *Current Opinion in Microbiology*, 5, pp.591–595.
- Herbert, K.M. et al., 2010. *E. coli* NusG Inhibits Backtracking and Accelerates Pause-Free Transcription by Promoting Forward Translocation of RNA Polymerase. *Journal of Molecular Biology*, 399(1), pp.17–30.
- Herna, A.M. et al., 2009. Repressor CopG prevents access of RNA polymerase to promoter and actively dissociates open complexes. *Nucleic Acids Research*, 37(14), pp.4799–4811.
- Hippe, P.H. Von, Mcswiggen, J.A. & Morgan, W.D., 1984. Protein-Nucleic Acid Interactions In Transcription : The Closed-to-Open Promoter Transition and Analysis of the Open. *Annual review of biochemistry*, 53, pp.389–446.
- Hocine, S. et al., 2013. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nature Methods*, 10(2), pp.119–122.
- Hollands, K., Sevostyanova, A. & Groisman, E.A., 2014. Unusually long-lived pause required for regulation of a Rho-dependent transcription terminator. *Proceedings of the National Academy of Sciences*, pp.1999–2007.

- Holmes, V.F. & Cozzarelli, N.R., 2000. Closing the ring: Links between SMC proteins and chromosome partitioning, condensation, and supercoiling. *Proceedings of the National Academy of Sciences*, 97(15), pp.1322–1324.
- Horan, L.H. & Noller, H.F., 2007. Intersubunit movement is required for ribosomal translocation. *Proceedings of the National Academy of Sciences*, 104(12), pp.4881–4885.
- Hsu, L.M., 2002. Promoter clearance and escape in prokaryotes. *Biochimica et Biophysica Acta*, 1577, pp.191–207.
- Huang, B., Bates, M. & Zhuang, X., 2009. Super-Resolution Fluorescence Microscopy. *Annu. Rev. Biochem*, 78, pp.993–1016.
- Huber, D., Voithenberg, L.V. Von & Kaigala, G. V, 2018. Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro and Nano Engineering*, 1, pp.15–24.
- Hufnagel, D.A., Depas, W.H. & Chapman, M.R., 2015. The Biology of the *Escherichia coli* Extracellular Matrix. *Microbiology Spectrum*, 3(3), pp.1–24.
- Huh, D. & Paulsson, J., 2011. Random partitioning of molecules at cell division. *Proceedings of the National Academy of Sciences*, 108(36), pp.15004–15009.
- Ishihama, A., 2000. Functional Modulation Of *Escherichia coli* RNA Polymerase. *Annual Review of Microbiology*, 54, pp.499–518.
- Jacob, F. & Monod, J., 1960. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *Journal of Molecular Biology*, (3), pp.318–356.
- John P Richardson, 1991. Preventing the synthesis of unused transcripts by rho factor. *Cell*, 64(6), pp.1047–1049.
- Jones, D.L., Brewster, R.C. & Phillips, R., 2014. Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346(6216), pp.1533–1536. Available at: <http://www.sciencemag.org/cgi/doi/10.1126/science.1255301>.
- Kaczanowska, M. & Ryde, M., 2007. Ribosome Biogenesis and the Translation Process in *Escherichia coli*. *Microbiology And Molecular Biology Reviews*, 71(3), pp.477–494.
- Kærn, M. et al., 2005. Stochasticity in gene expression: From theories to phenotypes. *Nature Reviews Genetics*, 6(6), pp.451–464.
- Kandavalli, V.K., Tran, H. & Ribeiro, A.S., 2016. Effects of  $\sigma$  factor competition are promoter initiation kinetics dependent. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1859(10), pp.1281–1288.
- Kane, M., Kerppola, K. & Rna, C.M., 1991. RNA polymerase: regulation of transcript elongation and elongation. *FASEB*, 5, pp.2833–2842.
- Kapanidis, A.N. et al., 2006. Initial Transcription by RNA Polymerase Proceeds Through a DNA-Scrunching Mechanism. *Science*, 314, pp.1144–1147.
- Kapanidis, A.N. et al., 2005. Retention of Transcription Initiation Factor  $\sigma^{70}$  in Transcription Elongation: Single-Molecule Analysis. *Molecular Cell*, 20, pp.347–356.
- Kireeva, M.L. & Kashlev, M., 2009. Mechanism of sequence-specific pausing of bacterial RNA polymerase. *Proceedings of the National Academy of Sciences*, 106(22), pp.1–6.
- Klumpp, S. & Hwa, T., 2008. Growth-rate-dependent partitioning of RNA polymerases in bacteria. *Proceedings of the National Academy of Sciences*, 105(51), pp.20245–20250.
- Komissarova, N. & Kashlev, M., 1997. Transcriptional arrest: *Escherichia coli* RNA polymerase translocates backward, leaving the 3' end of the RNA intact. *Proceedings of the National Academy of Sciences*, 94, pp.1755–1760.
- Koslover1, D.J. et al., 2012. Binding and Translocation of Termination Factor Rho Studied at the Single-Molecule Level. *Journal of General Microbiology*, 423(5), pp.664–676.
- Kubitschek, H.E., 1990. Cell Volume Increase in *Escherichia coli* after Shifts. *Journal of Bacteriology*, 172(1), pp.94–101.
- Landick, R., 2006. The regulatory roles and mechanism of transcriptional pausing. *Biochemical Society Transactions*, 34, pp.1062–1066.
- Landick, R., 2009. Transcriptional pausing without backtracking. *Proceedings of the National Academy of Sciences*, 106(22), pp.8797–8798.

- Larson, D.R. et al., 2011. Real-Time Observation of Transcription Initiation and Elongation on an Endogenous Yeast Gene. *Science*, 332(475).
- Lee, D.J., Minchin, S.D. & Busby, S.J.W., 2012. Activating Transcription in Bacteria. *Annual Review of Microbiology*, 66, pp.125–152.
- Leibler, S. & Kussell, E., 2010. Individual histories and selection in heterogeneous populations. *Proceedings of the National Academy of Sciences*, 107(29), pp. 13183–13188.
- Leng, F., Chen, B. & Dunlap, D.D., 2011. Dividing a supercoiled DNA molecule into two independent topological domains. *Proceedings of the National Academy of Sciences*, 108(50), pp.19973–19978.
- Lenstra, T.L. & Larson, D.R., 2016a. Single molecule mRNA detection in live yeast. *Current Protocol in Molecular Biology*, 113:14.24.1-14.24.15.
- Lenstra, T.L. et al., 2016b. Transcription Dynamics in Living cells. *Annual Review of Biophysics*, 45, pp.25–47.
- Leon, F.G. De et al., 2017. Tracking Low-Copy Transcription Factors in Living Bacteria : The Case of the lac Repressor. *Biophysical Journal*, 112(7), pp.1316–1327.
- Levsky, J.M. & Singer, R.H., 2003. Fluorescence in situ hybridization : past , present and future. *Journal of cell science*, 116, pp.2833–2838.
- Lewis, D.E.A. & Adhya, S., 2015. Molecular Mechanisms of Transcription Initiation at gal Promoters and their Multi-Level Regulation by GalR, CRP and DNA Loop. *biomolecules*, 5, pp.2782–2807.
- Lewis, M., 2005. The lac repressor. *C. R. Biologies*, 328, pp.521–548.
- Lim, F. & Peabody, D.S., 2002. RNA recognition site of PP7 coat protein. *Nucleic Acids Research*, 30(19), pp.4138–4144.
- Lineweaver, H. & Burk, D., 1934. The Determination of Enzyme Dissociation Constants. *Journal of the American Chemical Society*, (56), pp.658–666.
- Livak, K.J. & Schmittgen, T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2-DDCT method. *Methods*, 25(4), pp.402–408.
- Llopis, P.M. et al., 2010. Spatial organization of the flow of genetic information in bacteria. *Nature*, 466(7302), pp.77–81.
- Lloyd-Price, J. et al., 2016. Dissecting the stochastic transcription initiation process in live *Escherichia coli*. *DNA Research*, 23(3), pp.203–214.
- López-Maury, L., Marguerat, S. & Bähler, J., 2008. Tuning gene expression to changing environments: From rapid responses to evolutionary adaptation. *Nature Reviews Genetics*, 9(8), pp.583–593.
- Lutz, R. et al., 2001. Dissecting the functional program of *Escherichia coli* promoters: the combined mode of action of Lac repressor and AraC activator. *Nucleic acids research*, 29(18), pp.3873–3881.
- Lutz, R. & Bujard, H., 1997. Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR / O , the TetR / O and AraC / I 1 -I 2 regulatory elements. *Nucleic Acids Research*, 25(6), pp.1203–1210.
- Maeda, H., Fujita, N. & Ishihama, A., 2000. Competition among seven *Escherichia coli* sigma subunits: relative binding affinities to the core RNA polymerase. *Nucleic acids research*, 28(18), pp.3497–503.
- Mäkelä, J. et al., 2013. In vivo single-molecule kinetics of activation and subsequent activity of the arabinose promoter. *Nucleic Acids Research*, 41(13), pp.6544–6552.
- Mäkelä, J. et al., 2011. Stochastic sequence-level model of coupled transcription and translation in prokaryotes. *BMC Bioinformatics*, 12(1), p.121.
- Mäkelä, J., Kandavalli, V. & Ribeiro, A.S., 2017. Rate-limiting steps in transcription dictate sensitivity to variability in cellular components. *Scientific Reports*, 7, pp.1–10.
- Marbach, A. & Bettenbrock, K., 2012. lac operon induction in *Escherichia coli* : Systematic comparison of IPTG and TMG induction and influence of the transacetylase LacA. *Journal of Biotechnology*, 157(1), pp.82–88.
- Margeat, E. et al., 2006. Direct Observation of Abortive Initiation and Promoter Escape within Single Immobilized Transcription Complexes. *Biophysical Journal*, 90(4), pp.1419–1431.
- Martin, F.H. & Tinoco, I., 1980. DNA-RNA hybrid duplexes containing oligo ( dA : rU ) sequences are exceptionally unstable and may facilitate termination of transcription. *Nucleic Acids Research*, 8(10), pp.2295–2300.

- Martin, R.G. et al., 2002. Complex formation between activator and RNA polymerase as the basis for transcriptional activation by MarA and SoxS in *Escherichia coli*. *Molecular Microbiology*, 43(2), pp.355–370.
- Mauri, M. & Klumpp, S., 2014. A Model for Sigma Factor Competition in Bacterial Cells. *PLoS Computational Biology*, 10(10), pp.29–34.
- 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), pp.814–9.
- McClure, R., Cech, L. & David Johnston, E., 1978. State Assay for the RNA Polymerase Initiation Reaction. *Journal of biological chemistry*, 253(24), pp.8941–8948.
- McClure, W., 1985. Mechanism and Control of Transcription Initiation in Prokaryotes. *Annual Review of Biochemistry*, 54(1), pp.171–204.
- McClure, W.R., 1980. Rate-limiting steps in RNA chain initiation. *Proceedings of the National Academy of Sciences*, 77(10), pp.5634–5638.
- Mcgary, K. & Nudler, E., 2013. RNA polymerase and the ribosome : the close relationship. *Current Opinion in Microbiology*, 16(2), pp.112–117.
- Megerle, J.A. et al., 2008. Timing and dynamics of single cell gene expression in the arabinose utilization system. *Biophysical Journal*, 95(4), pp.2103–2115.
- Miller, J.H., 1992. *A Short Course in Bacterial Genetics – A Laboratory Manual and Handbook for Escherichia coli and Related Bacteria.*,
- Muller-Hill, B., 1998. The function of auxiliary operators. *Molecular Microbiology*, 29(1), pp.13–18.
- Murakami, K. et al., 1997. The two  $\sigma$  subunits of *Escherichia coli* RNA polymerase are asymmetrically arranged and contact different halves of the DNA upstream element. *Proceedings of the National Academy of Sciences*, 94, pp.1709–1714.
- Murakami, K.S. & Darst, S.A., 2003. Bacterial RNA polymerases: the whole story. *Current opinion in structural biology*, 13(1), pp.31–9.
- Murakawa, G.J. et al., 1991. Transcription and Decay of the lac Messenger : Role of an Intergenic Terminator. *Journal of Bacteriology*, 173(1), pp.28–36.
- Murray, P.R., Rosenthal, K.S. & Pfaller, M.A., 2009. *Medical Microbiology. 6th Edition Philadelphia: Mosby Elsevier.*
- 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), pp.8472–8483.
- Nakano, A., 2002. Spinning-disk Confocal Microscopy — A Cutting-Edge Tool for Imaging of Membrane Traffic Need for a rapid confocal scanning system. *Cell Structure And Function*, (27), pp.349–355.
- Navarre, W.W. et al., 2006. Selective Silencing of Foreign DNA Protein in Salmonella. *Science*, 682, pp.236–239.
- Neidhardt, F., 1987. *Escherichia coli* and Salmonella typhimurium: cellular and molecular biology. Washington, D.C., American Society for Microbiology.
- Neubauer, Z. & Calef, E., 1970. Immunity Phase-shift in Defective Lysogens : Hereditary Change of Early Regulation of  $\lambda$  Prophage. *Journal of Molecular Biology*, 51, pp.1–13.
- Noller, H.F., 2012. Evolution of Protein Synthesis from an RNA World. *Cold Spring Harbor Laboratory Press*, pp.1–14.
- Nudler, E. & Gottesman, M.E., 2002. Transcription termination and anti-termination in *E. coli*. *Genes to cells*, 7, pp.755–768.
- Oliveira, S.M.D. et al., 2016. Temperature-Dependent Model of Multi- step Transcription Initiation in *Escherichia coli* Based on Live Single-Cell Measurements. *PLoS Computational Biology*, pp.1–18.
- Osborn, A.E. & Field, B., 2009. Operons. *Cellular and Molecular Life Sciences*, 66, pp.3755–3775.
- Otsu, N., 1979. A Threshold Selection Method from Gray-Level Histograms. *IEEE*, 20(1), pp.62–66.
- Patrick, M. et al., 2015. Biochimie Free RNA polymerase in *Escherichia coli*. *Biochimie*, 119, pp.80–91.
- Paul, B.J., Berkmen, M.B. & Gourse, R.L., 2005. DksA potentiates direct activation of amino acid promoters by ppGpp. *Proceedings of the National Academy of Sciences*, 102(22), pp.7823–7828.
- Pawley, J.B., 2006. *Biological Confocal Microscopy Third.*, Springer.
- Peabody, D.S., 1993. The RNA binding site of bacteriophage MS2 coat protein. *EMBO Journal*, 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), pp.1838–1847.
- Peterson, J.R. et al., 2015. Effects of DNA replication on mRNA noise. *Proceedings of the National Academy of Sciences*, 112(52), pp.15886–15891.
- Pitchiaya, S. et al., 2014. Single Molecule Fluorescence Approaches Shed Light on Intracellular RNAs. *Chemical Reviews*.
- Proshkin, S. et al., 2010. Cooperation Between Translating Ribosomes and RNA Polymerase in Transcription Elongation. *Science*, 328, pp.504–507.
- Querido, E. & Chartrand, P., 2008. Using Fluorescent Proteins to Study mRNA Trafficking in Living Cells. *Methods in Cell Biology*, 85(08), pp.273–292.
- Raj, A. & Oudenaarden, A. Van, 2008. Review Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell*, 135, pp.216–225.
- Ramos, J.L. et al., 2001. Responses of Gram-negative bacteria to certain environmental stressors. *Current Opinion in Microbiology*, 4, pp.166–171.
- Raser, J.M. & O’Shea, E.K., 2005. Molecular biology - Noise in gene expression: Origins, consequences, and control. *Science*, 309(5743), pp.2010–2013.
- Revyakin, A. et al., 2006. Abortive Initiation and Productive Initiation by RNA Polymerase Involve DNA Scrunching. *Science*, 314(2006), pp.1139–1143.
- Reyes-lamothe, R. et al., 2014. High-copy bacterial plasmids diffuse in the nucleoid-free space, replicate stochastically and are randomly partitioned at cell division. *Nucleic Acids Research*, 42(2), pp.1042–1051.
- Ribeiro, A.S., 2010. Stochastic and delayed stochastic models of gene expression and regulation. *Mathematical Biosciences*, 223(1), pp.1–11.
- Ribeiro, A.S. & Lloyd-Price, J., 2007. SGN Sim, a Stochastic Genetic Networks Simulator. *Bioinformatics*, 23(6), pp.777–779.
- Ribeiro, A.S., Zhu, R. & Kauffman, S.A., 2006. A general modeling strategy for gene regulatory networks with stochastic dynamics. *Journal of Computational Biology*, 13(9), pp.1630–1639.
- Richardson, J.P., 2002. Rho-dependent termination and ATPases in transcript termination. *Biochimica et Biophysica Acta*, 1577, pp.251–260.
- Rosenberg, M. & Court, D., 1979. Regulatory Sequences Involved In The Promotion Transcription. *Ann. Rev. Genet.*, 13, pp.319–53.
- Ross, W. et al., 2013. Article The Magic Spot: A ppGpp Binding Site on E. coli RNA Polymerase Responsible for Regulation of Transcription Initiation. *Molecular Cell*, 50, pp.1–10.
- Roussel, M.R. & Zhu, R., 2006. Stochastic kinetics description of a simple transcription model. *Bulletin of Mathematical Biology*, 68(7), pp.1681–1713.
- Rovinskiy, N. et al., 2012. Rates of Gyrase Supercoiling and Transcription Elongation Control Supercoil Density in a Bacterial Chromosome. *PLoS*, 8(8).
- Russo, T.A. & Johnson, J.R., 2003. Medical and economic impact of extraintestinal infections due to *Escherichia coli*: focus on an increasingly important endemic problem. *Microbes and Infection*, 5(5), pp.449–56.
- Ruusuvuori, P. et al., 2010. Evaluation of methods for detection of fluorescence labeled subcellular objects in microscope images. *BMC Bioinformatics*, 11(248), pp.1–17.
- Saecker, R.M., Record, M.T. & Dehaseth, P.L., 2011. Mechanism of bacterial transcription initiation: RNA polymerase - Promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. *Journal of Molecular Biology*, 412(5), pp.754–771.
- Sambrook J & D W Russell, 2001. *Molecular cloning: a laboratory manual*. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.,
- Sanamrad, A. et al., 2014. Single-particle tracking reveals that free ribosomal subunits are not excluded from the *Escherichia coli* nucleoid. *Proceedings of the National Academy of Sciences*, 111(31), pp.11413–11418.
- Sanchez, A., Choubey, S. & Kondev, J., 2013. Regulation of Noise in Gene Expression. *Annual Review of Biophysics*, 42, pp.469–491.
- Sanchez, A. & Golding, I., 2013. Genetic Determinants and Cellular Constraints in Noisy Gene Expression. *Science*, 342(1188).

- Santangelo, P.J. et al., 2009. Single molecule – sensitive probes for imaging RNA in live cells. *Nature Methods*, 6(5), pp.10–14.
- Santangelo, T.J. & Artsimovitch, I., 2011. Termination and antitermination: RNA polymerase runs a stop sign. *Nature Reviews Microbiology*, 9(5), pp.319–329. Available at: <http://dx.doi.org/10.1038/nrmicro2560>.
- Schleif, R., 2010. AraC protein, regulation of the L-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action. *FEMS Microbiology Reviews*, pp.1–18.
- Schleif, R., 2000. Regulation of the L-arabinose operon of *Escherichia coli*. *Trends in Genetics*, 16(12), pp.559–565.
- Schmittgen, T.D. & Livak, K.J., 2008. Analyzing real-time PCR data by the comparative CT method. *Nature Protocols*, 3(6), pp.1101–1108.
- Schuwirth, B.S. et al., 2005. Structures of the Bacterial crystallographic restraints in the refinement to Ribosome at 3.5 Å Resolution. *Science*, 310, pp.827–835.
- Shaner, N.C. et al., 2004. Improved monomeric red , orange and yellow fluorescent proteins derived from *Discosoma sp.* red fluorescent protein. *Nature Biotechnology*, 22(12), pp.1567–1572.
- Sheridan, S.D., Benham, C.J. & Hatfield, G.W., 1998. Activation of Gene Expression by a Novel DNA Structural Transmission Mechanism That Requires Supercoiling-induced DNA Duplex Destabilization in an Upstream Activating Sequence. *Journal of bacteriology*, 273, pp.21298–21308.
- Shih, M. & Gussin, G.N., 1983. Mutations affecting two different steps in transcription initiation at the phage  $\lambda$  PRM promoter. *Proceedings of the National Academy of Sciences*, 80, pp.496–500.
- Shimomura, O., 1962. Extraction , Purification and Properties of Aequorin , a Bioluminescent Protein from the Luminous Hydromedusa , *Aequorea*. *Journal of Cellular and Comparative Physiology*, 59(3), p.1962.
- Skinner, S.O. et al., 2013. Measuring mRNA copy-number in individual *Escherichia coli* cells using single-molecule fluorescent in situ hybridization (smFISH). *Nature Protocols*, 8(6), pp.1100–1113.
- Sleator, R.D. & Hill, C., 2001. Bacterial osmoadaptation : the role of osmolytes in bacterial stress and virulence. , 26.
- So, L. et al., 2011. General properties of transcriptional time series in *Escherichia coli*. *Nature Genetics*, pp.1–9.
- Startceva, S. et al., 2019. Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1862, pp.119–128.
- Stoebel, D.M. et al., 2009. Compensatory Evolution of Gene Regulation in Response to Stress by *Escherichia coli* Lacking RpoS. , 5(10), pp.1–9.
- Stracy, M. et al., 2015. Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid. *Proceedings of the National Academy of Sciences*, p.E4390–E4399 |.
- Stracy, M. & Kapanidis, A.N., 2017. Single-molecule and super-resolution imaging of transcription in living bacteria. *Methods*, 120, pp.103–114.
- Straney, S.B. & Crothers, D.M., 1987. Kinetics of the Stages of Transcription Initiation at the. *Biochemistry*, pp.5063–5070.
- Susa, M., Kubori, T. & Shimamoto, N., 2006. A pathway branching in transcription initiation in *Escherichia coli*. *Molecular Microbiology*, 59, pp.1807–1817.
- Tagami, S., Sekine, S. & Yokoyama, S., 2011. A novel conformation of RNA polymerase sheds light on the mechanism of transcription. *Transcription*, 2(4), pp.162–167.
- Tanaka, K. et al., 1995. Promoter determinants for *Escherichia coli* RNA polymerase holoenzyme containing a sigma 38 ( the rpoS gene product ). *Nucleic Acids Research*, 23(5), pp.827–834.
- Taniguchi, Y. et al., 2010. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991), pp.533–538.
- Thattai, M. & van Oudenaarden, A., 2001. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15), pp.8614–8619.
- Thattai, M. & Oudenaarden, A. Van, 2004. Stochastic Gene Expression in Fluctuating Environments. *Genetics*, 530(May), pp.523–530.
- Tokunaga, M., Imamoto, N. & Sakata-sogawa, K., 2008. Highly inclined thin illumination enables clear single-molecule imaging in cells. *Nature Methods*, 5(2), pp.159–161.
- Tran, H. et al., 2015. Molecular BioSystems Kinetics of the cellular intake of a gene expression inducer at high concentrations. *Molecular BioSystems*, 11, pp.2579–2587.

- Tripathi, L., Zhang, Y. & Lin, Z., 2014. Bacterial sigma factors as targets for engineered or synthetic transcriptional control. *Frontiers in Bioengineering And Biotechnology Mini*, 2(33), pp.1–7.
- Tsien, R.Y., 1998. The Green Fluorescent Protein. *Annu. Rev. Biochem*, 67, pp.509–44.
- Uptain, S.M. & Kane, C.M., 1997. Basic Mechanisms Of Transcript Elongation. *Annu. Rev. Biochem*, 66, pp.117–72.
- Vogel, U. & Jensen, K.A.J.F., 1994. The RNA chain elongation rate in *Escherichia coli* depends on the growth rate . The RNA Chain Elongation Rate in *Escherichia coli* Depends on the Growth Rate. *Journal of Bacteriology*, 176(10), pp.2807–2812.
- Volkmer, B. & Heinemann, M., 2011. Condition-Dependent cell volume and concentration of *Escherichia coli* to facilitate data conversion for systems biology modeling. *PLoS ONE*, 6(7), pp.1–6.
- Wachter, R.M. & Remington, S.J., 1999. Sensitivity of the yellow variant of green fluorescent protein to halides and nitrate. *Current biology*, 9(17), pp.628–629.
- Wade, J.T. & Grainger, D.C., 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nature Publishing Group*, 12(9), pp.647–653. Available at: <http://dx.doi.org/10.1038/nrmicro3316>.
- Walter, G. et al., 1967. Initiation of DNA-Dependent RNA Synthesis and the Effect of Heparin on RNA Polymerase. *European J. Biochem*, 3, pp.194–201.
- Webb, D.J. & Brown, C.M., 2012. Epi-Fluorescence Microscopy. In D. J. Taatjes & J. Roth, eds. *Cell Imaging Techniques*. Humana Press, Totowa, NJ, p. 931.
- Weiss, A. et al., 2017. The sigma Subunit Governs RNA Polymerase Stability and Transcriptional Specificity in *Staphylococcus aureus*. *Journal of Bacteriology*, 199(2), pp.1–16.
- William S Reznikoff et al., 1985. The Regulation Of Transcription Initiation In Bacteria. *Ann. Rev. Genet.*, 19, pp.355–87.
- Wilson, K.S., 1995. Transcription termination at intrinsic terminators : The role of the RNA hairpin. *Proceedings of the National Academy of Sciences*, 92, pp.8793–8797.
- Wu, B. et al., 2011. Modern fluorescent proteins and imaging technologies to study gene expression, nuclear localization, and dynamics. *Curr Opin Cell Biol*, 23(3), pp.310–317.
- Xie, X.S. et al., 2008. Single-Molecule Approach to Molecular Biology in Living Bacterial Cells. *Annual Review of Biophysics*, 37(1), pp.417–444.
- Xu, J. & Liu, Y., 2018. Stochastic optical reconstruction microscopy (STORM). *Curr Protoc Cytom.*, 81(12), pp.1–38.
- Yakhnin, A. V, Murakami, K.S. & Babitzke, P., 2016. NusG Is a Sequence-specific RNA Polymerase Pause Factor That Binds to the Non-template DNA within the Paused Transcription Bubble. *Journal of biological chemistry*, 291(10), pp.5299–5308.
- Yamanaka, K., 1999. Cold Shock Response in *Escherichia coli* Real-Time PCR Bioinformatics and Data. *J. Mol. Microbiol. Biotechnol*, 1(2), pp.193–202.
- Yang, S. et al., 2014. variation to protein expression noise. *Nature Communications*, 5, pp.1–9.
- Yu, J. et al., 2006. Probing gene expression in live cells, one protein molecule at a time. *Science*, 311(5767), pp.1600–1603.
- Zernike F, 1942. Phase contrast, a new method for the microscopic observation of transparent objects part II. *Physica*, 9(10), p.80079.
- Zhao, K., Liu, M. & Burgess, R.R., 2010. Promoter and regulon analysis of nitrogen assimilation factor , p 54 , reveal alternative strategy for E . coli MG1655 flagellar biosynthesis. *Nucleic Acids Research*, 38(4), pp.1273–1283.



# PUBLICATION

I

**Dissecting the stochastic transcription initiation process in live**

***Escherichia coli***

J. Lloyd-Price, S. Startceva, V. Kandavalli, J.G. Chandraseelan, N. Goncalves, S.M.D  
Oliveira, A. Häkkinen and A.S. Ribeiro

DNA Research. 23 (3): 203-214, 2016

doi: 10.1093/dnares/dsw009

**Publication reprinted with the permission of the copyright holders.**



Full Paper

# Dissecting the stochastic transcription initiation process in live *Escherichia coli*

Jason Lloyd-Price, Sofia Startceva, Vinodh Kandavalli,  
Jerome G. Chandraseelan, Nadia Goncalves,  
Samuel M. D. Oliveira, Antti Häkkinen, and Andre S. Ribeiro\*

Laboratory of Biosystem Dynamics, Department of Signal Processing, Tampere University of Technology, PO Box 553, Office TC336, 33101 Tampere, Finland

\*To whom correspondence should be addressed. Tel. +358 408490736. Fax. +358 331154989. Email: andre.ribeiro@tut.fi

Edited by Prof. Kenta Nakai

Received 1 December 2015; Accepted 11 February 2016

## Abstract

We investigate the hypothesis that, in *Escherichia coli*, while the concentration of RNA polymerases differs in different growth conditions, the fraction of RNA polymerases free for transcription remains approximately constant within a certain range of these conditions. After establishing this, we apply a standard model-fitting procedure to fully characterize the *in vivo* kinetics of the rate-limiting steps in transcription initiation of the  $P_{lac/ara-1}$  promoter from distributions of intervals between transcription events in cells with different RNA polymerase concentrations. We find that, under full induction, the closed complex lasts  $\sim 788$  s while subsequent steps last  $\sim 193$  s, on average. We then establish that the closed complex formation usually occurs multiple times prior to each successful initiation event. Furthermore, the promoter intermittently switches to an inactive state that, on average, lasts  $\sim 87$  s. This is shown to arise from the intermittent repression of the promoter by LacI. The methods employed here should be of use to resolve the rate-limiting steps governing the *in vivo* dynamics of initiation of prokaryotic promoters, similar to established steady-state assays to resolve the *in vitro* dynamics.

**Key words:** free RNA polymerase, *in vivo* transcription dynamics, rate-limiting steps, reversible closed complex formation, repressor binding dynamics

## 1. Introduction

Gene expression has been intensively studied with the relatively new tools provided by fluorescent proteins and microscopy techniques with single-molecule resolution, in both prokaryotic<sup>1–5</sup> and eukaryotic<sup>6,7</sup> systems. These studies have established that this process cannot be fully characterized by the mean protein production rate,<sup>8–12</sup> since cells exhibit fluctuations (i.e. noise) over time and diversity in numbers across populations,<sup>13</sup> which, among other things, generates phenotypic diversity.<sup>8</sup> The noise has generally been investigated through indirect means, such as by observing the diversity in RNA and protein numbers in cell populations.<sup>2,3,10,11,14</sup> Other, more direct means

consist of observing the distribution of intervals between RNA productions<sup>2,4,5</sup> and between protein bursts in individual cells.<sup>3,15</sup>

From these observations, a wide range of gene expression behaviours have been reported and, therefore, significantly different probabilistic models of transcription have been proposed.<sup>2,4,16–18</sup> In general, higher-than-Poissonian variability in RNA numbers has been explained by models in which the promoter intermittently switched into an inactive state, resulting in bursty RNA production dynamics.<sup>2,16,19</sup> Meanwhile, lower-than-Poissonian variability appears to be more consistent with models assuming multiple rate-limiting steps.<sup>4,5,16,20,21</sup>

There is direct experimental evidence for the existence of both mechanisms. Recently, Chong et al.<sup>19</sup> showed that bursts of RNA production can emerge due to positive supercoiling build-up on a DNA segment, which eventually stops transcription initiation for a short period until the release of the supercoiling by gyrase. On the other hand, the existence of rate-limiting steps was established by studies using steady-state assays.<sup>22–24</sup> Also, more recently, by fitting a monotone piecewise-constant function to the fluorescence signal from MS2-GFP tagged RNAs in individual cells, it was shown that *in vivo* RNA production can be a sub-Poissonian process.<sup>4,5,20,21</sup>

Recent studies have considered the possibility that both mechanisms can be present in a single promoter.<sup>16,25</sup> In ref. 25, a model including both mechanisms was proposed, and statistical methods were developed to select the relevant components and estimate the kinetics of the intermediate steps in initiation based on empirical data. However, this method cannot distinguish the order of the steps which occur after the start of transcription initiation, nor can it determine their reversibility, which recent evidence suggests may play a significant role in the dynamics of RNA production.<sup>26</sup>

A complete model for transcription in prokaryotes must account, apart from the genome-wide variability in noise levels,<sup>17,27,28</sup> for the well-established genome-wide variability in mean transcription rate<sup>2,3,8</sup> and in fold change (ratio of production rate between zero and full induction)<sup>29</sup> in response to induction found, e.g. in *Escherichia coli* promoters. For example, *in vitro* measurements on fully induced variants of the *lar* promoter showed that the mean interval between transcription events of these variants differs by hundreds of seconds.<sup>29</sup> Promoters also differ widely in range of induction, even when differing only by a couple of nucleotides.<sup>29,30</sup> For example, while  $P_{larS17}$  has an induction range of 500 fold,  $P_{larconS17}$  has an induction range of 4.5-fold, even though it only differs by 3 point mutations.<sup>29</sup> This wide behavioural diversity is likely made possible by the sequence dependence of each step in transcription initiation.<sup>29</sup>

Thus far, the strategies used *in vitro* to characterize the kinetics of the steps involved in transcription initiation<sup>22,26</sup> have not been applied *in vivo* since they rely on measuring transcription for different RNA polymerase (RNAP) concentrations. Such a change in cells is expected to have a multitude of unforeseen effects<sup>31</sup> (in addition to the side effects of the means used to alter RNAP concentrations), which hampers the assessment of its consequences to the duration of the closed complex formation of a specific promoter. However, it is reasonable to hypothesize that, for certain small ranges of RNAP concentrations, these side effects will be negligible and thus, in such ranges, the inverse of the rate of transcription will be linear with respect to the inverse of the free RNAP concentration.

Importantly, in *E. coli*, RNAP concentrations have been shown to vary widely with differing growth conditions.<sup>32</sup> As such, here we make use of different media richness to achieve different RNAP concentrations and test whether within this range of conditions, the RNA production rate changes hyperbolically with the RNAP concentrations (i.e. if the inverse of this rate changes linearly with the inverse of the RNAP concentration). Having established this relationship, we make use of it to study the *in vivo* kinetics of transcription initiation of  $P_{lacIara-1}$ . In particular, we perform measurements of the time intervals between RNA productions at the single molecule level in different intracellular RNAP and inducer concentration conditions, which we use to derive a more detailed model of transcription initiation of  $P_{lacIara-1}$ . For this, we first extrapolate the mean interval between production events to the limit of infinite RNAP concentration, so as to estimate the *in vivo* durations of the open and closed complex formations of this promoter. Next, we examine the significance of

an intermittent inactive promoter state, and the role of LacI in the emergence of this state. Finally, for the first time *in vivo*, we determine the reversibility of the closed complex formation.

## 2. Materials and methods

### 2.1. Cells and plasmids

For single-cell RNAP fluorescence measurements, we used *E. coli* W3110 and RL1314,<sup>33</sup> generously provided by Robert Landick, University of Wisconsin-Madison. For single-cell transcription interval measurements, we used *E. coli* DH5 $\alpha$ -PRO (generously provided by Ido Golding, Baylor College of Medicine, Houston). The strain information is: *deoR*, *endA1*, *gyrA96*, *hsdR17(rK- mK+)*, *recA1*, *relA1*, *supE44*, *thi-1*,  $\Delta$ (*lacZYA-argF*)U169,  $\Phi$ 80 $\delta$ lacZ $\Delta$ M15, F-,  $\lambda$ -, PN25/*tetR*, *PlacIq/lacI* and *SpR*. This strain contains two constructs: a high-copy reporter plasmid vector PROTET-K133 (carrying MS2d-GFP under the control of  $P_{LtetO-1}$ ) and a single-copy plasmid vector pG-BAC carrying the target transcript (mRFP1 followed by 96 MS2-binding sites) under the control of  $P_{lacIara-1}$ .<sup>2</sup> This promoter is located approximately 2 and 9 kb from the origin of replication (Ori2) and the plasmid size is 11.5 kb.<sup>2</sup> This system has been used to measure the distribution of time intervals between RNA production events due to its ability to detect individual target RNA molecules consisting of numerous MS2 coat protein binding sites, which are rapidly bound by fluorescently tagged MS2 coat proteins. These can be seen as they are produced under a fluorescence microscope as fluorescent foci.<sup>2,4,5,20,21</sup> Finally, we used the plasmid pAB332 carrying *hupA-mCherry* to visualize nucleoids (generously provided by Nancy Kleckner, Harvard University, Cambridge, MA, USA). For our measurements, we inserted this plasmid into DH5 $\alpha$ -PRO cells so as to detect nucleoids in individual cells during the live cell microscopy sessions. HupA is a major nucleoid associated protein (NAP) that participates in its structural organization.<sup>34</sup>

### 2.2. Chemicals

The components of Lysogeny Broth (LB) were purchased from LabM (UK), and antibiotics from Sigma-Aldrich (USA). For RT-PCR, cells were fixed with RNAProtect bacteria reagent (Qiagen, USA). Tris and EDTA for lysis buffer were purchased from Sigma-Aldrich and lysozyme from Fermentas (USA). The total RNA extraction was done with RNeasy RNA purification kit (Qiagen). DNase I, RNase-free for RNA purification, was purchased from Promega (USA). iScript Reverse Transcription Supermix for cDNA synthesis and iQ SYBR Green supermix for RT-PCR were purchased from Biorad (USA). Agarose, isopropyl  $\beta$ -D-1-thiogalactopyranoside (IPTG), arabinose, and anhydrotetracycline (aTc) are from Sigma-Aldrich.

### 2.3. Growth media

To achieve different RNAP concentrations in cells, we altered their growth conditions as in.<sup>35</sup> For this, we used modified LB media which differed in the concentrations of some of their components. The media used are denoted as *m<sub>x</sub>*, where the composition per 100 ml are: *m* grams of tryptone, *m/2* gram of yeast extract and 1 g of NaCl (pH = 7.0). For example, 0.25 $\times$  media has 0.25 g of tryptone and 0.125 g of yeast extract per 100 ml.

### 2.4. Relative RNAP quantification

We measured relative RNAP concentrations in cells using four different methods. First, relative RNAP concentrations in the strains W3110 and DH5 $\alpha$ -PRO were measured from the relative *rpoC* transcript



levels obtained using RT-PCR. Cells containing the target plasmid with *P<sub>lacIara-1</sub>-mRFP1-96BS* and the reporter plasmids were grown overnight in respective media. Cells were diluted into fresh media to an  $OD_{600}$  of 0.05. After 110 min, cells were re-diluted to an  $OD_{600}$  of 0.05 into respective media containing IPTG (1 mM) and arabinose (1%). After 70 min, RNA protect reagent was added to fix the cells, followed by enzymatic lysis with Tris-EDTA lysozyme buffer (pH 8.3). RNA was isolated from cells using RNeasy mini-kit (Qiagen). One microgram of RNA was used as the starting material. The RNA samples were treated with DNase free of RNase to remove residual DNA. Next, RNA was reverse transcribed into cDNA using iScript reverse transcription super mix (Biorad). RT-PCR was performed using Power SYBR-green master mix (Life Technologies) with primers for the amplification of the target gene at a concentration of 200 nM. Reactions were carried out in triplicate with 500 nM per primer with a total reaction volume 20  $\mu$ l. The following primers were used for quantification: RpoC-F: CGTCAGATGCTGCGTAAAGC, RpoC-R: GCGATCTTGACGCGAGAGTA, mRFP1-F: TACGACG CCGAGGTCAAG, mRFP1-R: TTGTGGGAGGTGATGTCCA. Estimated relative RNAP concentrations  $\bar{R}_m$  in each condition  $m$ , and their standard uncertainties  $\sigma(\bar{R}_m)$ , were calculated according to the  $\Delta C_T$  method.<sup>36</sup>

Second, *E. coli* RL1314 cells with fluorescently tagged  $\beta'$  subunits were grown overnight in respective media. A pre-culture was prepared by diluting cells to an  $OD_{600}$  of 0.1 with fresh specific medium, and grown to an  $OD_{600}$  of 0.5 at 37°C at 250 rpm. Cells were pelleted by centrifugation and re-suspended in saline. Fluorescence from the cell population was measured using a fluorescent plate-reader (Thermo Scientific Fluoroskan Ascent Microplate Fluorometer).

Third, relative RNAP concentrations were also estimated based on the growth rates of DH5 $\alpha$ -PRO cells in Supplementary Fig. S1. First, we fit a power law function to the 'RNA polymerase molecules per cell' row of Table 3 from ref. 32, which we found to be  $R = 10^6 \mu^{-1.426}$ , where  $\mu$  is the cell doubling time. Relative RNAP concentrations were then estimated from the measured cell doubling times.

Lastly, we measured the relative RNAP concentrations in RL1314 cells under the microscope using fluorescently tagged RpoC (described in the next section).

## 2.5. Microscopy

DH5 $\alpha$ -PRO cells containing the target and the reporter plasmids were grown as described previously. Briefly, cells were grown overnight in respective media, diluted into fresh media to an  $OD_{600}$  of 0.1, and allowed to grow to an  $OD_{600}$  of  $\sim$ 0.3. For the reporter plasmid induction, aTc (100 ng/ml) was added 1 h before the start of the measurements. For the target plasmid, arabinose (1%) was added at the same time as aTc (following the protocol in ref. 2), and IPTG (1 mM) was added 10 min before the start of the measurements. Cells were pelleted and resuspended to fresh medium. A few microliters of cells were placed between a coverslip and an agarose gel pad (2%), which contains the respective inducers, in a thermal imaging chamber (FCS2, Biotech), heated to 37°C. The cells were visualized using a Nikon Eclipse (Ti-E, Nikon, Japan) inverted microscope with a C2+ confocal laser-scanning system using a 100 $\times$  Apo TIRF objective. Images were acquired using the Nikon Nis-Elements software. GFP fluorescence was measured using a 488 nm argon ion laser (Melles-Griot) and 514/30 nm emission filter. Phase-contrast images were acquired with the external phase contrast system and a Nikon DS-Fi2 camera. Fluorescence images were acquired every 1 min for a total duration of 2 h. Phase-contrast images were acquired simultaneously every 5 min during the measurements.

We tested for phototoxicity due to the fluorescence and the phase-contrast imaging in these measurements. Supplementary Results suggest that there is no significant phototoxicity. Additionally, we verified that the relative RNAP concentrations under the microscope are similar to those measured in the previous section by repeating the above procedure with RL1314 cells and imaging RpoC::GFP fluorescence, 1 h after being placed in the thermal imaging chamber (see Supplementary Fig. S4). The relative RNAP concentration was estimated from the mean fluorescence concentrations of cells growing in each media.

## 2.6. Image analysis

Cells were detected from the phase contrast images as described in ref. 37. First, the images were temporally aligned using cross-correlation. Next, an automatic segmentation of the cells was performed by MAMLE,<sup>38</sup> which was checked and corrected manually. Next, cell lineages were constructed by CellAging.<sup>39</sup> Alignment of the phase-contrast images with the confocal images was done by manually selecting 5–7 landmarks in both images, and using thin-plate spline interpolation for the registration transform. Fluorescent spots and their intensities were detected from the confocal images using the Gaussian surface-fitting algorithm from.<sup>40</sup>

Jumps were detected in each cell's spot intensity timeseries using a least-deviation jump-detection method.<sup>41</sup> Given the level of noise in the timeseries, jump sizes, i.e. the intensity of 'one RNA', were selected by manual inspection of the timeseries of total foreground spot intensities within cells of a given timeseries, and cross-referencing these values with the observed numbers of spots in the cells. After performing the jump detection process making use of the complete timeseries, jumps occurring within 5 min of the beginning or end of a cell's lifetime were disregarded due to our observation that the jump detection method tends to produce spurious jumps in these regions due to insufficient data. The remaining jumps were interpreted as RNA production times, from which intervals between transcription events were calculated. Finally, censored intervals were calculated as the time from the last RNA production in a cell until the last time at which a jump could have been observed (i.e. until 5 min prior to cell division or the end of the timeseries). This removes the possibility of false positives while not affecting the distribution of intervals.

This method, when first proposed, made two assumptions on the fluorescence of MS2-GFP tagged RNAs (named 'spots'). Importantly, both assumptions were recently shown to be valid.<sup>42</sup> First, an individual spot is bound sufficiently rapidly by MS2-GFPs such that its fluorescence intensity, when first detected, is already within the range of fluorescence of fully formed MS2-GFP-RNA spots (when taking one image per minute). In other words, the spot intensity of a newly transcribed RNA jumps from 0 to 'full' in <1 min, rather than slowly ramping up. Namely, since the transcription elongation rate of mRNA in *E. coli* is  $\sim$ 50 nt/s<sup>32</sup> and the target gene is  $\sim$ 3,200 bp long,<sup>1</sup> the time to elongate the MS2-binding site region of the target RNA is  $\sim$ 60 s. Provided that MS2-GFP binding to its RNA-binding sites is fast, there will therefore be a maximum of one timepoint at which the fully transcribed target RNA may have reduced fluorescence. Since MS2-GFP is produced in excess in the cell and its binding affinity is strong (dissociation constant of  $\sim$ 0.04 nM<sup>43</sup>), most binding sites will be saturated very shortly after being produced. In agreement with ref. 42, no gradual increase in spot fluorescence was observed around the time of the first appearance of a spot.

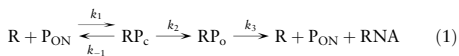
Second, once formed, MS2-GFP-RNA spots, as well as their fluorescence, are resistant to degradation for the duration of our

measurements (2 h). This was shown by measurements of the dissociation rate of MS2 coat proteins from their RNA binding sites (on the order of several hours<sup>43</sup>), and by measurements of the lifetimes of the fluorescence of MS2-GFP tagged RNAs kept under observation for more than 2 h.<sup>1,2,5,42,44</sup> Relevantly, no detectable decrease in fluorescence was observed during this time.<sup>42</sup>

## 2.7. Model of transcription initiation

We first consider a model that allows for RNA production dynamics to range from sub-Poissonian to super-Poissonian, given the results from genome-wide studies of the variability in RNA numbers<sup>27,45</sup> and from studies of the transcription dynamics of individual genes.<sup>2,4,5,17,20</sup> The features of the model that allow it to reproduce these numbers are based on processes known to occur during transcription initiation in *E. coli* (e.g. the open complex formation<sup>16,22,23</sup> and an ON/OFF mechanism<sup>16,19</sup>). Then, based on our novel empirical data and methodology, we aim to obtain the most parsimonious version of the model that fits the data for a given promoter. We expect this procedure to be applicable to any promoter, and to result in slightly different models due to their differing dynamics and regulatory mechanisms.

The full model of transcription initiation considered here consists of the following set of reactions:



Reaction (1) represents the multi-step process of transcription initiation of an active promoter in prokaryotes.<sup>23,24,46,47</sup> It begins with the formation of the closed complex ( $RP_c$ ), i.e. the binding of the RNA polymerase (R) to a free promoter ( $P_{ON}$ ). Once at the start site, the polymerase must open the DNA double helix, a process that includes several long-lived intermediate states,<sup>23,26,46,48</sup> resulting in the open complex ( $RP_o$ ). Finally, the polymerase begins RNA elongation, though before clearing the promoter, it may engage in abortive RNA synthesis in which short RNA transcripts (<10 nt) are produced.<sup>47,49</sup> The reactions in (1) should not be interpreted as elementary transitions. Rather, they represent the effective rates of the rate-limiting steps in the process, thus defining the promoter strength, and have been shown to be sequence-dependent.<sup>50</sup>

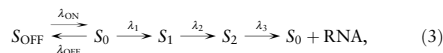
Specifically,  $k_1$  represents the rate at which polymerases find and bind to the promoter region, which is the overall result of the promoter search process which includes non-specific binding of the polymerases to the DNA, followed by a 1D diffusive search,<sup>51,52</sup> collectively referred to here as the closed complex formation. Subsequently, several rapid, possibly reversible isomerization reactions occur until the polymerase melts the DNA and forms the transcription ‘bubble’.<sup>51</sup> In Reaction (1), the  $RP_c$  state represents all substates until the first irreversible reaction in this chain. Consequently,  $k_2$  and  $k_{-1}$  should be interpreted as the product of the rates of the elementary reactions which exit from this group of substates, and the steady-state probability of being in the appropriate substates for these reactions to occur.

Similarly, the  $RP_o$  state may represent numerous substates between the first state after which the complex is committed to initiation, and successful initiation. However, after this point, we cannot distinguish the reversibility of any of the following steps, since the time-interval distribution of a sequence of elementary reversible reactions of arbitrary rates is observationally equivalent to a sequence of irreversible reactions.<sup>25</sup> The remaining steps (here, only  $k_3$ ) therefore represent

the rates of the slowest of these irreversible reactions. Such steps may include additional isomerization reactions, abortive RNA synthesis and promoter escape and clearance.<sup>35</sup>

Reaction (2) represents the promoter intermittently transitioning to a transcriptionally inactive state ( $P_{OFF}$ ). Experimentally verified mechanisms by which this can occur are the binding and unbinding of repressors and activators,<sup>29</sup> the accumulation of positive supercoiling in the DNA.<sup>19</sup> Additional mechanisms have also been hypothesized, such as transcriptional pausing<sup>53,54</sup> and others.<sup>55</sup>

For a given concentration of R, the interval distribution between transcription events described by Reactions (1) and (2) (i.e. the first-passage time distribution to reach the final state, starting in the  $P_{ON}$  state) is observationally equivalent to the interval distribution described by a model of the form:



where the system starts in state  $S_0$ . The relationship between the parameters of these two models is described in Supplementary Table S1. Note that the states  $S_i$  do not correspond to the promoter states in Reactions (1) and (2). For details on how to derive and evaluate the distribution function for this model, see Supplementary Material and.<sup>25</sup>

It is noted that this model assumes that only one copy of the promoter is present in each cell at any given time. In the experiments performed here, in all conditions tested, the bacteria divided sufficiently slowly such that they spent most lifetime with only one chromosome. Specifically, cells spent no more than  $11.4 \pm 1.0\%$  of their lifetime with two copies of the target promoter (Supplementary Material).

Finally, it is noted that the present model does not consider the influence of  $\sigma$  factors’ numbers on the dynamics of transcription initiation, focussing instead solely on the concentration of RNA polymerases (in particular, on the concentration of holoenzymes containing a  $\sigma^{70}$ , i.e.  $E\sigma^{70}$ , since our promoter of interest can only be transcribed by  $E\sigma^{70}$ ). This is based on the fact that, in all conditions tested, most RNA polymerases are occupied by  $\sigma$  factors.<sup>56,57</sup> Further, this occupation is made largely by  $\sigma^{70}$  since, first, when altering media richness, only  $\sigma^{32}$ ’s concentration is significantly altered<sup>56</sup> and, second, the binding affinity of  $\sigma^{70}$  to E is much higher than that of any other  $\sigma$  factor (e.g. it is approximately 9 times higher than that of  $\sigma^{32}$ ).<sup>57</sup>

## 2.8. Parameter estimation

Parameter estimates in Tables 1–3 were obtained by a maximum likelihood fit using the samples of the distribution of time intervals between production events obtained above (the intervals and censored intervals), as in.<sup>25</sup> The complete model-fitting procedure is detailed in the Supplementary Material. The uncertainty of the fit of the model parameters was estimated using the negative of the Hessian of the log-likelihood surface, evaluated at the maximum likelihood estimate.

The mean of the time interval distribution between transcription initiation events,  $I(R)$ , predicted by Reactions (1) and (2) is, for a given RNAP concentration R:

$$I(R) = \frac{(k_{ON} + k_{OFF})(k_{-1} + k_2)}{Rk_1k_2k_{ON}} + \frac{1}{k_2} + \frac{1}{k_3} = \tau_{CC}(R) + \tau_{CC} \quad (4)$$

where  $\tau_{CC}(R) = k_{CC}^{-1}R^{-1}$  is the mean time taken by the initial binding of RNAP for a given RNAP concentration, and  $\tau_{CC}$  is the mean time taken by the steps occurring after the polymerase has committed to transcription until the clearance of the promoter region (due to the

initiation of elongation). As such, we expect the majority of the duration of  $\tau_{CC}$  to consist of the open complex formation as defined in.<sup>46</sup> The remaining of its duration we attribute to failures in promoter escape.<sup>59</sup>

Estimates of  $\tau_{CC}$  and  $k_{CC}^{-1}$ , denoted  $\hat{\tau}_{CC}$  and  $\hat{k}_{CC}^{-1}$ , were obtained from the best-fit parameters of the most parsimonious model, as given in Table 3. The standard uncertainties of the estimators  $\hat{\tau}_{CC}$  and  $\hat{k}_{CC}^{-1}$ , denoted  $\sigma(\hat{\tau}_{CC})$  and  $\sigma(\hat{k}_{CC}^{-1})$ , were obtained using the Delta Method<sup>60</sup> from the uncertainties of the model parameters.

Finally, mean durations of intervals between transcription events for each media condition  $\hat{I}_m$ , were estimated by fitting the model in Reaction (3) to the data from only that condition, and taking the mean of the distribution. This procedure was followed to include the censored intervals in the estimate of  $\hat{I}_m$  to avoid underestimating the mean interval duration due to the limited observation times. The standard uncertainty  $\sigma(\hat{I}_m)$  was estimated using the Delta Method.<sup>60</sup>

## 2.9. Validation of the $\tau$ -plot slope

We verified the slope of the  $\tau$ -plot in Fig. 4 using the RT-PCR measurements from Fig. 3. These measurements are both linear with respect to  $\hat{R}_m^{-1}$ , but differ by an unknown scaling factor. We denote the estimated production rate as measured by RT-PCR in media condition  $m$  as  $\hat{S}_m$ , with standard uncertainty  $\sigma(\hat{S}_m)$ . We found this scaling factor by fitting the parameter  $c$  in  $\hat{I}_m = c\hat{S}_m^{-1}$  by weighted total least squares<sup>61</sup> (WTLS), with the measurements weighted by the inverse of their uncertainty (i.e.  $\sigma^{-2}(\hat{S}_m^{-1})$  and  $\sigma^{-2}(\hat{I}_m)$ ). This method was chosen since it accounts for the uncertainty in both of the measurements. It results in the estimate  $\hat{c}$ . The dashed line in Fig. 4 was obtained by fitting the scaled points  $\hat{c}\hat{S}_m^{-1}$  against  $\hat{R}_m^{-1}$  by WTLS. The uncertainty shown includes both the uncertainty in the WTLS fit of this line, as well as the uncertainty in  $\hat{c}$ .

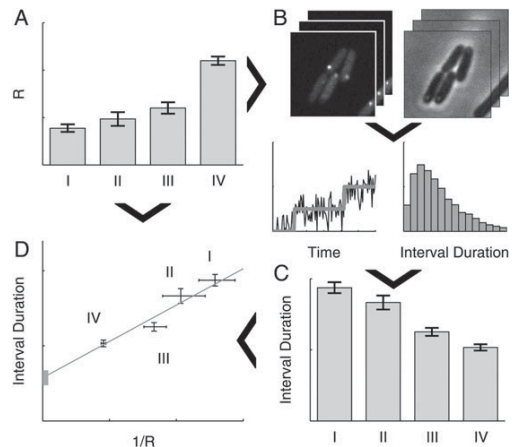
## 2.10. Method to infer the duration of the closed complex of a promoter

The method to infer the kinetics of transcription initiation *in vivo* is illustrated in Fig. 1. First, conditions are selected that differ widely in free intracellular RNAP concentrations (step A in Fig. 1). Next, an *in vivo* single-molecule detection technique is used to sample the time interval distribution between consecutive transcription events in individual cells in each of the conditions (step C in Fig. 1). To obtain these intervals, here we used the MS2d-GFP single RNA detection system<sup>4</sup> (step B in Fig. 1). Then, we fit a general model of transcription initiation to the empirical data (see above), which includes both the multi-step nature of transcription initiation as well as the possibility of an intermittently inactive promoter state<sup>2,5</sup> (Reactions (1) and (2)). From this fit, we obtain an estimate of the *in vivo* mean duration of the open complex formation by extrapolating the duration of intervals between transcription events to infinite RNAP concentrations, similar to the *in vitro* extrapolation presented in ref. 22 (step D in Fig. 1). The model fit will also assess the importance of an intermittent inactive promoter state and the reversibility and kinetics of the closed complex formation.

## 3. Results

### 3.1. Changing free RNA polymerase concentrations

We first verified that it is possible to change intracellular RNAP concentration by a wide range by changing the growth conditions of the cells.<sup>32,35,62</sup> As such, we grew cells in four media (described in the Materials and methods), labelled 1 $\times$ , 0.75 $\times$ , 0.5 $\times$ , and 0.25 $\times$ , which solely



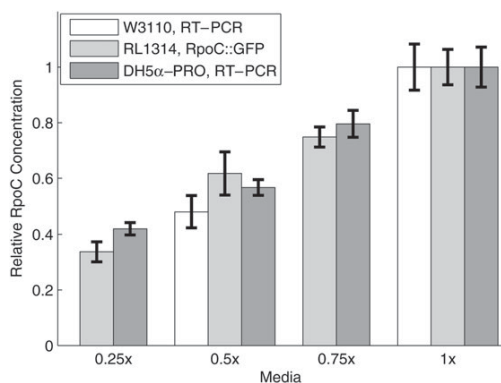
**Figure 1.** Schematic representation of the *in vivo* measurement of the initiation kinetics, using simulated data. (A) First, several conditions are selected, labelled I–IV, differing in intracellular RNAP concentration,  $R$ . (B) Next, we obtain timeseries of fluorescence and phase contrast (for cell segmentation purposes) images of cells expressing MS2d-GFP and target RNA under the control of the promoter of interest in each condition, from which time intervals between individual transcription events are determined. This is done by jump detection in the total RNA spot intensity of each cell (lower-left in B), from which the interval distribution is obtained (lower-right in B). (C) Mean interval durations are then estimated from these interval distributions for each condition. (D) Finally, the mean interval durations and measurements of  $R$  are combined into a  $\tau$ -plot,<sup>22</sup> from which estimates of the mean times taken by the closed complex and open complex formation are obtained for each condition. Arrows depict the flow of information in the measurement procedure.

differ in richness of two components (tryptone and yeast extract). We then measured the relative RNAP concentrations in cells grown in these four media using RT-PCR of the *rpoC* gene, i.e. the gene coding for the  $\beta'$  subunit, which is the limiting factor in the assembly of the RNAP holoenzyme.<sup>48,57,62</sup> Results in Fig. 2 (dark grey bars) show that, in the range tested, the RNAP concentration in the cells increases significantly with increasing media richness.

To validate this result, we measured the relative RNAP concentrations by plate reader in cells expressing fluorescently tagged RpoC in the strain RL1314 (derived from W3110),<sup>33</sup> in the same four media. In addition, we also measured the levels of the *rpoC* transcripts in the strain W3110 by RT-PCR in the 0.5 $\times$  and 1 $\times$  conditions. Results (Fig. 2) show that the relative changes in the protein and mRNA levels of *rpoC* match the measurements by RT-PCR of the *rpoC* gene in DH5 $\alpha$ -PRO.

Note that, even though the experimental procedures and strains differ, our measurements are in agreement with the relative changes in RNAP concentrations reported in ref. 32, for the difference in growth rates observed here between the 0.25 $\times$  and 1 $\times$  conditions (Supplementary Fig. S1), which we estimate to be  $\sim 0.48$  (Materials and methods). In this regard, given that the same result applies to (at least) three different strains, we expect it to be significantly strain-independent.

Finally, to verify that the relative RNAP concentrations measured in Fig. 2 are maintained under the microscope, we measured the relative RNAP concentration in the RL1314 cells expressing fluorescently



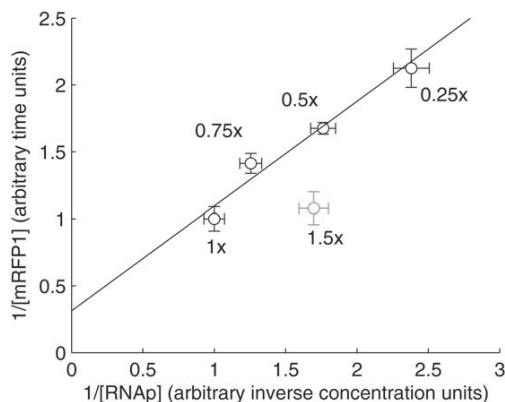
**Figure 2.** Measurements of the relative intracellular RNAP concentrations ( $\hat{R}_m$ ) for cells growing in the four different media. Bars show the standard uncertainties ( $\sigma(\hat{R}_m)$ ) of the measurements. Data is from two replicates with 3 technical replicates each (DH5 $\alpha$ -PRO, RT-PCR, and W3110, RT-PCR), and three replicates with three technical replicates each (RL1314, RpoC::GFP). All data are presented relative to the RNAP concentration at 1x. The media used are denoted as  $m$ x, where the composition per 100 ml is:  $m$  grams of tryptone,  $m/2$  grams of yeast extract and 1 g of NaCl (pH = 7.0). For example, 0.25x media has 0.25 g of tryptone and 0.125 g of yeast extract per 100 ml.

tagged RpoC under the microscope between the two extreme conditions (0.25x and 1x), after 1 h in the thermal imaging chamber (Materials and methods). The relative RNAP concentration between the conditions was measured to be  $0.367 \pm 0.012$ , which is consistent with the measurements in Fig. 2. Lastly, from these images, we did not observe significant cell-to-cell variability in the RNAP concentrations (Supplementary Fig. S4), indicating that the mean concentrations reported in Fig. 2 are representative of the populations.

These measurements show that the relative RNAP concentration changes widely between the selected growth conditions. However, the variable affecting transcription kinetics is the relative free RNAP concentration. As such, we must verify whether the relative total RNAP concentration can be used as a proxy for the relative free RNAP concentrations. If this holds true and there are no other factors affecting the production rate of the promoter of interest in these conditions, then the RNA production rate should be hyperbolic with respect to the RNAP concentration. That is, the reciprocal of the RNA production rate from this promoter should be linear when plotted against the reciprocal of the measured relative RNAP concentrations, and one should obtain a line on a Lineweaver–Burk plot.

There are several reasons why this plot may not be linear. If, for example, the ratio of free RNAP to total RNAP is not constant in this range of growth conditions, with a higher fraction of free RNAP in the poorer growth conditions due to increased ppGpp,<sup>31</sup> then we expect a curve with positive curvature on this plot. Meanwhile, a negative curvature would be obtained if the promoter of interest could be induced by increased cAMP in the poorer growth conditions, or if the cells spent, on average, a significantly increased amount of time with multiple copies of the plasmid in the richer growth conditions, among other possibilities. In these cases, to dissect the transcription initiation kinetics of such promoters, another method of modifying the free RNAP concentration will be required.

Given the above, we interpret a straight line on the Lineweaver–Burk plot as evidence that, for the conditions tested, (i) the relative free RNAP concentrations can be assessed from the total RNAP



**Figure 3.** Lineweaver–Burk plot of the inverse of the production rate of mRFP1 from the  $P_{lacIara-1}$  promoter against the inverse of the total RNAP concentrations for the same growth conditions as in Fig. 2 (black points), and for 1.50x media (grey point). Standard uncertainties are shown for both quantities (horizontal and vertical error bars). Relative production rates were measured by RT-PCR with two biological replicates with three technical replicates each.

concentrations, and (ii) no factors other than the changes in the free RNAP concentration affect the target promoter.

Here, we tested this by measuring the RNA production rate from  $P_{lacIara-1}$  in *E. coli* DH5 $\alpha$ -PRO by RT-PCR in the same four media conditions as in Fig. 2. We selected this promoter, since its dynamics has been extensively characterized<sup>2,21,29,63–67</sup> and because it has the same logical structure as the *lac* promoter, with an activator and a repressor.<sup>63</sup> The resulting Lineweaver–Burk plot is shown in Fig. 3 where a linear relationship is clearly observed between these points (black points). To determine whether the small deviations from linearity are statistically significant, we performed a likelihood ratio test between a linear fit by WTLS<sup>64</sup> (shown as a line in Fig. 3), and fits with higher order polynomials (also by WTLS by minimizing  $\chi^2$  as in<sup>64</sup>). No test rejected the linear model (all  $P > 0.25$ ). As noted earlier, this relationship is only expected to occur in a limited range of growth conditions. To illustrate this, we repeated the same measurements in 1.5x media (grey point in Fig. 3). The result shows that this hyperbolic relationship is lost in very rich media (including this point causes the likelihood ratio test to reject the linear model,  $P = 0.0014$ ). We conclude that, for the growth conditions in Fig. 2, the relative free RNAP concentrations are well-approximated by the total RNAP concentrations, and there are no significant other factors affecting the initiation dynamics of  $P_{lacIara-1}$ .

### 3.2. Interval distributions between consecutive RNA productions

Given this, it is possible to apply a standard model-fitting procedure to fully characterize the *in vivo* kinetics of the rate-limiting steps in transcription initiation of the  $P_{lacIara-1}$  promoter from distributions of intervals between transcription events in cells with different RNA polymerase concentrations.

We measured the distribution of time intervals between transcription events (hereafter referred to as ‘intervals’) for  $P_{lacIara-1}$  in each cell growth condition using the MS2d-GFP single-RNA detection system,<sup>1</sup> with a least-deviation jump-detection procedure<sup>41</sup> (Materials and

**Table 1.** Statistics of the measured distributions of intervals between transcription events from *lac/ara-1* promoters

Condition	Number of cells	Number of intervals	Number of censored intervals	Inferred interval mean and uncertainty (s)	Inferred CV <sup>2</sup>
0.25x	196	371	323	1,899 ± 105	1.08
0.5x	302	1,027	605	1,553 ± 50	1.06
0.75x	146	620	345	1,205 ± 51	1.09
1x	206	1,202	573	1,005 ± 112	1.21

Shown are the condition, the number of cells (which is the cell count at the start of the measurements), the numbers of whole and censored intervals extracted, and finally the inferred mean (and its standard uncertainty) and CV<sup>2</sup> of the interval distribution.

methods). This measurement results in samples from the interval distribution as well as ‘censored’ intervals, i.e. intervals for which we only observe the beginning due to cell division or the end of the time series. Both censored and uncensored intervals were accounted for in all parameter estimates to avoid biasing the estimates. For example, note that taking the mean of the uncensored intervals alone would underestimate the mean of the true interval distribution since long unobservable intervals would be absent from the estimate. Including the censored intervals balances this by considering long intervals that are at least as long as the censored interval length.<sup>25</sup>

From these distributions, we estimated the true mean and the squared coefficient of variation (CV<sup>2</sup>, defined as the variance over the squared mean) of the interval distributions (Materials and methods). We chose CV<sup>2</sup> for quantifying the noise in the interval distribution since, to a good approximation, this quantity reflects the level of noise in the protein levels regardless of the actual shape of the transcription interval distribution.<sup>68</sup> Further, this variable equals 1 for the interval distribution of a Poisson process (i.e. an exponential distribution), regardless of the mean rate. These results, along with the amount of empirical data used, are shown in Table 1.

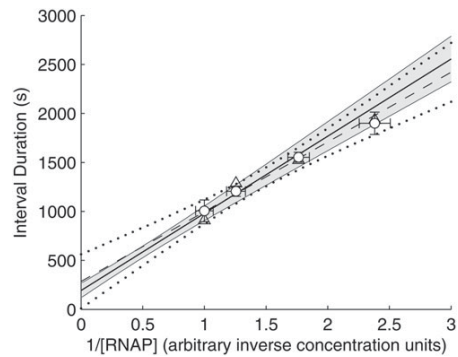
From Table 1, the mean interval decreases significantly with increasing media richness, as expected from the increased RNAP concentrations. Meanwhile, the CV<sup>2</sup> does not exhibit the same dependence on the media richness, and remains slightly >1 in all conditions tested.

### 3.3. Decomposition of the *in vivo* kinetics

From the data in Table 1, we next recreate the Lineweaver–Burk plot in Fig. 3 (white circles in Fig. 4), using the mean interval durations between RNA productions, as this quantity is an absolute measure of the inverse rate of RNA production (this plot is called a  $\tau$ -plot).

Previously, using *in vitro* techniques, it has only been possible to extract from a  $\tau$ -plot the mean duration of the open complex formation (the  $y$ -intercept of the plot, here denoted  $\tau_{\text{OC}}$ ), because the plot is based on the steady-state assay which only measures the mean rate of abortive transcription initiations. However, the distributions of time intervals between RNA productions contain information about the stochasticity of the process (i.e. the variability between intervals). As such, it is possible to extract a more complete model of the process of transcription. Namely, aside from the open complex formation, as mentioned in Materials and methods, it is possible to extract information on the closed complex and on an intermittent state prior to the closed complex formation.

In particular, we consider the detailed model of transcription initiation presented in Materials and methods (Reactions (1) and (2)),



**Figure 4.**  $\tau$ -plot for  $P_{lac/ara-1}$ , showing the mean interval between transcription events in individual cells for each media condition (white circles), with their standard uncertainties (vertical error bars) and the standard uncertainties of the relative RNAP concentrations (horizontal error bars). Also shown is the best-fit line (solid line), as determined by the intercept and slope obtained from the best-fitting model (Table 3), with one standard uncertainty estimated by Scheffé's method<sup>69</sup> combined with the Delta Method<sup>60</sup> (grey area). In addition, the figure shows the data from Fig. 3 (triangles), and the best-fitting line (dashed line, see Materials and methods) with one standard uncertainty estimated by Scheffé's method<sup>69</sup> (dotted black curves).

along with simplified models that can be considered if certain steps of the more detailed model do not influence the distribution of intervals. This model assumes that only one copy of the promoter is present in each cell at any given time, since in all conditions, the bacteria divided slowly, which suggests that they spent most lifetime with only one chromosome. We then consider three simplified models. First, if the time spent in the OFF state is very small, or if the system switches between OFF and ON very rapidly when compared with the forward reaction, then Reaction (2) will not affect the RNA production dynamics. A sufficient condition for both of these situations is that  $k_{\text{ON}} \gg k_1$ . The other two simplifications are two limits of the closed complex formation, first considered in<sup>22</sup>: (i)  $k_{-1} \gg k_2$ , i.e. it is reversible (Limiting Mechanism I), and (ii)  $k_2 \gg k_{-1}$ , i.e. irreversible (Limiting Mechanism II). Limiting Mechanism I was found to be more likely in several *in vitro* measurements of various promoters.<sup>22,23,26</sup>

While all three simplifications are consistent with a line on a  $\tau$ -plot, they produce significantly different distributions of intervals between RNA production events. For example, a significant ON/OFF mechanism will result in a more noisy distribution (a higher CV<sup>2</sup>).<sup>25</sup> Similarly, Limiting Mechanism I effectively eliminates one limiting step, which also results in higher noise when compared with Limiting Mechanism II (Supplementary Fig. S2).

We fit the full and simplified models of transcription initiation to the observed dynamics of  $P_{lac/ara-1}$  from all media conditions (Materials and methods). We used the Bayesian Information Criterion<sup>70</sup> (BIC) to compare the fits. The BIC is a model selection criterion which balances goodness-of-fit with the number of parameters to determine which model is most likely the ‘truth’. The difference between BIC values ( $\Delta\text{BIC}$ ) can be interpreted as evidence *against* the model with *higher* BIC, with a  $\Delta\text{BIC} > 5$  being interpreted as strong evidence.<sup>58</sup> Results are shown in Table 2. Since, for several of the models, the optimal fit was for  $k_3^{-1} = 0$ , we also considered models that do not include another rate-limiting step after the open complex formation.

From Table 2, the initiation kinetics of  $P_{lac/ara-1}$  is best-fit by Limiting Mechanism I (i.e. a reversible closed complex), with very high

**Table 2.** Fit parameters of the transcription initiation model in Reactions (1) and (2), and the models derived by applying the listed simplifying assumptions

Limiting mechanisms	Simplifications	$k_{ON}^{-1}$ (s)	$k_{OFF}^{-1}$ (s)	$k_1 k_{OFF}^{-1}$ (R <sup>-1</sup> )	$k_1^{-1}$ (Rs)	$k_1 k_2^{-1}$ (R <sup>-1</sup> )	$k_2^{-1}$ (s)	$k_1 k_2^{-1}$ (R <sup>-1</sup> )	$k_2^{-1}$ (s)	$k_2^{-1}$ (s)	$k_2^{-1}$ (s)	$\Delta BIC$	$\Delta BIC_C$
Full model		87	Fast	8,313	Fast	2,247	177	2,247	177	Fast	Fast	14.8	15.7
I	$k_{-1} \gg k_2, k_1 \gg k_{OFF}$	87	Fast <sup>a</sup>		Fast	7,446	192	7,446	192	Fast	Fast	8.1	8.5
I, $k_3 = \infty$	$k_{-1} \gg k_2, k_1 \gg k_{OFF}, k_3 = \infty$	87	Fast <sup>a</sup>		Fast	6,469	192	6,469	192	Fast	Fast	0.0	0.0
II	$k_2 \gg k_{-1}$	90	Fast	0.10	Fast		7		7			18.3	18.8
II, $k_3 = \infty$	$k_2 \gg k_{-1}, k_3 = \infty$	86	Fast	0.09	Fast		10		10			10.7	10.7
No ON/OFF	$k_{ON} \gg k_1$		Fast		Fast	0.49	326	0.49	326	Fast	Fast	188.1	188.1
No ON/OFF, I	$k_{ON} \gg k_1, k_{-1} \gg k_2$		Fast		Fast	0.50	328	0.50	328	Fast	Fast	180.1	179.6
No ON/OFF, I, $k_3 = \infty$	$k_{ON} \gg k_1, k_{-1} \gg k_2, k_3 = \infty$		Fast		Fast	0.50	328	0.50	328	Fast	Fast	172.0	171.1
No ON/OFF, II	$k_{ON} \gg k_1, k_2 \gg k_{-1}$				910		Fast		Fast	Fast	Fast	201.0	200.6
No ON/OFF, II, $k_3 = \infty$	$k_{ON} \gg k_1, k_2 \gg k_{-1}, k_3 = \infty$				910		Fast		Fast	Fast	Fast	192.9	192.0

Parameters denoted 'fast' are too fast to present on the timescale of seconds. When competing fast reactions occur, relevant ratios are given.  $\Delta BIC$  values are given as the difference of the model's BIC from the BIC of the best-fitting model (the one with  $\Delta BIC = 0$ ). Models with lower  $\Delta BIC$  are favoured over models with higher  $\Delta BIC$ .<sup>58</sup> Censored intervals were included in  $\Delta BIC_C$ , but not in  $\Delta BIC$ . The best-fitting model is shaded. Rates (and ratios) involving  $k_1^{-1}$  are given relative to the intracellular RNAP concentration in the 1x media.

<sup>a</sup> $k_1 k_2^{-1} k_{OFF}^{-1} = 0.11$ .

certainty ( $\Delta BIC$  of all other models  $> 8$ ). We also find evidence for a significant ON/OFF mechanism. Though the time spent in each OFF state is short ( $\sim 87$  s), it will turn OFF, on average,  $\sim 9.1$  times before committing to transcription in the 1x case (see Supplementary Material). This results in an interval distribution which is only slightly more noisy than what would be expected if the production process were Poissonian (i.e. a  $CV^2$  of the interval distribution of 1; see the  $CV^2$  values in Table 1). Interestingly, this implies that the noise in transcription of this promoter is representative of the behaviour of the majority of promoters in *E. coli*.<sup>27</sup> Finally, the steps after the commitment to transcription are fast, indicating that abortive initiation events do not play a significant role in the dynamics of RNA production by  $P_{lacIara-1}$ . This model is depicted graphically in Fig. 5.

In addition, from Table 2, we find that  $\tau_{CC}$  is  $193 \pm 49$  s. Meanwhile, the slope of the line on the  $\tau$ -plot, here denoted  $k_{CC}^{-1}$  is  $788 \pm 59$  R-s ( $R$  is the polymerase concentration such that  $R = 1$  is the polymerase concentration in 1x media). The line given by these values is shown in Fig. 4 (solid line). As a side note, the uncertainties of these estimates exaggerate the uncertainty of the inference, since the estimates are highly correlated (correlation coefficient of  $-0.6$ ). This correlation is responsible for the hyperbolic shape of the confidence bounds (grey region in Fig. 4).

We verified the slope of the solid line in Fig. 4 using the RT-PCR measurements presented in Fig. 3, scaled to match the timescale of the intervals (Materials and methods). The resulting line is shown in Fig. 4 (dashed line), and is in good agreement with both the line given by our estimates of  $\tau_{CC}$  and  $k_{CC}^{-1}$  (solid line), and the inferred interval means (white circles).

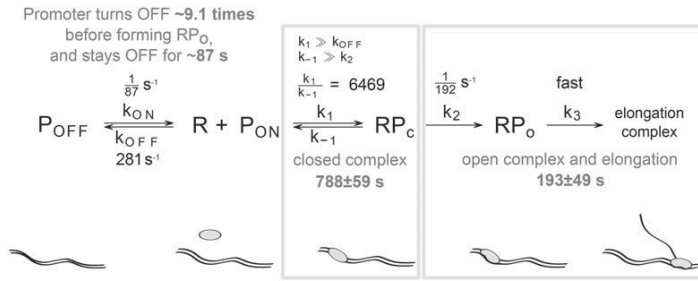
Lastly, we note that the BIC depends on the number of samples used to calculate the likelihood. Thus, BIC values calculated assuming that each censored interval is 'one sample' will over-penalize models with more parameters, while removing them will under-penalize them. Both sets of  $\Delta BIC$  values are presented in Table 2 and, in our case, both result in the same conclusion, and thus the distinction does not affect the results for  $P_{lacIara-1}$ . If, for another promoter, this turns out to be the case, additional measurements will be required to distinguish between the models.

Our results are in agreement with previous measurements of the kinetics of this and similar promoters. For example, a previous study reported that, under full induction in LB media (1x media here),  $P_{lacIara-1}$  expresses  $\sim 4$  RNA/h<sup>2</sup> (i.e. 1 RNA every  $\sim 900$  s), while we inferred the time between transcription events to be  $\sim 980$  s. Using the steady-state assay,  $\tau_{CC}$  was measured to be  $\sim 330$  s for  $P_{lac}$ <sup>71</sup> (with or without CRP-cAMP), while we obtained  $\sim 193$  s.

### 3.4. Determining the source of the intermittent inactive state for $P_{lacIara-1}$

We identified the presence of an ON/OFF mechanism in the dynamics of  $P_{lacIara-1}$ . It is worth noting that this ON/OFF phenomenon differs from the one reported in refs 2 and 19 since, first, we only observe OFF periods on the order of  $\sim 87$  s, while in ref. 2 the OFF periods reported for  $P_{lacIara-1}$  were on the order of 37 min. In addition, both here and in ref. 2, the promoter of interest is integrated in a single-copy plasmid, and thus the OFF periods cannot be explained by the buildup of positive supercoiling, since the plasmid is not topologically constrained.<sup>19</sup> We therefore hypothesized that the ON/OFF periods observed here more likely result from the intermittent formation of a DNA loop, due to the transient binding of LacI, which exists in high concentration in DH5 $\alpha$ -PRO ( $\sim 3,000$  copies vs.  $\sim 20$  in wild type<sup>63</sup>).

If LacI is responsible for the ON/OFF behaviour, then reducing the concentration of IPTG should affect the ON/OFF dynamics, and not



**Figure 5.** Best fitting model of transcription initiation (with ON/OFF mechanism and reversible close complex formation). The model parameters are specified in black and estimated durations of the transcription initiation steps for 1x LB media are shown in grey.

change the dynamics following the closed complex formation.<sup>29</sup> To test this prediction, and demonstrate the utility of the model-fitting approach, besides considering the interval measurements in 1x in Table 1, we also measured the interval distribution of  $P_{lacIara-1}$  using MS2d-GFP in the 1x media without induction by IPTG. From 130 cells, we extracted 57 intervals and 117 censored intervals between transcription events. From these, we inferred a mean interval of  $3,374 \pm 462$  s, and a CV<sup>2</sup> of 1.03. This mean is significantly greater than the mean measured in the fully induced condition ( $1,005 \pm 112$  s), consistent with the much stronger repression of the promoter by LacI in this condition.

Given the wide difference in dynamics of RNA production between the induced and non-induced cases, we used the model fitting procedure to determine which steps are significantly affected by LacI. For this, we performed independent fits of a reduced model of initiation to the induced and the non-induced conditions. This model is observationally equivalent to the full model of initiation (Reactions (1) and (2)) for a single value of  $R$ , and is presented in Reaction (3). This reduced model is necessary since we do not have measurements of the uninduced case at multiple values of  $R$  with which to fit all parameters of the full model. The reduced model's parameters are denoted by  $\lambda_x$ , which are related to, but are not equal to the values of  $k_x$ . Their relationship is presented in Supplementary Table S1. The fitting results are shown in Table 3 (labelled 'Independent'). We also considered joint models where parameters were fixed between conditions, and used the BIC to select the most likely model.

The first three models with joint parameters test for whether or not the parameters controlling the ON/OFF mechanism change with induction strength. Consistent with this hypothesis, the models with joint  $\lambda_{OFF}^{-1}$  are strongly rejected ( $\Delta$ BIC much higher than that of the Independent model). Surprisingly, the model with only joint  $\lambda_{ON}^{-1}$  was also rejected, implying that the mean OFF times might also vary with induction strength. Additional studies are needed to elucidate why such OFF times depend on the induction strength.

Having established that  $\lambda_{ON}^{-1}$  and  $\lambda_{OFF}^{-1}$  differ between conditions, we next assessed whether only these parameters differ. For that, we fixed  $\lambda_1^{-1}$  and  $\lambda_2^{-1}$ , and verified that this model is the most parsimonious model ( $\Delta$ BIC relative to the Independent model of  $-14.3$ ). We conclude that only  $\lambda_{ON}^{-1}$  and  $\lambda_{OFF}^{-1}$  differ between conditions, confirming the prediction that LacI is responsible for the ON/OFF mechanism affecting the RNA production dynamics.

Finally, other models were considered, e.g. the hypothesis that  $\lambda_1^{-1}$ ,  $\lambda_2^{-1}$ , and/or  $\lambda_{ON}^{-1}$  do not differ between conditions. These models were also strongly rejected in favour of the parsimonious model, and are not shown for brevity.

**Table 3.** Fit parameters of the transcription initiation model in Reaction (3) to the measured intervals in the 1x media with and without induction by IPTG

Joint parameters	Condition	$\lambda_{ON}^{-1}$ (s)	$\lambda_{OFF}^{-1}$ (s)	$\lambda_1 \lambda_{OFF}^{-1}$	$\lambda_1^{-1}$ (s)	$\lambda_2^{-1}$ (s)	$\Delta$ BIC
Independent	IPTG+	110	Fast	0.11	Fast	5	14.3
	IPTG-	48	Fast	0.01	Fast	Fast	
$\lambda_{ON}^{-1}$	IPTG+	4,444	Fast	11.50	Fast	964	120.3
	IPTG-		Fast	$\infty$	Fast	2,919	
$\lambda_{OFF}^{-1}$	IPTG+	7	Fast	$\infty$	Fast	964	152.9
	IPTG-	320		1.86	Fast	2,919	
$\lambda_{ON}^{-1}, \lambda_{OFF}^{-1}$	IPTG+	326	Fast	$\infty$	Fast	964	145.7
	IPTG-			1.94	Fast	2,918	
$\lambda_1^{-1}, \lambda_2^{-1}$	IPTG+	106	Fast	0.11	Fast	Fast	0.0
	IPTG-	48	Fast	0.01			

The relationship between these parameters and the parameters in Table 2 are discussed in the Materials and methods and Supplementary Material. Five models are considered, differing in which parameters are assumed to be the same between the two induction conditions. Parameters denoted 'fast' are too fast to present on the timescale of seconds. As  $\lambda_{OFF}^{-1}$  and  $\lambda_1^{-1}$  were found to be fast in all models, the  $\lambda_1 \lambda_{OFF}^{-1}$  ratio is also shown.  $\Delta$ BIC values are given as the difference of the model's BIC from the BIC of the best-fitting model (the one with  $\Delta$ BIC = 0). Models with lower  $\Delta$ BIC are favoured over models with higher  $\Delta$ BIC.<sup>58</sup>

### 3.5. Precision of the estimates

We define the precision of the estimates of  $\tau_{CC}^{-1}$  and  $k_{CC}^{-1}$  as the ratio between the timescale of the intervals (i.e. the mean interval in the condition with greatest  $R$ ) and the standard uncertainties of  $\hat{\tau}_{CC}^{-1}$  and  $\hat{k}_{CC}^{-1}$ , respectively. Specifically, the precision of  $\hat{\tau}_{CC}^{-1}$ 's estimate is  $P_{CC}^{-1} = \hat{I}_1 / \sigma(\hat{\tau}_{CC}^{-1})$ , and the precision of  $\hat{k}_{CC}^{-1}$ 's estimate is  $P_{CC} = \hat{I}_1 / \sigma(\hat{k}_{CC}^{-1})$ . Given this, here, with the volume of data in Table 1, we achieved  $P_{CC}^{-1} = 20.7$  and  $P_{CC} = 17.0$ , corresponding to errors of  $\sim 5$  and  $\sim 6\%$ , respectively.

In addition, we found that this precision is highly dependent on the dynamic range of RNAP concentrations. For example, for a small dynamic range of 1.5 (our measurements in Fig. 2 have a range of  $\sim 2.4$ ), the precisions  $P_{CC}^{-1}$  (in  $\hat{\tau}_{CC}^{-1}$ ) and  $P_{CC}$  (in  $\hat{k}_{CC}^{-1}$ ) would have been reduced to  $\sim 11.2$  and  $\sim 6.7$ , respectively. Losses in precision due to reduced dynamic ranges can, however, to some extent, be offset by collecting more samples for the interval distributions (see estimation of precision in Supplementary Material).

#### 4. Discussion

We established that, in *E. coli*, the concentration of free RNA polymerases differs significantly within a certain range of growth conditions, and that the inverse of the target RNA production rate under the control of  $P_{lacIara-1}$  varies linearly with the inverse of the free RNAP concentration (which are the conditions imposed in the *in vitro* measurements the open complex formation by steady state assays<sup>22,24,72</sup>). Thus, we were able to apply a standard model-fitting procedure to fully characterize the *in vivo* kinetics of the rate-limiting steps in transcription initiation of the  $P_{lacIara-1}$  promoter from distributions of intervals between transcription events in cells with different RNA polymerase concentrations. This revealed that this promoter has two rate-limiting steps: a reversible closed complex formation and a significant open complex formation. Further, it also intermittently switches to a short-lived inactive state. Based on the inferred timescale of this inactive state, we predicted that this state is the result of the intermittent binding of the repressor LacI, which we verified by measuring the interval distribution when the promoter is not induced by IPTG. We believe that the complexity of this process is the reason why it has not been reported before. Namely, previous studies only considered either multiple rate-limiting steps,<sup>4,5,22,23,66</sup> or an ON/OFF process,<sup>2,17,19,73,74</sup> while this promoter exhibits both.

We note that, provided that the promoter has a reversible closed complex formation, the model fitting procedure proposed here allows the duration and order of two steps following the closed complex to be obtained (specifically, the ratio between  $k_2$  and  $k_3$  can be determined from how the  $CV^2$  of the interval distribution changes with R; see Supplementary Fig. S2). Here, this additional step was not found. However, we expect that, for other promoters, or in different conditions (e.g. low temperatures<sup>72</sup>), this step may be significant. Meanwhile, if Limiting Mechanism II is found to be the best-fitting model, the order of the last two steps will remain ambiguous due to the lack of reversibility.

Finally, it is worth noting that in previous works, we have not found evidence for an ON/OFF mechanism for  $P_{lacIara-1}$ , due to the low levels of noise detected in the time intervals between transcription events.<sup>4,21,66</sup> This can be explained by, first, we did not consider censored intervals, which contribute significantly to the increase of the tail of the distribution of intervals.<sup>25</sup> Second, the OFF period is quite short, and thus its detection requires a large volume of data and a sensitive inference methodology.<sup>25</sup> Our results show that, by solving these two issues (by applying the methods in refs 41 and 25), our methodology can identify and characterize many relevant steps in transcription initiation, including those with lesser influence.

In the future, it would be of interest to extend the model to consider what occurs when more than one copy of a promoter is present in the cell. We expect that variations in the promoter copy numbers would, in that case, explain some of the variance of the data, instead of this variance being solely determined by the ON/OFF mechanism and the sequential steps.

We expect the methodology employed here to be applicable to promoters, native or synthetic, whose changes in the inverse of the transcription rate are linear with the inverse of the free RNAP concentrations. Also, it should be applicable to promoters evolved to interact with multiple transcription factors (TF), provided their fast binding and unbinding (compared with competing events), as they could be accounted for by tuning the rate constants of some of the reactions of the model. Further, multiple slow TFs, including activators, can be accounted for by adding appropriate TF-bound states, with differing production rates, in a similar manner to the ON/OFF

model. As such, the methodology should be applicable at a genome wide scale. It should also be applicable to eukaryotes, provided suitable means to alter polymerase concentrations. Lastly, it should be useful in detecting differences in transcription initiation kinetics of a promoter subject to different intra- or extra-cellular conditions.

#### Acknowledgements

We thank Axel Oikari, Abhishekh Gupta, and Antti Martikainen for valuable advice.

#### Supplementary Data

Supplementary Data are available at [www.dnaresearch.oxfordjournals.org](http://www.dnaresearch.oxfordjournals.org).

#### Funding

This work was supported by the Academy of Finland (257603 to A.S.R.); Centre of International Mobility (13.1.2014/TM-14-91361/CIMO to A.S.R.); Jenny and Antti Wihuri Foundation (to A.H.); and the Tampere University of Technology President's Graduate Programme (to J.L.-P. and S.S.). Funding to pay the Open Access publication charges for this article was provided by Academy of Finland (257603 to A.S.R.).

#### References

- Golding, I. and Cox, E.C. 2004, RNA dynamics in live *Escherichia coli* cells, *Proc. Natl. Acad. Sci. USA*, **101**, 11310–5.
- Golding, I., Paulsson, J., Zawilski, S.M. and Cox, E.C. 2005, Real-time kinetics of gene activity in individual bacteria, *Cell*, **123**, 1025–36.
- Yu, J., Xiao, J., Ren, X., Lao, K. and Xie, X.S. 2006, Probing gene expression in live cells, one protein molecule at a time, *Science*, **311**, 1600–3.
- Kandhavelu, M., Mannerström, H., Gupta, A., et al. 2011, *In vivo* kinetics of transcription initiation of the *lar* promoter in *Escherichia coli*. Evidence for a sequential mechanism with two rate-limiting steps, *BMC Syst. Biol.*, **5**, 149.
- 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.
- Fusco, D., Accornero, N., Lavoie, B., et al. 2003, Single mRNA Molecules Demonstrate Probabilistic Movement in Living Mammalian Cells, *Curr. Biol.*, **13**, 161–7.
- Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. and Tyagi, S. 2006, Stochastic mRNA synthesis in mammalian cells, *PLoS Biol.*, **4**, 1707–19.
- Kaern, M., Elston, T.C., Blake, W.J. and Collins, J.J. 2005, Stochasticity in gene expression: from theories to phenotypes, *Nat. Rev. Genet.*, **6**, 451–64.
- Arkin, A.P., Ross, J. and McAdams, H.H. 1998, Stochastic kinetic analysis of developmental pathway bifurcation in phage  $\lambda$ -infected *Escherichia coli* cells, *Genetics*, **149**, 1633–48.
- Elowitz, M.B., Levine, A.J., Siggia, E.D. and Swain, P.S. 2002, Stochastic gene expression in a single cell, *Science*, **297**, 1183–6.
- Raser, J.M. and O'Shea, E.K. 2005, Noise in gene expression: origins, consequences, and control, *Science*, **309**, 2010–3.
- Ozbudak, E.M., Thattai, M., Kurtser, I., Grossman, A.D. and van Oudenaarden, A. 2002, Regulation of noise in the expression of a single gene, *Nat. Genet.*, **31**, 69–73.
- McAdams, H.H. and Arkin, A.P. 1999, It's a noisy business! Genetic regulation at the nanomolar scale, *Trends Genet.*, **15**, 65–9.
- Süel, G.M., Garcia-Ojalvo, J., Liberman, L.M. and Elowitz, M.B. 2006, An excitable gene regulatory circuit induces transient cellular differentiation, *Nature*, **440**, 545–50.
- Cai, L., Friedman, N. and Xie, X.S. 2006, Stochastic protein expression in individual cells at the single molecule level, *Nature*, **440**, 358–62.
- 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**, e1000109.



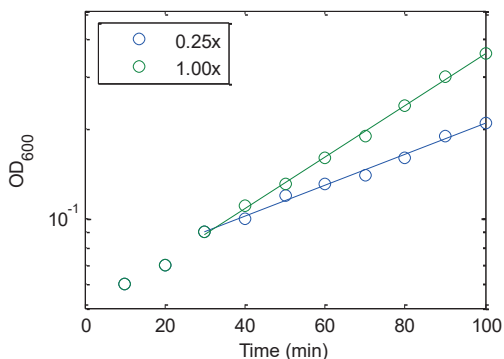
17. So, L.-H., Ghosh, A., Zong, C., Sepúlveda, L.A., Segev, R. and Golding, I. 2011, General properties of transcriptional time series in *Escherichia coli*, *Nat. Genet.*, **43**, 554–60.
18. Zhdanov, V.P. 2011, Kinetic models of gene expression including non-coding RNAs, *Phys. Rep.*, **500**, 1–42.
19. Chong, S., Chen, C., Ge, H. and Xie, X.S. 2014, Mechanism of transcriptional bursting in bacteria, *Cell*, **158**, 314–26.
20. Kandhavelu, M., Häkkinen, A., Yli-Harja, O. and Ribeiro, A.S. 2012, Single-molecule dynamics of transcription of the *lar* promoter, *Phys. Biol.*, **9**, 026004.
21. Kandhavelu, M., Lloyd-Price, J., Gupta, A., Muthukrishnan, A.-B., 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 *lacIara-1* promoter, *FEBS Lett.*, **586**, 3870–5.
22. McClure, W.R. 1980, Rate-limiting steps in RNA chain initiation, *Proc. Natl Acad. Sci. USA*, **77**, 5634–8.
23. McClure, W.R. 1985, Mechanism and control of transcription initiation in prokaryotes, *Annu. Rev. Biochem.*, **54**, 171–204.
24. Bertrand-Burggraf, E., Lefèvre, J.F. and Daune, M. 1984, A new experimental approach for studying the association between RNA polymerase and the *tet* promoter of pBR322, *Nucleic Acids Res.*, **12**, 1697–706.
25. Häkkinen, A. and Ribeiro, A.S. 2016, Characterizing rate limiting steps in transcription from RNA production times in live cells, *Bioinformatics*, <http://bioinformatics.oxfordjournals.org/content/early/2016/01/28/bioinformatics.btv744.abstract>.
26. Friedman, L.J. and Gelles, J. 2012, Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation, *Cell*, **148**, 679–89.
27. 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**, 533–8.
28. Sanchez, A., Garcia, H.G., Jones, D., Phillips, R. and Kondev, J. 2011, Effect of promoter architecture on the cell-to-cell variability in gene expression, *PLoS Comput. Biol.*, **7**, e1001100.
29. 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.
30. Garcia, H.G., Sanchez, A., Boedicker, J.Q., et al. 2012, Operator sequence alters gene expression independently of transcription factor occupancy in bacteria, *Cell Rep.*, **2**, 150–61.
31. Gummesson, B., Magnusson, L.U., Lovmar, M., et al. 2009, Increased RNA polymerase availability directs resources towards growth at the expense of maintenance, *EMBO J.*, **28**, 2209–19.
32. Bremer, H. and Dennis, P.P. 1996, Modulation of Chemical Composition and Other Parameters of the Cell by Growth Rate. In: Neidhardt, F.C., (ed.), *Escherichia Coli and Salmonella*, 2nd ed. ASM Press, Washington, DC, pp. 1553–69.
33. Bratton, B.P., Mooney, R.A. and Weisshaar, J.C. 2011, Spatial distribution and diffusive motion of RNA polymerase in live *Escherichia coli*, *J. Bacteriol.*, **193**, 5138–46.
34. Dillon, S.C. and Dorman, C.J. 2010, Bacterial nucleoid-associated proteins, nucleoid structure and gene expression, *Nat. Rev. Microbiol.*, **8**, 185–95.
35. Liang, S.-T., Bipatnath, M., Xu, Y.-C., et al. 1999, Activities of constitutive promoters in *Escherichia coli*, *J. Mol. Biol.*, **292**, 19–37.
36. Livak, K.J. and Schmittgen, T.D. 2001, Analysis of relative gene expression data using real-time quantitative PCR and the  $2^{-\Delta\Delta Ct}$  Method, *Methods*, **25**, 402–8.
37. Gupta, A., Lloyd-Price, J., Oliveira, S.M.D., Yli-Harja, O., Muthukrishnan, A.-B. and Ribeiro, A.S. 2014, Robustness of the division symmetry in *Escherichia coli* and functional consequences of symmetry breaking, *Phys. Biol.*, **11**, 066005.
38. Chowdhury, S., Kandhavelu, M., Yli-Harja, O. and Ribeiro, A.S. 2013, Cell segmentation by multi-resolution analysis and maximum likelihood estimation (MAMLE), *BMC Bioinformatics*, **14**, S8.
39. Häkkinen, A., Muthukrishnan, A.-B., Mora, A., Fonseca, J.M. and Ribeiro, A.S. 2013, CellAging: A tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*, *Bioinformatics*, **29**, 1708–9.
40. Häkkinen, A., Kandhavelu, M., Garasto, S. and Ribeiro, A.S. 2014, Estimation of fluorescence-tagged RNA numbers from spot intensities, *Bioinformatics*, **30**, 1146–53.
41. Häkkinen, A. and Ribeiro, A.S. 2015, Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data, *Bioinformatics*, **31**, 69–75.
42. 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**, 2579–87.
43. Johansson, H.E., Dertinger, D., LeCuyer, K.A., Behlen, L.S., Greef, C.H. and Uhlenbeck, O.C. 1998, A thermodynamic analysis of the sequence-specific binding of RNA by bacteriophage MS2 coat protein, *Proc. Natl Acad. Sci. USA*, **95**, 9244–9.
44. Golding, I. and Cox, E.C. 2006, Physical Nature of Bacterial Cytoplasm, *Phys. Rev. Lett.*, **96**, 98102.
45. 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.
46. 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.
47. DeHaseth, P.L., Zupancic, M.L. and Record, M.T. 1998, RNA polymerase-promoter interactions: The comings and goings of RNA polymerase, *J. Bacteriol.*, **180**, 3019–25.
48. Chamberlin, M.J. 1974, The selectivity of transcription, *Annu. Rev. Biochem.*, **43**, 721–75.
49. Hsu, L.M. 2009, Monitoring abortive initiation, *Methods*, **47**, 25–36.
50. Mulligan, M.E., Hawley, D.K., Enriken, R. and McClure, W.R. 1984, *Escherichia coli* promoter sequences predict in vitro RNA polymerase selectivity, *Nucleic Acids Res.*, **12**, 789–800.
51. Bai, L., Santangelo, T.J. and Wang, M.D. 2006, Single-molecule analysis of RNA polymerase transcription, *Annu. Rev. Biophys. Biomol. Struct.*, **35**, 343–60.
52. Wang, F. and Greene, E.C. 2011, Single-molecule studies of transcription: From one RNA polymerase at a time to the gene expression profile of a cell, *J. Mol. Biol.*, **412**, 814–31.
53. Artsimovitch, I. and Landick, R. 2000, Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals, *Proc. Natl Acad. Sci. USA*, **97**, 7090–5.
54. Rajala, T., Häkkinen, A., Healy, S., Yli-Harja, O. and Ribeiro, A.S. 2010, Effects of transcriptional pausing on gene expression dynamics, *PLoS Comput. Biol.*, **6**, e1000704.
55. Bar-Nahum, G. and Nudler, E. 2001, Isolation and characterization of  $\sigma^{70}$ -retaining transcription elongation complexes from *Escherichia coli*, *Cell*, **106**, 443–51.
56. Grigorova, I.L., Phleger, N.J., Mutalik, V.K. and Gross, C.a. 2006, Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA, *Proc. Natl Acad. Sci. USA*, **103**, 5332–7.
57. Maeda, H., Fujita, N. and Ishihama, A. 2000, Competition among seven *Escherichia coli*  $\sigma$  subunits: relative binding affinities to the core RNA polymerase, *Nucleic Acids Res.*, **28**, 3497–503.
58. Kass, R.E. and Raftery, A.E. 1995, Bayes Factors, *J. Am. Stat. Assoc.*, **90**, 773–95.
59. Hsu, L.M. 2002, Promoter clearance and escape in prokaryotes, *Biochim. Biophys. Acta*, **1577**, 191–207.
60. Casella, G. and Berger, R.L. 2001, *The Delta Method. Statistical Inference*, 2nd ed. Duxbury Press, Pacific Grove, CA, pp. 240–5.
61. Krystek, M. and Anton, M. 2008, A weighted total least-squares algorithm for fitting a straight line, *Meas. Sci. Technol.*, **19**, 79801.
62. Klumpp, S. and Hwa, T. 2008, Growth-rate-dependent partitioning of RNA polymerases in bacteria, *Proc. Natl Acad. Sci. USA*, **105**, 20245–50.

63. Lutz, R. and Bujard, H. 1997, Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I<sub>1</sub>-I<sub>2</sub> regulatory elements, *Nucleic Acids Res.*, **25**, 1203–10.
64. Stricker, J., Cookson, S., Bennett, M.R., Mather, W.H., Tsimring, L.S. and Hasty, J. 2008, A fast, robust and tunable synthetic gene oscillator, *Nature*, **456**, 516–9.
65. Martins, L., Mäkelä, J., Häkkinen, A., et al. 2012, Dynamics of transcription of closely spaced promoters in *Escherichia coli*, one event at a time, *J. Theor. Biol.*, **301**, 83–94.
66. Mäkelä, J., Kandhavelu, M., Oliveira, S.M.D., et al. 2013, *In vivo* single-molecule kinetics of activation and subsequent activity of the arabinose promoter, *Nucleic Acids Res.*, **41**, 6544–52.
67. Kandhavelu, M., Lihavainen, E., Muthukrishnan, A.B., Yli-Harja, O. and Ribeiro, A.S. 2012, Effects of Mg<sup>2+</sup> on *in vivo* transcriptional dynamics of the *lar* promoter, *BioSystems*, **107**, 129–34.
68. Pedraza, J.M. and Paulsson, J. 2008, Effects of molecular memory and bursting on fluctuations in gene expression, *Science*, **319**, 339–43.
69. Casella, G. and Berger, R.L. 2001, *Simultaneous Estimation and Confidence Bands. Statistical Inference*, 2nd ed. Duxbury Press, Pacific Grove, CA, USA, pp. 559–63.
70. Schwarz, G. 1978, Estimating the dimension of a model, *Ann. Stat.*, **6**, 461–4.
71. Malan, T.P., Kolb, A., Buc, H. and McClure, W.R. 1984, Mechanism of CRP-cAMP activation of *lac* operon transcription initiation activation of the *P1* promoter, *J. Mol. Biol.*, **180**, 881–909.
72. Buc, H. and McClure, W.R. 1985, Kinetics of open complex formation between *Escherichia coli* RNA polymerase and the *lac* UV5 promoter. Evidence for a sequential mechanism involving three steps, *Biochemistry*, **24**, 2712–23.
73. Sanchez, A., Choubey, S. and Kondev, J. 2013, Stochastic models of transcription: From single molecules to single cells, *Methods*, **62**, 13–25.
74. Schwabe, A., Rybakova, K.N. and Bruggeman, F.J. 2012, Transcription stochasticity of complex gene regulation models, *Biophys. J.*, **103**, 1152–61.

# Supplement to “Dissecting the stochastic transcription initiation process in live *Escherichia coli*”

Jason Lloyd-Price, Sofia Startceva, Vinodh Kandavalli, Jerome G. Chandraseelan, Nadia Goncalves, Samuel M. D. Oliveira, Antti Häkkinen and Andre S. Ribeiro

## I. Growth Curves



**Supplementary Figure S1:** Growth curves (OD<sub>600</sub>, measured with an Ultraspec 10 cell density meter) of cells in 1x and 0.25x media (circles) at 37 °C. DH5α-PRO cells were grown overnight in 1x media at 30 °C with aeration of 250 rpm, and diluted into fresh 1x media to an initial OD<sub>600</sub> of 0.05. Cells were incubated at 37 °C at 250 rpm until reaching the mid-log phase (~2 h), and re-diluted into the appropriate medium to an OD<sub>600</sub> of 0.05. Their OD<sub>600</sub> was measured every 10 minutes thereafter. At ~30 min, the cells in 0.25x media adjusted their growth rate (before this, the measurements overlap). Thus, growth rates were measured by least-squares fits (lines) from the data from 30 min onward. The slopes of the fits correspond to doubling times of 34.4 min (1.00x) and 57.9 min (0.25x).

## II. Models of transcription initiation

To evaluate the cumulative distribution function (CDF) of the distribution of time intervals between production events from the full model of transcription initiation for a given value of  $R$ , we first translate this model into an observationally equivalent model of the form in equation 3. For the full model, this translation is given in the first row of Supplementary Table S1. The translated model’s CDF can be evaluated using <sup>1</sup>. This CDF, when there are  $n$  steps after  $S_0$ , is referred to here as  $F_{\text{ON/OFF}+n}$ . This distribution has a mean and variance of:

$$\mu_{\text{ON/OFF}+n} = \frac{\lambda_{\text{OFF}}}{\lambda_1 \lambda_{\text{ON}}} + \sum_{i=1}^n \lambda_i^{-1} \quad (\text{S1})$$

$$\sigma_{\text{ON/OFF}+n}^2 = \frac{\lambda_{\text{OFF}}}{\lambda_1^2 \lambda_{\text{ON}}} \left( 2 + \frac{2\lambda_1 + \lambda_{\text{OFF}}}{\lambda_{\text{ON}}} \right) + \sum_{i=1}^n \lambda_i^{-2} \quad (\text{S2})$$

Assumptions	CDF	$\lambda_{\text{ON}}$	$\lambda_{\text{OFF}}$	$\lambda_1$	$\lambda_2$	$\lambda_3$
	$F_{\text{ON/OFF}+3}$	$k_{\text{ON}}$	$\frac{-k_{\text{ON}}}{(Q_0 - k_{\text{ON}})(Q_2 - k_{\text{ON}})}$	$\frac{Q_1}{k_1 k_2}$	$Q_1^{-1}$	$k_3$
$k_{-1} \gg k_2, k_1 \gg k_{\text{OFF}}$	$F_{\text{ON/OFF}+2}$	$k_{\text{ON}}$	$\frac{k_{\text{OFF}}}{RK_a + 1}$	$\frac{k_2 RK_a}{RK_a + 1}$	$k_3$	
$k_2 \gg k_{-1}$	$F_{\text{ON/OFF}+3}$	$k_{\text{ON}}$	$k_{\text{OFF}}$	$Rk_1$	$k_2$	$k_3$
$k_{\text{ON}} \gg k_1$	$F_{\text{Hypo}(3)}$			$\frac{u+v}{2}$	$\frac{u-v}{2}$	$k_3$
$k_{\text{ON}} \gg k_1, k_{-1} \gg k_2$	$F_{\text{Hypo}(2)}$			$\frac{k_2 RK_a}{RK_a + 1}$	$k_3$	
$k_{\text{ON}} \gg k_1, k_2 \gg k_{-1}$	$F_{\text{Hypo}(3)}$			$Rk_1$	$k_2$	$k_3$

**Supplementary Table S1:** Relation between kinetic parameters from equations (1) and (2) of the main manuscript with the parameters of the model from equation (3), for a given value of R. Here,  $K_a = k_1 k_{-1}^{-1}$ ,  $u = Rk_1 + k_{-1} + k_2$ ,  $v = \sqrt{(k_{-1} + k_2 - Rk_1)^2 + 4Rk_1 k_{-1}}$ , and  $Q_n$  are the roots of  $-x^3 + bx^2 - cx + d^*$ , where  $b = u + k_{\text{ON}} + k_{\text{OFF}}$ ,  $c = uk_{\text{ON}} + k_{\text{OFF}}(k_{-1} + k_2) + Rk_1 k_2$ ,  $d = Rk_1 k_2 k_{\text{ON}}$ , ordered such that  $\lambda_{\text{OFF}} \geq 0$ .

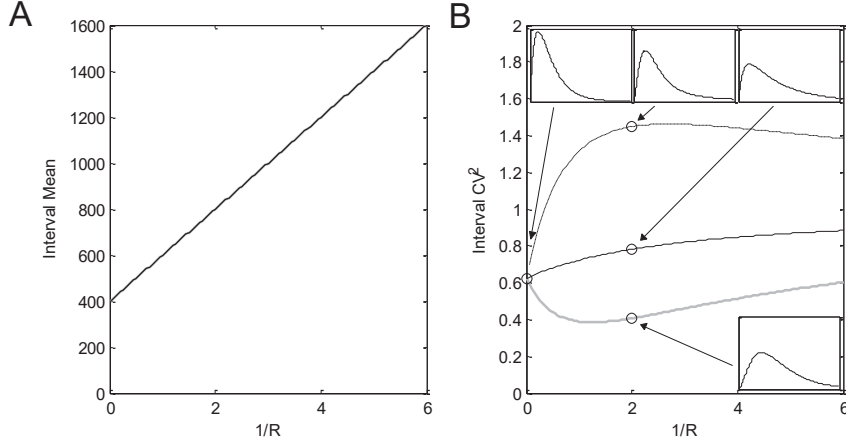
In the manuscript, several limiting cases of this model are considered. The first is that the ON/OFF mechanism is fast relative to initiation, i.e.  $k_{\text{ON}} \gg k_1$ . In this case, the model's CDF simplifies to that of a hypoexponential distribution with three exponentials with rates  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$ , which relate to the parameters of 0 as shown in the fourth row of Supplementary Table S1. The hypoexponential CDF with  $n$  exponentials is referred to here as  $F_{\text{Hypo}(n)}$ .

Two further simplifications are considered, referred to in the manuscript as Limiting Mechanisms I and II. Both of these result in models with CDFs that are equivalent to either  $F_{\text{ON/OFF}+n}$  or  $F_{\text{Hypo}(n)}$ . The parameters of the CDFs of the models derived from these three simplifying assumptions are presented in Supplementary Table S1. The final model simplification considered in the manuscript is when  $k_3 = \infty$ , i.e. when there is no rate-limiting third step in initiation, which removes the step parameterized by  $k_3$  from the model.

The model of transcription initiation predicts the same linear change in the mean interval duration with  $1/R$ , regardless of the model simplifications (Figure S2A). However, the different simplifications result in different distributions of intervals as a function of  $1/R$ , which will differ in, e.g., noise (Figure S2B).

---

\*  $Q_n$  can be evaluated with  $Q_n = -2\sqrt{p} \cos \left[ \frac{1}{3} \left( \cos^{-1} \left( \frac{-q}{2p^{3/2}} \right) - 2\pi n \right) \right] + \frac{b}{3}$ , where  $p = \frac{b^2 - 3c}{9}$  and  $q = \frac{b}{3} \left( \frac{9b^2}{2} - c \right) + d$ .



**Supplementary Figure S2:** Model prediction for (A) mean and (B)  $CV^2$  of intervals as a function of  $1/R$  with assumptions  $k_2 \gg k_{-1}$  (dashed black line,  $k_{ON}^{-1} = 1000$ ,  $k_{OFF}^{-1} = 200$ ,  $k_1^{-1} = 200k_{ON}(k_{ON} + k_{OFF})^{-1}$ ,  $k_2^{-1} = 300$ ,  $k_3^{-1} = 100$ ),  $k_{ON} \gg k_1$ ,  $k_{-1} \gg k_2$  (black lines,  $K_a = 1.5$ ,  $k_2^{-1} = 300$ ,  $k_3^{-1} = 100$ ), and  $k_{ON} \gg k_1$ ,  $k_2 \gg k_{-1}$  (grey lines,  $k_1^{-1} = 200$ ,  $k_2^{-1} = 300$ ,  $k_3^{-1} = 100$ ). Note that in (A), all three lines overlap. Interval distributions for several parameter sets are shown in the insets of (B) (the axes of the insets are the same).

### III. Parameter Estimation

Model parameter estimation was performed using a censored log-likelihood objective function as in <sup>1</sup>, which accounts for uncertainty in the measurement of  $R$ , and for the uncertainty in the interval durations that arises from the limited framerate of the measurements and from the limited observation time:

$$\log L(\boldsymbol{\theta}) = \sum_m \mathbb{E} \log L_m(\boldsymbol{\theta}; R^{-1}) \quad (S3)$$

where  $\mathbb{E}$  is the expectation over  $R^{-1}$ , and the conditional log-likelihood for condition  $m$  at relative RNAp concentration  $R$  is:

$$\begin{aligned} \log L_m(\boldsymbol{\theta}; R^{-1}) = & \sum_i \log [F_{\mathcal{M}}(t_{m,i} + T_M; \boldsymbol{\theta}, R^{-1}) - F_{\mathcal{M}}(\max(0, t_{m,i} - T_M); \boldsymbol{\theta}, R^{-1})] \\ & + \sum_i \log [1 - F_{\mathcal{M}}(c_{m,i}; \boldsymbol{\theta}, R^{-1})] \end{aligned} \quad (S4)$$

where  $F_{\mathcal{M}}(x; \boldsymbol{\theta}, R^{-1})$  is the CDF of the model being fit (either  $F_{ON/OFF+n}$  or  $F_{Hypo(n)}$ ) with parameters translated as appropriate using Supplementary Table S1,  $\boldsymbol{\theta}$  is the parameter vector,  $t_{m,i}$  are measured intervals in condition  $m$ ,  $T_M$  is the time between frames, and  $c_{m,i}$  are the right-censored intervals.

The expectation of  $\log L_m(\boldsymbol{\theta}; R^{-1})$  over  $R$  in equation (S3) accounts for the uncertainty in the measurement of  $R$ . This was performed with  $R^{-1} \sim \mathcal{N}(\hat{R}_m^{-1}, \sigma^2(\hat{R}_m^{-1}))$ , which was approximated by

evaluating the conditional log likelihood at 21 equally-spaced points in the interval  $\left[\hat{R}_m^{-1} - 3\sigma(\hat{R}_m^{-1}), \hat{R}_m^{-1} + 3\sigma(\hat{R}_m^{-1})\right]$ .

Fitting was performed using the ‘fminsearch’ function in Matlab, with multiple restarts, to ensure that a local minimum was not selected. Each restart was started randomly in the parameter subspace where the model’s mean interval at  $R=1$  matched the corresponding measured mean interval.

The Bayesian Information Criterion (BIC) was used to compare models. We selected it over other candidates, such as the Akaike Information Criterion (AIC), due to its consistency. That is, as the number of samples  $n \rightarrow \infty$ , the probability that the BIC will select the true model (assuming it is among the candidate models) approaches 1, while the AIC will tend to over-fit the data<sup>2</sup>. We note, however, that in the case of all model comparisons in the manuscript, none of the conclusions are altered by utilizing the AIC over the BIC.

The BIC is calculated as follows:

$$\text{BIC} = -2 \log L(\boldsymbol{\theta}_{\max}) + \log n \quad (\text{S5})$$

where  $\boldsymbol{\theta}_{\max}$  is the parameter set which maximizes  $\log L(\boldsymbol{\theta})$ .

#### IV. Number of transitions into the OFF state per RNA production event

In this section, we estimate the number of times that, on average, a promoter will transit into the OFF state for each time it commits to transcription. This estimation is made for the best fitting model (see Table 2 in the main manuscript).

For the best-fitting model (Limiting Mechanism I), the back-and-forward transitions between  $P_{ON}$  and  $RP_c$  states can be considered to be fast (since  $k_{-1} \gg k_2$  and  $k_1 \gg k_{OFF}$ ). We can therefore apply the slow-scale SSA to merge these two states<sup>3</sup>. In this limit, the probabilities  $P(P_{ON})$  and  $P(P_c)$  of being in  $P_{ON}$  and  $P_c$  states, respectively, are:

$$P(P_{ON}) = \frac{1}{K_a + 1} \text{ and } P(P_c) = \frac{K_a}{K_a + 1}, \text{ with } K_a = k_1 k_{-1}^{-1} \quad (\text{S6})$$

The propensity of changing from the merged state to  $RP_o$  is then  $[P(P_c)k_2]$ , while the propensity to move from the merged state to  $P_{OFF}$  equals  $[P(P_{ON})k_{OFF}]$ . The probability of moving into  $P_c$  instead of  $P_{OFF}$  is therefore given by:

$$P_{c/OFF} = \frac{P(P_c)k_2}{P(P_{ON})k_{OFF} + P(P_c)k_2} \quad (\text{S7})$$

Since each attempt at transcription is independent in the model, and has a constant probability of committing at each attempt, the number of times that the systems changes into the OFF state prior to committing to transcription follows a geometric distribution with a probability of success of  $P_{c/OFF}$ . The mean of this distribution is:

$$\mu = \frac{1 - P_{c/OFF}}{P_{c/OFF}} \quad (S8)$$

Converting this in terms of model parameters (and given  $k_1 k_2 k_{-1}^{-1} k_{OFF}^{-1} = 0.11$  from Table 2) one obtains:

$$\mu = \frac{k_{OFF}}{k_2 K_a} = \frac{1}{0.11} \quad (S9)$$

## V. Minimum samples required for a given precision

To estimate the number of samples required to obtain a given precision in the estimates of  $\tau_{CC}$  and  $\tau_{CC}$ , consider the following alternate method of measuring these values if we could sample the uncensored interval distribution between transcription events.

Let these measurements be at two RNAP concentrations  $\hat{R}_m$ , where  $m = \{1, 2\}$  such that  $D = \hat{R}_1 / \hat{R}_2 > 1$ . Let  $I_m$  be the population mean of the inter-transcription intervals in medium  $m$ , with corresponding standard deviation  $\sigma_m$ , and that we have  $n_m$  samples of this distribution (we assume, without significant loss of generality, that  $n_1 = n_2 = n$ ). For sufficient  $n$ , estimates of the population means  $\hat{I}_m$  will follow Normal distributions with  $\sigma^2(\hat{I}_m) = \sigma_m^2/n$ . The least-squares fit of a line to these points will thus result in:

$$\hat{\tau}_{CC} = \frac{\hat{I}_1 D - \hat{I}_2}{D-1}, \quad \sigma(\hat{\tau}_{CC}) = \frac{D^2 \sigma_1^2 + \sigma_2^2}{n(D-1)^2} \quad (S10)$$

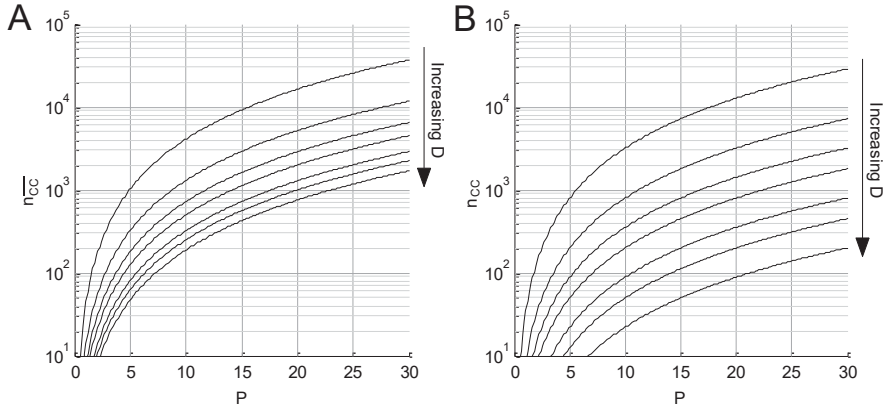
$$\hat{k}_{CC}^{-1} = \frac{\hat{I}_2 - \hat{I}_1}{D-1}, \quad \sigma(\hat{k}_{CC}^{-1}) = \frac{\sigma_1^2 + \sigma_2^2}{n(D-1)^2} \quad (S11)$$

Note that this method will overestimate the uncertainty in  $\hat{\tau}_{CC}$  and  $\hat{k}_{CC}^{-1}$  since these estimates are highly anti-correlated. We define the precision of the measurement as  $P = I_x / \sigma(\hat{\tau}_x)$ , where  $\hat{\tau}_x$  is  $\hat{\tau}_{CC}$  or  $\hat{k}_{CC}^{-1}$ . Intuitively, this definition relates the uncertainty in the estimate with the mean timescale of the intervals. For example, if the intervals are on a timescale of  $\sim 500$  s, to achieve a precision of 10 in  $\hat{\tau}_{CC}$ , we must know it to within 50 s. Assuming that  $\sigma_1^2 I_1^{-2} \approx \sigma_2^2 I_2^{-2} = \eta^2$ , i.e. that the CV<sup>2</sup> of the interval distribution is similar between the two RNAP concentrations, the number of samples required to achieve a given precisions in  $\hat{\tau}_{CC}$  and  $\hat{k}_{CC}^{-1}$  is:

$$n_{CC} = \eta^2 P^2 \frac{D^2 + 1}{(D-1)^2} \quad (S12)$$

$$n_{CC} = \eta^2 P^2 \frac{2}{(D-1)^2} \quad (S13)$$

Note that the above assumes that there is no variance in the estimate of the RNAP concentration, and that all  $n$  samples are uncensored. Equations (S12) and (S13) should therefore be considered as only a rough guide for the number of samples required. The number of samples required for a range of precisions and possible dynamic ranges in RNAP concentrations is shown in Supplementary Figure S3.



**Supplementary Figure S3:** Number of samples required in two conditions to achieve a given precision in (A)  $\hat{r}_{CC}$  and (B)  $\hat{k}_{CC}^{-1}$ , with production interval measurements at only two RNAP concentrations with ratio  $D$  and assuming  $\eta^2 = 1$ . Lines are shown for values of  $D$  of 1.25, 1.5, 1.75, 2, 2.5, 3, and 4 (from top to bottom).

## VI. Photo-toxicity measurements

To assess the level of phototoxicity from the imaging procedure under the microscope, we took the measurements in the 1.00x case (Table 1, main manuscript), and estimated the cells' doubling time under the microscope by counting the number of cells at the start and end of the two hour measurement period (first row of Table S2). In this case, cells were imaged by phase contrast every 5 minutes, and confocal microscopy every minute for two hours. We then imaged two new populations of cells, but in the first, we only imaged the cells with phase contrast (i.e. no confocal, row 2 of Table S2), while in the second, only two images were taken in total, one at the start and one at the end (row 3 of Table S2).

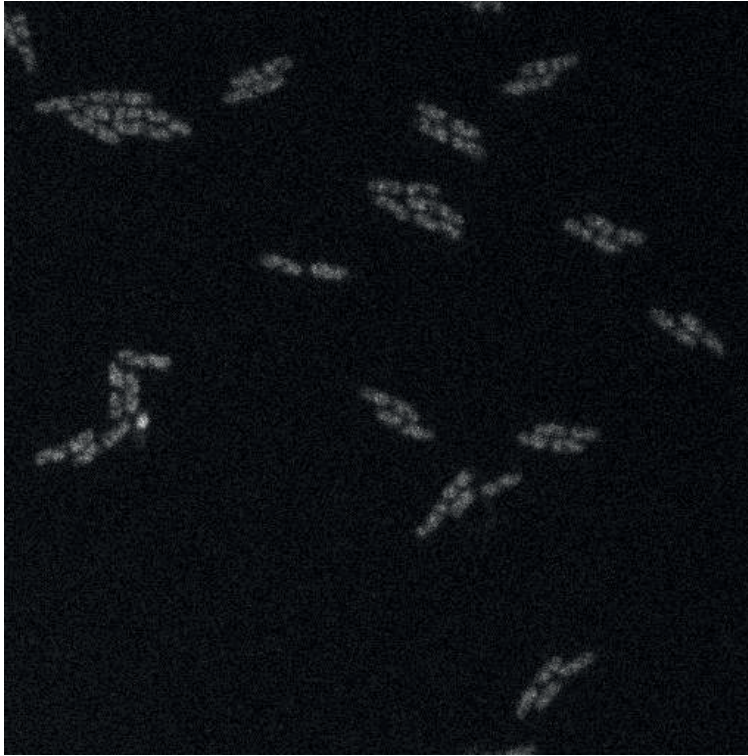
Phase Contrast	Confocal	Cells at start	Cells at end	Doubling Time
5 min	1 min	206	468	52.8 min
5 min	Not used	399	962	49.8 min
2 h	Not used	480	1189	48.4 min

**Supplementary Table S2:** Phototoxicity under the microscope for different imaging intervals and channels. All measurements took 2 hours. The first two columns of the table show the intervals at which images were taken. The subsequent columns show the number of cells at the start and end of the measurements, obtained from single phase contrast images. Finally, it is shown the estimated doubling time of the cells, which was determined from the fold change.

From Supplementary Table S2, the estimated doubling time while taking images with both channels is only 4.4 minutes longer than in the case with minimal imaging. Thus, while there is an observable effect on the doubling time, it is not expected to cause significant differences in the transcription initiation dynamics. In any case, any changes would affect all conditions similarly, and will not affect *relative* RNAP concentrations. Finally, we note that the effect from phase contrast imaging appears to be negligible.



## VII. Cell-to-cell variability in RNAP concentrations



**Supplementary Figure S4:** Confocal image of RL1314 cells expressing fluorescently-tagged RpoC in 1x media, one hour after being placed in the thermal imaging chamber at 37 °C. Contrast was enhanced for easier visualization.

## VIII. Number of promoter copies during the cell lifetime

The model fitting procedure employed in the main text assumes that there is only one copy of the target promoter in a cell at all times. To determine to what extent this assumption is not true in our experimental system, we measured the fraction of time cells contain two chromosomes. Since the F-plasmid replicates at the same time<sup>4</sup> or shortly after<sup>5</sup> the chromosome, this provides an upper bound for the fraction of time the cells spend with more than one promoter of interest (it is worth noting that, in our measurements, we did not observe cells with more than 2 nucleoids at any given point).

For this, *E. coli* DH5 $\alpha$ -PRO cells (see main text) were transformed with the pAB332 plasmid carrying the gene *hupA-mcherry* that encodes a fluorescent protein tag under the control of the *hupA* constitutive promoter<sup>6</sup>. This tagging protein, composed of a nucleoid-associated protein (HupA) fused with a red fluorescent protein (mCherry), can be used to assess the location and size of nucleoids in live cells<sup>7</sup> (see Methods).

Cells were diluted from overnight culture to an  $OD_{600}$  of 0.05 in fresh 1x media, supplemented with appropriate antibiotics, and kept at 37°C in a shaker at 250 rpm, until reaching an  $OD_{600}$  of 0.3. Cells were then placed in a thermal chamber (FCS2, Bioprotechs, USA), set to 37°C, and imaged once every minute for 1 hour (the red signal was too weak to continue after 1 hour) using a Nikon Eclipse (Ti-E, Nikon) inverted microscope equipped with C2+ (Nikon) confocal laser-scanning system. To visualise HupA-mCherry-tagged nucleoids, we used a 543 nm HeNe laser (Melles-Griot) and an emission filter (HQ585/65, Nikon). Phase contrast images of cells were captured every 5 minutes by a CCD camera (DS-Fi2, Nikon).

Cells were segmented from phase contrast images using CellAging<sup>8</sup>. Fluorescent nucleoids were segmented and quantified from confocal images as in <sup>7,9</sup>. Of the cells that were born and divided during the time series (124 cells), we found that the mean fraction of time points in which cells had two nucleoids was  $0.114 \pm 0.010$ .

Thus, we estimate the fraction of time spent with multiple target promoters to be at most  $11.4 \pm 1.0\%$  in 1x media. As this was the most nutrient-rich condition tested, other conditions should have even lower fractions<sup>5</sup>.

## References

1. Häkkinen, A., and Ribeiro, A. S. 2015, Characterizing rate limiting steps in transcription from RNA production times in live cells. *Bioinformatics*, in press. DOI: 10.1093/bioinformatics/btv744.
2. Burnham, K. P., and Anderson, D. R. 2004, Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociol. Methods Res.*, **33**, 261–304.
3. Cao, Y., Gillespie, D. T., and Petzold, L. R. 2005, The slow-scale stochastic simulation algorithm. *J. Chem. Phys.*, **122**, 14116.
4. Cooper, S., and Keasling, J. D. 1998, Cycle-specific replication of chromosomal and F plasmid origins. *FEMS Microbiol. Lett.*, **163**, 217–22.
5. Keasling, J. D., Palsson, B. Ø., and Cooper, S. 1991, Cell-cycle-specific F plasmid replication: Regulation by cell size control of initiation. *J. Bacteriol.*, **173**, 2673–80.
6. Fisher, J. K., Bourniquel, A., Witz, G., Weiner, B., Prentiss, M., and Kleckner, N. 2013, Four-dimensional imaging of *E. coli* nucleoid organization and dynamics in living cells. *Cell*, **153**, 882–95.
7. Oliveira, S. M. D., Neeli-Venkata, R., Goncalves, N. S. M., et al. 2016, Increased cytoplasm viscosity hampers aggregate polar segregation in *Escherichia coli*. *Mol. Microbiol.*, **99**, 686–99.
8. Häkkinen, A., Muthukrishnan, A.-B., Mora, A., Fonseca, J. M., and Ribeiro, A. S. 2013, CellAging: A tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*. *Bioinformatics*, **29**, 1708–9.
9. Mora, A. D., Vieira, P. M., Manivannan, A., and Fonseca, J. M. 2011, Automated drusen detection in retinal images using analytical modelling algorithms. *Biomed. Eng. Online*, **10**, 59.

# PUBLICATION II

**Effects of  $\sigma$  factor competition are promoter initiation kinetics dependent.**

Vinodh K. Kandavalli, Huy Tran, Andre S. Ribeiro

Biochimica et Biophysica Acta. Gene Regulatory Mechanisms. 1859, 1281–1288, 2016  
<http://dx.doi.org/10.1016/j.bbagr.2016.07.011>

**Publication reprinted with the permission of the copyright holders.**





Contents lists available at ScienceDirect

## Biochimica et Biophysica Acta

journal homepage: [www.elsevier.com/locate/bbagrm](http://www.elsevier.com/locate/bbagrm)Effects of  $\sigma$  factor competition are promoter initiation kinetics dependent

Vinodh K. Kandavalli, Huy Tran, Andre S. Ribeiro\*

Laboratory of Biosystem Dynamics, Department of Signal Processing, Tampere University of Technology, 33101 Tampere, Finland

## ARTICLE INFO

## Article history:

Received 3 May 2016

Received in revised form 5 July 2016

Accepted 7 July 2016

Available online 21 July 2016

## Keywords:

 $\sigma$  Factors competition*In vivo* Transcription dynamics

Single RNA detection

Open Complex Formation

Closed Complex Formation

## ABSTRACT

In *Escherichia coli*, the expression of a  $\sigma$  factor is expected to indirectly down-regulate the expression of genes recognized by another  $\sigma$  factor, due to  $\sigma$  factor competition for a limited pool of RNA polymerase core enzymes. Evidence suggests that the sensitivity of genes to indirect down-regulation differs widely. We studied the variability in this sensitivity in promoters primarily recognized by RNAP holoenzymes carrying  $\sigma^{70}$ . From qPCR and live single-cell, single-RNA measurements of the transcription kinetics of several  $\sigma^{70}$ -dependent promoters in various conditions and from the analysis of  $\sigma$  factors population-dependent models of transcription initiation, we find that, the smaller is the time-scale of the closed complex formation relative to the open complex formation, the weaker is a promoter's responsiveness to changes in  $\sigma^{38}$  numbers. We conclude that, in *E. coli*, a promoter's responsiveness to indirect regulation by  $\sigma$  factor competition is determined by the sequence-dependent kinetics of the rate limiting steps of transcription initiation.

© 2016 Elsevier B.V. All rights reserved.

## 1. Introduction

In *Escherichia coli*, metabolic changes are associated with changes in the numbers of the components of the transcription machinery, such as RNA polymerase (RNAP) core enzymes and  $\sigma$  factors [1]. For example, the regulation of  $\sigma$  factor numbers allows *E. coli* to implement genome-wide changes in expression rates [2–6] and consequently alter, among other, the cell growth phase. E.g. during stationary growth, some genes have much reduced activity compared to exponential growth, while others exhibit little to no differences or enhanced expression [2,3,7,8]. This diversity in responses is, to an extent, made possible by differences in the promoters' selectivity for  $\sigma$  factors [3,9,10] and/or are due to the action of transcription factors [3].

Another means by which changes in  $\sigma$  factor numbers cause genome-wide changes in expression rates is indirect regulation [3], which affects the rate of transcription initiation [11], where most gene expression regulation occurs [12–15]. E.g., in the stationary growth phase, *rpoS* expression is enhanced [10], while during exponential growth it is silenced [1,16]. Since RpoS ( $\sigma^{38}$ ) competes with the house-keeping RpoD ( $\sigma^{70}$ ) for a limited pool of RNAP core enzymes [1, 16,17], increasing its numbers decreases the fraction of RNAP holoenzymes carrying  $\sigma^{70}$  [3,18]. Consequently, the transcription rate of genes expressed by RNAP. $\sigma^{70}$  holoenzymes is expected to decrease. However, for unknown reasons, the degree by which the transcriptional

activity changes by indirect regulation varies widely between genes, from high degree of change to almost no change [2,3,7,8]. Previous studies of *in silico* models suggest that this diversity may be related to the strength of the promoter [17], or to the degree of RNAP saturation [19].

Transcription initiation is the stage of transcription where most regulation occurs [20]. Both *in vitro* as well as more recent *in vivo* measurements (e.g. by MS2-GFP tagging of RNA [21]), have informed that this is a sequential process that includes two major rate-limiting steps [22–27]. The first is the 'closed complex formation', which consists of trials of binding of a free-floating RNAP holoenzyme (carrying the appropriate  $\sigma$  factor) to the promoter region, until a stable complex is formed [20,28]. It follows the 'open complex formation' [22,23,29] (which usually is nearly irreversible [30]) that starts with the DNA unwinding [23,25] and ends when the RNAP escapes the promoter and elongation begins, which is expected to release the  $\sigma$  factor [31–33]. In the end of elongation, a fully formed RNA and an RNAP core enzyme are released [34,35].

The expected time-scale of the closed complex formation (i.e. the time taken by this process), is partially determined by the time it takes a free promoter to be bound by a free floating RNAP holoenzyme carrying the appropriate  $\sigma$  factor [23]. The frequency of occurrence of this event depends on the total numbers and fraction of those specific RNAP holoenzymes, which, due to limited numbers of RNAP core enzymes, should depend on the numbers of all types of  $\sigma$  factors in the cell. Meanwhile, the open complex formation should not depend on  $\sigma$  factor numbers since, at this stage of transcription, the appropriate RNAP holoenzyme necessary for it to occur is already 'in place'. Given this, we hypothesized that the ratio between the time scales of the

\* Corresponding author at: Office TC336, Department of Signal Processing, Tampere University of Technology, P.O Box 553, 33101 Tampere, Finland.  
E-mail address: [andre.ribeiro@tut.fi](mailto:andre.ribeiro@tut.fi) (A.S. Ribeiro).

two major rate-limiting steps in transcription should affect the sensitivity of a promoter to changes in  $\sigma$  factor numbers other than the  $\sigma$  factor by which it is preferentially transcribed by.

Here, we test this hypothesis. First, we analyze the dynamics of a stochastic model of transcription as a function of  $\sigma$  factors numbers in the cell and of the kinetics of the rate-limiting steps in transcription initiation of the promoter. From the results, we formulate a hypothesis of how the kinetics of the rate limiting steps in transcription initiation affects a promoter's response to changes in  $\sigma$  factors numbers. Next, we perform qPCR and live single-cell, single-RNA measurements of the transcription kinetics of a set of promoters primarily recognized by RNAP holoenzymes carrying  $\sigma^{70}$  in various conditions. These measurements and the comparison with the model predictions provide strong evidence that the kinetics of the sequence-dependent, multi-step transcription initiation process in promoters in *E. coli* is the key determining factor of the promoters' sensitivity to indirect regulation by changes in  $\sigma$  factors numbers, due to  $\sigma$  factor competition for limited RNAP core enzymes.

## 2. Materials and methods

### 2.1. Strains and plasmids

*E. coli* strains used are BW25113 [36] and its deletion mutant, JW5437–1 [37], lacking the *rpoS* gene [37], obtained from the Keio single-gene knockout collection. We denote BW25113 as *rpoS*<sup>+</sup> and JW5437–1 as *rpoS*<sup>−</sup>. These strains lack the *araB-araD*, *lacZ* genes, rendering inactive the negative feedback loop of the arabinose and lactose utilization system [38,39].

We inserted single copy target plasmids containing one of promoters of interest ( $P_{BAD}$ ,  $P_{tetA}$ ,  $P_{lac-O1}$ ,  $P_{lac-O1O3}$ , or  $P_{lac-ara-1}$ ) in the cells (Supplementary Table S2) so as to study their dynamics using qPCR.

In addition, to study  $P_{BAD}$  and  $P_{tetA}$  dynamics using MS2-GFP binding of RNA techniques [21,27,40], aside from the target plasmid, we inserted a low-copy reporter plasmid (pZS12MS2-GFP) carrying  $P_{lac-ms2-gfp}$  (generously provided by Phillips Cluzel, Harvard University, MA, USA) which produces the MS2-GFP proteins that bind the target RNA [41].

In the construct using  $P_{lac-O1O3}$  (Supplementary material), the native O2 operator site is missing as, in the natural construct, it is located in the native *lacZ* gene [42]. Also, the mCherry-48bs sequence is from [43], while the original promoter,  $P_{T7\phi 10}$ , was replaced by  $P_{lac-O1O3}$ . See Supplementary Section 2.2 for procedures.

In the construct using  $P_{lac-O1}$  (a kind gift from Ido Golding, Baylor College of Medicine, USA), the O3 and O2 operator sites are missing. Thus, it has only 1 operator site (O1). Finally, note that  $P_{lac-O1}$  and  $P_{lac-O1O3}$  have the same sequence in the −50 to +0 regions from the transcription start site (Supplementary Section 1.1).

For a detailed description of the single RNA detection method, estimation of time intervals between consecutive transcription events in individual cells, and microscopy techniques, see Supplementary Section 2.6. Meanwhile, qPCR measurement techniques are described in Supplementary Section 2.3.

### 2.2. Growth phase induction

The method used here to reach specific growth phases was proposed in [44]. Cells were grown in LB media (10 g/l tryptone, 5 g/l yeast extract and 10 g/l NaCl) overnight in an orbital shaker at 30 °C with aeration at 250 rpm. Afterwards, to induce exponential growth phase, cells were diluted in fresh LB media to reach an optical density ( $OD_{600}$ ) of ~0.05 and then grown at 37 °C with aeration at 250 rpm for 1 h. Meanwhile, to induce the stationary growth phase, cells were diluted in a stationary phase inducing media, i.e., they were placed on a media obtained by centrifuging the overnight cultured cells at 10,000 rpm for 10 min. Next, cells were allowed to growth at 37 °C with aeration at 250 rpm for 1 h. As shown by the growth curve analysis, this halts cell growth

(Supplementary Fig. S3). Finally, to assess whether the addition of arabinose alters cell growth rates, we measured the  $OD_{600}$  over time using a spectrophotometer. From the  $OD_{600}$  curves, we verified that the growth phases (exponential and stationary) of neither the wild-type nor deletion mutant strain were altered significantly for at least 2 h after adding arabinose to the media.

### 2.3. Intracellular RNA polymerase concentrations

Intracellular RNAP concentrations in *E. coli*, which include core enzymes and holoenzymes, can be made to differ by changing media composition in a specific manner within certain ranges [45]. Here, our aim was to obtain a set of media conditions where differences between intracellular RNAP concentrations are maximized between conditions, while differences in growth rates are minimized, so as to minimize differences in other cellular components. This was achieved in [46], using specially modified LB media conditions that differ in tryptone and yeast extract concentrations. In particular, starting from standard LB media (here denoted 1 × media), one can produce modified LB media with lower tryptone and yeast extract concentrations, which, within certain ranges, result in *E. coli* cells with gradually reduced intracellular RNAP concentrations [46]. The media used are denoted as 0.25 ×, 0.5 ×, and 1 ×, and their composition per 100 ml is, respectively: (0.25 ×) 0.25 g tryptone, 0.125 g yeast extract and 1 g NaCl (pH = 7.0); (0.5 ×) 0.5 g tryptone, 0.25 g yeast extract and 1 g NaCl (pH = 7.0); and (1 ×) 1 g tryptone, 0.5 g yeast extract and 1 g NaCl (pH = 7.0). The resulting RNAP concentrations were assessed by measuring the level of the RpoC protein, a core subunit of RNAP [47], by Western blot (Supplement). These measurements confirmed that the mean intracellular RNAP levels decrease with decreasing tryptone and yeast extract concentrations (Supplementary Fig. S1 and Supplementary Table S1) as first reported in [46]. Meanwhile, Supplementary Fig. S2 shows that cell growth rates are only mildly affected.

## 3. Results

### 3.1. Expected effects of changes in $\sigma$ factor numbers on the transcript production dynamics

We assume the model of transcription proposed in [48], but additionally account for competition between  $\sigma$  factors for binding to RNAP core enzymes, as these exist in limited numbers in *E. coli* [3, 17–19]. We study promoters primarily transcribed by  $\sigma^{70}$ , thus, the binding of  $\sigma^{70}$  to an RNAP core enzyme is modeled explicitly (reaction (A)). Other  $\sigma$  factors are referred to as  $\sigma^i$  and their interactions with RNAPs are modeled by a single reaction, (B), for simplicity, as they influence the model solely by limiting the number of RNAP core enzymes available to  $\sigma^{70}$ :



Reactions (A) and (B) describe the binding/unbinding of  $\sigma^{70}$  and other  $\sigma$  factors ( $\sigma^i$ ) to RNAP core enzymes (RNAP), respectively. These reactions therefore allow the formation of corresponding RNAP holoenzymes ( $RNAP.\sigma^{70}$  and  $RNAP.\sigma^i$ , respectively). The ratios between association and dissociation rate constants are  $K_{70}$  and  $K_i$ , respectively.

From (A) and (B), one can estimate an approximate expected number of  $RNAP.\sigma^{70}$ . Let the numbers of  $\sigma$  factors (either free floating or in a holoenzyme form) be  $[\sigma^{70}]$  and  $[\sigma^i]$  for  $\sigma^{70}$  and other factors, respectively, while the number of free floating RNAP core enzymes (i.e. not bound to a  $\sigma$  factor) is  $[RNAP]$ . In wild-type *E. coli*, due to a limited pool of core enzymes (i.e. there are more  $\sigma$  units than RNAP units), most RNAPs are expected to be in the holoenzyme form [17,18,49], i.e., in the form of

$RNAP.\sigma^{70}$  or  $RNAP.\sigma^i$ . Their numbers are limited only by the total amount of RNAP. Consequently, in equilibrium, the expected number of  $RNAP.\sigma^{70}$  is, in approximation, given by:

$$[RNAP.\sigma^{70}] \sim [RNAP] \frac{[\sigma^{70}]K_{70}}{[\sigma^{70}]K_{70} + [\sigma^i]K_i} \quad (C)$$

The model also includes explicitly a multi-step transcription process [48], following the empirical models proposed in [20,22,26]:



Reaction (D) models the closed complex formation, i.e. the finding and effective binding of RNAP holoenzymes with  $\sigma^{70}$  to the promoter ( $Pr$ ) region. Based on the assumption that  $Pr$  is preferentially transcribed by  $RNAP.\sigma^{70}$  holoenzymes, only  $\sigma^{70}$  is assumed to bind stably to  $Pr$ , i.e., other holoenzymes are assumed to either bind very rarely to the promoter region and/or to not remain bound for a significant amount of time.

Note that  $k_{cc}$  stands for the inverse of the mean time that  $Pr_{cc}$  stays in equilibrium with  $Pr$  and  $RNAP.\sigma^{70}$ , until it begins to form a stable open complex (i.e. the model does not represent explicitly the known instability of  $Pr_{cc}$ ) (see e.g. [23,26]). As such, reaction (D) should not be interpreted as elementary. Rather, its rate represents the inverse of the time-lapse until a stable open complex forms. This time-lapse depends on several steps in the process of closed complex formation, such as binding and unbinding of the polymerases to the DNA (i.e. reversibility), 1D diffusive searches, etc., along with the rate of commitment to the open complex formation [50,51].

The level of detail of our model of transcription is based on the level of detail that our measurements allow. In particular, we effectively measure time intervals between consecutive RNA productions along with the mean duration of the open complex formation (see below). From these two quantities, we obtain the difference between them, which is the time-lapse until reaching the stage of open complex formation. For this reason, reversibility is not modelled explicitly. Relevantly, there are several conditions under which the absence of reversibility is a valid approximation (e.g. when  $k_{oc}$  is much higher than the rate at which the close complex reverts to the previous state). In those cases, the difference between the time-lapses of open complex formation and intervals between consecutive RNA productions is effectively the time-lapse of the closed complex formation. For a detailed analysis of this and other such conditions, see e.g., [46].

Meanwhile, reaction (E) models the open complex formation, which is expected to be nearly irreversible [20], and that, once complete, is followed by RNAP escape from the promoter region [31], leading to elongation.

Finally, all steps following the open complex formation, including promoter escape [31], release of  $\sigma^{70}$  [33], elongation, termination and RNA and RNAP core enzyme release [33,35,52,53] are modeled by (F). Their time-scale is not accounted for since, in normal conditions, they are expected to be much faster than the closed and open complex formations [31,32]. Namely, only in the rare promoters, whose open complexes exhibit extremely short half-lives, is the promoter escape expected to be rate-limiting [31]. In this sense, similarly to (D), step (F) should not be interpreted as an elementary transition, but rather as representing a complex process whose effective rate is the rate at which a nearly-formed stable open complex will fully commit to elongation, thus defining the promoter's strength for RNA production once at this stage of initiation. Finally, in this model, the effects of repressors are accounted for explicitly in the values of the kinetic rates  $k_{cc}$  and  $k_{oc}$ .

Let  $\tau_{cc}$  be the time-scale of the closed complex formation and  $\tau_{oc}$  be the time-scale of the open complex formation (includes all rate-limiting steps prior to commitment to elongation). According to reactions (D-F), assuming one and only one  $Pr$  promoter in a cell, the mean time interval between consecutive RNA productions ( $\Delta t$ ) under the control of  $Pr$  is (Supplementary Section 2.5):

$$\Delta t = \tau_{cc} + \tau_{oc} = \frac{1}{k_{cc}[RNAP.\sigma^{70}]} + \frac{1}{k_{oc}} \quad (G)$$

From (C-G) it is deducible that increasing the numbers of a  $\sigma$  factor other than  $\sigma^{70}$  will increase  $\tau_{cc}$  and, consequently, the mean  $\Delta t$  of a promoter preferentially transcribed by  $RNAP.\sigma^{70}$ . The magnitude of this increase will depend on various rate constants (some unknown, e.g.  $k_{cc}$ ), and, thus, it can only be determined empirically.

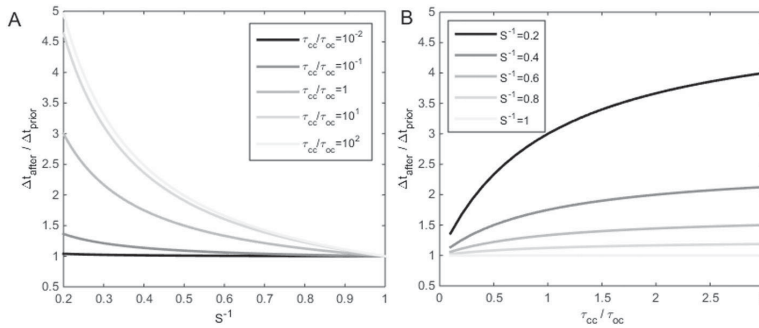
From (G), the model assumes that the mean  $\Delta t$  of a promoter preferentially transcribed by  $RNAP.\sigma^{70}$  holoenzymes is proportional to the inverse of the total  $RNAP.\sigma^{70}$  concentration (i.e. the RNA production rate changes hyperbolically with the RNAP concentration). This assumption was proposed in [46]. First, note that, in support of this assumption, even though most (approximately 85%) RNAP molecules are transiently bound to the DNA [54], this is expected to result solely in a reduced effective diffusion coefficient that reduces the upper bound of the binding rate constant of RNAPs to promoters [55–57] but not the amount of RNAP molecules capable of promoter binding. Also, in [46] it was shown that the mean  $\Delta t$  of a promoter preferentially transcribed by  $RNAP.\sigma^{70}$  (namely,  $P_{lac-ara-1}$ ) does in fact change linearly with the inverse of the total  $RNAP.\sigma^{70}$  concentration, within a certain range of RNAP concentrations (altered by altering media conditions). As such, we make use of the same media conditions to perform our measurements (see below). Further, in our measurements where media conditions are altered to change RNAP numbers, cells are in the exponential growth phase (Fig. S2), so as to not alter the means by which they regulate the functional  $RNAP.\sigma^{70}$  numbers (which could differ, e.g., in the stationary phase). Finally, we expect that, if there are biases in the stationary phase due to, e.g. sequestration of  $RNAP.\sigma^{70}$  by 6S RNA [58], they should be similar in all promoters, as they are all preferentially transcribed by  $RNAP.\sigma^{70}$ .

Finally, from the model, due to a limited pool of core enzymes when compared to the number of  $\sigma$  factors, for promoters whose transcription is primarily initiated by  $\sigma^{70}$ , one can predict, for a given ratio between  $\tau_{cc}/\tau_{oc}$  (which one can obtain empirically [46]), the expected ratio between the mean intervals between consecutive transcription intervals after ( $\Delta t_{after}$ ) and prior ( $\Delta t_{prior}$ ) a change in the amount of  $RNAP.\sigma^{70}$  in the cells. Namely, given that a change in  $RNAP.\sigma^{70}$  numbers should affect only the time-scale of the closed complex formation, defining  $S$  as the ratio between the number of  $RNAP.\sigma^{70}$  prior and after the change ( $S = [RNAP.\sigma^{70}]_{prior}/[RNAP.\sigma^{70}]_{after}$ ), given (G) one can write:

$$\frac{\Delta t_{after}}{\Delta t_{prior}} = \frac{S \times \tau_{cc} + \tau_{oc}}{\tau_{cc} + \tau_{oc}} = \frac{S \times (\tau_{cc}/\tau_{oc}) + 1}{(\tau_{cc}/\tau_{oc}) + 1} \quad (H)$$

From (H), the higher is  $\tau_{cc}/\tau_{oc}$  of a promoter, the more responsive will be that promoter dynamics' to changes in  $\sigma$  factor numbers. E.g., assume two promoters, X and Y, with identical mean  $\Delta t$  of 1000 s, but for X  $\tau_{cc} = 800$  s and  $\tau_{oc} = 200$  s while for Y  $\tau_{cc} = 200$  s and  $\tau_{oc} = 800$  s. Assuming that a change in  $\sigma$  factors numbers causes  $\tau_{cc}$  to be reduced to half, one finds that the mean  $\Delta t$  of X will equal 600 s, while the mean  $\Delta t$  of Y will equal 900 s, which indicates that the second promoter is less responsive.

Fig. 1 (left) shows the predicted results from equation (H) for a wide range of values of  $\tau_{cc}/\tau_{oc}$  (from  $10^{-2}$  to  $10^2$  in accordance with measurements reported in previous studies [20,22,23,25–27,40,43,46,60]), and assuming that  $S^{-1}$  varies between 0 and 1 (i.e. from none to the same number of  $RNAP.\sigma^{70}$  as in 'control' cells). Visibly, the model predicts a wide range of behavioral responses, gradually changing from highly



**Fig. 1.** Predictions from the model of the ratio between mean transcription intervals prior and after a change in the numbers of RNAP. $\sigma^{70}$  in the cells as a function of the inverse of the ratio between the number of RNAP. $\sigma^{70}$  prior and after that change. Predictions from the model assuming promoters with (A) different values of  $S^{-1}$  for five different values of  $\tau_{cc}/\tau_{oc}$  and (B) different values of  $\tau_{cc}/\tau_{oc}$  for five different values of  $S^{-1}$ .

sensitive to almost insensitive to changes in  $\sigma$  factor numbers other than  $\sigma^{70}$ . Meanwhile, Fig. 1 (right) shows the predicted results from equation (H) for a wide range of values of  $S^{-1}$  (from 0.2 times to the same number of RNAP. $\sigma^{70}$  as in ‘control’ cells), and assuming that  $\tau_{cc}/\tau_{oc}$  can vary between 0 and 3 (within the range of empirical values obtained in the measurements shown below).

To test these predictions, one needs to measure RNA production kinetics in conditions differing in  $S^{-1}$  and in the ratio  $\tau_{cc}/\tau_{oc}$ . The first is possible by, e.g., comparing RNA production kinetics in cells in the exponential and stationary growth phases, as shown below. The second is possible by measuring this kinetics in, e.g., different promoters (differing in  $\tau_{cc}/\tau_{oc}$  as shown below) and in promoters subject to different induction strengths (which can affect  $\tau_{cc}/\tau_{oc}$  as shown below).

### 3.2. $\sigma$ factor numbers in the stationary and exponential growth phase

From previous studies [1–3,10,16,18], we expect  $\sigma^{38}$  numbers to differ in cells in the exponential and stationary growth phases, but little differences are expected in the numbers of  $\sigma^{70}$  or RpoC (a subunit of RNAP) [1–3,16,18]. To test this, we make use of strain BW25113 (referred to as *rpoS*<sup>+</sup>) and its deletion mutant, JW5437–1 (referred to as *rpoS*<sup>-</sup>), that lacks the gene encoding for the RpoS protein [37].

First, for both strains, after inducing the specific growth phase (Methods and Supplementary Fig. S3) we performed Western blot to measure RpoS, RpoD, and RpoC protein levels in the exponential and stationary growth phases (Methods). Fig. 2 shows that, in *rpoS*<sup>+</sup> cells, while the RpoS levels are much higher in the stationary phase, the levels of RpoC and RpoD do not differ significantly between growth phases, as previously reported [1,10,16]. Aside from RpoS, other  $\sigma$  factors existing in some abundance in *E. coli* (specifically, RpoN, and RpoF) are not expected to differ in numbers between these growth phases [1,36,37,59]. As such, we conclude that the ratio between RpoS and RpoD numbers differ significantly between growth phases. Meanwhile, in *rpoS*<sup>-</sup> cells there are only relatively small differences in the numbers of any of the  $\sigma$  factors.

Given this, one can, by comparing the kinetics of transcription of a promoter in cells in the exponential and stationary growth phases, test the model predictions on the effects of changing  $S^{-1}$  on the ratio  $\Delta t_{after}/\Delta t_{prior}$ , which here after is referred to as  $\Delta t_{stat}/\Delta t_{exp}$ .

### 3.3. Validation of the model by comparing various promoters’ kinetics in different growth phases

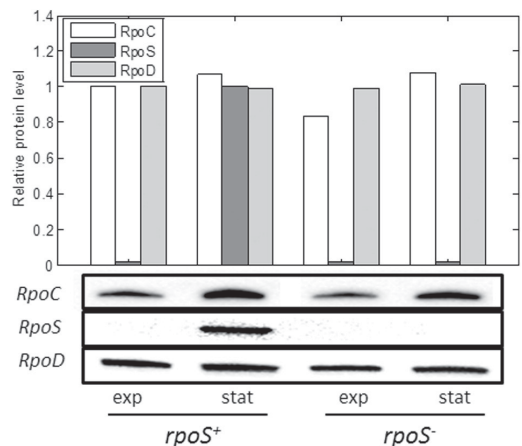
To test the model predictions on how  $\Delta t$  differs between the exponential and stationary growth phases ( $\Delta t_{stat}/\Delta t_{exp}$ ) as a function of the value of  $\tau_{cc}/\tau_{oc}$  when in the exponential growth phase, we selected a set of promoters preferentially transcribed by  $\sigma^{70}$ , namely,  $P_{BAD}$ ,  $P_{tetA}$ ,  $P_{lac-O1}$ ,  $P_{lac-O1O3}$ , and  $P_{lac-ara-1}$  (sequences in Supplement Section 1.1

and induction curves in Supplement Fig. S6) and measured by qPCR their  $\tau_{cc}/\tau_{oc}$  in cells in the exponential growth phase (Methods and Supplementary Fig. S4). In addition, for  $P_{lac-O1O3}$  and  $P_{BAD}$ , we also measured  $\tau_{cc}/\tau_{oc}$  when subject to different induction strengths.

To obtain empirical values of  $\tau_{cc}/\tau_{oc}$ , we used a slightly modified version of the methodology proposed in [45,46], which is an extension of a technique developed for *in vitro* studies [60]. Shortly, by establishing different RNAP concentrations (and, thus, different  $[RNAP \cdot \sigma^{70}]$ ) in live cells in the exponential growth phase (Methods, Supplementary Fig. S1, and Supplementary Table S1), which affects the time-scale of the closed complex formation but not of the open complex [45, 46], one can, from measurements of the transcription rate of a promoter in each such condition, obtain a ‘relative  $\tau$  plot’ (see Supplementary Section 2.8 for detailed explanation), from which  $\tau_{cc}/\tau_{oc}$  can be estimated.

Meanwhile, we cannot measure  $[RNAP \cdot \sigma^{70}]$  directly. However, this quantity approximately equals the total number of RNAP holoenzymes in the cells ( $[RNAP]$ ) in cells in the exponential growth phase, due to the absence of other competing  $\sigma$  factors, particularly  $\sigma^{38}$  [17,18,61]. Such  $[RNAP]$  amounts can be measured by Western Blot (Fig. 2).

We obtained nine ‘relative  $\tau$  plots’ (Supplementary Fig. S4A–I), from which we extracted  $\tau_{cc}/\tau_{oc}$  (Table 1). Visibly, all promoters under full



**Fig. 2.** Quantification of RNAP subunits by Western Blot. Protein levels of rpoC, rpoS, and rpoD genes in BW25113 (*rpoS*<sup>+</sup>) and JW5437–1 (*rpoS*<sup>-</sup>) cells in the exponential (‘exp’) and stationary (‘stat’) growth phases, as measured by Western blot. Values for RpoC and RpoD genes are relative to those of the *rpoS*<sup>+</sup> strain in the exponential phase. Values for RpoS gene are relative to those of the *rpoS*<sup>-</sup> strain in the stationary phase.



induction differ in  $\tau_{cc}/\tau_{oc}$ . Also, the induction strength affects  $\tau_{cc}/\tau_{oc}$  in  $P_{BAD}$  and  $P_{lacO1O3}$  similarly, in that  $\tau_{cc}/\tau_{oc}$  decreases significantly as the induction strength is decreased (which shows that induction acts differently on the closed and open complex formations in these promoters, as expected). Given these results, we can make use of this set of promoters, under various induction regimes, to survey how the value of  $\tau_{cc}/\tau_{oc}$  affects  $\Delta t_{stat}/\Delta t_{exp}$ .

For this, for each condition, we measured by qPCR the RNA production rate in cells in the exponential and in the stationary growth phase and then calculated the fold change in the transcription rate between these growth phases ( $\mu_{stat}/\mu_{exp}$ , where  $\mu$  is the transcription rate as measured by qPCR). From the inverse of  $\mu_{stat}/\mu_{exp}$ , we obtained  $\Delta t_{stat}/\Delta t_{exp}$  (see Section 2.8 in Supplement).

Note that, even though cells will grow and thus dilute RNA numbers at different rates when in the exponential and stationary phases, these ratios can be compared between cells in different growth phases because, first, RNA degradation rates are not growth phase dependent [15], and, second, all qPCR values were normalized by the productions rates of an internal reference gene (16S RNA).

Results in Table 1 indicate that  $\Delta t_{stat}/\Delta t_{exp}$  differs from 1 in nearly all conditions, and that it differs significantly between conditions. Importantly, there is a strong positive correlation between  $\Delta t_{stat}/\Delta t_{exp}$  and  $\tau_{cc}/\tau_{oc}$  (Spearman's rank correlation coefficient of 0.9) that is statistically significant ( $p$ -value 0.002).

Finally, we fitted the model described by equation (H) using weighted least square fit to the empirical data (Fig. 3). Visibly, the model fits the data.

The best fit is for  $S^{-1}$  of  $0.25 \pm 0.02$ , which is close to expectations given the wide differences in transcript production kinetics between the exponential and stationary growth phases shown in Table 1 (in agreement with genome wide measurements in recent works [6]) and also given the significant, measured differences in  $\sigma^{38}$  numbers as well (Fig. 2). Nevertheless, note that such value of  $S^{-1}$  is not necessarily accounted for solely by changes in  $\sigma^{38}$  numbers. Other changes in, e.g., numbers of regulatory molecules influencing transcription [62–65] may also play a role.

### 3.4. Validation of the model by live, single-RNA detection measurements in $rpoS^+$ and $rpoS^-$ cells

Next, we selected the two promoters in Table 1 with more 'extreme' initiation kinetics:  $P_{BAD}$  under full induction (largest  $\tau_{cc}/\tau_{oc} \sim 2.40$ ) and  $P_{tetA}$  (smallest  $\tau_{cc}/\tau_{oc} \sim 0.08$ ). For these, we conducted microscopy measurements of RNA production dynamics at the single molecule level by multiple MS2-GFP tagging of individual RNA molecules [21,27] (Supplementary Material) in cells ( $rpoS^+$ ) in the exponential and stationary growth phase. According to the model and the qPCR measurements

**Table 1**  
Ratios between the transcription activity of various promoters under various induction schemes between cells in the exponential and stationary growth phases ( $\Delta t_{stat}/\Delta t_{exp}$ ) along with the ratio  $\tau_{cc}/\tau_{oc}$  measured from cells under exponential growth.

Promoter	Induction level	$\tau_{cc}/\tau_{oc}$ (90% CI)	$\Delta t_{stat}/\Delta t_{exp}$ (90% CI)
$P_{BAD}$	0.1% ara	2.40 (0.49–>10)	3.74 (3.12–4.49)
$P_{BAD}$	0.01% ara	1.20 (0.30–5.4)	2.54 (1.30–5.02)
$P_{BAD}$	0.001% ara	0.20 (0.06–0.35)	1.77 (1.29–2.40)
$P_{lac-O1O3}$	1 mM IPTG	1.20 (0.50–2.89)	2.30 (1.77–3.00)
$P_{lac-O1O3}$	0.05 mM IPTG	0.84 (0.26–2.26)	1.54 (0.89–2.70)
$P_{lac-O1O3}$	0.005 mM IPTG	0.13 (0.01–0.23)	1.39 (1.02–1.90)
$P_{tetA}$	–	0.08 (0.01–0.18)	1.09 (0.81–1.46)
$P_{lac-O1}$	1 mM IPTG	0.50 (0.31–0.75)	1.51 (1.20–1.91)
$P_{lac-ara1}$	0.1% ara + 1 mM IPTG	0.97 (0.28–3.07)	2.75 (1.72–4.45)

Shown in each line are the different promoters, the induction level, the value of  $\tau_{cc}/\tau_{oc}$  as measured by qPCR in cells in the exponential growth phase, and, finally, the ratio (fold change) between the mean duration of the intervals between transcription events in cells in the stationary and in the exponential growth phases as estimated from qPCR measurements (Methods). Also shown are their 90% confidence intervals (CI), calculated by assuming the threshold cycle measured by qPCR following a Gaussian distribution.

(Table 1), we should observe significant differences in the dynamics between the two growth phases in  $P_{BAD}$  ( $\Delta t_{stat}/\Delta t_{exp} = 3.74$ ) but not in  $P_{tetA}$  ( $\Delta t_{stat}/\Delta t_{exp} = 1.09$ ). In addition, we performed the same measurements in cells lacking the  $rpoS$  gene ( $rpoS^-$ ), for which the model predicts little changes with growth phase in either promoter kinetics due to the lack of differences in  $\sigma^{38}$  factor numbers.

For each condition, we performed time-lapse microscopy measurements (2 h long, 30 s intervals between consecutive images). The photo-toxicity caused by confocal microscopy was assessed and found to be negligible (Supplementary Material). From the images, we extracted RNA productions in individual cells (Supplementary Material) as in [66]. Distributions of the durations of these intervals are shown in Fig. 4. From each distribution, we calculated the intervals mean duration ( $\Delta t$ ). Results (Table 2) are in agreement with previous studies [27,40].

From Table 2, in  $rpoS^+$  cells, the RNA production dynamics of  $P_{BAD}$  differs widely in the exponential and stationary growth phases ( $p$ -value  $< 10^{-5}$ ), as predicted by the model, given the measured high value of  $\tau_{cc}/\tau_{oc}$  (Table 1). Also, while in  $rpoS^+$  cells the intervals between consecutive RNA productions are  $\sim 2.28$  times longer in mean duration in the exponential phase, in  $rpoS^-$  cells this difference is much reduced ( $\sim 1.55$  fold), as predicted by the model, since the numbers of  $\sigma^{38}$  should differ much less between growth phases in  $rpoS^-$  cells (Fig. 2).

Meanwhile, the dynamics of  $P_{tetA}$  does not exhibit statistically significant differences between conditions in  $rpoS^+$  cells, as predicted by the model given small value of  $\tau_{cc}/\tau_{oc}$  of this promoter (Table 1), nor in  $rpoS^-$  cells, as expected.

Finally, we performed qPCR measurements of  $P_{BAD}$  and  $P_{tetA}$  activity in the two growth conditions in  $rpoS^-$  cells and compared to  $rpoS^+$  cells. The differences are in qualitative agreement with those found by microscopy (Table 2 and Supplementary Fig. S5).

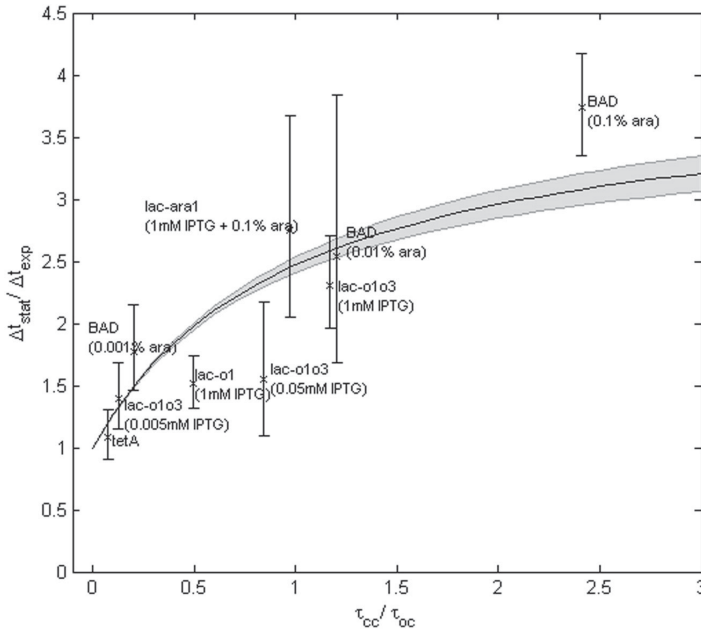
## 4. Discussion and conclusions

We investigated the selectivity and sensitivity mechanisms to indirect regulation by global changes in  $\sigma$  factor numbers of promoters primarily transcribed by RNAP- $\sigma^{70}$  holoenzymes. Analysis of our model of transcription suggests that, given that changes in  $\sigma$  factors numbers affect the closed complex formation but not the open complex formation, the degree of sensitivity of a promoter to indirect regulation due to changes in  $\sigma$  factors numbers is determined by the ratio between the time-scales of closed and open complex formation. In particular, this sensitivity increases for increasing ratio between the time-scales of closed and open complex formation.

To validate this model-based prediction, from qPCR measurements, we compared the kinetics of the rate-limiting steps in transcription initiation of several promoters primarily transcribed by RNAP- $\sigma^{70}$  in  $rpoS^+$  and  $rpoS^-$  cells when in the exponential and stationary growth phase, and for various induction schemes. The measured differences in transcription kinetics between growth phases in the various cases support the model predictions, given the measured differences in  $\sigma^{38}$  numbers in cells in the two growth phases and the measured differences in  $\tau_{cc}/\tau_{oc}$  between promoters and as a function of induction strength.

To provide additional validation, live single-RNA time-lapse microscopy measurements of time intervals between consecutive RNA productions in individual cells from two promoters differing widely in  $\tau_{cc}/\tau_{oc}$ , were conducted in  $rpoS^+$  cells and  $rpoS^-$  cells. Only for  $rpoS^+$  cells did we find differences in the transcription dynamics at difference growth phases (as expected from the model), and only for the promoter with higher  $\tau_{cc}/\tau_{oc}$  (as expected given the small value of  $\tau_{cc}/\tau_{oc}$  for the other promoter).

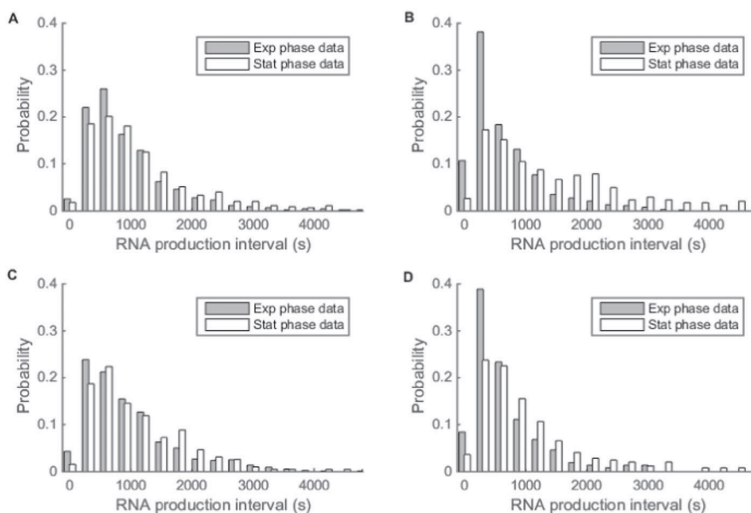
From these measurements, we conclude that a promoter's sensitivity to changes in other  $\sigma$  factor numbers is both promoter-specific (due to the promoter sequence dependence of  $\tau_{cc}$  and  $\tau_{oc}$ ) as well as induction strength dependent, as both these factors affect  $\tau_{cc}$  and  $\tau_{oc}$ .



**Fig. 3.** Ratio between mean transcription rates in cells in the stationary and in the exponential growth phase ( $\Delta t_{\text{stat}}^+ / \Delta t_{\text{exp}}^+$ ) for varying  $\tau_{\text{cc}} / \tau_{\text{oc}}$ , as inferred from the empirical data obtained by qPCR. The best-fit model is shown by the black solid line while standard errors are shown by the gray area. Also shown are the mean (crosses) and standard errors (vertical bars) of the empirical data to which the model was fitted to.

Our results imply that promoters' sensitivity to indirect regulation by changes in  $\sigma$  factor numbers is both evolvable (given the sequence dependence of  $\tau_{\text{cc}} / \tau_{\text{oc}}$ ) as well as adaptable to stress conditions (due to  $\tau_{\text{cc}} / \tau_{\text{oc}}$  being subject to regulation by, e.g., transcription factors). In this regard, we expect the sensitivity of essential genes to be weaker (i.e. we predict low  $\tau_{\text{cc}} / \tau_{\text{oc}}$  in these genes) as their expression products are essential for survival even under stress. Meanwhile, stress-response genes are expected to

have high sensitivity, so as to be responsive to alterations in  $\sigma$  factor numbers upon the emergence of stress conditions. Future studies are needed to test these hypotheses. In addition, we expect the ability to tune this sensitivity at the single gene level to be of importance as a means of regulating specific functionalities of *E. coli*. Namely, it should be possible to use this system to tune also the sensitivity of genetic motifs to changes in  $\sigma$  factor numbers, by tuning the sensitivity of its component genes.



**Fig. 4.** Distributions of intervals between consecutive RNA productions in individual cells detected by multiple MS2-GFP tagging of individual RNA molecules from (A)  $P_{\text{tetA}}$  in  $rpoS^+$  cells, (B)  $P_{\text{BAD}}$  in  $rpoS^-$  cells, (C)  $P_{\text{tetA}}$  in  $rpoS^+$  cells and (D)  $P_{\text{BAD}}$  in  $rpoS^-$  cells. Results from cells in the exponential ("Exp", dark grey bars) and stationary growth phase ("Stat", white bars).

**Table 2***In vivo* transcription dynamics of  $P_{BAD}$  and  $P_{tetA}$  in  $rpoS^+$  and  $rpoS^-$  cells in exponential and stationary growth phases.

Strain – Phase	No. samples	$\Delta t$ (s)	Ratio between mean transcription intervals ( $\Delta t_{stat}/\Delta t_{exp}$ ) (90% CI)	p-value ( $rpoS^+$ vs $rpoS^-$ )	p-value (Exp. vs Stat.)
$P_{BAD}$					
$rpoS^+$ – Exp	624	700		0.548	$<10^{-5}$
$rpoS^+$ – Stat	342	1595	2.28 (2.07–2.50)	$<10^{-5}$	
$rpoS^-$ – Exp	368	679			$<10^{-5}$
$rpoS^-$ – Stat	244	1053	1.55 (1.38–1.74)		
$P_{tetA}$					
$rpoS^+$ – Exp	435	982		0.188	0.014
$rpoS^+$ – Stat	447	1157	1.18 (1.08–1.28)	0.994	
$rpoS^-$ – Exp	160	996			0.142
$rpoS^-$ – Stat	192	1156	1.16 (1.01–1.34)		

Shown are the numbers of samples, the mean duration ( $\Delta t$ ) of the intervals between consecutive transcription events in individual cells, the ratio  $\Delta t_{stat}/\Delta t_{exp}$ , and the mean duration of the intervals in the stationary growth phase relative to mean duration in the exponential growth phase. Also shown are the p-values of Kolmogorov-Smirnov (KS) tests comparing the intervals' distributions between strains and between growth phases. In the statistical tests, for p values smaller than 0.01, the null hypothesis that the two sets of data are from the same distribution is rejected. Also shown are the 90% confidence intervals (CI) of the transcription intervals, calculated from the empirical data.

Our findings further suggest that previously suggested mechanisms, such as promoter strength [17], cannot alone explain the present results. E.g., our measurements show that two promoters of similar strength ( $P_{BAD}$  and  $P_{tetA}$  exhibit similar values of  $\Delta t$ ) can have very different degrees of change in transcription rate as a function of changes in  $\sigma^{38}$  numbers. Meanwhile, our results agree with a prediction that saturated promoters should only be weakly affected by  $\sigma$  factor competition [19], as this saturation ought to occur at the stage of closed complex formation and, thus the degree of response to changes in  $\sigma$  factor numbers of saturated promoters should be similar to that of promoters with relatively short-length closed complex formations.

As a side note, while assumed by the model for simplicity, we do not expect the differences in  $\sigma^{38}$  numbers between growth phases to be the sole cause for the observed differences in transcription dynamics. In the future, it should be of interest to study what other changes in other factor(s) (e.g., ppGpp [65], cAMP [63,64], 6S RNA [58,67] etc.) contribute to these differences.

We find plausible the existence of a similar mechanism in eukaryotic cells in the case of promoters containing a TATA-box. In these, for transcription to start, a TFIID factor must first bind to a TATA box. This factor, composed of a TBP (TATA binding protein) and of 1 out of at least 15 different TAFs (TBP associated factors), has been described as being a 'relative' of  $\sigma$  factors [68]. It has been proposed that the needed association between TAFs and TBP allows for coordinated regulation of transcription in eukaryotes, similar to  $\sigma$  factors [69], as distinct sets of TAFs likely dictate the type of promoter at which a given TFIID will function. By regulating the intracellular numbers of one TAF, it should be possible to indirectly regulate the speed of transcription initiation of a promoter not transcribed by that TAF, provided competition in the cell for TBP factors. Evidence for such competition exists, as TBP is not found in isolated form *in vivo* [70]. As in the case of *E. coli*, the effectiveness of this mechanism should be promoter-dependent in that it should depend on the ratio between the time-scales of the first and subsequent rate-limiting steps in initiation.

Finally, the present findings should be of use in studying *in vivo* transcription kinetics. For example, a recent work proposed a method for, from time-lapsed measurements of time intervals between RNA productions in live cells subject to different media, extract the number, time-scale, and order of occurrence of rate-limiting steps in transcription initiation [46]. We proposed here an alternative method to obtain this information, based on comparing transcription kinetics in cells with differing  $\sigma$  factors numbers, which can be used as a means of validation.

## Transparency document

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

## Acknowledgments

The authors thank Antti Häkkinen, Jarno Mäkelä, and Jerome Chandraseelan for valuable advices. This work was supported by Academy of Finland [257603 to ASR] and Centre of International Mobility (13.1.2014/TM-14-91361/CIMO, VK). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.bbaggm.2016.07.011>.

## References

- [1] M. Jishage, A. Iwata, S. Ueda, A. Ishihama, Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of four species of sigma subunit under various growth conditions, *J. Bacteriol.* 178 (1996) 5447–5451.
- [2] M. Rahman, M.R. Hasan, T. Oba, K. Shimizu, 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 (2006) 585–595, <http://dx.doi.org/10.1002/bit.20858>.
- [3] A. Farewell, K. Kvint, T. Nyström, Negative regulation by RpoS: A case of sigma factor competition, *Mol. Microbiol.* 29 (1998) 1039–1051, <http://dx.doi.org/10.1046/j.1365-2958.1998.00990.x>.
- [4] 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 (1995) 1032–1042.
- [5] T. Dong, H.E. Schellhorn, Control of RpoS in global gene expression of *Escherichia coli* in minimal media, *Mol. Genomics.* 281 (2009) 19–33, <http://dx.doi.org/10.1007/s00438-008-0389-3>.
- [6] 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 (2014) 4, <http://dx.doi.org/10.1186/1741-7007-12-4>.
- [7] T.H. Tani, A. Khodursky, R.M. Blumenthal, P.O. Brown, R.G. Matthews, Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis, *Proc. Natl. Acad. Sci. U. S. A.* 99 (2002) 13471–13476, <http://dx.doi.org/10.1073/pnas.212510999>.
- [8] D. Chang, D.J. Smalley, T. Conway, Gene expression profiling of *Escherichia coli* growth transitions: an expanded stringent response model, *Mol. Microbiol.* 45 (2002) 289–306.
- [9] R. Hengge-Aronis, Recent insights into the general stress response regulatory network in *Escherichia coli*, *J. Mol. Microbiol. Biotechnol.* 4 (2002) 341–346.
- [10] R.P. Lange, R. Hengge-Aronis, Identification of a central regulator of stationary-phase gene expression in *Escherichia coli*, *Mol. Microbiol.* 5 (1991) 49–59, <http://dx.doi.org/10.1111/j.1365-2958.1991.tb01825.x>.
- [11] T.M. Gruber, C.A. Gross, Multiple sigma subunits and the partitioning of bacterial transcription space, *Annu. Rev. Microbiol.* 57 (2003) 441–466, <http://dx.doi.org/10.1146/annurev.micro.57.030502.090913>.
- [12] Y. Taniguchi, P.J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emilii, X.S. Xie, Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells, *Science* 329 (2010) 533–538, <http://dx.doi.org/10.2142/biophys.51.136>.
- [13] O. Yarchuk, J. Guillerez, M. Dreyfus, Interdependence of Translation, Transcription and mRNA Degradation in the *ZacZ* Gene, *J. Mol. Biol.* 226 (1992) 581–596.

- [14] H. Lodish, A. Berk, S.L. Zipursky, P. Matsudaira, D. Baltimore, J. Darnell, *Molecular Cell Biology*, fourth ed. W.H. Freeman, New York, 2000.
- [15] H. Chen, K. Shiroguchi, H. Ge, X.S. Xie, Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*, *Mol. Syst. Biol.* 11 (2015) 781, <http://dx.doi.org/10.15252/msb.20145794>.
- [16] M. Jishage, A. Ishihama, Regulation of RNA Polymerase Sigma Subunit Synthesis in *Escherichia coli*: Intracellular Levels of  $\sigma^{70}$  and  $\sigma^{38}$ , *J. Bacteriol.* 177 (1995) 6832–6835.
- [17] I.L. Grigorova, N.J. Phleger, V.K. Mutalik, C.A. Gross, Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA, *Proc. Natl. Acad. Sci. U. S. A.* 103 (2006) 5332–5337, <http://dx.doi.org/10.1073/pnas.0600828103>.
- [18] H. Maeda, N. Fujita, A. Ishihama, Competition among seven *Escherichia coli* sigma subunits: relative binding affinities to the core RNA polymerase, *Nucleic Acids Res.* 28 (2000) 3497–3503, <http://dx.doi.org/10.1093/nar/28.18.3497>.
- [19] M. Mauri, S. Klumpp, A Model for Sigma Factor Competition in Bacterial Cells, *PLoS Comput. Biol.* 10 (2014) 1–16, <http://dx.doi.org/10.1371/journal.pcbi.1003845>.
- [20] W.R. McClure, Mechanism and control of transcription initiation in prokaryotes, *Annu. Rev. Biochem.* 54 (1985) 171–204, <http://dx.doi.org/10.1146/annurev.bi.54.070185.001131>.
- [21] I. Golding, J. Paulsson, S.M. Zawilski, E.C. Cox, Real-time kinetics of gene activity in individual bacteria, *Cell* 123 (2005) 1025–1036, <http://dx.doi.org/10.1016/j.cell.2005.09.031>.
- [22] W.R. McClure, Rate-limiting steps in RNA chain initiation, *Proc. Natl. Acad. Sci. U. S. A.* 77 (1980) 5634–5638, <http://dx.doi.org/10.1073/pnas.77.10.5634>.
- [23] R.M. Saecker, M.T. Record, P.L. Dehaseth, Mechanism of bacterial transcription initiation: RNA polymerase - Promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis, *J. Mol. Biol.* 412 (2011) 754–771, <http://dx.doi.org/10.1016/j.jmb.2011.01.018>.
- [24] D.F. Browning, S.J.W. Busby, The regulation of bacterial transcription initiation, *Nat. Rev. Microbiol.* 2 (2004) 57–65, <http://dx.doi.org/10.1038/nrmicro787>.
- [25] R. Lutz, T. Lozinski, T. Ellinger, H. Bujard, Dissecting the functional program of *Escherichia coli* promoters: the combined mode of action of lac repressor and AraC activator, *Nucleic Acids Res.* 29 (2001) 3873–3881, <http://dx.doi.org/10.1093/nar/29.18.3873>.
- [26] H. Buc, W.R. McClure, Kinetics of open complex formation between *Escherichia coli* RNA polymerase and the lac UV5 promoter. Evidence for a sequential mechanism involving three steps, *Biochemistry* 24 (1985) 2712–2723, <http://dx.doi.org/10.1021/bi00332a018>.
- [27] A.B. Muthukrishnan, M. Kandhavelu, J. Lloyd-Price, F. Kudasov, S. Chowdhury, O. Yli-Harja, A.S. Ribeiro, Dynamics of transcription driven by the tetA promoter, one event at a time, in live *Escherichia coli* cells, *Nucleic Acids Res.* 40 (2012) 8472–8483, <http://dx.doi.org/10.1093/nar/gks583>.
- [28] I.O. Nivedenskaya, H. Vahedian-Movahed, Y. Zhang, D.M. Taylor, R.H. Ebright, B.E. Nickels, Interactions between RNA polymerase and the core recognition element are a determinant of transcription start site selection, *Proc. Natl. Acad. Sci. U. S. A.* (2016) E2899–E2905, <http://dx.doi.org/10.1073/pnas.1603271113>.
- [29] P.L. DeHaseth, J.D. Helmann, Open complex formation by *Escherichia coli* RNA polymerase: The mechanism of polymerase-induced strand separation of double helical DNA, *Mol. Microbiol.* 16 (1995) 817–824, <http://dx.doi.org/10.1111/j.1365-2958.1995.tb02309.x>.
- [30] R. Schleif, AraC protein, regulation of the l-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action, *FEMS Microbiol. Rev.* 34 (2010) 779–796, <http://dx.doi.org/10.1111/j.1574-6976.2010.00226.x>.
- [31] L.M. Hsu, Promoter clearance and escape in prokaryotes, *Biochim. Biophys. Acta* 1577 (2002) 191–207.
- [32] K.M. Herbert, A. La Porta, B.J. Wong, R.A. Mooney, K.C. Neuman, R. Landick, S.M. Block, Sequence-resolved detection of pausing by single RNA polymerase molecules, *Cell* 125 (2006) 1083–1094, <http://dx.doi.org/10.1016/j.cell.2006.04.032>.
- [33] P.L. DeHaseth, T.M. Lohman, R.R. Burgess, M.T. Record, Nonspecific interactions of *Escherichia coli* RNA polymerase with native and denatured DNA: differences in the binding behavior of core and holoenzyme, *Biochemistry* 17 (1978) 1612–1622, <http://dx.doi.org/10.1021/bi00602a006>.
- [34] S.M. Uptain, C.M. Kane, M.J. Chamberlin, Basic mechanisms of transcript elongation and its regulation, *Annu. Rev. Biochem.* 66 (1997) 117–172, <http://dx.doi.org/10.1146/annurev.biochem.66.1.117>.
- [35] M. Raffaele, E.I. Kanin, J. Vogt, R.R. Burgess, A.Z. Ansari, Holoenzyme switching and stochastic release of sigma factors from RNA polymerase in vivo, *Mol. Cell* 20 (2005) 357–366, <http://dx.doi.org/10.1016/j.molcel.2005.10.011>.
- [36] K.A. Datsenko, B.L. Wanner, One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products, *Proc. Natl. Acad. Sci. U. S. A.* 97 (2000) 6640–6645.
- [37] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K.A. Datsenko, M. Tomita, B.L. Wanner, H. Mori, Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection, *Mol. Syst. Biol.* 2 (2006) 1–11, <http://dx.doi.org/10.1038/msb4100050>.
- [38] G. Fritz, J.A. Megerle, S.A. Westermayer, D. Brick, R. Heermann, K. Jung, J.O. Rädler, U. Gerland, Single cell kinetics of phenotypic switching in the Arabinose utilization system of *E. coli*, *PLoS One* 9 (2014).
- [39] E.M. Ozbudak, M. Thattai, H.N. Lim, B.I. Shraiman, A. Van Oudenaarden, Multistability in the lactose utilization network of *Escherichia coli*, *Nature* 427 (2004) 737–740, <http://dx.doi.org/10.1038/nature02298>.
- [40] J. Mäkelä, M. Kandhavelu, S.M.D. Oliveira, J.G. Chandraseelan, J. Lloyd-Price, J. Peltonen, O. Yli-Harja, A.S. Ribeiro, In vivo single-molecule kinetics of activation and subsequent activity of the arabinose promoter, *Nucleic Acids Res.* 41 (2013) 6544–6552, <http://dx.doi.org/10.1093/nar/gkt350>.
- [41] T.T. Le, S. Harlepp, C.C. Guet, K. Dittmar, T. Emonet, T. Pan, P. Cluzel, Real-time RNA profiling within a single bacterium, *Proc. Natl. Acad. Sci. U. S. A.* 102 (2005) 9160–9164, <http://dx.doi.org/10.1073/pnas.0503311102>.
- [42] S. Oehler, M. Amouyal, P. Kolkhof, B. von Wilcken-Bergmann, B. Muller-Hill, Quality and position of the three lac operators of *E. coli* define efficiency of repression, *EMBO J.* 13 (1994) 3348–3355.
- [43] N. Concalves, L. Martins, H. Tran, S. Oliveira, R. Neeli-venkata, J. Fonseca, A. Ribeiro, In vivo single-molecule dynamics of transcription of the viral T7 Phi 10 promoter in *Escherichia coli*, in: The 8th International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2016), June 26–30, 2016, Lisbon, Portugal 2016, pp. 9–15 ISBN: 978-1-61208-488-6.
- [44] G. Sezonov, D. Joseleau-Petit, R. D'Ari, *Escherichia coli* physiology in Luria-Bertani broth, *J. Bacteriol.* 189 (2007) 8746–8749, <http://dx.doi.org/10.1128/JB.01368-07>.
- [45] S.-T. Liang, M. Bipatnath, Y. Xu, S. Chen, P. Dennis, M. Ehrenberg, H. Bremer, Activities of constitutive promoters in *Escherichia coli*, *J. Mol. Biol.* 292 (1999) 19–37, <http://dx.doi.org/10.1006/jmbi.1999.3056>.
- [46] J. Lloyd-Price, S. Startceva, V. Kandavalli, J. Chandraseelan, N. Concalves, 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, <http://dx.doi.org/10.1093/dnares/dsw009>.
- [47] P. Cramer, Multisubunit RNA polymerases, *Curr. Opin. Struct. Biol.* 12 (2002) 89–97, [http://dx.doi.org/10.1016/S0959-440X\(02\)00294-4](http://dx.doi.org/10.1016/S0959-440X(02)00294-4).
- [48] A.S. Ribeiro, R. Zhu, S.A. Kauffman, A general modeling strategy for gene regulatory networks with stochastic dynamics, *J. Comput. Biol.* 13 (2006) 1630–1639, <http://dx.doi.org/10.1089/cmb.2006.13.1630>.
- [49] N. Shepherd, P. Dennis, H. Bremer, Cytoplasmic RNA Polymerase in *Escherichia coli*, *J. Bacteriol.* 183 (2001) 2527–2534, <http://dx.doi.org/10.1128/JB.183.8.2527>.
- [50] L. Bai, T.J. Santangelo, M.D. Wang, Single-Molecule Analysis of RNA Polymerase Transcription, *Annu. Rev. Biophys. Biomol. Struct.* 35 (2006) 343–360, <http://dx.doi.org/10.1146/annurev.biophys.35.010406.150153>.
- [51] F. Wang, E.C. Greene, Single-molecule studies of transcription: From one RNA polymerase at a time to the gene expression profile of a cell, *J. Mol. Biol.* 412 (2011) 814–831, <http://dx.doi.org/10.1016/j.jmb.2011.01.024>.
- [52] R.A. Mooney, S.A. Darst, R. Landick, Sigma and RNA polymerase: An on-again, off-again relationship? *Mol. Cell* 20 (2005) 335–345, <http://dx.doi.org/10.1016/j.molcel.2005.10.015>.
- [53] T.T. Harden, C.D. Wells, L.J. Friedman, R. Landick, A. Hochschild, J. Kondev, J. Gelles, Bacterial RNA polymerase can retain  $\sigma$  70 throughout transcription, *Proc. Natl. Acad. Sci. U. S. A.* 113 (2016) 602–607, <http://dx.doi.org/10.1073/pnas.1513899113>.
- [54] 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. U. S. A.* 112 (2015) E4390–E4399, <http://dx.doi.org/10.1073/pnas.1507592112>.
- [55] M. Smoluchowski, Attempt for a mathematical theory of kinetic coagulation of colloid solutions, *Z. Phys. Chem.* 92 (1917) 129.
- [56] S.S. Andrews, D. Bray, Stochastic simulation of chemical reactions with spatial resolution and single molecule detail, *Phys. Biol.* 1 (2004) 137–151, <http://dx.doi.org/10.1088/1478-3967/1/3/001>.
- [57] S.A. Rice, in: C.H. Bamford, C.F.H. Tipper, R.G. Compton (Eds.), *Diffusion-Limited Reactions*, Comprehensive Chemical Kinetics, vol. 25, Elsevier Science, Amsterdam, 1985.
- [58] K.B. Decker, D.M. Hinton, The secret to 6S: Regulating RNA polymerase by ribosequestration, *Mol. Microbiol.* 73 (2009) 137–140, <http://dx.doi.org/10.1111/j.1365-2958.2009.06759.x>.
- [59] T. Dong, R. Yu, H. Schellhorn, Antagonistic regulation of motility and transcriptome expression by RpoN and RpoH in *Escherichia coli*, *Mol. Microbiol.* 79 (2011) 375–386, <http://dx.doi.org/10.1111/j.1365-2958.2010.07449.x>.
- [60] M.T. Record, W.S. Reznikoff, M.L. Craig, K.L. McQuade, P.J. Schlax, *Escherichia coli* RNA polymerase (Es70), promoters, and the kinetics of the steps of transcription initiation, second ed. American Society for Microbiology, Washington, DC, 1996.
- [61] M. Jishage, A. Ishihama, A stationary phase protein in *Escherichia coli* with binding activity to the major  $\sigma$  subunit of RNA polymerase, *Proc. Natl. Acad. Sci. U. S. A.* 95 (1998) 4953–4958.
- [62] T.P. Malan, A. Kolb, H. Buc, W.R. McClure, Mechanism of CRP-cAMP activation of lac operon transcription initiation with action of the P1 promoter, *J. Mol. Biol.* 180 (1984) 881–909, [http://dx.doi.org/10.1016/0022-2836\(84\)90262-6](http://dx.doi.org/10.1016/0022-2836(84)90262-6).
- [63] C.M. Johnson, R.F. Schleif, In vivo induction kinetics of the arabinose promoters in *Escherichia coli*, *J. Bacteriol.* 177 (1995) 3438–3442.
- [64] S. Ogden, D. Haggerty, C.M. Stoner, D. Kolodrubetz, R. Schleif, The *Escherichia coli* L-arabinose operon: Binding sites of the regulatory proteins and a mechanism of positive and negative regulation, *Proc. Natl. Acad. Sci. U. S. A.* 77 (1980) 3346–3350.
- [65] C. Condon, C. Squires, C.L. Squires, Control of rRNA transcription in *Escherichia coli*, *Microbiol. Rev.* 59 (1995) 623–645.
- [66] A. Häkkinen, M. Kandhavelu, S. Garasto, A.S. Ribeiro, Estimation of fluorescence-tagged RNA numbers from spot intensities, *Bioinformatics* 30 (2014) 1146–1153, <http://dx.doi.org/10.1093/bioinformatics/btt766>.
- [67] K.M. Wassarman, C. Storz, 6S RNA regulates *E. coli* RNA polymerase activity, *Cell* 101 (2000) 613–623, [http://dx.doi.org/10.1016/S0092-8674\(00\)80873-9](http://dx.doi.org/10.1016/S0092-8674(00)80873-9).
- [68] J.A. Jaehning, Sigma factor relatives in eukaryotes, *Science* 253 (1991) 859, <http://dx.doi.org/10.1126/science.1876846>.
- [69] W.P. Tansey, W. Herr, TAFs: Guilt by association? *Cell* 88 (1997) 729–732, [http://dx.doi.org/10.1016/S0092-8674\(00\)81916-9](http://dx.doi.org/10.1016/S0092-8674(00)81916-9).
- [70] J.A. Goodrich, R. Tjian, TBP-TAF complexes: Selectivity factors for eukaryotic transcription, *Curr. Opin. Cell Biol.* 6 (1994) 403–409, [http://dx.doi.org/10.1016/0955-0674\(94\)90033-7](http://dx.doi.org/10.1016/0955-0674(94)90033-7).

# Supplementary Data for “Effects of $\sigma$ factor competition are promoter initiation kinetics dependent”

by Vinodh K. Kandavalli, Huy Tran, and Andre S. Ribeiro

## 1. Supplementary Results

### 1.1 Promoter sequences and their affinity to $\sigma$ factors

Sequences of promoters  $P_{BAD}$  [1],  $P_{tetA}$  [2,3],  $P_{lac-O1O3}$  and  $P_{lac-O1}$  [4], and  $P_{lac-ara1}$  [5] with consensus boxes (in red) at the -10 and -35 elements:

#### $P_{BAD}$ :

```
ccataagattagcggatcctacctgacgcttttatcgcaactctctactgtttctccatA
                        -35                               -10           +1
```

#### $P_{tetA}$ :

```
ccagatgattaattcctaatttttgttgacactctatcattgatagagttattttaccacT
                        -35                               -10           +1
```

#### $P_{lac-O1}$ and $P_{lac-O1O3}$ :

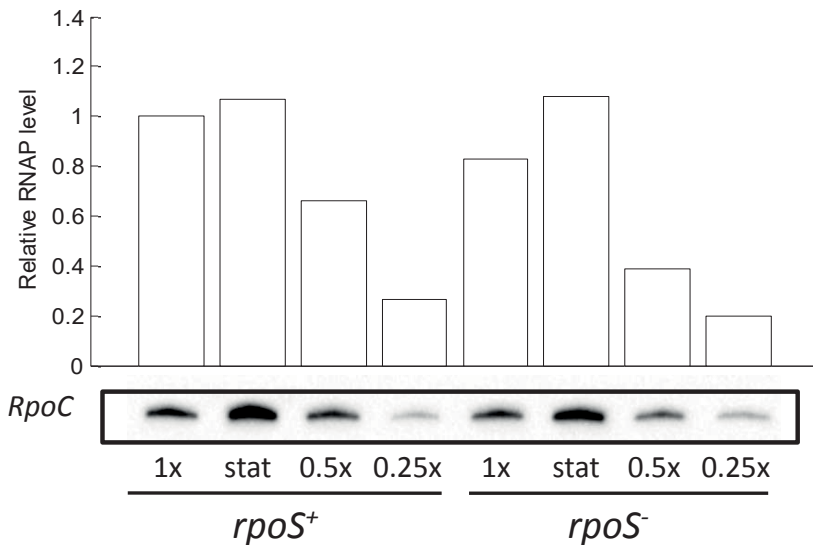
```
gctcactcattagccaccccaggctttacactttatgcttcggctcgtatgttgtgtggA
                        -35                               -10           +1
```

#### $P_{lac-ara1}$ :

```
ccataagattagcggatcctaagctttacaattgtgagcgctcacaattatgatagattcA
                        -35                               -10           +1
```

All promoters have very conserved consensus at position -10, that allows transcription by RNAP holoenzymes carrying factors of the  $\sigma^{70}$  family alone [6–8]. Further, the conserved consensus at position -35 is also associated to  $\sigma^{70}$  recognition [9,10], by causing the binding affinity to holoenzymes carrying  $\sigma^{70}$  to be much higher than to holoenzymes carrying  $\sigma^{38}$  [6,11] (present in the stationary phase [12]). Furthermore, the promoters lack consensus for recognition of the  $\sigma^{54}$  family (present in the exponential phase or under nitrogen stress [12] at positions -12 and -24 [13]).

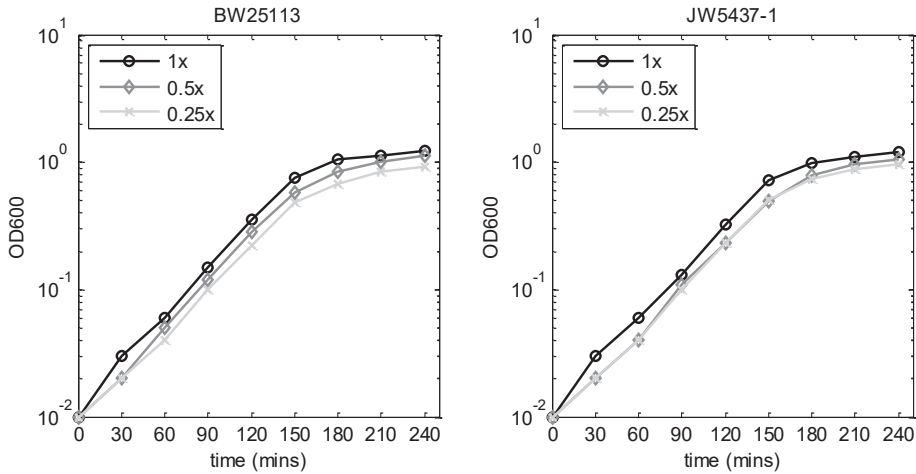
## 1.2 Measurements of RpoC protein levels in the exponential growth phase in 0.25x, 0.5x, and 1.0x media and in the stationary phase in 1.0x media



**Figure S1.** *RpoC* levels in BW25113 (*rpoS*<sup>+</sup>) and JW5437-1 (*rpoS*<sup>-</sup>) cells in the exponential growth phase when in 0.25x, 0.5x, and 1.0x media (section 2.3 of the main manuscript) and in the stationary phase in 1.0x media ('stat'), as measured by protein immunoblot. Protein levels are shown in relative to the 1x media for *rpoS*<sup>+</sup> cells. The normalized intensity volumes for RpoC proteins are also shown in Supplementary Table S1. The values were extracted from blot images by the 'Image Lab' software (version 5.2.1).

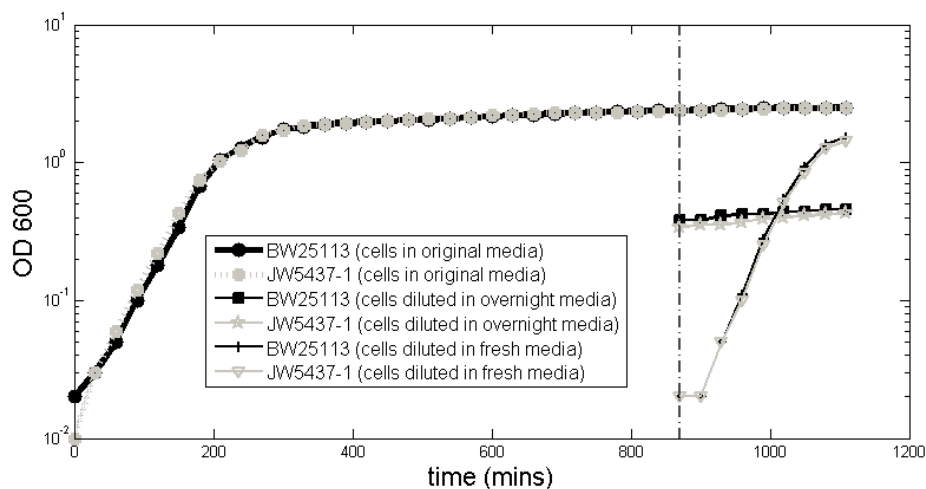
## 1.3 Bacterial growth rates for differing media, growth phase, and photo-toxicity.

We first measured growth rates of cells when in 1x, 0.5x and 0.25x fresh media (thus, differing in RNA polymerase concentrations, see main manuscript). For this, we measured the cultures' optical density (OD<sub>600</sub>) every 30 minutes, following dilution to OD<sub>600</sub> of 0.01, using the spectrophotometer (Figure S2). The strains are BW25113 (wild type, or WT) and a deletion mutant, JW5437-1. Both strains were obtained from the Keio single-gene knockout collection [14]. From Figure S2, for both WT and deletion mutants, one finds that the mean cell doubling time (~27 mins) is not heavily affected by the media within the range of 1x to 0.25x (neither it differs significantly between strains), in agreement with the results reported in [15].



**Figure S2.** OD curves of bacterial populations in 0.25x, 0.5x, and 1x fresh media conditions. The *E. coli* strains used are (A) BW25113 and (B) JW5437-1.

To obtain exponential and stationary-like growth rates for the bacterial cells (BW25113 and JW5437-1), we used the method proposed in [16]. Briefly, it consists of growing cells overnight in an initially fresh LB media (cells in ‘original media’ in Figure S3). At about 15 hours (dashed vertical line in Figure S3), the cells are in stationary phase (Figure S3) in agreement with [16]. Next, we diluted some of these cells and either kept them in the same media (‘cells diluted in overnight media’) or pre inoculated them in fresh media (‘cells diluted in fresh media’). From that moment (vertical line) onwards, while the set of cells diluted in original media continues to exhibit stationary-like growth (Figure S3, in agreement with [15][17]), those diluted in fresh media quickly change to exponential-like growth (Figure S3), as reported in [16]. As a control, we also continued to observe the growth of cells of the overnight culture not subject to dilution. These exhibited a very similar (lack of) growth to those diluted into the same overnight media, which supports the conclusion that the latter ones are in ‘stationary-like growth phase’.

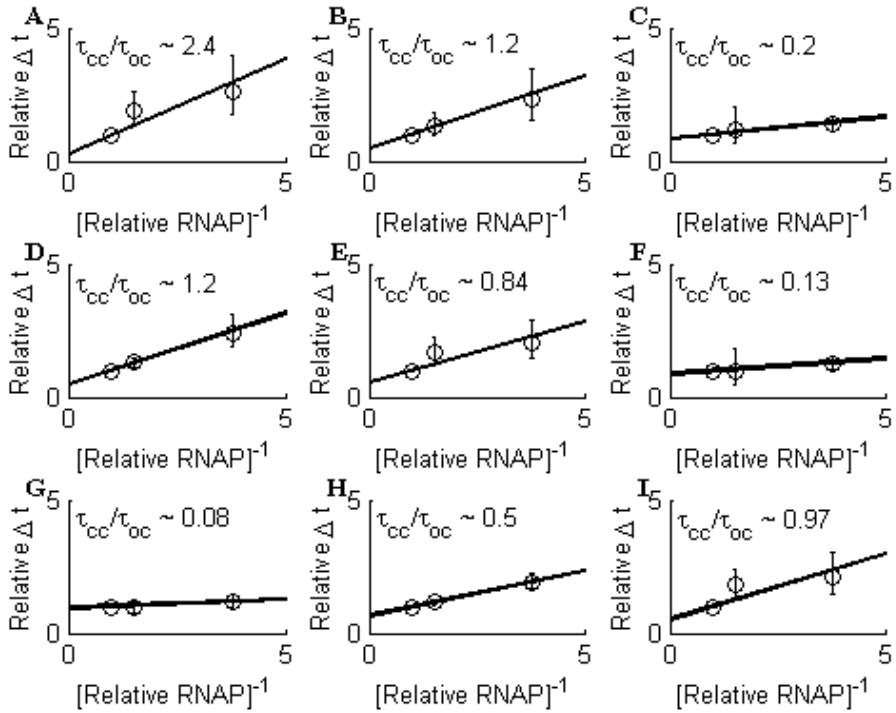


**Figure S3.** OD curves of *E. coli* strains BW25113 (black lines) and JW5437-1 (grey lines) cells. Growth rates were measured every 30 minutes

We also assessed whether the confocal imaging of cells (for 2 hours with images taken every 30 seconds), caused significant photo-toxicity. For this, we compared division times of BW25113 cells in exponential growth phase under the microscope, when and when not exposed to confocal imaging (cells imaged by phase contrast in both cases). We observed doubling times of 72 min. and 68 min. with and without confocal imaging, respectively, from which we conclude that the confocal imaging does not introduce significant photo-toxicity. In these experiments, cells were in fresh 1x media (Methods).

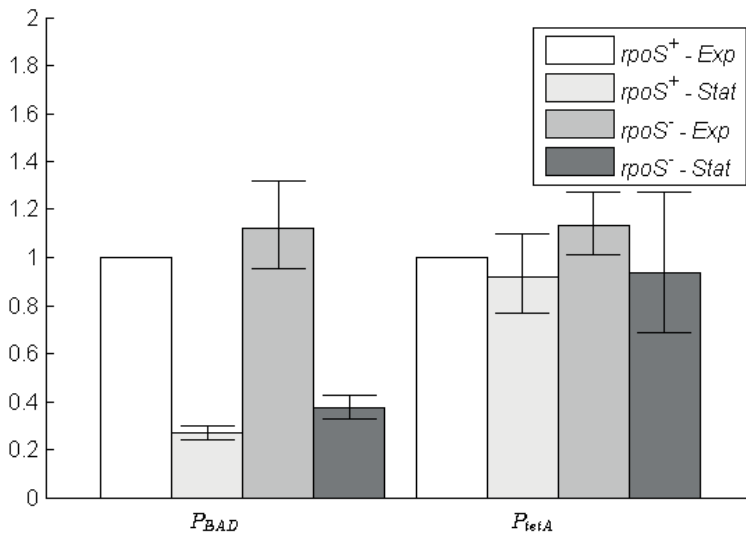


### 1.4 Relative $\tau$ plots for various promoters subject to different induction levels



**Figure S4. Relative  $\tau$  plots for various promoters subject to various induction levels. (A)**  $P_{BAD}$  with 0.1% arabinose. **(B)**  $P_{BAD}$  with 0.01% arabinose. **(C)**  $P_{BAD}$  with 0.001% arabinose. **(D)**  $P_{lac-O103}$  with 1 mM IPTG. **(E)**  $P_{lac-O103}$  with 0.05 mM IPTG. **(F)**  $P_{lac-O103}$  with 0.005 mM IPTG. **(G)**  $P_{tetA}$  with no inducers. **(H)**  $P_{lac-O1}$  with 1 mM IPTG. **(I)**  $P_{lac-ara1}$  with 1 mM IPTG and 0.1% arabinose (full induction). Also shown is the value of  $\tau_{cc}/\tau_{oc}$  in each case, extracted from the intersection of the linear fit with the y-axis (corresponding to a condition with  $[\text{RNAP}] \sim \infty$ , Methods).

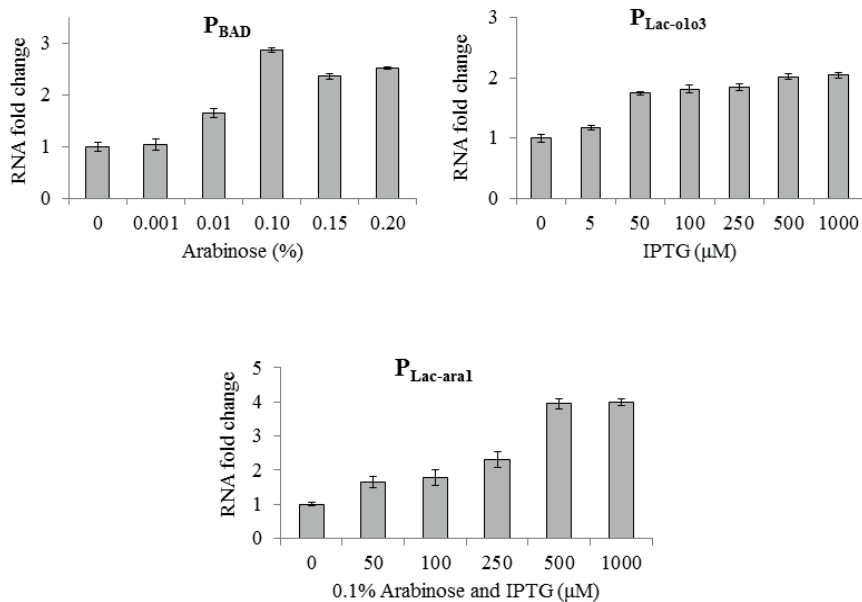
## 1.5 qPCR measurements of relative RNA levels under the control of $P_{BAD}$ and of $P_{tetA}$ in $rpoS^+$ and $rpoS^-$ cells



**Figure S5.** Relative RNA levels under the control of  $P_{BAD}$  (0.1% arabinose) and of  $P_{tetA}$  in BW25113 ( $rpoS^+$ ) and JW5437-1 ( $rpoS^-$ ) cells in the exponential growth phase (“Exp”) and in the stationary growth phase (“Stat”), as measured by qPCR using the 16S RNA housekeeping gene for internal reference.

## 1.6 Induction curves

We obtained induction curves for  $P_{BAD}$ ,  $P_{lac-O103}$ , and  $P_{lac-ara-1}$  from qPCR measurements. Results are shown in Figure S6. In the case of  $P_{lac-O1}$ , we observed no significant induction when adding IPTG (data not shown), which could be expected given the absence in this promoter of 2 of the 3 operator binding sites for the LacI repressors. For  $P_{tetA}$ , we did not obtain an induction curve, as this promoter does not require induction, because the cells (BW25113) lack the gene coding for TetR, the repressor of  $P_{tetA}$ .



**Figure S6.** Induction curves of P<sub>BAD</sub>, P<sub>Lac0103</sub>, and P<sub>Lac-ara1</sub> obtained by qPCR. The mean and standard error of RNA numbers per cell in each condition were extracted from 3 replicates.

### 1.7 Intensity values for RpoC protein identified in the chemiluminescent blot

In Supplementary Table S1, we present the values for all measurements, as reported by the software ‘ImageLab’ after analyzing the images obtained by Western Blot.

Sample	Channel	Band No.	Relative Front	Volume (Int)	Band %	Norm. Factor	Norm. Vol. (Int)
<i>rpoS</i> <sup>+</sup> 1X	Chemi	1	0,186928	13042740	100,0	1,00000	13042740
<i>rpoS</i> <sup>+</sup> stat	Chemi	1	0,188235	30918800	100,0	0,44952	13898773
<i>rpoS</i> <sup>+</sup> 0.5X	Chemi	1	0,190850	11238084	100,0	0,76600	8608410
<i>rpoS</i> <sup>+</sup> 0.25X	Chemi	1	0,193464	2004464	100,0	1,72060	3448882
<i>rpoS</i> <sup>-</sup> 1X	Chemi	1	0,192157	10965204	100,0	0,98637	10815743
<i>rpoS</i> <sup>-</sup> stat	Chemi	1	0,194771	21550900	100,0	0,65403	14095053
<i>rpoS</i> <sup>-</sup> 0.5X	Chemi	1	0,194771	7248384	100,0	0,69351	5026869
<i>rpoS</i> <sup>-</sup> 0.25X	Chemi	1	0,196078	3065088	100,0	0,84078	2577086

Table S1. *RpoC* protein normalized volumes in BW25113 (*rpoS*<sup>+</sup>) and JW5437-1 (*rpoS*<sup>-</sup>) cells in the exponential growth phase when in 0.25x, 0.5x, and 1.0x media and in the stationary phase (‘stat’) in 1.0x media, as measured by western blot and analyzed by ‘Image Lab’ software (version 5.2.1).

## 2. Supplementary Materials and Methods

### 2.1 Plasmids

Target promoter	Target Construct	Plasmid name	Reference	Note
P <sub>BAD</sub>	P <sub>BAD</sub> - <i>mrfp1-96bs</i>	pTRUEBLUE	[18]	Constructed in our lab.
P <sub>tetA</sub>	P <sub>tetA</sub> - <i>mrfp1-96bs</i>	pTRUEBLUE	[3]	Constructed in our lab.
P <sub>lac-O1O3</sub>	P <sub>lac-O1O3</sub> - <i>mCherry-48bs</i>	pBELO	[19]	Constructed in our lab.
P <sub>lac-O1</sub>	P <sub>lacO1</sub> - <i>lacZα-96bs</i>	pTRUEBLUE	[20]	Kind gift from I. Golding
P <sub>lac-ara1</sub>	P <sub>lac-ara1</sub> - <i>mrfp1-96bs</i>	pTRUEBLUE	[4]	Kind gift from I. Golding

Table S2. List of target constructs inserted into cells.

### 2.2 Genetic construct of P<sub>lac-O1O3</sub>-*mCherry-48bs*

To construct P<sub>lac-O1O3</sub>-*mCherry-48* binding sites (bs), we used a plasmid carrying *mCherry* followed by a 48bs array in the pBELO vector backbone, originally constructed in [19]. To amplify the target gene containing P<sub>lac-O1O3</sub> with both operator sites (O1 and O3), we used the chromosomal native *lacZ* gene. A primer set was designed as follows:

P<sub>lac-O1O3</sub> Forward: 5'GCTCACCATCCTCCTCGTAATCATGGTCATAGCTGTTTCCTG 3'

P<sub>lac-O1O3</sub> Reverse: 5'CGACAGGTTTCCCGACGCGTTGGCCGATTCATTAATG 3'

P<sub>lac-O1O3</sub> was amplified and inserted into the pBELO vector backbone by Gibson Assembly [21], to obtain a single copy F-based plasmid carrying the target region P<sub>lac-O1O3</sub>-*mCherry-48bs*. This product was transferred into competent *E. coli* host cells. The recombinants were selected with antibiotic screening and further confirmed with sequence analysis.

### 2.3 Quantitative PCR

Cells (5 ml) at different growth phases were harvested as described in the main manuscript, followed by the addition of 10 ml of RNA protect bacteria reagent and immediate mixing by vortexing for 5 seconds. Samples were incubated for 5 minutes at room temperature, and then centrifuged at 5000 × g for 10 minutes. The supernatant was discarded and any residual supernatant was removed by inverting the tube once onto a paper towel. The entire RNA content was isolated by using the RNeasy kit (Qiagen) according to the instructions of the manufacturer. Samples were quantified using a Nanovue plus spectrophotometer (GE Healthcare life sciences) and the quality of

the isolated RNA was assessed by measuring the ratio of absorbance at 260 and 280 nm (A<sub>260</sub>/A<sub>280</sub> ratio) of the sample (2.0–2.1). DNaseI treatment was then performed to avoid DNA contamination. cDNA was synthesized (Fermentas, Finland) from 1 µg of RNA with iScript Reverse Transcription Supermix. The cDNA templates with a final concentration of 10 ng/µl were added to the qPCR master mix containing iQ SYBR Green supermix (Fermentas, Finland) with primers for the target and reference genes at a final concentration of 200 nM. We used the 16S RNA housekeeping gene for internal reference.

The primers set for the target RNAs and the reference gene (16S RNA) are as follow:

mRPF1 (for the study of P<sub>BAD</sub>, P<sub>tetA</sub> and P<sub>lac-ara1</sub>)

Forward: 5' TACGACGCCGAGGTCAAG 3'

Reverse: 5' TTGTGGGAGGTGATGTCCA 3'

lacZα (for the study of P<sub>lacO1</sub>)

Forward: 5' CCGGATCCTCGAGAGCTTAG 3'

Reverse: 5' CTAATCGATTCAATTGGGTAACG 3'

mCherry (for the study of P<sub>lac-O103</sub>)

Forward: 5' CACCTACAAGGCCAAGAAGC 3'

Reverse: 5' TGGTGTAGTCCTCGTTGTGG 3'

16S RNA:

Forward: 5' CGTCAGCTCGTGTGTGAA 3'

Reverse: 5' GGACCGCTGGCAACAAAG 3'.

The qPCR experiments were performed using a Biorad MiniOpticon Real time PCR system (Biorad, Finland). The following thermal cycling protocol was used: 40 cycles of 95 °C for 10 s, 52 °C for 30 s, and 72 °C for 30 s for each cDNA replicate. These reactions were performed in three replicates for each condition, with a final reaction volume of 25 µl. We use no-RT controls and no-template controls to crosscheck non-specific signals and contamination. PCR efficiencies of these reactions were greater than 95%. The data from CFX Manager™ Software was used to calculate the relative gene expression and its standard error [22].

## 2.4 Western blotting

Cells were harvested by centrifuging and then lysed with the B-PER Bacterial protein extraction reagent (Thermo Scientific) containing the protease inhibitors. Cell lysate was incubated at room temperature for 10 mins and then centrifuged at 15000 ×g for 5 mins to remove debris and collect the supernatant. The samples containing the total protein were diluted with the 4X lamella sample

loading buffer containing the  $\beta$  mercaptomethanol and boiled for 5 mins at 95°C. Approximately 30  $\mu$ g of the total protein in each sample was resolved by 4 – 20 % TGX gels (Biorad). Proteins were separated by electrophoresis and then electro-transferred on the PVDF membrane (Biorad). Membranes were incubated with respective primary antibodies for RpoC, RpoS and RpoD (Biologend) of 1:2000 dilution overnight at 4 °C, followed by HRP-secondary antibodies (Sigma Aldrich) 1:5000 dilution for 1 hour at room temperature. Detection was done by the chemilumiscence reagent (Biorad). Images were generated by the chemidoc XRS system (Biorad). Band quantification was done using Image Lab software version 5.2.1.

## 2.5 Estimation of the ratio between the closed and open complex formation

Given the model described by equations (A-B) and (D-F) in the main manuscript, the mean time-scale of the closed complex formation ( $\tau_{cc}$ ) (i.e. for the occurrence of a reaction D) is inversely proportional to the abundance of  $RNAP \cdot \sigma^{70}$ , and equals:

$$\tau_{cc} = \frac{1}{k_{cc}[RNAP \cdot \sigma^{70}]} \quad (S1)$$

In (D),  $k_{cc}$  is the rate constant of the closed complex formation. Provided that the closed and open complex formations are the main rate-limiting steps in transcription [23–26], as assumed by the model, then one finds that the mean interval between consecutive transcription events ( $\Delta t$ ) equals approximately:

$$\Delta t = \tau_{cc} + \tau_{oc} = \frac{1}{k_{cc}[RNAP \cdot \sigma^{70}]} + \frac{1}{k_{oc}} \quad (S2)$$

where  $\tau_{oc}$  is the mean time-scale of the open complex formation.

The variables in equation [S2], whose values can be obtained from measurements, are  $\Delta t$  (see section 2.7 below) and  $[RNAP \cdot \sigma^{70}]$  (by measuring  $[RNAP]$ ; see main manuscript). In particular, as  $\Delta t$  is inversely proportional to the target RNA production rate, it can be extracted by qPCR measurements [27,28]. Else, it can be directly measured from time-lapse microscopy measurements (section 2.6 of this document). Meanwhile, relative values of  $[RNAP]$  can be obtained by protein immunoblot (previous section of this document). Given these two quantities, it is possible to estimate  $\tau_{cc}/\tau_{oc}$  by weighted least square fit of a line to the measured mean values of  $\Delta t$  when plotted against the relative numbers of  $[RNAP]^{-1}$ .

In practice, here we obtain  $\tau_{cc}/\tau_{oc}$  by measuring both  $[RNAP]$  and  $1/\Delta t$ , in three conditions, differing in the RNAP abundance in the cells, following the strategy proposed in [29]. To alter the intracellular abundance of RNAP, we place cells in modified LB media, with chemical

compositions of 1x, 0.5x and 0.25x (see above). As the RNA degradation rates are not affected by the growth conditions used here [27], the transcription rates (proportional to  $1/\Delta t$ ) at 0.5x and 0.25x relative to that in 1x media can be assessed from the relative RNA levels measured by qPCR.

## 2.6 Time lapse microscopy

Time lapse microscopy measurements of the dynamics of RNA production in individual cells, at the single RNA level, were conducted for  $P_{BAD}$  and  $P_{tetA}$  so as to compare their activity when cells were in the exponential and in the stationary phases. For this, in general, the strain has to contain two constructs [4]: a reporter plasmid (carrying MS2-GFP, here under the control of  $P_{Lac}$ ) and a single-copy plasmid vector pIG-BAC carrying the target transcript (mRFP1 followed by 96 MS2-binding sites) under the control of  $P_{BAD}$  or  $P_{tetA}$ .

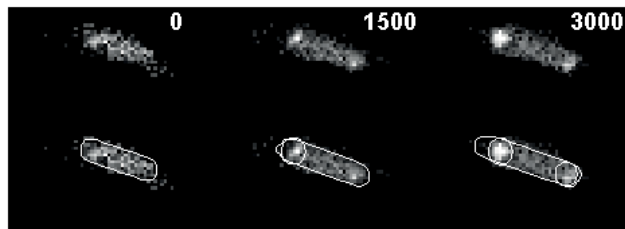
This system has been used to measure the distribution of time intervals between RNA production events due to its ability to detect individual target RNA molecules consisting of the numerous MS2 coat protein binding sites, which are rapidly bound by fluorescently tagged MS2 coat proteins. These tagged RNAs can be seen as soon as they are produced under a fluorescence microscope as fluorescent foci [3,4,18,30–32].

One hour after incubation in the respective phase-inducing media, for both cells containing  $P_{BAD}$  or  $P_{tetA}$ , we first induced the MS2-GFP reporter (under the control of  $P_{lac}$  in both cases) by adding 1 mM IPTG for 45 minutes in liquid culture. We verified by visual inspection that, at this stage, cells contained sufficient, uniformly distributed MS2-GFP in the cytoplasm to detect any target RNA [3,4,18,30–32]. Next, to activate  $P_{BAD}$  controlling the production of the target RNA, we added 0.1% of arabinose for 5 minutes while in liquid culture. Cells were then placed under microscope observation. Meanwhile,  $P_{tetA}$  does not require induction, as both BW25113 and JW5437-1 strains lack the gene coding for TetR, the repressor of  $P_{tetA}$  [14]. Note that this does not interfere with the comparison of the dynamics of the two promoters, as we only compare distributions of time intervals between consecutive RNAs in each cell, not the mean number of RNAs produced.

In all cases, we used an inverted microscope body Nikon Eclipse (Ti-E, Nikon, Japan). In the left port of the body, we used a confocal C2+ scanner connected to a LU3 laser system (Nikon) with a 488 nm argon ion laser. The laser shutter was open only during exposure time, to minimize photobleaching. For phase-contrast, we used an external phase-contrast setting (Nikon) with a DS-Fi2 CCD-camera. For both phase-contrast and confocal imaging, we used a 100x oil-immersion objective (Apo TIRF, Nikon). The software used to capture images and control the microscope was NIS-Elements (Nikon). To maintain, during microscopy, stable growth conditions and induction of

the promoters controlling the production of the RNA target for MS2-GFP and of MS2-GFP reporter proteins, a peristaltic pump introduced a constant flow of the appropriate media and inducers.

For imaging, a few  $\mu\text{l}$  of cells were placed between a glass coverslip and a slab of 3% agarose containing the respective media and inducers. During image acquisition, the slide was kept in a temperature-controlled chamber (Bioprotechs, FCS2) at 37 °C. Cells were imaged every 30 seconds for 2 hours, for both cell segmentation (phase contrast images) and detecting MS2-GFP tagged RNAs (confocal microscopy images). Example images of a cell over time, along with the results of cell segmentation and detection of spots (which appear as a bright spots in the green fluorescent channel) are shown in Figure S7.



**Figure S7.** MS2-GFP tagged RNAs in an *E. coli* cell over time. Unprocessed frames (top) along with the segmented cell and RNA spots (bottom). The moment images were captured is shown at the top of each frame.

## 2.7 Image processing, temporal fluorescence intensity of MS2-GFP tagged RNA molecules, and estimation of time intervals between consecutive RNA production events in individual cells

Image analysis of microscopy time-series was performed as in [3,18,30]. First, we used a semi-automated cell segmentation strategy as in [33], using the software MAMLE [34] followed by manual correction. Afterwards, fluorescent spots in each segmented cell, at each moment, are detected automatically as in [32], by estimating the cell background intensity distribution using its median and median absolute deviation, and then performing thresholding with a given confidence level and assuming that this distribution is Gaussian. From this, one obtains information on each cell, at each time point, along with the spots and their intensity.

Next, as in [33], we establish the relationships between cells in sequential frames (lineage construction). The segments overlapping the most are associated with the segments of the next frame. If the association is one-to-one, it is assumed that it must be the same cell (no division occurred). Else, if the association is one-to-many, it is interpreted as a cell division. For zero-to-one



or one-to-zero associations, no relationships between segments are established. Manual correction of the segmentation removes cases where cells disappear in the middle of a time series.

It follows the estimation of time intervals between consecutive RNA productions in individual cells. This estimation relies on the determination of when new RNA molecules appear in the cells and, from there, it obtains absolute time intervals. This technique of detecting, in individual cells, the moment when a novel RNA molecule ('spot') appears from time lapse microscopy images using the multiple MS2-GFP RNA-tagging system [4], has been previously used in [3,18,21,22,29-31,40-41].

First, the moments of appearance of novel target RNAs in each cell are obtained, as in [32], by least squares fitting a monotonically increasing piecewise-constant function to the corrected total spot intensity in that cell over time. The number of terms for the fitting is selected by an F-test with a p-value of 0.01. Each discontinuity, i.e. jump, corresponds to the production of one target RNA (Figure S7) [3,18,30–32]. Finally, the time intervals between consecutive RNA production events in each cell are extracted (events separated by cell divisions are not considered). This method, first proposed in [32] and subsequently improved in [36], allows estimating the accuracy of the estimation based on the number of cells observed and the level of noise of the 'spot fluorescence' signal from individual cells. For a set of more than 100 cells and a noise level of 1 in the fluorescent signal (as measured by the coefficient of variation and in agreement with the present measurements), we expect an accuracy of 80% [36]. An example application of this method to the signal from a cell is shown in Figure S7.

For this method to count accurately the RNA production events, first, new RNA molecules need to appear nearly fully-tagged when first detected, so as to cause a significant "jump" of standard size in the "total spots fluorescence intensity" of the cell. This will occur, provided that the speed of transcription elongation (expected to be ~60 and ~90 base pairs per second, at 37°C [37–39]) and the speed of MS2-GFP binding to the target RNA are such that a 'complete spot formation' (i.e. the occupation of nearly all MS2-GFP binding sites) does not take much longer than the interval between consecutive images, which in our measurements is 30 seconds long. Second, it is necessary that an MS2-GFP tagged RNA, once tagged, does not degrade significantly (neither abruptly nor gradually) during the measurement period (to allow using a step-increasing function).

Both assumptions were recently tested by observing the fluorescence intensity of individual, tagged RNAs for 30 minutes (1 image per minute) in individual cells that contained a single target RNA [35]. From the data in [35], first, new RNA molecules are fully or nearly fully-tagged when first detected, as no significant increases in tagged RNA fluorescence are observed after detection

of the tagged RNA. Second, by fitting the intensity of each tagged RNA over time with a decaying exponential function and inferring its intensity degradation rate, a mean decaying rate of  $\sim 8.1 \times 10^{-5} \text{ s}^{-1}$  was measured, corresponding to a mean half-life of  $\sim 144$  mins, which is longer than our observation window (120 mins). Given these results, we conclude that the fluorescence of tagged RNAs does not decrease significantly over time (gradually or abruptly) during the measurement period. These results are supported by previous studies of the properties of the MS2 coat protein of bacteriophage [40,41], and by studies that showed that most MS2 binding sites of the target RNA are constantly occupied by MS2-GFP proteins, resulting in the ‘immortalization’ of tagged RNAs due to the isolation from RNA-degrading enzymes [4,20].

Finally, note that the RNA ‘jump detection’ method can tolerate infrequent “blinking” of existing tagged RNAs, due to moving out of focus transiently, without loss of information [32,36].

## 2.8 ‘Relative’ and ‘absolute’ $\tau$ plots

The ‘relative’  $\tau$  plots of data from *in vivo* cells here presented are based on the assumption that free intracellular RNAP concentrations can be changed within a significant range (similar to *in vitro*  $\tau$  plot measurements). This assumption was shown to be valid in [29] when changing the media composition in a specific way (see Supplementary Figure S1 and Table S1, with RpoC numbers in each media condition).

Once setting the media conditions that establish specific free intracellular RNAP concentrations (shown to correspond to the total RNAP concentrations [29]), one can measure by qPCR, for each condition, the rate of RNA production of the promoter of interest (relative to the reference gene), which is inversely *proportional* to the mean duration of the intervals between consecutive RNA productions in live cells. Next, from these measurements, one can obtain the ratio between relative rates of RNA production in different conditions (e.g. exponential and stationary phases). Note that this ratio *equals* the inverse of the ratio between time intervals between RNA productions (see equation H in the main manuscript).

Then, one can fit the general model of transcription initiation to the empirical data (reactions A-B and D-F, from which H is derived), which accounts for the multi-step nature of transcription initiation and the need for a  $\sigma$  factor to initiate transcription. That is, from equation (H) for which we obtain empirical values for two of its three terms, we estimate the mean *in vivo* time-scale of the closed complex formation relative to the time-scale of the open complex formation (ratio  $\tau_{cc}/\tau_{oc}$ ), by extrapolating it for an infinite RNAP concentration. This methodology follows the one presented in [29] (which is based on measurements of time intervals between consecutive RNA productions in

individual cells and thus allowed estimating the absolute value of  $\tau_{oc}$ ), and is similar to the extrapolation for *in vitro* data presented in [23].

Meanwhile, ‘absolute’  $\tau$  plots, which are based on the same assumptions as above, are obtained directly from measurements of the mean time-scale of intervals between consecutive RNA productions in individual cells and mean RNAP concentrations. By measuring the former in conditions that cause the latter to differ [29], one can then plot the mean absolute time-scale of intervals as a function of the inverse of RNAP concentrations and estimate the mean absolute *in vivo* time-scale of the open complex formation  $\tau_{oc}$ , as in [29], as the value of the mean absolute time-scale of the intervals when the RNAP concentration is infinitely large.

In the present study, only relative  $\tau$  plots are presented (section 1.4 of this document).

## REFERENCES

- [1] B.R. Smith, R. Schleif, Nucleotide sequence of the L-arabinose regulatory region of *Escherichia coli* K12., *J. Biol. Chem.* 253 (1978) 6931–6933.
- [2] M.T. Korpela, J.S. Kurittu, J.T. Karvinen, M.T. Karp, A recombinant *Escherichia coli* sensor strain for the detection of tetracyclines, *Anal. Chem.* 70 (1998) 4457–4462. doi:10.1021/ac980740e.
- [3] A.B. Muthukrishnan, M. Kandhavelu, J. Lloyd-Price, F. Kudasov, S. Chowdhury, O. Yli-Harja, A.S. Ribeiro, Dynamics of transcription driven by the tetA promoter, one event at a time, in live *Escherichia coli* cells, *Nucleic Acids Res.* 40 (2012) 8472–8483. doi:10.1093/nar/gks583.
- [4] I. Golding, J. Paulsson, S.M. Zawilski, E.C. Cox, Real-time kinetics of gene activity in individual bacteria., *Cell.* 123 (2005) 1025–36. doi:10.1016/j.cell.2005.09.031.
- [5] R. Lutz, H. Bujard, Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements, *Nucleic Acids Res.* 25 (1997) 1203–10. doi:10.1093/nar/25.6.1203.
- [6] F. Colland, N. Fujita, D. Kotlarz, J. a. Bown, C.F. Meares, A. Ishihama, A. Kolb, Positioning of  $\sigma(S)$ , the stationary phase  $\sigma$  factor, in *Escherichia coli* RNA polymerase-promoter open complexes, *EMBO J.* 18 (1999) 4049–4059. doi:10.1093/emboj/18.14.4049.
- [7] A. Wise, R. Brems, V. Ramakrishnan, M. Villarejo, Sequences in the -35 region of *Escherichia coli* rpoS-dependent genes promote transcription by E $\sigma$ S, *J. Bacteriol.* 178

(1996) 2785–2793.

- [8] T.M. Gruber, C.A. Gross, Multiple sigma subunits and the partitioning of bacterial transcription space., *Annu. Rev. Microbiol.* 57 (2003) 441–66.  
doi:10.1146/annurev.micro.57.030502.090913.
- [9] S. Lissner, H. Margalit, Compilation of *E. coli* mRNA promoter sequences, *Nucleic Acids Res.* 21 (1993) 1507–1516. doi:10.1093/nar/21.7.1507.
- [10] D.K. Hawley, W.R. McClure, Nucleic Compilation and analysis of *Escherichia coli* promoter DNA sequences, *Nucleic Acids Res.* 11 (1983) 2237–2255.
- [11] G. Becker, R. Hengge-Aronis, What makes an *Escherichia coli* promoter sigma(S) dependent? Role of the -13/-14 nucleotide promoter positions and region 2.5 of sigma(S)., *Mol. Microbiol.* 39 (2001) 1153–65.
- [12] M. Jishage, A. Iwata, S. Ueda, A. Ishihama, Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli* : intracellular levels of four species of sigma subunit under various growth conditions., *J. Bacteriol.* 178 (1996) 5447–5451.
- [13] H. Barrios, B. Valderrama, E. Morett, Compilation and analysis of sigma(54)-dependent promoter sequences., *Nucleic Acids Res.* 27 (1999) 4305–4313. doi:gkc653 [pii].
- [14] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K.A. Datsenko, M. Tomita, B.L. Wanner, H. Mori, Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection., *Mol. Syst. Biol.* 2 (2006) 1–11.  
doi:10.1038/msb4100050.
- [15] T.E. Shehata, A.G. Marr, Effect of nutrient concentration on the growth of *Escherichia coli*., *J. Bacteriol.* 107 (1971) 210–216.
- [16] E.O. Powell, Growth rate and generation time of bacteria, with special reference to continuous culture., *J. Gen. Microbiol.* 15 (1956) 492–511. doi:10.1099/00221287-15-3-492.
- [17] T. Dong, H.E. Schellhorn, Global effect of RpoS on gene expression in pathogenic *Escherichia*, 17 (2009) 1–17. doi:10.1186/1471-2164-10-349.
- [18] J. Mäkelä, M. Kandhavelu, S.M.D. Oliveira, J.G. Chandraseelan, J. Lloyd-Price, J. Peltonen, O. Yli-Harja, A.S. Ribeiro, In vivo single-molecule kinetics of activation and subsequent activity of the arabinose promoter, *Nucleic Acids Res.* 41 (2013) 6544–6552.  
doi:10.1093/nar/gkt350.

- [19] N. Goncalves, L. Martins, H. Tran, S. Oliveira, R. Neeli-venkata, J. Fonseca, A. Ribeiro, In vivo single-molecule dynamics of transcription of the viral T7 Phi 10 promoter in *Escherichia coli*, in: The 8th International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2016), June 26-30, 2016, Lisbon, Portugal. ISBN: 978-1-61208-488-6 (pp. 9-15).
- [20] I. Golding, E.C. Cox, RNA dynamics in live *Escherichia coli* cells., Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 11310–5. doi:10.1073/pnas.0404443101.
- [21] D.G. Gibson, L. Young, R.-Y. Chuang, J.C. Venter, C. a Hutchison, H.O. Smith, Enzymatic assembly of DNA molecules up to several hundred kilobases., Nat. Methods. 6 (2009) 343–345. doi:10.1038/nmeth.1318.
- [22] K.J. Livak, T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method., Methods. 25 (2001) 402–408. doi:10.1006/meth.2001.1262.
- [23] W.R. McClure, Rate-limiting steps in RNA chain initiation, Proc. Natl. Acad. Sci. U. S. A. 77 (1980) 5634–5638. doi:10.1073/pnas.77.10.5634.
- [24] W.R. McClure, Mechanism and control of transcription initiation in prokaryotes., Annu. Rev. Biochem. 54 (1985) 171–204. doi:10.1146/annurev.bi.54.070185.001131.
- [25] R. Lutz, T. Lozinski, T. Ellinger, H. Bujard, Dissecting the functional program of *Escherichia coli* promoters: the combined mode of action of Lac repressor and AraC activator., Nucleic Acids Res. 29 (2001) 3873–3881. doi:10.1093/nar/29.18.3873.
- [26] H. Buc, W.R. McClure, Kinetics of open complex formation between *Escherichia coli* RNA polymerase and the lac UV5 promoter. Evidence for a sequential mechanism involving three steps., Biochemistry. 24 (1985) 2712–2723. doi:10.1021/bi00332a018.
- [27] H. Chen, K. Shiroguchi, H. Ge, X.S. Xie, Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*., Mol. Syst. Biol. 11 (2015) 781. doi:10.15252/msb.20145794.
- [28] 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., Science. 329 (2010) 533–538. doi:10.2142/biophys.51.136.
- [29] J. Lloyd-Price, S. Startceva, V. Kandavalli, J. Chandraseelan, N. Goncalves, S.M.D. Oliveira,

- A. Häkkinen, A.S. Ribeiro, Dissecting the stochastic transcription initiation process in live *Escherichia coli*, DNA Res. (2016) 1–12. doi:doi: 10.1093/dnares/dsw009.
- [30] M. Kandhavelu, J. Lloyd-Price, A. Gupta, A.B. Muthukrishnan, O. Yli-Harja, A.S. Ribeiro, Regulation of mean and noise of the in vivo kinetics of transcription under the control of the lac/ara-1 promoter, FEBS Lett. 586 (2012) 3870–3875. doi:10.1016/j.febslet.2012.09.014.
- [31] A.B. Muthukrishnan, A. Martikainen, R. Neeli-Venkata, A.S. Ribeiro, In Vivo Transcription Kinetics of a Synthetic Gene Uninvolved in Stress-Response Pathways in Stressed *Escherichia coli* Cells., PLoS One. 9 (2014) 1–11. doi:10.1371/journal.pone.0109005.
- [32] M. Kandhavelu, A. Häkkinen, O. Yli-Harja, A.S. Ribeiro, Single-molecule dynamics of transcription of the lar promoter, Phys. Biol. 9 (2012) 1–12. doi:10.1088/1478-3975/9/2/026004.
- [33] A. Häkkinen, 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 (2013) 1708–1709. doi:10.1093/bioinformatics/btt194.
- [34] S. Chowdhury, M. Kandhavelu, O. Yli-Harja, A.S. Ribeiro, An interacting multiple model filter-based autofocus strategy for confocal time-lapse microscopy., J. Microsc. 245 (2012) 265–75. doi:10.1111/j.1365-2818.2011.03568.x.
- [35] 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 (2015) 2579–2587. doi:10.1039/C5MB00244C.
- [36] A. Häkkinen, A.S. Ribeiro, Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data, Bioinformatics. 31 (2014) 69–75. doi:10.1093/bioinformatics/btu592.
- [37] U. Vogel, K.F. Jensen, The RNA chain elongation rate in *Escherichia coli* depends on the growth rate, J. Bacteriol. 176 (1994) 2807–2813.
- [38] P.P. Dennis, M. Ehrenberg, D. Fange, H. Bremer, Varying rate of RNA chain elongation during rrn transcription in *Escherichia coli*, J. Bacteriol. 191 (2009) 3740–3746. doi:10.1128/JB.00128-09.
- [39] J. Ryals, R. Little, H. Bremer, Control of rRNA and tRNA syntheses in *Escherichia coli* by guanosine tetraphosphate, J. Bacteriol. 151 (1982) 1261–1268.

- [40] S.J. Talbot, S. Goodman, S.R. Bates, C.W. Fishwick, P.G. Stockley, Use of synthetic oligoribonucleotides to probe RNA-protein interactions in the MS2 translational operator complex., *Nucleic Acids Res.* 18 (1990) 3521–3528.
- [41] D. Fusco, N. Accornero, B. Lavoie, S.M. Shenoy, J.-M. Blanchard, R.H. Singer, E. Bertrand, Single mRNA Molecules Demonstrate Probabilistic Movement in Living Mammalian Cells, *Curr. Biol.* 13 (2003) 161–167. doi:10.1016/S0960-9822(02)01436-7.





# PUBLICATION III

**Rate-limiting steps in transcription dictate sensitivity to variability in cellular components.**

Jarno Mäkelä, Vinodh Kandavalli and Andre S. Ribeiro


Scientific Reports. 7:10588, 2(10), 2017.

doi: [10.1038/s41598-017-11257-2](https://doi.org/10.1038/s41598-017-11257-2)

**Publication reprinted with the permission of the copyright holders.**



# SCIENTIFIC REPORTS



OPEN

## Rate-limiting steps in transcription dictate sensitivity to variability in cellular components

Jarno Mäkelä<sup>1,4</sup>, Vinodh Kandavalli<sup>1</sup> & Andre S. Ribeiro<sup>1,2,3</sup>

Cell-to-cell variability in cellular components generates cell-to-cell diversity in RNA and protein production dynamics. As these components are inherited, this should also cause lineage-to-lineage variability in these dynamics. We conjectured that these effects on transcription are promoter initiation kinetics dependent. To test this, first we used stochastic models to predict that variability in the numbers of molecules involved in upstream processes, such as the intake of inducers from the environment, acts only as a transient source of variability in RNA production numbers, while variability in the numbers of a molecular species controlling transcription of an active promoter acts as a constant source. Next, from single-cell, single-RNA level time-lapse microscopy of independent lineages of *Escherichia coli* cells, we demonstrate the existence of lineage-to-lineage variability in gene activation times and mean RNA production rates, and that these variabilities differ between promoters and inducers used. Finally, we provide evidence that this can be explained by differences in the kinetics of the rate-limiting steps in transcription between promoters and induction schemes. We conclude that cell-to-cell and consequent lineage-to-lineage variability in RNA and protein numbers are both promoter sequence-dependent and subject to regulation.

Single-cell measurements have shown that, even in monoclonal bacterial populations, cells differ widely in component numbers<sup>1–6</sup>. Most cell-to-cell variability in, e.g. RNA and protein numbers, in the regime of low molecule numbers, can be explained by the stochastic nature of biochemical reactions. Meanwhile, in the high molecule numbers regime, most variability is due to cell-to-cell variability in the numbers of molecules involved in gene expression<sup>1</sup>.

Fluctuations in molecular species numbers in a cell propagate through direct and indirect interactions between species<sup>7,8</sup>. Also, noise from cellular processes such as DNA replication, and partitioning of molecules in cell division, also contribute significantly<sup>9,10</sup>. Importantly, these fluctuations have non-negligible timescales, often longer than cells' lifetime<sup>1,11,12</sup>, causing differences between sister cells to propagate to the timescale of cell lineages<sup>13–15</sup>.

Molecule number fluctuations likely affect most cellular processes. One process susceptible to these fluctuations is gene expression, as it depends on molecular species existing in small numbers (e.g. transcription factors) as well as on a cell's abundance of polymerases, ribosomes, and  $\sigma$  factors<sup>3,14–19</sup>.

At the single gene level, fluctuations in specific regulatory or uptake molecule numbers generate noise in the rates and timing of gene expression<sup>4,5,13</sup>. For example, gene expression activation rates by external inducers depend on the number of uptake membrane proteins<sup>5</sup>. As these differ in number between cells, so will intake times. Meanwhile, active transcription initiation rates (i.e. the main regulator of RNA production kinetics) differ due to, e.g., differences in the number of available RNA polymerases. It is expected that the effects of these noise sources in transcription will differ with the stage of gene expression affected.

Relevantly, the cell-to-cell variability in the kinetics of a chemical process depends not only on the variability in the numbers of the molecules involved, but also on the complexity of the process. For example, in a multi-step process such as transcription<sup>6,20–23</sup>, the degree to which the cell-to-cell variability in RNA polymerase numbers

<sup>1</sup>Laboratory of Biosystem Dynamics, BioMediTech Institute and Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, 33101, Tampere, Finland. <sup>2</sup>Multi-scaled biodata analysis and modelling Research Community, Tampere University of Technology, 33101, Tampere, Finland. <sup>3</sup>CA3 CTS/UNINOVA. Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516, Caparica, Portugal. <sup>4</sup>Present address: Department of Biochemistry, University of Oxford, South Parks Road, Oxford, OX1 3QU, UK. Correspondence and requests for materials should be addressed to A.S.R. (email: [andre.ribeiro@tut.fi](mailto:andre.ribeiro@tut.fi))

(or another molecule involved in the process) affects the RNA numbers' cell-to-cell variability, depends on the kinetics of all steps of the process. In particular, it is expected that only the duration of the first step (closed complex formation) will depend on the RNA polymerase numbers. As such, the larger the fraction of time in transcription initiation taken by the closed complex formation, the higher will be the effects of cell-to-cell variability in RNA polymerase numbers on the variability in RNA production kinetics. For example, if the closed complex formation takes only a small fraction of the overall duration of the process, even large deviations in its kinetics due to high variability in the numbers of the molecules involved (RNA polymerase, transcription factors, etc.) will not to cause major variability in the overall RNA production kinetics.

Thus, we hypothesize that promoters that differ in their sequence-dependent rate-limiting steps kinetics<sup>21, 23–26</sup>, will differ in their susceptibility to variability in molecule numbers. In addition, as the kinetics of the rate-limiting steps in transcription initiation are usually subject to regulation, e.g., by transcription factors<sup>21, 27, 28</sup>, we further hypothesize that the effects of cell-to-cell variability in molecule numbers can be tuned. Finally, as the time scale of fluctuations in molecule numbers and, thus cell-to-cell differences, can last longer than cell lifetimes and therefore propagate to cell lineages<sup>1, 12, 13</sup>, we expect that different promoters and different induction schemes will result in different lineage-to-lineage variability in RNA numbers.

To test these hypotheses, we combine stochastic modeling and time-lapse, single-cell, single-RNA level measurements of cell lineages to analyze the effects of variability in cellular components on transcription dynamics. Namely, we dissect the variability at each stage, from the external intake of inducers to the production of RNA molecules. For this, we first model transcription in cells accounting for the variability in numbers of the molecules involved in inducers intake and in transcription initiation rate constants, and study how these sources of variability contribute to the RNA variability over time. Next, to validate the model predictions, we measure differences in transcription dynamics between cell lineages. For this, we follow independent lineages for several generations under the microscope and measure RNA production in each lineage with single-cell, single-RNA sensitivity, to assess how the variability in gene activation rates following the introduction of inducers and in RNA production intervals in active promoters contribute to the lineage-to-lineage variability in RNA numbers over time. This variability is assessed and compared when inducing the same promoter,  $P_{lac/ara-1}$ , with different inducers (IPTG and arabinose), and when inducing different promoters ( $P_{lac/ara-1}$  and  $P_{lac}$ ) with the same inducer (IPTG). Finally, we use different inducer concentrations to regulate the kinetics of the rate-limiting steps in transcription initiation, and study how this can be used to tune the propagation of noise in cellular component numbers into RNA numbers.

## Results

**Cell-to-cell variability in cellular components are expected to generate cell-to-cell variability in gene activation times and in active transcription kinetics.** As in ref. 29, in each cell, we model gene activation and subsequent active transcription as stochastic multistep processes. Here, in addition, we impose that the rate of each step is dependent on the molecule number of specific molecular species (Fig. 1A and B). Specifically, the inducers' intake kinetics from the environment differs with the number of uptake proteins<sup>5</sup>, while the rate of closed complex formation in transcription initiation differs with the numbers of free RNA polymerases (RNAP), as most active promoters are not saturated with holoenzymes<sup>17, 30</sup>. Thus, in this model, the cell-to-cell variability in uptake protein and RNAP numbers affect the variability in gene activation and subsequent transcription initiation rates, respectively, thus contributing to the cell-to-cell variability in RNA numbers.

Gene activation is the passage of a promoter from a non-producing to a producing state, following the appearance of an inducer in the media. It includes subsequent events such as diffusion of inducers in the extracellular and intracellular environments, crossing of the cell membranes, and finding and binding to a promoter or its repressor.<sup>4, 31–33</sup> As these steps differ widely between genes, to model the dynamics of activation, we consider only the rate-limiting steps and model it as a two-step stochastic process as in refs 4, 29 (Supplementary Information):



Here,  $I_1$  is a promoter in a non-producing state,  $I_2$  is an intermediate state, and  $S_0$  is a producing state, in which the promoter is available for transcription.

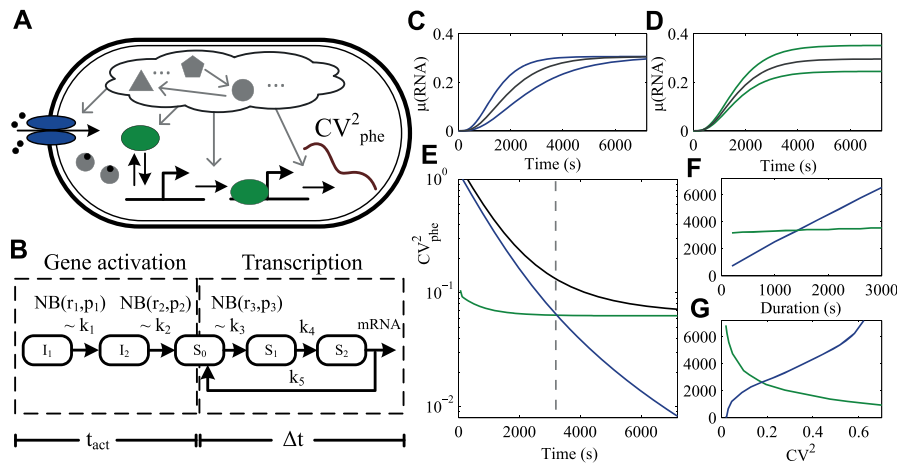
Active transcription in *E. coli* is a multi-step process, with the closed complex and open complex formation being, in most promoters, the most rate-limiting steps<sup>21–23</sup>. Transcription can thus be formulated as<sup>22</sup>:



In (2), transcription initiates when an RNA polymerase holoenzyme (RNAP) binds to a promoter (Pro) and forms a closed complex ( $\text{RP}_c$ ). This step is reversible and thus, it takes several attempts, until one of them eventually successfully forms a stable open complex ( $\text{RP}_o$ ). Finally, the holoenzyme forms an elongation complex and synthesizes an RNA. The first-passage time distribution to produce an RNA is observationally equivalent to the distribution generated by a simplified version of the models in (1) and (2), shown in Fig. 1B (Supplementary Information)<sup>26, 34</sup>.

Each model cell contains a number of uptake proteins and RNAPs that are drawn from negative binomial distributions of measured molecular species numbers' (Supplementary Information). To attain RNA production dynamics in each cell, we used the finite state projection algorithm<sup>35</sup>, in which a finite set of linear ordinary differential equations is formulated for the truncated state space of the system to predict the time-varying probability distributions. From this, we obtain the RNA number distribution of a cell population over time.

To quantify and compare the effects of cell-to-cell variability in uptake protein and RNAP numbers, the variability in RNA numbers is described as<sup>36</sup>:



**Figure 1.** *In Silico* prediction of variability in RNA numbers from variability in molecule numbers in gene activation and in active transcription. **(A)** Schematic representation of unspecified intracellular processes affecting the kinetics of gene activation by external inducers and subsequent transcription that generate cell-to-cell variability in RNA numbers over time ( $CV^2_{phe}$ ). **(B)** Gene activation (whose duration is represented by  $t_{act}$ ) is modeled as a stochastic 2-step process, while subsequent transcription events (whose overall duration is represented by  $\Delta t$ ) are modeled as a stochastic 3-step process. The rates  $k_1, k_2,$  and  $k_3$  are proportional to the molecule numbers drawn from negative binomial distributions. **(C, D)** show the resulting median (gray) and the quartiles (blue in **(C)** and green in **(D)**) of the RNA numbers over time in cells differing in **(C)** uptake molecule numbers or **(D)** RNAP numbers. **(E)**  $CV^2_{phe}$  resulting from differences in RNAP (green) or in uptake protein numbers (blue), and from differences in both (black). The dashed vertical line is the crossing time. From this figure, we find that cell-to-cell variability in uptake protein numbers contributes to RNA numbers diversity mostly at the early stages of a time series and then gradually dissipates, while noise in transcription is a constant source to RNA numbers diversity that dominates the latter stages of a time series. **(F, G)** show the effects on the crossing time of changing **(F)** the mean duration of the activation period (blue) and subsequent transcription events (green) and **(G)** the  $CV^2$  of uptake proteins (blue) and RNAP (green) numbers.

$$CV^2 = CV^2_{proc} + CV^2_{phe} \quad (3)$$

where

$$CV^2_{proc} = \frac{\overline{\langle n_i^2 \rangle} - \langle \langle n_i \rangle \rangle^2}{\langle \langle n_i \rangle \rangle^2}, \quad CV^2_{phe} = \frac{\overline{\langle n_i \rangle^2} - \langle \langle n_i \rangle \rangle^2}{\langle \langle n_i \rangle \rangle^2} \quad (4)$$

Here,  $n_i$  is the number of RNAs in cells of a sub-population of cells with parameter values  $i$  (i.e. number of uptake proteins and RNAPs); the bracket operator  $\langle (\cdot) \rangle$  represents averaging over all cells with parameter values  $i$ ; and the bar operator  $\overline{(\cdot)}$  represents averaging over all values of  $i$ .

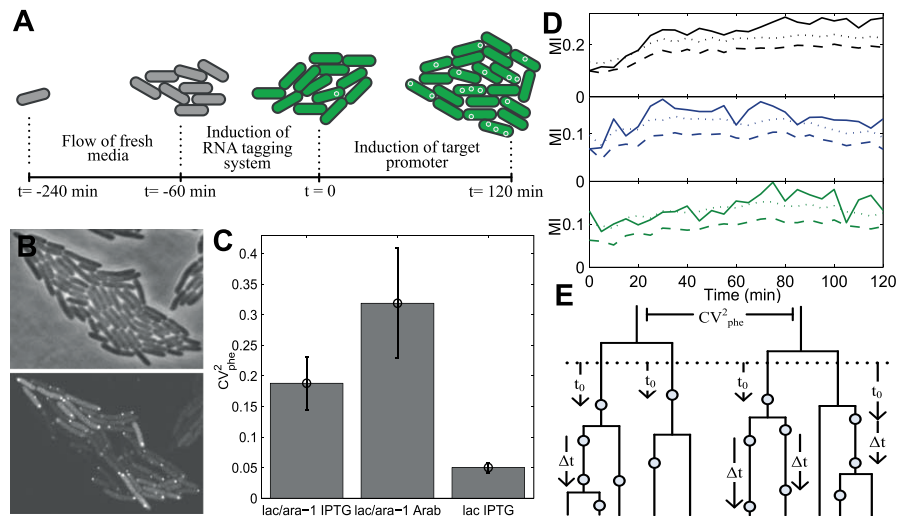
As the number of uptake proteins and RNAPs are the features that can differ between cells, they are used here as the features that define the ‘phenotype’ of a cell. Overall variability in RNA numbers is generated by the process’ stochasticity ( $CV^2_{proc}$ ) and by the differences in the cells’ propensities to produce RNAs ( $CV^2_{phe}$ ), due to ‘phenotypic’ variability.

Note that the kinetics of gene activation and transcription do not differ between the cells. Effects of variability in these processes were studied in<sup>26,29</sup>. Here, we focus on the effects of the ‘phenotypic’ variability ( $CV^2_{phe}$ ) on the kinetics of activation and active transcription.

First, we studied the effects of cell-to-cell variability solely in uptake protein numbers. For that, the model cells do not differ in RNAP numbers. From Fig. 1C and E, this source of variability contributes to RNA numbers diversity mostly at the early stages of a time series. Once transcription becomes active in most cells, the uniform process of RNA degradation across the cell population causes its effects to gradually dissipate.

Next, we assumed no variability in numbers of uptake proteins and studied the effects of variability in RNAP numbers. Here, the initial stages of the time series exhibit much less cell-to-cell variability in RNA numbers ( $CV^2_{phe}$ ) than the previous model. However, as transcription is activated throughout the cell population, its contribution to RNA numbers diversity becomes evident (Fig. 1D and E), being maximized when equilibrium is reached between RNA production and degradation.

Finally, we considered model cells where cell-to-cell diversity in both uptake protein and RNAP numbers are present. In these, in agreement with the above, the early stage of the time series is dominated by the variability in the gene activation process, while the latter stages are dominated by the variability in the transcription process



**Figure 2.** Variability in RNA production between lineages. (A) Cells are placed under the microscope at  $t = -240$  min and continuously supplemented with fresh medium. At  $t = -60$  min, the induction of the reporter system (MS2d-GFP) is initiated. At  $t = 0$  min, with the cells already flooded with MS2d-GFP proteins for accurate RNA detection, the induction of the target RNA for MS2d-GFP is initiated. (B) Phase contrast image of an induced lineage and corresponding fluorescence image with tagged RNA molecules. (C)  $CV^2_{\text{phe}}$  of the RNA numbers between lineages, 2 hours after induction. Shown are  $P_{\text{lac/ara-1}}$  induced with 1 mM IPTG (29 lineages) and with 1% Arabinose (14 lineages), and  $P_{\text{lac}}$  induced with 1 mM IPTG (60 lineages). Error bars are the standard errors determined by bootstrapping of the cells in the lineages (Fig. S3). Differences between conditions suggest that promoter sequence and transcription factors can regulate the  $CV^2_{\text{phe}}$  in RNA production. (D) MI (solid line), sMI (dashed line) and 1-tailed 0.01 p-value (dotted line) between a cell's lineage and the number of RNAs of each cell for  $P_{\text{lac/ara-1}}$  induced with 1 mM IPTG (black),  $P_{\text{lac/ara-1}}$  induced with 1% Arabinose (blue), and  $P_{\text{lac}}$  induced with 1 mM IPTG (green). In all conditions, the significant variability in the  $CV^2_{\text{phe}}$  in RNA numbers arises during the induction process. (E) Illustration of RNA production events (circles) over time in individual cells of lineages. The waiting times for the first RNAs to appear in lineages ( $t_0$ ) and the subsequent time intervals between consecutive RNA production events ( $\Delta t$ ) in single cells are shown. The dotted line depicts the start of induction of the target promoter.

(Fig. 1E). The moment when the latter overtakes the former is defined here as ‘crossing time’, and provides information about the duration of the influence from upstream processes. Importantly, the crossing time is often greater than a cell's generation time, as shown in previous studies<sup>4,29</sup>.

In addition, we quantified the dependence of the crossing time on the dynamics of activation and subsequent active transcription (Fig. 1F). We find that increasing the mean duration of gene activation increases the crossing time, as expected, while changing the active transcription initiation rate has only minimal effects. Also, the variability in RNAP and uptake protein numbers (measured by the  $CV^2$ ) affects the crossing time (Fig. 1G). Namely, increasing the  $CV^2$  of RNAP numbers decreases the crossing time, while increasing the  $CV^2$  of uptake protein numbers increases it.

**Variability in RNA numbers between lineages differs between promoters and their induction scheme.** *E. coli* cells have been shown to behave more similarly in protein production kinetics when sharing a common ancestor due to inheritable epigenetic factors<sup>13</sup>. These factors are propagated to the progeny for several generations<sup>1,11,12</sup>, and thus cell lineages are expected to differ in these factors.

Given this, here we consider each independent lineage as a distinct phenotype, with a specific RNA production rate and inducer intake rate. To validate this assumption, we studied how individual cell lineages respond to transcription induction by measuring, over the course of several generations, the RNA production in each cell with single molecule sensitivity following the introduction of an inducer in the media.

We grew lineages from individual cells under the microscope, induced the reporter and target gene, and then measured the RNA production dynamics in each cell once the lineages reached a size of 40–50 cells (Fig. 2A). All data of each condition is from the same experiment to avoid differences between overnight cultures, gel properties, etc. We detected production of RNA molecules by MS2-GFP tagging method (Fig. 2B, Fig. S1, and Supplementary Information), which protects the target RNA from degradation for the duration of the measurements<sup>37–39</sup>. Parameters for the detection of the target RNA were kept the same between lineages to avoid biases in detection.

Measurements were conducted for differing inducers and promoters. Namely, we used a single copy  $P_{lac/ara-1}$  (inducible by arabinose and/or IPTG)<sup>20</sup> and a single copy  $P_{lac}$  (inducible by IPTG)<sup>37</sup>. For  $P_{lac/ara-1}$  induced by 1 mM IPTG,  $P_{lac/ara-1}$  induced by 1% arabinose, and  $P_{lac}$  induced by 1 mM IPTG (in all cases for 2 hours), the cells exhibited, after 2 hours of induction, on average, 2.3, 0.4, and 3.0 RNAs, respectively, in agreement with previous *in vivo* measurements<sup>1,6</sup> (Supplementary Information, section 'RNA numbers in cells'). It is noted that the strain used here was modified to contain a very high copy number of lac repressors (~3000 vs. ~20 in wild type)<sup>20</sup> and to not code for lactose permease, which transports lactose into the cell. The first feature allows greatly increasing the fold change with induction when compared to the natural system. The second feature allows studying this system without the interference of feedback systems. In  $P_{lac/ara-1}$  promoter, the CRP/cAMP site has been replaced by the AraC binding sites of the  $P_{BAD}$  promoter to avoid pleiotropic effects and allow further activation of transcription<sup>20</sup>. Fig. S2 shows the topologies and sequences of the mentioned promoters.

To quantify the variability in RNA production dynamics between lineages, we obtained the  $CV^2_{phe}$  of the lineages in each condition (Fig. 2C, Fig. S3). We find differences between all conditions, indicating that possibly both the intake (which differs with the inducer molecule) and the active transcription (which differs with the promoter sequence) processes affect the  $CV^2_{phe}$  in RNA production of the lineages. Note that the  $CV^2_{phe}$  is independent of the mean transcription initiation rate (Fig. S4).

Due to being limited to observe a finite number of cells and lineages, it is possible that these values differ solely due to random chance. To test this, we measured the mutual information (MI)<sup>40</sup>, which quantifies how much a variable informs about another, between the lineage and the RNA numbers of each cell. For comparison, we randomly permuted cells between lineages for  $10^5$  times and calculated the average spurious MI (sMI), along with the 1-tailed p-value. The results are:  $P_{lac/ara-1}$  induced by IPTG (MI: 0.336, sMI: 0.258, p-value <  $10^{-5}$ );  $P_{lac/ara-1}$  induced by arabinose (MI: 0.138, sMI: 0.072, p-value <  $10^{-5}$ );  $P_{lac}$  induced by IPTG (MI: 0.185, sMI: 0.120, p-value <  $10^{-5}$ ). Thus, in all conditions, the hypothesis of having obtained the measured variability in RNA numbers between lineages by random chance can be rejected. Also, to test whether the difference between the MI and sMI increases during the activation period of transcription following the addition of inducers, we obtained the MIs for each condition every 5 min for 2 hours (Fig. 2D). Initially, the MI and sMI are very similar but, as time advances, the MI increases rapidly, becoming significantly above the average sMI (and 1-tailed p-value of 0.01) (see also mean values for lineages in Fig. S5).

To test for the possibility that the inducer was not reaching all cells under observation, we calculated the correlation between the distance between a cell and the colony edge and its RNA numbers. In all conditions, we found only very weak, not statistically significant, spatial correlations (Table S1), meaning that the induction is approximately uniform in space. Also, we tested for reproducibility of the lineage variability from independent measurements by conducting three independent measurements for cells with  $P_{lac/ara-1}$  induced by IPTG. We observed no statistically significant differences between the measurements (Figs S6 and S7).

We conclude that, in all conditions, the variability between lineages in mean RNA numbers is significantly above chance. Further, it differs with both the promoter, which should affect the kinetics of active transcription, as well as with the inducer, which should affect the kinetics of both intake and active transcription.

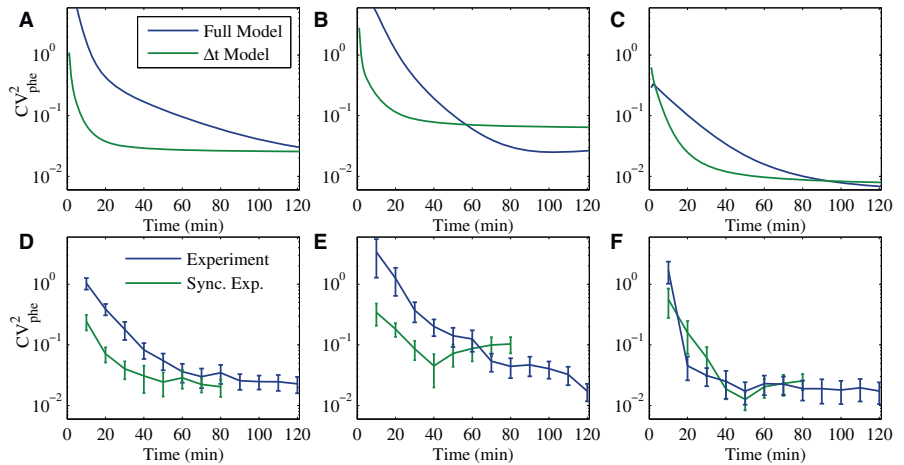
### Contributions of gene activation and active transcription to lineage variability differ over time, with the former being transient and the latter being a constant source of variability.

The observed lineage-to-lineage variability in RNA numbers can arise from gene activation, active transcription, or both. To assess the contribution of each process over time, we observed the waiting times for the first target RNA appearance ( $t_0$ ; which includes both  $t_{act}$  and  $\Delta t$ ) in each cell present at the start of induction<sup>29</sup>, along with the time intervals between consecutive RNA production events in each cell ( $\Delta t$ )<sup>29</sup> (Fig. 2E, Supplementary Information, Figs S1, S8 and S9). We extracted information from the same time-lapse experiment so as to minimize potential differences in environmental conditions. We also limited the observations to ~10 lineages per experiment to obtain sufficient time sampling. Results show that the  $CV^2_{phe}$  in both gene activation times and transcription intervals between lineages differs between conditions (Table S2).

To validate that the time series data are representative of large populations of lineages, we compared the lineage-to-lineage variability in mean RNA numbers of the time series measurements to that of two independent measurements for the condition of  $P_{lac/ara-1}$  induced by IPTG. We observed no statistically significant differences (Figs S6 and S7).

To estimate the contributions of each process to the observed lineage variability in RNA numbers over time, we fitted the measured  $t_0$  and  $\Delta t$  to the model of gene activation and transcription (Fig. 1B, Supplementary Information). We show results when assuming both activation ( $t_{act}$ ) and active transcription ( $\Delta t$ ) (referred to as 'full model'), and when assuming only active transcription (' $\Delta t$  model') (Fig. 3A–C). In all conditions, the  $\Delta t$  model reaches a plateau, i.e. a constant  $CV^2_{phe}$  faster than the full model. The height of this plateau is determined by the  $CV^2_{phe}$  of  $\Delta t$  and is independent of the mean transcription initiation rate (Fig. S4, Table S2). The two conditions that differ the most in the time to reach the plateau are  $P_{lac/ara-1}$  induced by IPTG and  $P_{lac/ara-1}$  induced by arabinose. Further, under arabinose induction, the  $CV^2_{phe}$  of the  $\Delta t$  model is initially higher, due to differences in the mean values of  $t_{act}$  and  $\Delta t$ . Over time, the two quantities will become similar (Fig. S10).

To compare with the model predictions, we calculated the empirical  $CV^2_{phe}$  in RNA numbers over time. For this, we only considered branches of lineages where RNA productions occurred. The outcomes of the full models are expected to be representative of these measurements. Meanwhile, to obtain empirical values comparable with the  $\Delta t$  model, we synchronized the first production moment of RNA in each lineage to  $t = 0$  and then disregarded that first production event. To avoid biases due to the reduced number of cells in the later parts of the time series, we only considered the first 80 minutes of the synchronized time series.



**Figure 3.** Lineages  $CV^2_{\text{phe}}$  in RNA numbers over time. (A–C)  $CV^2_{\text{phe}}$  in RNA numbers between lineages predicted for the full model (both  $t_0$  and  $\Delta t$  processes) and the  $\Delta t$  model over time. (A) model  $P_{\text{lac/ara-1}}$  induced with IPTG, (B) model  $P_{\text{lac/ara-1}}$  induced with arabinose and (C) model  $P_{\text{lac}}$  induced with IPTG. (D–F)  $CV^2_{\text{phe}}$  in RNA numbers of measured and synchronized (sync) lineages. Branches of lineages without RNA production are discarded and, in the sync data, the last 40 minutes are not used due to the need for synchronization (D)  $P_{\text{lac/ara-1}}$  induced with IPTG (15 lineages), (E)  $P_{\text{lac/ara-1}}$  induced with arabinose (10 lineages), and (F)  $P_{\text{lac}}$  induced with IPTG (8 lineages). Error bars are standard errors determined by bootstrapping of the lineages. As predicted by the models, in all cases, the contributions from gene activation kinetics and from active transcription dynamics to the  $CV^2_{\text{phe}}$  in RNA numbers differ over time, that the former has only a transient effect.

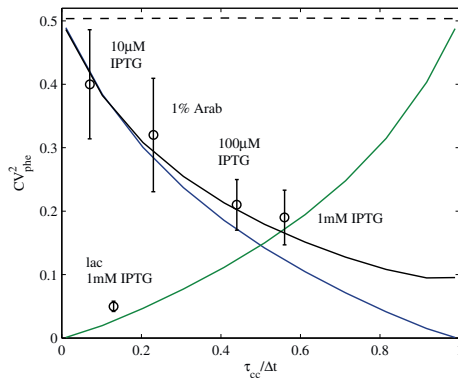
The empirical lineages  $CV^2_{\text{phe}}$  are shown for each condition, with and without synchronization (Fig. 3D–F). As predicted by the models, the  $CV^2_{\text{phe}}$  of the synchronized lineages exhibits a plateau. Also, in  $P_{\text{lac/ara-1}}$  the  $CV^2_{\text{phe}}$  of synchronized lineages reaches the plateau faster than the  $CV^2_{\text{phe}}$  of non-synchronized lineages. Meanwhile,  $P_{\text{lac}}$  does not exhibit significant influence by the gene activation process on the lineages'  $CV^2_{\text{phe}}$ . We expect that this is due to the higher leakiness of this promoter (Table S2). To test this notion, we studied the expected impact of leakiness on  $CV^2_{\text{phe}}$  using a model that allows transcription in the absence of inducers. This leakiness was modelled as a Poisson process, and various rates of leakiness were tested. The results show that increasing leakiness decreases the lineages'  $CV^2_{\text{phe}}$  (Fig. S11).

Overall, these results confirm that the contributions from gene activation kinetics and from active transcription dynamics to the lineages'  $CV^2_{\text{phe}}$  in RNA numbers differ over time, and that the former has only a transient effect. Importantly, fluctuations in transcription kinetics act as a constant source of variability in RNA numbers between lineages that differs between conditions (i.e. between promoters and between induction mechanisms of the same promoter).

**Rate-limiting steps in transcription regulate the effects of cell-to-cell variability in cellular components on transcription kinetics variability.** Why do the three conditions differ in variability between lineages ( $CV^2_{\text{phe}}$ ) in the same strain? Promoter sequences have been shown to differ widely in the kinetics of the rate-limiting steps in transcription initiation<sup>6,21,23,24</sup>. Also, depending on the molecular species whose numbers fluctuate, different stages of transcription are expected to be affected. For example, different transcription factors act at different stages and variability in their numbers affect mostly the variability in the kinetics of those stages alone.

Given this, we hypothesized that differences in the kinetics of the rate limiting steps as well as in which rate limiting steps are affected by differences in the numbers of transcription factors could be the source for the observed differences in  $CV^2_{\text{phe}}$  between the conditions studied here. Let  $\tau_{\text{cc}}$  represent the stages of transcription initiation whose kinetics depends on RNAP concentration, while  $\tau_{\text{oc}}$  represents subsequent stages, which are independent of RNAP concentration<sup>22,26,30,41</sup>. Given these definitions, we considered 4 different stochastic multi-step models of transcription (Fig. 1B) with the variability in molecule numbers affecting different rate-limiting steps: (1) variability in molecule numbers affecting only  $\tau_{\text{cc}}$ ; (2) variability in molecule numbers affecting only  $\tau_{\text{oc}}$ ; (3) variability in molecule numbers affecting both  $\tau_{\text{cc}}$  and  $\tau_{\text{oc}}$  equally; (4) variability in numbers of two molecular species (with different variabilities) affecting  $\tau_{\text{cc}}$  and  $\tau_{\text{oc}}$  independently. The extent of variability was set to be the same in all models ( $CV^2 = 0.5$ ) (except model 4, in which one molecular species has lower variability ( $CV^2 = 0.1$ )) to reflect the empirical values reported in<sup>7</sup>. The overall RNA production rate was identical in all cases and does not affect the  $CV^2_{\text{phe}}$  (Fig. S4). We studied the effects on  $CV^2_{\text{phe}}$  of RNA numbers as a function of  $\tau_{\text{cc}}$  relative to the overall duration of the transcription intervals,  $\Delta t$ .





**Figure 4.**  $CV^2_{\text{phe}}$  in RNA numbers as a function of  $\tau_{\text{cc}}/\Delta t$  as predicted by models and assessed by measurements. Lines are  $CV^2_{\text{phe}}$  from stochastic models with variability in molecule numbers affecting  $\tau_{\text{cc}}$  (green),  $\tau_{\text{oc}}$  (blue), and both simultaneously (dashed line), as a function of  $\tau_{\text{cc}}/\Delta t$ . Also shown is a model with variability in numbers of two molecular species (with different variabilities) affecting  $\tau_{\text{oc}}$  and  $\tau_{\text{cc}}$  (black). Circles are the measured lineages  $CV^2_{\text{phe}}$  as a function of  $\tau_{\text{cc}}/\Delta t$ .  $P_{\text{lac/ara-1}}$  induced with 10  $\mu\text{M}$  IPTG (61 lineages), 100  $\mu\text{M}$  IPTG (54 lineages), 1 mM IPTG (29 lineages), and 1% Arabinose (14 lineages). Also shown is  $P_{\text{lac}}$  induced with 1 mM IPTG (60 lineages). Error bars are standard errors determined by bootstrapping of the lineages. The same promoter subject to different induction levels influences its  $\tau_{\text{cc}}/\Delta t$  and will consequently differ in  $CV^2_{\text{phe}}$  in a way that is predictable by our model of transcription. Also, different transcription factors result in different  $CV^2_{\text{phe}}$ .

The results (Fig. 4) show that  $CV^2_{\text{phe}}$  varies with  $\tau_{\text{cc}}/\Delta t$  in models 1, 2, and 4, where the variability in molecule numbers affect  $\tau_{\text{cc}}$  and  $\tau_{\text{oc}}$  differently. In general, if variability in molecule numbers affects the longer lasting step, it results in higher  $CV^2_{\text{phe}}$  in RNA numbers. This does not occur in model 3, because the variability in molecule numbers affects both rate-limiting steps equally. Overall, we conclude that it is possible to tune the effects of variability in molecular species affecting transcription by tuning the ratio between the durations of the rate-limiting steps in transcription initiation.

To provide empirical validation, we first measured the extent to which  $\tau_{\text{cc}}/\Delta t$  of  $P_{\text{lac/ara-1}}$  can be tuned by varying the IPTG concentration, as it has been shown that the kinetics of the rate-limiting steps can be regulated by inducers<sup>21,26</sup>. The  $\tau_{\text{cc}}/\Delta t$  is obtained from  $\tau$ -plots, as in ref. 26. For that, the inverse of the RNA production rate is plotted as a function of inverse of the relative RNAP concentration. Next, it is extrapolated for an “infinite” RNAP concentration, so as to obtain the relative value of  $\tau_{\text{cc}}$  (Supplementary Information).

To alter RNAP concentrations in live cells, we used media with different concentrations of specific components, as described in ref. 26, and measured relative RpoC levels (i.e. the  $\beta'$  subunit, which is the limiting factor in the assembly of the RNAP holoenzyme) in each condition by Western Blotting (Fig. S12, Supplementary Information). Importantly, it has been shown by qPCR and plate reader measurements that the inverse of the RNA production rate of  $P_{\text{lac/ara-1}}$  change linearly with the inverse of the total RNAP concentration within the range of media richness used in our measurements<sup>26</sup>.

Next, we measured by qPCR the fold-change in RNA production rates in each media compared to the control condition. Following this,  $\tau_{\text{cc}}/\Delta t$  was extracted from the  $\tau$ -plot for each inducer condition (Fig. S13). Finally, for each condition, from microscopy measurements, we measured the lineages  $CV^2_{\text{phe}}$  in RNA numbers after 2 hours of induction.

We show (Fig. 4) the experimental lineages  $CV^2_{\text{ext}}$  for  $P_{\text{lac/ara-1}}$  for different IPTG concentrations (10  $\mu\text{M}$ , 100  $\mu\text{M}$ , and 1 mM) as a function of  $\tau_{\text{cc}}/\Delta t$ . Also shown are the results for  $P_{\text{lac/ara-1}}$  induced with 1% arabinose and  $P_{\text{lac}}$  induced with 1 mM IPTG. Notably, in  $P_{\text{lac/ara-1}}$ , as  $\tau_{\text{cc}}/\Delta t$  increases, the lineages  $CV^2_{\text{phe}}$  decreases. This behavior fits models 2 and 4, i.e., in this case the variability in molecule numbers influences mostly  $\tau_{\text{oc}}$ . Interestingly, in this regard, it is known that a bound lac repressor prevents open complex formation<sup>27</sup>. Similarly, AraC also affects the open complex formation<sup>21</sup>. This suggests that, in  $P_{\text{lac/ara-1}}$ , the cell-to-cell variability in lac repressor and AraC numbers might be the sources of the lineages  $CV^2_{\text{phe}}$  in RNA numbers.

$P_{\text{lac}}$  on the other hand, exhibits much lower lineages  $CV^2_{\text{phe}}$  (Fig. 4.) than those of  $P_{\text{lac/ara-1}}$  suggesting that its regulatory mechanisms and/or noise sources differ significantly from  $P_{\text{lac/ara-1}}$ . Congruently,  $P_{\text{lac}}$  has fewer LacI binding sites than  $P_{\text{lac/ara-1}}$  and a CAP binding site, which facilitates closed complex formation<sup>20,21,28,42</sup> (Fig. S2). As such,  $P_{\text{lac}}$  is expected to have different contributions to transcriptional variability from the transcription factor.

We conclude that transcription factors can be used to indirectly control the propagation of variability from molecular species numbers, given their ability to tune the kinetics of the rate-limiting steps in transcription initiation. In addition, we expect that different promoters, differing in regulatory mechanism and/or noise sources<sup>21–23</sup>, will differ in responsiveness to molecular fluctuations.

## Discussion

It is well-known that the variability in cellular components, particularly in core regulators of gene expression, such as RNA polymerases, transcription factors, and ribosomes does not affect all genes uniformly (see e.g. ref. 19). i.e., the resulting degree of phenotypic variability is known to be genetic-background dependent. However, the causes for this dependency remain unclear. Here, we provided one likely molecular mechanism responsible for the gene-specific phenotypic variability. In particular, we considered that gene expression is a multi-step process, that genes differ in the duration of each step, and that each step is affected differently by changes in the numbers of the core regulators. Based on this, we hypothesized that genes have unique, tunable levels of susceptibility to the variability in cellular components and, particularly, to variability in the core regulators numbers.

Moreover, as the molecular components affecting transcription are inherited, cell-to-cell variability in RNA numbers should result in lineage-to-lineage variability in the same numbers. Consequently, transcription dynamics diversity between cells should result in transcription dynamics diversity between lineages whose degree, similarly to the cell-to-cell diversity, should differ between genes and with induction schemes.

In support of our hypothesis, we first showed that the lineage-to-lineage variability in mean RNA numbers differs between promoters and when inducing the same promoter with different inducers. Also, we showed that the former is due to differences in initiation kinetics between promoters, while the latter is due to different inducers leading to different active transcription initiation kinetics.

Aside from these sources of lineage-to-lineage variability, which have a constant effect over time, we further showed that the process of gene activation by an inducer acts as a transient source. Namely, we showed that differences in the kinetics of inducer intake during gene activation causes tangible differences in the lineage-to-lineage variability in mean RNA numbers, which gradually dissipate as all cells of the lineages become activated.

Next, to support our hypothesis that differences in the kinetics of the rate-limiting steps in transcription initiation allow genes to be affected differently by fluctuations in the numbers of molecular species involved in transcription, we showed that changing the inducer or its concentration, which changes the initiation kinetics of a promoter, changes the lineage-to-lineage variability. Also, different promoters subject to the same inducer exhibit different lineage-to-lineage variability. In particular, we showed that a source acting on the first step alone will have weak effects on promoters where this step is relatively fast, but will have strong effects on promoters where this step is the most rate-limiting one. These results indicate that the effects of variability in molecular species in the dynamics of transcription at the single cell level are subject to regulation and, in agreement with previous studies<sup>7</sup>, are evolvable at the single gene level.

In this regard, it is of interest to mention a recent study showed that selection on expression noise can have a stronger impact on sequence variation than mean expression level<sup>43</sup>. As such, it is of importance to identify which mechanisms cells can use to evolve noise levels of individual genes. The main contribution of our study, aside from the direct quantification and better understanding of the degree of diversity in RNA production kinetics between cells and lineages, is the identification of a mechanism, namely, the multi-step nature of transcription initiation, that allows the effects of extrinsic noise sources to be tunable by transcription factors and by the promoter sequence, which makes it both adaptable and evolvable.

Given the substantial fluctuations and cell-to-cell diversity known to exist in cellular components in *E. coli* cells<sup>1</sup>, we expect the promoter-level sensitivity to molecule number fluctuations to be a key factor for a reliable dynamics of small genetic circuits and cellular functioning in general. Also, given the evolvability and adaptability of the kinetics of the rate-limiting steps of transcription initiation, we expect that *E. coli* is constantly adjusting these features at the single gene level in order to reach optimal levels of functioning. Namely, we expect a global reduction of cell-to-cell and lineage-to-lineage diversity in RNA numbers when in stable environments, and, following a bet-hedging strategy, its rapid enhancement when exploring new environments.

In addition to this, since, in general, the intake kinetics of gene expression regulators is itself subject to regulation, it may be that this and the above regulatory mechanisms act and evolve in a combined fashion. Variability in molecules responsible for gene activation and activity can be generalized as a “signaling” level of regulation in individual cells that can affect the response and sensitivity of the transcriptional circuits to perturbation. Importantly, the differences in the initiation kinetics of the promoters of a small circuit, should allow these circuits to exhibit ‘circuit state-dependent’ or signal-specific reactions. For example, consider a genetic switch where the initiation kinetics of promoter 1 is mostly spent in closed complex formation, while in promoter 2 it is mostly spent in open complex formation. In such a system, the outcome of fluctuations in RNA polymerase numbers (or transcription factors controlling closed complex formation) will depend on the switch’s present state. I.e. if the gene 2 is ‘ON’, the effects will be weak, but if it is gene 1 that is ‘ON’, the effects will be strong (more likely cause a switch in dynamics to occur). Future studies are needed to investigate how properties of genetic switches and genetic circuits are differentially sensitive to particular changes in the cellular composition.

Finally, we expect our results to be of value in the field of synthetic biology, which aims to engineer genetic networks with desired level of responsiveness to environmental cues by, among other, tuning the sensitivity to fluctuations in cellular component numbers at the single gene level. We expect our results to provide valuable information in this effort. For example, we believe that our results provide valuable clues on how to reduce present toggle switches<sup>244</sup> susceptibility to perturbations in cell physiology or in how to, alternatively, make the dynamics of a genetic circuit more responsive to changes in cellular physiology, in order to incorporate a cell’s current state into the circuit’s decision making process<sup>13</sup>.

## Materials and Methods

**Strains and plasmids.** Experiments were conducted in *E. coli* strain DH5 $\alpha$ -PRO, generously provided by I. Golding (Baylor College of Medicine, Houston, TX). It contains two genetic constructs: (a) pPROTet-K133 carrying P<sub>LacO1</sub>-MS2d-GFP, and (b) a single-copy F-based vector, pIG-BAC with a P<sub>lac/ara-1</sub> promoter controlling

the production of mRFP1 followed by a 96 MS2d binding site array ( $P_{lac/ara-1}$ -mRFP1-MS2d-96BS)<sup>37</sup>. We also use a modified system, with  $P_{lac}$  controlling the expression of an RNA with the 96 MS2d binding site array (named ' $P_{lac}$ -MS2d-96BS')<sup>42</sup>. Detailed information is provided in the supplementary information.

**Growth-conditions and microscopy.** Cells were grown overnight at 30 °C with aeration and shaking in lysogeny broth (LB) medium, supplemented with appropriate antibiotics, diluted 1:1000 fold into fresh LB medium and allowed to grow at 37 °C at 250 RPM until an optical density of  $OD_{600} \approx 0.3$ . Afterwards, a few  $\mu$ L of cells were placed between a 3% agarose gel pad and a glass coverslip, before assembling the FCS2 imaging chamber (Biopetechs). Cells were dispersed on the agarose gel pad, to give each the progeny of each cell enough space grow in numbers during the experiment. Prior to starting the experiment, the chamber was heated to 37 °C and placed under the microscope.

A flow of fresh (pre-warmed to 37 °C) LB medium containing the appropriate antibiotics was provided to cells under microscope observation by a peristaltic pump (Biopetechs) at a rate of  $0.5 \text{ mL min}^{-1}$ . At first, cells were perfused with media for ~4 hours to grow colonies from individual cells. Next, we perfused the cells with  $100 \text{ ng ml}^{-1}$  anhydrotetracycline (aTc) to induce  $P_{LtetO1}$  for MS2d-GFP production. Finally, after 1 hour (usually, at this stage, each colony, i.e. lineage, reached a size of ~40 cells), we perfused cells with 1 mM IPTG (or 1% L-arabinose) and  $100 \text{ ng ml}^{-1}$  aTc.

Cells were visualized in a Nikon Eclipse (Ti-E, Nikon) inverted microscope with C2+ (Nikon), a point scanning confocal microscope system, using a 100x Apo TIRF (1.49 NA, oil) objective. Fluorescence images were acquired using a 488 nm argon ion laser (Melles-Griot) and a 514/30 nm emission filter (Nikon). The fluorescence images were acquired once per minute during the last 2 hours of the microscopy measurements. The laser shutter was open only during the exposure time to minimize photobleaching. Meanwhile, an external phase contrast system (Nikon) was used with a DS-Fi2 CCD camera (Nikon) to obtain phase contrast images once per every 5 minutes. All images were acquired with NIS-Elements software (Nikon).

**Data and image analysis.** Data was analyzed using custom software written in MATLAB 2014a (MathWorks). Cells in phase contrast images were segmented using 'CellAging' (Fig. S1A)<sup>45</sup>. Alignment of the phase contrast images with the confocal images was done by selecting several landmarks in both images and using thin-plate spline interpolation for the registration transform. Fluorescent MS2d-GFP-RNA spots in each cell, at each frame, were detected with the Kernel Density Estimation (KDE) method using a Gaussian kernel (Fig. S1B)<sup>46</sup>. Cell background corrected spot intensities were then calculated by subtracting the mean cell background intensity multiplied by the area of the spots from the total fluorescence intensity of the spots. RNA numbers of individual cells at the different time moments as in<sup>37</sup>. From the distribution of background-corrected total spots intensity in cells, the first peak is set to correspond to the intensity of a single RNA molecule and the number of tagged RNAs in each spot is estimated by dividing its intensity by that of the first peak (Fig. S1C, Supplementary Information). To calculate the waiting times for the first production, the time intervals between consecutive production events and the total number of production events in lineages, the background-corrected total spots intensity over time in each cell was fitted to a monotone piecewise-constant function by least squares<sup>46</sup>. The number of terms was selected using the F-test with a p-value of 0.01. Each jump corresponds to the production of a single RNA (Fig. S1D). This method relies on the fact that, once tagged with MS2d-GFP, the RNA does not degrade and its fluorescence does not decay for several hours<sup>39</sup>. Waiting times for the first production of RNAs in each lineage were calculated by selecting cells without spots at the beginning of induction (i.e., without leaky expression), and detecting when the first production occurred in each branch of each lineage. Time intervals between consecutive RNA productions in individual cells were obtained by extracting the time between consecutive jumps in the total spots intensity (Fig. S1)<sup>46</sup>.

## References

1. Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).
2. Bakshi, S., Siryaporn, A., Goulian, M. & Weisshaar, J. C. Superresolution imaging of ribosomes and RNA polymerase in live Escherichia coli cells. *Mol. Microbiol.* **85**, 21–38 (2012).
3. Yang, S. *et al.* Contribution of RNA polymerase concentration variation to protein expression noise. *Nat. Commun.* **5**, 1–9 (2014).
4. Megerle, J. A., Fritz, G., Gerland, U., Jung, K. & Rädler, J. O. Timing and dynamics of single cell gene expression in the arabinose utilization system. *Biophys. J.* **95**, 2103–2115 (2008).
5. Choi, P. J., Cai, L., Frieda, K. & Xie, X. S. A stochastic single-molecule event triggers phenotype switching of a bacterial cell. *Science* **322**, 442–446 (2008).
6. Jones, D. L., Brewster, R. C. & Phillips, R. Promoter architecture dictates cell-to-cell variability in gene expression. *Science* **346**, 1533–1537 (2014).
7. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
8. Paulsson, J. Models of stochastic gene expression. *Phys. Life Rev.* **2**, 157–175 (2005).
9. Huh, D. & Paulsson, J. Non-genetic heterogeneity from stochastic partitioning at cell division. *Nat. Genet.* **43**, 95–100 (2011).
10. Peterson, J. R., Cole, J. A., Fei, J., Ha, T. & Luthey-Schulten, Z. A. Effects of DNA replication on mRNA noise. *Proc. Natl. Acad. Sci. USA* **112**, 15886–15891 (2015).
11. Hensel, Z. *et al.* Stochastic expression dynamics of a transcription factor revealed by single-molecule noise analysis. *Nat. Struct. Mol. Biol.* **19**, 797–802 (2012).
12. Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* **307**, 1962–1965 (2005).
13. Robert, L. *et al.* Pre-dispositions and epigenetic inheritance in the Escherichia coli lactose operon bistable switch. *Mol. Syst. Biol.* **6**, 357 (2010).
14. Kiviet, D. J. *et al.* Stochasticity of metabolism and growth at the single-cell level. *Nature* **514**, 376–379 (2014).
15. Yun, H. S., Hong, J. & Lim, H. C. Regulation of Ribosome Synthesis in Escherichia coli Effects of Temperature and Dilution Rate Changes. *Biotechnol. Bioeng.* **52**, 615–624 (1996).

16. Klumpp, S., Zhang, Z. & Hwa, T. Growth Rate-Dependent Global Effects on Gene Expression in Bacteria. *Cell* **139**, 1366–1375 (2009).
17. Liang, S. *et al.* Activities of constitutive promoters in Escherichia coli. *J. Mol. Biol.* **292**, 19–37 (1999).
18. Bremer, H. & Dennis, P. Modulation of chemical composition and other parameters of the cell by growth rate. *Neidhardt, F. (ed.). Washington, DC Am. Soc. Microbiol. Press* 1553 (1996).
19. Kandavalli, V. K., Tran, H. & Ribeiro, A. S. Effects of  $\sigma$  factor competition are promoter initiation kinetics dependent. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1859**, 1281–1288 (2016).
20. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-2 regulatory elements. *Nucleic Acids Res.* **25**, 1203–1210 (1997).
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.* **29**, 3873–3881 (2001).
22. McClure, W. R. Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.* **54**, 171–204 (1985).
23. Saecker, R. M., Record, M. T. & DeHaseth, P. L. Mechanism of Bacterial Transcription Initiation: RNA Polymerase - Promoter Binding, Isomerization to Initiation-Competent Open Complexes, and Initiation of RNA Synthesis. *J. Mol. Biol.* **412**, 754–771 (2011).
24. McClure, W. R. Rate-limiting steps in RNA chain initiation. *Proc. Natl. Acad. Sci. USA* **77**, 5634–5638 (1980).
25. Friedman, L. J. & Gelles, J. Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation. *Cell* **148**, 679–689 (2012).
26. Lloyd-Price, J. *et al.* Dissecting the stochastic transcription initiation process in live Escherichia coli. *DNA Res.* **23**, 203–214 (2016).
27. Sanchez, A., Osborne, M. L., Friedman, L. J., Kondev, J. & Gelles, J. Mechanism of transcriptional repression at a bacterial promoter by analysis of single molecules. *EMBO J.* **30**, 3940–3946 (2011).
28. Busby, S. & Ebright, R. H. Transcription activation by catabolite activator protein (CAP). *J. Mol. Biol.* **293**, 199–213 (1999).
29. Mäkelä, J. *et al.* *In vivo* single-molecule kinetics of activation and subsequent activity of the arabinose promoter. *Nucleic Acids Res.* **41**, 6544–6552 (2013).
30. Ehrenberg, M., Bremer, H. & Dennis, P. P. Medium-dependent control of the bacterial growth rate. *Biochimie* **95**, 643–658 (2013).
31. Schleif, R. Regulation of the L-arabinose operon of Escherichia coli. *Trends Genet.* **16**, 559–565 (2000).
32. Skerra, A. Use of the tetracycline promoter for the tightly regulated production of a murine antibody fragment in Escherichia coli. *Gene* **151**, 131–135 (1994).
33. Weickert, M. J. & Adhya, S. The galactose regulon of Escherichia coli. *Mol. Microbiol.* **10**, 245–251 (1993).
34. Moffitt, J. R. & Bustamante, C. Extracting signal from noise: Kinetic mechanisms from a Michaelis-Menten-like expression for enzymatic fluctuations. *FEBS J.* **281**, 498–517 (2014).
35. Minsky, B. & Khammash, M. The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.* **124**, 44104 (2006).
36. Lu, T., Shen, T., Bennett, M. R., Wolynes, P. G. & Hasty, J. Phenotypic variability of growing cellular populations. *Proc. Natl. Acad. Sci. USA* **104**, 18982–18987 (2007).
37. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell* **123**, 1025–1036 (2005).
38. Peabody, D. S. The RNA binding site of bacteriophage MS2 coat protein. *EMBO J.* **12**, 595–600 (1993).
39. Tran, H., Oliveira, S. M. D., Goncalves, N. & Ribeiro, A. S. Kinetics of the cellular intake of a gene expression inducer at high concentrations. *Mol. Biosyst.* **11**, 2579–2587 (2015).
40. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**(379–423), 623–656 (1948).
41. Patrick, M., Dennis, P. P., Ehrenberg, M. & Bremer, H. Free RNA polymerase in E. coli. *Biochimie* **119**, 80–91 (2015).
42. Golding, I. & Cox, E. C. RNA dynamics in live Escherichia coli cells. *Proc. Natl. Acad. Sci. USA* **101**, 11310–11315 (2004).
43. Metzger, B. P. H., Yuan, D. C., Gruber, J. D., Duveau, F. & Wittkopp, P. J. Selection on noise constrains variation in a eukaryotic promoter. *Nature* **521**, 344–347 (2015).
44. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403**, 339–342 (2000).
45. Häkkinen, A., Muthukrishnan, A.-B., Mora, A., Fonseca, J. M. & Ribeiro, A. S. CellAging: a tool to study segregation and partitioning in division in cell lineages of Escherichia coli. *Bioinformatics* **29**, 1708–9 (2013).
46. Häkkinen, A. & Ribeiro, A. S. Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data. *Bioinformatics* **31**, 69–75 (2015).

## Acknowledgements

Work supported by Academy of Finland (295027 and 305342 to ASR), Jane and Aatos Erkkö Foundation (610536 to ASR), and TUT President's Graduate Programme (JM).

## Author Contributions

J.M. and A.S.R. conceived the study. J.M. and V.K. performed the microscopy experiments. J.M. performed the modeling and analysis. V.K. executed qPCR and Western Blotting. All authors performed research. J.M. and A.S.R. drafted the manuscript which was revised by all authors.

## Additional Information

**Supplementary information** accompanies this paper at doi:10.1038/s41598-017-11257-2

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017

# **Supplementary Material for: “Rate-limiting steps in transcription dictate sensitivity to variability in cellular components”**

Jarno Mäkelä<sup>1,2</sup>, Vinodh Kandavalli<sup>1</sup> and Andre S. Ribeiro<sup>1,3,4,\*</sup>

<sup>1</sup>Laboratory of Biosystem Dynamics, BioMediTech Institute and Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, 33101, Tampere, Finland.

<sup>2</sup>Present address: Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK.

<sup>3</sup>Multi-scaled biodata analysis and modelling Research Community, Tampere University of Technology, 33101, Tampere, Finland.

<sup>4</sup>CA3 CTS/UNINOVA. Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2829-516, Caparica, Portugal.

\*Correspondence: [andre.ribeiro@tut.fi](mailto:andre.ribeiro@tut.fi)

Keywords: gene expression, in vivo single-RNA detection, lineage-to-lineage variability, rate-limiting steps in transcription

## SI Materials and Methods

### *Strains and plasmids*

The strain information of *E. coli* DH5 $\alpha$ -PRO, generously provided by I. Golding (Baylor College of Medicine, Houston, TX) is: deoR, endA1, gyrA96, hsdR17(rK- mK+), recA1, relA1, supE44, thi-1,  $\Delta$ (lacZYA-argF)U169,  $\Phi$ 80 $\delta$ lacZ $\Delta$ M15, F-,  $\lambda$ -, PN25/tetR, PlacIq/lacI, and SpR. It contains two genetic constructs: (a) pPROTet-K133 carrying P<sub>LtetO1</sub>-MS2d-GFP, and (b) a single-copy F-based vector, pIG-BAC with a P<sub>lac/ara-1</sub> promoter controlling the production of mRFP1 followed by a 96 MS2d binding site array (P<sub>lac/ara-1</sub>-mRFP1-MS2d-96BS) (Golding et al. 2005). We also use a modified system, with P<sub>lac</sub> controlling the expression of an RNA with the 96 MS2d binding site array (named 'P<sub>lac</sub>-MS2d-96BS')(Golding & Cox 2004). It was implemented in the same strain and uses the same reporter.

The strain produces necessary regulatory proteins for these constructs, namely, LacI, TetR and AraC, from the chromosome (Lutz & Bujard 1997). The MS2d-GFP single RNA detection system has been shown to detect individual target RNAs a few seconds after their transcription, provided that sufficient MS2d-GFP proteins are present in the cells (Golding & Cox 2004). Once tagged with MS2d-GFP, the RNA molecules do not degrade and their fluorescence does not decay significantly for a few hours (Tran et al. 2015). Also, it was shown that, in standard time-lapse microscopy measurements with consecutive images separate by 1 minute intervals, once appearing, each tagged RNA spot already exhibits 'full' fluorescence (Tran et al. 2015).

### *RNA numbers in cells*

Estimation of RNA numbers in cells from the distribution of background-corrected total spots intensity in cells is only accurate if cells produce a small number of RNAs (note that the variability of each peak is expected to double from one peak to the next). We applied it here since this condition holds (see main manuscript). In this regard, these numbers are in agreement with several previous works. E.g., recently, RNA production *in vivo* from wild-type (WT) P<sub>lac</sub>, P<sub>lacUV5</sub> and a library of synthetic promoters was measured at the single molecule level using fluorescence *in situ* hybridization (FISH) (Jones et al. 2014). The mean mRNA numbers, under constitutive expression, varied between 0.04 and 10 per cell. E.g., WT P<sub>lac</sub> exhibited a mean

mRNA number per cell of 0.4, while the stronger  $P_{lacUV5}$  had a mean RNA numbers per cell of 10 (Jones et al. 2014). Furthermore, it has been shown that the mean mRNA numbers per cell at the transcriptome level ranges from  $10^{-4}$  to 10 mRNA per cell (Taniguchi et al. 2010). We find our measured RNA numbers to be in full accordance with these results.

### ***qPCR***

Target gene quantification was also done by qPCR. Cells containing the target plasmid were grown and induced with the respective inducers (1% arabinose for  $P_{lac/ara-1}$ -mRFP1-96BS, 10 $\mu$ M IPTG, 100 $\mu$ M IPTG and 1mM IPTG for  $P_{lac/ara-1}$ -mRFP1-96BS and 1mM IPTG for  $P_{lac}$ -lacZ-96BS) as described in the methods, and cells were harvested by centrifuging them at 8000 $\times$ g for 5 minutes. To the pelleted cells, twice the amount of RNA protect reagent (Qiagen) was added and followed by enzymatic lysis with Tris EDTA lysozyme buffer (pH 8.0). The total RNA was isolated by the RNeasy kit (Qiagen) according to the manufacturer instructions and the concentration of RNA was quantified by a Nanovue plus spectrophotometer (GE Healthcare). The RNA samples were treated with DNase to remove residual DNA, followed by cDNA synthesis using the iSCRIPT reverse transcription super mix. The cDNA samples were mixed with the qPCR master mix containing iQ SYBR Green supermix (Biorad) with primers for the target and reference genes. The reaction was carried out in triplicates with a total reaction volume of 20  $\mu$ l. For quantifying the target gene, we used mRFP1 primers (Forward: 5' TACGACGCCGAGGTCAAG 3' and Reverse: 5' TTGTGGGAGGTGATGTCCA 3'), and lacZ primers (Forward: 5' CCGGATCCTCGAGAGCTTAG 3' and Reverse: 5' CTAATCGATTCAATTGGGTAACG 3'). For the reference gene, we used 16S RNA primers (Forward: 5' CGTCAGCTCGTGTGTGAA 3' and Reverse: 5' GGACCGCTGGCAACAAAG 3') were used. The qPCR experiments were performed by a MiniOpticon Real time PCR system (Biorad). The following conditions were used during the reaction: 40 cycles of 95°C for 10 s, 52°C for 30 s and 72°C for 30 s for each cDNA replicate. We used no-RT controls and no-template controls to crosscheck non-specific signals and contamination. PCR efficiencies of these reactions were greater than 95%. The data from CFX Manager TM Software was used to calculate the relative gene expression and its standard error (Livak & Schmittgen 2001).

### ***Western Blotting***

Cultures of *E. coli* DH5 $\alpha$ -PRO strain were grown in media of different richness (“1x”, “0.5x” and “0.25x”), as described in materials and methods. Cells were harvested at OD<sub>600</sub> of 0.3 and lysed with the B-PER bacterial protein extraction reagent (Thermo scientific) in the presence of protease inhibitors for 10 min. Subsequently, the lysed cells were centrifuged at 15000 $\times$ g for 10 mins, supernatants collected, diluted in the 4X laemmli sample loading buffer containing  $\beta$ -mercaptoethanol and boiled for 5 mins at 95 °C. The samples from all the cultures, each containing ~30  $\mu$ g of total soluble proteins, were resolved by 4 to 20 % TGX stain free precast gels (Biorad). Proteins were separated by electrophoresis and then electro-transferred to the PVDF membrane. Membranes were blocked with 5% non-fat milk and incubated with respective primary RpoC antibodies of 1:2000 dilutions (Biolegend) overnight at 4 °C, followed by the appropriate HRP-secondary antibodies 1:5000 dilutions (Sigma Aldrich) for 1 h at room temperature. For detection, chemilumiscence reagent (Biorad) was used. Images were generated by the Chemidoc *XRS* system (Biorad) (Fig. S12). Band intensity quantification was done by the Image lab software (version 5.2.1).

### ***$\tau$ -plot***

Mean transcription rates (i.e. mean RNA production rates) depend on the free RNAP concentration of the cells (McClure 1985; Liang et al. 1999; Ehrenberg et al. 2013). These concentrations can be tuned by altering media composition in a specific manner (Liang et al. 1999; Patrick et al. 2015). This was achieved in (Lloyd-Price et al. 2016) by modifying components of LB media. It was further shown that in a certain range of media compositions, the relative free RNAP concentration can be assessed from the total RNAP concentration, because it varies in a linear fashion with the changes in media composition (Lloyd-Price et al. 2016). Also, no evidence for factors other than the free RNAP concentration affecting the rate of production from target promoter was found. Following this methodology, we used media compositions per 100 ml as follows: (“0.25x” condition) 0.25 g tryptone, 0.125 g yeast extract and 1 g NaCl (pH 7.0); (“0.5x” condition) 0.5 g tryptone, 0.25 g yeast extract and 1 g NaCl (pH 7.0); (“1x” condition) 1 g tryptone, 0.5 g yeast extract and 1 g NaCl (pH 7.0). The relative RNAP concentrations in each condition were assessed by measuring the level of the RpoC protein by Western blot (Fig. S12). These measurements confirmed that the relative RNAP levels change



linearly with media compositions, as reported in (Lloyd-Price et al. 2016). For each condition, the target RNA production rates in different media were measured by qPCR.

Next, based on the premise that the RNAP concentration only affects the duration of closed complex formation but not the duration of the subsequent rate-limiting steps (Liang et al. 1999; Lloyd-Price et al. 2016), we extracted the ratio between the RNAP-dependent fraction ( $\tau_{cc}$ ) of the mean duration of the time intervals between transcription events and the overall mean duration of the time intervals between transcription events ( $\Delta t$ ),  $\tau_{cc}/\Delta t$ , following the methodology proposed in (Liang et al. 1999; Lloyd-Price et al. 2016). I.e., the inverse of RNA production rates were plotted against the inverse of relative RNAP levels and fitted by a line using weighted total least squares (Krystek & Anton 2007). From this fit,  $\tau_{cc}/\Delta t$  was estimated by extrapolating the inverse of RNA production rate to an infinite RNAP concentration (Lloyd-Price et al. 2016; Liang et al. 1999; Patrick et al. 2015). The results are shown in (Fig. S13).

### ***Stochastic model of gene activation and transcription***

The stochastic model considers both gene activation following the appearance of inducers in the media, and transcription following the activation step. Gene activation is the process by which a gene that is in a non-producing state enters a producing state, via a multi-step process. This process includes events such as diffusion of the activator molecules in the periplasm and cytoplasm, binding to a transcription factor, protein-protein interactions etc. These events differ with the induction system of each particular gene (Schleif 2000; Megerle et al. 2008; Skerra 1994; Weickert, M.J. and Adhya 1993).

In the case of the promoters studied in this work, the waiting times for gene activation by the external inducers have been measured at the single cell level and shown to exhibit dynamics of activation of the target gene that can be well modelled by a 2-step stochastic process (Fig. 1B) (Mäkelä et al. 2013; Megerle et al. 2008; Fritz et al. 2014; Tran et al. 2015). The first-passage time distribution, which corresponds to the total time spent in each of the states of the process, can be thus described by a general model of the form (Mäkelä et al. 2013; Moffitt & Bustamante 2014):



Here,  $I_1$  is the non-producing state of the system,  $I_2$  is an intermediate state, and  $S_0$  is the producing state of the system, in which the promoter is available for transcription. These

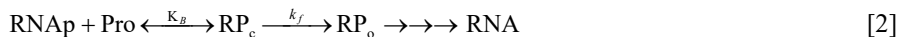
reactions occur at rates  $k_1$  and  $k_2$ , respectively, and are both catalyzed by an uptake protein (Upt). These rates correspond to the total time spent on each state (Moffitt & Bustamante 2014). It is noted that the number of uptake proteins is set to affect the rates of both steps, as the shape of the distribution was found not to change with inducer concentration (Megerle et al. 2008).

Note that the dependence of the reactions on the numbers of Upt proteins allows the model to exhibit cell-to-cell diversity in the kinetics of gene activation, provided that these numbers differ between cells (similarly, the dynamics could differ over time).

For parameter values, we made use of measurements of the arabinose utilization system, which has been reported to take, on average,  $\sim 23$  min to activate each cell, with a standard deviation of  $\sim 10$  min (Megerle et al. 2008). As an independent validation, we used a plate reader to measure the production kinetics over time following induction (Fig. S9). Given the maturation time of RFP1 of 0.7 h (Campbell et al. 2002), the results are in agreement with the single RNA measurements.

It is worth noting that previous studies on gene activation (Johnson & Schleif 1995; Daruwalla et al. 1981) reported faster kinetics than those measured here and also reported in (Mäkelä et al. 2013). This is likely due to several reasons. Namely, as mentioned in the main manuscript, first, different strains were used. E.g., DH5 $\alpha$ -PRO, the strain used here, was modified to contain a very high copy number of lac repressors ( $\sim 3000$  vs.  $\sim 20$  in wild type) (Lutz & Bujard 1997). Second, in the case of the  $P_{lac/ara-1}$  promoter, note that our cells do not code for lactose permease, which transports lactose into the cell. Finally, for the case of induction with arabinose, we do not de-repress the promoter with IPTG, which is expected to delay RNA production significantly. We note that, aside from the reduced speed in RNA production, we do not expect these differences to cause additional significant functional differences.

Next, we describe the process of active transcription also included in the model. In *E. coli*, this process consists of a sequence of steps, with the formations of the closed complex and open complex being, in most promoters, the rate-limiting ones (McClure 1985; Saecker et al. 2011; Lutz et al. 2001). Transcription, as a dynamic process, can thus be formulated as (McClure 1985):



where transcription initiates by RNA polymerase holoenzyme (RNAP) binding to a promoter (Pro) and forming the closed complex (RP<sub>c</sub>). This step is usually reversible. Following several

attempts, the holoenzyme will eventually succeed in opening the DNA strands, thus creating a transcription bubble, and assemble the polymerase clamp through several intermediate steps to form a stable open complex (RP<sub>o</sub>). Finally, the holoenzyme will form an elongation complex and synthesize the nascent RNA molecule.

From [2], it is possible to extract a time interval distribution between transcription events ( $\Delta t$ ). For this, we use the fact that the first-passage time distribution to produce an RNA is observationally equivalent to the distribution described by a model of the form (Lloyd-Price et al. 2016; Moffitt & Bustamante 2014):



Here,  $S_0$  is a state in which the promoter is available for transcription (following induction). Transition to state  $S_1$  occurs at the rate  $R \cdot k_3$  ( $R$  being the number of RNAP molecules) and, following this, transition to state  $S_2$  occurs at the rate  $k_4$ . Finally, an RNA is produced and the promoter returns to state  $S_0$  at the rate  $k_5$ . RNA degradation is modelled as an exponential process with a rate of  $5 \text{ min}^{-1}$  (Bernstein et al. 2002; Chen et al. 2015). Models of this form have been shown to fit recent *in vivo* measurements of  $\Delta t$  at the single RNA level (see e.g. (Lloyd-Price et al. 2016; Tran et al. 2015)).

A recent study has quantified that, for  $P_{lac/ara-1}$ , the RNAP-dependent fraction of time of transcription initiation ( $R^{-1} \cdot k_3^{-1}$ ) lasts  $\sim 788 \text{ s}$ , while the non-RNAP dependent fraction lasts  $\sim 193 \text{ s}$  ( $k_4^{-1}$ ) (Lloyd-Price et al. 2016). The RNAP dependent stage of initiation ( $R^{-1} \cdot k_3^{-1}$ ) includes the reversible closed complex formation and transcriptionally inactive promoter states, which occur, e.g., due to binding and unbinding of the repressor (Lutz et al. 2001) and accumulation of negative supercoiling in the DNA (Chong et al. 2014).

Meanwhile, the steps following open complex formation have been found to be fast (here, this is modeled by setting  $k_5 = \infty$ ), indicating that abortive initiation events do not play a major role in the dynamics of RNA production in  $P_{lac/ara-1}$  (Lloyd-Price et al. 2016). This is expected since only in rare promoters, whose open complexes exhibit extremely short half-lives, is promoter escape expected to be rate-limiting (Hsu 2002).

In (Taniguchi et al. 2010), a global characterization of cell-to-cell variability in protein numbers showed a noise limit that is independent of the mean. The distribution of protein

numbers in a population was found to be well fitted by a discrete negative binomial distribution (Taniguchi et al. 2010). Here, we model the variability in uptake protein numbers (for the induction process) and RNAP (for the transcription process) taking this into account. Parameter values for the negative binomial distribution for the RNAP variability ( $CV^2 = 0.1$ ) were obtained from (Jones et al. 2014; Taniguchi et al. 2010) and for the uptake protein ( $CV^2 = 0.27$ ) from (Megerle et al. 2008).

The variability in RNAP numbers affects the rate of closed complex formation (McClure 1980; McClure 1985; Saecker et al. 2011). The variability in uptake protein numbers affects the rates of both steps in initiation, as the shape of the distribution does not change with inducer concentration (Megerle et al. 2008). Since fluctuations in protein numbers were shown to have a time scale of several cell cycles (up to 5 hours) (Taniguchi et al. 2010; Hensel et al. 2012; Rosenfeld et al. 2005), we assume fixed protein numbers for each cell in the models (but differing between cells as noted above).

### ***CME solution***

To predict the time-varying probability distributions from the models, we make use of a direct integration of the Chemical Master Equation (CME) of the sum of  $d$ -exponential variates model for gene activation and transcription using the Finite State Projection algorithm (Munsky & Khammash 2006). This method truncates the infinite state space of the CME, so that the amount of probability outside the truncated region is negligible, and formulates a finite set of linear ordinary differential equations for each possible state of the system. The state space was truncated at 100 RNA molecules. This means that this space contains virtually all of the total probability in the system (we never observed a cell to have more than 20 RNAs). The probability mass vector at each time moment is then obtained for all phenotypes. Next, the population distribution is obtained by utilizing the negative binomial distribution to assign weight for each combination of molecule numbers. From this distribution, we calculate mean and variance of RNA molecules between phenotypes at each time moment.

### ***Fitting empirical distributions to a sum of $d$ -exponential variates***

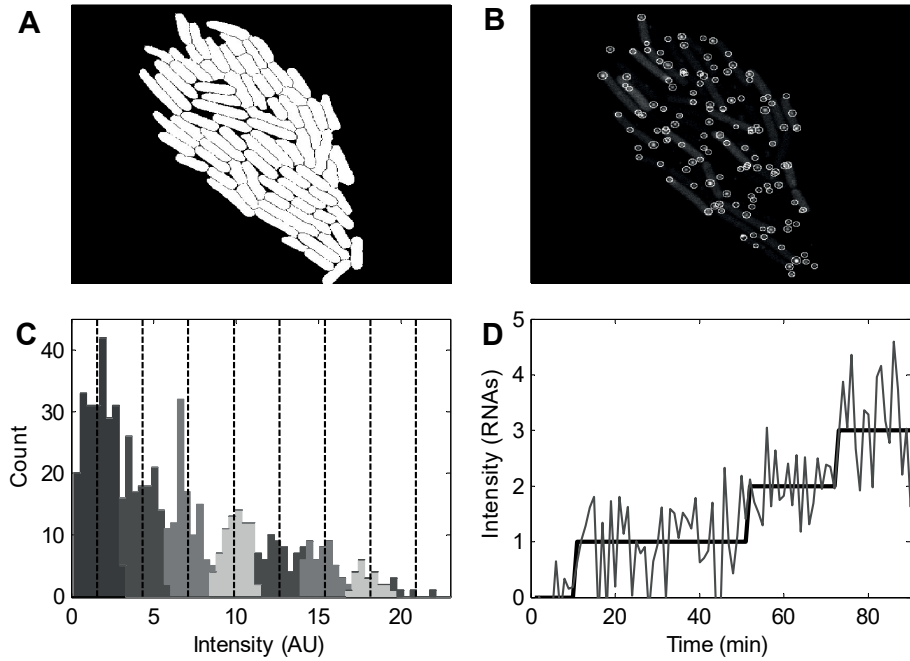
To fit an empirical distribution to a sum of  $d$ -exponential variates (of possibly unequal rates), as in (Mäkelä et al. 2013), we select the exponential rate parameters  $\lambda_1, \dots, \lambda_d$  so that the

Kolmogorov-Smirnov (K-S) statistics is minimized. I.e., parameters are selected as  $\hat{\theta} = \arg \max_{\theta=\lambda_1, \dots, \lambda_d} \sup_x |F_\theta(x) - G(x)|$ , where  $F_\theta(x)$  is the cumulative distribution function (CDF) of a sum of  $d$  exponentials with parameters  $\theta = (\lambda_1, \dots, \lambda_d)$ , and  $G(x)$  is the CDF of the empirical distribution.

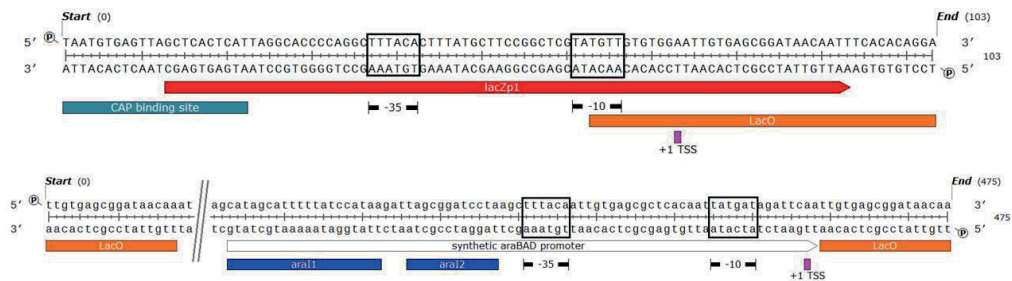
$$F_{\theta=L_1, \dots, L_d}(x) := \sum_{i=1}^d \left( (1 - e^{-L_i x}) \prod_{\substack{j=1 \\ j \neq i}}^d \frac{L_j}{L_j - L_i} \right) \quad [4]$$

The parameter values  $\theta$  are found using a nonlinear numerical optimizer. This method is convenient, since if the K-S test is rejected for the parameters  $\hat{\theta}$ , it would also be rejected for any other set of parameters  $\theta$  in this family of fitted distributions, indicating that these distributions are inappropriate models of the data.

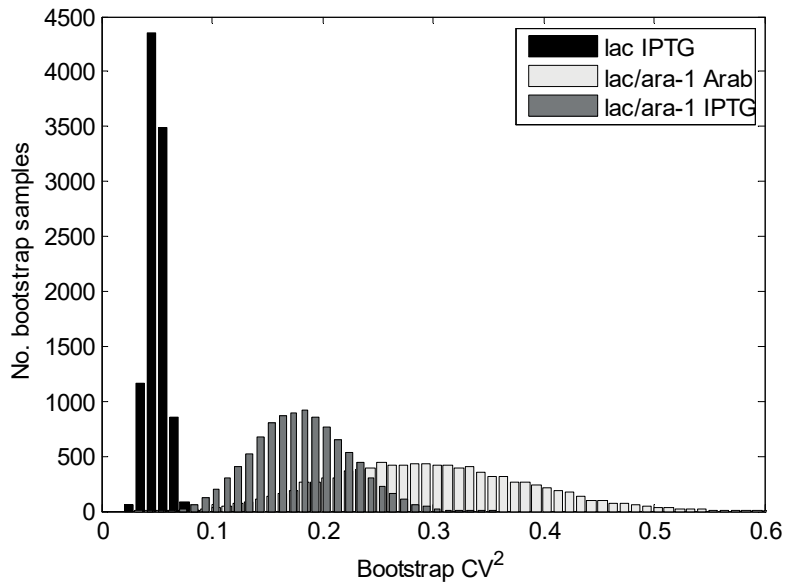
## Supplementary Figures



**Figure S1.** Image analysis and RNA quantification. **(A)** Segmented cell backgrounds **(B)** Detection of RNA spots using the Kernel Density Estimation. **(C)** Analysis of RNA numbers from single cells. **(D)** Example of the results of the method of detection of novel RNA appearance events in a single cell from time series data on total RNA-spot fluorescence in a cell.

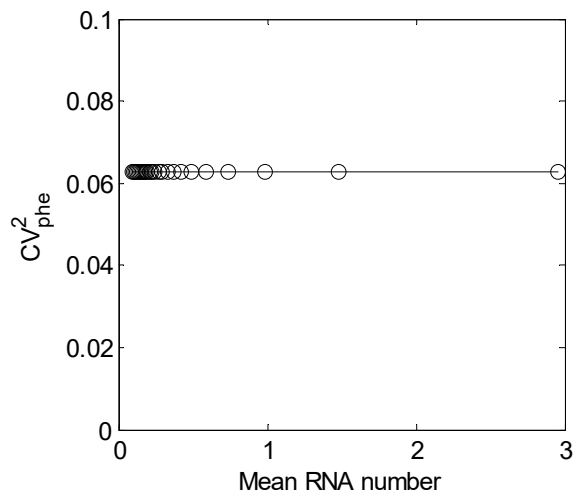


**Figure S2.** Topography and sequences of promoters (top) *P<sub>lac</sub>* (Golding & Cox 2004) and (bottom) *P<sub>lac/ara-1</sub>* (Lutz & Bujard 1997). RNA polymerase binding sites are boxed. The two small pink boxes show the transcriptional start site. Blue boxes show the operator binding sites of *araI1* and *araI2*, and orange boxes show the operator sites of *lacO*. White and red arrows show the *araBAD* and the *LacZp1* promoters, respectively. The figures were produced using the SnapGene Software (GSL Biotech, Chicago, IL, USA).

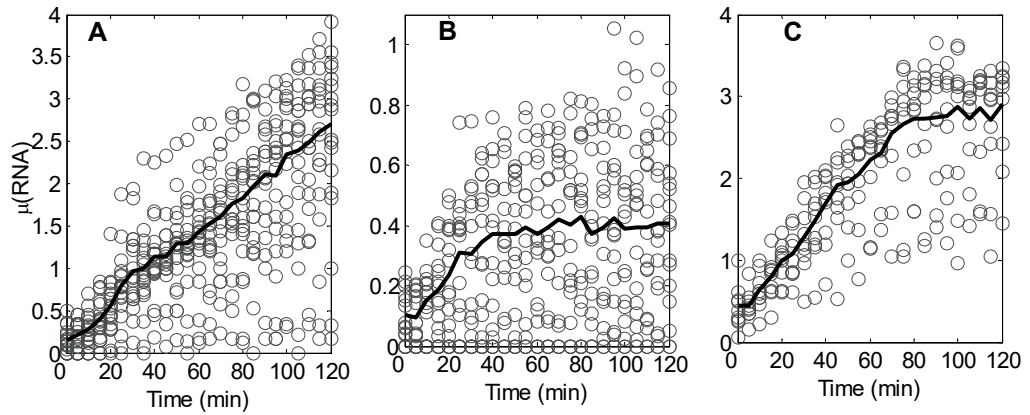


**Figure S3.** Bootstrapping of cells in the lineages.  $CV^2_{\text{phe}}$  of the RNA numbers between lineages. Std of bootstrapping samples corresponding to standard errors for lac IPTG, lac/ara-1 IPTG and lac/ara-1 Arab are 0.008, 0.043 and 0.089, respectively.

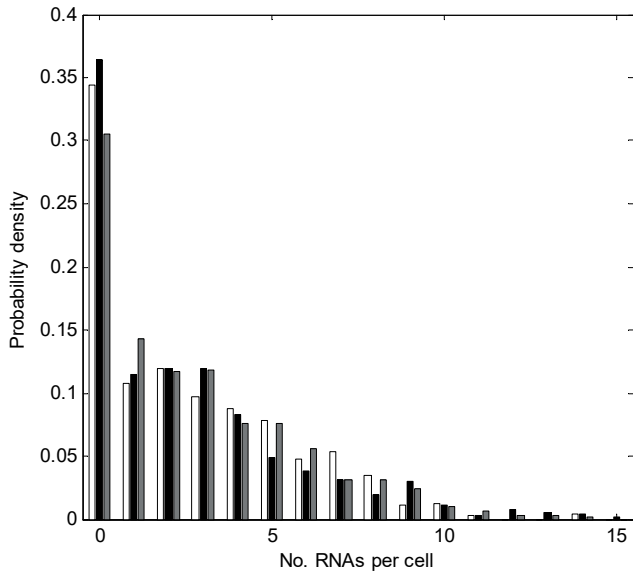




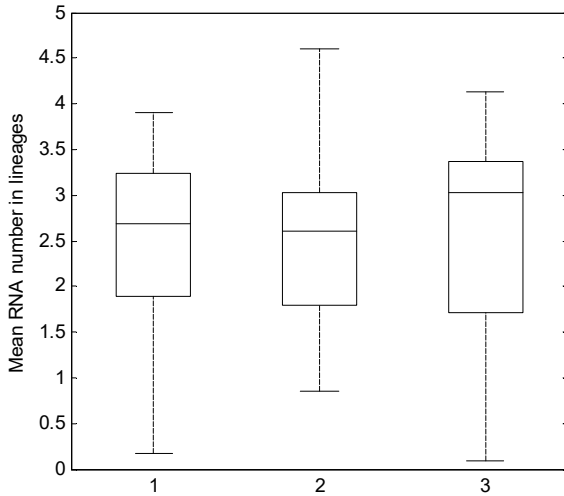
**Figure S4.**  $CV_{\text{phe}}^2$  as a function of mean RNA number. Different mean RNA numbers were achieved by changing the overall duration of transcription. Time-lengths between consecutive transcription events were modeled to be between 100 s and 3000 s, resulting in different mean RNA levels (while maintaining constant the ratio  $\tau_{\text{cc}}/\Delta t$ ). From these results, we conclude that  $CV_{\text{phe}}^2$  of RNA numbers is independent of the mean production rate of that RNA.



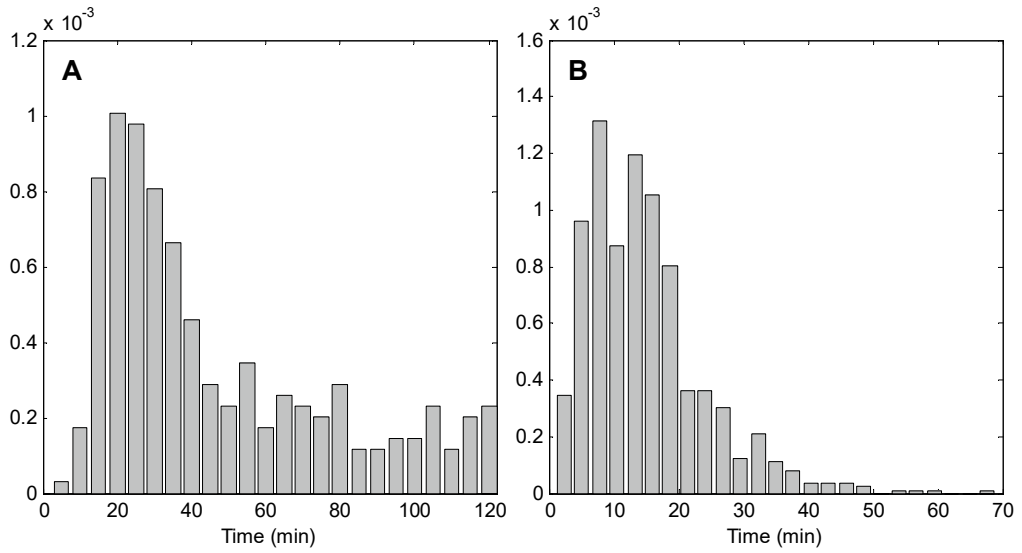
**Figure S5.** Mean RNA numbers per cell in each lineage over time, following induction. **(A)**  $P_{lac/ara-1}$  induced with 1 mM IPTG (1468 cells). **(B)**  $P_{lac/ara-1}$  induced with 1 % L-arabinose (1296 cells). **(C)**  $P_{lac}$  induced with 1 mM IPTG (1665 cells). The degree of lineage-to-lineage variability in each condition is expected to be a consequence of, among other causes, the lineage-to-lineage variability in cellular components affecting the kinetics of active transcription and inducers intake (Mäkelä et al, 2013). The contribution on the variability from these two processes is also expected to change over time in each condition, and seems to differ between the 3 conditions.



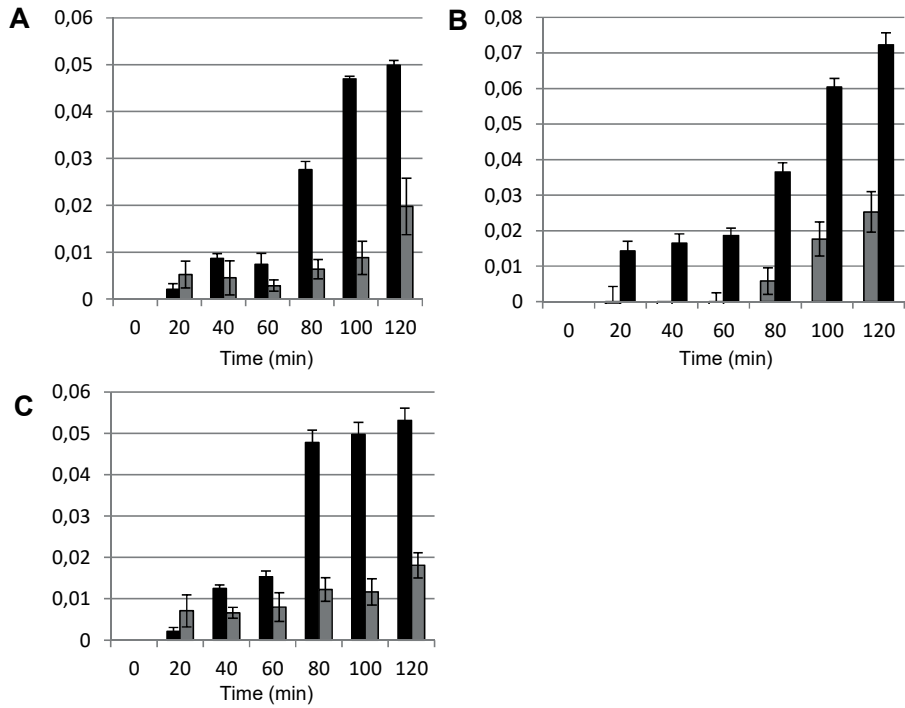
**Figure S6.** RNA numbers in individual cells after 2 hours of induction from 3 independent experiments of  $P_{lac/ara-1}$  induced with IPTG. Experiment 1 is a time series measurement and experiments 2 and 3 are cell population measurements. To compare the RNA distributions from different experiments, we used the two-sample Kolmogorov-Smirnov test to test the null hypothesis that the samples are drawn from the same distribution. We obtained p-values of 0.24 (between experiments 1 and 2) and 0.58 (between experiments 1 and 3) and, thus, the null hypothesis cannot be rejected (for p-value < 0.01, it is generally accepted that the hypothesis that the two distributions are the same should be rejected). The number of cells observed in experiments 1, 2 and 3 were 924, 1219 and 764, respectively. No statistically significant differences between the experiments are visible.



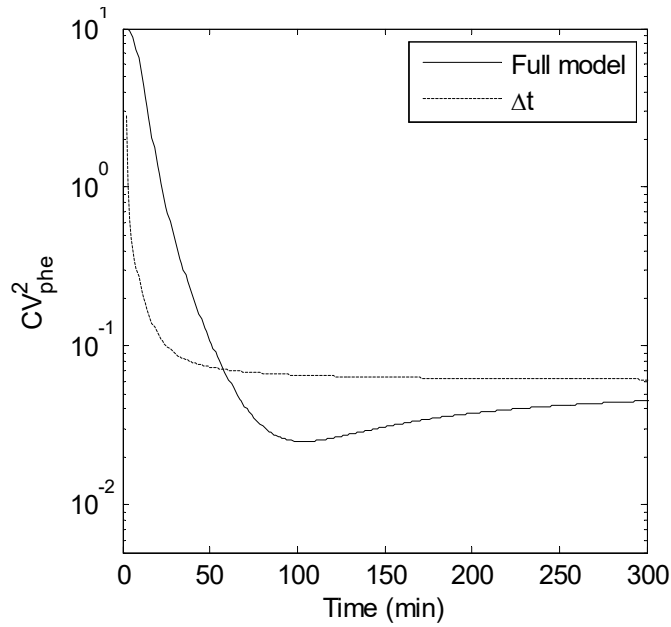
**Figure S7.** Independent measurements of lineage-to-lineage variability in mean RNA numbers for  $P_{lac/ara-1}$  after 2h of induction by 1mM IPTG. Experiment 1 is a time series and experiments 2 and 3 are cell population measurements. We show the boxplots of mean RNA numbers for each experiment. To test the null hypothesis that the samples are drawn from the same distribution, we used the two-sample Kolmogorov-Smirnov test and obtained p-values of 0.77 (experiments 1 and 2) and 0.82 (experiments 1 and 3). For p-value  $< 0.01$  it is generally accepted that the hypothesis that the two distributions are the same should be rejected. Given this, the null hypothesis cannot be rejected. In all conditions, the variability between lineages in mean RNA numbers is above chance. Relevantly, this variability differs with the promoter as well as with the inducer.



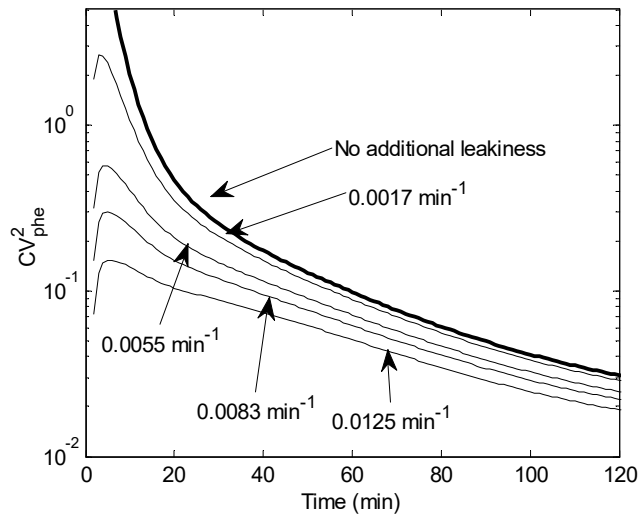
**Figure S8.** Production kinetics of RNAs by  $P_{lac/ara-1}$  induced with IPTG. **(A)** Lineage waiting times for the first production event, and **(B)** time intervals between consecutive production events in individual cells. The y-axis is the probability density.



**Figure S9.** Plate reader measurements of target promoter expression. **(A)**  $P_{lac/ara-1}$  induced with 1mM IPTG (black). **(B)**  $P_{lac/ara-1}$  induced with 1% l-arabinose (black). **(C)**  $P_{lac}$  induced with 1mM IPTG (black). The grey bar is the control (without induction). Data is normalized with the first time moment of the times series. y-axis is the normalized fluorescence intensity.

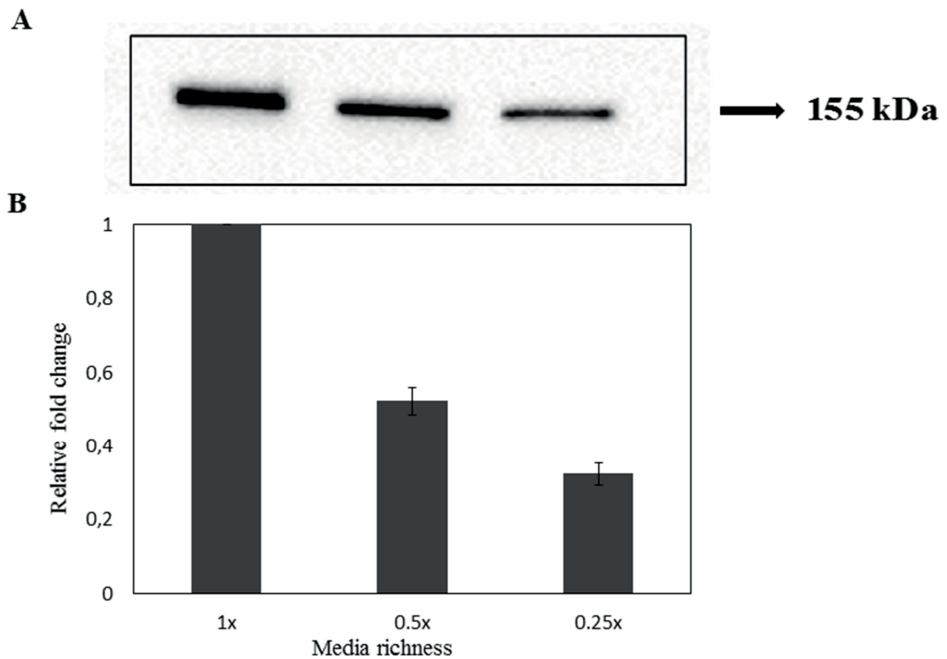


**Figure S10.**  $CV_{\text{phe}}^2$  for models of  $P_{\text{lac/ara-1}}$  induced with arabinose. The models consist of both processes (Full model:  $\Delta t$  and  $t_0$ ), and only the  $\Delta t$  process. Over time, the  $CV_{\text{phe}}^2$  of the full model and the  $CV_{\text{phe}}^2$  of the model accounting for active transcription ( $\Delta t$  process) become similar, as the activation events become more rare as time progresses.

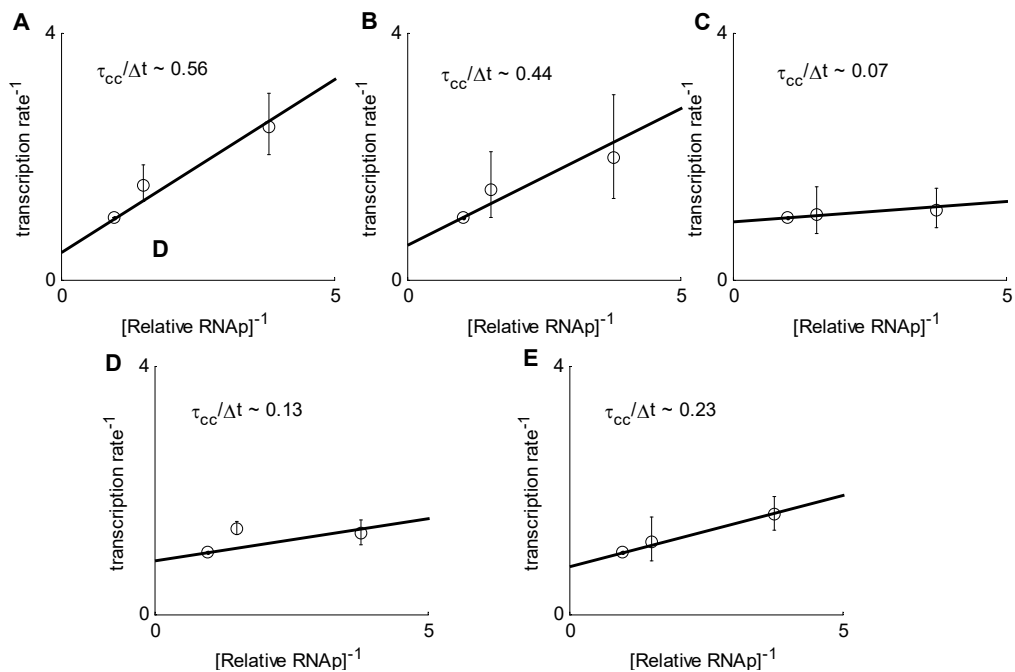


**Figure S11.** Model of  $P_{lac/ara-1}$  induced with IPTG (full model with different rates of leakiness). Leakiness is modeled as an extra reaction of production whose dynamics is that of a Poisson process (see rates of leakiness in each case). Visibly, increasing the rate of leakiness decreases the lineages'  $CV^2_{\text{phe}}$ .





**Figure S12.** Relative RpoC protein levels of *E. coli* cells of the strain DH5 $\alpha$ -PRO when grown in different media richness (1x, 0.5x, and 0.25x) measured by Western blotting. **(A)** A replicate of Western blot image. **(B)** The observed levels of the RpoC protein compared to the protein level in 1x medium (treated as a 100% reference) were 52% in 0.5x medium and 32% in 0.25x medium. The error bars (0.5x: 3.8%, 0.25x: 3.1%) reflect 90% confidence intervals in the differences for 4 biological replicates. The relative band intensities were quantified by the Image lab software (version 5.2.1) from the chemiluminescence blots. We find large differences in RpoC protein levels for different media richness.



**Figure S13.**  $\tau$ -plots. (A)  $P_{lac/ara-1}$  with 1 mM IPTG. (B)  $P_{lac/ara-1}$  with 100  $\mu$ M IPTG. (C)  $P_{lac/ara-1}$  with 10  $\mu$ M IPTG. (D)  $P_{lac}$  with 1 mM IPTG. (E)  $P_{lac/ara-1}$  with 1 % Arabinose. Also shown is the resulting ratio  $\tau_{cc}/\Delta t$  in each case. Visibly,  $\tau_{cc}/\Delta t$  differs with inducer concentration (compare A, B, and C results), type of inducer (compare, e.g., A and E), and between different promoters (compare A and D).

## Supplementary Tables

**Table S1.** Correlation between the distance to the center of a colony of a cell and its number of RNA molecules in each of the three conditions studied. We calculated the correlation between the distance to colony center and the number of produced RNA molecules in individual cells to assess whether the extracellular environment (i.e. inducer concentration or media richness) had sufficient local variability in conditions to generate tangible differences in the RNA production rates of individual cells. In all conditions, there are only weak, not statistically significant correlations, indicating that the induction level of individual cells is not location-dependent.

	$\rho$	p-value
$P_{lac/ara-1}$ IPTG	0.0394	0.27
$P_{lac/ara-1}$ Arab	-0.0230	0.59
$P_{lac}$ IPTG	-0.0175	0.63

**Table S2.** Measured mean values ( $\mu$ ) and  $CV^2_{phe}$  of  $t_{act}$  and  $\Delta t$  for  $P_{lac/ara-1}$  with IPTG (15 lineages),  $P_{lac/ara-1}$  with arabinose (10 lineages), and  $P_{lac}$  with IPTG (8 lineages). Also shown is the leakiness (percentage of cells with RNAs prior to induction) and the total number of RNA production events observed. Error estimates are from bootstrap sampling of the cells in the lineages. The  $CV^2_{phe}$  in  $t_{act}$  as well as in  $\Delta t$  between lineages differ between all conditions.

	$P_{lac/ara-1}$ IPTG	$P_{lac/ara-1}$ Arab	$P_{lac}$ IPTG
No. RNA prod. events	1799	391	1388
Leakiness	9.8%	7.2%	34.3%
$\mu$ ( $t_{act}$ ) (s)	2030 $\pm$ 191	2488 $\pm$ 209	1085 $\pm$ 126
$CV^2_{phe}$ ( $t_{act}$ )	0.141 $\pm$ 0.041	0.078 $\pm$ 0.023	0.124 $\pm$ 0.056
$\mu$ ( $\Delta t$ ) (s)	889 $\pm$ 25	1254 $\pm$ 85	1365 $\pm$ 39
$CV^2_{phe}$ ( $\Delta t$ )	0.014 $\pm$ 0.004	0.051 $\pm$ 0.013	0.008 $\pm$ 0.004

## References

- Bernstein, J.A. et al., 2002. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 99(15), pp.9697–9702.
- Campbell, R.E. et al., 2002. A monomeric red fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12), pp.7877–7882.
- Chen, H. et al., 2015. Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. *Molecular Systems Biology*, 11(781), pp.1–11.
- Chong, S. et al., 2014. Mechanism of Transcriptional Bursting in Bacteria. *Cell*, 158(2), pp.314–326.
- Daruwalla, K.R., Paxton, A.T. & Henderson, P.J.F., 1981. Energization of the transport systems for arabinose and comparison with galactose transport in Escherichia coli. *Biochemical Journal*, 200(3), pp.611–627.
- Ehrenberg, M., Bremer, H. & Dennis, P.P., 2013. Medium-dependent control of the bacterial growth rate. *Biochimie*, 95(4), pp.643–658.
- Fritz, G. et al., 2014. Single cell kinetics of phenotypic switching in the arabinose utilization system of E. coli. *PLoS ONE*, 9(2), p.e89532.
- Golding, I. et al., 2005. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6), pp.1025–1036.
- Golding, I. & Cox, E.C., 2004. RNA dynamics in live Escherichia coli cells. *Proceedings of the National Academy of Sciences of the United States of America*, 101(31), pp.11310–11315.
- Hensel, Z. et al., 2012. Stochastic expression dynamics of a transcription factor revealed by single-molecule noise analysis. *Nature Structural and Molecular Biology*, 19(8), pp.797–802.
- Hsu, L.M., 2002. Promoter clearance and escape in prokaryotes. *Biochimica et Biophysica Acta*, 1577(2), pp.191–207.
- Johnson, C.M. & Schleif, R.F., 1995. In vivo induction kinetics of the arabinose promoters in Escherichia coli. *Journal Of Bacteriology*, 177(12), pp.3438–3442.
- Jones, D.L., Brewster, R.C. & Phillips, R., 2014. Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346(6216), pp.1533–1537.

- Krystek, M. & Anton, M., 2007. A weighted total least-squares algorithm for fitting a straight line. *Measurement Science and Technology*, 18, pp.3438–3442.
- Liang, S. et al., 1999. Activities of constitutive promoters in Escherichia coli. *Journal of molecular biology*, 292(1), pp.19–37.
- Livak, K.J. & Schmittgen, T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2-DDCT method. *Methods*, 25(4), pp.402–408.
- Lloyd-Price, J. et al., 2016. Dissecting the stochastic transcription initiation process in live Escherichia coli. *DNA Research*, 23(3), pp.203–214.
- Lutz, R. et al., 2001. Dissecting the functional program of Escherichia coli promoters: the combined mode of action of Lac repressor and AraC activator. *Nucleic Acids Research*, 29(18), pp.3873–3881.
- Lutz, R. & Bujard, H., 1997. Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Research*, 25(6), pp.1203–1210.
- McClure, W.R., 1985. Mechanism and control of transcription initiation in prokaryotes. *Annual Review of Biochemistry*, 54, pp.171–204.
- McClure, W.R., 1980. Rate-limiting steps in RNA chain initiation. *Proceedings of the National Academy of Sciences of the United States of America*, 77(10), pp.5634–5638.
- Megerle, J.A. et al., 2008. Timing and dynamics of single cell gene expression in the arabinose utilization system. *Biophysical Journal*, 95(4), pp.2103–2115.
- Moffitt, J.R. & Bustamante, C., 2014. Extracting signal from noise: Kinetic mechanisms from a Michaelis-Menten-like expression for enzymatic fluctuations. *FEBS Journal*, 281(2), pp.498–517.
- Munsky, B. & Khammash, M., 2006. The finite state projection algorithm for the solution of the chemical master equation. *Journal of Chemical Physics*, 124(4), p.44104.
- Mäkelä, J. et al., 2013. In vivo single-molecule kinetics of activation and subsequent activity of the arabinose promoter. *Nucleic Acids Research*, 41(13), pp.6544–6552.
- Patrick, M. et al., 2015. Free RNA polymerase in E. coli. *Biochimie*, 119, pp.80–91.
- Rosenfeld, N. et al., 2005. Gene regulation at the single-cell level. *Science*, 307(5717), pp.1962–1965.
- Saecker, R.M., Record, M.T. & DeHaset, P.L., 2011. Mechanism of Bacterial Transcription

- Initiation: RNA Polymerase - Promoter Binding, Isomerization to Initiation-Competent Open Complexes, and Initiation of RNA Synthesis. *Journal of Molecular Biology*, 412(5), pp.754–771.
- Schleif, R., 2000. Regulation of the L-arabinose operon of Escherichia coli. *Trends in Genetics*, 16(12), pp.559–565.
- Skerra, A., 1994. Use of the tetracycline promoter for the tightly regulated production of a murine antibody fragment in Escherichia coli. *Gene*, 151(1–2), pp.131–135.
- Taniguchi, Y. et al., 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991), pp.533–538.
- Tran, H. et al., 2015. Kinetics of the cellular intake of a gene expression inducer at high concentrations. *Molecular Biosystems*, 11(9), pp.2579–2587.
- Weickert, M.J. and Adhya, S., 1993. The galactose regulon of Escherichia coli. *Mol. Microbiol*, 10, pp.245–251.

# PUBLICATION IV

## **Transcription Initiation Controls Skewness of the Distribution of Intervals Between RNA Productions**

Vinodh K. Kandavalli, Sofia Startceva, and Andre S. Ribeiro.

In Proceedings of the 31<sup>st</sup> European Simulation and Modelling (ESM-2017), October 25-27, 2017, pp. 418-421 ISBN: 978-9492859-00-6. Paulo J.S, G. (Ed.).

**Publication reprinted with the permission of the copyright holders.**





# TRANSCRIPTION INITIATION CONTROLS SKEWNESS OF THE DISTRIBUTION OF INTERVALS BETWEEN RNA PRODUCTIONS

Vinodh K. Kandavalli, Sofia Startceva, and Andre S. Ribeiro  
Laboratory of Biosystem Dynamics, BioMediTech Institute  
and Faculty of Biomedical Sciences and Engineering,  
Tampere University of Technology, Finland  
E-mail: andre.ribeiro@tut.fi

## KEYWORDS

Transcription Initiation, Skewness in RNA production; Stochastic Models; Single-RNA measurements.

## ABSTRACT

Most regulation in transcription controls when and with which intensity genes are expressed. However, recent evidence suggests that control is also exerted on the noisiness of this process. Here, we use an empirically validated stochastic multi-step model of transcription to explore how its steps kinetics affect the skewness of the distribution of intervals between consecutive RNA productions in individual cells. From the simulations, we show that skewness is independent of the mean transcription rate, but differs widely with the fraction of time the RNA polymerase spends in the steps following open complex formation. Next, from qPCR and live, time-lapse, single-RNA microscopy measurements of multiple promoters, we validate our model predictions. Using the validated model, we then show that skewness affects, e.g., the fraction of time protein numbers are below a threshold. We conclude that skewness in transcription kinetics can be tuned by the rate-limiting steps in initiation and, thus, may be an evolvable decision-making parameter of genetic circuits.

## INTRODUCTION

In prokaryotes, e.g. *Escherichia coli*, transcription is the critical process where most regulation of the metabolism and responses to environment changes occur (López-Maury et al. 2008). It is thus not surprising that *E. coli* possesses a plethora of repression and activation molecules, along with other means to silence and activate specific genes (McClure 1985; Lutz et al. 2001). There are also various global regulation mechanisms, such as  $\sigma$  factors (Jishage et al. 1996) and DNA super-coiling (Menzel & Gellert 1983).

Nevertheless, bacterial cell populations exhibit single-cell heterogeneity in gene expression profiles (Leibler & Kussell 2010). This diversity has two sources. One is the stochastic nature of the chemical processes involving gene expression, due to the low number of regulatory molecules involved. The other is differences between cells in their numbers of various components, age, cycle stage, etc. (Elowitz et al. 2002). This noise was found to affect multiple cellular functions,

including stress response, metabolism, cell cycle, circadian rhythms and aging (Raj & van Oudenaarden 2008).

Similarly to noise, asymmetries in RNA and protein kinetics might play a role in cells metabolism, etc., as they can determine if the number of RNA or proteins crosses a threshold ‘used’ by a genetic circuit in decision making. So far, such asymmetries have not been quantified, but recent measurements of time intervals between RNA productions in individual cells of various promoters under various conditions suggest that they are not negligible (Tran et al. 2015; Häkkinen & Ribeiro 2015; Häkkinen & Ribeiro 2016).

Here, we investigate if and by which degree the rate-limiting steps in transcription initiation can tune asymmetries in the distribution of intervals between productions of RNAs in individual cells. For this, we consider a stochastic model of transcription initiation and investigate how the asymmetries differ within the realistic ranges of parameter values. Next, we experimentally validate these predictions by qPCR and live, time-lapse, single-molecule RNA microscopy measurements of the transcription kinetics of multiple promoters. Finally, we investigate whether changes in skewness of a gene’s transcription kinetics can have tangible consequences in the crossing of thresholds of protein numbers over time.

## MATERIALS AND METHODS

### Bacterial Strains and Plasmids, Media and Cell Growth, Microscopy, and Image Analysis

Microscopy data are from (Kandavalli et al. 2016; Oliveira et al. 2016). Briefly, *E. coli* cells carrying 2 plasmids were used: a low copy reporter plasmid expressing MS2-GFP controlled by the promoter  $P_{Lac}$  or  $P_{Tet}$ , and a single copy F-based plasmid, expressing the RNA with a 96 MS2-GFP binding site array followed by mRFP1 controlled by  $P_{BAD}$ ,  $P_{TetA}$  or  $P_{Lac-ara-1}$ . Cultures were grown in LB media overnight at 30 °C in an orbital shaker with aeration of 250 rpm and diluted to fresh LB media to initial  $OD_{600}$  of 0.05 (measured with Ultraspec 10 cell density meter). Next, they were incubated at 37 °C at 250 rpm until reaching an  $OD_{600}$  of 0.25. To produce MS2-GFP, e.g. when under the control of  $P_{Lac}$ , cells are induced with 1 mM IPTG (for  $P_{Tet}$  we induce with 100 ng of aTc) and allowed to grow until  $OD_{600}$  of 0.5. For the target induction, 0.1% arabinose and 1 mM IPTG for  $P_{Lac-ara-1(Full)}$ , 0.1% arabinose alone for  $P_{Lac-ara-1(ara)}$ , 1 mM IPTG alone for  $P_{Lac-ara-1(IPTG)}$  and 0.1% arabinose for  $P_{BAD}$  is used. For  $P_{TetA}$ , no induction is required, as the cells lack the gene coding for the repressor, TetR (Kandavalli et al. 2016).

For microscopy, a few  $\mu$ l of cells with the reporter and target plasmids were sandwiched between a coverslip and an agarose gel pad (2.5%), also containing the inducers. Prior to this, the chamber (FCS2, Biopetechs) was heated to 37 °C and placed under the microscope. Cells were visualized using a Nikon Eclipse (Ti-E, Nikon) inverted microscope, equipped with a 100x Apo TIRF (1.49 NA, oil) objective. Confocal

images were obtained by a C2+ (Nikon) confocal laser-scanning system. To visualize fluorescence ‘spots’, we used a 488 nm laser (Melles-Griot) and an emission filter (HQ514/30, Nikon). Confocal images were taken every 1 min for 2 h, and phase contrast images were obtained every 5 min by an external phase contrast system and CCD camera (DS-Fi2, Nikon), using Nikon Nis-Elements software.

Analysis of the images was performed by the software ‘CellAging’ in four steps (Häkkinen et al. 2013): (i) cell segmentation from phase-contrast images, (ii) fluorescent RNA intensity detection from the confocal images, (iii) cell lineage construction, and (iv) RNA production estimation from the single-cell RNA intensity time series. We used the software to perform an automated segmentation of phase-contrast images, followed by manual correction. Next, from each segmented cell, at each time point, fluorescent spots are detected automatically. Finally, cell lineages are constructed, by establishing the relationships between cell masks in sequential frames. In these, time-series of fluorescent spots intensity were obtained for each cell. From those, the time points when novel RNA molecules (‘spots’) appear in each cell were estimated (Häkkinen & Ribeiro 2015). Finally, the time intervals between consecutive RNA productions in individual cells were estimated.

## qPCR

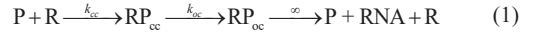
In the case of  $P_{BAD}$  and  $P_{TetA}$ , qPCR data was obtained from (Kandavalli et al. 2016). In the case of  $P_{Lac-ara-1}$ , the data is from measurements performed here. To measure gene expression by qPCR, we grew cells as in (Kandavalli et al. 2016). Total RNA was isolated and quantified. The RNA samples were treated with DNase to remove residual DNA, followed by cDNA synthesis. cDNA samples were mixed with qPCR master mix containing iQ SYBR Green supermix (Biorad), with primers for the target and reference genes. The reaction was carried out in triplicates. For quantifying the target gene, we used the following primers: for mRFP1 (Forward: 5’ TACGACGCCGAGGTCAAG 3’ and Reverse: 5’ TTGTGGGAGGTGATGTCCA 3’), and for the 16S RNA reference gene (Forward: 5’ CGTCAGCTCGTGTGTGAA 3’ and Reverse: 5’ GGACCGCTGGCAACAAAG 3’). The following conditions were used: 40 cycles of 95°C for 10 s, 52°C for 30 s and 72°C for 30 s for each cDNA replicate. We used no-RT controls and no-template controls to crosscheck non-specific signals and contamination. PCR efficiencies of these reactions were greater than 95%. The data from CFX Manager TM Software was used to calculate the relative gene expression and its standard error (Livak & Schmittgen 2001).

## Model of Transcription

We model transcription as a multi-step process, represented in (1), following the empirically validated models in (McClure 1985). The level of detail of our model is based on the one we

can reach in the measurements (i.e. intervals between RNA production events with measurements taken every minute).

From these intervals one can dissect the fraction of time of those intervals that is spent prior and after commitment to the open complex formation (Häkkinen & Ribeiro 2015). As such, transcription is modeled by the following multi-step process (McClure 1985):



where  $k_{cc}^* = R \cdot k_{cc}$ .

The process starts with RNAP, R, binding to a free, active promoter, P, and forming a closed complex,  $RP_{cc}$ , at the rate  $k_{cc}$ .  $k_{cc}^*$  stands for the inverse of the mean time that  $RP_{cc}$  remains in equilibrium with P and RNAP, until it starts forming a stable open complex. That is, this model does not explicitly represent the instability of  $RP_{cc}$ . As such, the first step is not an elementary chemical process. Rather, its rate represents the inverse of the time until a stable open complex forms, which depends on preceding events, such as binding and unbinding of the RNAP to the promoter (i.e. reversibility), 1D diffusive searches, etc. (Bai et al. 2006).

The second step in (1) represents the open complex formation,  $RP_{oc}$ , which is a nearly irreversible step (McClure 1985; Lloyd-Price et al. 2016; Kandavalli et al. 2016) and requires the RNAP to open the DNA double helix (Chamberlin.MJ 1974; McClure 1985). This is followed by promoter escape (after which P is released into the system), elongation and termination (i.e., the release of RNA and R). These latter steps are expected to be much shorter-length than the events in initiation (Herbert et al. 2008) and, thus, are not represented. Further, they would only affect the variance and not the mean duration of the intervals between transcription events.

Regardless of the complexity of the steps, recent studies suggest that, to a degree, the process can be well-modelled by two consecutive, independent exponential steps (Tran et al. 2015; Kandavalli et al. 2016). Thus, the probability density function (pdf) of the distribution of intervals between transcriptions is the convolution of their pdfs:

$$f_{\Delta t}(t) = \frac{k_{cc}^* k_{oc}}{k_{oc} - k_{cc}^*} (e^{-k_{cc}^* t} - e^{-k_{oc} t}) \quad (2)$$

We assume also a first-order reaction modelling RNA degradation with a rate  $k_{d,rna} = 0.0033 \text{ s}^{-1}$ , which is the median of the RNA degradation rate in *E. coli* (Bernstein et al. 2002):



For simplicity, we model translation as a single-step event which produces an unfolded protein  $Pro_{un}$  with a rate of  $k_{tr} = 0.0637$  (Jones et al. 2007):



Finally, proteins fold into functional at the rate  $k_{fold} = 0.0024$ , and degrade at the rate  $k_{d,pro} = 0.0017$  (Cormack et al. 1996):



## Skewness as a Measure of Asymmetry of the Intervals Distribution.

As a measure of asymmetry of the distribution of transcription intervals, we use its skewness,  $S$ , as in (MacGillivray 1986):

$$S = \frac{m_3}{m_2^{3/2}}, \text{ where } m_r = \frac{1}{n} \sum (x_i - \bar{x})^r \quad (7)$$

More precisely, we estimate the sample skewness ( $S_s$ ) of measured and simulated data distributions by applying a correction to increase the estimates precision for samples from asymmetric distributions (8) (Joanes & Gill 1998):

$$S_s = \frac{\sqrt{n(n-1)}}{n-2} \cdot S \quad (8)$$

To estimate the standard uncertainty of  $S_s$ , we performed non-parametric bootstrap as in (Carpenter & Bithell 2000). Namely, for each data set, we resampled the data randomly with replacement (using the original amount of samples)  $10^5$  times, and calculated the bootstrap sample skewness,  $S_{sb}$ . As the obtained  $S_{sb}$  distributions were well-approximated by a normal distribution, we estimated the standard uncertainty as the 68% percentile confidence interval of the  $S_{sb}$  distribution.

## $\tau$ plots

*In vitro* and *in vivo* studies have demonstrated that the mean time between transcription events ( $\Delta t$ ) can be altered by changing the free RNAP concentration (Shehata & Marr 1971; Lloyd-Price et al. 2016). Also, assuming (1), only the closed complex formation duration changes with the free RNAP concentration (McClure 1985). This change was shown to be linear for a given range of cell growth conditions (Lloyd-Price et al. 2016). As such, it is possible, within this range of conditions, to produce a  $\tau$  plot by placing the inverse of the relative RNAP concentration in the x-axis, and the inverse of the relative rate of RNA production in the y-axis.

The relative rate of RNA production can be measured by qPCR. The relative RNAP concentration can be measured by Western Blot (Kandavalli et al. 2016). The data points are then fitted with a line. Scalling the production rates to the condition of interest, the intercept of the line with the y-axis equals ( $\tau_{oc}/\Delta t$ ) of this condition, as it represents the media condition with infinite RNAP (and, thus, with infinitely fast closed complex formation).

## Stochastic Simulations

To simulate the model (1)-(6), we use SGNS2 (Lloyd-Price et al. 2012), which is driven by the Stochastic Simulation Algorithm (Gillespie 1977), but allows also for multi-time-delayed reactions (Roussel & Zhu 2006). We accounted for individual cell observation times, as these affect the measured intervals (unlike in the theoretical predictions of the pdf of the distributions of intervals between RNA production events (2)). This single-cell observation time-lengths depend on (a) cell doubling time, (b) duration of the measurement, and (c) the degree of overlap between measurement time and cell doubling time. We measured such distribution of single-cell observation windows for each studied condition. Next, we set

$k_{cc}$  and  $k_{oc}$  according to the  $\tau_{oc}/\Delta t$  obtain from qPCR and Western Blot. The absolute values of these rates are then fitted to match the mean of the measured distribution.

## Truncated Gaussian Distribution of Intervals Between Consecutive RNA Productions

To obtain Gaussian distributions truncated at zero and with a given mean  $\mu$  and squared coefficient of variation,  $CV^2$ , we use the following procedure. First, we obtained the best fit value of the standard deviation  $\hat{\sigma}$  of the Gaussian with mean  $\mu$ , which minimizes the difference between  $CV^2$  and  $CV_{tr}^2$  of the truncated distribution. Next, we calculated a scaling coefficient  $\alpha = \mu/\mu_{tr}$ . Finally, we truncated at zero a Gaussian distribution with mean of  $\alpha\mu$  and standard deviation of  $\alpha\hat{\sigma}$ , which results in the desired distribution.

## RESULTS AND CONCLUSIONS

### Skewness is Controlled by $\tau_{oc}/\Delta t$ , but is Independent of the Mean Interval Between Transcription Events

We consider the model of transcription in (1). Its RNA production kinetics is determined by  $k_{cc}^*$  and  $k_{oc}$ .  $k_{cc}^*$  defines the inverse of the mean time for the RNAP to find the promoter and complete the closed complex ( $\tau_{cc}$ ).  $k_{oc}$  defines the inverse of the mean time for the completion of the open complex formation ( $\tau_{oc}$ ). Thus:

$$\Delta t = \tau_{cc} + \tau_{oc} \quad (9)$$

Given measurements of  $\Delta t$  in live *E. coli* cells of various active promoters (Häkkinen & Ribeiro 2016; Kandavalli et al. 2016; Lloyd-Price et al. 2016; Tran et al. 2015; Oliveira et al. 2016), we assume its realistic range of values to be between 10 and 2500 seconds. To study how changing the kinetics of the two rate limiting steps of the model allows tuning the asymmetry (as measured by  $S$ ) of the  $\Delta t$  distribution, we vary  $\tau_{oc}/\Delta t$  (by changing  $\tau_{cc}$  and  $\tau_{oc}$ ) while maintaining  $\Delta t$  constant. For each such combination of  $\Delta t$  and  $\tau_{oc}/\Delta t$  values, from (7), we calculated  $S$  of the pdf of the  $\Delta t$  distribution. From this, we find that, first,  $S$  changes significantly with  $\tau_{oc}/\Delta t$ , being symmetric around 0.5, where it is minimal. On the other hand, it is independent from the mean value  $\Delta t$ , for any given constant value of  $\tau_{oc}/\Delta t$  (Fig. 1).

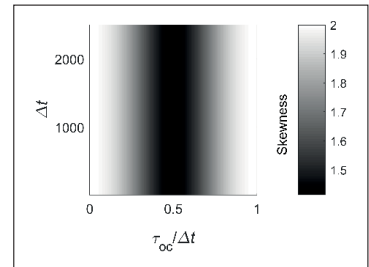


Figure 1: 2D plot of the model-based analytical prediction of the skewness of the distribution of intervals between consecutive RNA production events as a function of  $\Delta t$  and  $\tau_{oc}/\Delta t$ .

## Empirical Validation of the Model Predictions

To validate the above, we attained empirical data on the  $\Delta t$  distribution for various promoters and induction schemes (Kandavalli et al. 2016; Oliveira et al. 2016). Also, for each condition, by qPCR (Methods), we measured  $\tau_{oc}/\Delta t$ . In Fig. 2, we confront the empirical values of  $S$  as a function of  $\tau_{oc}/\Delta t$  with the simulated predictions (Methods). Visibly, the model fits the data for a wide range of possible values of  $\tau_{oc}/\Delta t$ .

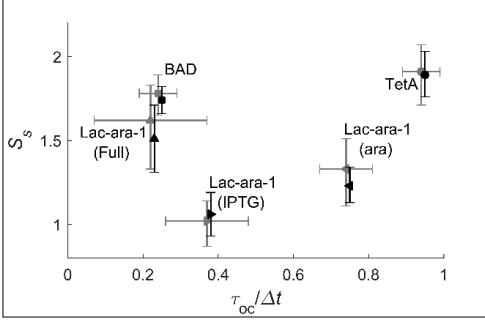


Figure 2: Sample skewness ( $S_s$ ) as a function of  $\tau_{oc}/\Delta t$  for measured data (grey points) and data obtained from simulations of the model (black points) along with standard uncertainties (error bars). In both sets of data, for each condition, 100 or more  $\Delta t$  intervals were extracted from a total of 100 or more cells. The data from simulations is shifted along the x-axis by 0.01, to assist visualization.

## Skewness in Transcription Initiation Kinetics Affects Threshold Crossing by Protein Numbers

We next explore *in silico* the potential role of  $S$  in tuning protein numbers over time, in particular, we study protein number threshold-crossing. For that, we quantify, as a function of  $\tau_{oc}/\Delta t$ , the fraction of time during the course of an *in silico* experiment that the protein numbers equal zero.

In addition to reactions (1)-(6), we accounted for RNA and protein dilution due to cell division as in (Goncalves et al. 2016), assuming mean cell lifetimes of 1h. We modeled 7 conditions, with the same mean RNA and protein numbers but differing in  $S$ , and measured the fraction of time that the protein numbers equal zero during a time series (each series being 100 hours long). We simulated 1000 time series per condition. In each, data from the first hour was omitted to exclude the transient state from the subsequent data analysis. In Fig. 3, we present the results per condition, averaged over all simulations. Namely, we show the value set for  $\tau_{oc}/\Delta t$  in each condition, along with the resulting  $S$  and  $CV^2$  (Fig. 3) of the distribution of time intervals between consecutive RNA productions in individual cells. Also shown is the fraction of time that the model cells are absent of proteins produced by the gene of interest. This quantity is used here as a quantifier of the propensity of this gene expression system to cross a lower-bound threshold in protein numbers over time.

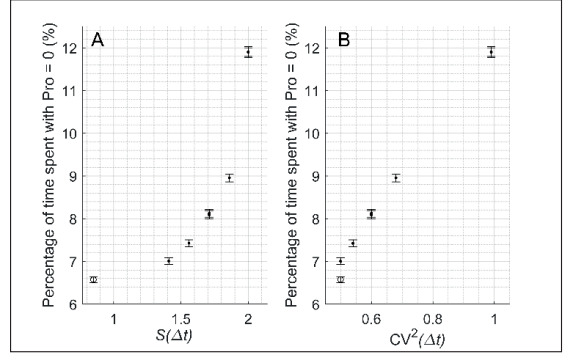


Figure 3: Skewness ( $S$ ) and squared coefficient of variation ( $CV^2$ ) of the distributions of intervals between consecutive productions of RNA molecules in individual model cells differing in  $\tau_{oc}/\Delta t$ . Shown is the relative time that cells are absent of the protein of interest as a function of (A)  $S$  and (B)  $CV^2$  of the time intervals distribution between consecutive RNA productions. Data from stochastic simulations (1000 cells per condition) and from a truncated Gaussian. The error bars are the 90% confidence intervals.

From Fig. 3, as  $\tau_{oc}/\Delta t$  decreases within realistic parameter values (Fig. 2), causing  $S_s$  and  $CV^2$  to change, we find that the fraction of time that proteins are absent from the model cells differs significantly between neighboring conditions.

As the model of transcription does not allow varying  $S$  and  $CV^2$  independently, these results do not suffice to show that it was the change in  $S$  that caused the change in the fraction of time that the model cells spent without proteins produced by the gene of interest. To show that the threshold crossing can be affected by  $S$  alone, we performed an additional set of simulations where  $\Delta t$  follows a Gaussian distribution (truncated at zero) with the same mean and  $CV^2$  as the  $\Delta t$  distribution of condition  $\tau_{oc}/\Delta t = 0.5$ . Results in Fig. 3 show that due to its lower  $S$  (0.85), as predicted, the fraction of time that proteins are absent in model cells with ‘Gaussian-like’ RNA production dynamics differs significantly from the control model cells with RNA production dynamics following (1), including when having the same  $CV^2$ .

We conclude that tuning  $S$  of the  $\Delta t$  distribution has tangible effects in RNA and protein numbers over time, even if the  $CV^2$  is not or is only weakly affected. Importantly, according to model (1), this tuning can occur by regulating  $\tau_{cc}$  and  $\tau_{oc}$ , which are physical properties of the promoter that are both sequence-dependent (McClure 1985) and subject to external regulation, e.g., by transcription factors (Lutz et al. 2001) or global regulatory molecules, such as  $\sigma$  factors.

## DISCUSSION

Here, based on a 2-step stochastic model of transcription and empirical data on the time intervals between consecutive RNA productions in individual cells from various promoters and induction schemes, we first made use of the model to investigate how tuning the relative duration of the steps prior

and after commitment to the open complex formation allows tuning the skewness of the RNA production kinetics. We determined how this skewness changes as a function of these rate-limiting steps and, most interestingly, that it is minimized for equal duration of the two rate-limiting steps, and made use of the empirical data to validate these predictions. Finally, by tuning the skewness of the transcription initiation kinetics within realistic parameter value intervals we observed modifications in protein numbers dynamics strong enough to likely affect the behaviour of small genetic circuits.

Importantly, we expect  $S$  to be tunable via the regulation of  $\tau_{cc}$  and  $\tau_{oc}$ , which are sequence dependent and subject to external regulation, e.g., by transcription factors or global regulatory molecules, such as  $\sigma$  factors. Thus, this regulatory mechanism is expected to be both evolvable as well as adaptable to environmental changes.

In the future, we aim to expand our research, first, by studying more complex models of transcription and investigate how each rate-limiting factor influences the degree of skewness in RNA production kinetics. Second, we aim to investigate on how the tuning of skewness of the component genes allows attaining desired macro dynamics in various genetic circuits.

## REFERENCES

- Bai, L., et al., 2006. "Single-Molecule Analysis of RNA Polymerase Transcription". *Annual Review of Biophysics and Biomolecular Structure*, 35, pp.343–60.
- Bernstein, J. A et al., 2002. "Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays". *Proc. Natl. Acad. Sci. U.S.A.*, 99(15), pp.9697–702.
- Carpenter, J. & Bithell, J., 2000. "Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians". *Statistics in Medicine*, 19, pp.1141–64.
- Chamberlin. M. J, 1974. "The selectivity of preparation". *Psychological Review*, 81(5), pp.442–64.
- Cormack, B.P. et al., 1996. "FACS-optimized mutants of the green fluorescent protein (GFP)". *Gene*, 173, pp.33–8.
- Elowitz, M.B. et al., 2002. "Stochastic Gene Expression in a Single Cell". *Science*, 297, pp.1183–6.
- Gillespie, D.T., 1977. "Concerning the validity of the stochastic approach to chemical kinetics". *Journal of Statistical Physics*, 16(3), pp.311–8.
- Goncalves, N. et al., 2016 "Temperature Dependence of Leakiness of Transcription Repression Mechanisms of *Escherichia coli*". *Proc. of the Computational Methods in Sys. Biol.*, Sept. 21-23, Cambridge, U.K., pp 341-2.
- Häkkinen, A. & Ribeiro, A.S., 2016. "Characterizing rate limiting steps in transcription from RNA production times in live cells". *Bioinformatics*, 32(9), pp.1346–52.
- Häkkinen, A. et al., 2013. "CellAging: A tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*". *Bioinformatics*, 29, pp.1708–9.
- Häkkinen, A. & Ribeiro, A.S., 2015. "Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data". *Bioinformatics*, 31(1), pp.69–75.
- Herbert, K.M., et al., 2008. "Single-molecule studies of RNA polymerase: motoring along". *Annual Review of Biochemistry*, 77, pp.149–76.
- Jishage, M. et al., 1996. "Regulation of RNA polymerase sigma subunit synthesis in *Escherichia coli*: intracellular levels of four species of sigma subunit under various growth conditions". *Journal of Bacteriology*, 178(18), pp.5447–51.
- Joanes, D.N & Gill, C.A. 1998. "Comparing Measures of Sample Skewness and Kurtosis". *Royal Statistical Society*, 47(1), pp.183–9.
- Jones, B. et al 2007. "Is there a Liquid State Machine in the Bacterium *Escherichia Coli*?" *Proceedings of the 2007 IEEE Symposium on Artificial Life*, pp. 187-91.
- Kandavalli, V.K. et al. 2016. "Effects of  $\sigma$  factor competition are promoter initiation kinetics dependent". *BBA - Gene Regulatory Mechanisms*, 1859(10), pp.1281–8.
- Leibler, S. & Kussell, E., 2010. "Individual histories and selection in heterogeneous populations". *Proc. Natl. Acad. Sci. U.S.A.*, 107, pp.13183–8.
- Livak, K.J. & Schmittgen, T.D., 2001. "Analysis of relative gene expression data using real-time quantitative PCR and the 2<sup>-</sup>(Delta Delta C(T)) Method". *Methods*, 25, pp.402–8.
- Lloyd-Price, J. et al., 2016. "Dissecting the stochastic transcription initiation process in live *Escherichia coli*". *DNA Research*, 23(3), pp.203–14.
- Lloyd-Price, J. et al., 2012. "SGNS2: A compartmentalized stochastic chemical kinetics simulator for dynamic cell populations". *Bioinformatics*, 28(22), pp.3004–5.
- López-Maury, L. et al., 2008. "Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation". *Nature Reviews Genetics*, 9(8), pp.583–93.
- Lutz, R. et al., 2001. "Dissecting the functional program of *Escherichia coli* promoters: the combined mode of action of Lac repressor and AraC activator". *Nucleic Acids Research*, 29(18), pp.3873–81.
- MacGillivray, H.L., 1986. "Skewness and Asymmetry: Measures and Orderings". *The Annals of Statistics*, 14(3), pp.994–1011.
- McClure, W.R., 1985. "Mechanism and control of transcription initiation in prokaryotes". *Annual Review of Biochemistry*, 54, pp.171–204.
- Menzel, R. & Gellert, M., 1983. "Regulation of the genes for *E. coli* DNA gyrase: homeostatic control of DNA supercoiling". *Cell*, 34(1), pp.105–13.
- Oliveira, S.M.D. et al., 2016. "Temperature-Dependent Model of Multi-step Transcription Initiation in *Escherichia coli* Based on Live Single-Cell Measurements". *PLoS Computational Biology*, 12(10).
- Raj, A. & van Oudenaarden, A., 2008. "Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences". *Cell*, 135(2), pp.216–26.
- Roussel, M.R. & Zhu, R., 2006. "Stochastic kinetics description of a simple transcription model". *Bulletin of Mathematical Biology*, 68(7), pp.1681–713.
- Shehata, T.E. & Marr, A.G., 1971. "Effect of nutrient concentration on the growth of *Escherichia coli*". *Journal of Bacteriology*, 107(1), pp.210–6.
- Tran, H. et al., 2015. "Kinetics of the cellular intake of a gene expression inducer at high concentrations". *Molecular BioSystems*, 11, pp.2579–87.



# PUBLICATION V

## **Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression**

Sofia Startceva, Vinodh K. Kandavalli, Ari Visa, and Andre S. Ribeiro

Biochimica et Biophysica Acta. Gene Regulatory Mechanisms. 1862, 119–128, 2019  
<https://doi.org/10.1016/j.bbagr.2018.12.005>.

**Publication reprinted with the permission of the copyright holders.**

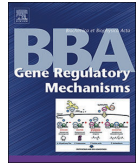






Contents lists available at ScienceDirect

## BBA - Gene Regulatory Mechanisms

journal homepage: [www.elsevier.com/locate/bbagrm](http://www.elsevier.com/locate/bbagrm)

## Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression

Sofia Startceva<sup>a</sup>, Vinodh K. Kandavalli<sup>a</sup>, Ari Visa<sup>b</sup>, Andre S. Ribeiro<sup>a,\*</sup><sup>a</sup> *Laboratory of Biosystem Dynamics, BioMediTech Institute and Faculty of Biomedical Sciences and Engineering, Tampere University of Technology, 33101 Tampere, Finland*<sup>b</sup> *Faculty of Computing and Electrical Engineering, Tampere University of Technology, Tampere 33101, Finland*

## ARTICLE INFO

## Keywords:

Single-cell time-lapse microscopy  
 Transcription initiation  
 RNA and protein numbers  
 Asymmetry and tailedness  
 Threshold crossing

## ABSTRACT

Genetic circuits change the *status quo* of cellular processes when their protein numbers cross thresholds. We investigate the regulation of RNA and protein threshold crossing propensities in *Escherichia coli*. From in vivo single RNA time-lapse microscopy data from multiple promoters, mutants, induction schemes and media, we study the asymmetry and tailedness (quantified by the skewness and kurtosis, respectively) of the distributions of time intervals between transcription events. We find that higher thresholds can be reached by increasing the skewness and kurtosis, which is shown to be achievable without affecting mean and coefficient of variation, by regulating the rate-limiting steps in transcription initiation. Also, they propagate to the skewness and kurtosis of the distributions of protein expression levels in cell populations. The results suggest that the asymmetry and tailedness of RNA and protein numbers in cell populations, by controlling the propensity for threshold crossing, and due to being sequence dependent and subject to regulation, may be key regulatory variables of decision-making processes in *E. coli*.

## 1. Introduction

The gene regulatory networks of bacteria, such as *Escherichia coli*, include network motifs [1,2]. Some of these are responsible for decision-making processes that assist cells in adapting to environmental changes [3,4]. Significant behavioural changes in these motifs usually occur when the numbers of one or more of the component proteins cross thresholds [3]. The underlying mechanisms that define the propensity for the protein numbers of a given gene to cross a specific threshold are not yet fully understood.

In *E. coli*, it is common for the protein numbers to follow the corresponding RNA numbers [5,6]. These are determined by the rates of RNA production and degradation. Interestingly, RNA degradation in *E. coli* appears to be largely independent from the RNA sequence, abundance and metabolic function [7–9], suggesting that little regulation occurs at this stage. Meanwhile, various regulatory mechanisms of transcription have been identified, which usually act at the stage of initiation, suggesting that control over the RNA numbers is exerted at this stage [10–12].

From the dynamics point of view, the regulation of transcription initiation kinetics occurs via the tuning of the time-length of the rate-limiting steps of initiation, respectively, the events prior and after

committing to open complex formation [13–17]. In particular, recent studies [14,16–18] have shown that, under full induction, the in vivo kinetics of these rate-limiting steps, along with supercoiling buildups [19], define, to a great extent, the distribution of time intervals between consecutive RNA production events (here referred to as ‘ $\Delta t$  distribution’). Further, it was shown that not only the first moment (mean), but also the second moment of this distribution (variance) can be tuned by the kinetics of these steps [16,18].

Given this, we hypothesise that, by tuning the kinetics of these rate-limiting steps, one can also tune the third and fourth moments of the  $\Delta t$  distribution (respectively, the skewness and kurtosis). Further, we hypothesise that these two moments can be tuned independently from the mean and coefficient of variation. To test these hypotheses, we perform in vivo time-lapse microscopy employing single-RNA detection by MS2-GFP tagging [20–22], from which we extract the  $\Delta t$  distributions for various promoters, media, induction schemes, growth phases, mutants and a stress condition. Next, for each condition, we estimate their mean, coefficient of variation, skewness and kurtosis. Subsequently, we estimate the kinetics of the rate-limiting steps in each condition and assess their influence on the skewness and kurtosis. Finally, to test whether changing the skewness and kurtosis of the  $\Delta t$  distribution has functional consequences, we measure the corresponding values of the

\* Corresponding author.

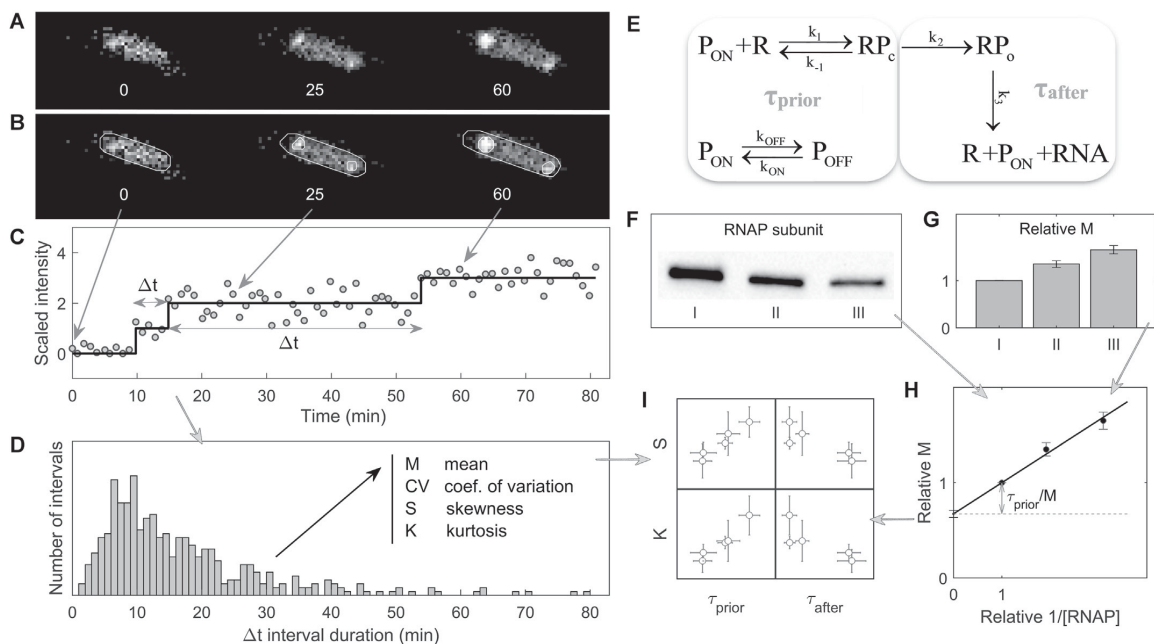
E-mail address: [andre.ribeiro@tut.fi](mailto:andre.ribeiro@tut.fi) (A.S. Ribeiro).<https://doi.org/10.1016/j.bbagrm.2018.12.005>

Received 29 October 2018; Received in revised form 4 December 2018; Accepted 5 December 2018

Available online 14 December 2018

1874-9399/© 2018 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license

<http://creativecommons.org/licenses/by-nc-nd/4.0/>.



**Fig. 1.** Schematic representation of the steps for the analysis of the dynamics of RNA production in individual cells, from in vivo single-RNA, single-cell measurements. (A) Example confocal microscopy images over time of a cell expressing MS2-GFP and the target RNAs. (B) Segmentation of a cell and the MS2-GFP tagged RNA spots within (white lines). (C) Scaled RNA spots intensity over time (grey circles) of the example cell, along with the best-fitting monotonic piecewise-constant curve (black line) from which  $\Delta t$  intervals are estimated. (D) The distribution of time intervals between consecutive RNA production events in individual cells ( $\Delta t$ ) from which mean (M), coefficient of variation (CV), skewness (S) and kurtosis (K) are extracted. (E) Model of transcription initiation. The first box contains the reactions occurring before commitment to open complex formation, with their mean time-length denoted as  $\tau_{\text{prior}}$ . The second box contains the reactions occurring after commitment to open complex formation, with their mean time-length equals  $\tau_{\text{after}}$ . For a detailed description of these reactions and parameters see Supplementary materials and methods, Section 1.6. (F) Western blot image of the RNA polymerase (RNAP) subunit in different media richness. (G) Relative inverse transcription rate of the target gene, measured by qPCR. (H) Relative  $\tau$  plot (Lineweaver–Burk plot [25]) of the inverse of the RNA production rate versus the inverse of the RNAP concentration, [RNAP] for estimating  $\tau_{\text{prior}}$  relative to M. (I) S and K versus  $\tau_{\text{prior}}$  and  $\tau_{\text{after}}$  in different conditions.

skewness and kurtosis of the distributions of single-cell protein expression levels.

## 2. Materials and methods

**Fig. 1** informs on the models and methods used. In short, the main empirical data ( $\Delta t$  distributions) are obtained by measuring when each RNA appears in each cell. Also, we measure the average intracellular RNAP concentration. From these concentrations and the corresponding mean of the  $\Delta t$  distribution in each condition, we estimate the time spent in transcription initiation prior and after commitment to open complex formation ( $\tau_{\text{prior}}$  and  $\tau_{\text{after}}$ , respectively, with their sum equalling  $\Delta t$ ) (model in **Fig. 1E**).

In summary, we first estimate  $\tau_{\text{prior}}/M$  from  $\tau$  plots [23]. For this, the inverse of the RNA production rate relative to the control (as measured by qPCR) is plotted against the inverse of the RNAP concentration relative to the control (as measured by Western blot, Supplementary materials and methods, Section 1.4). Next, a line is fitted to the data. The point where this line intersects the Y axis equals the extrapolated value of the inverse of the transcription rate for an ‘infinite’ RNAP concentration. As such it should equal  $\tau_{\text{after}}/M$ , according to the model in **Fig. 1E**. From this and the value of M, one can calculate  $\tau_{\text{after}}$  and  $\tau_{\text{prior}}$  (Supplementary materials and methods, Section 1.5). Next, from the same  $\Delta t$  distributions, we extract the coefficient of variation, skewness and kurtosis in each condition.

Note that, although genes replicate during the cells lifetime by a process that is not absent of noise and many variables control when

each specific gene is replicated [24], we assume that the rate constants controlling the kinetics of RNA production of our gene of interest (**Fig. 1E**), which is on a single-copy F-plasmid, do not change significantly during the lifetime of the cells. To validate this assumption we compared the distributions of time intervals (between consecutive RNA production events) that started and ended in the first half of the lifetime with intervals that started and ended in the second half (Supplementary results, Section 2.1). From the comparisons of these distributions in each condition (**Table 1**) we conclude that the assumption is sufficiently accurate.

### 2.1. Bacterial strains, plasmids, growth conditions, MS2-GFP tagging system, induction of the reporter and target genes, and measurement conditions

The *E. coli* strain used was DH5 $\alpha$ -PRO (identical to DH5 $\alpha$ Z1 [26] whose genotype is: deoR, endA1, gyrA96, hsdR17(rK – mK+), recA1, relA1, supE44, thi-1,  $\Delta$ (lacZYA-argF)U169,  $\Phi$ 80 $\delta$ lacZ $\Delta$ M15, F-,  $\lambda$ -, PN25/tetR, PlacIq/lacI and SpR. This strain produces, from the chromosome and in abundance, the necessary regulatory proteins for their constructs, namely, LacI, AraC and TetR [26]. E.g. LacI, the main repressor of the control promoter ( $P_{\text{lac/ara-1}}$ ), exists in a concentration much higher than the wild type ( $\sim 3000$  copies vs  $\sim 20$  in wild type [26]). These characteristics allow tight regulation of both target and reporter genes, ensuring that the observed RNAs are due to active transcription and not the result of transcription leakiness (i.e. in the absence of activation). In particular, we measured leaky expression of

**Table 1**

Description of conditions. Shown are the name by which the condition is identified, the target plasmid and corresponding inducer, the reporter plasmid and corresponding inducer, and the media.

Conditions	Target promoter	Target inducers	Reporter promoter	Reporter inducer	Growth media
LA	$P_{lac/ara-1}$	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	1 ×
LA(75)	$P_{lac/ara-1}$	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	0.75 ×
LA(50)	$P_{lac/ara-1}$	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	0.5 ×
LA(ara)	$P_{lac/ara-1}$	1% ara	$P_{LtetO-1}$	100 ng aTc	1 ×
LA(IPTG)	$P_{lac/ara-1}$	1 mM IPTG	$P_{LtetO-1}$	100 ng aTc	1 ×
LA(oxi)	$P_{lac/ara-1}$	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	1 × + 0.6 mM H <sub>2</sub> O <sub>2</sub>
Mut1	$P_{lac/ara-1}$ (Mut-1)	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	1 ×
Mut2	$P_{lac/ara-1}$ (Mut-2)	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	1 ×
Mut3	$P_{lac/ara-1}$ (Mut-3)	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	1 ×
Mut4	$P_{lac/ara-1}$ (Mut-4)	1 mM IPTG + 1% ara	$P_{LtetO-1}$	100 ng aTc	1 ×
tetA	$P_{tetA}$	–	$P_{lac}$	1 mM IPTG	1 ×
tetA(st)	$P_{tetA}$	–	$P_{lac}$	1 mM IPTG	Stationary phase
BAD	$P_{BAD}$	0.1% ara	$P_{lac}$	1 mM IPTG	1 ×
BAD(st)	$P_{BAD}$	0.1% ara	$P_{lac}$	1 mM IPTG	Stationary phase

$P_{lac/ara-1}$ , in the absence of IPTG and arabinose, and found only ~5% or less cells with an MS2-GFP tagged RNA, 2 h after inducing the reporter expressing MS2-GFP.

We also use BW25113, whose genotype is F-, DE(araD-araB)567, lacZ4787(del)::rrnB-3, LAM-, rph-1, DE(rhaD-rhaB)568, hsdR514, which expresses LacI and AraC from the genotype. The absence of TetR allows the Tet promoter to express constitutively.

All cells carry two plasmids: a multi-copy reporter plasmid coding for MS2-GFP under the control of an inducible promoter and a single-copy F-based target plasmid coding for the transcript with multiple MS2-GFP binding sites under the control of another promoter (Table 1). Also, in all target plasmids, we inserted a sequence coding for a red fluorescent protein, between the target promoter and MS2 binding sites. Promoter sequences are specified in Supplementary Fig. S1. Tagged RNAs can be visualized as fluorescent spots [14,20–23] (Fig. 1A).

In general, to observe RNAs tagged by MS2-GFP proteins, cells were grown overnight in LB media with the respective antibiotics at 30 °C in an orbital shaker with aeration of 250 rpm. From the overnight culture, cells were diluted using fresh LB media (unless stated otherwise in Table 1) to an initial OD<sub>600</sub> of 0.05 (measured with a spectrophotometer, Ultrospec 10; GE Healthcare) and incubated at 37 °C at 250 rpm to allow growth until reaching an OD<sub>600</sub> of 0.25. In general, the reporter gene was induced 1 h prior to the target gene, to allow for sufficient MS2-GFP proteins to be produced prior to the appearance of the target RNAs. For a detailed description, see Supplementary materials and methods, Section 1.1. Inducers of target and reporter genes are described in Table 1.

The MS2-GFP RNA tagging technique, proposed in [27], is at present the only direct method to measure time intervals between RNA production events in live, individual cells [14,16,21,22]. This is possible because, first, once appearing, each tagged RNA spot exhibits ‘full’ fluorescence (assuming 1 min interval between microscopy images) [22]. This removes uncertainty in the process of RNA counting as it reduces the possibility for ‘partially fluorescent RNAs’. This uncertainty is further reduced in that, once tagged, the fluorescence of the spots remains near constant for longer than our measurement time (2 h or more) [22]. This provides significant reliability to the quantification of the time-length of intervals between consecutive RNA production events [21].

MS2-GFP tagging affects the spatial organization of the RNAs inside the cell [28]. However, this does not affect the precision of quantification of the intervals between consecutive RNA production events, which are based solely on the total intensity of the MS2-GFP tagged RNAs in a cell, not on their location.

To assess whether this technique has a negative impact on cell physiology, we compared cell growth rates and morphology with and without activating the expression of the MS2-GFP reporter.

Supplementary results in Section 2.2 show that growth rates and cell morphology are not significantly affected by expression of MS2-GFP, in agreement with previous studies [14,23].

Finally, it is also reasonable to assume that MS2-GFP tagging could affect the protein expression levels of the target gene, due to partially interfering with the target RNA (albeit in a different region from the one coding for the red fluorescent protein). We tested this by comparing protein expression levels when and when not activating the expression on MS2-GFP (Supplementary results, Section 2.3). The results confirm that the expression levels of the red fluorescent protein are not perturbed significantly by MS2-GFP tagging (Fig. S9).

Meanwhile, to measure the single cell distributions of RNAP concentration, we used *E. coli* RL1314 strain with fluorescently tagged β' subunits (a kind gift from Robert Landick, University of Wisconsin-Madison) [29]. From the overnight culture, we diluted the cells to an OD<sub>600</sub> of 0.1 in various media richness (Materials and methods) and allowed them to grow to an OD<sub>600</sub> of 0.5 at 37 °C at 250 rpm. Cells were then pelleted by centrifugation and visualized under the microscope.

The plasmids (Table 1) construction and transformation were performed using standard molecular cloning techniques [30]. To construct  $P_{lac/ara-1}$ -mCherry-48 binding sites (bs) mutants, we used a plasmid carrying mCherry followed by a 48bs array in the pBELO vector backbone, originally constructed in [31]. To obtain the mutant promoters (Supplementary Fig. S1), we synthesized new promoter sequences of  $P_{lac/ara-1}$  with specific point mutants with support from Gene Script, USA. Next, we inserted them into the pBELO vector backbone by Gibson Assembly [32], to obtain a single copy F-based plasmid carrying the target region  $P_{lac/ara-1}$ -mCherry-48bs mutants. This product was transferred into competent *E. coli* host cells. The recombinants were selected by antibiotic screening and confirmed with sequence analysis. It is noted that the mutant promoters were selected solely based on that their Δ distributions differed from the one of  $P_{lac/ara-1}$ .

## 2.2. Chemicals

The chemical components of LB media are Tryptone, Yeast extract and NaCl, purchased from LabM (Topley House, Bury, Lancashire, UK). The antibiotics used are Kanamycin 34 µg/ml, Ampicillin 50 µg/ml and Chloramphenicol 35 µg/ml, purchased from Sigma-Aldrich (St. Louis, MO). The inducers used are isopropyl β-D-1-thiogalactopyranoside (IPTG), anhydrotetracycline (aTc) and arabinose (ara), purchased from Sigma-Aldrich. Agarose (Sigma-Aldrich) was used for preparing the microscope gel pads. For PCR, Phusion high-fidelity polymerase and other PCR reagents were purchased from Finnzymes (Finland). Qiagen kits (USA) were used for plasmid isolation. For qPCR, cells were treated with RNA protect bacteria reagent (Qiagen, USA). iScript Reverse Transcription Supermix for cDNA synthesis and iQ SYBR green

supermix for qPCR were purchased from Biorad (USA).

### 2.3. Growth media

In all experiments, we used the LB media and its altered versions, first described in [14]. Namely, we used the following media compositions per 100 ml: 1 g tryptone, 0.5 g yeast extract and 1 g NaCl (pH 7.0), referred to as '1×' (Table 1); 0.75 g tryptone, 0.375 g yeast extract and 1 g NaCl (pH 7.0), referred to as '0.75×'; 0.5 g tryptone, 0.25 g yeast extract and 1 g NaCl (pH 7.0), referred to as '0.5×'; 0.25 g tryptone, 0.125 g yeast extract and 1 g NaCl (pH 7.0), referred to as '0.25×'. These four media are used to attain various mean intracellular RNA polymerase concentrations ([RNAP]) in cell populations, while not affecting normal cell physiology and morphology [14,16,23] (Supplementary Fig. S2A). Additionally, in two conditions, as in [23], we used the stationary phase media obtained by centrifuging the overnight culture of LB media at 10000 rpm for 10 min followed by filtration [23] (growth rates shown in Supplementary Fig. S2B).

### 2.4. qPCR measurements

Cells with target plasmids were harvested by centrifuging them at 8000 ×g for 5 min. To the pelleted cells, twice the amount of RNA protect reagent (Qiagen) was added, followed by the enzymatic lysis with Tris EDTA lysis buffer (pH 8.0). Total RNA was isolated using RNeasy kit (Qiagen) according to the kit instructions. The concentration of RNA was quantified using the Nanovue plus spectrophotometer (GE Healthcare). The RNA samples were treated with DNase to remove the residual DNA, followed by cDNA synthesis, using the iSCRIPT reverse transcription super mix. The cDNA samples were mixed with the qPCR master mix containing iQ SYBR Green Supermix (Biorad) with primers for the target and reference genes. The reaction was carried out in triplicates with the total reaction volume of 20 µl. For quantifying the target gene we used following primers: for mRFP1 (Forward: 5' TACG ACGCCGAGGTCAAG 3' and Reverse: 5' TTGTGGGAGGTGATGCCA 3'), for mCherry (Forward: 5' CACCTACAAGGCCAAGAAGC 3' Reverse: 5' TGGTGTAGTCCTCGTTGTGG 3'). For the reference gene, 16S RNA primers (Forward: 5' CGTCAGTCGTGTTGTGAA 3' and Reverse: 5' GGACCGCTGGCAACAAAG 3') were used. The qPCR experiments were performed by a MiniOpticon Real-time PCR system (Biorad). The following conditions were used during the reaction: 40 cycles of 95 °C for 10 s, 52 °C for 30 s and 72 °C for 30 s for each cDNA replicate. We used no-RT controls and no-template controls to crosscheck non-specific signals and contamination. PCR efficiencies of these reactions were > 95%. The data from CFX Manager TM Software was used to calculate the relative gene expression and its standard error [33].

### 2.5. Microscopy

Measurements of integer-valued numbers of RNAs or of the moments when a new RNA appears in individual cells were conducted using microscopy. For this, a few µl of cells carrying the induced reporter and target plasmids were placed between a coverslip and agarose gel pad (2.5%), with the respective inducers and antibiotics. Next, an FCS2 chamber (Bioptechs) was heated to 37 °C and placed under the microscope. Cells were visualized using a Nikon Eclipse (Ti-E, Nikon) inverted microscope, equipped with a 100× Apo TIRF (1.49 NA, oil) objective. Confocal images were obtained by a C2+ (Nikon) confocal laser-scanning system. For measuring GFP fluorescence (to visualize MS2-GFP 'spots' or RNAP-GFP), we used a 488 nm laser (Melles-Griot) and an emission filter (HQ514/30, Nikon). For time series, confocal images were taken every 1 min for 2 h. Previous studies [14] have shown that these microscopy settings do not cause significant phototoxicity in this strain. Finally, phase-contrast images were obtained simultaneously, with an external phase-contrast system and CCD camera (DS-Fi2, Nikon), every 5 min. Images were extracted using Nikon Nis-

Elements software.

### 2.6. Image and data analysis

Microscopy images were analysed using the software 'CellAging' [34]. For details see Supplementary materials and methods, Section 1.2. From these analysed time-lapse images, we extracted intervals between consecutive RNA production events in individual cells, from which empirical distributions of these intervals ( $\Delta t$  distributions) were obtained (Fig. 1A–D). Data analysis was conducted using tailored algorithms implemented in MATLAB R2017b (MathWorks).

### 2.7. Flow cytometry

Measurements of protein expression levels were conducted using flow cytometry (FC). For this, cells from 5 ml of bacterial culture were diluted 1:10,000 into 1 ml PBS vortexed for 10 s. We performed measurements under various conditions. In each condition, a total of 50,000 cells were observed. Measurements were performed using an ACEA NovoCyte Flow Cytometer (ACEA Biosciences Inc., San Diego, USA) with a yellow laser (561 nm) for excitation and the PE-Texas Red (mCherry) fluorescence detection channel (615/20 nm filter) for emission, at a flow rate of 14 µl/min and a core diameter of 7.7 µm. The PMT voltage of 584 was used for mCherry. To avoid background signal from particles smaller than bacteria, the detection threshold was set to 5000 in FSC-H analyses.

We applied unsupervised gating [35] (implemented in Python 3.6) to the flow cytometry data. We set the fraction of the cells whose data is used in the analysis ( $\alpha$ ) to 0.9, as it was sufficient to remove data points produced by debris, cell doublets and other undesired events. Reducing  $\alpha$  further did not change the results qualitatively.

## 3. Results

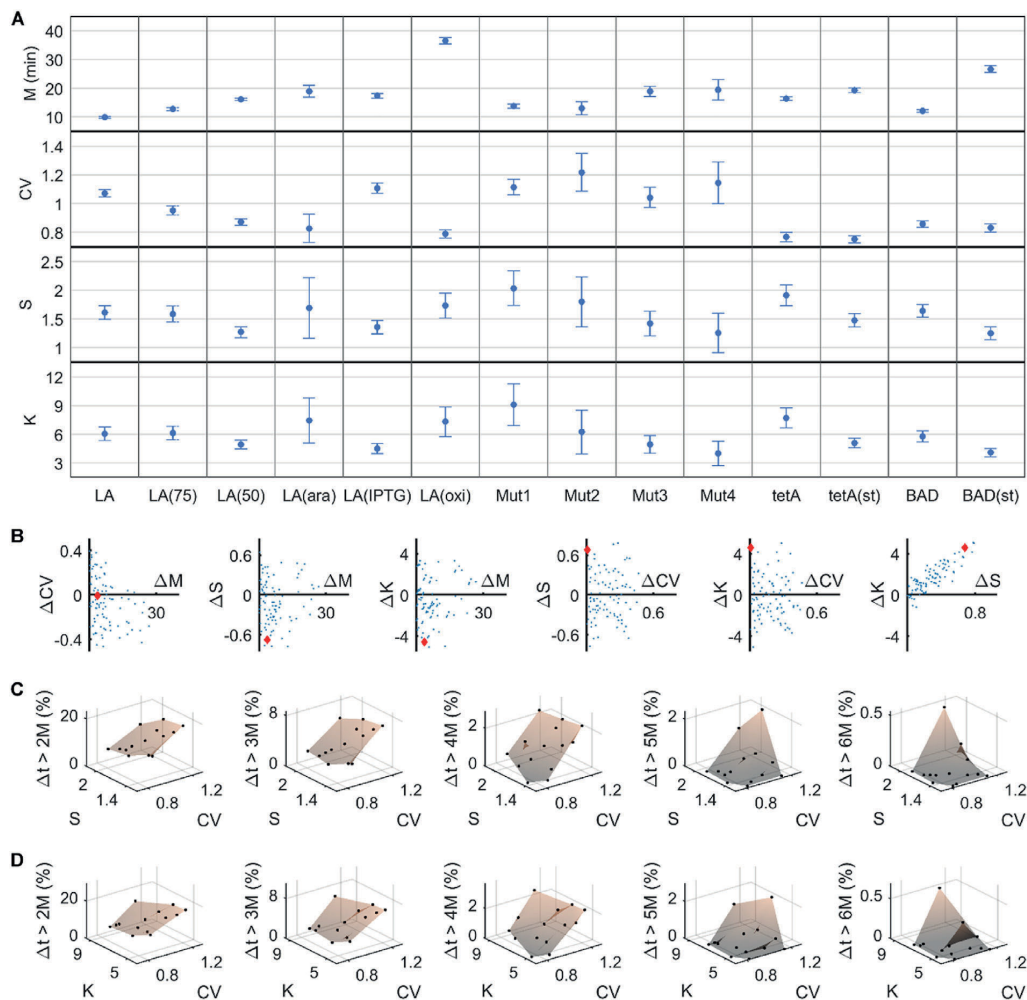
### 3.1. Mean, coefficient of variation, skewness and kurtosis of the distributions of time intervals between consecutive RNA productions in individual cells differ with promoter sequence, regulatory factors and growth conditions

First, we obtained empirical data on the  $\Delta t$  distributions in 14 conditions (see Table 1 for details). These conditions were selected so as to test if the promoter sequence (conditions LA, Mut1, Mut2, Mut3, and Mut4, see Supplementary Fig. S1), regulatory factors such as RNAP and inducer concentrations (conditions LA, LA(75), LA(50), LA(ara), LA(IPTG)), and variables associated to the environment (e.g. media and stress) affect the skewness and kurtosis of the  $\Delta t$  distribution.

Results are shown in Supplementary Fig. S3. From these distributions, we estimated their mean (M), coefficient of variation (CV), skewness (S) and kurtosis (K) (Supplementary materials and methods, Section 1.3). The data was produced from at least 3 repeats per condition. Since no significant differences were found between repeats, the data for each condition were merged. Noteworthy, all target genes used have identical sequences upstream and downstream of the promoter region (Materials and methods). Also, as noted above, as they are integrated into single-copy F-plasmids, not anchored to the membrane, they are not expected to be significantly influenced by transcription halting due to positive supercoiling buildup [19,36].

From Fig. 2A, M and CV differ between conditions. S and K also differ between conditions, but do so following a similar trend to one another. Importantly, changes in S and K seem uncorrelated with the values of M and CV. These results suggest that altering the promoter sequence and/or the active regulation allows altering M, CV and S independently.

Observing only subsets of this data, we find it to be in accordance with the model considered (Fig. 1E). E.g., consider the conditions LA, LA(75) and LA(50), which differ only in [RNAP] [14]. In these, as



**Fig. 2.** Skewness (S) and kurtosis (K) affect the probability of crossing upper-bound thresholds in the time length of the intervals between consecutive RNA production events in individual cells ( $\Delta t$ ). (A) Mean (M), coefficient of variation (CV), S and K of the distribution of  $\Delta t$  intervals (~600 cells per condition). S and K vary independently from M and CV. Error bars denote SEM. (B) Pairwise differences ( $\Delta$ ) in M, CV, S and K between conditions (blue dots). The red diamond is the difference between LA(IPTG) and Mut1 conditions that illustrates how changes in S and K can be independent from changes in M and CV. (C and D) Percentage of  $\Delta t$  intervals (black dots) that are longer than a given threshold (from 2M to 6M) against (C) CV and S, and (D) CV and K. Also shown is the natural neighbour interpolation surface.

[RNAP] decreases, M increases and CV decreases. Meanwhile, S and K decrease (weakly) as [RNAP] decreases. This change is weak enough so that, as shown in the next section, the only significant difference in S is between the two extreme conditions, LA and LA(50), and differences in K are not statistically significant (Supplementary Table S1).

Mutations in  $P_{lac/ara-1}$  (Supplementary Fig. S1) also cause significant behavioural changes. Namely, M, CV and S differ between the mutants independently from each other, and only changes in S and K appear to be correlated. The same is observed when considering only the induction schemes of  $P_{lac/ara-1}$  (LA, LA(ara) and LA(IPTG) conditions). Oxidative stress also affects M, CV, S and K significantly, when compared to the control. Further, comparing the three promoters tested here ( $P_{lac/ara-1}$ ,  $P_{tetA}$  and  $P_{BAD}$ ), again M, CV and S differ in an independent way, and only the differences between conditions in S and K exhibit a similar trend.

Finally, comparing  $P_{tetA}$  and  $P_{BAD}$  in the exponential and stationary

growth phases (Supplementary Fig. S2A,B), we find that both differ significantly in M, S and K with the growth phase. This agrees with the findings in [23], which reported that the kinetics of rate-limiting steps in transcription changes with  $\sigma^{38}$  numbers (even in  $\sigma^{70}$ -dependent promoters). Interestingly, the differences in M, CV, S and K between growth phases are, qualitatively, the same in both promoters, supporting that they have the same cause.

We also tested whether the differences in M, CV, S and K between conditions could be explained by differences between the distributions of cell lifetimes or between the distributions of intracellular RNAP concentrations. The results of this test indicate that the features of the  $\Delta t$  distribution cannot be explained by the features of either these distributions (Supplementary results, Section 2.4; Supplementary Figs. S4A and S5).

**Table 2**

Pearson's correlation coefficient  $r$  (with the corresponding two-tailed  $p$ -value) for all conditions, for the subset 'Mutants', where only the promoter sequence differs between conditions, and for the subset 'Regulatory factors', where only the inducers or RNA polymerase concentrations differ between conditions. For  $p$ -values  $\leq 0.05$ , the null hypothesis that there is no correlation is rejected.

	M vs CV	M vs S	M vs K	CV vs S	CV vs K	S vs K
All conditions	-0.44 (0.12)	-0.19 (0.52)	-0.08 (0.80)	0.01 (0.98)	-0.10 (0.73)	0.94 (< 0.001)
Mutants	-0.12 (0.85)	-0.64 (0.24)	-0.56 (0.32)	0.27 (0.66)	0.07 (0.91)	0.96 (< 0.01)
Regulatory factors	-0.47 (0.43)	-0.24 (0.70)	0.02 (0.98)	-0.17 (0.79)	-0.54 (0.34)	0.91 (0.03)

### 3.2. Promoter sequence and regulatory factors suffice to alter skewness and kurtosis of RNA production kinetics independently from its mean and coefficient of variation

To determine whether changes in M, CV, S and K between conditions are uncorrelated in a statistical sense, we first calculated linear correlations between each pair of these features when considering all 14 conditions (Fig. 2A). Results in Table 2 show no significant correlation between all pairs, except between S and K. The result holds also when applying the Bonferroni-Holm correction for multiple comparisons (the corrected  $p$ -value in the case of S and K is  $< 0.001$ ). Tests for non-linear correlations (Kendall's and Spearman's rank correlation coefficients) give the same qualitative results. While this could be due to the lack of significant changes in M and CV, results in Fig. 2A reject this hypothesis. We thus conclude that all features can differ between conditions in an uncorrelated way, aside from S and K.

We also performed pairwise comparisons of M, CV, S and K between each pair of the 14 conditions. The results (Supplementary Table S1) show statistically significant differences between many pairs of conditions, indicating that all features differ widely between conditions. In detail, one observes that it is possible to alter S and K significantly, while CV is kept unchanged (e.g. between LA(IPTG) and Mut1). Similarly, the same is possible keeping M unchanged (e.g. between LA(50) and tetA).

Next, we quantified the degree with which each feature can differ between conditions while another feature is kept constant. In Fig. 2B we show all pairwise differences in M, CV, S and K between conditions. In all cases, we find that a feature can differ widely while the others remain mostly unchanged, except between S and K.

Finally, we investigated how S and K change as a function of the promoter sequence and the regulatory factors. For this, we considered two subsets of the data above. The first subset ('Mutants') includes the original  $P_{lac/ara-1}$  promoter (LA) and the 4 mutants, specifically 1 single-point mutant (Mut1) and 3 three-point mutants (Mut2, Mut3 and Mut4) (Supplementary Fig. S1). The second subset ('Regulatory factors') includes the control (LA), two conditions with different [RNAP] (LA(75) and LA(50)) and two induction schemes (LA(IPTG) and LA(ara)). From Table 2, we conclude that changes in S (and K), due to point mutations and/or due to altering the concentrations of the regulatory factors, are not correlated to the changes in CV and M.

As before, for both subsets, we tested whether the differences in M, CV, S and K between conditions could be explained by differences between the distributions of cell lifetimes. Again, the results showed that the features of the cell lifetimes distributions cannot explain the features of the  $\Delta t$  distribution (Supplementary results, Section 2.4; Supplementary Fig. S4B,C).

### 3.3. Increasing the skewness and kurtosis of RNA production kinetics enhances the probability of crossing upper bound thresholds in intervals between consecutive RNA production events

Stochastic models of gene expression assuming transcription initiation as a two-step process predict that changing these steps' kinetics can alter the noise in RNA production without changing the mean rate of RNA production [37]. If the intrinsic noise in transcription changes,

so will the probability of crossing thresholds based on RNA numbers. Here we quantify this noise by the CV of the  $\Delta t$  distribution [17,18], because this distribution is not affected by noise in RNA degradation.

If this noise was symmetric around the mean of the  $\Delta t$  distribution, the CV would suffice to estimate the probability of threshold crossing. However, recent results [16,17] suggest that it can be significantly asymmetric. As such, a more accurate estimation of threshold crossing probabilities in RNA numbers requires calculating S and K of the  $\Delta t$  distribution.

To test whether S and K differ significantly between the conditions (Supplementary materials and methods, Section 1.3), we first obtained, for each condition, the fraction of individual  $\Delta t$  intervals that are longer than a given threshold. We considered the thresholds 2 M, 3 M, 4 M, 5 M and 6 M, to eliminate influences by the value of M. Results in Supplementary Table S2 indicate that the fraction of intervals that cross a specific threshold differ between conditions, particularly for higher thresholds.

Next, to determine whether it is CV or S (and K) that is responsible for the differences in threshold crossing probabilities between conditions, we plotted the percentage of intervals in each condition that crossed each threshold against CV and S. We also calculated the natural neighbour interpolation surfaces (using MATLAB R2017b function `scatteredInterpolant` [38]).

Results in Fig. 2C show that for the lower thresholds (2 M and 3 M), varying S does not alter significantly the chance of threshold crossing, while changing CV does. For higher thresholds (4 M and 5 M), both S and CV are relevant. For the highest threshold (6 M), the relevance of S further increases. Equivalent conclusions are reached when considering K instead of S (Fig. 2D).

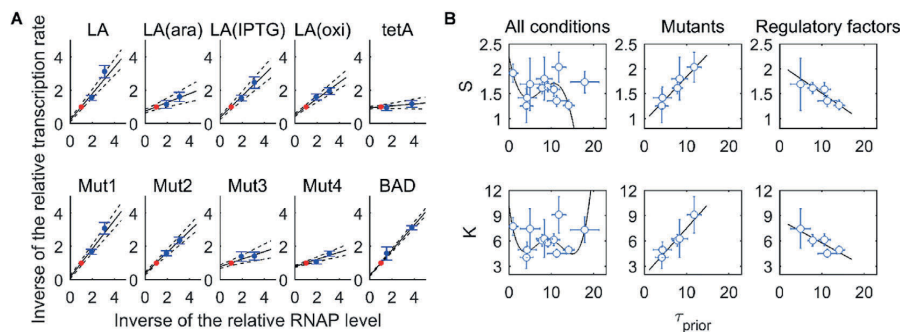
Overall, tuning S and K of the  $\Delta t$  distribution allows altering significantly the probability of crossing upper-bound thresholds in  $\Delta t$  values and, thus, of crossing lower-bound thresholds of RNA numbers in individual cells.

### 3.4. Skewness and kurtosis of RNA production kinetics can be tuned by the rate-limiting steps in transcription initiation

Previous studies have established that CV can be tuned by changing the kinetics of the rate-limiting steps in transcription initiation [14,16,17]. In particular, for example, changing the average time spent in the events prior ( $\tau_{\text{prior}}$ ) and after ( $\tau_{\text{after}}$ ) commitment to open complex formation without changing M, allows tuning noise in RNA production without affecting the rate of this production [16]. We hypothesised that S and K could be similarly regulated.

To test this, for each condition, we first estimated the mean fraction of time spent in the events prior to commitment to open complex formation ( $\tau_{\text{prior}}/M$ ) from  $\tau$  plots (Materials and methods, paragraphs 1–2). Namely, we plotted the inverse of the relative RNA production rate, as measured by qPCR, against the inverse of the relative RNAP concentration, as measured by Western blot (Supplementary materials and methods, Section 1.4). Then, we fitted a line to the data from which we obtain  $\tau_{\text{prior}}/M$  (Fig. 3A and Supplementary Table S3). Finally, from this and the value of M (Fig. 2A), we obtained the absolute values of  $\tau_{\text{prior}}$  and  $\tau_{\text{after}}$  for each condition (Supplementary Table S3).

Cells in the stationary phase (conditions tetA(st) and BAD(st)) are



**Fig. 3.** Skewness ( $S$ ) and kurtosis ( $K$ ) of the distribution of intervals between consecutive RNA production events in individual cells change linearly with the fraction of time spent in events prior to commitment to the open complex formation ( $\tau_{\text{prior}}$ ). (A) Relative  $\tau$  plots. Transcription rates are measured by qPCR, and RNA polymerase (RNAP) levels are measured by Western blot (Supplementary Fig. S2C). Values are shown relative to the control condition (red dot). Error bars denote the standard error. The solid line is the best-fitting line, and the dashed lines denote the standard error of the fit. (B)  $S$  and  $K$  plotted against  $\tau_{\text{prior}}$ . Values plotted for all conditions and for subsets ('Mutants' and 'Regulatory factors'). Error bars denote SEM. The black line is the best-fitting model. The linear relationships are statistically significant when the set of variables allowed to change between conditions is restricted to either the sequence of the promoter or the regulatory factors. When all variables are allowed to differ simultaneously, the best-fitting model is a polynomial of the third or fourth degree.

not considered since, in these conditions,  $\sigma^{38}$  numbers are sufficiently high for the amount of core RNAP enzymes to become a less accurate proxy of the RNAP- $\sigma^{70}$  holoenzymes levels [23]. Additional factors that may differ include potential sRNA regulation [39,40], ppGpp [41], cAMP (see e.g. [42]) contribute to these differences.

We assessed whether  $S$  and  $K$  change with  $\tau_{\text{prior}}$ . For this, we plotted  $S$  and  $K$  against  $\tau_{\text{prior}}$  in each condition (Fig. 3B) and performed likelihood ratio tests (at significance level of 0.05) between the best-fit polynomial models (using weighted total least squares approach [14,43]) with degrees ranging from 0 to  $N-1$ , with  $N$  being the number of conditions ( $p$ -values are shown in Supplementary Table S4). We also tested whether the data can be better explained by a model where  $\tau_{\text{prior}}$  does not differ between conditions, by performing a likelihood ratio test between this model and the selected best-fitting polynomial (Supplementary Table S4). For both  $S$  and  $K$ , the zero-degree and the first-degree polynomial models, as well as the models where  $\tau_{\text{prior}}$  is constant, are rejected in favour of higher-degree polynomials.

The fact that  $S$  and  $K$  are best fit by, respectively, third and fourth degree polynomials (that still do not explain all data points) illustrates the level of complexity of the data. This is likely due to the conditions differing in several factors (promoter, induction scheme, etc.). We thus next consider, as above, the subsets 'Mutants' and 'Regulatory factors'. For each, we perform, also as above, likelihood ratio tests to determine the best fitting models (Supplementary Table S4). In both subsets, a 1st degree model is preferred.

Meanwhile, from the Pearson's correlation coefficient (with the corresponding two-tailed  $p$ -value) between  $\tau_{\text{prior}}$  and skewness ( $S$ ) and kurtosis ( $K$ ), for the subsets 'Mutants' and 'Regulatory factors', we find a significant correlation in all cases (absolute correlation values above 0.85 and  $p$ -values  $\leq 0.05$ ), except for  $K$  in 'Regulatory factors', where the  $p$ -value equals 0.06. Overall, the results suggest that, similarly to  $M$  and  $CV$ , tuning  $\tau_{\text{prior}}$  can regulate  $S$  and  $K$ . This implies that the lower bound threshold crossing probability of RNA numbers over time can be tuned.

Next, we performed the same analysis for changing  $\tau_{\text{after}}$  and  $\tau_{\text{prior}}/M$ . Contrary to when considering  $\tau_{\text{prior}}$ , the results (Supplementary Fig. S6 and Supplementary Tables S5-S6) do not allow establishing statistically significant relationships (also the  $p$ -values from the Pearson's correlation were larger than 0.05).

Interestingly, the linear relationships of  $S$  and  $K$  with  $\tau_{\text{prior}}$  are positive in the subset 'Mutants' and negative in the subset 'Regulatory factors'. This strongly indicates that  $\tau_{\text{prior}}$  is not the only parameter defining these features. Namely, we hypothesise that these relationships

may depend on what causes  $\tau_{\text{prior}}$  to differ between the conditions. For instance, in one subset, the difference may be due to differences in the mean time required by the RNAP to complete a closed complex formation, while in the other subset the differences may be in the number of times that the RNAP fails to commit to the open complex formation. These potential differences could be accounted for in the model by tuning  $k_1$ ,  $k_{-1}$  and  $k_2$  (Supplementary materials and methods, Section 1.6), but cannot be detected by the measurements conducted here. Future work is needed to test this hypothesis.

### 3.5. Skewness and kurtosis of the RNA production kinetics and of the distribution of protein expression levels in individual cells are negatively correlated

To assess if changes in  $S$  and  $K$  of the  $\Delta t$  distribution could affect the phenotypic distribution of cell populations, we next investigate whether these changes result in significant changes in the distribution of protein expression levels of a cell population. This is expected given the known coupling between transcription and translation in prokaryotes [44–46]. Nevertheless, it is reasonable to assume that noise in the stochastic process of translation (e.g. on the time to be completed once initiated) would render changes in  $S$  and  $K$  ineffectual on protein expression levels. A model of gene expression in prokaryotes accounting for the coupling between the two processes is shown in Supplementary materials and methods, Section 1.6.

We first tested whether the mean protein expression levels of the cell populations follow their mean RNA numbers. For that, we measured RNA numbers (by microscopy) and protein mean expression levels (by flow cytometry) produced under the control of  $P_{\text{lac/ara-1}}$  for various induction conditions. We expect the same relationship in all other constructs used here, as they have identical sequences following the promoter sequence. Results in Supplementary Fig. S7 show that the average number of proteins in a cell population follows the average RNA numbers.

Given this, since  $M$  of the  $\Delta t$  distribution is negatively correlated with the mean RNA numbers of the cell population, one can expect it to also be negatively correlated to the mean number of proteins. Using the same promoter as a case-study, we tested whether the skewness and kurtosis of the distribution of protein expression levels of a cell population are sensitive to the induction strength. For this, we measured the total fluorescence intensity level of the proteins expressed by  $P_{\text{lac/ara-1}}$  in individual cells for various induction levels using flow cytometry (Materials and methods). From these, for each induction level, we

obtained the distribution of fluorescence of individual cells (in arbitrary units). For each of these distributions, we estimated the mean ( $M_p$ ), skewness ( $S_p$ ) and kurtosis ( $K_p$ ) as previously (Supplementary materials and methods, Section 1.3). From Supplementary Fig. S8, we find that  $S_p$  and  $K_p$  can differ with induction strength. Also, it is possible to have, for similar values of  $M_p$ , significantly different values of  $S_p$  and  $K_p$  (e.g. conditions 0 to 25  $\mu$ M). Further, conditions differing in  $M_p$  can have similar values of  $S_p$  and  $K_p$  (beyond 100  $\mu$ M). Overall, we find that, as for the  $\Delta t$  distributions,  $S_p$  and  $K_p$  can change independently from  $M_p$  and vice versa.

Next, we investigate whether changes in  $S$  and  $K$  of the  $\Delta t$  distribution due to changing the promoter sequence or its regulation reflect on the distribution of protein expression levels, as expected from the model. For this, we consider, respectively, the subsets ‘Mutations’ and ‘Induction schemes’. We note that, within these subsets, the cells are grown under identical culture conditions and do not differ in their fundamental physiology, and are therefore not expected to differ in, e.g., ribosome population and/or in any other global gene expression regulators, such as [RNAP] or  $\sigma$  factors. For these reasons, here we do not consider the other conditions in Table 1, as the translation rate or protein maturation time may differ significantly from the control.

For each condition considered, we measured the fluorescence intensity from the target proteins by flow cytometry (Materials and methods) and obtained the single-cell distributions of protein fluorescence intensity. Next, we estimated its  $M_p$  (in arbitrary units),  $S_p$  and  $K_p$ , as previously. We also measured  $M_p$  for cells with an uninduced  $P_{lac/ara-1}$  to obtain a reference point for the values of  $M_p$ . In this regard, the LA(ara) condition was not included in the subsequent analysis since, for unknown reasons, its protein expression levels were not significantly above those of the uninduced  $P_{lac/ara-1}$  (Fig. 4).

In Fig. 4, we show  $M_p$ ,  $S_p$  and  $K_p$  plotted against  $M$ ,  $S$  and  $K$ , respectively, along with the best-fitting models obtained by likelihood ratio tests (Supplementary Table S7). In all cases, the linear model is preferred. We also calculated the Pearson’s correlation coefficient for each case. The results agree with the likelihood ratio tests. Namely, there are strong, statistically significant ( $p$ -values  $\leq 0.05$ ), negative correlations between  $M$  and  $M_p$  ( $-0.82$ ) and between  $S$  and  $S_p$  ( $-0.86$ ). Between  $K$  vs  $K_p$  the negative correlation is also strong ( $-0.70$ ), but the  $p$ -value is 0.12, likely due to higher uncertainty. From the statistically significant linear relationships, we conclude that the differences in skewness and kurtosis of the  $\Delta t$  distribution between conditions result in statistically significant differences between the skewness and kurtosis of the corresponding protein distributions, in a manner that is consistent with the model. As a side note, our data does not allow investigating whether a similar (expected) correlation exist in the case of CV and  $CV_p$ , since LA, LA(IPTG), and the mutant promoters have CV values that cannot be distinguished in a statistical sense (Supplementary Table S1).

Finally, to assess if the values of  $M$  could explain the values of  $S_p$  and  $K_p$ , we performed likelihood ratio tests (as above) between  $M$  and  $S_p$  and between  $M$  and  $K_p$ . A polynomial model of the 1st order was

rejected in both cases ( $p$ -values equal 0.04 and 0.02, respectively). Also, we failed to find linear correlations ( $p$ -values equal 0.06 and 0.25, respectively). We conclude that  $M$  is not correlated with either  $S_p$  or  $K_p$ , as expected from the lack of the correlation between  $M$  and  $S$  or  $K$ .

#### 4. Discussion and conclusions

Previous research have established that bacterial transcription is mostly regulated at the stage of initiation [10–12,47]. This regulation, e.g. by transcription factors and  $\sigma$  factors, affects the mean and variance in RNA and protein numbers [10–12,19,48]. From the dynamics point of view, these and similar regulatory molecules were shown to have direct effect on the kinetics of the rate-limiting steps in transcription initiation of a gene (assessed here by  $\tau_{prior}$  and  $\tau_{after}$ ), resulting in changes in the mean and variance of its distribution of intervals between consecutive RNA production events in individual cells ( $\Delta t$  distribution) [14,23].

Here we provided evidence that the fraction of cells that reach high thresholds in RNA and protein numbers of an externally regulated gene can be tuned by altering the skewness and kurtosis of its  $\Delta t$  distribution. Also, we showed that this can be achieved without significantly altering the mean and CV of this distribution. Further, this regulation is possible by tuning  $\tau_{prior}$  and  $\tau_{after}$  alone which can be altered by changing the promoter sequence, the induction scheme, or the intracellular RNAP concentration.

On the other hand, we did not find significant evidence that the skewness and kurtosis could be altered independently from one another. Instead, they exhibit a strong positive correlation (Fig. 2B,  $\Delta K$  vs  $\Delta S$ , and Table 2). We suggest that this may be due to the variability of the time length between transcription events along with the existence of mechanical constraints imposed by the transcription machinery. This variability is visible in Fig. S3, which shows that the distributions of intervals between transcriptions are broad, with several intervals having a short time-length. This limits how much the kurtosis of this distribution can increase by increasing the tail on the left side. This limit does not exist on the right side. Thus, increasing the kurtosis of one of these distributions by increasing the size of the right tail cannot be easily compensated on the left side so that the skewness remains unaltered.

Regulation of asymmetry and tailedness of gene expression, so far, has only been considered in the context of small genetic circuits or complex regulatory pathways (e.g. [3]). Given the above, our findings suggest that regulatory mechanisms of individual genes suffice for this regulation as well. In particular, based on the data from the conditions in Table 1, we found statistically significant linear relationships between  $\tau_{prior}$  and the skewness and kurtosis of the  $\Delta t$  distribution, provided that either only the promoter sequence or the regulatory factors (i.e. inducers and RNAP concentrations) differ between the conditions. We hypothesise that relationships more complex than linear are also possible, if more than one parameter is allowed to change. E.g. in the future it would be of interest to investigate whether the data in Fig. 3B

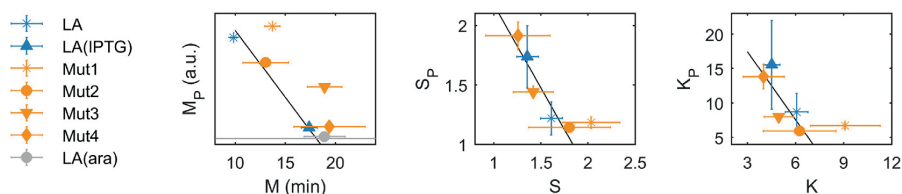


Fig. 4. Mean ( $M$ ), skewness ( $S$ ), and kurtosis ( $K$ ) of the distribution of protein expression levels in individual cells change linearly with the corresponding features of the distribution of time intervals between consecutive RNA production events ( $\Delta t$  distribution). (From left to right)  $M_p$ ,  $S_p$  and  $K_p$  of the single-cell distributions of protein levels against the corresponding feature of the  $\Delta t$  distributions (extracted from Fig. 2A). Error bars denote SEM (in some cases, the SEM is too small to produce visible error bars). The solid line is the best fitting model. On the left plot, the horizontal grey line corresponds to  $M_p$  for an uninduced  $P_{lac/ara-1}$  which is used as a reference point (SEM is too small to be represented).  $M_p$  of LA(ara) is not considered in model fitting.



could be better explained by consider both  $\tau_{\text{prior}}$  and  $\tau_{\text{after}}$  simultaneously. Nevertheless, the linear relationships found here are evidence that the skewness and kurtosis are evolvable (i.e. sequence dependent) and adaptable (i.e. subject to regulation). Meanwhile, the strong correlation between RNA production kinetics and single-cell distribution of protein levels suggests that tuning these skewness and kurtosis can have a significant impact on the phenotypic distribution of the cell population.

It is well known that the two rate-limiting steps of transcription initiation here considered (i.e. the events prior and after commitment to open complex formation) are composed of specific ‘sub-steps’, such as promoter escape [49–51], reversibility of the closed complex formation and isomerization [13,52,53]. Further developments in the dissection techniques of the *in vivo* kinetics of these sub-steps during transcription initiation should allow characterising, in greater detail, their contributions to the regulation of the skewness and kurtosis of the distributions of RNA production kinetics and corresponding protein numbers. This should also allow establishing precise methods for tuning the skewness and kurtosis of these distributions.

It is worth noting that the findings here reported do not discard the importance of other mechanisms of regulation of protein numbers in *E. coli*, such as regulation by sRNAs [39,40,54]. Here we did not consider this mechanism since all target genes studied shared the same elongation region. It will be of interest to study whether this post-transcription regulation process also allows tuning the skewness and kurtosis of single-cell distributions of protein numbers, particularly given its known effects on the cell-to-cell variability in protein numbers [55,56] and protein numbers’ threshold-crossing propensities [39,57].

Finally, while a strict relationship between the skewness and kurtosis in the RNA and protein numbers was established here, the implications of these findings in the context of the qualitative behaviour of genetic circuits remain to be demonstrated. We expect the amplitude of these effects to differ with the circuit topology, as in the case of mean and variance [58–60]. If the effects are significant, direct regulation of these features in genetic circuits (by tuning the rate limiting steps of the component genes) should allow a more precise control of their kinetics, towards enhancing their robustness to fluctuations in molecular numbers or environmental changes, and sensitivity to external signals.

## Funding

Work supported by Tampere University of Technology Graduate School Grant (Finland) [to S.S.]; Pirkanmaa Regional Fund [to V.K.K.]; Academy of Finland [295027, 305342 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. Funding for open access charge: Academy of Finland [295027].

## Conflict of interest

The authors declare that they have no conflict of interest.

## Transparency document

The [Transparency document](#) associated with this article can be found, in online version.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.bbagem.2018.12.005>.

## References

[1] S.S. Shen-Orr, R. Milo, S. Mangan, U. Alon, Network motifs in the transcriptional

- regulation network of *Escherichia coli*, *Nat. Genet.* 31 (2002) 64–68, <https://doi.org/10.1038/ng881>.
- [2] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, *Science* 298 (2002) 824–827, <https://doi.org/10.1126/science.298.5594.824>.
- [3] U. Alon, Network motifs: theory and experimental approaches, *Nat. Rev. Genet.* 8 (2007) 450–461, <https://doi.org/10.1038/nrg2102>.
- [4] E. Kussell, S. Leibler, Phenotypic diversity, population growth, and information in fluctuating environments, *Science* 309 (2005) 2075–2078, <https://doi.org/10.1126/science.1114383>.
- [5] Y. Liu, A. Beyer, R. Aebersold, On the dependency of cellular protein levels on mRNA abundance, *Cell* 165 (2016) 535–550, <https://doi.org/10.1016/j.cell.2016.03.014>.
- [6] C. Vogel, E.M. Marcotte, Insights into the regulation of protein abundance from proteomic and transcriptomic analyses, *Nat. Rev. Genet.* 13 (2012) 227–232, <https://doi.org/10.1038/nrg3185>.
- [7] J.A. Bernstein, A.B. Khodursky, P.-H. Lin, S. Lin-Chao, S.N. Cohen, Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays, *Proc. Natl. Acad. Sci. U. S. A.* 99 (2002) 9697–9702, <https://doi.org/10.1073/pnas.112318199>.
- [8] H. Chen, K. Shiroguchi, H. Ge, X.S. Xie, Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*, *Mol. Syst. Biol.* 11 (2015) 781, <https://doi.org/10.15252/msb.20145794>.
- [9] M.P. Deutscher, Degradation of RNA in bacteria: comparison of mRNA and stable RNA, *Nucleic Acids Res.* 34 (2006) 659–666, <https://doi.org/10.1093/nar/gkj472>.
- [10] S.M. McLeod, R.C. Johnson, Control of transcription by nucleoid proteins, *Curr. Opin. Microbiol.* 4 (2001) 152–159, [https://doi.org/10.1016/S1369-5274\(00\)00181-8](https://doi.org/10.1016/S1369-5274(00)00181-8).
- [11] E.F. Ruff, A.C. Drennan, M.W. Capp, M.A. Poulos, I. Artsimovitch, T.M. Record Jr., *E. coli* RNA polymerase determinants of open complex lifetime and structure, *J. Mol. Biol.* 247 (2015) 2435–2450, <https://doi.org/10.1016/j.jmb.2015.05.024>.
- [12] D.F. Browning, S.J.W. Busby, Local and global regulation of transcription initiation in bacteria, *Nat. Rev. Microbiol.* 14 (2016) 638–650, <https://doi.org/10.1038/nrmicro.2016.103>.
- [13] P.L. deHaseth, M.L. Zupancic, T.M. Record Jr., RNA polymerase-promoter interactions: the comings and goings of RNA polymerase, *J. Bacteriol.* 180 (1998) 3019–3025 (PMID: 9620948).
- [14] 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 (2016) 203–214, <https://doi.org/10.1093/dnares/dsw009>.
- [15] W.R. McClure, Rate-limiting steps in RNA chain initiation, *Proc. Natl. Acad. Sci. U. S. A.* 77 (1980) 5634–5638, <https://doi.org/10.1073/pnas.77.10.5634>.
- [16] J. Mäkelä, V. Kandavalli, A.S. Ribeiro, Rate-limiting steps in transcription dictate sensitivity to variability in cellular components, *Sci. Rep.* 7 (2017) 10588, <https://doi.org/10.1038/s41598-017-11257-2>.
- [17] S.M.D. Oliveira, A. Häkkinen, J. Lloyd-Price, H. Tran, V. Kandavalli, A.S. Ribeiro, Temperature-dependent model of multi-step transcription initiation in *Escherichia coli* based on live single-cell measurements, *PLoS Comput. Biol.* 12 (2016) e1005174, <https://doi.org/10.1371/journal.pcbi.1005174>.
- [18] A. Häkkinen, A.S. Ribeiro, Characterizing rate limiting steps in transcription from RNA production times in live cells, *Bioinformatics* 32 (2016) 1346–1352, <https://doi.org/10.1093/bioinformatics/btv744>.
- [19] S. Chong, C. Chen, H. Ge, X.S. Xie, Mechanism of transcriptional bursting in bacteria, *Cell* 158 (2014) 314–326, <https://doi.org/10.1016/j.cell.2014.05.038>.
- [20] I. Golding, J. Paulsson, S.M. Zawilski, E.C. Cox, Real-time kinetics of gene activity in individual bacteria, *Cell* 123 (2005) 1025–1036, <https://doi.org/10.1016/j.cell.2005.09.031>.
- [21] A. Häkkinen, A.S. Ribeiro, Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data, *Bioinformatics* 31 (2015) 69–75, <https://doi.org/10.1093/bioinformatics/btu592>.
- [22] 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 (2015) 2579–2587, <https://doi.org/10.1039/C5MB00244C>.
- [23] V.K. Kandavalli, H. Tran, A.S. Ribeiro, Effects of  $\sigma$  factor competition are promoter initiation kinetics dependent, *Biochim. Biophys. Acta* 1859 (2016) 1281–1288, <https://doi.org/10.1016/j.bbagem.2016.07.011>.
- [24] J.R. Peterson, J.A. Cole, J. Fei, T. Ha, Z.A. Luthey-Schulten, Effects of DNA replication on mRNA noise, *Proc. Natl. Acad. Sci. U. S. A.* 112 (2015) 15886–15891, <https://doi.org/10.1073/pnas.1516246112>.
- [25] H. Lineweaver, D. Burk, The determination of enzyme dissociation constants, *J. Am. Chem. Soc.* 56 (1934) 658–666, <https://doi.org/10.1021/ja01318a036>.
- [26] R. Lutz, H. Bujard, Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and AraC/I<sub>1</sub>-I<sub>2</sub> regulatory elements, *Nucleic Acids Res.* 25 (1997) 1203–1210, <https://doi.org/10.1093/nar/25.6.1203>.
- [27] I. Golding, E.C. Cox, RNA dynamics in live *Escherichia coli* cells, *Proc. Natl. Acad. Sci. U. S. A.* 101 (2004) 11310–11315, <https://doi.org/10.1073/pnas.0404443101>.
- [28] A. Gupta, J. Lloyd-Price, R. Neeli-Venkata, S.M.D. Oliveira, A.S. Ribeiro, *In vivo* kinetics of segregation and polar retention of MS2-GFP-RNA complexes in *Escherichia coli*, *Biophys. J.* 106 (2014) 1928–1937, <https://doi.org/10.1016/j.bpj.2014.03.035>.
- [29] B.P. Bratton, R.A. Mooney, J.C. Weisshaar, Spatial distribution and diffusive motion of RNA polymerase in live *Escherichia coli*, *J. Bacteriol.* 193 (2011) 5138–5146, <https://doi.org/10.1128/JB.00198-11>.
- [30] J. Sambrook, E.F. Fritsch, T. Maniatis, *Molecular Cloning: A Laboratory Manual*, 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York,

- 0879693096, 1989.
- [31] N.S.M. Goncalves, L. Martins, H. Tran, S.M.D. Oliveira, R. Neeli-Venkata, J.M. Fonseca, A.S. Ribeiro, *In vivo* single-molecule dynamics of transcription of the viral T7 P1 10 promoter in *Escherichia coli*, The 8th International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2016), 2016 978-1-61208-488-6, pp. 9–15.
- [32] D.G. Gibson, L. Young, R.-Y. Chuang, J.C. Venter, C.A. Hutchison III, H.O. Smith, Enzymatic assembly of DNA molecules up to several hundred kilobases, *Nat. Methods* 6 (2009) 343–345, <https://doi.org/10.1038/nmeth.1318>.
- [33] K.J. Livak, T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the  $2^{-\Delta\Delta C_t}$  method, *Methods* 25 (2001) 402–408, <https://doi.org/10.1006/meth.2001.1262>.
- [34] A. Häkkinen, 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 (2013) 1708–1709, <https://doi.org/10.1093/bioinformatics/btt194>.
- [35] M. Razo-Mejia, S.L. Barnes, N.M. Belliveau, G. Chure, T. Einav, M. Lewis, R. Phillips, Tuning transcriptional regulation through signaling: a predictive theory of allosteric induction, *Cell Syst.* 6 (2018) 456–469.e10, <https://doi.org/10.1016/j.cels.2018.02.004>.
- [36] J.D. Boeke, P. Model, A prokaryotic membrane anchor sequence: carboxyl terminus of bacteriophage  $\phi 1$  gene III protein retains it in the membrane, *Proc. Natl. Acad. Sci. U. S. A.* 79 (1982) 5200–5204, <https://doi.org/10.1073/pnas.79.17.5200>.
- [37] S.M.D. Oliveira, M.N.M. Bahrudeen, S. Startceva, A.S. Ribeiro, Estimating effects of extrinsic noise on model genes and circuits with empirically validated kinetics, in: M. Pelillo, I. Poli, A. Roli, R. Serra, D. Slanzi, M. Villani (Eds.), *Artificial Life and Evolutionary Computation. WIVACE 2017*, Springer, 2018, pp. 181–193, [https://doi.org/10.1007/978-3-319-78658-2\\_14](https://doi.org/10.1007/978-3-319-78658-2_14).
- [38] I. Amidror, Scattered data interpolation methods for electronic imaging systems: a survey, *J. Electron. Imaging* 11 (2002) 157–176, <https://doi.org/10.1117/1.1455013>.
- [39] E. Levine, Z. Zhang, T. Kuhlman, T. Hwa, Quantitative characteristics of gene regulation by small RNA, *PLoS Biol.* 5 (2007) e229, <https://doi.org/10.1371/journal.pbio.0050229>.
- [40] E.G.H. Wagner, P. Romby, Small RNAs in bacteria and archaea: who they are, what they do, and how they do it, *Adv. Genet.* 90 (2015) 133–208, <https://doi.org/10.1016/bs.adgen.2015.05.001>.
- [41] C. Condon, C. Squires, C.L. Squires, Control of rRNA transcription in *Escherichia coli*, *Microbiol. Rev.* 59 (1995) 623–645 (PMID: 8531889).
- [42] C.M. Johnson, R.F. Schleif, *In vivo* induction kinetics of the arabinose promoters in *Escherichia coli*, *J. Bacteriol.* 177 (1995) 3438–3442 (PMID: 7768852).
- [43] M. Krystek, M. Anton, A weighted total least-squares algorithm for fitting a straight line, *Meas. Sci. Technol.* 18 (2007) 3438–3442, <https://doi.org/10.1088/0957-0233/18/11/025>.
- [44] O. Dahan, H. Gingold, Y. Pilpel, Regulatory mechanisms and networks couple the different phases of gene expression, *Trends Genet.* 27 (2011) 316–322, <https://doi.org/10.1016/j.tig.2011.05.008>.
- [45] C. Yanofsky, Attenuation in the control of expression of bacterial operons, *Nature* 289 (1981) 751–758, <https://doi.org/10.1038/289751a0>.
- [46] S. Proshkin, A.R. Rahmouni, A. Mironov, E. Nudler, Cooperation between translating ribosomes and RNA polymerase in transcription elongation, *Science* 328 (2010) 504–508, <https://doi.org/10.1126/science.1184939>.
- [47] D.F. Browning, S.J.W. Busby, The regulation of bacterial transcription initiation, *Nat. Rev. Microbiol.* 2 (2004) 57–65, <https://doi.org/10.1038/nrmicro787>.
- [48] W.R. McClure, Mechanism and control of transcription initiation in prokaryotes, *Annu. Rev. Biochem.* 54 (1985) 171–204, <https://doi.org/10.1146/annurev.bi.54.070185.001131>.
- [49] D. Duchi, D.L.V. Bauer, L. Fernandez, G. Evans, N. Robb, L.C. Hwang, K. Gryte, A. Tomescu, P. Zawadzki, Z. Morichaud, et al., RNA polymerase pausing during initial transcription, *Mol. Cell* 63 (2016) 939–950, <https://doi.org/10.1016/j.molcel.2016.08.011>.
- [50] L.M. Hsu, Promoter escape by *Escherichia coli* RNA polymerase, *EcoSal Plus* 3 (2008) 1–16, <https://doi.org/10.1128/ecosalplus.4.5.2.2>.
- [51] A.N. Kapanidis, E. Margeat, S.O. Ho, E. Kortkhonja, S. Weiss, R.H. Ebright, Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism, *Science* 314 (2006) 1144–1147, <https://doi.org/10.1126/science.1131399>.
- [52] L.M. Hsu, Promoter clearance and escape in prokaryotes, *Biochim. Biophys. Acta* 1577 (2002) 191–207, [https://doi.org/10.1016/S0167-4781\(02\)00452-9](https://doi.org/10.1016/S0167-4781(02)00452-9).
- [53] T.M. Record Jr., W.S. Reznikoff, M.L. Craig, K.L. McQuade, P.J. Schlax, *Escherichia coli* RNA polymerase ( $E\sigma^{70}$ ), promoters, and the kinetics of the steps of transcription initiation, in: F.C. Neidhardt, R. Curtiss, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Reznikoff, M. Riley, D. Schneider, H.E. Umberger (Eds.), *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, 2nd ed., ASM press, Washington, DC, 1555810845, 1996, pp. 792–821.
- [54] G. Storz, J. Vogel, K.M. Wassarman, Regulation by small RNAs in bacteria: expanding frontiers, *Mol. Cell* 43 (2011) 880–891, <https://doi.org/10.1016/j.molcel.2011.08.022>.
- [55] R. Arbel-Goren, A. Tal, T. Friedlander, S. Meshner, N. Costantino, D.L. Court, J. Stavans, Effects of post-transcriptional regulation on phenotypic noise in *Escherichia coli*, *Nucleic Acids Res.* 41 (2013) 4825–4834, <https://doi.org/10.1093/nar/gkt184>.
- [56] R. Arbel-Goren, A. Tal, B. Parasar, A. Dym, N. Costantino, J. Muñoz-García, D.L. Court, J. Stavans, Transcript degradation and noise of small RNA-controlled genes in a switch activated network in *Escherichia coli*, *Nucleic Acids Res.* 44 (2016) 6707–6720, <https://doi.org/10.1093/nar/gkw273>.
- [57] P. Mehta, S. Goyal, N.S. Wingreen, A quantitative comparison of sRNA-based and protein-based gene regulation, *Mol. Syst. Biol.* 4 (2016) 221, <https://doi.org/10.1038/msb.2008.58>.
- [58] E.D. Cameron, J.J. Collins, Tunable protein degradation in bacteria, *Nat. Biotechnol.* 32 (2014) 1276–1281, <https://doi.org/10.1038/nbt.3053>.
- [59] L.G. Morelli, F. Jülicher, Precision of genetic oscillators and clocks, *Phys. Rev. Lett.* 98 (2007) 228101, <https://doi.org/10.1103/PhysRevLett.98.228101>.
- [60] R. Zhu, A.S. Ribeiro, D. Salahub, S.A. Kauffman, Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models, *J. Theor. Biol.* 246 (2007) 725–745, <https://doi.org/10.1016/j.jtbi.2007.01.021>.

**Supplementary Material for:**

# **Regulation of asymmetries in the kinetics and protein numbers of bacterial gene expression**

Sofia Startceva, Vinodh K. Kandavalli, Ari Visa and Andre S. Ribeiro

## **1. Supplementary Materials and Methods**

### **1.1. Measuring times of RNA and proteins following induction**

When measuring the integer-valued number of RNAs or the moment when a new RNA appears in a cell, for  $P_{lac/ara-1}$  and its variants, following the procedure above, we induce the reporter gene with aTc and the target gene with arabinose (when appropriate, as in [1]). Next, 50 min later, we induce the target gene with a given amount of IPTG (Table 1). Images of cells are taken 1 h after that, from which RNA numbers are quantified. In time series measurements, imaging starts 10 min after induction with IPTG (for details, see Materials and Methods, section 2.5). For other promoters ( $P_{tetA}$  and  $P_{BAD}$ ), the reporter gene, under the control of a  $P_{lac}$ , is induced with IPTG. Next, 50 min later, we induce the target gene (using the inducer specified in Table 1).

When measuring protein expression levels, we followed the same protocols as for measuring RNA numbers (aside from inducing MS2-GFP production), but we waited 90 min after induction of the target before performing the flow cytometry measurements. The additional 30 min compared to the RNA measurements are to account for the time for protein translation and maturation, in agreement with [2]. We also tested other waiting times (15, 45 and 60 min), but 30 min was the time interval that generated more consistent results between RNA and protein numbers in all conditions.

### **1.2. Image analysis of microscopy data**

We used the software ‘CellAging’ [3]. It performs automated segmentation of phase-contrast images, followed by a manual correction. Next, confocal images are semi-automatically aligned with the phase-contrast images using thin-plate spline interpolation for the registration transform (for that, we manually select 5-8 landmarks that adjust the cell masks to the borders of the corresponding cells from the confocal images). After alignment, cell lineages are constructed (when applicable), by establishing the relationships between cell masks in sequential frames. Next, from each segmented cell, at each time point, fluorescent spots are detected automatically by the Gaussian surface-fitting algorithm [4]. From these data, time-series of fluorescent spots intensity were obtained for each cell, and the time points when novel RNA molecules appear in each cells were estimated [4]. This allows obtaining the time between consecutive RNA production events in individual cells (see Materials and Methods, section 2.6).

### 1.3. Analysing the mean, coefficient of variation, skewness and kurtosis of the $\Delta t$ distribution

From the  $\Delta t$  distributions, we calculated M, CV, S and K in accordance with the definitions below, where  $\langle \Delta t \rangle$ ,  $\sigma_{\Delta t}$  and  $n$  denote the average, SD and sample size of the  $\Delta t$  distribution, respectively. In the case of S and K, we also applied the sample size correction [5]).

Feature	M	CV	S	K
Definition	$\langle \Delta t \rangle$	$\frac{\sigma_{\Delta t}}{\langle \Delta t \rangle}$	$\frac{\langle (\Delta t - \langle \Delta t \rangle)^3 \rangle}{\sigma_{\Delta t}^3}$	$\frac{\langle (\Delta t - \langle \Delta t \rangle)^4 \rangle}{\sigma_{\Delta t}^4}$
Corrected value	-	-	$\frac{\sqrt{n(n-1)}}{n-2} S$	$\frac{(n-1)}{(n-2)(n-3)} ((n+1)K - 3(n-1)) + 3$

Next, we estimated the standard error of the mean (SEM) of these features using a non-parametric bootstrap method [6,7]. Namely, for each  $\Delta t$  distribution, we performed  $10^5$  random resamples with replacement and obtained the bootstrapped distributions of M, CV, S and K values. Since a bootstrapped distribution is expected to converge to Gaussian according to the central limit theorem, the standard deviation (SD) of each bootstrapped distribution is equivalent to the SEM of the corresponding feature. This allows using a 2-sample z-test to compare the estimated features between conditions.

The same methodology was also applied when extracting mean, coefficient of variation, skewness and kurtosis from other distributions, such as the distribution of protein expression levels in single cells.

### 1.4. Western blot measurements

Mean RNA production rates differ with the free RNAP concentration in the cells [8,9]. The RNAP concentrations in each condition, relative to the control, were assessed by measuring the level of the RpoC protein by Western blot. The results confirmed that the relative RNAP levels change linearly with media richness as first reported in [1] and then confirmed in [10–12]. To attain different concentrations of intracellular RNAP without altering significantly the growth rates of the cells, we grow the cultures in media of different richness (1x, 0.5x and 0.25x), as described above. Results are shown in Supplementary Figure S2C.

Pelleted cells were lysed with B-PER bacterial protein extraction reagent supplemented with a protease inhibitor for 10 min, at room temperature. Afterwards, the lysed cells were centrifuged at 15000xg for 10 min, and the supernatant was collected and diluted in the 4X laemmli sample loading buffer containing  $\beta$ -mercaptoethanol, after which it was boiled for 5 min at 95 °C. Each sample containing ~30  $\mu$ g of total soluble proteins, were resolved by 4% to 20% TGX stain free precast gels (Biorad). Proteins were separated by electrophoresis and then electro-transferred to the PVDF membrane. Membranes were blocked with 5% non-fat milk for 1 h at room temperature and incubated with respective primary RpoC antibodies of 1:2000 dilutions (Biolegend) overnight at 4 °C, followed by the appropriate HRP-secondary antibodies 1:5000 dilutions (Sigma Aldrich) for 1 h at room temperature. For detection, chemiluminescence reagent (Biorad) was

used. Images were generated by the Chemidoc XRS system (Biorad). Quantification of the band intensity was done using Image lab software (version 5.2.1).

### 1.5. Estimating the time spent in transcription initiation prior and after commitment to open complex formation

To estimate  $\tau_{\text{prior}}$  and  $\tau_{\text{after}}$ , we use a methodology based on measuring RNA production rates at different intracellular RNAP concentrations in live cells. The method follows a similar protocol, established using *in vitro* techniques [13], and was adapted for *in vivo*, single-cell, single-RNA detection measurement techniques [1].

From the *in vitro* measurements one can directly measure the time-length of the closed and open complex formations since one can limit which components are in the reaction vessels and which reactions can take place during transcription initiation [13]. This is not possible in live cells. Also, one can only measure (by microscopy and single-RNA detection by MS2-GFP tagging) the time intervals between consecutive RNA production events in individual cells ( $\Delta t$ ) at different intracellular RNAP concentrations [1]. As such, all normal events during transcription initiation can occur, unlike when using *in vitro* techniques. Consequently,  $\tau_{\text{prior}}$  (Figure 1) is not the mean time-length of the closed complex formation since, among other, it also is affected by transient promoter locking events. Similarly,  $\tau_{\text{after}}$  is not the mean time-length of the open complex formation since it is affected by other events, such as promoter escape. Rather,  $\tau_{\text{prior}}$  is the mean time-length of all events preceding commitment to open complex formation, while  $\tau_{\text{after}}$  is the mean time-length of all events subsequent to this commitment.

According to the model (Figure 1E),  $\tau_{\text{prior}}$  depends on the intracellular concentration of RNAP while  $\tau_{\text{after}}$  does not. Thus, provided knowledge on  $M$  (mean of the  $\Delta t$  distribution, which equals the inverse of the mean RNA production rate),  $\tau_{\text{prior}}$  and  $\tau_{\text{after}}$  can be estimated from measurements of the rates of RNA production at different RNAP concentrations [1,10,13–15] (Materials and Methods, section 2.3). For that, one can use a Lineweaver–Burk plot [16] of the inverse of the RNA production rate versus the inverse of the RNAP concentration ( $[RNAP]$ ) (also named ‘ $\tau$  plot’). From this, one can estimate  $\tau_{\text{after}}$  (which equals the inverse of the rate of RNA production for infinite  $[RNAP]$ ). Next,  $\tau_{\text{prior}}$  at a given  $[RNAP]$  can be obtained by subtracting  $\tau_{\text{after}}$  from  $M$  at that  $[RNAP]$ .

Here, we measure  $[RNAP]$  by Western blot [10,11] and RNA production rates by qPCR [10], relative to the control condition (1x LB media) (Figure 1F-G). From these, we estimate  $\tau_{\text{prior}}/M$  (Figure 1H), where the line is obtained by a maximum likelihood fit [17]. We also calculate the standard error of the estimate using the Delta Method [18]. Next, given  $M$  for each condition, we calculate the absolute values of  $\tau_{\text{prior}}$  and  $\tau_{\text{after}}$  for that condition (Figure 1I).

### 1.6. Stochastic model of transcription

*In vitro* studies have shown that, in normal conditions, the kinetics of active transcription initiation in *E. coli* can be well described as a stochastic, two rate-limiting steps process [1,14,15,19–21]. The kinetics of these steps can be regulated separately from one another [1,10,13,15,19,21–24]. The first rate-limiting step is the set of events that take place from the freeing of a promoter from a preceding RNAP until the successful binding of

the ‘next’ RNAP to the promoter and commitment to open complex formation (including, among other, the sporadic repression states and the finding of the transcription start site by the RNAP). The average time-length of these events is here denoted as  $\tau_{\text{prior}}$ . We note that this time includes the fractions of time that the promoter may be under the influence of a repressor molecule.

The second step is the set of events (e.g. isomerization) that occur from the commitment to open complex formation up to its completion and promoter escape [22,25–30]. The average time-length of these events is here denoted as  $\tau_{\text{after}}$ . The sum of these two average time-lengths ( $\tau_{\text{prior}}$  and  $\tau_{\text{after}}$ ) is denoted as  $M$ , which corresponds to the mean time-length between two consecutive transcription events.

Given this, the empirical data is analysed assuming that transcription is well modelled by a two rate-limiting steps stochastic process [1] (depicted in Figure 1E). In detail, in this model, an active promoter ( $P_{\text{ON}}$ ) can participate in either of two competing processes. The first is a transition with the rate  $k_{\text{OFF}}$  of  $P_{\text{ON}}$  to an intermittent inactive state ( $P_{\text{OFF}}$ ), e.g. due to repression. This step is reversible (e.g. due to the unbinding of the repressor) with the rate  $k_{\text{ON}}$ .

The other competing step is  $P_{\text{ON}}$  being bound by an RNAP ( $R$ ) at the rate  $k_1$  and forming a closed complex ( $RP_c$ ). This step is also reversible [1,13,15] at the rate  $k_{-1}$  and competes with the formation of an open complex ( $RP_o$ ) whose rate constant is  $k_2$ . Once committed to the open complex formation, it is assumed that, in normal conditions, transcription is no longer reversible [13]. The subsequent steps are accounted for by a single-step reaction with the rate  $k_3$  [14,31,32] (also see [33] and references within). These steps include, among other, promoter escape (freeing the promoter for new events), transcription elongation, and termination, at which point the RNA and RNAP are also released.

This stochastic model does not consider positive supercoiling buildups, as we do not model genes exhibiting particularly high expression levels [34], in accordance with the empirical data (Figure 2A).

### 1.7. Stochastic model of coupled transcription and translation

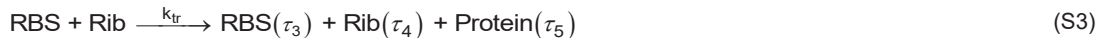
To model the dynamic coupling between transcription and translation, one needs a more complex stochastic model of transcription than the one considered in Figure 1E. For this, we model explicitly the ribosome binding site (RBS) region of the RNA, while still also modelling the complete RNA molecule. This is because the production of the RBS occurs soon after promoter escape (following the completion of transcription initiation) and, once this occurs, translation can begin (but not be completed before the transcript is complete). For a detailed description of this modelling strategy see e.g. [35,36] and references within. The multi-delayed stochastic model on RNA production considered here is:



In reactions S1 and S2, aside from the rate constants defined in the previous section,  $k_e$  is the rate of promoter escape. Meanwhile  $\tau_1$  is the time for the RNAP to move 30 to 60 base pairs (bp) downstream of the transcription start site. This allows a new RNAP to bind [37]. At approximately the same time, an RBS is produced (since this region of the RNA is up to  $\sim 40$  nucleotides long [38]). As such, and given the much longer time-length of the intervals between consecutive RNA production events, we assume that this process time-length also equals  $\tau_1$ . Finally,  $\tau_2$  is the time-length of completion of transcription elongation along with RNAP and RNA release.

As a side note, this model can also account for elongation along with alternative pathways, such as pausing, arrests, editing, pyrophosphorolysis and RNA polymerase traffic. Namely, the effects of such events can be accounted for by the distribution from which the values of  $\tau_1$  and  $\tau_2$  are randomly extracted [39].

Next, translation is modelled by reaction S3, using the RBS above as a reactant (thus allowing it to initiate prior to the complete production of the corresponding RNA). The other reactant is a ribosome (Rib) [35]:



In (S3),  $k_{\text{tr}}$  is the binding rate of a ribosome to the RBS of the target RNA. Meanwhile,  $\tau_3$  is the time for the RBS to be available for a new ribosome to bind and  $\tau_4$  is the time for a polypeptide to be produced and the ribosome to be released. Finally,  $\tau_5$  includes the time for the previous events plus the time for protein folding and maturation. As above, one can consider in the distribution from which  $\tau_4$  and  $\tau_5$  are extracted, events such as variable codon translation rates, ribosome traffic, back-translocation and trans-translation.

Known events not accounted in this model are premature termination during transcription and drop-off in translation, whose occurrence is rare in normal growth conditions [39].

Based on this model, while affected by noise, we expect a positive correlation between the mean number of proteins and the RNA numbers. This correlation should be maximal if the moments when RNA and proteins numbers are counted are distanced by the mean time taken to produce a functional protein from the RNA.

## 2. Supplementary Results

### 2.1. RNA production kinetics during the lifetime of the cells

It is reasonable to hypothesize that the kinetics of RNA production of the target gene may differ following gene replication. Meanwhile, we interpret our measurements of RNA production intervals assuming that in each cell there is only one gene active coding for this target RNA. For this to be valid, on average, there should not exist a significant difference in the kinetics of RNA production (e.g. mean rate) between the first and second half of the cells lifetime.

To test this, we compared distributions of  $\Delta t$  intervals extracted from cells during the first half of their lifetime and during the second half of their lifetime (during which the DNA replicates). In particular, we compared the

distribution of intervals that started and ended in the first half of the lifetime with the distribution of intervals that started and ended in the second half of the lifetime. For this, we performed 2-sample Kolmogorov-Smirnov tests for each condition (see Table 1 for the list of conditions), and applied a Bonferroni-Holm correction for multiple comparisons to the  $p$ -values obtained. We found that, at the significance level of 0.05, the two distributions cannot be distinguished ( $p$ -values  $> 0.31$ ) except for the LA(75) condition ( $p$ -value = 0.04). As it is unlikely that DNA replication would affect this condition differently when compared to the other conditions, we conclude that there are no significant differences in the kinetics of RNA production of the target gene during the cell lifetime. This suggests that, in our measurements, DNA replication does not disturb significantly the RNA production kinetics of our target genes.

## 2.2. Cell growth rates and morphology

We tested whether the expression of MS2-GFP proteins, at the induction levels employed in this study, affects cell growth rates and/or cell morphology. For this, first, we measured mean cell division times. Their mean and standard error were found to equal  $44.3 \pm 1.4$  min, when expressing, and  $43.2 \pm 1.3$  min, when not expressing MS2-GFP, from which we conclude that they do not differ significantly. Next, using phase-contrast microscopy and image analysis [3], we compared the morphology of the cells with and without the expression of the MS2-GFP proteins, and found no significant differences.

## 2.3 Distribution of protein expression levels in individual cells is not affected by MS2-GFP tagging

To test if the MS2-GFP tagging system could affect the protein expression levels of the target gene, we measured the distribution of single-cell protein expression levels (by flow cytometry) under the control of  $P_{lac/ara-1}$  (LA condition, Table 1 in main manuscript) when and when not activating the expression of MS2-GFP. From the distributions, we extract  $M$ ,  $CV$ ,  $S$ , and  $K$ , as these are the features of interest.

To quantify the degree to which two distributions differ (i.e. the distance  $D$  between them), we obtained the distance between the values of  $M$ ,  $CV$ ,  $S$  and  $K$  of these distributions, and normalized them by dividing by the mean value of that feature in the conditions considered. Assuming that  $\Delta$  is the difference between two features, this distance between two distributions equals:

$$D = \frac{|\Delta M_p|}{\langle M_p \rangle} + \frac{|\Delta CV_p|}{\langle CV_p \rangle} + \frac{|\Delta S_p|}{\langle S_p \rangle} + \frac{|\Delta K_p|}{\langle K_p \rangle} \quad (S4)$$

In order to determine whether this distance is significant, we also considered the distances between pairs of distributions obtained in different conditions. Shortly, if the distance  $D$  between the LA conditions expressing and not expressing MS2-GFP is smaller than the distances between different conditions, we can conclude that the expression of MS2-GFP followed by tagging of the target RNA does not perturb significantly the relationship between RNA and protein numbers of the target gene.

For this, we make use of the single-cell distributions of protein expression levels of the control condition (LA) along with the subset 'Mutants' (Mut1, Mut2, Mut3, Mut4) and the LA(IPTG) condition, since those are the conditions used in Figure 4. Since we make use of more than two conditions, the normalization in equation



(S4) is performed by dividing the difference in each feature between a pair of conditions by the mean of all conditions considered.

In Figure S9, it is visible that, in general, the pair of conditions LA, differing in whether MS2-GFP is expressed, exhibits one of the smallest differences in each of the features considered. More importantly, when considering the four features together (using distance  $D$  as defined above), they are the pair of conditions whose distributions of single-cell protein expression levels are most similar. We thus conclude that the expression of MS2-GFP does not affect significantly the observed protein expression levels.

This result can be explained by the location of the coding regions of the RNA target for MS2-GFP in the plasmid, relative to the transcription start site. Namely, it starts with a ribosome binding site (RBS), followed by the region coding for the red fluorescent protein. Only afterwards is the region coding for the MS2-GFP binding sites, thus minimizing interference with the RBS activity and with the degradation rate of the region coding for the red fluorescent protein.

#### **2.4. Skewness and kurtosis of RNA production kinetics are not correlated to the distributions of cell lifetimes or to the distributions of intracellular RNAP concentrations**

It is reasonable to assume that differences in the shapes of the distributions of cell lifetimes between the conditions considered above could also affect the  $\Delta t$  distributions. To test this, we measured cell lifetimes in the conditions where cells are in the exponential growth phase (Table 1). Next, we calculated the mean, coefficient of variation, skewness and kurtosis of each of these distributions of cell lifetimes (here named  $M_L$ ,  $CV_L$ ,  $S_L$ , and  $K_L$ , respectively) and plotted them against the corresponding  $M$ ,  $CV$ ,  $S$ , and  $K$  of the  $\Delta t$  distribution (Figure S4A). In each case, we calculated the Pearson's correlation coefficient (with the corresponding two-tailed  $p$ -value), and found no significant correlation (all  $p$ -values  $> 0.05$ ).

We conclude that, in our data,  $S$  and  $K$  of the  $\Delta t$  distribution are not correlated with any feature of the distribution of cell lifetimes. This is in agreement with our observation that the cell morphology (Materials and Methods) and physiology do not differ significantly between the conditions considered.

Further, as in the main manuscript, section 3.2, we applied the same calculations when considering the subsets 'Mutations' and 'Regulatory factors' separately (Figure S4B,C). Again, when applying the Bonferroni-Holm correction for multiple comparisons, the only potential correlation ( $K$  vs.  $K_L$  in the subset 'Regulatory factors') is not statistically significant. These results show that even when reducing the number of variables differing between conditions, there is no visible significant correlation between the features of the distributions of cell lifetimes and the features of the  $\Delta t$  distribution.

Finally, we obtained the single-cell distributions of RNAP concentrations using a cell strain where RNAPs are fluorescently tagged with GFP (Materials and Methods) in the media richness conditions 1x, 0.75x and 0.5x (Supplementary Figure S5). We found no relationship between the skewness of these distributions and  $S$  of the corresponding  $\Delta t$  distributions.

The results on the various conditions differing in target promoter or regulatory factors are expected since the cells are from the same strain and in the same media conditions. Similarly, the results on the conditions

differing in medium are expected given that these media (1x, 0.75x and 0.5x) were specially tuned for having cells with differing RNAP levels but similar average growth rates [1]. In this regard, it is worth mentioning that when observing the lifetimes of a small number of cells (values of  $M_L$  in Figure S4) there are visible differences between the conditions. However, the growth curves (Supplementary Figure S2A) indicate that this variability is due to the small number of cells that are observed during their entire lifetime by microscopy (when compared to the growth curves).

## Supplementary References

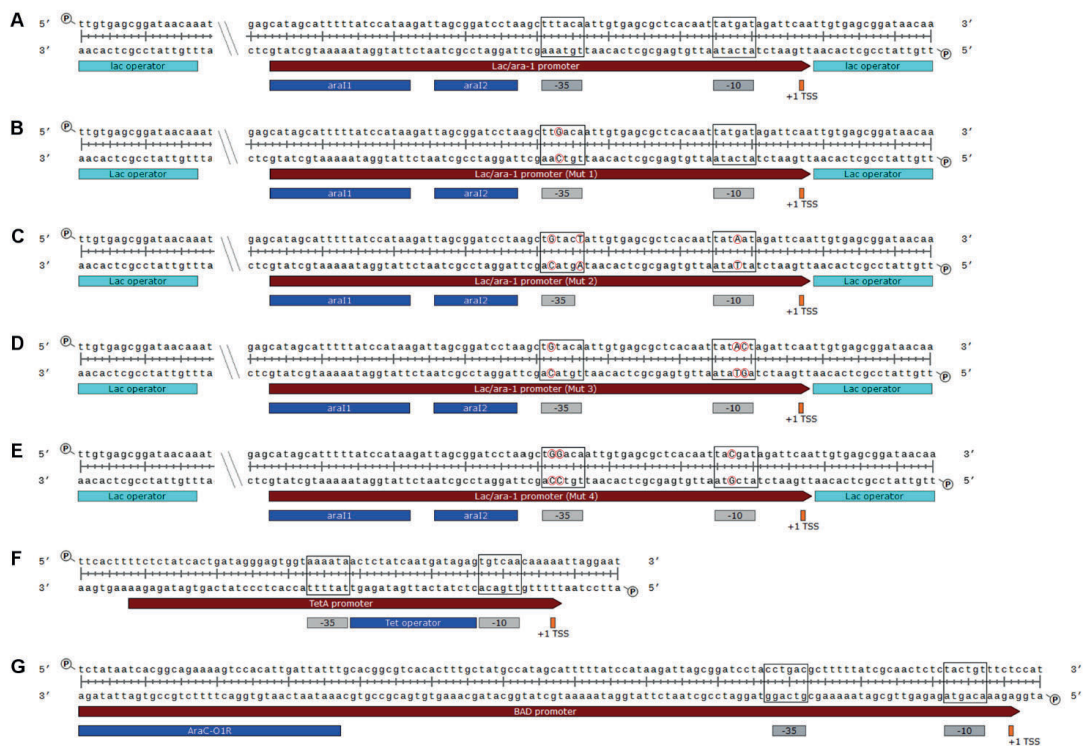
- [1] Lloyd-Price J., Startceva S., Kandavalli V., Chandraseelan J.G., Goncalves N., Oliveira S.M.D., Häkkinen A., and Ribeiro A.S. (2016). Dissecting the stochastic transcription initiation process in live *Escherichia coli*. *DNA Res.*, 23: 203–214. DOI: <https://doi.org/10.1093/dnares/dsw009>
- [2] Hebisch E., Knebel J., Landsberg J., Frey E., and Leisner M. (2013). High variation of fluorescence protein maturation times in closely related *Escherichia coli* strains. *PLoS One*, 8: e75991. DOI: <http://dx.doi.org/10.1371/journal.pone.0075991>
- [3] Häkkinen A., Muthukrishnan A.-B., Mora A., Fonseca J.M., and Ribeiro A.S. (2013). CellAging: A tool to study segregation and partitioning in division in cell lineages of *Escherichia coli*. *Bioinformatics*, 29: 1708–1709. DOI: <https://doi.org/10.1093/bioinformatics/btt194>
- [4] Häkkinen A., and Ribeiro A.S. (2015). Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data. *Bioinformatics*, 31: 69–75. DOI: <https://doi.org/10.1093/bioinformatics/btu592>
- [5] Joanes D.N., and Gill C.A. (1998). Comparing Measures of Sample Skewness and Kurtosis. *J. R. Stat. Soc., Ser. D (Stat.)*, 47: 183–189. DOI: <http://dx.doi.org/10.1111/1467-9884.00122>
- [6] Carpenter J., and Bithell J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.*, 19: 1141–1164. DOI: [http://dx.doi.org/10.1002/\(SICI\)1097-0258\(20000515\)19:9%3C1141::AID-SIM479%3E3.0.CO;2-F](http://dx.doi.org/10.1002/(SICI)1097-0258(20000515)19:9%3C1141::AID-SIM479%3E3.0.CO;2-F)
- [7] DiCiccio T.J., and Efron B. (1996). Bootstrap Confidence Intervals. *Stat. Sci.*, 11: 189–228. DOI: <http://dx.doi.org/10.1214/ss/1032280214>
- [8] Liang S.-T., Bipatnath M., Xu Y.-C., Chen S.-L., Dennis P., Ehrenberg M., and Bremer H. (1999). Activities of constitutive promoters in *Escherichia coli*. *J. Mol. Biol.*, 292: 19–37. DOI: <https://doi.org/10.1006/jmbi.1999.3056>
- [9] Ehrenberg M., Bremer H., and Dennis P.P. (2013). Medium-dependent control of the bacterial growth rate. *Biochimie*, 95: 643–658. DOI: <http://dx.doi.org/10.1016/j.biochi.2012.11.012>
- [10] Kandavalli V.K., Tran H., and Ribeiro A.S. (2016). Effects of  $\sigma$  factor competition are promoter initiation kinetics dependent. *Biochim Biophys. Acta*, 1859: 1281–1288. DOI: <http://dx.doi.org/10.1016/j.bbagr.2016.07.011>

- [11] Mäkelä J., Kandavalli V., and Ribeiro A.S. (2017). Rate-limiting steps in transcription dictate sensitivity to variability in cellular components. *Sci. Rep.*, 7: 10588. DOI: <https://doi.org/10.1038/s41598-017-11257-2>
- [12] Oliveira S.M.D., Häkkinen A., Lloyd-Price J., Tran H., Kandavalli V., and 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: e1005174. DOI: <http://dx.doi.org/10.1371/journal.pcbi.1005174>
- [13] McClure W.R. (1985). Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.*, 54: 171–204. DOI: <http://dx.doi.org/10.1146/annurev.bi.54.070185.001131>
- [14] Lutz R., Lozinski T., Ellinger T., and Bujard H. (2001). Dissecting the functional program of *Escherichia coli* promoters: the combined mode of action of Lac repressor and AraC activator. *Nucleic Acids Res.*, 29: 3873–3881. DOI: <http://dx.doi.org/10.1093/nar/29.18.3873>
- [15] McClure W.R. (1980). Rate-limiting steps in RNA chain initiation. *Proc. Natl. Acad. Sci. U.S.A.*, 77: 5634–5638. DOI: <http://dx.doi.org/10.1073/pnas.77.10.5634>
- [16] Lineweaver H., and Burk D. (1934). The Determination of Enzyme Dissociation Constants. *J. Am. Chem. Soc.*, 56: 658–666. DOI: <http://dx.doi.org/10.1021/ja01318a036>
- [17] Bevington P.R., and Robinson D.K. (2003). Least-Squares Fit to a Straight Line. In *Data Reduction and Error Analysis for the Physical Sciences* (New York: McGraw-Hill), pp. 98–115. ISBN: 0-07-247227-8
- [18] Casella G., and Berger R.L. (2001). The Delta Method. In *Statistical Inference*, 2<sup>nd</sup> ed (Pacific Grove, CA: Duxbury Press), pp. 240–245. ISBN: 0-534-24312-6
- [19] Buc H., and McClure W.R. (1985). Kinetics of open complex formation between *Escherichia coli* RNA polymerase and the *lac* UV5 promoter. Evidence for a sequential mechanism involving three steps. *Biochemistry*, 24: 2712–2723. DOI: <http://dx.doi.org/10.1021/bi00332a018>
- [20] Chamberlin M.J. (1974). The selectivity of transcription. *Annu. Rev. Biochem.*, 43: 721–775. DOI: <http://dx.doi.org/10.1146/annurev.bi.43.070174.003445>
- [21] Browning D.F., and Busby S.J.W. (2016). Local and global regulation of transcription initiation in bacteria. *Nat. Rev. Microbiol.*, 14: 638–650. DOI: <http://dx.doi.org/10.1038/nrmicro.2016.103>
- [22] deHaseth P.L., Zupancic M.L., and Record T.M. Jr. (1998). RNA polymerase-promoter interactions: the comings and goings of RNA polymerase. *J. Bacteriol.*, 180: 3019–3025. PMID: 9620948
- [23] Jones D.L., Brewster R.C., and Phillips R. (2014). Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346: 1533–1536. DOI: <http://dx.doi.org/10.1126/science.1255301>

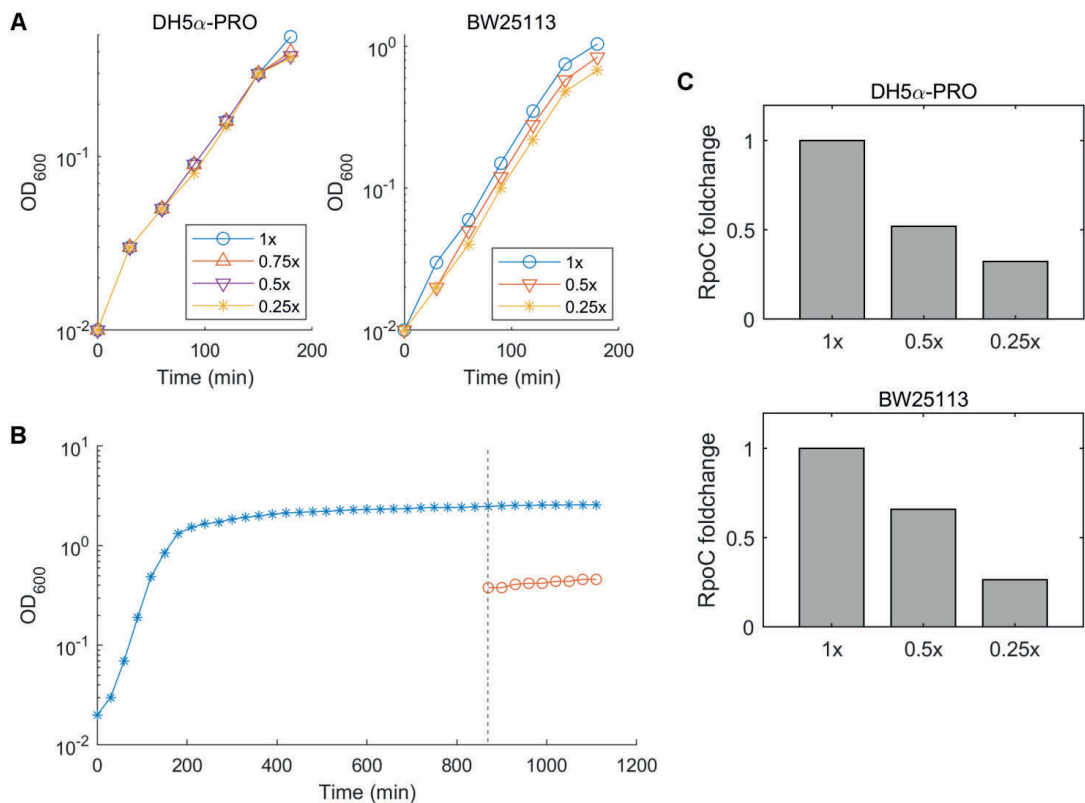
- [24] Feklistov A., Sharon B.D., Darst S.A., and Gross C.A. (2014) Bacterial Sigma Factors: A Historical, Structural, and Genomic Perspective. *Annu. Rev. Microbiol.*, 68: 357–376. DOI: <http://dx.doi.org/10.1146/annurev-micro-092412-155737>
- [25] Duchi D., Bauer D.L.V., Fernandez L., Evans G., Robb N., Hwang L.C., Gryte K., Tomescu A., Zawadzki P., Morichaud Z., *et al.* (2016). RNA polymerase pausing during initial transcription. *Mol. Cell*, 63: 939–950. DOI: <http://dx.doi.org/10.1016/j.molcel.2016.08.011>
- [26] Duchi D., Gryte K., Robb N.C., Morichaud Z., Sheppard C., Brodolin K., Wigneshweraraj S., and Kapanidis A.N. (2018). Conformational heterogeneity and bubble dynamics in single bacterial transcription initiation complexes. *Nucleic Acids Res.*, 46: 677–688. DOI: <http://dx.doi.org/10.1093/nar/gkx1146>
- [27] Hsu L.M. (2002). Promoter clearance and escape in prokaryotes. *Biochim Biophys. Acta*, 1577: 191–207. DOI: [http://dx.doi.org/10.1016/S0167-4781\(02\)00452-9](http://dx.doi.org/10.1016/S0167-4781(02)00452-9)
- [28] Hsu L.M. (2008). Promoter escape by *Escherichia coli* RNA polymerase. *EcoSal Plus*, 3: 1–16. DOI: <http://dx.doi.org/10.1128/ecosalplus.4.5.2.2>
- [29] Kapanidis A.N., Margeat E., Ho S.O., Kortkhonjia E., Weiss S., and Ebright R.H. (2006). Initial transcription by RNA polymerase proceeds through a DNA-scrunching mechanism. *Science*, 314: 1144–1147. DOI: <http://dx.doi.org/10.1126/science.1131399>
- [30] Lerner E., Chung S., Allen B.L., Wang S., Lee J., Lu S.W., Grimaud L.W., Ingargiola A., Michalet X., Alhadid Y., *et al.* (2016). Backtracked and paused transcription initiation intermediate of *Escherichia coli* RNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.*, 113: E6562–E6571. DOI: <https://doi.org/10.1073/pnas.1605038113>
- [31] Greive S.J., and von Hippel P.H. (2005). Thinking quantitatively about transcriptional regulation. *Nat. Rev. Mol. Cell Biol.*, 6: 221–232. DOI: <http://dx.doi.org/10.1038/nrm1588>
- [32] Herbert K.M., La Porta A., Wong B.J., Mooney R.A., Neuman K.C., Landick R., and Block S.M. (2006). Sequence-resolved detection of pausing by single RNA polymerase molecules. *Cell*, 125: 1083–1094. DOI: <http://dx.doi.org/10.1016/j.cell.2006.04.032>
- [33] Rajala T., Häkkinen A., Healy S., Yli-Harja O., and Ribeiro A.S. (2010). Effects of Transcriptional Pausing on Gene Expression Dynamics. *PLoS Comput. Biol.*, 6: e1000704. DOI: <https://doi.org/10.1371/journal.pcbi.1000704>
- [34] Chong S., Chen C., Ge H., and Xie X.S. (2014). Mechanism of transcriptional bursting in bacteria. *Cell*, 158: 314–326. DOI: <http://dx.doi.org/10.1016/j.cell.2014.05.038>
- [35] Ribeiro A.S. (2010). Stochastic and delayed stochastic models of gene expression and regulation. *Math. Biosci.*, 223: 1–11. DOI: <http://dx.doi.org/10.1016/j.mbs.2009.10.007>

- [36] 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: 725–745. DOI: <http://dx.doi.org/10.1016/j.jtbi.2007.01.021>
- [37] Record T.M. Jr., Reznikoff W.S., Craig M.L., McQuade K.L., Schlax P.J. (1996). *Escherichia coli* RNA polymerase ( $E\sigma^{70}$ ), promoters, and the kinetics of the steps of transcription initiation. In *Escherichia coli* and *Salmonella typhimurium*: Cellular and Molecular Biology, 2<sup>nd</sup> ed, F.C. Neidhardt, R. Curtiss, J.L. Ingraham, E.C.C. Lin, K.B. Low, B. Magasanik, W.S. Reznikoff, M. Riley, D. Schneider, and H.E. Umbarger, eds, (Washington, DC: ASM press), pp. 792–821. ISBN: 1555810845
- [38] Shultzaberger R.K., Bucheimer R.E., Rudd K.E., and Schneider T.D. (2001). Anatomy of *Escherichia coli* ribosome binding sites. *J. Mol. Biol.*, 313: 215–228. DOI: <https://doi.org/10.1006/jmbi.2001.5040>
- [39] 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: 121. DOI: <https://doi.org/10.1186/1471-2105-12-121>

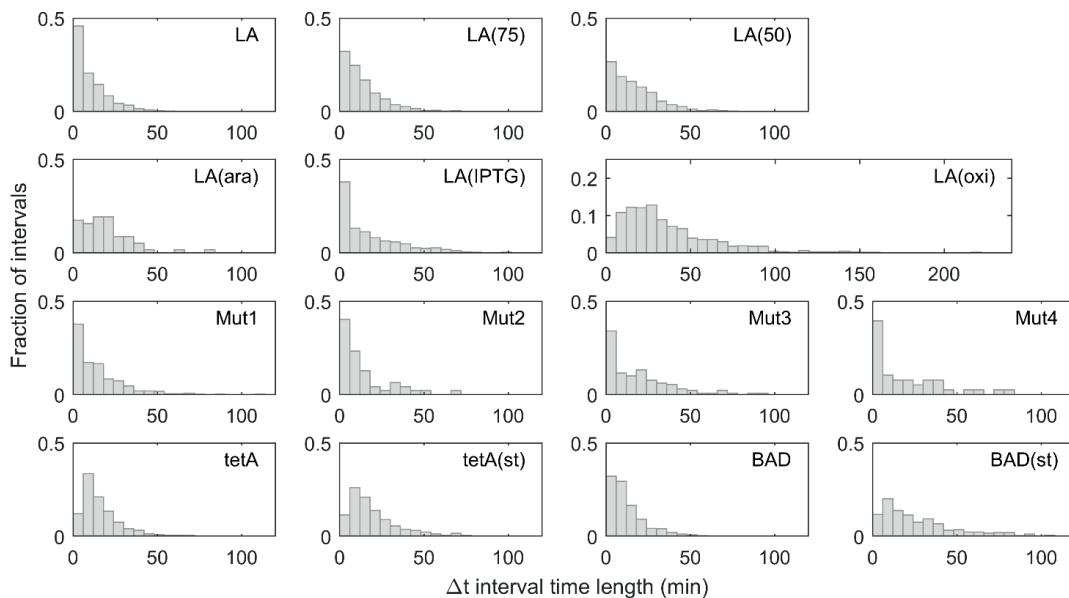
## Supplementary Figures



**Figure S1.** Related to Figure 2A and Tables 1 and 2. Schematic representation of the target promoter's sequences. The -35 and -10 promoter elements are shown in black boxes. The transcription start sites (+1 TSS) are marked in orange. Operator sites are marked in cyan and blue. In the mutants, specific nucleotide changes in the -35 and -10 region are marked by red circles. These promoters were used in the studied conditions (Table 1) as follows: **(A)** LA, LA(75), LA(50), LA(ara), LA(IPTG) and LA(oxi); **(B)** Mut1; **(C)** Mut2; **(D)** Mut3; **(E)** Mut4; **(F)** tetA and tetA(st); **(G)** BAD and BAD(st).

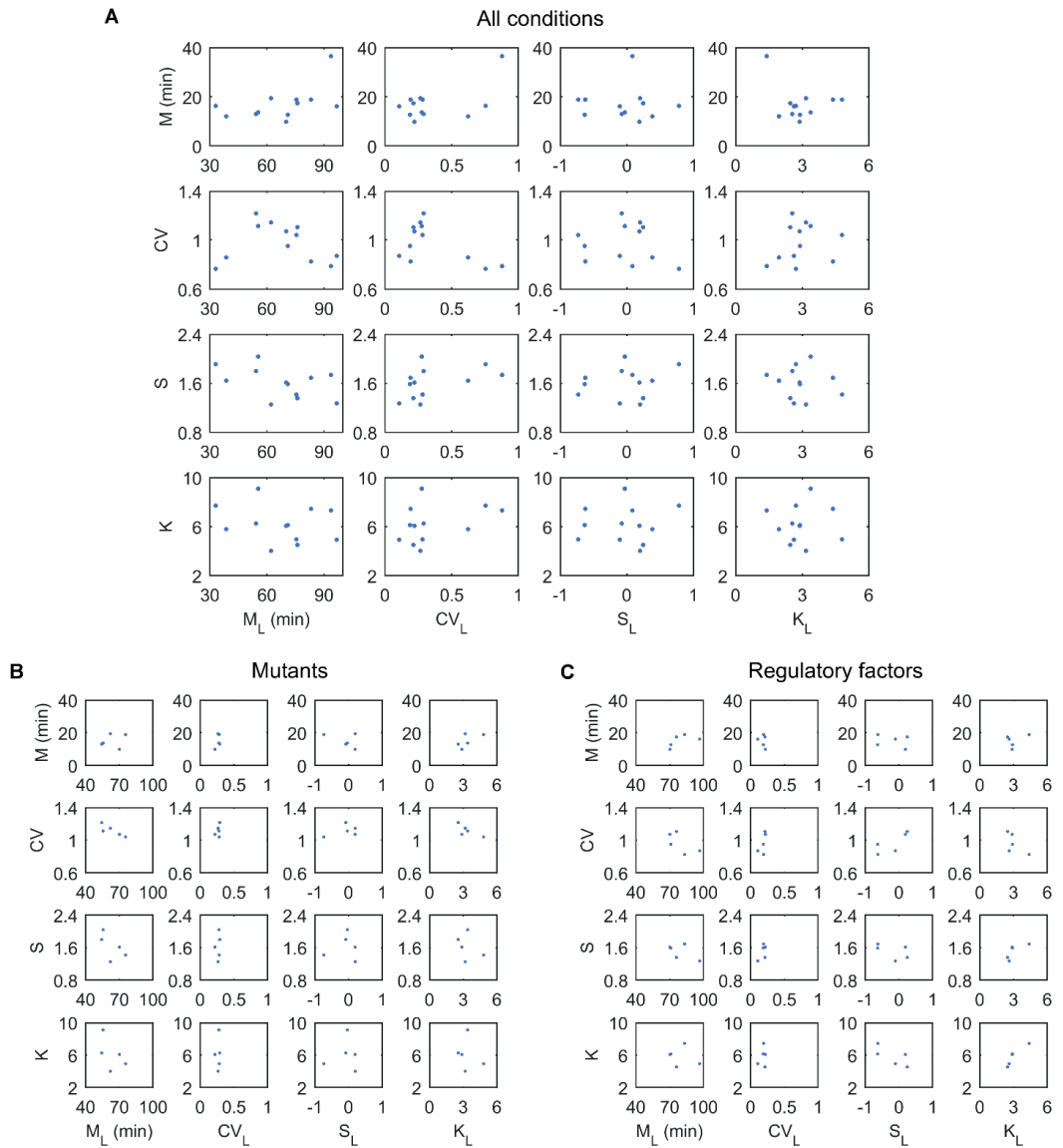


**Figure S2.** Related to Figures 2A and 3A. Bacterial growth curves and RNAP levels as a function of media richness, relative to the control condition. **(A)** Bacterial growth curves of the DH5 $\alpha$ -PRO and BW25113 *E. coli* strains when grown in LB media with different richness (1x, 0.75x, 0.5x and 0.25x, see Materials and Methods for a detailed description). The optical density at the wavelength of 600 nm (OD<sub>600</sub>) was measured every 30 min for 3 h. **(B)** Bacterial growth curve of the BW25113 strain reaching the stationary phase. Cells were grown in 1x LB media (see Materials and Methods for a detailed description) at 37 °C with shaking at 250 rpm, and the OD<sub>600</sub> values were monitored every 30 min (blue stars). After the cells reached the stationary phase, we diluted them in stationary phase media (Materials and Methods) and monitored the OD<sub>600</sub> every 30 min (red circles) for 4 h. The vertical dashed line shows the time of the dilution. **(C)** RNAP levels (relative to the 1x condition) of the DH5 $\alpha$ -PRO and BW25113 *E. coli* strains grown in LB media with different richness (1x, 0.5x, and 0.25x) as assessed by Western blot measurements of the RpoC protein (Supplementary Materials and Methods, section 1.4).

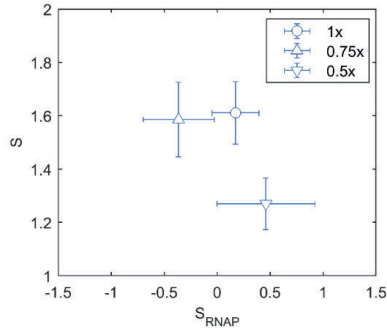


**Figure S3.** Related to Figure 2A. The shape of the distribution of time intervals between consecutive RNA production events in individual cells ( $\Delta t$  distribution) changes significantly between mutants as well as with the promoter, induction scheme and media. See Table 1 for a detailed description of each condition and Supplementary Table S8 for the statistical tests to assess whether the distributions normalized by the mean differ significantly. Data were collected from approximately 600 cells per condition.

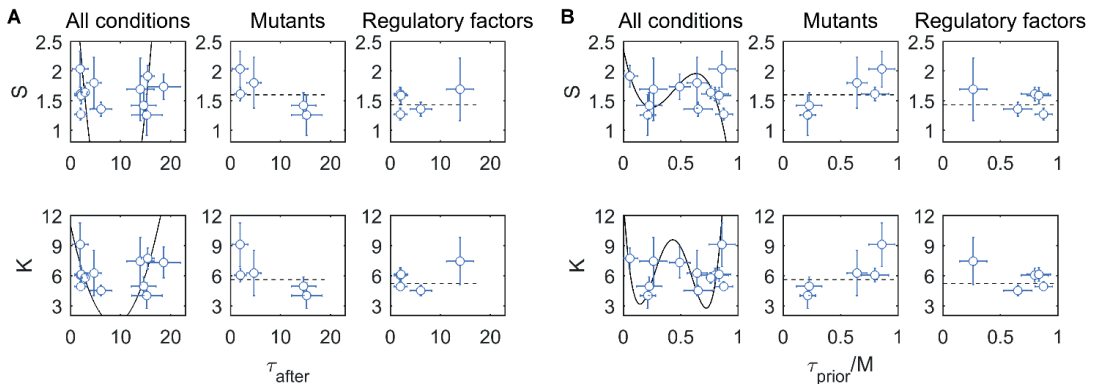




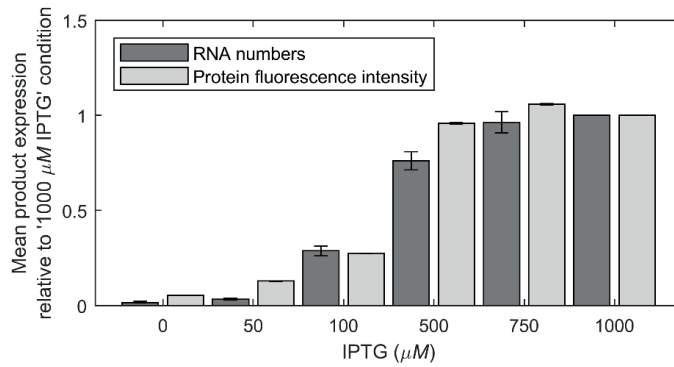
**Figure S4.** Related to Figure 1. Mean ( $M$ ), coefficient of variation ( $CV$ ), skewness ( $S$ ) and kurtosis ( $K$ ) of the distribution of intervals between consecutive RNA production events in individual cells plotted against the mean ( $M_L$ ), coefficient of variation ( $CV_L$ ), skewness ( $S_L$ ) and kurtosis ( $K_L$ ) of the corresponding distributions of single-cell lifetimes. Shown are **(A)** all conditions, **(B)** the ‘Mutants’ subset, **(C)** the ‘Regulatory factors’ subset.



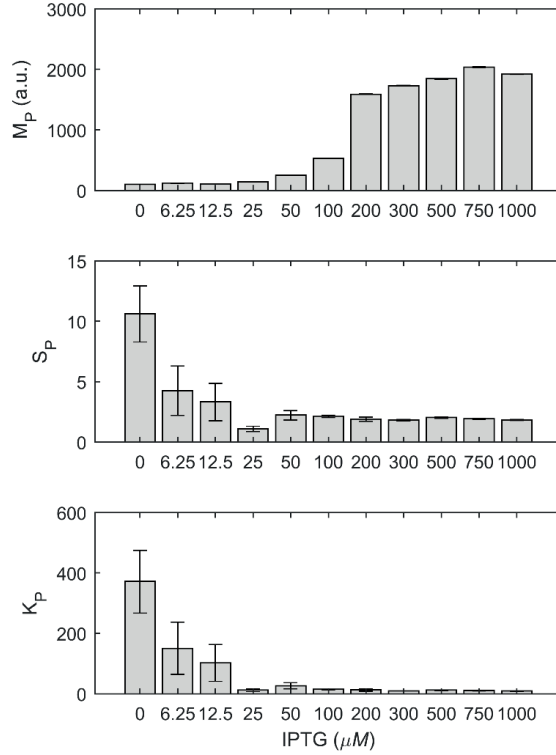
**Figure S5.** Related to Figure 1. Skewness of the  $\Delta t$  distribution ( $S$ ) measured from a fully induced  $P_{lac/ara-1}$  promoter in various media conditions (see section 1.3 in main manuscript) plotted against the skewness of the single-cell RNAP fluorescence distribution. The RNAP fluorescence distributions are measured by microscopy ( $\sim 400$  cells per condition). Error bars denote SEM.



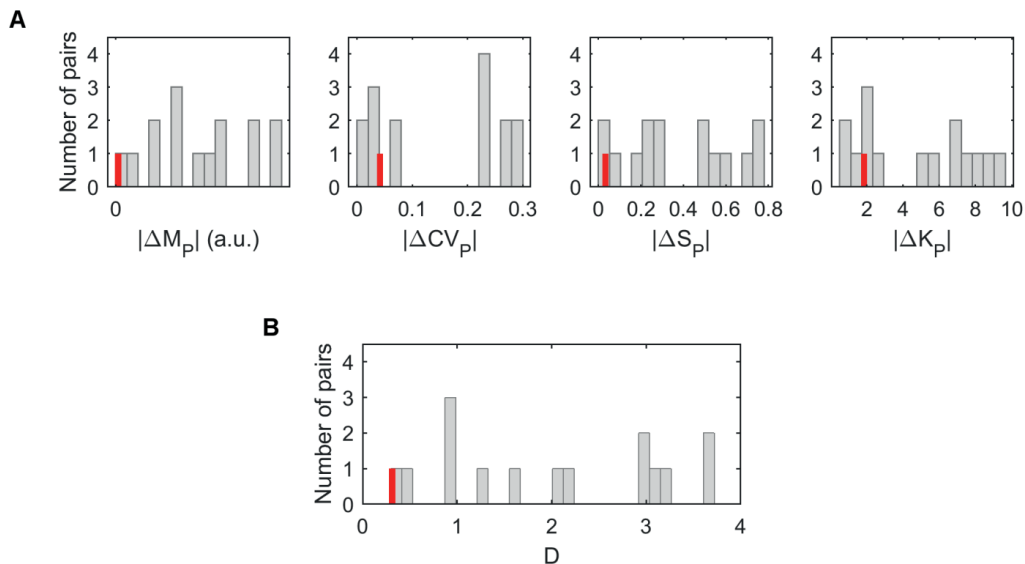
**Figure S6.** Related to Figure 3B and Supplementary Tables S5 and S6. Skewness ( $S$ ) and kurtosis ( $K$ ) of the distribution of intervals between consecutive RNA production events in individual cells do not show linear relationships with the fraction of time spent in events after commitment to the open complex formation ( $\tau_{\text{after}}$ ) nor with the mean fraction of time spent in the events prior to commitment to open complex formation ( $\tau_{\text{prior}}/M$ , where  $M$  is the mean time between transcription events). Shown is **(A)**  $S$  and  $K$  as a function of  $\tau_{\text{after}}$  and **(B)**  $S$  and  $K$  as a function of  $\tau_{\text{prior}}/M$ , for all conditions and for the subsets of conditions 'Mutants' and 'Regulatory factors'. Error bars denote SEM. The best-fitting model is shown as a dashed line if it is a zero-degree polynomial and as a solid line if it is a polynomial of a higher degree. In plots where two separate lines are visible, the best fitting model is partially outside of the plot borders on the y-axis.



**Figure S7.** Related to Figure 4. Mean protein numbers of the target gene under the control of  $P_{lac/ara-1}$  follow the corresponding average RNA numbers for increasing induction levels. Induction curve of  $P_{lac/ara-1}$  as seen by observing the mean RNA numbers produced by the target promoter ( $P_{lac/ara-1}$ ) in individual cells using microscopy (dark grey), and by observing the mean fluorescent intensity of proteins in individual cells from the same promoter using flow cytometry (light grey). In all conditions, cells are subject to 1% of arabinose. Data obtained by microscopy was collected 60 min after induction of the target gene, while data obtained by flow cytometry were collected 90 min after induction of the target gene. In both measurements, the values are shown relative to the value obtained in the condition '1000 μM IPTG' of the corresponding measurement. Error bars denote the standard error of the ratio.



**Figure S8.** Related to Figure 4. The skewness ( $S_P$ ) and kurtosis ( $K_P$ ) of the single-cell distributions of protein expression levels differ with induction strength but can also have significantly different values for similar mean ( $M_P$ ) expression levels. (Top)  $M_P$ , (middle)  $S_P$  and (bottom)  $K_P$  of the single-cell distributions of protein expression levels, as expressed under the control of  $P_{lac/ara-1}$  for changing induction strength. Data were collected by flow cytometry, 90 min after induction of the target gene. Error bars denote SEM. In the top figure, the error bars are too small to be visible. In the regime of weak induction (25 or less  $\mu M$  IPTG),  $S_P$  and  $K_P$  show significant, consistent changes, although  $M_P$  does not exhibit significant changes. Meanwhile, above 25  $\mu M$  IPTG concentration, the opposite occurs.



**Figure S9.** Related to Figure 4. Activation of the MS2-GFP reporter does not affect significantly the single-cell distribution of protein expression levels. (A) Numbers of pairs of conditions (grey bars) with given values of, respectively, the absolute differences in mean ( $|\Delta M_p|$ ), coefficient of variation ( $|\Delta CV_p|$ ), skewness ( $|\Delta S_p|$ ) and kurtosis ( $|\Delta K_p|$ ) of the single-cell distributions of protein expression levels. The conditions considered are LA, LA(IPTG), Mut1, Mut2, Mut3 and Mut4 (see Table 1 in main manuscript). Meanwhile, the red bar marks the values for these differences between the pair of measurements in the LA condition with and without activating the reporter. (B) Distance D between the values of M, CV, S and K (equation S4) for the same pairs of conditions as in (A). The red bar holds the value 0.31, while the grey bar further to the left holds the value 0.33.

## Supplementary Tables

**Table S1.** Related to Figure 2B and Table 2. Two-tailed  $p$ -values obtained by testing, for each pair of conditions, the null hypothesis ( $H_0$ ) that the difference in the mean (M), coefficient of variation (CV), skewness (S) and kurtosis (K) of the distribution of time intervals between consecutive RNA production events between the two conditions equals zero, using a 2-sample z-test. In cases where the  $p$ -value  $\leq 0.05$ , the  $H_0$  is rejected (highlighted with italics). In cases where the  $p$ -value  $> 0.05$ , the  $H_0$  cannot be rejected.

<b>M</b>	LA(75)	LA(50)	LA(ara)	LA(IPTG)	LA(oxi)	Mut1	Mut2	Mut3	Mut4	tetA	tetA(st)	BAD	BAD(st)
LA	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	0.16	<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>
LA(75)		<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	0.28	0.90	<i>&lt; 0.001</i>	0.06	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	0.28	<i>&lt; 0.001</i>
LA(50)			0.19	0.17	<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	0.18	0.12	0.36	0.75	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>
LA(ara)				0.49	<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	0.06	1.00	0.90	0.24	0.86	<i>&lt; 0.01</i>	<i>&lt; 0.01</i>
LA(IPTG)					<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	0.07	0.42	0.58	0.31	0.07	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>
LA(oxi)						<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>
Mut1							0.77	<i>&lt; 0.01</i>	0.12	<i>&lt; 0.01</i>	<i>&lt; 0.001</i>	0.05	<i>&lt; 0.001</i>
Mut2								<i>&lt; 0.01</i>	0.13	0.16	<i>&lt; 0.01</i>	0.68	<i>&lt; 0.001</i>
Mut3									0.90	0.17	0.84	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>
Mut4										0.40	0.98	<i>&lt; 0.01</i>	0.06
tetA												<i>&lt; 0.01</i>	<i>&lt; 0.001</i>
tetA(st)												<i>&lt; 0.001</i>	<i>&lt; 0.001</i>
BAD													<i>&lt; 0.001</i>
<b>CV</b>	LA(75)	LA(50)	LA(ara)	LA(IPTG)	LA(oxi)	Mut1	Mut2	Mut3	Mut4	tetA	tetA(st)	BAD	BAD(st)
LA	<i>&lt; 0.01</i>	<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	0.43	<i>&lt; 0.001</i>	0.49	0.28	0.68	0.62	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>
LA(75)		<i>&lt; 0.01</i>	0.23	<i>&lt; 0.01</i>	<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	0.05	0.25	0.20	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	<i>&lt; 0.01</i>
LA(50)			0.66	<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	<i>&lt; 0.01</i>	0.06	<i>&lt; 0.01</i>	<i>&lt; 0.001</i>	0.65	0.25
LA(ara)				<i>&lt; 0.01</i>	0.71	<i>&lt; 0.01</i>	<i>&lt; 0.01</i>	0.08	0.07	0.56	0.46	0.76	0.97
LA(IPTG)					<i>&lt; 0.001</i>	0.91	0.42	0.41	0.80	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>
LA(oxi)						<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	0.60	0.31	0.06	0.31
Mut1							0.47	0.42	0.84	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>
Mut2								0.24	0.71	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	<i>&lt; 0.01</i>
Mut3									0.52	<i>&lt; 0.001</i>	<i>&lt; 0.001</i>	<i>&lt; 0.01</i>	<i>&lt; 0.01</i>
Mut4										<i>&lt; 0.01</i>	<i>&lt; 0.01</i>	0.05	<i>&lt; 0.01</i>
tetA											0.70	<i>&lt; 0.01</i>	0.14
tetA(st)												<i>&lt; 0.01</i>	<i>&lt; 0.01</i>
BAD													0.47

(see the rest of the table on the next page)

<b>S</b>	LA(75)	LA(50)	LA(ara)	LA(IPTG)	LA(oxi)	Mut1	Mut2	Mut3	Mut4	tetA	tetA(st)	BAD	BAD(st)
LA	0.89	< 0.01	0.88	0.12	0.62	0.19	0.67	0.43	0.33	0.17	0.40	0.85	< 0.01
LA(75)		0.06	0.85	0.20	0.57	0.18	0.64	0.51	0.37	0.16	0.54	0.75	0.06
LA(50)			0.43	0.57	0.05	< 0.01	0.23	0.53	0.97	< 0.01	0.17	< 0.01	0.89
LA(ara)				0.53	0.94	0.57	0.87	0.63	0.49	0.69	0.69	0.93	0.41
LA(IPTG)					0.13	< 0.01	0.32	0.79	0.78	< 0.01	0.46	0.07	0.52
LA(oxi)						0.42	0.89	0.30	0.24	0.53	0.29	0.70	0.05
Mut1							0.66	0.10	0.09	0.73	0.08	0.22	< 0.01
Mut2								0.43	0.32	0.81	0.47	0.72	0.22
Mut3									0.69	0.08	0.81	0.35	0.49
Mut4										0.09	0.54	0.29	0.99
tetA											< 0.01	0.20	< 0.01
tetA(st)												0.29	0.16
BAD													< 0.01
<b>K</b>	LA(75)	LA(50)	LA(ara)	LA(IPTG)	LA(oxi)	Mut1	Mut2	Mut3	Mut4	tetA	tetA(st)	BAD	BAD(st)
LA	0.94	0.18	0.57	0.08	0.46	0.19	0.93	0.34	0.16	0.19	0.26	0.76	< 0.01
LA(75)		0.15	0.59	0.07	0.49	0.20	0.96	0.31	0.15	0.21	0.22	0.70	< 0.01
LA(50)			0.30	0.55	0.14	0.06	0.57	0.99	0.50	< 0.01	0.82	0.23	0.19
LA(ara)				0.23	0.96	0.61	0.72	0.32	0.20	0.92	0.33	0.49	0.16
LA(IPTG)					0.09	< 0.01	0.46	0.68	0.71	< 0.01	0.42	0.10	0.53
LA(oxi)						0.51	0.70	0.19	0.10	0.83	0.17	0.35	< 0.01
Mut1							0.37	0.08	< 0.01	0.57	0.07	0.14	< 0.01
Mut2								0.59	0.39	0.56	0.61	0.84	0.35
Mut3									0.55	< 0.01	0.89	0.43	0.40
Mut4										< 0.01	0.43	0.20	0.95
tetA											< 0.01	0.10	< 0.01
tetA(st)												0.35	0.13
BAD													< 0.01

**Table S2.** Related to Figure 2C-D. Percentage of the time intervals between consecutive RNA production events in individual cells ( $\Delta t$  intervals) longer than a given threshold in each condition.

Threshold \ Condition	2M	3M	4M	5M	6M
LA	16.3	6.3	2.2	0.7	0.3
LA(75)	13.9	4.7	1.5	0.7	0
LA(50)	11.8	3.2	0.8	0.1	0
LA(ara)	10.6	3.5	1.8	0	0
LA(IPTG)	17.7	7.1	1.7	0.5	0
LA(oxi)	11.0	3.0	0.5	0.2	0
Mut1	13.8	6.2	2.5	1.2	0.5
Mut2	17.0	6.4	2.1	2.1	0
Mut3	16.3	6.2	1.6	0	0
Mut4	21.1	7.9	2.6	0	0
tetA	9.4	3.2	0.9	0	0
tetA(st)	11.4	2.9	0	0	0
BAD	12.1	4.0	0.8	0.1	0
BAD(st)	12.9	3.5	0.3	0	0



**Table S3.** Related to Figure 3A. Mean time length spent in the events prior to commitment to open complex formation ( $\tau_{\text{prior}}$ ) and in the events following the commitment to open complex formation ( $\tau_{\text{after}}$ ) for each condition, along with their SEM. Also shown, for each condition, is the mean fraction of time between transcription events that is spent in the events prior to commitment to open complex formation ( $\tau_{\text{prior}}/M$ , where M is the mean time between transcription events), along with its SEM.

Condition	$\tau_{\text{prior}} \pm \text{SEM}$ (min)	$\tau_{\text{after}} \pm \text{SEM}$ (min)	$\tau_{\text{prior}}/M \pm \text{SEM}$
LA	7.8 $\pm$ 1.2	2.0 $\pm$ 1.2	0.80 $\pm$ 0.12
LA(75)	10.6 $\pm$ 1.4	2.1 $\pm$ 1.3	0.83 $\pm$ 0.10
LA(50)	14.1 $\pm$ 1.3	2.0 $\pm$ 1.2	0.87 $\pm$ 0.08
LA(ara)	5.0 $\pm$ 2.3	13.9 $\pm$ 2.7	0.26 $\pm$ 0.12
LA(IPTG)	11.3 $\pm$ 2.2	6.1 $\pm$ 2.2	0.65 $\pm$ 0.12
LA(oxi)	17.9 $\pm$ 3.4	18.6 $\pm$ 3.4	0.49 $\pm$ 0.09
Mut1	11.8 $\pm$ 1.7	1.9 $\pm$ 1.6	0.86 $\pm$ 0.11
Mut2	8.3 $\pm$ 1.9	4.7 $\pm$ 1.4	0.64 $\pm$ 0.09
Mut3	4.3 $\pm$ 2.1	14.6 $\pm$ 2.5	0.23 $\pm$ 0.11
Mut4	4.2 $\pm$ 1.5	15.2 $\pm$ 3.1	0.21 $\pm$ 0.07
tetA	0.9 $\pm$ 1.2	15.4 $\pm$ 1.3	0.06 $\pm$ 0.07
BAD	9.2 $\pm$ 0.6	2.9 $\pm$ 0.5	0.76 $\pm$ 0.04

**Table S4.** Related to Figure 3B. One-tailed  $p$ -values obtained from likelihood ratio tests between the pairs of the polynomial models of degrees  $n$  and  $m$ . The models are best-fitted to the values of skewness (S) and kurtosis (K) as a function of  $\tau_{\text{prior}}$ , estimated in all studied conditions (excluding tetA(st) and BAD(st)) and in various subsets of these conditions. A model where  $\tau_{\text{prior}}$  does not change between conditions is denoted as  $n = 0_{\text{inv}}$ . For  $p$ -values  $\leq 0.05$ , we assumed that the model of degree  $m$  fits the data significantly better than the model of degree  $n$ .

Data set	S			K		
	n	m	$p$ -value	n	m	$p$ -value
All conditions	0	1	0.01	0	1	0.11
	1	2	0.54	0	2	0.15
	1	3	0.03	0	3	0.25
	3	4	0.98	0	4	0.02
	3	5	0.66	4	5	0.93
	3	6	0.84	4	6	0.94
	3	7	0.65	4	7	0.70
	3	8	0.68	4	8	0.83
	3	9	0.65	4	9	0.92
	3	10	0.75	4	10	0.96
	3	11	0.83	4	11	0.98
$0_{\text{inv}}$	3	< 0.001	$0_{\text{inv}}$	4	< 0.001	
Mutants	0	1	0.05	0	1	0.03
	1	2	0.86	1	2	0.77
	1	3	0.98	1	3	0.86
	1	4	0.97	1	4	0.95
	$0_{\text{inv}}$	1	< 0.001	$0_{\text{inv}}$	1	< 0.001
Regulatory factors	0	1	0.01	0	1	0.04
	1	2	0.70	1	2	0.85
	1	3	0.92	1	3	0.72
	1	4	0.90	1	4	0.70
	$0_{\text{inv}}$	1	< 0.001	$0_{\text{inv}}$	1	< 0.001

**Table S5.** Related to Supplementary Figure S6A. One-tailed  $p$ -values obtained from likelihood ratio tests between the pairs of the polynomial models of degrees  $n$  and  $m$ . The models are best-fitted to the values of skewness (S) and kurtosis (K) as a function of  $\tau_{\text{after}}$ , estimated in all studied conditions (excluding tetA(st) and BAD(st)) and in various subsets of these conditions. A model where  $\tau_{\text{prior}}$  does not change between conditions is denoted as  $n = 0_{\text{inv}}$ . For  $p$ -values  $\leq 0.05$ , we assumed that the model of degree  $m$  fits the data significantly better than the model of degree  $n$ .

Data set	S			K		
	n	m	$p$ -value	n	m	$p$ -value
All conditions	0	1	0.17	0	1	0.53
	0	2	< 0.01	0	2	< 0.01
	2	3	0.99	2	3	0.75
	2	4	1.00	2	4	0.30
	2	5	1.00	2	5	0.43
	2	6	1.00	2	6	0.59
	2	7	1.00	2	7	0.56
	2	8	1.00	2	8	0.68
	2	9	1.00	2	9	0.78
	2	10	1.00	2	10	0.86
	2	11	1.00	2	11	0.91
	$0_{\text{inv}}$	2	< 0.001	$0_{\text{inv}}$	2	< 0.001
Mutants	0	1	0.15	0	1	0.08
	0	2	0.24	0	2	0.19
	0	3	0.41	0	3	0.34
	0	4	0.41	0	4	0.30
		$0_{\text{inv}}$	0	< 0.001	$0_{\text{inv}}$	0
Regulatory factors	0	1	0.65	0	1	0.38
	0	2	0.13	0	2	0.09
	0	3	0.21	0	3	0.10
	0	4	0.14	0	4	0.19
		$0_{\text{inv}}$	0	< 0.001	$0_{\text{inv}}$	0

**Table S6.** Related to Supplementary Figure S6B. One-tailed  $p$ -values obtained from likelihood ratio tests between the pairs of the polynomial models of degrees  $n$  and  $m$ . The models are best-fitted to the values of skewness (S) and kurtosis (K) as a function of  $\tau_{\text{prior}}/M$ , estimated in all studied conditions (excluding tetA(st) and BAD(st)) and in various subsets of these conditions. A model where  $\tau_{\text{prior}}$  does not change between conditions is denoted as  $n = 0_{\text{inv}}$ . For  $p$ -values  $\leq 0.05$ , we assumed that the model of degree  $m$  fits the data significantly better than the model of degree  $n$ .

Data set	S			K		
	n	m	$p$ -value	n	m	$p$ -value
All conditions	0	1	0.06	0	1	0.41
	0	2	0.04	0	2	0.04
	2	3	< 0.01	2	3	0.76
	3	4	0.77	2	4	0.02
	3	5	0.95	4	5	0.77
	3	6	0.52	4	6	0.96
	3	7	0.69	4	7	0.99
	3	8	0.81	4	8	1.00
	3	9	0.89	4	9	1.00
	3	10	0.94	4	10	1.00
	3	11	0.97	4	11	1.00
	$0_{\text{inv}}$	3	< 0.001	$0_{\text{inv}}$	4	< 0.001
Mutants	0	1	0.13	0	1	0.07
	0	2	0.23	0	2	0.12
	0	3	0.26	0	3	0.23
	0	4	0.40	0	4	0.28
	$0_{\text{inv}}$	0	< 0.001	$0_{\text{inv}}$	0	< 0.001
Regulatory factors	0	1	0.12	0	1	0.55
	0	2	0.10	0	2	0.12
	0	3	0.16	0	3	0.14
	0	4	0.18	0	4	0.22
	$0_{\text{inv}}$	0	< 0.001	$0_{\text{inv}}$	0	< 0.001

**Table S7.** Related to Figure 4. One-tailed  $p$ -values obtained from likelihood ratio tests between the pairs of the polynomial models of degrees  $n$  and  $m$ . The models are best-fitted to the values of mean ( $M_P$ ), skewness ( $S_P$ ) and kurtosis ( $K_P$ ) of the distribution of protein numbers as a function of the corresponding features ( $M$ ,  $S$  and  $K$ ) of the distribution of time intervals between consecutive RNA production events in individual cells ( $\Delta t$  distribution), estimated in the conditions from the subset 'Mutants' and LA(IPTG) condition. A model where a feature of the  $\Delta t$  distribution does not change between conditions is denoted as  $n = 0_{inv}$ . For  $p$ -values  $\leq 0.05$ , we assumed that the model of degree  $m$  fits the data significantly better than the model of degree  $n$ .

M <sub>P</sub> vs M			S <sub>P</sub> vs S			K <sub>P</sub> vs K		
n	m	$p$ -value	n	m	$p$ -value	n	m	$p$ -value
0	1	< 0.001	0	1	< 0.001	0	1	< 0.001
1	2	0.99	1	2	0.21	1	2	0.17
1	3	0.97	1	3	0.45	1	3	0.38
1	4	1.00	1	4	0.66	1	4	0.58
1	5	1.00	1	5	0.81	1	5	0.58
0 <sub>inv</sub>	1	< 0.001	0 <sub>inv</sub>	1	0.02	0 <sub>inv</sub>	1	0.04

**Table S8.** Related to Figure S3. Comparisons of the  $\Delta t$  distributions (normalized by the mean) of pairs of conditions (see Table 1) by a two-tailed Kolmogorov-Smirnov test. The table shows the  $p$ -values obtained from these tests, for each pair of conditions. In cases where the  $p$ -value  $\leq 0.05$ , the  $H_0$  that the  $\Delta t$  values normalized by the mean are from the same distribution is rejected (highlighted with italics).

	LA(75)	LA(50)	LA(ara)	LA(IPTG)	LA(oxi)	Mut1	Mut2	Mut3	Mut4	tetA	tetA(st)	BAD	BAD(st)
LA	< 0.001	< 0.001	< 0.01	< 0.01	< 0.001	< 0.01	0.78	0.67	0.18	< 0.001	< 0.001	< 0.001	< 0.001
LA(75)		< 0.01	0.28	< 0.001	< 0.001	< 0.01	0.23	0.18	0.05	< 0.001	< 0.001	< 0.001	< 0.001
LA(50)			0.56	< 0.001	< 0.001	< 0.01	0.15	< 0.01	< 0.01	< 0.001	< 0.001	< 0.001	< 0.01
LA(ara)				< 0.01	0.39	0.06	0.14	0.07	0.09	0.31	0.31	0.18	0.43
LA(IPTG)					< 0.001	0.09	0.74	0.67	0.49	< 0.001	< 0.001	< 0.001	< 0.001
LA(oxi)						< 0.001	< 0.01	< 0.001	< 0.001	0.23	0.83	< 0.01	< 0.01
Mut1							0.75	0.92	0.30	< 0.001	< 0.001	< 0.001	< 0.001
Mut2								0.65	0.78	< 0.01	< 0.01	< 0.01	< 0.01
Mut3									0.25	< 0.001	< 0.001	< 0.001	< 0.001
Mut4										< 0.001	< 0.001	< 0.001	< 0.01
tetA											0.41	< 0.01	< 0.001
tetA(st)												< 0.01	< 0.01
BAD													0.14





