© 2011 Lok-hang So

GENE REGULATION IN *ESCHERICHIA COLI* BEYOND THE "RATE" APPROXIMATION

BY

LOK-HANG SO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Robert M. Clegg, Chair
Adjunct Assistant Professor Ido Golding, Director of Research
Assistant Professor Aleksei Aksimentiev
Professor John D. Stack

ABSTRACT

The blueprint of a living cell is inscribed in its DNA. A region of DNA encoding a protein is called a gene. The cell reads the DNA and makes molecular machines made up of proteins to carry out all cellular functions required for survival. All cells live in ever-changing environments, and have different needs at different times. The control of when and how often each protein is produced from a gene is called gene regulation.

Transcription, the copying of a DNA sequence into a complementary mRNA molecule, is the first step in the information flow from DNA to proteins, and most regulation is already done at the transcription level to avoid the production of superfluous intermediates. A living cell takes environmental stimuli as input, and regulates the activity of genes through DNA-binding proteins called transcription factors.

The activity of a gene is described by its time-series of discrete mRNA production events. The events constituting this transcriptional time-series are stochastic and exhibit intermittent, bursty behavior, in bacteria as well as higher organisms. Thus the transcriptional time-series cannot be fully described by a simple chemical "rate"—the probability per unit time of transcribing an mRNA molecule. An important consequence of this temporal complexity is that gene expression level can be tuned by varying different features of the time-series. It is then natural to ask: What modulation scheme is used by the cell to change expression levels of genes? Furthermore, if we look at the transcriptional time-series of multiple genes, would we see different modulation schemes for different genes, or a common modulation scheme shared by all genes? Last but not least, what is the molecular mechanism leading to bursty transcriptional time-series? What are the biophysical states that correspond to the active and inactive periods in a bursty transcriptional time-series?

To answer these questions, I characterized the mRNA copy-number statistics from multiple promoters in the model organism *Escherichia coli* under various growth conditions using single-molecule fluorescence *in situ* hybridization. The kinetics of the underlying transcriptional time-series was then inferred using the two-state model, a simple stochastic mathematical model that describes bursty transcription time-series. I found that the degree of burstiness depends only on the gene expression level, while being independent of the details of gene regulation. The observed behavior is explained by the underlying variation in the duration of bursting events.

At this stage, there is no mechanistic, molecular-level understanding of what gives rise to the bursty behavior of gene activity in bacteria. However, my finding here, that the properties of the transcriptional time-series are gene-independent rather than gene-specific, is contrary to the most common theoretical model used to explain bursty transcriptional time-series in bacteria, which involves the binding and unbinding of transcription factors at the promoter. My data suggests that the observed bursty kinetics arises from gene-nonspecific mechanisms such as DNA topology modulation, RNA polymerase dynamics, or regulation by broad-target DNA-binding proteins. Further investigation would narrow down the source of bursty transcriptional time-series.

To my loving Parents
and my lovely fiancée Angela

大學之道，在明明德，在親民，在止於至善。

—《禮記・大學》

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Ido Golding, for his support, guidance, and enthusiasm throughout the course of my research. Being instructed by Ido, an exceptional scientist, a caring mentor, and a fearless leader, is definitely an eye-opening and a brain-rewiring experience. I would like to thank my dissertation and preliminary exam committee members, Dr. Aleksei Aksimentiev, Dr. Robert Clegg, Dr. Paul Goldbart, and Dr. John Stack, for their insightful advice.

I am cordially thankful to Dr. Ming-chung Chu and Dr. Kwok-leung Liu of the Department of Physics at the Chinese University of Hong Kong, and Dr. Jen-Chieh Peng of the Department of Physics at the University of Illinois at Urbana-Champaign, who gave me opportunities to taste scientific research during my undergraduate years. These experiences had a significant influence on my deciding to pursue a Ph.D. degree.

I would like to give special thanks to my loving parents, Doris and Ting-fai So, and my lovely fiancée, Angela Chan, for their unconditional love, incessant support, stimulating encouragement, and constant understanding over the long years of my Ph.D. program, without which I would not be able to stay sane in such a foreign country as the USA, which is still alien to me in every aspect after having been here for five years.

I sincerely appreciate Christina and Andrew Lai, and Nicole and Po-kin Leung, for helping me adapt to local life and providing valuable advice on all aspects of life.

I am heartily grateful to Yu-Jing Huang and King-lam Hui for keeping their doors, hearts, and ears open whenever I am depressed. I am continually comforted and encouraged by their sincere friendship.

I have benefited greatly from the outpouring generosity, love and support from many friends, without whom I would not be able to stay healthy, both physically and psychologically: My Houston *Badminton Life* friends Chi-Hong Chao, Suet Nee Chen, Yu-Jing Huang, Kai Li Tan, Tsung-Huang Tsai, Wenyi Wang, and Chih-Chao Yang, whose company is always warmly refreshing; My Physics Illinois comrades Kwok-yan Chan, Ting-pong Choy, Po-kin Leung, Ki Lie, Ka-ho Lo, Tak-kei Lui, Yin-yee Ng, Wah-kai Ngai, and Man-hong Yung, with whom I spent numerous energizing (yet exhausting at times) and fun (though usually nerdy) weekends; And my bosom friends Chun-lok Au, Kelvin Chan, Karen Chow, Camay Fung, Daniel Ho, King-lam Hui, and King-yi Lee, with whom I share a reassuring friendship.

My sincere gratitude goes to everyone in the Golding lab, members of the Center for the Physics of Living Cell, the Department of Physics at the University of Illinois at Urbana-Champaign, and the Verna and Marrs McLean Department of Biochemistry and Molecular Biology at Baylor College of Medicine for providing a supportive learning and research environment.

I am grateful to S. Adhya, W. Boos, M. Cashel, V. Chakravartty, Y. Chemla, L. Chubiz, D. Court, J. Cronan, M. Dreyfus, M. Elowitz, Y. Feng, H. Garcia, E. Grace, T. Ha, J. Halliday, C. Herman, T. Hwa, J. Imlay, S. Jang, J. Kondev, T. Kuhlman, J. Little, R. Milo, A. van Oudenaarden, R. Phillips, A. Raj, C. Rao, Á. Sánchez, D. Satory, S. Sawai, P. Selvin, V. Shahrezaei, A. Sokac, P. Swain, S. Syed, and G.Tkačik for advice and providing reagents.

This work was supported by grants from the National Institutes of Health (R01GM082837), the National Science Foundation (082265, PFC: Center for the Physics of Living Cells), and the Human Frontier Science Program (RGY 70/2008).

TABLE OF CONTENTS

1 Background 1

1.1 The Central Dogma of Molecular Biology 1

The blueprint of a cell is stored as a linear genetic code 1

The genetic code is transcribed into a short-lived intermediate, which is then translated into proteins ..1

1.2 *Escherichia coli*, our model organism 2

1.3 Gene regulation and cellular information processing 2

Gene regulation means making the right amount of the right set of proteins at the right time 2

Information flow from DNA to protein is mostly regulated at the transcription level 3

Transcription factors are the regulators of information flow 3

1.4 The "rate" approximation and beyond 4

The Boolean approximation and the "rate" description 4

Beyond the "rate" approximation: bursty transcriptional time-series 5

The two-state model 5

Questions addressed in this work 6

2 Quantifying mRNA Statistics in *E. coli* Using Single-Molecule Fluorescence *in situ* Hybridization 8

2.1 Basics of single-molecule fluorescence *in situ* hybridization 8

Labeling target mRNA with fluorescently tagged oligonucleotide probes 8

mRNA counting at the single-cell level with single-molecule resolution 9

2.2 Obtaining mRNA copy-number statistics from images 10

Automated image analysis 10

Calibration of mRNA counting 10

2.3 Accuracy and dynamic range of smFISH measurements 11

Comparison with quantitative reverse transcription polymerase chain reaction 11

Comparison with data in literature 12

2.4 Limitations of smFISH 15

2.5 Summary 15

3 From mRNA Statistics to Transcription Kinetics 16

3.1 The two-state model and burstiness 16

Predicted population statistics for the two-state model 16

The burstiness of transcriptional time-series can be quantified by the Fano factor 17

3.2 Correcting for gene copy-number 18

An *E. coli* cell typically has multiple copies of its genome 18

Gene copy-number effect can be accounted for by scaling the mean and variance of the mRNA distribution 19

3.3 Correcting for mRNA lifetime 20

mRNA lifetime is affected by numerous factors 20

Differences in mRNA lifetimes can be accounted for by scaling the mean and variance of the mRNA distribution 23

3.4 Summary 23

4 Gene-Independent Behavior of Transcription Burstiness in *E. coli* 24

4.1 The degree of burstiness depends on the mean expression level alone 24

The burstiness of the lactose promoter depends only on the mean expression level, irrespective of whether expression is induced by de-repressor or activator 24

The burstiness of multiple promoters of *E. coli* shows the same dependence on the mean expression level 25

Activity of identical copies of the same gene in a cell exhibit a positive covariance 27

Cell-to-cell variability is dominated by intrinsic noise 27

4.2 Gene expression level is changed by modulating the gene off-rate 28

The dependence of the burstiness on the expression level in different modulation schemes is different 28

Gene off-rate modulation scheme is consistent with the observed behavior 28

4.3 Gene-independent behavior suggests transcription bursts arise from global regulation mechanisms 30

4.4 Summary 31

5 Support from Live-Cell Data 36

5.1 Real-time gene activity measurement in individual live cells 36
Labeling mRNA transcribed from the promoter-of-interest with MS2-GFP 36
Real-time mRNA counting in individual live cells with single-molecule resolution 37
Automated image analysis, lineage reconstruction, and detection of transcription events 38
Calibration of mRNA counting 38
5.2 Live-cell data is consistent with the gene off-rate modulation scheme 38
5.3 Limitations of the live-cell method 39
5.4 Summary 40

6 Summary and Outlook 41

6.1 Gene-independent behavior of transcription burstiness 41
6.2 Proposed molecular mechanisms that could give rise to transcription bursts 42
6.3 Extending live-cell measurements to endogenous promoters 42
6.4 Improving the absolute calibration of mRNA counting 43

Appendix A. Strains, Growth Media, and Growth Conditions 45

A.1 Growth media and conditions 45
A.2 Strains 46
Bacterial strains 46
Phage strains 47
Promoters 48

Appendix B. Detailed Description of Experimental Protocols 52

B.1 Single-molecule fluorescence *in situ* hybridization (smFISH) 52
Probe design and labeling 52
Sample fixation and permeabilization 55
Hybridization 55
Washing 56
Microscopy 56
B.2 Quantitative reverse transcription polymerase chain reaction 57

Total RNA extraction 57
qRT-PCR experiment 58
qRT-PCR data analysis 59
RNA lifetime measurement 59
B.3 Measurement of gene activity in individual live cells 59
Strains and plasmid 59
Time-lapse microscopy 60
Image processing and data analysis 60

Appendix C. Data Analysis 61

C.1 Spot recognition 61
C.2 Step detection 61

Appendix D. Modeling Transcription Kinetics 63

D.1 Gillespie simulation of the two-state model 63
D.2 Simultaneous modulation of all three kinetic parameters 63

References 66

1 Background

In this chapter, the basic biology upon which this work is built will be introduced. I will first describe the Central Dogma of Molecular Biology, and introduce the organism studied in this work, *Escherichia coli*. I will then explain the concepts of gene regulation and cellular information processing. Finally, I will introduce the concept of bursty transcriptional time-series and the phenomenological two-state model, and bring up the questions addressed in this work.

1.1 The Central Dogma of Molecular Biology

The blueprint of a cell is stored as a linear genetic code

The Central Dogma of Molecular Biology (Crick, 1970) is the universal scheme of all living cells for storing, propagating, and utilizing genetic information[1]. The blueprint (genome) of a living cell is inscribed in double-stranded linear molecules of deoxyribonucleic acid (DNA) (Alberts, 2008). A DNA molecule is made up of monomers called nucleotides. There are four types of nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C). The linear sequences of these four letters, A, T, G, and C, on DNA molecules encode all hereditary information about the living cell, in the same way sequences of 0's and 1's encode electronic data.

A cell has one or more copies of its DNA. DNA is replicated (**Figure 1.1**, looping arrow) and passed on to progeny cells during cell division.

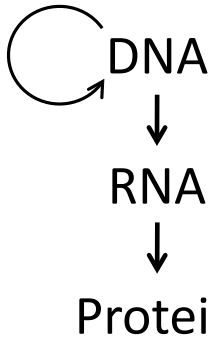


Figure 1.1. A simplified version of the Central Dogma of Molecular Biology.

DNA is replicated and passed on to progeny cells. Genetic information stored on DNA is transcribed into RNA, which is then translated to proteins, the building blocks of the molecular machines of the cell. Adapted from Crick, 1970.

The genetic code is transcribed into a short-lived intermediate, which is then translated into proteins

To carry out specific tasks required for survival, cells need molecular machines made up of proteins. Proteins are linear molecules made up of monomers called amino acids, folded into specific three-dimensional conformations, capable of catalyzing biochemical reactions in a living cell (Alberts, 2008). The amino acid sequence of a protein is stored as a gene on a DNA molecule. Every amino acid can be represented by a three-letter combination of the genetic code. The sequence of letters on a gene, read in triplets, uniquely defines an amino acid sequence (Alberts, 2008).

To make a protein, information stored in a gene on a DNA molecule is first transcribed into a short-lived intermediate called messenger ribonucleic acid (messenger RNA, or simply mRNA) by an enzyme

[1] Note that there are exceptions to rules in biology most of the time. Here, processes like the reverse transcription of RNA into DNA (Alberts, 2008) are not included. Importantly, this simplified picture suffices for the purpose of this work.

called RNA polymerase (**Figure 1.1**, arrow from DNA to RNA) (Alberts, 2008). RNA molecules are linear molecules similar to DNA. The genetic information on an mRNA molecule is then translated into proteins by the ribosome (**Figure 1.1**, arrow from RNA to Protein). The sequence of letters on an mRNA is read in triplets, and translated into a sequence of amino acids (Alberts, 2008).

1.2 *Escherichia coli*, our model organism

Before going deeper into the ideas of gene regulation and cellular information processing, I introduce *E. coli*, our model organism[2]. *E. coli* is a rod-shaped bacterium about 2 μm long and 0.5 μm wide (**Figure 1.2**) (Neidhardt, 1987). It serves as a model organism in biology (Campbell and Reece, 2005). In particular, the laboratory strain K-12 grows rapidly, has undemanding growth requirements, is metabolically versatile, and has tractable genetics (Hobman, Penn *et al.*, 2007). A lot is known about the physiology, genetics, molecular biology, and biochemistry of *E. coli* (Higgins, 2005; Neidhardt, Ingraham *et al.*, 1990; Neidhardt, 1987). Its complete genome sequence is known (Blattner, Plunkett *et al.*, 1997). Standard experimental methods for manipulating *E. coli* are readily available, including procedures for modifying the DNA of *E. coli* and performing biochemical analysis of *E. coli* cells (Ausubel, 1987; Miller, 1992; Sambrook and Russell, 2001). It is an ideal experimental platform for quantitative studies of living cells (Alon, 2007; Locke and Elowitz, 2009; Pearson, 2008; Phillips, Kondev *et al.*, 2009), and will be the focus of this work.

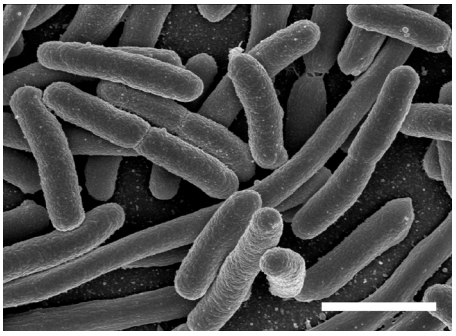


Figure 1.2. Scanning electron micrograph of *E. coli* cells.

E. coli is a rod-shaped bacterium. It serves as a model organism and is one of the most commonly used organisms in biological research. Scale bar equals 2 μm. (Credit: Rocky Mountain Laboratories, NIAID, NIH.)

1.3 Gene regulation and cellular information processing

Gene regulation means making the right amount of the right set of proteins at the right time

The DNA of a cell contains the information needed to produce all the proteins the cell needs. However, all living cells face an ever-changing environment. For instance, different food might be available to an *E. coli* cell at different times, and the cell would need different proteins to digest them (Neidhardt, Ingraham *et al.*, 1990). The cell might at times need molecular propellers that consist of proteins to move towards nutrients or away from toxins (Neidhardt, Ingraham *et al.*, 1990). When the cell

[2] A *model organism* is an organism that has been studied intensively over a long period to understand particular biological phenomena, with the expectation that discoveries made in the organism will provide insight into the working of other organisms.

suffers from DNA damage, it would need DNA repair proteins (Neidhardt, Ingraham *et al.*, 1990). To optimize the utilization of energy and materials, a cell gathers information about its environment, makes the proteins it needs, and shuts down the production of unnecessary proteins (Alberts, 2008). The control of the production of the right amount of the right set of proteins at the right moment of time is called *gene regulation*. Gene regulation is essential to all living cells, allowing them to survive and thrive in changing environments.

Information flow from DNA to protein is mostly regulated at the transcription level

The information flow from DNA to protein can be controlled by regulating any step along the flow. For example, a cell can change the rate of transcription or the rate of translation. The transcription of genes from DNA into mRNA is the first step in the information flow from genes to proteins, so it is natural that most regulation is done at the transcription level, avoiding the production of superfluous intermediates (Alberts, 2008).

Transcription factors are the regulators of information flow

Transcription factors are protein molecules that monitor the environment and help the cell regulate all of its genes. Transcription factors bind to specific DNA sites. Upon binding, transcription factors can either activate or repress the transcription of specific genes by modulating the effectiveness of RNA polymerase binding to and initiating transcription from *promoters*[3]. Environmental signals change the DNA-binding affinities, activating abilities, and repressing abilities of transcription factors, thereby regulating transcription. Information flows from the environment to transcription factors, whose activities constitute an intra-cellular representation of the environment signals, then to the genes (**Figure 1.3**). In this sense, a cell works just like a microscopic information processor. It takes environmental signals as inputs, uses transcription factors to "calculate" the types and amounts of proteins to make, and responds to the environmental signals by making the appropriate proteins (Alon, 2007).

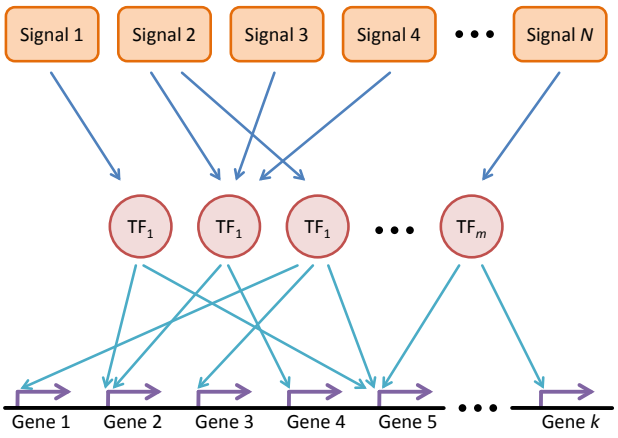


Figure 1.3. Transcription factors are the regulators of information flow.

Transcription factors (TF's) are protein molecules that monitor the environment and regulate genes accordingly. Adapted from Alon, 2007.

As an example, the Lac repressor, one of the first transcription factors that was identified, isolated, and characterized, monitors the presence of lactose in the environment (Müller-Hill, 1996). When there is no lactose, Lac repressor binds to specific sites on the *E. coli* DNA called *lac* operators to repress

[3] A *promoter* is a region on the DNA upstream of a gene, where RNA polymerase binds and initiates the transcription of mRNA.

transcription of the genes involved in lactose metabolism. When lactose is present, however, Lac repressor is inactivated, and is no longer able to bind DNA. Lactose metabolism genes are no longer repressed, and the proteins responsible for transporting and digesting lactose are produced (Müller-Hill, 1996).

1.4 The "rate" approximation and beyond

The Boolean approximation and the "rate" description

To capture the essence of simple gene regulation schemes, the Boolean approximation is sometimes used. In this description, genes are either maximally "on" or completely "off" (Alon, 2007). Interacting networks of transcription factors and genes can be conveniently treated as combinations of logic gates and analyzed using Boolean algebra. In some cases the Boolean description suffices if a qualitative picture is all we need. However, for quantitative studies, the Boolean approximation might not suffice. To improve the fidelity we can use the "rate" approximation, where we consider gene activity as a Poisson process characterized by a "rate"—the probability of transcribing an mRNA molecule per unit time. **Figure 1.4A** shows a Poisson *transcriptional time-series*[4], where each event (vertical bar) in the time-series is the production of one mRNA molecule. The waiting times between the events are exponentially distributed. This is a significant step forward, and can yield a lot more quantitative information about the system.

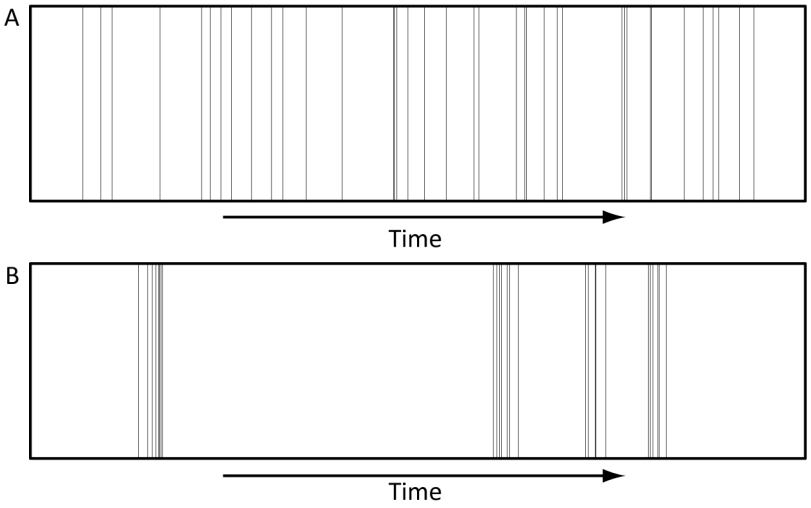


Figure 1.4. Poisson and bursty transcriptional time-series.

The horizontal axis is time in arbitrary units. Each vertical bar represents one transcription event, in which one mRNA molecule is produced.

(A) A Poisson transcriptional time-series is characterized by a constant probability of transcription per unit time, resulting in exponentially distributed waiting times between events.

(B) A bursty transcriptional time-series is characterized by short, intense bursts of events, separated by extended periods of quiescence. The two time-series give the same mean mRNA molecules per cell. Adapted from Golding and Cox, 2006.

[4] A *transcriptional time-series* is a discrete series of transcription events in time. Each event in the time-series is the production of one mRNA molecule.

Beyond the "rate" approximation: bursty transcriptional time-series

We now know that the transcription in *E. coli* is bursty, characterized by short, intense bursts of mRNA production events, separated by extended periods of quiescence (**Figure 1.4B**) (Golding and Cox, 2006; Golding, Paulsson *et al.*, 2005). This is also true for other organisms studied so far, including yeast (Zenklusen, Larson *et al.*, 2008), social amoeba (Chubb, Trcek *et al.*, 2006), and mammalian cells (Raj, Peskin *et al.*, 2006; Yunger, Rosenfeld *et al.*, 2010). This means a single "rate" would not be enough to describe the bursty transcriptional time-series of mRNA production. In other words, instead of just varying the "rate" of transcription as in the Poisson scenario, there are multiple possible ways to change the mean expression level (mRNA molecules per cell) of a gene.

The two-state model

A simple mathematical model suitable for describing bursty transcriptional time-series is the phenomenological two-state model (**Figure 1.5A**) (Peccoud and Ycart, 1995; Raj, Peskin *et al.*, 2006; Shahrezaei and Swain, 2008), where a gene stochastically switches between "off" and "on" states, with rates k_{on} and k_{off}, and mRNA is produced stochastically only in the "on" state, with rate k_{TX}. mRNA is degraded stochastically with rate k_d. This model naturally leads to bursty transcriptional time-series, with the "off" state corresponding to the quiescence periods and the "on" state corresponding to the periods of activity. The two-state model can be solved analytically to give expressions for the mean, variance, and all higher moments of the steady-state mRNA distribution in terms of the kinetic parameters k_{on}, k_{off}, k_{TX}, and k_d (see **Section 3.1** for details) (Peccoud and Ycart, 1995; Raj, Peskin *et al.*, 2006; Shahrezaei and Swain, 2008). Measured mRNA kinetics (Chubb, Trcek *et al.*, 2006; Golding, Paulsson *et al.*, 2005) and copy-number statistics (Golding, Paulsson *et al.*, 2005; Raj, Peskin *et al.*, 2006; Tan and van Oudenaarden; Zenklusen, Larson *et al.*, 2008) have been shown to be consistent with the two-state picture in a variety of model systems. However, this picture is still purely

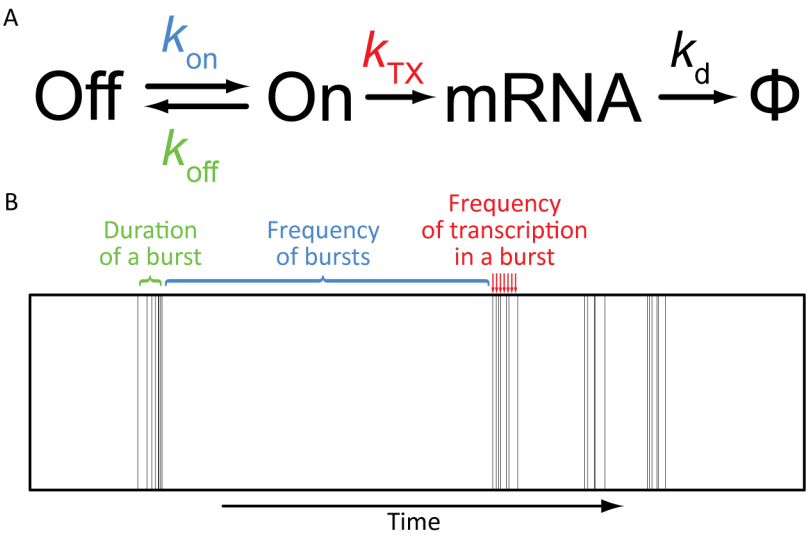


Figure 1.5. The two-state model.

(A) The two-state model is a simple, phenomenological, probabilistic mathematical model for describing bursty transcriptional time-series. A gene switches between "on" and "off" states with rates k_{on} and k_{off}. mRNA is produced with rate k_{TX} only when the gene is "on" and existing mRNA molecules are degraded with rate k_d.

(B) Modulating each of the three kinetic parameters of mRNA production in the two-state model corresponds to changing one of the three features of the bursty transcriptional time-series.

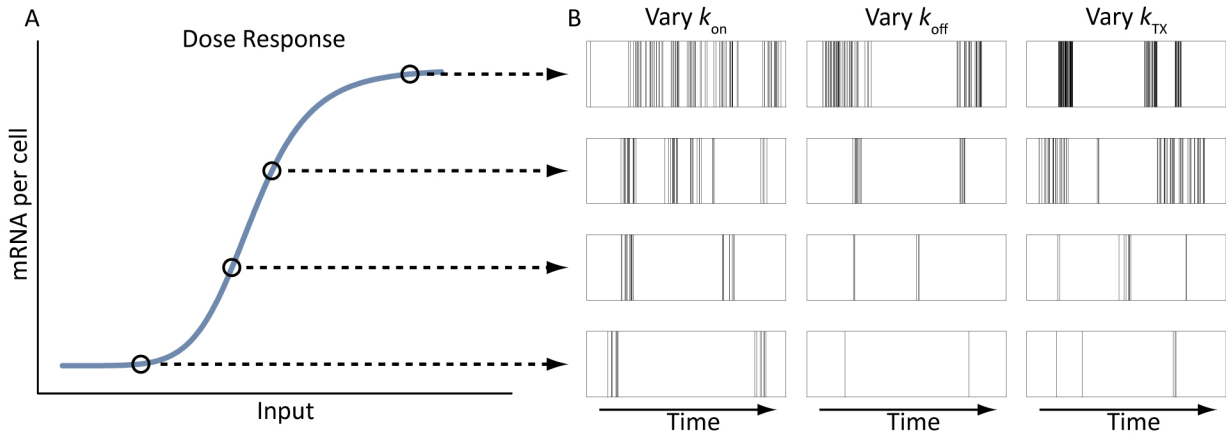


Figure 1.6. Different features of the transcriptional time-series can be modulated to vary gene expression level.

(A) Schematic representation of the gene-activity (dose-response) curve for a typical bacterial promoter. The expression level (mRNA per cell) as a function of the external stimulus is shown. The curve shown is arbitrary, but is typical of the sigmoidal response exhibited by many bacterial promoters (Alon, 2007; Kuhlman, Zhang *et al.*, 2007), for example see the experimentally measured curve for the lactose promoter in **Figure 2.2A**.

(B) Different modulation schemes of the transcriptional time-series, all capable of creating the same gene activity curve in panel A. Each plot shows a time-series of producing mRNA molecules (vertical bars). In the three cases shown, only a single parameter of gene activity was varied in each scheme: (i) the rate of switching the gene on, k_{on} (left); (ii) the rate of switching the gene off, k_{off} (middle); and (iii) the rate of transcription when in the "on" state, k_{TX} (right). Modifying each of these parameters leads to a modulation of a different feature in the transcriptional time-series: the rate of transcription bursts (left), the duration of bursts (middle) and the number of mRNA molecules produced in each burst (right). Notice that all time-series in the same row lead to the same mean mRNA level.

phenomenological. Despite considerable theoretical attention (Blake, M *et al.*, 2003; Dobrzynski and Bruggeman, 2009; Fang, 2005; Golding, Paulsson *et al.*, 2005; Mitarai, Dodd *et al.*, 2008; van Zon, Morelli *et al.*, 2006) we do not possess a biophysical understanding of the nature of the "on" and "off" states and what governs the transitions between them.

An important consequence of the burstiness of transcriptional time-series is that the expression level of a gene does not uniquely determine the parameters of the time-series. In other words, an ensemble of many different time-series can lead to the same mRNA level. Similarly, a change in the level of expression (as, for example, in response to different stimulus levels) can in principle occur by varying the different properties of the transcriptional time-series (**Figure 1.5B**) (Raj, Peskin *et al.*, 2006). Changes in the mRNA level can be obtained by modifying any of the three kinetic parameters characterizing mRNA production: k_{on}, the rate of switching to the "on" state ("on rate"), which determines the rate of transcription bursts; k_{off}, the rate of switching back to the "off" state ("off rate"), which determines the duration of transcription bursts; and k_{TX}, the rate of producing mRNA while in the "on" state, which determines how frequently mRNA molecules are produced during each transcription burst (**Figure 1.6**). We call the different ways of changing the expression level *modulation schemes*.

Questions addressed in this work

Now that we know the transcriptional time-series in *E. coli* is bursty, a natural question arises: What modulation scheme is used by the cell to change the expression levels of genes? Furthermore, if we look at

the transcriptional time-series of multiple genes in *E. coli*, would we see different modulation schemes for different genes, or a common modulation scheme shared by all genes? Last but not least, what is the molecular mechanism leading to bursty transcriptional time-series in *E. coli*? What are the biophysical states that correspond to the “on” and “off” states in the phenomenological two-state model?

2 Quantifying mRNA Statistics in *E. coli* Using Single-Molecule Fluorescence *in situ* Hybridization

To address the questions raised in the previous chapter, I need to quantify the transcription kinetics in *E. coli* cells. To do so, the copy-number statistics of mRNA molecules was first obtained using single-molecule fluorescence *in situ* hybridization (smFISH), an experimental method that will be introduced in this chapter. The image analysis procedures for obtaining mRNA statistics from smFISH experiments will be described, and tests that validate the accuracy and dynamic range of smFISH will be presented. In the next chapter, I will explain how the underlying transcription kinetics was inferred from the mRNA copy-number statistics obtained from smFISH experiments.

2.1 Basics of single-molecule fluorescence *in situ* hybridization

Labeling target mRNA with fluorescently tagged oligonucleotide probes

smFISH was first developed by Raj *et al.* (Raj, van den Bogaard *et al.*, 2008) based on Femino *et al.* (Femino, Fay *et al.*, 1998), and later adapted to *E. coli* in our lab (Zong, So *et al.*, 2010). Briefly, A set of ~50–70 fluorescently labeled *oligonucleotide* [5] probes, each 20 bases in length and complementary to the target mRNA sequence, was designed and synthesized. Cells were grown in desired conditions, harvested in the *exponential phase*[6], and chemically fixed. The fixed cells were then washed and permeabilized. The fluorescent probes were then hybridized to target mRNA molecules, and the cells were washed and imaged under a fluorescence microscope. The target mRNA

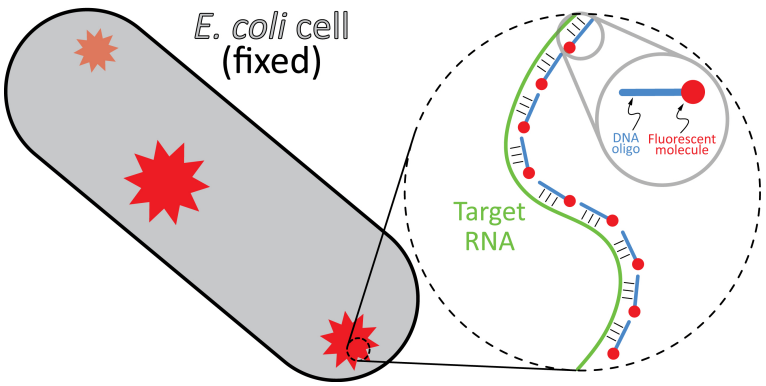


Figure 2.1. A schematic of the smFISH technique.

A set of ~50–70 fluorescently labeled oligonucleotide probes, each 20 bases in length, was designed against the target mRNA. The probes were hybridized to target mRNA molecules, and the target mRNA molecules appeared as fluorescence spots under the microscope. Integrating the total fluorescence from all spots allows estimation of the mRNA number in each cell.

[5] An *oligonucleotide* is a single-stranded DNA molecule with ~50 bases or less.

[6] *Exponential phase* of bacterial growth is characterized by cells dividing into two daughter cells at a typical doubling time. The number of new cells appearing per unit time is proportional to the number of existing cells, and the total number of cells increases exponentially.

molecules, decorated with fluorescent probes, would appear as localized fluorescence spots in cells (**Figure 2.1** and **Figure 2.2B**). The experimental protocol is detailed in **Appendix B.1**.

mRNA counting at the single-cell level with single-molecule resolution

To estimate the number of mRNA molecules from the gene-of-interest in a given cell, the total intensity of the fluorescent spots in the cell was measured, yielding an estimate of the number of bound probes, in turn indicating the number of target mRNA molecules (see the next section for more details). This approach follows the one previously used in live-cell studies of mRNA kinetics using the MS2-GFP system (Golding, Paulsson *et al.*, 2005). mRNA levels covering almost three orders of magnitude (~0.1–60 mRNA molecules per cell) could be measured using smFISH (**Figure 2.2A**, red curve). The smFISH-based measurements allowed us to obtain the probability distribution of mRNA copy-number from a gene-of-interest under a given growth condition (e.g. at a specific concentration of a specific nutrient). The mRNA histograms were well described by negative binomial distributions (**Figure 2.2C**) (Paulsson and Ehrenberg, 2000), consistent with the prediction of the two-state model (see **Section 3.1** for more

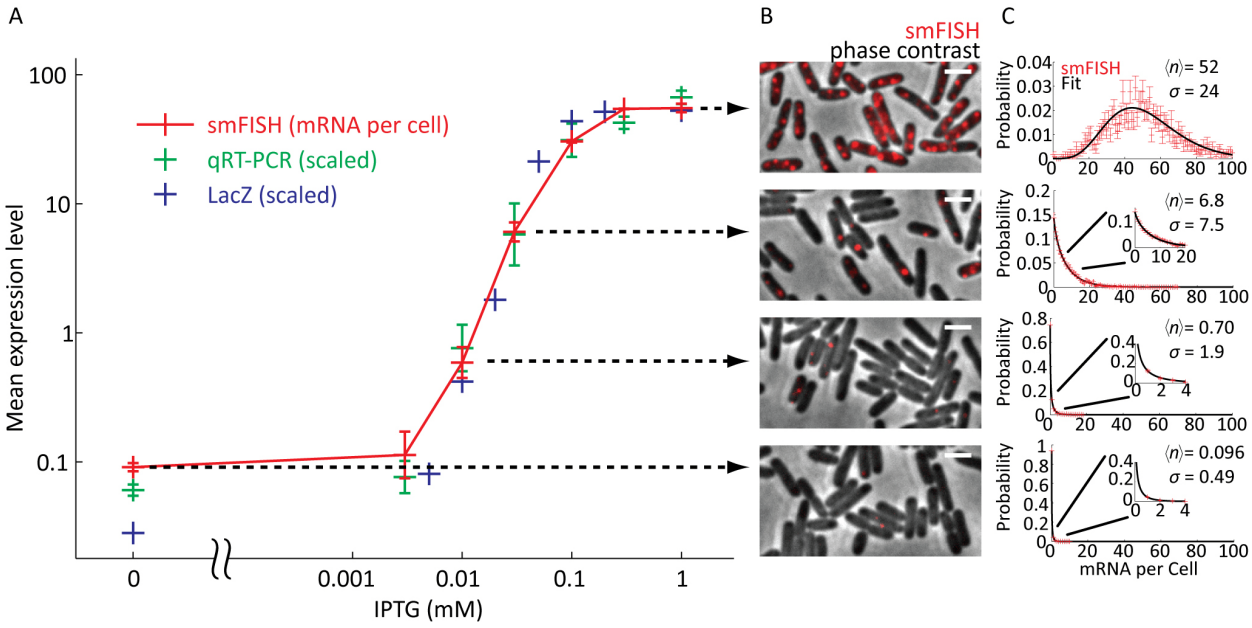


Figure 2.2. Characterizing mRNA copy-number statistics using smFISH.

(A) The mean expression level of the *E. coli* lactose promoter P_{lac}, as a function of inducer (isopropyl β-D-1-thiogalactopyranoside, IPTG) concentration. The expression level as measured by smFISH (in units of mRNA per cell, average of two independent experiments) is shown, as well as the results of quantitative reverse transcription polymerase chain reaction (qRT-PCR; average of two independent experiments; normalized by the mean smFISH level; see **Section 2.3**) and β-galactosidase (LacZ) activity assay, as reported in the literature (Kuhlman, Zhang *et al.*, 2007, normalized by the mean smFISH level). Error bars denote the standard errors from duplicate experiments. The good agreement between the three assays, over almost three orders of expression level, demonstrates the accuracy and dynamic range of the smFISH method.

(B) Typical images of smFISH-labeled cells at different induction levels. An overlay of the phase contrast (grayscale) and smFISH probes targeting the *lacZ* gene (red) is shown. Each image corresponds to the expression level designated by the horizontal arrow. Scale bars equal 2 μm.

(C) *lacZ* mRNA copy-number histograms obtained from smFISH at different induction levels. The experimental data (red) and the fit to a negative binomial distribution (black) are shown, as well as the estimated values for the mean mRNA number $\langle n \rangle$ and standard deviation σ in each sample. Each plot corresponds to the expression level designated by the horizontal arrow.

details) (Raj, Peskin *et al.*, 2006; Shahrezaei and Swain, 2008). In particular, the smFISH data allowed us to accurately measure both the mean and variance of mRNA copy-number.

2.2 Obtaining mRNA copy-number statistics from images

Automated image analysis

Hundreds of microscope images were produced from each run of smFISH experiment. Homemade MATLAB (MathWorks) programs were used to automate image processing, yielding mRNA copy-number statistics of populations of cells.

Cell recognition was performed on phase contrast images of cells using the *Schnitzcell* MATLAB module (gift of Michael Elowitz, California Institute of Technology) (**Figure 2.3A**). The program utilizes edge detection and other morphological operations from the MATLAB Image Processing Toolbox. The output was checked and corrected using the manual graphical user interface in *Schnitzcell*. The output consisted of label matrices representing areas occupied by cells, from which quantities such as the length, position, and area of cells were readily extracted.

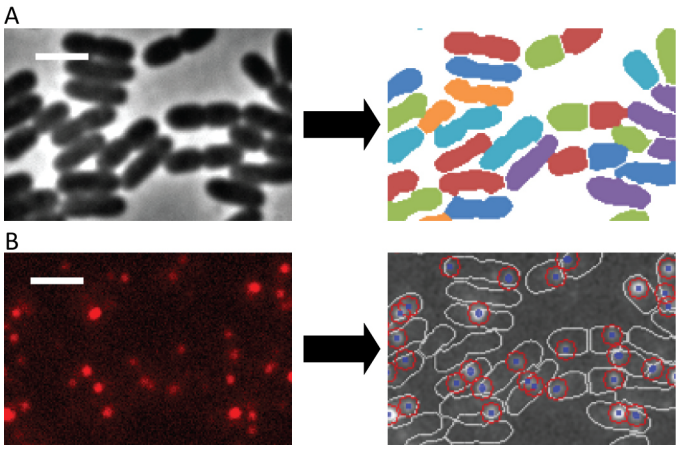


Figure 2.3. Automated image analysis using MATLAB.

(A) Cell recognition based on phase contrast images using *Schnitzcell*. The user gets label matrices of areas occupied by single cells as output.

(B) Spot recognition and quantification program identifies the spots, allocates them to the cells, and performs relevant measurements based on fluorescence images. White outlines represent the recognized cells. Scale bars equal 2 μm.

A spot recognition program developed in our lab was used to automatically identify and quantify localized fluorescence signals (**Figure 2.3B**). A Gaussian filter was first applied to smooth out noise, and spots were recognized by the presence of a local maximum in both x- and y-directions. This was done at each z-position in the stack of images, and each spot was quantified at the z-position where it had the highest fluorescence intensity (where the spot is in focus). See **Appendix C.1** for further details.

Calibration of mRNA counting

A fluorescence spot could consist of multiple mRNA molecules in close proximity. The integrated intensity arising from a single mRNA molecule needed to be estimated for each smFISH experiment so that the fluorescence intensities could be normalized to give the absolute number of mRNA molecules. The typical intensity of the "false positives" in an experiment was first estimated from the histogram of individual spot intensities of a negative control (i.e. samples with no target mRNA at all) (**Figure 2.4**, red

curves). The histograms of individual spot intensities from relatively low expression samples were then examined (**Figure 2.4**, bar graphs). Because most spots in these samples were expected to contain a single mRNA molecule, the first peak that emerged above the false positive intensity range in each of these histograms served as an estimate for the intensity of a single mRNA molecule. The mean intensity of the first peaks from multiple such histograms was taken as the single mRNA intensity for that particular experiment (**Figure 2.4**, orange line). The sum of the intensities of all the spots in each cell was then normalized to give the absolute number of mRNA molecules. Note that this algorithm is far from perfect, and based on the reproducibility between experiments, I estimate its accuracy at ±25%, but it is possible that additional systematic errors make the accuracy even lower. Possible ways of improving the calibration will be discussed in **Section 6.4**.

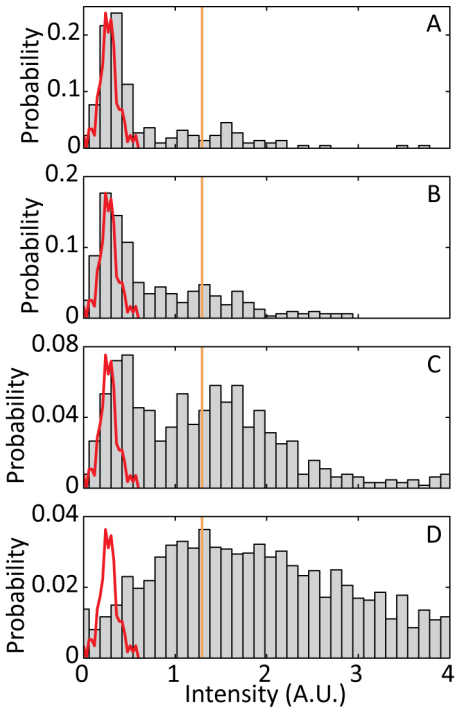


Figure 2.4. Calibration of mRNA counting. The red curve in each of the four plots is the histogram of individual spot intensities from a negative control. The bar graphs are histograms of the spot intensities from samples with (A) 0.10, (B) 0.15, (C) 0.73, and (D) 7.04 mRNA molecules per cell (numbers estimated from smFISH). The orange line denotes the mean of the first peaks from the grey bar graphs above the false positive level, which is taken as the intensity that corresponds to one mRNA molecule.

2.3 Accuracy and dynamic range of smFISH measurements

For a subset of the promoters used in this study, I compared the mRNA levels measured using smFISH to numbers obtained using other well-established methods. I was thus able to establish the accuracy and dynamic range of smFISH.

Comparison with quantitative reverse transcription polymerase chain reaction

Quantitative reverse transcription polymerase chain reaction (qRT-PCR) is a well-established standard assay for measuring the population average of target RNA levels (Ausubel, 1987; Bustin and Nolan, 2004; Nolan, Hands *et al.*, 2006; Schmittgen and Livak, 2008). In a typical qRT-PCR reaction, the target RNA is first reverse-transcribed into complementary DNA (cDNA), and the cDNA is replicated in the form of double-stranded DNA (dsDNA), as in a conventional PCR reaction. The presence of dsDNA product is detected by SYBR Green, a dye that fluoresces when bound to dsDNA. A threshold fluorescence is chosen at where all the PCR reactions are in the exponentially replicating phase (**Figure B.4A**). The larger the initial amount of RNA in a sample, the earlier the fluorescence reaches the threshold. Defining Ct_i to be the cycle number at which the fluorescence of sample i reaches the threshold, and $N_{0,i}$ to be the initial amount of target RNA in sample i, I have

$$N_{0,i}E^{\mathrm{Ct}_i} = K, \tag{2.1}$$

where E is the efficiency of the PCR (defined as the fold-increase of the amount of dsDNA product per PCR cycle), and K is a constant determined by where you set the threshold. E is obtainable from the dilution series, because the relative amounts of target RNA (i.e. ratios of $N_{0,i}$) in each sample in the dilution series is known (**Figure B.4C**). The relative initial amount of RNA is then given by

$$N_{0,i} \propto E^{-\mathrm{Ct}_i}. \tag{2.2}$$

For details about qRT-PCR data analysis see **Appendix B.2**.

The expression level of the lactose promoter P_{lac} as a function of IPTG concentration in the strain TK310 (**Appendix A**) (Kuhlman, Zhang *et al.*, 2007), obtained from smFISH experiments, were compared to the expression levels measured using qRT-PCR. Expression levels covering almost three orders of magnitude (~0.1–60 mRNA molecules per cell) could be measured using smFISH. The smFISH-based estimation of mRNA numbers was in excellent agreement with measurements using qRT-PCR (**Figure 2.2A**, green data points), demonstrating the accuracy and dynamic range of smFISH.

Comparison with data in literature

I compared data obtained using smFISH and data from the literature whenever available. The data show excellent agreements in all instances.

1. P_{lac}

P_{lac} is the promoter expressing the lactose metabolism genes. P_{lac} expression level as a function of IPTG (a lactose analogue) concentration in the strain TK310 (Kuhlman, Zhang *et al.*, 2007), obtained from smFISH experiments, were compared to the expression levels obtained using the β-galactosidase (LacZ) activity assay from the literature (**Figure 2.2A**, blue data points) (Kuhlman, Zhang *et al.*, 2007). The results obtained from the two assays agree very well.

2. *rrnB*P1

*rrnB*P1 is the promoter expressing ribosomal RNA (rRNA), the building blocks of ribosomes. *rrnB*P1 expression level as a function of the growth rate in the strain CF7753 (**Appendix A**) (gift of Michael Cashel, National Institutes of Health) (Hernandez and Bremer, 1990; Potrykus, Vinella *et al.*, 2006), obtained from smFISH experiments, were compared to the known numbers of ribosomes per cell (Neidhardt, 1987). In making the comparison, a correction factor was included, to compensate for the difference in lifetime between endogenous ribosomal RNA (which is stable) and the *lacZ* reporter used (estimated using qRT-PCR, see **Appendix B.2**): If n is the number of *lacZ* mRNA molecules from my *rrnBP*1 reporter strain and N is the number of rRNA molecules in a cell, I have

$$\begin{cases} \dot{n} = P_{\mathrm{smFISH}} - nk_{\mathrm{d,mRNA}} \\ \dot{N} = P_{\mathrm{rib}} - Nk_{\mathrm{d,rRNA}} \end{cases}, \tag{2.3}$$

where P_{smFISH} is the *rrnB*P1 promoter activity (transcripts produced per unit time) to be estimated using smFISH, P_{rib} is the *rrnB*P1 promoter activity to be estimated using known numbers from the literature, $k_{\text{d,mRNA}}$ is the mRNA degradation rate, and $k_{\text{d,rRNA}}$ is the rRNA degradation rate. In steady-state, $\dot{n}$ and $\dot{N}$ vanish, so

$$\begin{cases} n = P_{\text{smFISH}}/k_{\text{d,mRNA}} = P\tau_{\text{d,mRNA}} \\ N = P_{\text{rib}}/k_{\text{d,rRNA}} = P\tau_{\text{d,rRNA}} \end{cases}, \tag{2.4}$$

where $\tau_{\text{d,mRNA}}$ is the mRNA lifetime and $\tau_{\text{d,rRNA}}$ is the rRNA lifetime. rRNA is stable (Neidhardt, 1987), so it is only diluted by cell growth and division. Therefore $\tau_{\text{d,rRNA}}$ is equal to the cell generation time τ_{g} (numbers from Neidhardt, 1987). On the other hand, I assume that the variation in $\tau_{\text{d,mRNA}}$ is small compared to the variation in $\tau_{\text{d,rRNA}}$ when the cell generation time is varied. Thus all the variables but $\tau_{\text{d,mRNA}}$ in **Equation (2.4)** changes with the cell growth rate. As a result, I can write

$$\begin{cases} P_{\text{smFISH}} = n/\tau_{\text{d,mRNA}} \propto n \\ P_{\text{rib}} = N/\tau_{\text{d,rRNA}} \propto N/\tau_{\text{g}} \end{cases}. \tag{2.5}$$

That means I can scale the number of ribosomes with the cell generation time, and compare it with the number of mRNA molecules obtained using smFISH. As shown in **Figure 2.5A**, the results obtained from the two methods agree very well.

3. P_{marII}

P_{marII} is the promoter expressing the genes responsible for resistance to multiple antibiotics and toxic chemicals. P_{marII} expression level in the strain LC544 (**Appendix A**) (gift of Christopher Rao, University of Illinois at Urbana-Champaign) as a function of sodium salicylate concentration, obtained from smFISH experiments, was compared to the data obtained using a yellow fluorescence protein reporter (unpublished data from Lon Chubiz, Christopher Rao Lab, University of Illinois at Urbana-Champaign). The results of the two assays agree very well (**Figure 2.5B**).

4. $P_{bioBFCD}$

$P_{bioBFCD}$ is the promoter expressing the biotin metabolism genes. $P_{bioBFCD}$ expression level in the strain CY481 (**Appendix A**) (gift of John Cronan, University of Illinois at Urbana-Champaign) (Abdel-Hamid and Cronan, 2007; Barker and Campbell, 1980) as a function of biotin concentration, obtained from smFISH experiments, was compared to the data obtained using the LacZ activity assay from the literature (Abdel-Hamid and Cronan, 2007). The results of the two assays agree very well (**Figure 2.5C**).

5. P_{RM}

P_{RM} is the promoter expressing the bacteriophage λ repressor CI, a protein responsible for the maintenance of lysogeny. The expression of P_{RM} is autoregulated by CI. P_{RM} expression level of lysogens carrying the *cI*D38N mutation and various P_{RM} mutations in the host strain JL2497 ($RecA^+$), obtained

from smFISH experiments, were compared to the data obtained using the LacZ activity assay from the literature (Michalowski and Little, 2005). The two data sets agree very well[7] (**Figure 2.5D**).

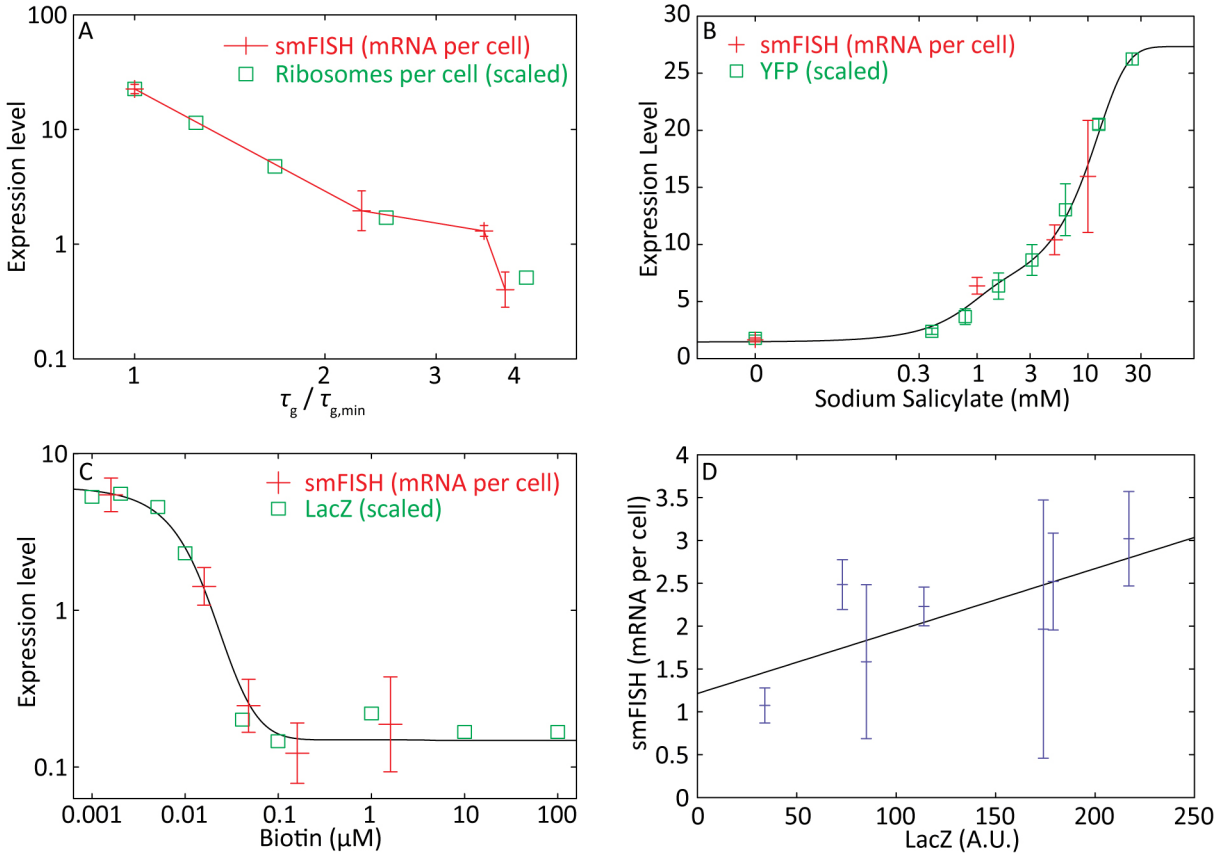


Figure 2.5. Comparison of the expression levels obtained using smFISH with the expression levels obtained using other methods.

(A) The expression level of the ribosomal promoter *rrnB*P1 as a function of growth rate obtained using smFISH and from the known number of ribosomes per cell (Neidhardt, 1987), scaled to compensate for the difference in lifetime between stable endogenous ribosomal RNA and *lacZ* reporter mRNA. The red error bars represent the standard deviation of the data obtained from two independent repeats of smFISH experiments. The x-axis is the generation time normalized by the minimum generation time achievable by the strain studied.

(B) The expression level of the P_{marII} promoter as a function of sodium salicylate concentration obtained using smFISH, compared to the data obtained using a yellow fluorescence protein (YFP) reporter (unpublished data, Lon Chubiz, Christopher Rao Lab, University of Illinois at Urbana-Champaign). The red error bars represent the standard deviation of the data obtained from two independent repeats of smFISH experiments. The black curve serves as a visual guide.

(C) The expression level of the $P_{bioBFCD}$ promoter as a function of biotin concentration obtained using smFISH, compared to the data obtained using the LacZ activity assay from the literature (Abdel-Hamid and Cronan, 2007). The red error bars represent the standard deviation of the data obtained from two independent repeats of smFISH experiments. The black curve serves as a visual guide.

(D) The expression level of the P_{RM} promoter in lysogens carrying the *cI*D38N mutation and various P_{RM} mutations (Michalowski and Little, 2005) obtained using smFISH, compared to the data obtained using the LacZ activity assay from the literature (Michalowski and Little, 2005). The vertical error bars represent the standard deviation of the data obtained from two independent repeats of smFISH experiments. The black line is a liner fit. The data for the λ_{NP6} mutant lies outside the plotted range.

[7] Specifically, the lysogens used in the comparison were MG1655(λ_{IG831}), JL2497(λ_{JL815}), JL2497(λ_{NP2}), JL2497(λ_{NP4}), JL2497(λ_{NP5}), JL2497(λ_{NP6}), JL2497(λ_{NP8}), and JL2497(λ_{NP10}) (Michalowski and Little, 2005). See **Appendix A** for details about bacterial and phage strains.

2.4 Limitations of smFISH

smFISH is a very powerful technique, allowing us to count specific target mRNA molecules at single-cell level with single-molecule sensitivity, and obtain statistics of a fixed population of cells. No genetic modification of the cells is required in order to detect the mRNA molecules of interest. The biggest shortcoming of this method is that only chemically fixed (i.e. dead) cells can be assayed, and the mRNA copy-number statistics was only obtained from a snapshot of a population of cells at a certain time. Transcription events were not directly observed, and we were not able to quantify the properties of the transcriptional time-series directly.

To get a more complete picture of the transcriptional time-series, I complemented smFISH experiments with live-cell experiments that allow us to observe mRNA production events in real-time in growing cells, which will be discussed in detail in **Chapter 5**.

2.5 Summary

In this chapter I described how to quantify the copy-number statistics of mRNA molecules in *E. coli* cells using smFISH. The analysis routines for extracting data from smFISH experiments were delineated, and tests establishing the accuracy and dynamic range of the smFISH method were presented. In the next chapter I will explain how the transcription kinetics was inferred from data obtained from smFISH experiments.

3 From mRNA Statistics to Transcription Kinetics

In the previous chapter I explained how the mRNA copy-number statistics of *E. coli* cells was obtained using smFISH. In this chapter, I will revisit the two-state model introduced in **Chapter 1**, and describe how this model allows us to connect the mRNA copy-number statistics to the stochastic transcription kinetics. I will then explain how the burstiness of the transcriptional time-series is quantified. **Section 3.2** and **Section 3.3** contain technical details on how the smFISH data was normalized to compensate for the different gene copy-number and mRNA lifetimes across different data sets. In the next chapter, I will use the tools introduced here to analyze smFISH data, and I will present the resulting findings.

3.1 The two-state model and burstiness

Predicted population statistics for the two-state model

As described in **Chapter 1** the two-state model is a simple phenomenological model for describing bursty transcriptional time-series. A gene switches stochastically between an "on" state and an "off" state, and mRNA is produced only in the "on" state (**Figure 1.5A**). The steady-state distribution of mRNA copy-number for the two-state model is given by

$$P(m) = \frac{m_s^m e^{-m_s}}{m!} \frac{\Gamma(\zeta_0+m)\Gamma(\zeta_0+\zeta_1)}{\Gamma(\zeta_0+\zeta_1+m)\Gamma(\zeta_0)} {}_1F_1(\zeta_1, \zeta_0+\zeta_1+m; m_s), \tag{3.1}$$

(Peccoud and Ycart, 1995; Raj, Peskin *et al.*, 2006; Shahrezaei and Swain, 2008), where $P(m)$ is the probability of having m mRNA molecules per cell, $m_s = k_{\mathrm{TX}}/k_{\mathrm{d}}$, $\zeta_0 = k_{\mathrm{on}}/k_{\mathrm{d}}$, $\zeta_1 = k_{\mathrm{off}}/k_{\mathrm{d}}$, $\Gamma(z)$ is the gamma function (Abramowitz and Stegun, 1972), and ${}_1F_1(a,b;z)$ is the confluent hypergeometric function of the first kind (Abramowitz and Stegun, 1972). The mean of this distribution is

$$\langle n \rangle = \frac{k_{\mathrm{on}} k_{\mathrm{TX}}}{(k_{\mathrm{on}}+k_{\mathrm{off}})k_{\mathrm{d}}}, \tag{3.2}$$

while the variance is

$$\sigma^2 = \frac{k_{\mathrm{on}} k_{\mathrm{TX}}}{(k_{\mathrm{on}}+k_{\mathrm{off}})k_{\mathrm{d}}} + \frac{k_{\mathrm{on}} k_{\mathrm{off}}}{(k_{\mathrm{on}}+k_{\mathrm{off}})^2} \frac{k_{\mathrm{TX}}^2}{k_{\mathrm{d}}(k_{\mathrm{on}}+k_{\mathrm{off}}+k_{\mathrm{d}})} \tag{3.3}$$

(Peccoud and Ycart, 1995; Sanchez and Kondev, 2008).

In the "rapid burst" regime where $k_{\mathrm{off}} \gg k_{\mathrm{d}}$, **Equation (3.1)** tends to a negative binomial distribution

$$P(m) = \frac{\Gamma(r+m)}{m!\,\Gamma(r)} p^r (1-p)^m, \tag{3.4}$$

where $r = k_{\mathrm{on}}/k_{\mathrm{d}}$ is the number of times the gene turns "on" per mRNA lifetime and $p = k_{\mathrm{off}}/(k_{\mathrm{off}} + k_{\mathrm{TX}})$ is the probability that the gene will turn "off" when it is "on". In the Poisson limit where $k_1 = 0$, i.e. the gene is always "on", **Equation (3.1)** reduces to a Poisson distribution

$$P(m) = \frac{m_s^m e^{-m_s}}{m!}. \quad (3.5)$$

The two-state model allows us to connect population statistics ($P(m)$, σ^2 , and $\langle n \rangle$) and transcription kinetics (k_{on}, k_{off}, k_{TX}, and k_d). After obtaining the mRNA statistics from smFISH experiments, I can constrain the underlying kinetics using the two-state model.

The burstiness of transcriptional time-series can be quantified by the Fano factor

The bursty kinetics of a transcriptional time-series is reflected in the cell-to-cell variability in the mRNA statistics. As demonstrated in **Figure 1.4B**, a bursty transcriptional time-series is characterized by short, intense bursts of mRNA production events when the gene is "on", separated by extended periods of quiescence when the gene is "off". Comparing that to the Poisson transcriptional time-series in **Figure 1.4A**, we can imagine that cells experiencing extended periods of quiescence would have mRNA numbers much lower than the mean, while cells experiencing bursts would have mRNA numbers much higher than the mean. The bursty transcriptional time-series would give rise to a population with a variance higher than what would arise from a Poisson transcriptional time-series (**Figure 3.1**). In general, the more bursty the transcriptional time-series, the bigger the cell-to-cell variability in mRNA copy-number in a population of cells.

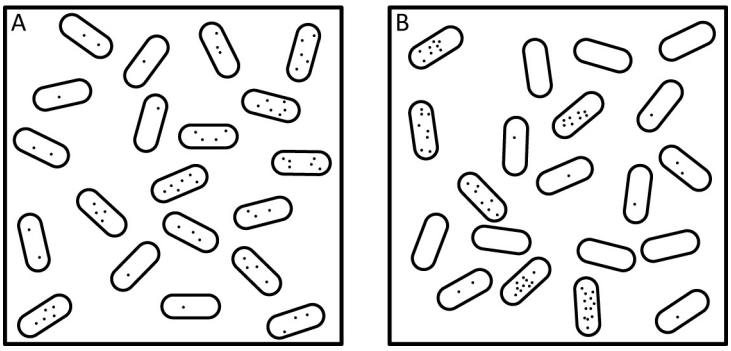


Figure 3.1. The burstiness of a transcriptional time-series is reflected in the cell-to-cell variability in mRNA numbers.

Each panel shows a population of 20 cells. Each black dot in the cells represents one mRNA molecule. The two populations have the same mean number of mRNA molecules per cell.

(A) mRNA copy-number distribution arising from a Poisson transcriptional time-series. The variance (3.15) is very close to the mean (3.10).

(B) mRNA copy-number distribution arising from a bursty transcriptional time-series. The variance (16.1) is much bigger than the mean (3.10).

To make the above observation quantitative, I use the Fano factor to quantify the burstiness of transcriptional time-series (Golding, Paulsson *et al.*, 2005). It is defined as the ratio of the variance σ^2 to the mean $\langle n \rangle$:

$$b = \frac{\sigma^2}{\langle n \rangle}. \quad (3.6)$$

b indicates how bursty a time-series is relative to a Poisson process (**Figure 3.2**) (Goh and Barabasi, 2008), with $b = 1$ corresponding to non-bursty, Poissonian mRNA production. In the "rapid burst" regime where $k_{off} \gg k_d$, b is equal to the mRNA burst size (Thattai and van Oudenaarden, 2001). For simplicity, I refer to b as the *burstiness* of the transcriptional time-series. The two-state transcription model allows us to calculate the mean and variance (**Equations (3.2)** and **(3.3)**), and therefore b, for

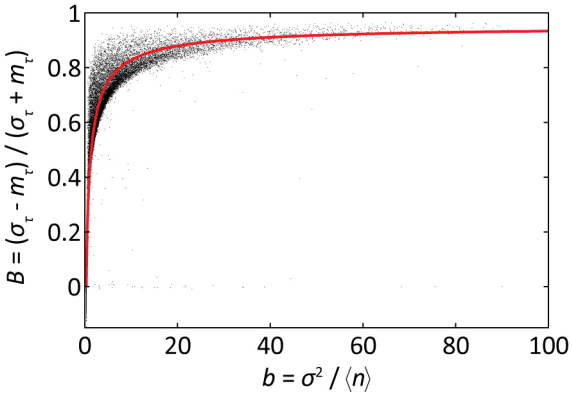


Figure 3.2. Comparing the measured burstiness B and the Fano factor b.

The measured burstiness of a simulated transcriptional time-series $B = \frac{\sigma_\tau - m_\tau}{\sigma_\tau + m_\tau}$ (Goh and Barabasi, 2008) is plotted against the Fano factor b. σ_τ and m_τ are the standard deviation and mean of the inter-event times in the simulated time series, respectively. B was calculated from simulation data. b was calculated from the kinetic parameters using **Equations (3.2)** and **(3.3)**. Each data point is based on a specific combination of transcription kinetics parameters, and about 48,000 distinct combinations are represented in this plot. The monotonically increasing relationship demonstrates that b is an appropriate measure of burstiness. The red curve is a fit to $y = \frac{A(x-1)^C}{D+(x-1)^C}$, giving $A = 0.95$, $C = 0.81$, and $D = 2.4$, $R^2 = 0.97$.

any set of kinetic parameters (Kepler and Elston, 2001; Peccoud and Ycart, 1995; Raj, Peskin *et al.*, 2006; Shahrezaei and Swain, 2008; Simpson, Cox *et al.*, 2004).

3.2 Correcting for gene copy-number

An *E. coli* cell typically has multiple copies of its genome

The two-state model only takes into account a single copy of the gene-of-interest. In reality, bacterial cells could have multiple copies of their DNA, and the copy-number of a given gene can range from one to eight, depending on growth conditions and cell age (Neidhardt, 1987). This is due to the fact that cells have to finish replicating their DNA before each cell division, and fast-growing cells would even start new rounds of replication even before the previous one is complete. Since I am interested in the kinetics of mRNA production from a single gene copy, I have to normalize my data appropriately.

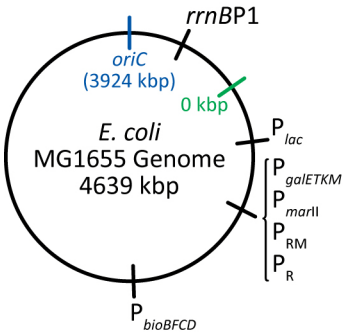
I first need to have an estimate of the population average of the copy-number of my gene-of-interest $\langle X \rangle$. To do that, I used the formula[8]

$$\langle X \rangle = 2^{[C(1-m')+D]/\tau_g}, \tag{3.7}$$

(Bremer and Churchward, 1977) where C and D are the durations of the C- and D-periods (the time needed for replicating the genome and the time between termination of replication and cell division, respectively) (Neidhardt, 1987) in minutes, τ_g is the cell doubling time in minutes[9], and m' is the normalized distance of the gene from *oriC*, which I know for each of the constructs I used (**Figure 3.3**). C

[8] This formula was derived assuming exponential increase of the number of replication origins, number of replication termini, and number of cells in an exponentially growing population of cells. With these assumptions, the number of replication termini in the culture always lags the number of replication origins by C minutes, and the number of cells in the culture always lags the number of replication termini by D minutes.

[9] Cells were grown in 1 ml of culture in a well of a 24-well plate (Corning, #3526) at 37°C in a plate reader (Tecan Infinite F200), with kinetic cycles settings: 1 mm orbital shaking for 600 seconds, then measure OD_{600} with 10 flashes after 1 second of settle time. OD_{600} as a function of time is then fitted to $\log_2 y = Ax + B$, and cell doubling time was calculated from $\tau_g = -1/A$.



Promoter	Location (kbp)	Distance from *oriC* (kbp)	*m'*
*rrnB*P1	4246	322	0.14
P_{lac}	366	1077	0.46
$P_{galETKM}$	791	1506	0.65
P_{marII}	799	1514	0.65
P_{RM}	799	1514	0.65
P_{R}	799	1514	0.65
$P_{bioBFCD}$	1576	2291	0.99

Figure 3.3. Locations of the genes studied in this work.

The locations were determined from the construction of the strains described in literature. See **Appendix A** for details of the strains and promoters, and relevant references. Coordinates are based on the *E. coli* strain MG1655 genome sequence on the National Center for Biotechnology Information website. m' is the normalized distance of the promoter, with $m' = 0$ being *oriC*, where DNA replication is initiated, and $m' = 1$ being directly opposite to *oriC*, where replication terminates.

and D as a function of τ_g are shown in **Figure 3.4** (Neidhardt, 1987). Values of $\langle X \rangle$ calculated for m' ranging from zero to one are shown in **Figure 3.5**.

Gene copy-number effect can be accounted for by scaling the mean and variance of the mRNA distribution

To quantify the effect of gene copy-number on transcription kinetics, I carried out Gillespie simulations (see **Appendix D.1** for details) (Gillespie, 1977) of the two-state model with varying gene copy-numbers, and calculated the variance and mean of the resulting mRNA number distribution. I found that, as the gene copy-number X changes, the burstiness b (calculated using **Equation (3.6)**) stays approximately constant (within 20% variation relative to the value at one copy of the gene) (**Figure 3.6A**), while the mean mRNA molecules per cell $\langle n \rangle$ scales linearly with X (**Figure 3.6B**). This result can be understood intuitively in the "rapid burst" regime. Each copy of the gene gives rise to short, rapid bursts of mRNA molecules that do not overlap in time. The net effect of the presence of multiple copies of the gene would be increasing the frequency of occurrence of bursts, but each burst still has the same burst size.

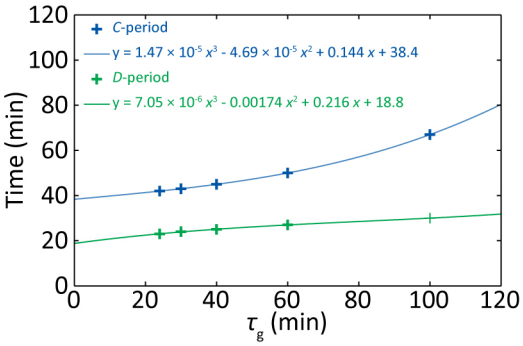


Figure 3.4. *C*- and *D*-periods as a function of generation time τ_g.

The crosses are numbers taken from the literature (Neidhardt, 1987), and the lines are fits to a cubic polynomial.

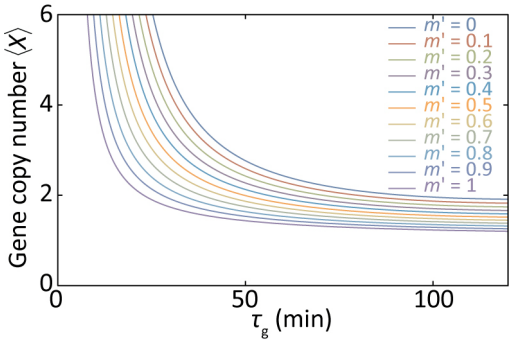


Figure 3.5. Estimating the average gene copy-number in a population.

The average gene copy-number was calculated using **Equation (3.7)**, with m' ranging from zero to one. Values of C and D in **Figure 3.4** were used.

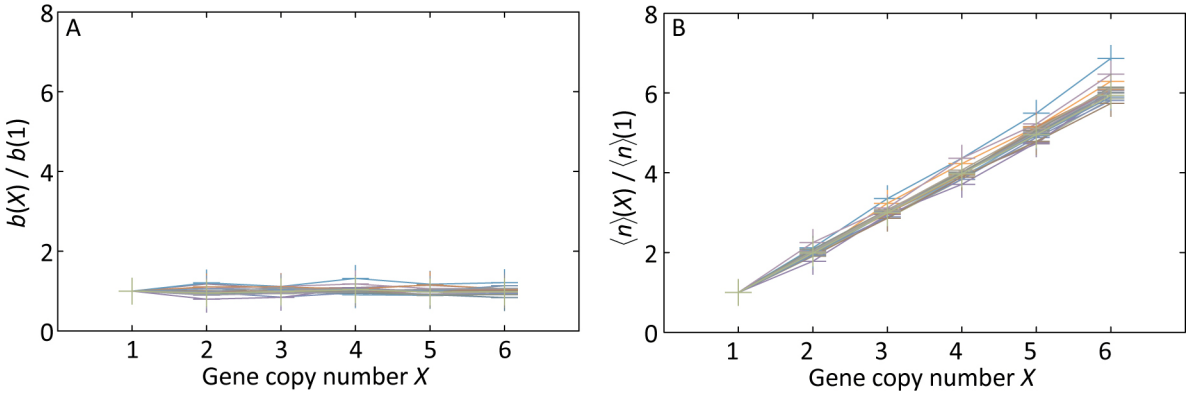


Figure 3.6. The effect of gene copy-number on the transcriptional time-series.

The effect of gene copy-number on transcription kinetics was quantified using Gillespie simulations. As the gene copy-number X was increased from one to six, (A) the burstiness b remains approximately constant, while (B) the mean number of mRNA molecules per cell $\langle n \rangle$ scales linearly with X. Each color represents a distinct combination of k_{on}, k_{off}, and k_{TX}.

Thus, in analyzing smFISH data, I scaled the mean mRNA molecules per cell by $n \to n/\langle X \rangle$ and the burstiness by $b \to b$, where $\langle X \rangle$ is the estimated average gene copy-number in the cell population. By doing so I am able to extract the single-gene behavior from my data.

3.3 Correcting for mRNA lifetime

mRNA lifetime is affected by numerous factors

Different mRNA transcripts are subject to different degradation rates. Although the exact mechanism of mRNA degradation in *E. coli* and the factors affecting the degradation rate are still largely unknown, previous studies in the literature suggest that factors affecting mRNA lifetime include growth condition, temperature, sequence of the transcript, abundance of the transcript, secondary structure of the transcript, translation efficiency of the transcript, position of the gene in the *E. coli* genome, and the function of the gene product (Arraiano, Andrade *et al.*, 2010; Bernstein, Khodursky *et al.*, 2002; Coburn and Mackie, 1999; Iost and Dreyfus, 1995; Neidhardt, 1987; Pedersen and Reeh, 1978; Selinger, Saxena *et al.*, 2003; Singh, Chang *et al.*, 2009).

In this study, however, I would like to focus on the kinetics of mRNA production, or more specifically k_{on}, k_{off}, and k_{TX} of the two-state model, but not k_{d}. Therefore I need to normalize my data properly to take away the effect of the different mRNA lifetimes across different experiments. I estimated the lifetimes of the target mRNA in different strains in different growth conditions. This was done by inhibiting transcription with rifampicin[10], and estimating the relative amount of target mRNA left using qRT-PCR at subsequent time points (see **Appendix B.3** for experimental details) (Bernstein, Khodursky *et al.*, 2002; Selinger, Saxena *et al.*, 2003). The results are shown in **Figure 3.7** and **Table 3.1**.

[10] Rifampicin is an antibiotic that halts transcription in bacterial cells by inhibiting RNA polymerase.

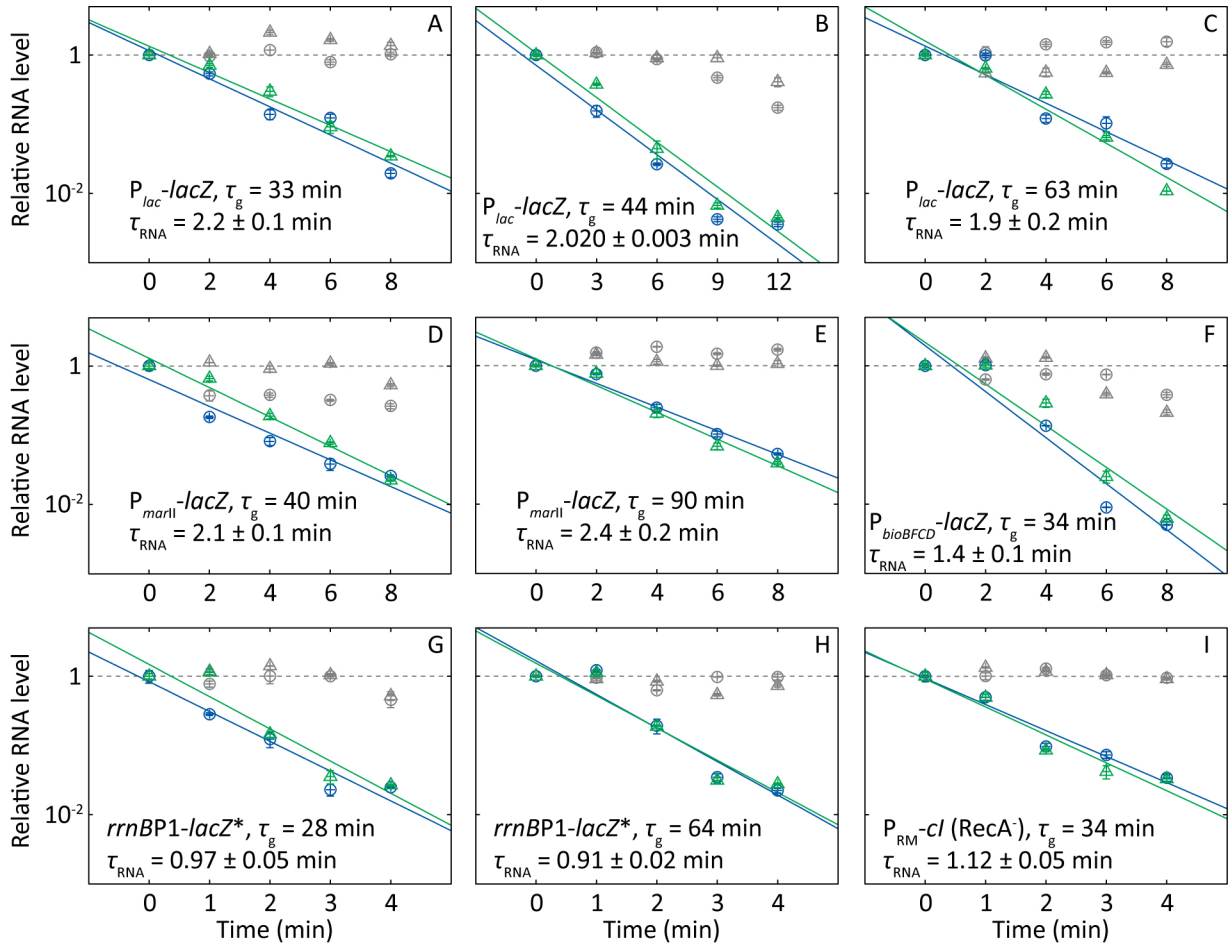


Figure 3.7. mRNA lifetime measurements.

The plots depict the relative mRNA level, measured using qRT-PCR, as a function of time after adding rifampicin, a transcription inhibitor. Each panel corresponds to a specific transcript under a specific growth condition (τ_g is the cell doubling time). See **Table 3.1**, **Appendix A**, and **Appendix B.3** for further details. Two independent repeats were performed in each case (green triangles and blue circles), as well as negative controls with no rifampicin added (grey triangles and circles). The green and blue lines are fits to $\ln y = Ax + B$, and mRNA lifetime was calculated from $\tau_{RNA} = -1/A$. The values shown are the mean ± standard deviation from the two independent repeats.

The majority of the promoters examined in this study (P_{lac}, $P_{galETKM}$, P_{marII}, and $P_{bioBFCD}$; see **Table 3.1** and **Table A.1**) expressed the full *lacZ* gene as a reporter[11] (**Table A.1**). Thus, although the parameters of mRNA production varied largely between promoters, the actual transcripts produced were identical, and were subject to similar post-transcriptional regulation, including mRNA degradation (Bernstein, Khodursky *et al.*, 2002). To verify that the mRNA degradation rate was determined by the transcript rather than the promoter, I measured the mRNA lifetime of *lacZ* expressed from three promoters (P_{lac}, P_{marII}, and $P_{bioBFCD}$). The values found showed a standard deviation of ~23%, significantly less than the differences in mRNA lifetime between different transcripts (**Table 3.1**). I also verified that

[11] P_{lac} expressed the endogenous *lacZ* gene. $P_{galETKM}$, P_{marII}, and $P_{bioBFCD}$ had the *lacZ* gene genetically engineered downstream of them, and the promoters expressed the *lacZ* gene when active (**Table A.1**).

the differences in growth rate did not have a strong effect on mRNA lifetime (**Table 3.1**) (~9% standard deviation when comparing growth at doubling times of 40 minutes versus 90 minutes).

In addition to *lacZ*, some strains used in this study expressed other transcripts (27 strains expressing variants of *cI* and *cro* genes; one strain expressing a modified version of the *lacZ* transcript in which the ribosome binding site has been mutated; see **Table 3.1**, **Table A.1**, and **Table A.3**). I measured the mRNA lifetimes for these reporters as well (**Figure 3.7** and **Table 3.1**). As expected, different transcripts exhibited different degradation rates, with more than a three-fold range of mRNA lifetimes (from ~0.9 minute to ~2.8 minutes; see **Table 3.1**).

Table 3.1. The mRNA lifetimes of transcripts used in this work.

No.[12]	Strain	Promoter	Transcript[13]	Expression level (mRNA/cell)	Doubling time (min)	RNA lifetime[14] (min)	Source
E1	TK310	P_{lac}	*lacZ*	~4.9	~33	2.2 ± 0.1	This work
E2	TK310	P_{lac}	*lacZ*	~55	~44	2.020 ± 0.003	This work
E3	TK310	P_{lac}	*lacZ*	~0.59	~63	1.9 ± 0.2	This work
E4	LC544	P_{marI}	*lacZ*	~4.1	~40	2.1 ± 0.1	This work
E5	LC544	P_{marI}	*lacZ*	~6.4	~90	2.4 ± 0.2	This work
E6	CY481	$P_{bioBFCD}$	*lacZ*	~1.4	~34	1.4 ± 0.1	This work
E7	CF7753	*rrnB*P1	*lacZ**	~22	~28	0.97 ± 0.05	This work
E8	CF7753	*rrnB*P1	*lacZ**	~1.9	~64	0.91 ± 0.02	This work
E9	NC416	P_{RM}	*cI*	~10	~34	2.8 ± 0.8	Zong, So *et al.*, 2010
E10	JL5902 (λ_{NP2})	P_{RM} (*cI*D38N)	*cI*	~6.9	~34	1.12 ± 0.05	This work
E11	NC416	P_R	*cro*	~1.1	~34	1.9 ± 0.5	Zong, So *et al.*, 2010

[12] Comments on the experiments:

a. E1 through E3; E4 and E5; and E7 and E8:
Comparing mRNA lifetimes for the same transcript at different growth rates.

b. E1 through E3; E4 and E5; and E7 and E8:
Comparing mRNA lifetimes for the same transcript at different expression levels.

c. E1, E4, and E6:
Comparing mRNA lifetimes for the same transcript expressed from different promoters.

d. E1, E7, E9, and E11:
Comparing mRNA lifetimes of different transcripts.

e. E9 and E10:
Comparing mRNA lifetimes for the same transcript in $RecA^+$ and $RecA^-$ backgrounds.

[13] *lacZ** has an inefficient ribosome binding site to avoid toxicity from excessive LacZ levels (Hernandez and Bremer, 1990; Potrykus, Vinella *et al.*, 2006) (Michael Cashel, personal communication).

[14] The values of mRNA lifetime shown here are the mean ± standard deviation from two independent experiments.

Differences in mRNA lifetimes can be accounted for by scaling the mean and variance of the mRNA distribution

Once the mRNA lifetimes (and hence the degradation rates k_d) were known, I was able to correct for the differences in k_d between different data sets, such that my examination of mRNA kinetics will only reflect differences on the production side. This normalization was done by multiplying the mean expression level $\langle n \rangle$ by the degradation rate k_d relative to the value for *lacZ* (chosen as an arbitrary standard). This calibration procedure allowed the comparison of data sets with different k_d's. It is justified by the scaling properties of the two-state model, in which the mean expression level is proportional to the mRNA lifetime, while the burstiness exhibits only a weak dependence on k_d (see **Equations (3.2)** and **(3.3)**). I verified numerically that the normalization procedure above is justified for the range of parameters observed in my experiments (**Figure 3.8**).

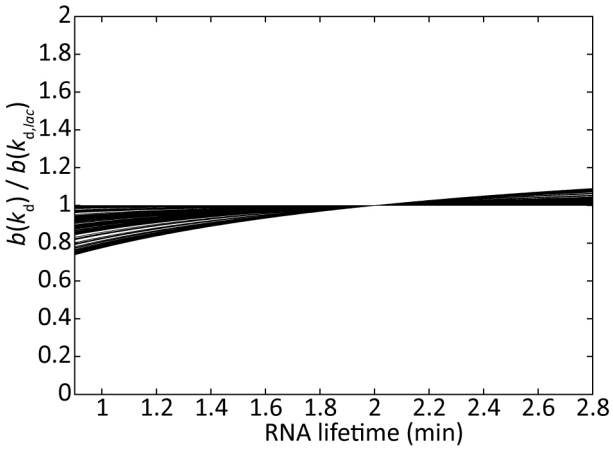


Figure 3.8. The effect of differences in mRNA lifetime on the degree of burstiness.

The burstiness, b, was calculated for a range of mRNA lifetimes and transcription parameters corresponding to the data in this study. Specifically, mRNA lifetimes cover the range found in **Table 3.1**; the values of k_{on}, k_{off}, and k_{TX} range from ten-fold below to ten-fold above typical values in our data set (see **Figure 4.6B**). b was calculated from the analytical expressions (**Equations (3.2)** and **(3.3)**) and was normalized by the value obtained for *lacZ* mRNA lifetime (**Table 3.1**). b does not deviate by more than ~20% from the value for *lacZ*.

3.4 Summary

The two-state model allows us to connect mRNA copy-number statistics to stochastic transcription kinetics, and the burstiness of a transcriptional time-series can be quantified by the Fano factor. The data obtained from smFISH experiments was properly normalized to compensate for the different gene copy-number and mRNA lifetimes across different data sets. In the next chapter, I will use the tools introduced here to analyze smFISH data, and I will present the resulting findings.

4 Gene-Independent Behavior of Transcription Burstiness in *E. coli*

In the previous chapter, we saw how the transcription burstiness can be quantified from the mRNA copy-number statistics measured using smFISH. Here, results from my experiments will be presented. I measured the burstiness of multiple *E. coli* promoters under various growth conditions. The burstiness of the lactose promoter showed the same dependence on the expression level, when the induction is achieved by activation or de-repression. When I extended the analysis to multiple promoters with different molecular mechanisms of regulation, the observed burstiness still followed the same trend. This gene-independent behavior is contrary to the common theoretical model ascribing bursty transcriptional time-series to the kinetics of binding and unbinding of transcription factors at the promoter. Instead, this behavior suggests that transcription bursts arise from gene-nonspecific mechanisms.

4.1 The degree of burstiness depends on the mean expression level alone

The burstiness of the lactose promoter depends only on the mean expression level, irrespective of whether expression is induced by de-repressor or activator

I first used smFISH to quantify the mRNA copy-number statistics of the *E. coli* lactose promoter P_{lac} (**Table A.1** and **Table A.3**). I scanned a range of expression levels by growing the strain TK310 (**Table A.1**) (Kuhlman, Zhang et al., 2007) with various amount of β-d-1-thiogalactopyranoside (IPTG) and adenosine 3',5'-cyclic monophosphate (cAMP) (see **Appendix A.1** for details). IPTG induces P_{lac} by de-repression. It deactivates the Lac repressor, a transcription factor that represses the expression from P_{lac} (Müller-Hill, 1996). cAMP, on the other hand, induces P_{lac} by activation. It activates the cAMP-receptor-protein (CRP). Activated CRP is a transcription factor that promotes the expression from P_{lac} by recruiting RNA polymerase (Müller-Hill, 1996). I plotted the burstiness b as a function of the mean expression level $\langle n \rangle$ (corrected for gene copy-number and RNA lifetime, see **Chapter 3**) for a range of expression levels (**Figure 4.1A**). Induction was achieved by adding various amounts of IPTG or various amounts of cAMP.

Bursty transcriptional time-series has been proposed to arise from the kinetics of binding and unbinding of transcription factors at the promoter, and regulation by different transcription factors acting with different molecular mechanisms would give rise to different burstiness characteristics (**Figure 4.2**) (Cox, McCollum *et al.*, 2008; Garcia, Sanchez *et al.*, 2010; Sanchez, Garcia *et al.*, 2011). One would thus expect the two different regulation mechanisms of P_{lac} to give distinct burstiness behaviors. Surprisingly, all data points lie on the same curve, with the burstiness depending on the mean expression level alone,

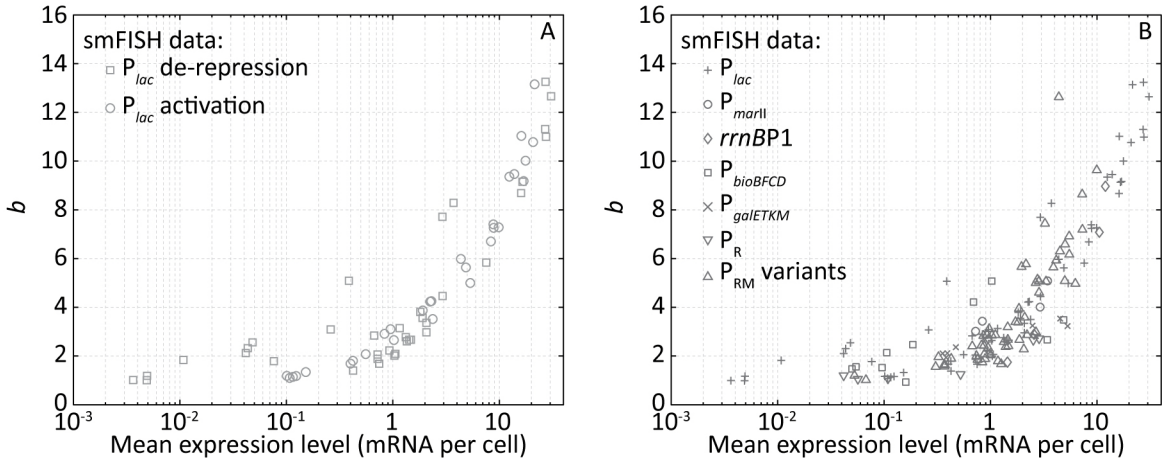


Figure 4.1. Gene-independent behavior of the burstiness of transcriptional time-series.

(A) The burstiness of the lactose promoter P_{lac} as a function of the mean expression level. The dependence of the burstiness on the mean expression level is the same when induction of P_{lac} is achieved by de-repression by IPTG or activation by cAMP.

(B) The burstiness of 20 promoters under more than 50 growth conditions as a function of the mean expression level. The dependence of the burstiness on the mean expression level is the same for all the promoters under all growth conditions.

irrespective of the mode of regulation. This suggests the possibility that the burstiness of transcriptional time-series might arise from factors other than the molecular details of regulation of a specific promoter. To further examine this possibility, I extended the analysis to a variety of promoters that have very different regulation mechanisms.

The burstiness of multiple promoters of *E. coli* shows the same dependence on the mean expression level

In addition to the lactose promoter P_{lac}, I used smFISH to quantify mRNA statistics from 19 other *E. coli* promoters: the galactose promoter $P_{galETKM}$ (Tokeson, Garges *et al.*, 1991; Weickert and Adhya, 1993), the multiple antibiotic resistance promoter P_{marII} (Alekshun and Levy, 1999), the ribosomal RNA promoter *rrnB*P1 (Hernandez and Bremer, 1990; Potrykus, Vinella *et al.*, 2006), the biotin promoter $P_{bioBFCD}$ (Abdel-Hamid and Cronan, 2007; Barker and Campbell, 1980), the bacteriophage λ promoter P_R, and 13 variants of the bacteriophage λ promoter P_{RM} (Lim and Sauer, 1989;

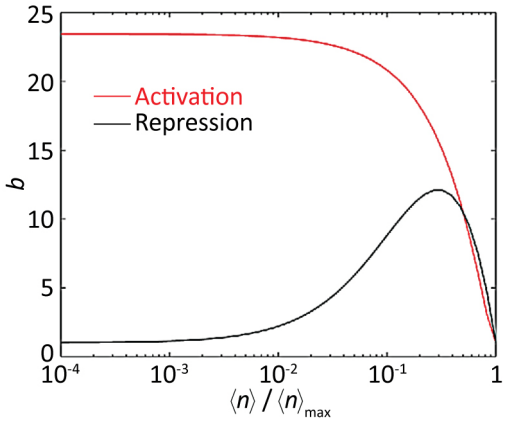


Figure 4.2. Theoretical prediction of the burstiness due to binding kinetics of transcription factor as a function of the expression level.

The burstiness b is plotted against the normalized expression level $\langle n\rangle/\langle n\rangle_{max}$, calculated based on a theoretical model in which the burstiness arises from binding and unbinding of transcription factors (Sanchez, Garcia *et al.*, 2011). The red line represents simple activation and the black line represents simple repression, both having the same transcription factor affinity. They show distinctly different dependences of the burstiness on the expression level. Figure adapted from Figure 5B of Sanchez, Garcia *et al.*, 2011.

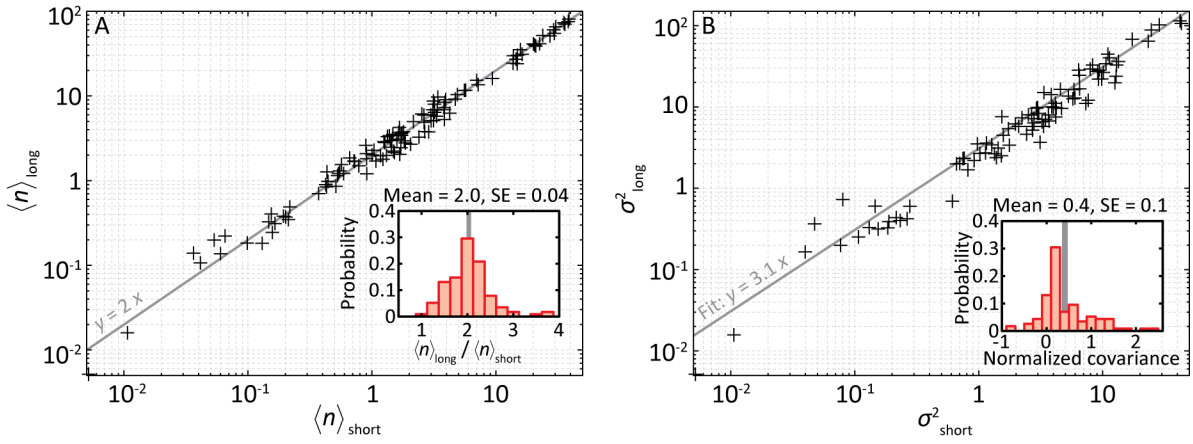


Figure 4.3. Positive covariance between the activities of multiple gene copies in the cell.

(A) mRNA statistics of cells having different gene copy-numbers can be obtained by examining cells at the beginning and end of the life cycle, as indicated by the cell length. The mean number of mRNA per cell from the 20% longest cells $\langle n \rangle_{\text{long}}$ is plotted against that of the 20% shortest cells $\langle n \rangle_{\text{short}}$. The data (black crosses) is from 115 independent experiments with different promoters and expression levels. The points fall on the $y = 2x$ line (grey line), consistent with the expectation that cells at the end of the life cycle will have twice the gene copies of newly divided cells (Neidhardt, 1987). The inset shows a histogram of $\frac{\langle n \rangle_{\text{long}}}{\langle n \rangle_{\text{short}}}$, with a mean 2.0 and standard error 0.04 (shaded region).

(B) The variance of mRNA number per cell from the 20% longest cells σ^2_{long} is plotted against the variance of the 20% shortest cells σ^2_{short}. Fitting to $y = Ax$ gives $A = 3.1 \pm 0.1$ (grey line). The value obtained is bigger than two, which would be expected if the activities of individual gene copies were independent (Paulsson, 2004). In other words, a positive covariance between identical copies of the same gene was found, consistent with the idea that idetical copies tend to turn "on" and "off" together. The inset shows a histogram of the normalized covariance (Paulsson, 2004), defined as $\frac{\sigma^2_{\text{long}} - 2\sigma^2_{\text{short}}}{2\langle n \rangle^2_{\text{short}}}$. The normalized covariance has a mean 0.4 and standard error 0.1 (shaded region).

Michalowski and Little, 2005; Sauer, Jordan *et al.*, 1990) (**Table A.1** and **Table A.3**). In cases where promoter activity is regulated by growth conditions (e.g. the presence of a specific sugar or amino acid), a range of growth conditions was used so that the full range of mRNA levels could be achieved (see **Appendix A.1**). Together with P_{lac}, this ensemble of promoters allowed us to scan almost three orders of magnitude of expression levels (~0.01–60 mRNA molecules per cell), different molecular mechanisms of transcription regulation (activation, repression and combinations thereof), and topologies of gene networks controlling gene activity, such as the presence or absence of feedback (Shen-Orr, Milo *et al.*, 2002) (**Table A.3**). All of these factors have been suggested to affect the observed fluctuations in gene activity (Cox, McCollum *et al.*, 2008; Garcia, Sanchez *et al.*, 2010; Kaern, Elston *et al.*, 2005; Kittisopikul and Suel, 2010; Pedraza and van Oudenaarden, 2005; Sanchez, Garcia *et al.*, 2011). In total, more than 150 independent experiments were performed, each one yielding the mRNA copy-number distribution from a given gene at a given stimulus level.

I plotted the burstiness b as a function of the mean expression level $\langle n \rangle$ (**Figure 4.1B**) (corrected for the gene copy-number and RNA lifetime, see **Chapter 3**). Strikingly, b exhibited a gene-independent behavior; that is, the values from different genes and growth conditions show a clear trend, with a

dependence on the expression level $\langle n \rangle$ alone. The properties of the time-series seem to depend primarily on the mean mRNA level, not on the specific gene or stimulus[15].

Activity of identical copies of the same gene in a cell exhibit a positive covariance

Even though the burstiness of multiple promoters in *E. coli* shows the same dependence on the expression level, this similarity in the time-series characteristics does not necessarily mean that the actual activity of the different genes is coordinated in time, i.e. that genes turn "on" and "off" in unison. It is interesting to note, however, that multiple copies of the same gene, present when the bacterial chromosome replicates, exhibit a positive, non-zero covariance (**Figure 4.3**), suggesting that their temporal activity may indeed be correlated.

Cell-to-cell variability is dominated by intrinsic noise

There are two sources of cell-to-cell variability in mRNA numbers: the inherent fluctuations of the two-state process ("intrinsic noise") and fluctuations in other cellular components ("extrinsic noise") (Elowitz, Levine *et al.*, 2002; Swain, Elowitz *et al.*, 2002). Based on our findings, the cell-to-cell variability in mRNA numbers is dominated by intrinsic noise. This is suggested by the following observations:

(i) The noise, defined as the squared coefficient of variation $\eta^2 = \sigma^2/\langle n \rangle^2$, decreases monotonically with $\langle n \rangle$ (**Figure 4.4**), which is the typical behavior of intrinsic noise, but not extrinsic noise (Elowitz, Levine *et al.*, 2002).

(ii) In the limit of low $\langle n \rangle$, $b \approx 1$ (**Figure 4.1B**), as expected for the intrinsic noise of a Poisson process. Transcription has been shown previously to be Poissonian at very low expression level (Cai, Friedman *et al.*, 2006; Yu, Xiao *et al.*, 2006).

(iii) In the limit of high $\langle n \rangle$, η^2 decreases sharply rather than approaches a plateau (**Figure 4.4**). Such a plateau would be expected in the presence of

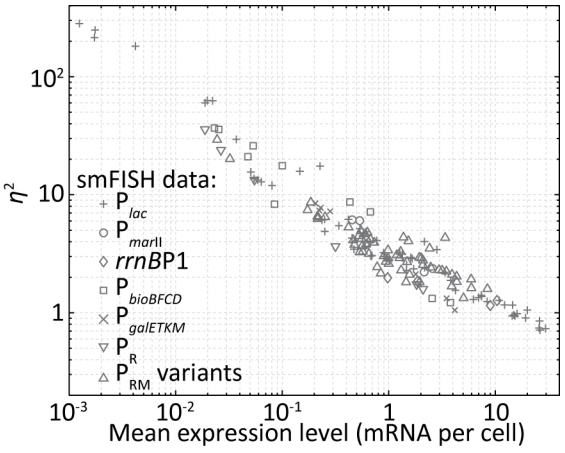


Figure 4.4. Noise as a function of mean expression level.

The data set shown in **Figure 4.1** is plotted as the noise $\eta^2 = \frac{\sigma^2}{\langle n \rangle^2}$ as a function of the mean expression level $\langle n \rangle$. η^2 decreases monotonically with $\langle n \rangle$, typical of intrinsic noise.

[15] The only instance in which a different behavior is suggested is the bacteriophage T7 promoter P_{T7} in strain MC1061(λDE3) with the P_{T7}-*lacZ*(RBS701) reporter (Proux and Dreyfus, 2008). However this observation needs to be taken with caution. The T7 RNA polymerase (RNAP) gene is under the control of the IPTG-inducible promoter P_{lacUV5}, and there is a small piece of *lacZ* leftover downstream of the promoter, which spans about one-sixth of the complete smFISH probe set. When transcribed this leftover piece would contribute to extraneous smFISH signal. Moreover, the artificially constructed genetic cascade (consisting of IPTG inducing the expression of T7 RNAP and then T7 RNAP transcribing the *lacZ* reporter) and the resulting decoupled transcription and translation (Iost and Dreyfus, 1995) might give rise to artifacts not characteristic of the natural behavior of *E. coli*.

extrinsic noise (Elowitz, Levine *et al.*, 2002; Taniguchi, Choi *et al.*, 2010). The observed dominance of intrinsic noise in mRNA number fluctuations is consistent with previous observations, that extrinsic noise is an important factor at the level of the protein species (Elowitz, Levine *et al.*, 2002; Taniguchi, Choi *et al.*, 2010), but not mRNA (Golding, Paulsson *et al.*, 2005).

4.2 Gene expression level is changed by modulating the gene off-rate

The dependence of the burstiness on the expression level in different modulation schemes is different

One of the questions raised in **Chapter 1** is that what modulation scheme is used by the cell to change the expression levels of genes. As I showed in **Figure 1.6B**, each of the modulation schemes creates a transcriptional time-series of different characteristics at a given gene expression level. Even though these different time-series produce the same mean level of mRNA, the different characteristics of the time-series are in turn reflected in the degree of cell-to-cell variability in mRNA numbers. Each of the modulation schemes described above yields a typical curve for b as a function of the mean mRNA level $\langle n \rangle$ (**Figure 4.5**). These curves are distinct from each other; thus, measuring $b(\langle n \rangle)$ experimentally using smFISH would in principle allow us to discriminate among the different scenarios and identify which kinetic parameter of the transcriptional time-series is varied. Similar analysis can be performed on the noise in the time-series, quantified by the squared coefficient of variation $\eta^2 = \frac{\sigma^2}{\langle n \rangle^2}$ (Bar-Even, Paulsson *et al.*, 2006).

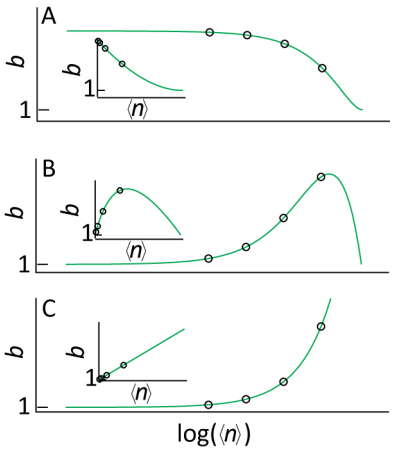


Figure 4.5. The effect of the different modulation schemes on burstiness.

The Fano factor b, which quantifies the burstiness of the transcriptional time-series, is plotted as a function of the mean expression level $\langle n \rangle$ for the (A) k_{on}; (B) k_{off}; and (C) k_{TX} modulation schemes. The main panels show $b(\langle n \rangle)$ on a semi-log scale, while the insets show the same data on a linear scale. The three modulation schemes lead to distinct functional forms of $b(\langle n \rangle)$. $b(\langle n \rangle)$ was calculated using **Equations (3.2)**, **(3.3)**, and **(3.6)**, with the same parameters used in **Figure 1.6**. The circles correspond to the four expression levels explicitly shown in **Figure 1.6**.

Gene off-rate modulation scheme is consistent with the observed behavior

To find out which modulation scheme is used by the cell, I compared the experimental plot in **Figure 4.1B** to the theoretical curves in **Figure 4.5**. The observed mRNA statistics is consistent with the assumption that expression level is changed by varying the rate at which the gene switches back to the "off" state (off-rate k_{off}), or in other words the duration of transcription bursts. Specifically, note

that $b(\langle n\rangle)$ starts with a Poisson-like behavior ($b\sim1$) when $\langle n\rangle$ is small and then increases as a sub-linear function of $\langle n\rangle$.

This observation can be made quantitative by fitting the experimental data for $b(\langle n\rangle)$ to the analytical expressions for the two-state model (**Section 3.1**) (Kepler and Elston, 2001; Peccoud and Ycart, 1995; Raj, Peskin *et al.*, 2006; Shahrezaei and Swain, 2008; Simpson, Cox *et al.*, 2004). Specifically, if I assume k_{on} is modulated, I get (by eliminating k_{on} from **Equations (3.2)** and **(3.3)**)

$$\frac{\sigma^2}{\langle n\rangle} = 1 + \frac{(k_{TX}-k_d\langle n\rangle)^2}{k_{TX}k_{off}+k_d(k_{TX}-k_d\langle n\rangle)}. \quad (4.1)$$

Similarly, if I assume that k_{off} is modulated, I get

$$\frac{\sigma^2}{\langle n\rangle} = 1 + \frac{k_d(k_{TX}-k_d\langle n\rangle)\langle n\rangle}{k_{on}k_{TX}+{k_d}^2\langle n\rangle}, \quad (4.2)$$

and if I assume k_{TX} is modulated, I get

$$\frac{\sigma^2}{\langle n\rangle} = 1 + \frac{k_d k_{off}\langle n\rangle}{k_{on}(k_{on}+k_{off}+k_d)}. \quad (4.3)$$

If I assume k_{off} is modulated (**Equation (4.2)**), a good fit is obtained ($R^2 = 0.81$), as shown in **Figure 4.6A**. For comparison, trying to fit the observed data with the two alternative scenarios, modulating the gene on-rate k_{on} (**Equation (4.1)**) or the transcription rate k_{TX} (**Equation (4.3)**), yielded inferior fits ($R^2 = -7.9\times10^{-6}$ and 0.58, respectively). Moreover, the scenario of varying k_{off} yields a fit superior to the alternatives when compared on a promoter-by-promoter basis (**Figure 4.7** and **Figure 4.8**). As a control, trying to fit a simulated collection of promoters with randomly selected kinetic parameters using the k_{off}-modulation description also yielded a poor fit ($R^2 = 0.085$, **Figure 4.9**). As an additional test for

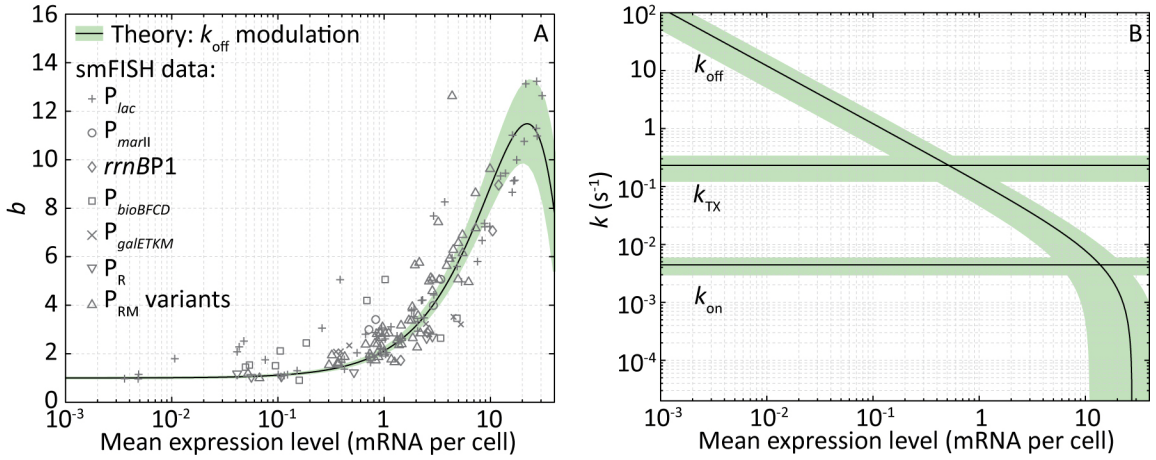


Figure 4.6. The gene expression level in *E. coli* is varied by changing the gene off-rate.

(A) The data points here represent the same data set as in **Figure 4.1B**. The black curve is a fit of the analytic expression of the k_{off}-modulation scheme (**Equation (4.2)**) to the smFISH data, using k_{on} and k_{TX} as fitting parameters. The green shade represents the 95% confidence interval of the fit.

(B) The estimated rate parameters for gene activity in *E. coli*. These were obtained from fitting $b(\langle n\rangle)$ in panel (A) to the case of varying k_{off} in the two-state model. When changing the expression level $\langle n\rangle$ for over three orders of magnitude, the rates of switching "on" (k_{on}) and of transcription in the "on" state (k_{TX}) remain approximately constant, while the rate of switching off (k_{off}) varies over four orders of magnitude. The errors in k_{on} and k_{TX} (green shade) are based on the variability in estimates between individual promoters (**Figure 4.7**). The error in k_{off} (green shade) is calculated from the resulting fit.

the validity of my parameter estimation, I performed Gillespie simulations (see **Appendix D.1** for details) (Gillespie, 1977) of mRNA kinetics and verified that the theoretical and experimental copy-number histograms are in agreement, beyond the mere values of $\langle n \rangle$ and σ^2 (**Figure 4.10**). The theoretical fit allows us to make the observation of gene-independence more quantitative: When comparing the data from individual promoters to the universal fit, I found that the average deviation of a single-promoter data from the universal fit is ~33% (**Figure 4.11A**). Six of seven data sets exhibit a correlation coefficient above 0.85 between the data and the theoretical prediction (**Figure 4.11B**).

Fitting the experimental data to the scenario of k_{off}-modulation allowed us to estimate the values of the three kinetic parameters governing mRNA production: k_{on}, the rate of switching to the "on" state, which determines the frequency of bursts, k_{TX}, the rate of producing mRNA while the gene is "on", and k_{off}, the rate of switching back to the "off" state, which determines the duration of bursts. k_{on} and k_{TX} are approximately constant for different genes and expression levels, while k_{off} changes over more than three orders of magnitude when expression level is varied (**Figure 4.6B**). Note that, of these three parameters, the only one which has been estimated in the past is k_{TX}, which corresponds to the maximal transcription initiation rate possible (when a gene is constantly "on"). The value obtained from my single-cell measurements ($k_{\text{TX}} = 0.23 \pm 0.11\ \text{s}^{-1}$) is in good agreement with values from the literature (Kennell and Riezman, 1977; Liang, Bipatnath *et al.*, 1999; Neidhardt, 1987). Interestingly, I also noted a dependence of k_{on} and k_{TX} on the bacterial growth rate (**Figure 4.8**).

4.3 Gene-independent behavior suggests transcription bursts arise from global regulation mechanisms

At this stage, there is no mechanistic, molecular-level understanding of what gives rise to the bursty behavior of gene activity in bacteria; specifically, what the physiological nature of the gene "on" and "off" states is, and how the rates of switching between states can be varied in the individual cell or over the time course of evolution. However, the gene-independent behavior I observed immediately suggests that the rate parameters in the two-state picture are not determined by the details of molecular regulation of an individual promoter, such as the binding and unbinding kinetics of a specific transcription factor, or the topology of the individual gene network, such as the presence or absence of feedback, as commonly proposed in literature (Garcia, Sanchez *et al.*, 2010; Golding, Paulsson *et al.*, 2005; Kepler and Elston, 2001; Mitarai, Dodd *et al.*, 2008; Sanchez, Garcia *et al.*, 2011; Simpson, Cox *et al.*, 2004). Instead, my finding here, that the properties of the transcriptional time-series are gene-independent rather than gene-specific, suggests that gene on/off switching is dominated by a process that acts in a similar manner on different genes, possibly exerting its influence at a genome-wide level, such as DNA topology dynamics, RNA polymerase dynamics, or regulation by broad-target DNA-binding proteins (Mitarai, Dodd *et al.*, 2008; Mooney, Davis *et al.*, 2009; Reppas, Wade *et al.*, 2006). Thus, all genes expressed at a given level exhibit a similar transcriptional time-series. Interestingly, these types of mechanisms are reminiscent of

those suggested to underlie non-Poissonian transcription kinetics in eukaryotes, where the burstiness is broadly ascribed to "chromatin modifications" (Blake, M *et al.*, 2003; Raj, Peskin *et al.*, 2006; Whitelaw, Chong *et al.*, 2010).

4.4 Summary

In this chapter I described the gene-independent behavior of burstiness in *E. coli*. The burstiness of 20 promoters in a range of growth conditions follows the same dependence on the mean expression level. This behavior is best explained by the gene off-rate modulation scheme, in which the expression level is changed by varying the duration of transcription bursts. This observation rules out theoretical models attributing transcription bursts to specific details of regulation of an individual promoter, and suggests that transcription bursts arise from global mechanisms common to all promoters. In the next chapter, I will present results from live-cell experiments, which provides additional evidence for the gene off-rate modulation picture.

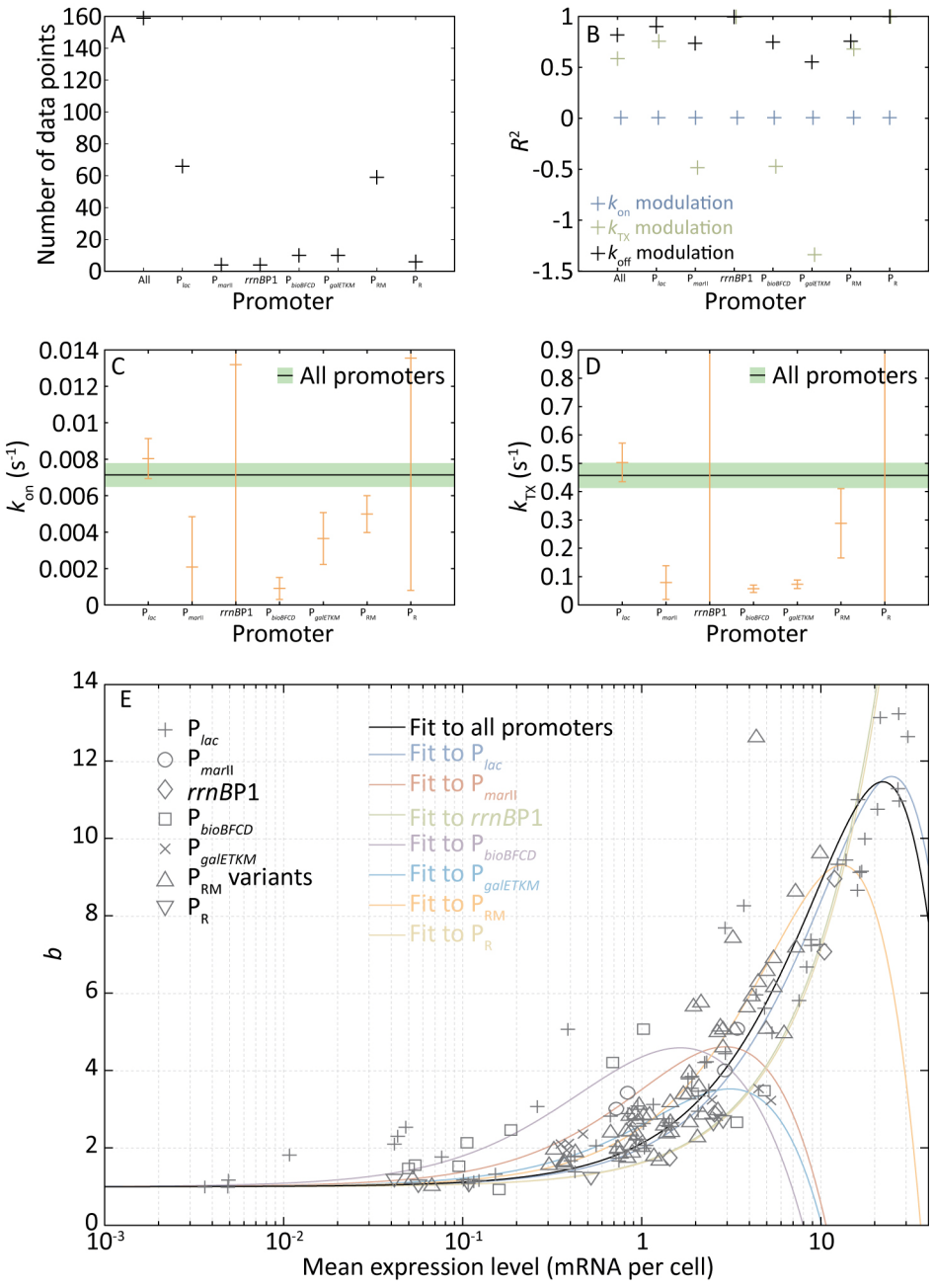


Figure 4.7. Fitting the modulation schemes to smFISH data from individual promoters.

(A) The number of smFISH data points obtained from individual promoters.

(B) The R^2 of fits yielded from fitting the three modulation schemes to smFISH data obtained from individual promoters. The R^2 of fitting the k_{off}-modulation scheme is consistently higher than the R^2 for the other two schemes. The actual fits of the k_{off}-modulation scheme are shown in panel (E).

(C and D) k_{on} and k_{TX} values obtained from fitting the k_{off}-modulation scheme to smFISH data obtained from individual promoters. The black line is the value obtained from a fit to all data points, with the 95% confidence interval shown in green shading. The actual fits of the k_{off}-modulation scheme are shown in panel (E).

(E) Fits of the k_{off}-modulation scheme to the smFISH data obtained from individual promoters. The black line is a fit to all data points.

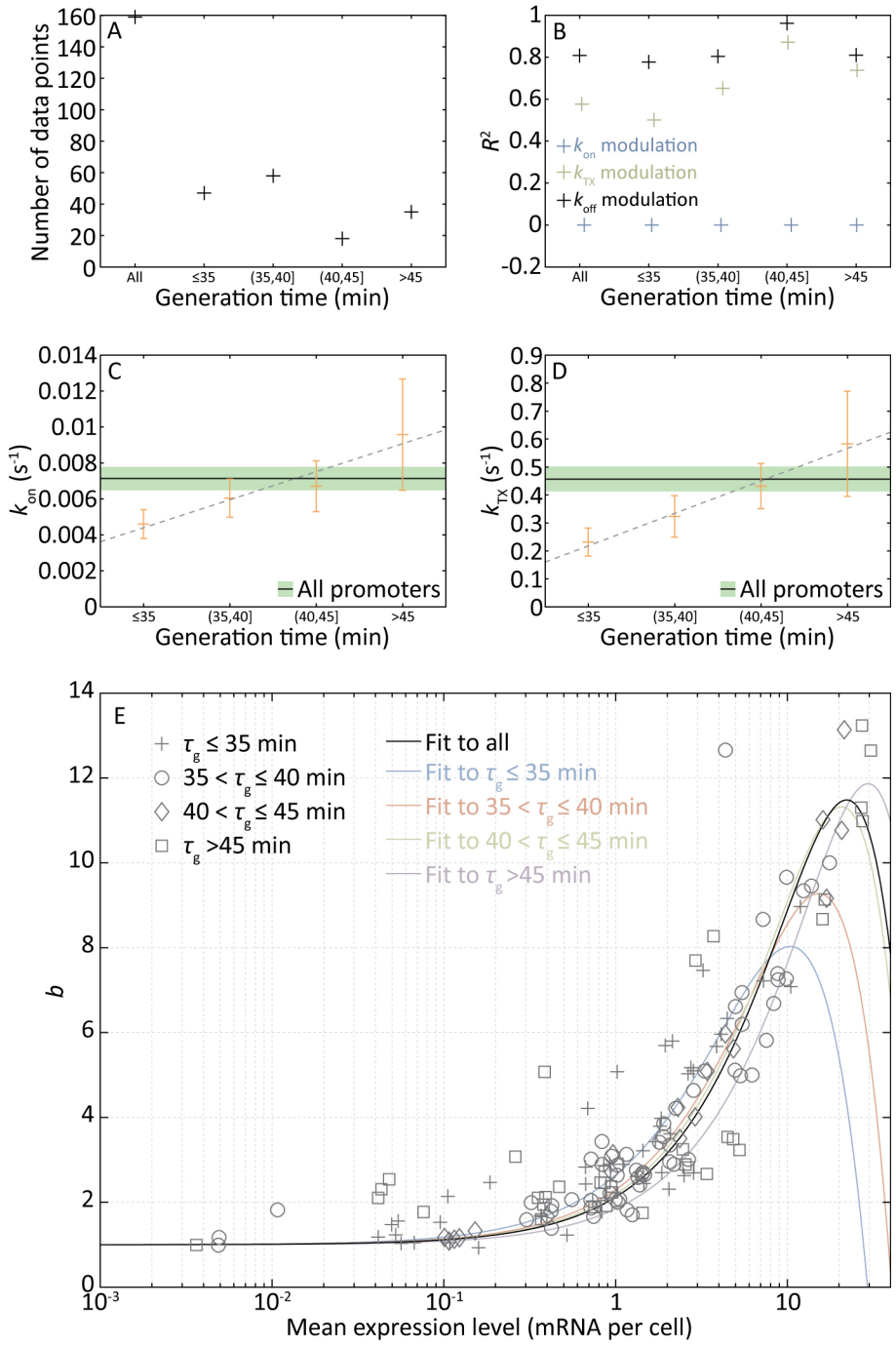


Figure 4.8. Fitting the modulation schemes to subsets of smFISH data separated by generation time.

(A) The number of smFISH data points in each subset.

(B) The R^2 of fits yielded from fitting the three modulation schemes to subsets of the smFISH data separated by generation time. The R^2 of fitting the k_{off}-modulation scheme is consistently higher than the R^2 for the other two schemes. The actual fits of the k_{off}-modulation scheme are shown in panel (E).

(C and D) k_{on} and k_{TX} values obtained from fitting the k_{off}-modulation scheme to subsets of smFISH data separated by generation time. There is an obvious trend of k_{on} increasing with generation time. Grey line is a linear fit. The black line is the value obtained from a fit to all data points, with the 95% confidence interval shown in green shading. The actual fits of the k_{off}-modulation scheme are shown in panel (E).

(E) Fits of the k_{off}-modulation scheme to subsets of the smFISH data separated by generation time. The black line is a fit to all data points.

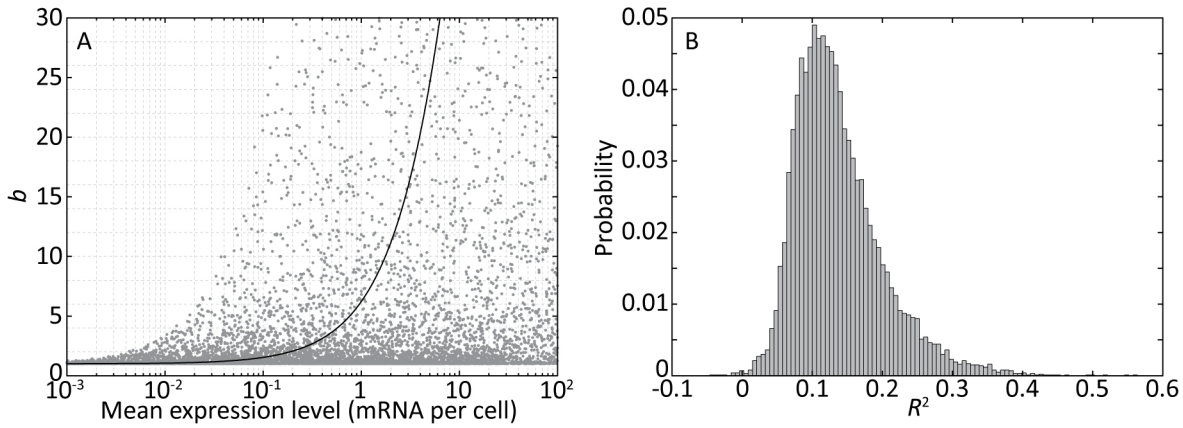


Figure 4.9. k_{off}-modulation fails to describe the behavior of randomly generated promoters.

(A) The grey markers are b and $\langle n \rangle$ calculated from randomly generated k_{on}, k_{off}, and k_{TX} using analytic expressions (**Equations (3.2)**, **(3.3)**, and **(3.6)**). k_{on}, k_{off}, and k_{TX} were chosen randomly from ranges spanning four orders of magnitude, and data with $\langle n \rangle \geq 100$ were discarded. The black curve is a fit of the k_{off}-modulation curve to this data. The R^2 obtained from this fit is 0.08, significantly lower than that obtained from fitting to smFISH data ($R^2 = 0.81$, see **Figure 4.6A**).

(B) The histogram of the R^2 values obtained from fits of the k_{off}-modulation curve to 10,000 sets of data from randomly generated k_{on}, k_{off}, and k_{TX}. Each set has the same number of points as in our smFISH data. The mean R^2 is 0.14, significantly lower than that obtained from fitting to the smFISH data ($R^2 = 0.81$, see **Figure 4.6A**).

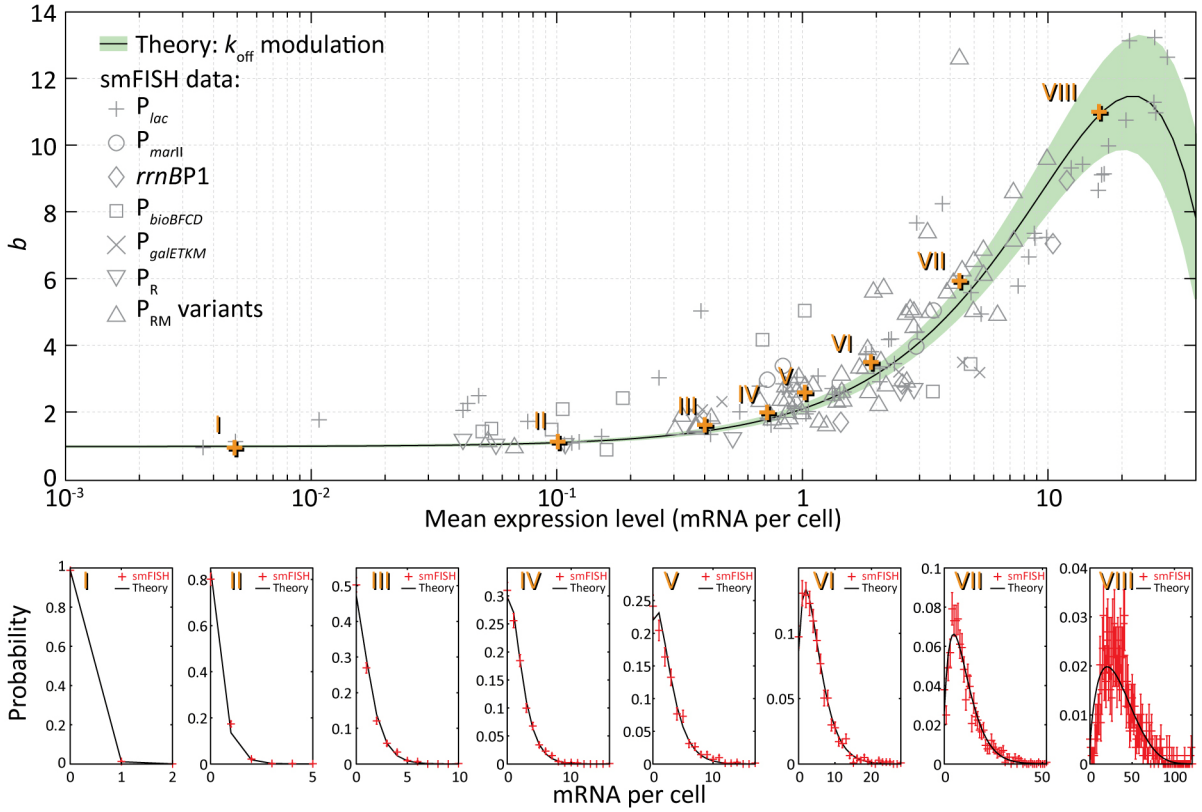


Figure 4.10. Comparing mRNA distributions from experiments and theory.

The distributions of mRNA copy-number (red markers, lower panels) from eight smFISH experiments (marked I to VIII in top panel) were compared to the theoretically predicted distributions (black lines, lower panels) (**Equation (3.1)**). To take into account the effects of the gene copy number X, the theoretical distributions were calculated using scaled transcription kinetics parameters: $k_{on} \rightarrow k_{on} \times \langle X \rangle$, $k_{off} \rightarrow k_{off} \times \langle X \rangle$, and $k_{TX} \rightarrow k_{TX} \times \langle X \rangle$, where $\langle X \rangle$ is the estimated average gene copy number of the population (**Section 3.2**).

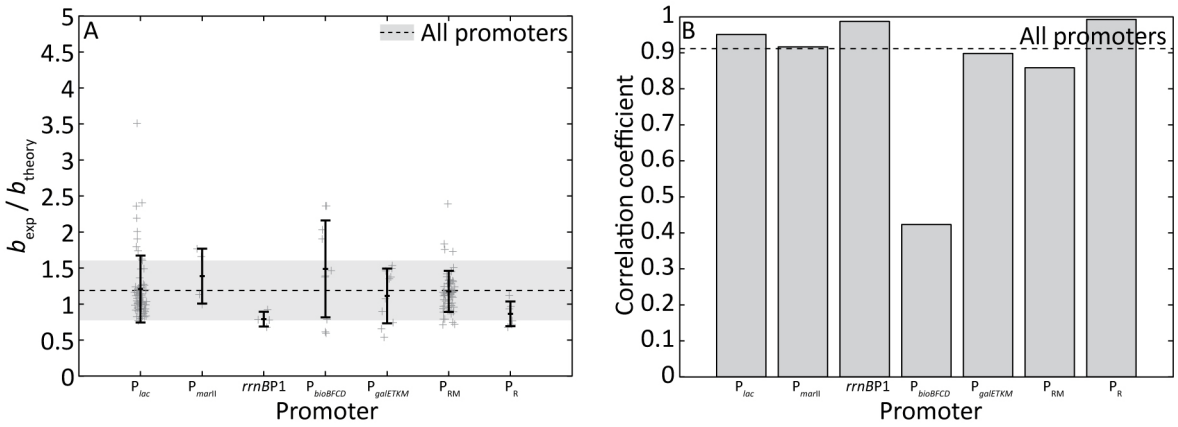


Figure 4.11. Comparing individual promoter data to the universal theoretical fit.

(A) The ratio of experimental to theoretical burstiness for individual promoters. b_{exp} is the burstiness estimated from smFISH data. b_{theory} is the burstiness predicted from the theoretical, gene-independent fit in **Figure 4.6A** (at the same expression level). Grey crosses denote the individual experiments. Black crosses mark the mean ± standard deviation of the data from each promoter. The dashed horizontal line and the horizontal grey patch denote the mean ± standard deviation for all the data sets pooled together. The average deviation of an individual promoter from the universal theoretical fit is ~33%.

(B) The correlation coefficient between the experimental and theoretical values of the burstiness. Data sets are the same as in Panel (A). The dashed horizontal line marks the correlation coefficient corresponding to all data sets pooled together. The correlation coefficients between six out of the seven promoters and the universal theoretical fit are bigger than 0.85.

5 Support from Live-Cell Data

In the previous chapter, I presented the gene-independent behavior of burstiness discovered using smFISH. Although smFISH is a powerful technique for studying the transcription kinetics in *E. coli*, it suffers from the fact that the data is only obtained from a snapshot of a population of cells at a certain time. Transcription events were not directly observed, and the properties of the transcriptional time-series were not quantified directly. The live-cell technique introduced in this chapter, in contrast, enables us to observe transcription events in growing cells in real-time, and complements smFISH experiments in transcription kinetics studies. The results obtained from the live-cell technique are consistent with the gene off-rate modulation scheme.

5.1 Real-time gene activity measurement in individual live cells

Labeling mRNA transcribed from the promoter-of-interest with MS2-GFP

The MS2-GFP system was developed by Golding *et al.* (Golding and Cox, 2004; Golding, Paulsson *et al.*, 2005) based on Bertrand *et al.* (Bertrand, Chartrand *et al.*, 1998; Fusco, Accornero *et al.*, 2003) for labeling mRNA in individual growing *E. coli* cells in real-time. It was previously used to demonstrate transcriptional bursting in *E. coli* (Golding and Cox, 2004; Golding, Paulsson *et al.*, 2005). **Figure 5.1** shows a schematic of this construct. Covalently linked bacteriophage MS2 coat protein dimer (MS2) (Peabody, 1997; Peabody and Lim, 1996) fused to the green fluorescent protein GFPmut3 (GFP) (Cormack, Valdivia *et al.*, 1996) is expressed in excess from a high-copy plasmid. On a BAC-based single-copy plasmid, the red fluorescent protein mRFP1 (Campbell, Tour *et al.*, 2002) followed by an array of 96 MS2 binding sites (MS2bs) is placed under the control of the promoter-of-interest. When transcription is

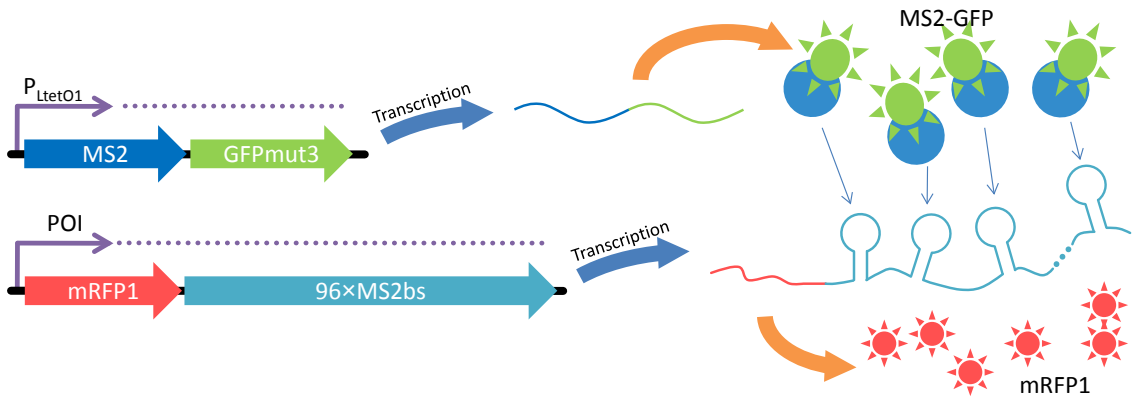


Figure 5.1. Schematic of the genetic construct for real-time gene activity measurement in individual live cells.

MS2-GFP is expressed in excess from a high-copy plasmid, under the control of an anhydrotetracycline-inducible promoter (P_{LtetO1}). mRFP1 followed by an array of 96 MS2bs is placed under the control of a promoter-of-interest (POI) in a BAC-based single-copy plasmid. When transcription is initiated from the promoter-of-interest, an mRNA molecule carrying 96 MS2bs emerges, and MS2-GFP quickly binds to the MS2bs, resulting in localized fluorescence, which can be imaged and quantified.

initiated from the promoter-of-interest, an mRNA encoding for mRFP1 together with 96 MS2bs emerges, and MS2-GFP quickly binds to the MS2bs. As a result, each mRNA would appear as a diffraction-limited green fluorescent spot in the cell when imaged using a fluorescence microscope (**Figure 5.2A**). The mRNA is then translated into mRFP1, and the red fluorescence level of the cell reflects the amount of protein produced.

Real-time mRNA counting in individual live cells with single-molecule resolution

To observe transcription events in growing cells in real-time, *E. coli* cells were grown under a fluorescence microscope in desired growth conditions, and time-lapse movies were made by imaging the cells at regular intervals, both in phase contrast and fluorescence channels (**Figure 5.2A**) (see **Appendix B.3** for details). As target mRNA molecules were transcribed and labeled with MS2-GFP, they appear as green fluorescence spots. As in the case of smFISH (**Section 2.2**), the number of mRNA molecules transcribed from the promoter-of-interest in a given cell at a given time point is estimated by measuring the total intensity of fluorescent spots in the cell, yielding an estimate of the number of bound MS2-GFP, which in turn indicates the number of target mRNA molecules (**Appendix B.3**) (Golding and Cox, 2008; Golding, Paulsson *et al.*, 2005). In this way I am able to get the number of mRNA molecules in a growing cell as a function of time for several hours.

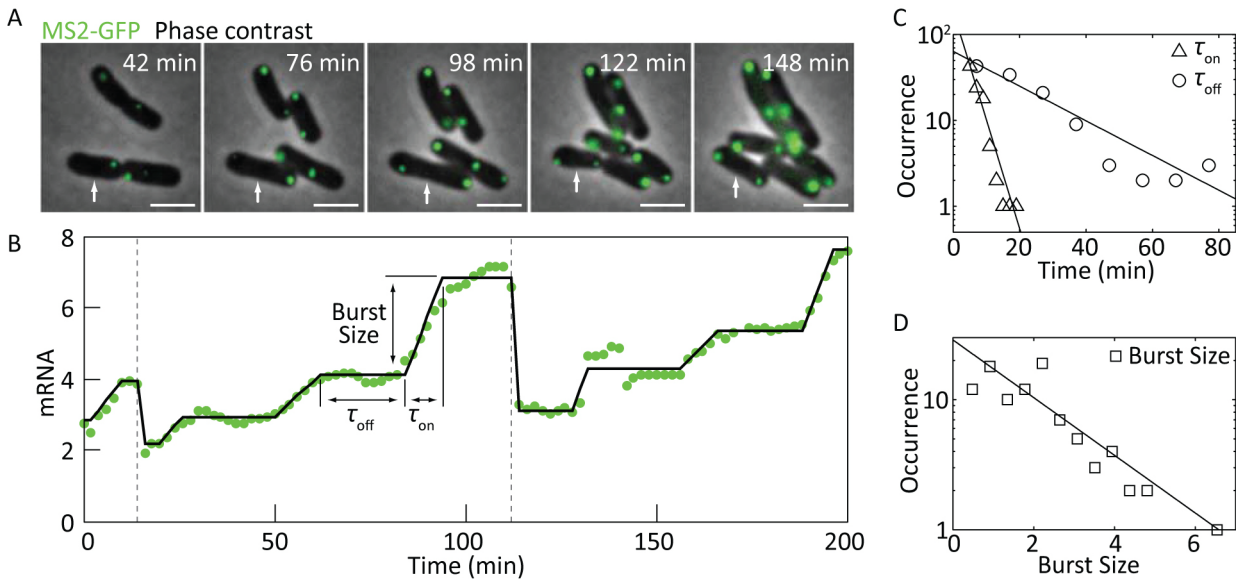


Figure 5.2. Extracting transcription kinetic parameters from real-time data in individual live cells.

(A) Snapshots from a time-lapse movie are shown. Each image is an overlay of the phase contrast and green fluorescence channels. Green fluorescence spots are mRNA molecules labeled by MS2-GFP. Scale bars equal 2 μm.

(B) The number of mRNA molecules as a function of time in the cell indicated by the white arrows in panel (A). Green data points are smoothed data, and the black line is the discretized trajectory obtained using a step detection program (**Appendix C.2**). Each step is a transcription bursting event, with the step height being the burst size and the step width being the "on" time τ_{on}. The duration of a constant level is the "off" time τ_{off}. Grey vertical dashed lines indicate cell divisions, where mRNA molecules segregate into daughter cells, causing a decrease in the mRNA number.

(C and D) The distribution of τ_{on}, τ_{off}, and burst size, extracted from a time-lapse movie consisting of many trajectories like the one shown in panel (B). All three parameters are exponentially distributed.

Automated image analysis, lineage reconstruction, and detection of transcription events

I obtained quantitative data from time-lapse movies with the help of MATLAB programs. Cell and spot recognition were performed the same manner as in smFISH experiments (**Section 2.2**). Cell recognition was performed on phase contrast images of cells using the *Schnitzcell* MATLAB module (gift of Michael Elowitz, California Institute of Technology) (**Figure 2.3A**). A spot recognition program developed in our lab was used to automatically identify and quantify localized fluorescence signals, yielding the number of mRNA molecules in each cell at each time-point of the time-lapse movie (**Figure 2.3B**).

A program developed in our lab was used to automatically track the cells and reconstruct the lineages of growing and dividing cells, based on positions of the recognized cells at each time-point of the time-lapse movie. The recognized spots were then assigned to the cell lineages accordingly, yielding trajectories of number of mRNA molecules as a function of time. A step detection program based on a *t*-test algorithm I developed was used to identify discrete mRNA levels and locate steps connecting the discrete mRNA levels, which represent transcription events (**Appendix C.2**). Transcription kinetic parameters were then extracted from the "discretized" mRNA trajectories (**Figure 5.2B**).

Calibration of mRNA counting

As in the case of smFISH, a fluorescence spot could consist of multiple mRNAs in close proximity. The integrated intensity arising from a single mRNA molecule needed to be estimated for each time-lapse movie so that fluorescence intensities could be normalized to give the absolute number of mRNA molecules (**Section 2.2**). The mapping between fluorescence and mRNA numbers is currently obtained by manually examining the histograms of intensities from all fluorescence spots in the time-lapse movie, and asking which intensity peak corresponds to a single molecule (Golding and Cox, 2008; Golding, Paulsson *et al.*, 2005). However, this method is inaccurate and in some experiments fails altogether, and the absolute number of mRNA molecules is not always obtainable. Possible ways of improving the calibration will be discussed in **Section 6.4**.

5.2 Live-cell data is consistent with the gene off-rate modulation scheme

I quantified the kinetics of mRNA production from one promoter, $P_{lac/ara}$ (Lutz and Bujard, 1997), in individual living cells. Cells were grown under the microscope in the presence of different levels of the inducers, IPTG and L-(+)-arabinose (see **Appendix B.3** for details). As expected, the mRNA kinetics consisted of periods of activity, where a random number of transcripts are produced, separated by periods of inactivity (**Figure 5.2B**) (Golding, Paulsson *et al.*, 2005). Measuring the mean durations of "off" and

"on" periods, τ_{off} and τ_{on} (both in minutes), as well as the amount of mRNA produced within each "on" period B (in arbitrary fluorescence units), allowed us to estimate k_{on}, k_{off}, and k_{TX} at a given gene activity level. To extract k_{TX}, the relationship $B \propto \tau_{on} k_{TX}$ was used, which could be written as $k_{TX} \propto B/\tau_{on}$. To extract k_{on}, I used the relationship $k_{on} \propto 1/\tau_{off}$. To extract k_{off}, I noted that k_{TX} was constant, and used the relationship $k_{off} \propto 1/\tau_{on} \propto 1/B$.

As seen in **Figure 5.3**, the behavior of these kinetic parameters is consistent with the observations based on smFISH experiments: Changing the level of mRNA $\langle n \rangle$ is achieved by varying k_{off}, while k_{on} and k_{TX} are kept approximately constant.

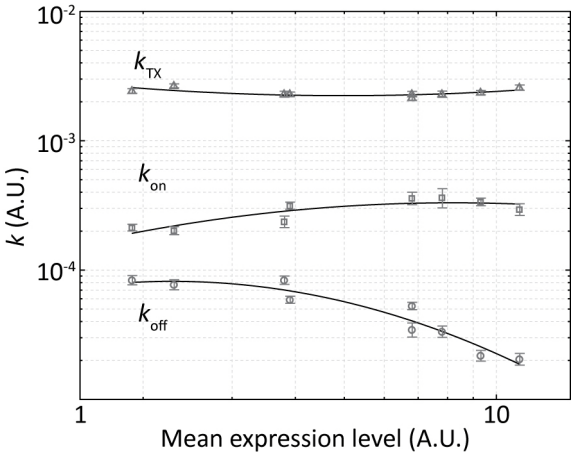


Figure 5.3. Live-cell data is consistent with gene off-rate modulation.

mRNA production from the promoter $P_{lac/ara}$ was quantified using the MS2-GFP method. Cells were grown under the microscope in the presence of varying levels of the inducers (IPTG and L-(+)-arabinose), leading to a range of expression levels. In each experiment, we measured the mean duration of inactive and active periods, as well as the amount of mRNA produced during active periods (**Figure 5.2**). These measurements allowed us to estimate the rate parameters in the two-state model. The data (markers) is from nine independent experiments (> 400 cells). The error bars represent the standard errors within each experiment. The solid lines are fits to second degree polynomials. The behavior of the rate parameters as a function of gene level was consistent with the observation that k_{off} is varied to control expression level, while k_{on} and k_{TX} remain approximately constant (less than two-fold change over the full set of experiments).

5.3 Limitations of the live-cell method

Being able to measure the transcription kinetics directly in live cells supplements the shortcoming of smFISH. Transcription kinetics is directly observed, eliminating the need to infer the kinetic parameters from population statistics. However, this method has its own drawbacks.

The genetic construct consisting of MS2-GFP and 96 MS2bs needs to be introduced into the cell (**Figure 5.1**), and the system is no longer endogenous. This perturbation would affect the physiology of the cells in many ways, and care needs to be taken when analyzing the data. MS2-GFP molecules tend to form clumps when expressed in large amounts, which not only impair the health of cells, but also affect mRNA counting (**Figure 5.4**). Moreover, the mRNA molecule with the 96 MS2bs becomes "immortalized" when tagged by MS2-GFP. It

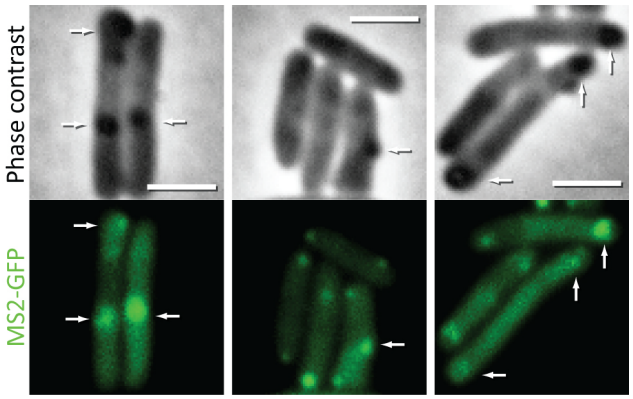


Figure 5.4. Clumping of MS2-GFP.

When expressed in excess, MS2-GFP molecules tend to form clumps (white arrows), which impair the health of cells and affect mRNA counting. Images are taken from a time-lapse movie. Scale bars equal 2 μm.

never gets degraded, and stays in the cell until it is diluted by cell growth and division (**Figure 5.5**; see also the mRNA trajectory in **Figure 5.2B**, which shows decreases in mRNA number only at cell divisions). A possible explanation is that bound MS2-GFP protects the mRNA molecule from RNA-degrading enzymes (Golding, Paulsson *et al.*, 2005). Phototoxicity arising from prolonged exposure of cells to fluorescence excitation light in the course of a time-lapse movie would also lead to suboptimal growth.

The live-cell technique discussed in this chapter and smFISH are complementary to each other. They each have their own strengths and weaknesses, and offer two different perspectives on the same system. To get a complete picture, the two methods should be used in conjunction.

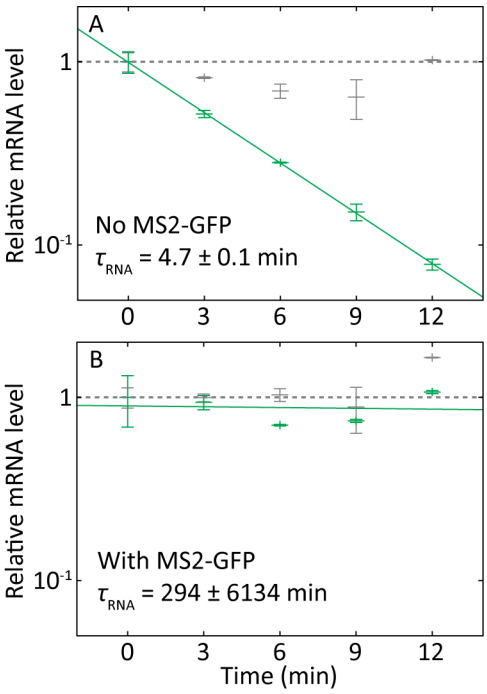


Figure 5.5. mRNA gets immortalized when tagged by MS2-GFP.

The relative mRNA level, measured using qRT-PCR, as a function of time after adding rifampicin (green data points) (A) without MS2-GFP; or (B) with MS2-GFP. The green line is a fit of an exponential decay to the green data points. The grey data points are negative controls with no rifampicin added. With MS2-GFP, the mRNA level essentially remained constant throughout the course of the experiment.

5.4 Summary

Even though smFISH is a powerful technique for studying transcription kinetics in *E. coli*, transcription events are not directly observed. The live-cell technique introduced in this chapter enables us to observe transcription events in growing cells in real-time, and complements smFISH experiments. Transcription kinetics in one *E. coli* promoter was quantified, and the results were consistent with the gene off-rate modulation scheme.

6 Summary and Outlook

6.1 Gene-independent behavior of transcription burstiness

Multiple studies in recent years have demonstrated that the transcriptional time-series is often bursty rather than Poissonian (Chubb and Liverpool, 2010; Chubb, Trcek *et al.*, 2006; Golding, Paulsson *et al.*, 2005; Pare, Lemons *et al.*, 2009; Raj, Peskin *et al.*, 2006; Yunger, Rosenfeld *et al.*, 2010), and can be described by the phenomenological two-state model (Peccoud and Ycart, 1995; Raj, Peskin *et al.*, 2006; Shahrezaei and Swain, 2008). In this work I have extended and generalized these observations by describing how the transcriptional time-series in *E. coli* is modulated when the gene expression level is varied. I found that the transcriptional time-series tends to be Poissonian at low expression levels (at or below one mRNA molecule per cell); the degree of burstiness, as characterized by the Fano factor $b = \sigma^2/\langle n \rangle$, then rises in a sub-linear manner with increasing gene activity. The observed behavior is consistent with varying the gene off-rate as the means to change the expression level, while maintaining the gene on-rate and transcription rate constant. In other words, the duration of the transcription bursts is the main feature that changes as the expression level is varied. Importantly, this behavior is not gene- or input-specific (although it can also be observed when examining a single gene, see **Figure 4.7**); rather, it was observed in the complete ensemble of promoters and stimuli examined. Note that a more complex scenario, where multiple kinetic parameters are simultaneously varied, is also consistent with the observed smFISH data (see **Appendix D.2** and **Figure D.1**). However, such a scenario does not need to be invoked in order to explain the experimental data. The multi-parameter modulation scenario also appears inconsistent with the live-cell data (**Figure 5.3**).

A number of past studies have characterized the noise, defined as the squared coefficient of variation $\eta^2 = \sigma^2/\langle n \rangle^2$, of multiple genes, using a library of fluorescent protein fusions (Bar-Even, Paulsson *et al.*, 2006; Geva-Zatorsky, Dekel *et al.*, 2010; Newman, Ghaemmaghami *et al.*, 2006; Taniguchi, Choi *et al.*, 2010). A study in yeast (Bar-Even, Paulsson *et al.*, 2006) found that the noise η^2 displayed a genome-wide trend of a power-law dependence on the mean expression level (a similar trend was recently observed when examining different mutants of a single yeast promoter (Mao, Brown *et al.*, 2010)). This gene-independent behavior is consistent with my findings here. Moreover, by modeling the underlying kinetics, it was shown in the same study that the protein fluctuations were likely dominated by the mRNA species (Bar-Even, Paulsson *et al.*, 2006), as assumed in this work. A recent genome-wide study in *E. coli* (Taniguchi, Choi *et al.*, 2010) found that the Fano factor increased monotonically with the mean protein level. This observation is most easily explained by my findings of a gene-independent behavior of the transcriptional burst size (**Figure 4.1B** and **Figure 4.6A**).

6.2 Proposed molecular mechanisms that could give rise to transcription bursts

At this stage, there is no mechanistic, molecular-level understanding of what gives rise to the bursty behavior of gene activity in bacteria; specifically, what the physiological nature of the gene "on" and "off" states is, and how the rates of switching between states can be varied in the individual cell or over the time course of evolution. The most common theoretical model used to explain the two-state gene activity in bacteria involves the binding and unbinding of transcription factors at the promoter (Garcia, Sanchez *et al.*, 2010; Golding, Paulsson *et al.*, 2005; Kepler and Elston, 2001; Mitarai, Dodd *et al.*, 2008; Sanchez, Garcia *et al.*, 2011; Simpson, Cox *et al.*, 2004). However, my finding here, that the properties of the transcriptional time-series are gene-independent rather than gene-specific, suggests that the observed two-state kinetics involves gene-nonspecific mechanisms such as DNA topology modulation, RNA polymerase dynamics, or regulation by broad-target DNA-binding proteins (Mitarai, Dodd *et al.*, 2008; Mooney, Davis *et al.*, 2009; Reppas, Wade *et al.*, 2006). Interestingly, these types of mechanisms are reminiscent of those suggested to underlie the non-Poissonian transcription kinetics in eukaryotes, where burstiness is broadly ascribed to "chromatin modifications" (Blake, M *et al.*, 2003; Raj, Peskin *et al.*, 2006; Whitelaw, Chong *et al.*, 2010). Future studies will have to reveal whether the fact that transcription burstiness appears in both kingdoms reflects a similarity in the underlying mechanisms, or instead results from the selection of an advantageous phenotype in different systems.

6.3 Extending live-cell measurements to endogenous promoters

An important limitation of this work is that by mainly using smFISH to count mRNA, I was only able to obtain snapshots of cell populations, but was naturally unable to follow the time-course of gene activity in individual cells (with the exception of a single promoter). This limitation prevented me from examining temporal correlations in the transcriptional time-series. Correlations in the gene-activity trajectories of individual cells have been shown to contain important information about the underlying gene regulatory network (Austin, Allen *et al.*, 2006; Cox, McCollum *et al.*, 2008). Such correlations are likely to be affected by the bursty behavior described here. The MS2-based system (**Chapter 5**) (Golding, Paulsson *et al.*, 2005) was used to study the transcription kinetics of one promoter, but the promoter-of-interest was on a plasmid. The observed transcription kinetics could be different from that of a chromosomal promoter, because plasmid DNA is physiologically different from the chromosomal DNA. Extending the use of the MS2-based system to multiple endogenous *E. coli* promoters should allow the direct observation of transcription kinetics and the characterization of the temporal correlations in the transcription time-series in the future.

I am planning to extend the MS2-based system to multiple endogenous *E. coli* promoters by moving the MS2bs array (**Figure 5.1**) into the *E. coli* chromosome. This will be done using recombineering

(Ausubel, 1987; Datsenko and Wanner, 2000), a genetic engineering technique in bacteria utilizing homologous recombination. Briefly, the DNA fragment to be inserted into the chromosome will be amplified using PCR, with homologies to the chromosome added to both ends. In this case the fragment will be the MS2bs array together with an antibiotic resistance gene for selecting successful recombinants. Using electroporation, this DNA fragment will then be transformed into a bacterial strain expressing a bacteriophage recombination system (the bacteriophage λ Red system consisting of the *exo*, *bet*, and *gam* genes). The fragment will be incorporated into the chromosome by homologous recombination and successful recombinants will be selected with antibiotics. Using P1 transduction (Ausubel, 1987), the chromosomal region with the MS2bs array under the control of a promoter-of-interest will then be moved to a "clean" strain without the bacteriophage recombination machinery and suitable for experiments.

With the new strains I will be able to follow the transcription kinetics of endogenous promoters in real-time, allowing further characterization of the temporal aspects of bursty transcriptional time-series.

6.4 Improving the absolute calibration of mRNA counting

The calibration of mRNA counting is of utmost importance to the accuracy of the data from smFISH and MS2-GFP-based experiments. The measured fluorescence needs to be reliably mapped to absolute number of mRNA molecules. Calibration is the weakest link of my current mRNA counting algorithm, both in the smFISH and MS2-based live cell technique. Here are some possible ways to improve the calibration:

(i) *In vitro* transcribed mRNA can be used as an absolute standard for calibrating smFISH experiments. The mRNA corresponding to the gene-of-interest can be transcribed using a standard *in vitro* transcription system (Ausubel, 1987), and the amount of mRNA can be quantified using spectrophotometry. The *in vitro* transcribed mRNA can then be added to a qRT-PCR reaction at a known number of molecules per cell, and the result can be compared against the results from positive samples of qRT-PCR and smFISH experiments.

(ii) Individual smFISH probes can be imaged (preliminary results shown in **Figure 6.1**). The photon count from a single dye molecule can be measured and used for calibration of smFISH experiments. By comparing the total fluorescence intensity of individual spots in smFISH samples to that of single probes, the number of mRNA molecules can be estimated using the known number of probes hybridized to each transcript-of-interest.

(iii) Individual MS2-GFP molecules can be imaged. MS2-GFP can be purified using an affinity tag in the expression vector (Golding and Cox, 2004), and individual molecules can be imaged. The photon count from a single MS2-GFP molecule can be measured and used for calibration of live-cell experiments. By comparing the total fluorescence intensity of individual spots in time-lapse movies of live cells to that of single molecules, the number of mRNA molecules can be estimated using the known number of MS2-GFP molecules attached to each MS2bs array.

(iv) Purified MS2-GFP molecules can also be used in *in vitro* labeling of purified target mRNA molecules, and the intensity of individual labeled mRNA molecule can be measured.

(v) Single MS2-GFP fluorescence intensity can also be estimated using the binomial statistics characterizing the partitioning of cytoplasmic components at cell division (Rosenfeld, Perkins *et al.*, 2006; Rosenfeld, Young *et al.*, 2005) and used for calibration of live-cell data.

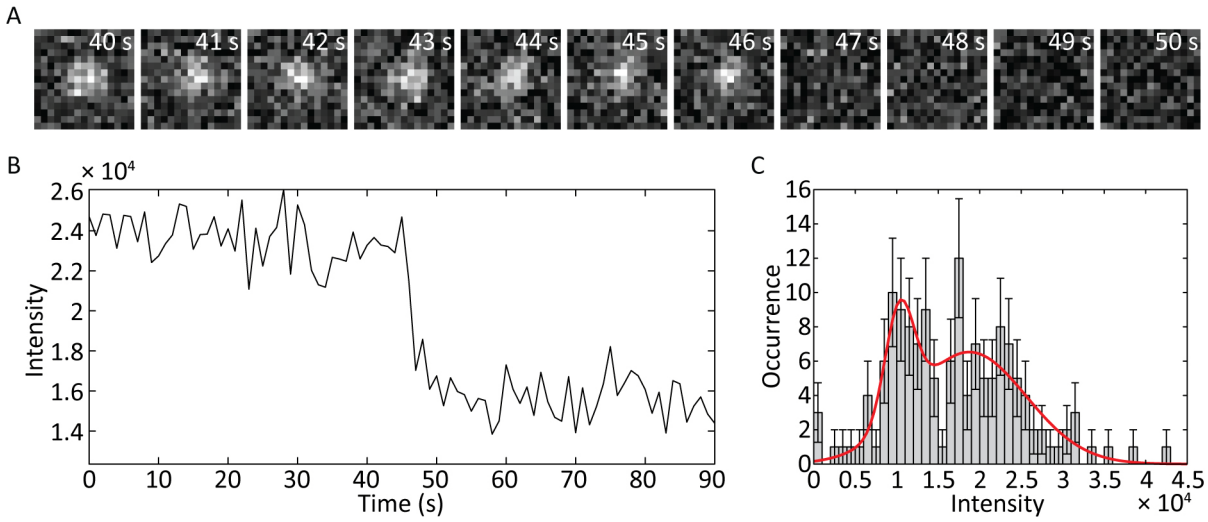


Figure 6.1. Imaging individual smFISH probes.

(A) An excerpt from a time-lapse movie of an individual smFISH probe. The TAMRA dye molecule was bleached after 46 seconds of exposure to excitation light.

(B) The intensity of the single smFISH probe in panel (A) as a function of exposure time.

(C) The histogram of intensities of fluorescence spots in an image of smFISH probe solution. The error bars represent Poisson shot noise. The red curve is a fit to a sum of two Gaussians. The peaks of the two Gaussians are located at 1.0×10^4 and 1.9×10^4, approximately a two-fold difference, suggesting that the spots were either single probes or clumps of two probes.

Appendix A. Strains, Growth Media, and Growth Conditions

A.1 Growth media and conditions

All strains were grown in M9 minimal media with thiamine and casamino acids, and glucose as a carbon source (M9CAgluc, Teknova, #M8010) unless otherwise stated. All strains were grown at 37°C with shaking unless otherwise stated.

To achieve different expression levels from P_{lac}, strain TK310 (gift of Terrence Hwa, University of California, San Diego) was grown with 0 to 1 mM of isopropyl β-D-1-thiogalactopyranoside (IPTG, Sigma, #I6758) and 0 to 10 mM of adenosine 3',5'-cyclic monophosphate (cAMP) (Sigma, #A9501).

To achieve different expression levels from $P_{galETKM}$, strain JT235$_S$ (gift of Sankar Adhya, National Institutes of Health) was grown with 10 mM cAMP and 0 to 30 mM of D-fucose (Sigma, #F8150).

To achieve different expression levels from P_{mar1I}, strain LC544 (gift of Christopher Rao, University of Illinois at Urbana-Champaign) was grown with 0 to 10 mM of sodium salicylate (Sigma, #S3007).

To achieve different expression levels from *rrnB*P1, strain CF7753 (gift of Michael Cashel, National Institutes of Health) was grown in minimal media with different carbon sources, which resulted in different growth rates. Specifically, M9 minimal media (with thiamine) with glucose, glycerol, succinate, and acetate as carbon source were used. These are prepared following standard molecular biology protocols (Ausubel, 1987; Miller, 1992; Sambrook and Russell, 2001).

To achieve different expression levels from $P_{bioBFCD}$, strain CY481 (gift of John Cronan, University of Illinois at Urbana-Champaign) was grown in M9 minimal media with thiamine, vitamin-free casamino acids (Difco, #228830), and glucose, prepared following standard molecular biology protocols (Ausubel, 1987; Miller, 1992; Sambrook and Russell, 2001), with 0.0016 to 1.6 μM of biotin (Sigma, #B4501). To get rid of traces of biotin from the LB plate on which the cells grew, cells from overnight cultures were washed with media with the same conditions as the experimental conditions three times before diluted into fresh media.

To achieve different expression levels from P_{RM}, lysogens of various P_{RM} and/or *cI* mutants were used (P_{RM} mutants are gift of John Little, University of Arizona). When strain JL5902 (RecA$^-$) (gift of John Little, University of Arizona) was used as host, cells were grown in LBGM (LB (Ausubel, 1987) with 10 mM $MgSO_4$ and 0.2% glucose). When strain JL2497 (RecA$^+$) (gift of John Little, University of Arizona) was used as host, cells were grown in M9CAgluc.

To achieve different expression levels from P_{RM} and P_R with the temperature sensitive allele *cI*857, reporter strain NC416 was used (gift of Donald Court, National Institutes of Health). Cells were grown in LB at 30, 32, 34, 36, 38, and 40°C.

A.2 Strains

Bacterial strains

Bacterial strains used are listed in **Table A.1**. All of them are *E. coli* K-12 derivatives. BW14894 was used as a negative control for smFISH experiments using *lacZ* probes. MG1655 was used as a negative control for smFISH experiments using *cI* and *cro* probes. MG1655, JL5902, and JL2497 were used as hosts for bacteriophage λ in smFISH experiments using *cI* probes to study the P_{RM} promoter.

Table A.1. Bacterial strains used in this work.

Strain	Relevant genotype	Promoter	Reporter[16]	Reference	Source
BW14894	Δ*lacIZYA*	-	-	Yakovleva, Kim *et al.*, 1998	CGSC #8280
MG1655	Wild-type, λ^-	-	-		Lab stock
JL5902	$RecA^-$	-	-	Little, Shepley *et al.*, 1999	John Little
JL2497	$RecA^+$	-	-	Little, Shepley *et al.*, 1999	John Little
TK310	Δ*cyaA* Δ*cpdA* Δ*lacY*	P_{lac}	Endogenous *lacZ* mRNA	Kuhlman, Zhang *et al.*, 2007	Terrence Hwa
JT235$_S$	Δ*lacIZY galE*::*lacZ F' gal*$^+$	$P_{galETKM}$	*lacZ*	Tokeson, Garges *et al.*, 1991	Sankar Adhya
LC544	Δ*lacIZY* P_{marII}-*lacZ*	P_{marII}	*lacZ*		Christopher Rao
LC543	Δ*lacIZY*	-	-		Christopher Rao
CF7753	*ΔlacZ(mluI) rrnB*P1''::*lacZ-kan*	*rrnB*P1	*lacZ**	Hernandez and Bremer, 1990; Potrykus, Vinella *et al.*, 2006	Michael Cashel
CY481	Δ*lacIZY bioF-lacZ*	$P_{bioBFCD}$	*lacZ*	Abdel-Hamid and Cronan, 2007; Barker and Campbell, 1980	John Cronan
NC416	*luc-N'...cro+ cI1'-lacZYA*	P_{RM}, P_R	*cI*, *cro*	Svenningsen, Costantino *et al.*, 2005	Donald Court

[16] *lacZ** has an inefficient ribosome binding site to avoid toxicity from excessive LacZ levels (Hernandez and Bremer, 1990; Potrykus, Vinella *et al.*, 2006) (Michael Cashel, personal communication).

Phage strains

Phage strains used are listed in **Table A.2**.

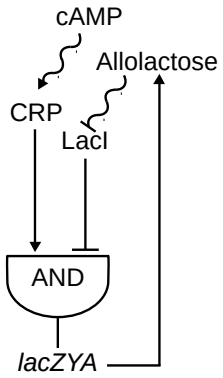
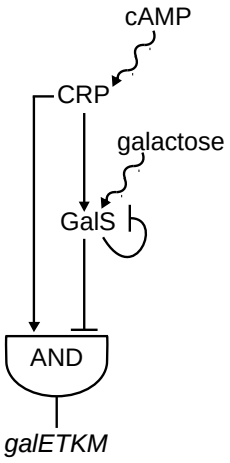
Table A.2. Phage strains used in this work.

Strain	Relevant genotype	Reference	Source
λ_{IG831}	Wild-type		Lab stock
λ_{IG2504}	*c*I T88C	Sauer, Jordan *et al.*, 1990	Lab stock
$\lambda_{IG28061}$	*c*I V36I	Lim and Sauer, 1989	Lab stock
$\lambda_{IG28062}$	*c*I L18V	Lim and Sauer, 1989	Lab stock
λ_{JL815}	*c* I D38N	Michalowski and Little, 2005	John Little
λ_{NP2}	*c*I D38N P_{RM} -35 TAGA→ GCTG P_{RM} -10 GATT→TATT	Michalowski and Little, 2005	John Little
λ_{NP3}	*c*I D38N P_{RM} -35 TAGA→CATT P_{RM} -10 GATT→GAAT	Michalowski and Little, 2005	John Little
λ_{NP4}	*c*I D38N P_{RM} -35 TAGA→CCTT P_{RM} -10 GATT→CCAT	Michalowski and Little, 2005	John Little
λ_{NP5}	*c*I D38N P_{RM} -35 TAGA→CTAA P_{RM} -10 GATT→GAAT	Michalowski and Little, 2005	John Little
λ_{NP6}	*c*I D38N P_{RM} -35 TAGA→CCCA P_{RM} -10 GATT→TGAT	Michalowski and Little, 2005	John Little
λ_{NP7}	*c*I D38N P_{RM} -35 TAGA→TACC P_{RM} -10 GATT→TACT	Michalowski and Little, 2005	John Little
λ_{NP8}	*c*I D38N P_{RM} -35 TAGA→GTGT P_{RM} -10 GATT→GTAT	Michalowski and Little, 2005	John Little
λ_{NP10}	*c*I D38N P_{RM} -35 TAGA→CCAA P_{RM} -10 GATT→GAAT	Michalowski and Little, 2005	John Little
λ_{NP11}	*c*I D38N P_{RM} -35 TAGA→CTCA P_{RM} -10 GATT→TGCT	Michalowski and Little, 2005	John Little

Promoters

Table A.3 provides additional details of the promoters used in this study.

Table A.3. Characteristics of promoters used in this work.

Promoter	Expression level (mRNA/cell)	Molecular mode(s) of regulation	Network motif	References
P_{lac}	~0.01–60	Activated by CRP[17] Repressed by LacI[18]	 Simple regulation with double negative feedback	Alon, 2007; Müller-Hill, 1996
$P_{galETKM}$	~0.1–10	Activated by CRP Repressed by GalR[19] and GalS[20]	 Incoherent feed-forward loop	Alon, 2007; Weickert and Adhya, 1993

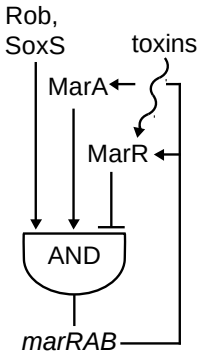
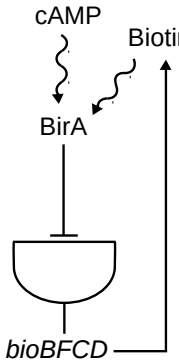
[17] CRP: cAMP (3'-5'-cyclic adenosine monophosphate) receptor protein, activated in the presence of cAMP.

[18] LacI: *lac* repressor, inactivated in the presence of lactose or lactose analogues.

[19] GalR: *gal* repressor, inactivated in the presence of D-galactose or D-fucose.

[20] GalS: *gal* isorepressor, inactivated in the presence of D-galactose or D-fucose.

Table A.3. (continued)

Promoter	Expression level (mRNA/cell)	Molecular mode(s) of regulation	Network motif	References
P_{marRAB}	~2–10	Activated by MarA[21], Rob[22] and SoxS[23] Repressed by MarR[24]	 Simple regulation with positive and negative feedback	Alekshun and Levy, 1999
***rrnB*P1**	~1–20	Constitutive	Not fully characterized	Liang, Bipatnath *et al.*, 1999
$P_{bioBFCD}$	~0.1–10	Repressed by BirA[25]	 Simple regulation with negative feedback	Abdel-Hamid and Cronan, 2007

[21] MarA: *mar* activator.

[22] Rob: Rob transcriptional activator.

[23] SoxS: transcriptional dual regulator of superoxide response regulon.

[24] MarR: *mar* repressor, inactivated in the presence of sodium salicylate and a variety of antibiotics and toxic chemicals.

[25] BirA: biotin transcriptional repressor, activated in the presence of biotin, sequestered by unbiotinylated AccB (biotin carboxyl carrier protein).

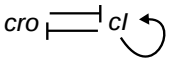
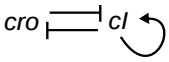
Table A.3. (continued)

Promoter	Expression level (mRNA/cell)	Molecular mode(s) of regulation	Network motif	References
P_{RM} (*c*/D38N)	~2	No positive autoregulation by CI[26], reduced P_{RM} activity		Michalowski and Little, 2005
P_{RM} (*c*/D38N P_{RM} -35 TAGA→ GCTG P_{RM} -10 GATT→TATT)	~7	No positive autoregulation by CI, reduced P_{RM} activity		Michalowski and Little, 2005
P_{RM} (*c*/D38N P_{RM} -35 TAGA→ CATT P_{RM} -10 GATT→GAAT)	~1	No positive autoregulation by CI, reduced P_{RM} activity		Michalowski and Little, 2005
P_{RM} (*c*/D38N P_{RM} -35 TAGA→ CCTT P_{RM} -10 GATT→CCAT)	~2	No positive autoregulation by CI, reduced P_{RM} activity		Michalowski and Little, 2005
P_{RM} (*c*/D38N P_{RM} -35 TAGA→ CTAA P_{RM} -10 GATT→GAAT)	~2	No positive autoregulation by CI, reduced P_{RM} activity	*cro* ⊣ *cI* ⊣ *cro* Two-node double negative feedback loop	Michalowski and Little, 2005
P_{RM} (*c*/D38N P_{RM} -35 TAGA→ CCCA P_{RM} -10 GATT→TGAT)	~2	No positive autoregulation by CI, reduced P_{RM} activity		Michalowski and Little, 2005
P_{RM} (*c*/D38N P_{RM} -35 TAGA→ TACC P_{RM} -10 GATT→TACT)	~2	No positive autoregulation by CI, reduced P_{RM} activity		Michalowski and Little, 2005
P_{RM} (*c*/D38N P_{RM} -35 TAGA→ GTGT P_{RM} -10 GATT→GTAT)	~2	No positive autoregulation by CI, reduced P_{RM} activity		Michalowski and Little, 2005
P_{RM} (*c*/D38N P_{RM} -35 TAGA→ CCAA P_{RM} -10 GATT→GAAT)	~2	No positive autoregulation by CI, reduced P_{RM} activity		Michalowski and Little, 2005
P_{RM} (*c*/D38N P_{RM} -35 TAGA→ CTCA	~3	No positive autoregulation by CI, reduced P_{RM} activity		Michalowski and Little, 2005

[26] λ repressor, activates P_{RM} at lower concentrations and represses P_{RM} at higher concentrations.

P_{RM} -10 GATT→TGCT)

Table A.3. (continued)

Promoter	Expression level (mRNA/cell)	Molecular mode(s) of regulation	Network motif	References
P_{RM}	~6	Autoregulated by CI Repressed by Cro[27]		Ptashne, 2004
P_{RM} (*c*/T88C)	~6	Covalent dimerization of CI	Two-node double negative feedback loop with single node positive autoregulation	Sauer, Jordan *et al.*, 1990
P_{RM} (*c*/V36I)	~4	Misfolding of CI		Lim and Sauer, 1989
P_{RM} (*c*/L18V)	~9	Misfolding of CI		Lim and Sauer, 1989
P_R	~0.1–6	Autoregulated by Cro Repressed by CI	 Two-node double negative feedback loop with single node positive autoregulation	Ptashne, 2004

[27] λ "anti-repressor", represses P_{RM} and P_R.

Appendix B. Detailed Description of Experimental Protocols

B.1 Single-molecule fluorescence *in situ* hybridization (smFISH)

The procedures are based on Raj *et al.* (Raj, van den Bogaard *et al.*, 2008). Modifications were made to adapt the protocol to *E. coli* (Zong, So *et al.*, 2010). Sterile, nuclease-free, aerosol-barrier pipette tips were used in all procedures. Nuclease-free reagents were used whenever available. Diethylpyrocarbonate (DEPC)-treated water (Ambion, #AM9922) was used wherever the protocol calls for water.

Probe design and labeling

DNA oligonucleotide probes were designed using the online program developed by Arjun Raj[28] (van Oudenaarden Lab, Massachusettes Institute of Technology). DNA oligos with 3'-end amine modification were ordered from Biosearch Technologies[29]. The 96-well cartridge in which the oligos were delivered was spun briefly to get all liquid to the bottom of the wells. The content of each well (100 µl) was then transferred to a microcentrifuge tube and stored at -20°C.

To label the probes with fluorescent dyes, 10 µl of each DNA oligo was first pooled into a microcentrifuge tube. 1/9 volume of 1 M sodium bicarbonate solution (filter sterilized) (Fisher Scientific, #BP328) was added to the pool to get a final concentration of 0.1 M sodium bicarbonate. 2.5 mg of fluorescent dye (6-TAMRA, SE (Invitrogen, #C6123) for *lacZ* and *cI* probes, Cy5 Bis NHS ester (GE Healthcare Life Sciences #PA15000) for *cro* probes) was dissolved in 5 µl of dimethyl sulfoxide (Fisher Scientific, #BP231), and 50 µl of 0.1 M sodium bicarbonate solution was added to the solution. The DNA oligo solution was then added to the dye solution. The tube was wrapped in aluminum foil and incubated overnight at 37°C in the dark.

The labeled probes were purified by ethanol precipitation. The solution was transferred to a 50 ml centrifuge tube. 1/9 volume of 3 M sodium acetate solution (pH 5.2, autoclaved) (Fisher Scientific, #BP333) was added to get to a final concentration of 0.3 M sodium acetate. 2.5 volumes of 100% ethanol was added and mixed well. The solution was aliquoted to microcentrifuge tubes and incubated at -70°C for at least 3 hours (up to overnight). It was then centrifuged in a bench-top microcentrifuge at maximum speed for 30 minutes. The supernatant was discarded and any remaining liquid was absorbed with Kimwipe. The pellet was dissolved in 90 µl of water. 10 µl of 3 M sodium acetate solution (pH 5.2) was added to obtain a final concentration of 0.3 M sodium acetate. This was followed by two more rounds of

[28] http://singlemoleculefish.com/designer.html

[29] Oligos were ordered with a synthesis scale of 10 nmol delivered, cartridge purification, and water as the buffer solution.

ethanol precipitation. Finally, the pellets were dissolved in a total of 500 µl Tris-EDTA (pH 8.0) (Fisher Scientific, #BP2473). The tube was wrapped in aluminum foil and stored at -20°C.

After labeling and ethanol precipitations, the absorbance spectrum of a ten-fold diluted probe solution was measured using a spectrophotometer (Thermo Scientific Nanodrop 2000) (**Figure B.1**). The concentration of DNA oligos in the solution can be estimated from the absorbance at 260 nm with the well known extinction coefficient of DNA at 260 nm. On the other hand, the absorption peak of the fluorescent dye and its extinction coefficient at this peak are available from the manufacturer, and the concentration of dye molecules in the solution can again be estimated. The molar concentrations of DNA oligos and fluorescent dye molecules were close to one, meaning there are no significant amounts of excess DNA oligos or dye molecules in the solution.

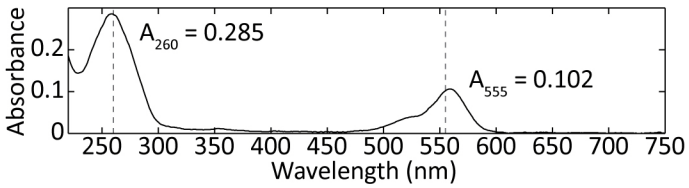


Figure B.1. Measuring the absorbance of smFISH probes using a spectrophotometer.

The absorbance spectrum of a ten-fold diluted probe solution was measured using the Nanodrop spectrophotometer. The DNA oligo concentration was estimated from the absorbance at 260 nm (A_{260}) and the fluorescence dye (TAMRA) concentration was estimated from the absorbance at 555 nm (A_{555}).

Sequences of *lacZ* probes (5' to 3'):

GTGAATCCGTAATCATGGTC; TCACGACGTTGTAAAACGAC; ATTAAGTTGGGTAACGCCAG;
TATTACGCCAGCTGGCGAAA; ATTCAGGCTGCGCAACTGTT; AAACCAGGCAAAGCGCCATT;
AGTATCGGCCTCAGGAAGAT; AACCGTGCATCTGCCAGTTT; TAGGTCACGTTGGTGTAGAT;
AATGTGAGCGAGTAACAACC; GTAGCCAGCTTTCATCAACA; AATAATTCGCGTCTGGCCTT;
AGATGAAACGCCGAGTTAAC; AATTCAGACGGCAAACGACT; TTTCTCCGGCGCGTAAAAAT;
ATCTTCCAGATAACTGCCGT; AACGAGACGTCACGGAAAAT; GCTGATTTGTGTAGTCGGTT;
TTAAAGCGAGTGGCAACATG; AACTGTTACCCGTAGGTAGT; ATAATTTCACCGCCGAAAGG;
TTTCGACGTTCAGACGTAGT; ATAGAGATTCGGGATTTCGG; TTCTGCTTCAATCAGCGTGC;
ACCATTTTCAATCCGCACCT; TTAACGCCTCGAATCAGCAA; ATGCAGAGGATGATGCTCGT;
TCTGCTCATCCATGACCTGA; TTCATCAGCAGGATATCCTG; CACGGCGTTAAAGTTGTTCT;
TGGTTCGGATAATGCGAACA; TTCATCCACCACATACAGGC; TGCCGTGGGTTTCAATATTG;
ATCGGTCAGACGATTCATTG; TGATCACACTCGGGTGATTA; ATACAGCGCGTCGTGATTAG;
GATCGACAGATTTGATCCAG; AAATAATATCGGTGGCCGTG; TTTGATGGACCATTTCGGCA;
TATTCGCAAAGGATCAGCGG; AAGACTGTTACCCATCGCGT; TGCCAGTATTTAGCGAAACC;
AAACGGGGATACTGACGAAA; TAATCAGCGACTGATCCACC; GGGTTGCCGTTTTCATCATA;
TCGGCGTATCGCCAAAATCA; TTCATACAGAACTGGCGATC; TGGTGTTTTGCTTCCGTCAG;
ACGGAACTGGAAAAACTGCT; TATTCGCTGGTCACTTCGAT; GTTATCGCTATGACGGAACA;
TTTACCTTGTGGAGCGACAT; GTTCAGGCAGTTCAATCAAC; TTGCACTACGCGTACTGTGA;
AGCGTCACACTGAGGTTTTC; ATTTCGCTGGTGGTCAGATG; ACCCAGCTCGATGCAAAAAT;
CGGTTAAATTGCCAACGCTT; CTGTGAAAGAAAGCCTGACT; GGCGTCAGCAGTTGTTTTTT;
TACGCCAATGTCGTTATCCA; TAAGGTTTTCCCCTGATGCT; ATCAATCCGGTAGGTTTTCC;

GTAATCGCCATTTGACCACT; AGTTTTCTTGCGGCCCTAAT; ATGTCTGACAATGGCAGATC;
ATAATTCAATTCGCGCGTCC; TGATGTTGAACTGGAAGTCG; TCAGTTGCTGTTGACTGTAG;
ATTCAGCCATGTGCCTTCTT; AATCCCCATATGGAAACCGT; AGACCAACTGGTAATGGTAG.

Sequences of *cI* probes (5' to 3'):

GGTTTCTTTTTTGTGCTCAT; CTCAAGCTGCTCTTGTGTTA; AATTGCTTTAAGGCGACGTG;
GGGATAAGCCAAGTTCATTT; ATCTTGTCTGCGACAGATTC; AATAAAGCACCAACGCCTGA;
GCATTTAATGCATTGATGCC; TGCAAGCAATGCGGCGTTAT; CTTCAACGCTAACTTTGAGA;
CTGGCGATTGAAGGGCTAAA; CGCTTCATACATCTCGTAGA; TAAGTGACGGCTGCATACTA;
ACAGGGTACTCATACTCACT; CCCTGCCTGAACATGAGAAA; TTCTAAGCTCAGGTGAGAAC;
TCCGCATCACCTTTGGTAAA; TTTGGTTGTGCTTACCCATC; AGAATGCAGAATCACTGGCT;
CGGTCATGGAATTACCTTCA; AGCTTGGCTTGGAGCCTGTT; AGAATTAACATTCCGTCAGG;
AACAGCCTGCTCAGGGTCAA; CTATGCAGAAATCACCTGGC; AACTCATCACCCCCAAGTCT;
CCTGATCAGTTTCTTGAAGG; GTAAAAACACCTGACCGCTA; TTGGGTACTGTGGGTTTAGT;
CAACTCTCATTGCATGGGAT; AGCGATAACTTTCCCCACAA; AAACGTCTCTTCAGGCCACT;
ATTGTTATCAGCTATGCGCC; GGGAGTGAAAATTCCCCTAA; CGGTAAGTCGCATAAAAACC;
GAGCGCTTATCTTTCCCTTT; GCCAGCAGAGAATTAAGGAA; CAACCTGCAGGTGATGATTA;
CTGAACCAGACTCTTGTCAT; CAACTGAAGCTTTAGAGCGA; GCGAGGCTGTTCTTAATATC;
GGAATCCCAATGATTCGTCA; GTGTCGCCTTCAACAAACAA; CCGAAAACAGTTCTGGCAAA;
AATCTGTCAGATCGGATGTG; CCACTGCTTAATGACATTCC; ATCAGTGGCTCTATCTGAAC;
AACATCGTCTTTGGTGGTTC; TCGGCCGATGAAATGCATAT; TAGCTTGGCTTCTACCTTCA.

Sequences of *cro* probes (5' to 3'):

AGGGTTATGCGTTGTTCCAT; AAAGCGCATTGCATAATCTT; CTTTAGCTGTCTTGGTTTGC;
GCGCTTTGATATACGCCGAG; TGCATGAATGGCCTTGTTGA; TAGTTAAAAAAATCTTTCGG;
TAAACGCTTCCATCAGCGTT; GAAGGGCTTTACCTCTTCCG; CTGTTGTTTTTTTGTTACTC;
GTCACGACGTTGTAAAACGA; ATTAAGTTGGGTAACGCCAG; TATTACGCCAGCTGGCGAAA;
AGGTTACGTTGGTGTAGATG; AATGTGAGCGAGTAACAACC; GTAGCCAGCTTTCATCAACA;
AATAATTCGCGTCTGGCCTT; AGATGAAACGCCGAGTTAAC; AATTCAGACGGCAAACGACT;
TTTCTCCGGCGCGTAAAAAT; ATCTTCCAGATAACTGCCGT; AACGAGACGTCACGGAAAAT;
GCTGATTTGTGTAGTCGGTT; TTAAAGCGAGTGGCAACATG; AACTGTTACCCGTAGGTAGT;
ATAATTTCACCGCCGAAAGG; TTTCGACGTTCAGACGTAGT; ATAGAGATTCGGGATTTCGG;
ACCATTTTCAATCCGCACCT; TTAACGCCTCGAATCAGCAA; TTCATCAGCAGGATATCCTG;
CACGGCGTTAAAGTTGTTCT; TGGTTCGGATAATGCGAACA; TTGGCTTCATCCACCACATA;
AGACGATTCATTGGCACCAT; TGATCACACTCGGGTGATTA; ATTTGATCCAGCGATACAGC;
AAATAATATCGGTGGCCGTG; TTTGATGGACCATTTCGGCA; ATTTAGCGAAACCGCCAAGA;
AAACGGGGATACTGACGAAA; TTAATCAGCGACTGATCCAC; ATACAGAACTGGCGATCGTT;
AAACTGCTGCTGGTGTTTTG; TATTCGCTGGTCACTTCGAT; GTTATCGCTATGACGGAACA;
TTTACCTTGTGGAGCGACAT; GTTCAGGCAGTTCAATCAAC; AAATCCATTTCGCTGGTGGT;
TTGCCAACGCTTATTACCCA.

Sample fixation and permeabilization

An overnight culture was diluted 100- to 1000-fold into 30 ml of growth medium in a baffled flask. The culture was incubated at 37°C with shaking. When the OD_{600} of the culture reached 0.3–0.4, it was cooled in an ice-water bath. A volume with number of cells equivalent to 15 ml of OD_{600} = 0.4 was transferred to an ice-cold 50 ml Corning centrifuge tube and the cells were pelleted by centrifugation (5 minutes, 4500×*g*, 4°C). The supernatant was removed and the cells were resuspended in 1 ml freshly prepared 3.7% formaldehyde (Fisher, #BP531) in 1× PBS (diluted from 10× PBS, Ambion, #AM9625). The cells were then transferred to a microcentrifuge tube and mixed on a rotator at room temperature for 30 minutes. The cells were pelleted by centrifugation (8 minutes, 400×*g*). The supernatant was removed and the cells were washed in 1 ml 1× PBS twice (i.e. resuspended in 1 ml 1× PBS, centrifuged at 600×*g* for 3.5 minutes, and supernatant removed). The cells were resuspended in 300 µl water, and then 350 µl of 100% ethanol was added and mixed twice to get to a final concentration of 70% ethanol. The cells were left at room temperature with mixing on a rotator for at least one hour (or alternatively, at 4°C for at least a week) to permeabilize the cell membrane. The cells can be stored at 4°C for at least several weeks after this step.

I did not use lysozyme to treat the cells because the probes can get into cells without lysozyme treatment. Ethanol treatment alone suffices in this case. Also, lysozyme treatment would lead to bad morphology and insufficient contrast in phase contrast images of the cells, making automated cell recognition very unreliable, if not impossible.

Hybridization

After permeabilization, cells were centrifuged (7 minutes, 600×*g*) and the supernatant was removed. The cells were resuspended in 1 ml of 40% wash solution (see below) and the tube was left standing at room temperature for a few minutes. The cells were then centrifuged (7 minutes, 600×*g*) and the supernatant was removed. 3.3 µl of a ten-fold diluted probe solution was added to 50 µl of 40% hybridization solution. The cells were then resuspended in this hybridization mix and left at 30°C overnight. Hybridized samples could be stored at 4°C for at least six months.

A range of formamide concentrations was initially tested, and 40% gave the best results in that it is high enough so that background noise due to non-specific binding is low, while still low enough so that the fluorescence signal from target mRNA molecules is not impaired.

A range of probe solution amount to add to each hybridization reaction was initially tested for each lot of smFISH probes. The fluorescence signal would initially increase with the amount of probes added, and then reach a plateau. Three times as much probes as when the signal first reached the plateau was eventually used in actual experiments.

10 ml of 40% wash solution contains 4 g of formamide (Ambion, #AM9342) and 1 ml of 20× SSC (Ambion, #AM9763).

10 ml of 40% hybridization solution contains 1 g of dextran sulfate (Sigma, #D8906), 4 g of formamide, 10 mg of *E. coli* tRNA (Sigma, #R4251), 1 ml of 20× SSC, 40 µl of 50 mg/ml BSA (Ambion, #AM2616), and 100 µl of 200 mM ribonucleoside vanadyl complex (New England Biolabs, #S1402S). The solution was filter sterilized, aliquoted, and stored at -20°C.

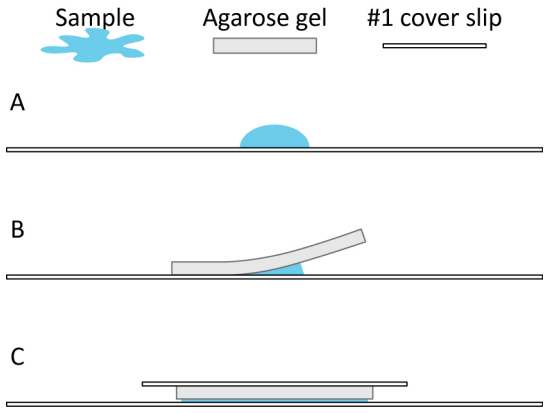


Figure B.2. Preparing sample for microscopy.

(A) 1 µl of sample was pipetted onto a 24 × 50 mm #1 coverslip.

(B) A 1 mm thick 1.5% agarose gel pad was laid on the sample.

(C) A 22 × 22 mm #1 coverslip was placed on top of the agarose gel pad.

Washing

10–25 µl of hybridized sample was transferred to a microcentrifuge tube. The rest was stored at 4°C. 200 µl of 40% wash solution was added to the tube and mixed well. Cells were pelleted by centrifugation (3.5 minutes, 600×*g*) and the supernatant was removed. The cells were washed three more times (i.e. resuspended in 200 µl of 40% wash solution, incubated at 30°C for 30 minutes, centrifuged at 600×*g* for 3.5 minutes, and supernatant removed). 4',6-diamidino-2-phenylindole (DAPI, Fisher Scientific, #PI-46190) was added to the wash solution to a final concentration of 10 µg/ml in the last wash. The cells were resuspended in 10 µl of 2× SSC and kept on ice.

If the cell density is too low, it would take a lot of time to image a sufficient number of cells. If the cell density is too high, automated image processing would start becoming unreliable. The volume of 2× SSC used for resuspending the cells could be increased or decreased to optimize the cell concentration for microscopy and automated image processing.

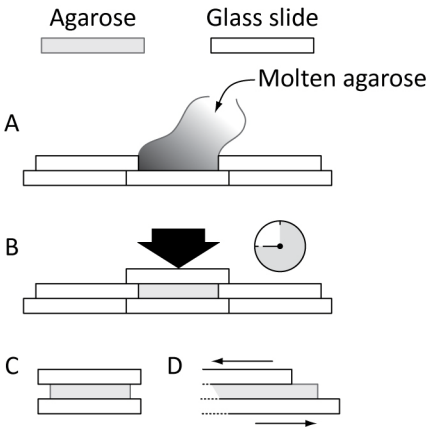


Figure B.3. Preparing agarose gel pad for microscopy.

(A) Molten 1.5% agarose is poured into a well created by an assembly of five glass microscope slides.

(B) The well is closed by a sixth slide. A small weight is placed on top. The agarose was let stand on the bench for about 45 minutes.

(C) The four slides on the sides are removed. Extra agarose outside the slides is trimmed away. The gel pad is stored between the two remaining slides, wrapped in food wrap.

(D) The two slides are slid against each other along their lengths to expose the gel pad. A piece of appropriate size is cut out with a sharp blade.

Microscopy

1 µl of sample was pipetted onto a 24 × 50 mm #1 coverslip (Fisher Scientific, #12-545F) (**Figure B.2A**). A 1 mm thick 1.5% agarose gel pad (in 1× PBS) (**Figure B.3**) was laid on the sample (**Figure B.2B**). A 22 × 22 mm #1 coverslip (Fisher Scientific, #12-545B) was placed on top of the agarose gel pad (**Figure B.2C**). The sample was imaged using an inverted epifluorescence microscope (Nikon Instruments Eclipse TE2000-E) and a cooled EMCCD camera (Photometrics Cascade 512). A 100× N.A. 1.40 oil

immersion phase contrast objective (Nikon Instruments Plan Apo 100× / 1.40 Oil) was used in conjunction with a 2.5× lens in front of the camera. The microscope and camera were controlled using the Metamorph software (Molecular Devices). TexasRed filter set (Nikon Instruments, #96365) was used for imaging mRNA tagged by 6-TAMRA-labeled smFISH probes, Cy5 filter set (Nikon Instruments, #96324) was used for imaging mRNA tagged by Cy5-labeled smFISH probes, and DAPI filter set (Nikon Instruments, #96310) was used to image DNA stained by DAPI. *z*-stacks with nine slices and 250 nm spacing were acquired for phase contrast and TexasRed images. Each sample was imaged at multiple locations to get a total of at least ~1000 cells.

B.2 Quantitative reverse transcription polymerase chain reaction

Sterile, nuclease-free, aerosol-barrier pipette tips were used in all procedures. Nuclease-free reagents were used whenever available. DEPC-treated water (Ambion, #AM9922) was used wherever the protocol calls for water.

Total RNA extraction

Cells were grown in the same conditions as in smFISH experiments. When the OD_{600} of the culture reached 0.3–0.4, a volume of culture with number of cells equivalent to 2 ml of OD_{600} = 0.3 was added to 2 volumes of RNAprotect Bacteria Reagent (Qiagen, #76506). Total RNA was then extracted using the RNeasy Mini Kit (Qiagen, #74104). The optional DNase I digestion step (RNase-free DNase Set, Qiagen, #79254) was included in the protocol, and the samples were treated with DNase I for one hour at room temperature. The eluate from the final step was passed through the spin column one additional time to increase total RNA concentration.

Each total RNA sample was diluted ten-fold (in a final buffer concentration of 10 mM Tris·Cl, pH 7.5 (Fisher Scientific, #BP1757)) and scanned using a spectrophotometer (Thermo Scientific Nanodrop 2000). The purity and concentration of the total RNA samples were estimated from the absorbance spectrum between 220 and 320 nm. The absorbance of RNA in solution depends strongly on the pH and ionic strength of the solution the RNA sample is in (Wilfinger, Mackey *et al.*, 1997), so it is very important to dilute the sample into proper buffer conditions before performing spectrophotometric measurements.

Part of each total RNA sample was used as template in qRT-PCR reactions, and the rest was stored at -80°C.

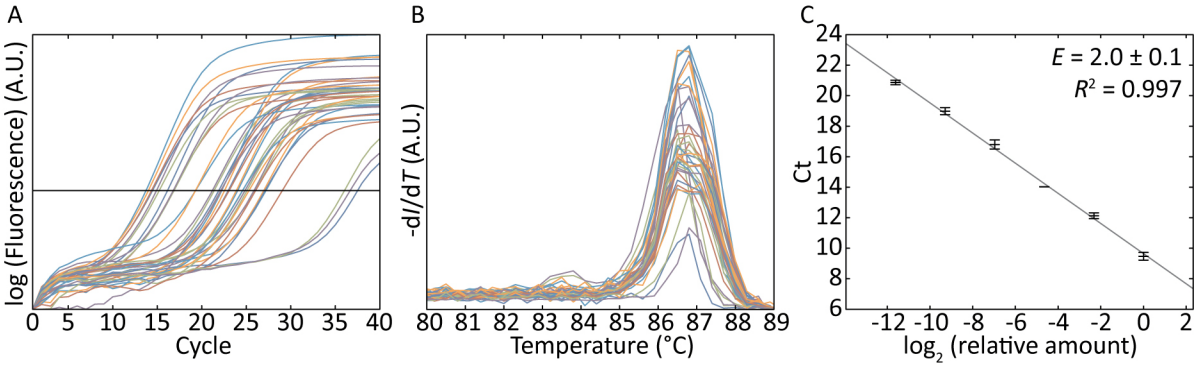


Figure B.4. Quality control in qRT-PCR experiments.

(A) SYBR Green fluorescence signal as a function of cycle number in an array of qRT-PCR reactions. The threshold (horizontal black line) is chosen at the exponential region of replication, and the Ct for a sample is the cycle number at the threshold.

(B) The first derivatives of melting curves of the final products of an array of qRT-PCR reactions. All samples give a peak at around 87°C.

(C) Ct versus relative amount for a dilution series (black points), with the fit (black line) giving the efficiency E of the PCR reaction.

qRT-PCR experiment

All total RNA samples were diluted to 50 μg/ml. HotStart-IT SYBR Green One-Step qRT-PCR Master Mix Kit (USB, #75770) was used. For each reaction, 10 μl of HotStart-IT SYBR Green[30] Master Mix (2×), 0.16 μl of MMLV-RT, 0.16 μl of RNase inhibitor, 1 μl of 10 μM forward primer, 1 μl of 10 μM reverse primer, 2 μl of 50 μg/ml total RNA sample, and 5.68 μl of RNase-free water were mixed on ice. The primers used for amplifying *lacZ* transcripts were 5′-ACGACATTGGCGTAAGTGAAGCG and 5′-CGTGAGCGGTCGTAATCAGC. The primers used for amplifying *cI* transcripts were 5′-CAACAGCCTGCTCAGGGTCAAC and 5′-GGTGATGCGGAGAGATGGGTAAGC. Primers were ordered from Integrated DNA Technologies[31]. White 8-tube strips (Bio-Rad, #TLS0851) and flat optical 8-cap strips (Bio-Rad, #TCS0803) were used. All reactions were carried out in duplicates. A dilution series of the sample with the highest expected expression level was included in the reactions for calibration. Negative controls with water instead of reverse transcriptase and/or RNA template were also included in the reactions to ensure the reactions were free of contaminants. Thermal cycles and fluorescence measurements were carried out in a MiniOpticon Real-Time PCR System (Bio-Rad). Details of thermal cycles are as follows: 50°C for 10 minutes; 95°C for 2 minutes; and then 40 cycles of 95°C for 15 seconds, 60°C for 30 seconds, 72°C for 40 seconds, followed by fluorescence measurement (**Figure B.4A**); and finally a melting curve from 90°C through 70°C at 0.3°C intervals and 1 second dwell time. After the protocol was completed, first derivatives of melting curves were visually inspected to ensure no unexpected product was present (**Figure B.4B**).

[30] SYBR Green is a fluorescence dye that fluoresces when bound to double-stranded DNA. SYBR Green fluorescence signal in a qRT-PCR reaction would be proportional to the amount of double-stranded DNA product present.

[31] PCR primers were ordered with a synthesis scale of 25 nmol, standard desalting purification, and as a lyophilized powder.

qRT-PCR data analysis

Amplification curves were viewed in semi-logarithmic scale. A fluorescence threshold was chosen at a level where amplification appeared exponential for every reaction (**Figure B.4A**). The Ct value for each reaction is the cycle number at which the fluorescence equals the threshold. Ct values of the dilution series were plotted against the logarithm (base 2) of the dilution ratios. The plot was fitted to a linear curve, and the efficiency E of the PCR reactions was calculated as $E = 2^{-1/\text{slope}}$ (**Figure B.4C**). The coefficient of determination R^2 of the linear fit was verified to be at least 0.99, and E was verified to be close to 2. Finally, the relative expression levels of the samples were estimated as $E^{-\text{Ct}}$.

RNA lifetime measurement

RNA lifetime measurements were carried out following the method of Bernstein *et al.* (Bernstein, Khodursky *et al.*, 2002). Cells were diluted from overnight cultures and grown in a 25 ml volume, under the same growth conditions as in smFISH experiments. When OD_{600} reached ~0.3, the culture was separated into two halves (~12 ml + 12 ml), each half being transferred to a new culture flask, and grown at the same conditions for 10–15 minutes in a water-bath shaker. 1.5 ml of culture was then extracted from each flask and mixed with 3 ml of Qiagen RNAprotect Bacterial reagent to stabilize cellular RNA (see **Total RNA extraction** above for details). These samples were treated as $t = 0$ samples. Rifampicin (Fisher Scientific, #BP26795) was then added to one of the remaining cultures, at a final concentration of 500 μg/ml, to inhibit transcription. The culture without rifampicin served as a control. 1.5 ml of culture was extracted from each flask at regular intervals after addition of rifampicin (e.g. at $t = 2, 4, 6, 8$ minutes) and mixed with 3 ml of Qiagen RNAprotect Bacterial reagent to stabilize cellular RNA. Subsequent total RNA extraction and qRT-PCR were performed as described previously in this section. Target RNA level of the rifampicin-treated sample as a function of time (in minutes) was fitted to an exponential, $\ln y = Ax + B$, and RNA lifetime (in minutes) estimated as $\tau_{\text{RNA}} = -1/A$.

B.3 Measurement of gene activity in individual live cells

Protocols and genetic constructs were based on Golding *et al.* (Golding and Cox, 2004; Golding and Cox, 2008; Golding, Paulsson *et al.*, 2005).

Strains and plasmid

The strain used was CZ071, harboring the two plasmids constructed by Golding *et al.* (Golding, Paulsson *et al.*, 2005) (**Figure 5.1**), pIG-K133(2cTG) and pIG-BAC($P_{lac/ara}$-mRFP1-96bs). CZ071 is wild type MG1655, into which the Z1 cassette (Lutz and Bujard, 1997) was integrated. pIG-K133(2cTG) is a ColE1-based plasmid, with the anhydrotetracycline (aTc) inducible promoter P_{LtetO1} controlling expression

of MS2-GFP. pIG-BAC($P_{lac/ara}$-mRFP1-96bs) is an F-based plasmid, with the IPTG and L-(+)-arabinose inducible promoter $P_{lac/ara}$ controlling expression of mRFP1 followed by an array of 96 MS2d binding sites.

Time-lapse microscopy

Cells were grown overnight in LB with the required antibiotics at 37°C with shaking, and diluted 1000-fold into LB with antibiotics and various amounts of IPTG and/or L-(+)-arabinose. Cells were grown at 37°C with shaking until OD_{600} reached about 0.2. aTc was then added to a final concentration of 10 ng/ml[32]. The cells were then grown at 27°C for one hour[33]. 1 μl of cell culture was pipetted onto a 24 × 50 mm #1 coverslip (**Figure B.2A**). A 1 mm thick 1.5% agarose gel pad (in LB, with inducers[34]) (**Figure B.3**) was laid on the sample (**Figure B.2B**). A 22 × 22 mm #1 coverslip was placed on top of the agarose gel pad (**Figure B.2C**). A time-lapse movie of the growing cells was recorded using an inverted epifluorescence microscope (Nikon Instruments Eclipse TE2000-E) and a cooled EMCCD camera (Photometrics Cascade 512). A 100× N.A. 1.40 oil immersion phase contrast objective (Nikon Instruments Plan Apo 100× / 1.40 Oil) was used in conjunction with a 2.5× lens in front of the camera. The microscope and camera were controlled using the Metamorph software (Molecular Devices). TexasRed filter set (Nikon Instruments, #96365) was used for imaging mRFP1. GFP filter set (Nikon Instruments, #96362) was used to image MS2-GFP. Nine *z*-stacks at 250 nm spacing were obtained for phase contrast images. The cells were imaged at two-minute intervals.

Image processing and data analysis

Images from time-lapse movies were processed and analyzed using MATLAB programs. Cell recognition was performed on phase contrast images of cells using the *Schnitzcell* MATLAB module (gift of Michael Elowitz, California Institute of Technology) as described in **Section 2.2**. Cell lineages were reconstructed using a cell tracking program developed in our lab. RNA signals (localized MS2-GFP fluorescence) at each time point were detected and quantified using a spot recognition program developed in our lab as described in **Appendix C.1**. RNA time series of each cell lineage was analyzed by a step detection program developed in our lab, as described in **Appendix C.2**. Quantities such as gene "on" and "off" times, average mRNA per cell, and transcription burst sizes were then extracted.

[32] A range of aTc concentrations were initially tested, and 10 ng/ml gave the best consistency. Cells became unhealthy at higher concentrations because of the toxicity of aTc, while lower concentrations were not able to give enough MS2-GFP.

[33] The temperature shift gives cells time to adapt to the lower temperature, because time-lapse microscopy was done at ambient temperature, which is about 27°C.

[34] Agarose in LB is melted by heating, cooled to 55°C, and 10 ng/ml aTc and the same concentrations of IPTG and L-(+)-arabinose as in the culture were added to the molten agarose.

Appendix C. Data Analysis

All image processing and data analysis were performed using MATLAB (MathWorks).

C.1 Spot recognition

A spot recognition program developed in our lab was used to automatically identify and quantify localized fluorescence signals (**Figure 2.3B**). A Gaussian filter was first applied to smooth out noise. The smoothed image was only used for spot recognition, but not quantification. Peaks in both x- and y-directions were then located. The threshold of the peak height, below which peaks were disregarded, is a parameter chosen by the user. Spots were recognized as where peaks in both x- and y-directions are present, and the spots were allocated into cells based on spot positions and the label matrix from the cell recognition program *Schnitzcell*.

The threshold peak height was chosen to be low enough so that most spots were recognized in positive samples and still high enough so that very few spots were recognized in negative controls.

Total intensity of each spot was measured by summing the intensities of all pixels inside a circle centered at the spot. The radius of such circles is a parameter chosen by the user. The background intensity per pixel was then calculated as the median intensity of background pixels. Background pixels are pixels within the boundary of the cell the spot was located in, minus pixels covered by the circles around spots. The total background intensity was then subtracted from the total intensity of each spot. Other quantities such as fluorescence peak height, position of spot relative to the cell major and minor axes, and number of recognized spots in each cell were also measured and recorded.

In smFISH experiments, spot recognition was done at each z-position in the stack of images. Each spot was traced along z based on position, and was quantified at the z where it had the highest fluorescence intensity (where the spot is in best focus).

C.2 Step detection

Transcription is a discrete event, where an integer number of mRNA molecules are produced in an "on" period of a gene. Ideally, mRNA trajectories should appear as discrete, constant mRNA levels connected by steps—the constant levels correspond to "off" periods and the steps correspond to transcription events during "on" periods (black line in **Figure 5.2B**). However, the live-cell data I collected from time-lapse movies is inevitably noisy. To extract useful information from the noisy data, I need to reconstruct the discrete mRNA levels and the steps connecting them.

A step detection program was developed to serve this purpose. The program was adapted from an algorithm by Sheyum Syed (Selvin lab, University of Illinois at Urbana-Champaign). The raw data was first smoothed (**Figure C.1A**) using a moving average filter with a span of five time points. Then at every time point T, a two-sample t-test was performed between two goups of data points: a group consisting of six time points immediately before T and the time point at T, and a group consisting of the time point at T and six time points immediately after T. This would yeild a p-value for every time point (**Figure C.1B**). If I now specify a p-value threshold, steps would occur in the mRNA trajectory where the p-value is lower than this threshold (**Figure C.1B**). A range of values for the p-value threshold was scanned, and the threshold giving the smallest χ^2 between the discretized mRNA trajectory and the raw data was used.

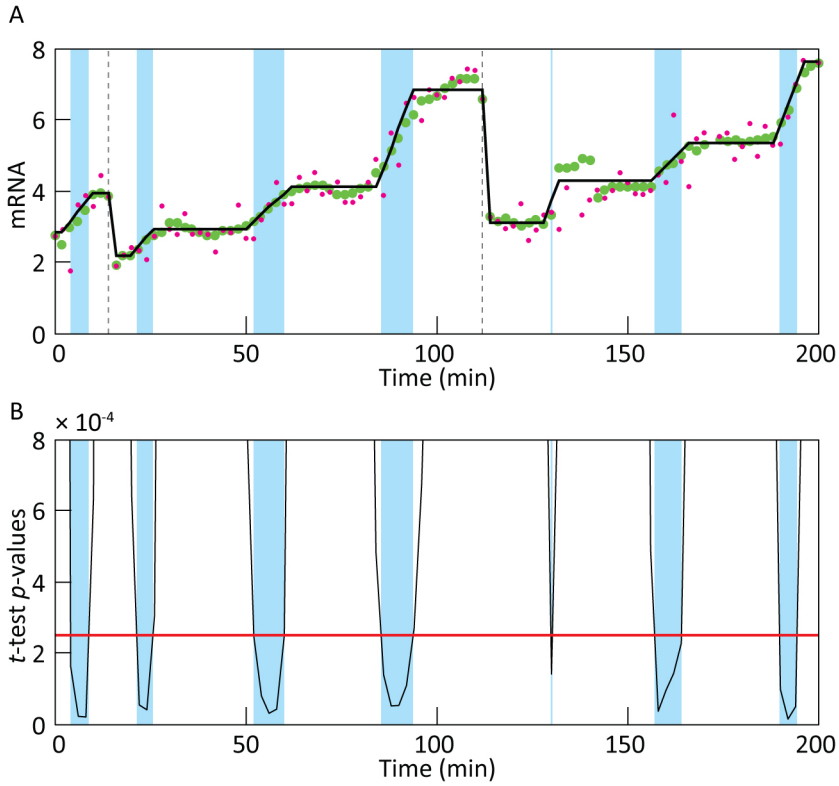


Figure C.1. Step detection algorithm based on t-test.

(A) This is the same data set shown in **Figure 5.2B**. The raw data (red dots) was first smoothed (green circles). A t-test was then performed on the smoothed data, and steps in the trajectory were recognized as places where the resulting p-value is below a threshold (light blue regions). The black line represents the resulting discretized trajectory, with constant mRNA levels when the p-value is above the threshold, and steps connecting the constant levels when the p-value is below the threshold. The grey vertical dashed lines indicate cell divisions.

(B) t-test p-values obtained from the data in panel (A). The red horizontal line is the threshold that minimizes the χ^2 between the discretized mRNA trajectory (black line in panel (A)) and the raw data (red dots in panel (A)). Steps occur when the p-value is lower than this threshold (light blue regions).

Appendix D. Modeling Transcription Kinetics

D.1 Gillespie simulation of the two-state model

The Gillespie algorithm (Gillespie, 1977) was used to simulate the stochastic kinetics of the two-state model (**Chapter 1** and **Figure 1.5A**) (Raj, Peskin *et al.*, 2006; Shahrezaei and Swain, 2008), which can give rise to bursty transcriptional time-series. In this model, each copy of a gene can either be in an active ("on") or inactive ("off") state. It switches from "on" state to "off" state with rate k_{off}, and from "off" state to "on" state with rate k_{on}. When it is "on", mRNA is produced at a rate k_{TX}. mRNA is degraded at a rate k_{d}. Probability per unit time of each reaction occurring was calculated from these reaction rates, weighted by the number of reactant species present, and the reaction trajectory was simulated stochastically.

Cell division and gene replication were incorporated as optional components in the simulations. Reactant species segregate according to binomial statistics upon cell division. Gene copy number doubles at a time specified by the user. To mimic FISH experiments, a random time point was chosen from each of 1000 trajectories at which number of mRNA is "measured".

D.2 Simultaneous modulation of all three kinetic parameters

Here I consider a modulation scheme which allows all three kinetic parameters, k_{on}, k_{off}, and k_{TX} to change. This scenario does not yield a single trajectory in k-space. Instead, an ensemble of possible trajectories can be defined. To form this ensemble, I first approximate the experimental $\sigma^2/\langle n\rangle$ versus $\langle n\rangle$ data using a Hill function (**Figure D.1A**, black line). Each point $(\sigma^2, \langle n\rangle)$ defines a curve in $(k_{\text{on}}, k_{\text{off}}, k_{\text{TX}})$ space. Thus, each colored marker in **Figure D.1A** corresponds to a curve of the same color in **Figure D.1B**. From **Equations (3.2)** and **(3.3)**, I can write down the equations of these curves with k_{TX} as a parameter:

$$\mathbf{k} = \begin{pmatrix} \frac{k_{\text{d}}\langle n\rangle}{k_{\text{TX}}}\left(\frac{\langle n\rangle(k_{\text{TX}}-k_{\text{d}}\langle n\rangle)}{\sigma^2-\langle n\rangle} - k_{\text{d}}\right) \\ \left(\frac{k_{\text{d}}\langle n\rangle}{k_{\text{TX}}} - 1\right)\left(k_{\text{d}} + \frac{(k_{\text{d}}\langle n\rangle - k_{\text{TX}})\langle n\rangle}{\sigma^2-\langle n\rangle}\right) \\ k_{\text{TX}} \end{pmatrix}. \tag{D.1}$$

The direction tangent to this line at a given k_{TX} is given by the derivative

$$\mathbf{v} = \frac{\partial \mathbf{k}}{\partial k_{\mathrm{TX}}} = \begin{pmatrix} \frac{k_{\mathrm{d}}^2\langle n\rangle}{k_{\mathrm{TX}}^2}\left(\frac{\langle n\rangle^2}{\sigma^2-\langle n\rangle}+1\right) \\ \frac{\langle n\rangle}{\sigma^2-\langle n\rangle} - \frac{k_{\mathrm{d}}^2\langle n\rangle}{k_{\mathrm{TX}}^2}\left(\frac{\langle n\rangle^2}{\sigma^2-\langle n\rangle}+1\right) \\ 1 \end{pmatrix}. \quad \text{(D.2)}$$

To go from a certain point along $\mathbf{k}$ given by $(\sigma^2, \langle n\rangle)$ and k_{TX} to a point along an adjacent line $\mathbf{k}'$ given by $(\sigma'^2, \langle n'\rangle)$ and k'_{TX}, such that the line joining these two points is perpendicular to $\mathbf{v}$, I require $(\mathbf{k}' - \mathbf{k}) \cdot \mathbf{v} = 0$. This leads to

$$Ak'^2_{\mathrm{TX}} + \mathrm{B}k'_{\mathrm{TX}} + \mathrm{C} = 0, \quad \text{(D.3)}$$

where

$$A = \frac{\langle n'\rangle}{\sigma'^2-\langle n'\rangle}v_2 + v_3,$$

$$B = \frac{k_{\mathrm{d}}\langle n'\rangle^2}{\sigma'^2-\langle n'\rangle}(v_1 - 2v_2) - k_{\mathrm{d}}v_2 - \mathbf{k}\cdot\mathbf{v},$$

$$C = k_{\mathrm{d}}\langle n'\rangle\left(1 + \frac{\langle n'\rangle^2}{\sigma'^2-\langle n'\rangle}\right)(v_2 - v_1),$$

and v_1, v_2, and v_3 are the three components of $\mathbf{v}$. If I start at a certain point along $\mathbf{k}$ given by $(\sigma^2, \langle n\rangle)$ and $k_{\mathrm{TX},i}$, and move to the adjacent line sequentially in this manner, I obtain a trajectory in $(k_{\mathrm{on}}, k_{\mathrm{off}}, k_{\mathrm{TX}})$ space (**Figure D.1B**, black lines). Such a trajectory gives a modulation scheme in which k_{on}, k_{off}, and k_{TX} are all changed, as shown in **Figure D.1C**.

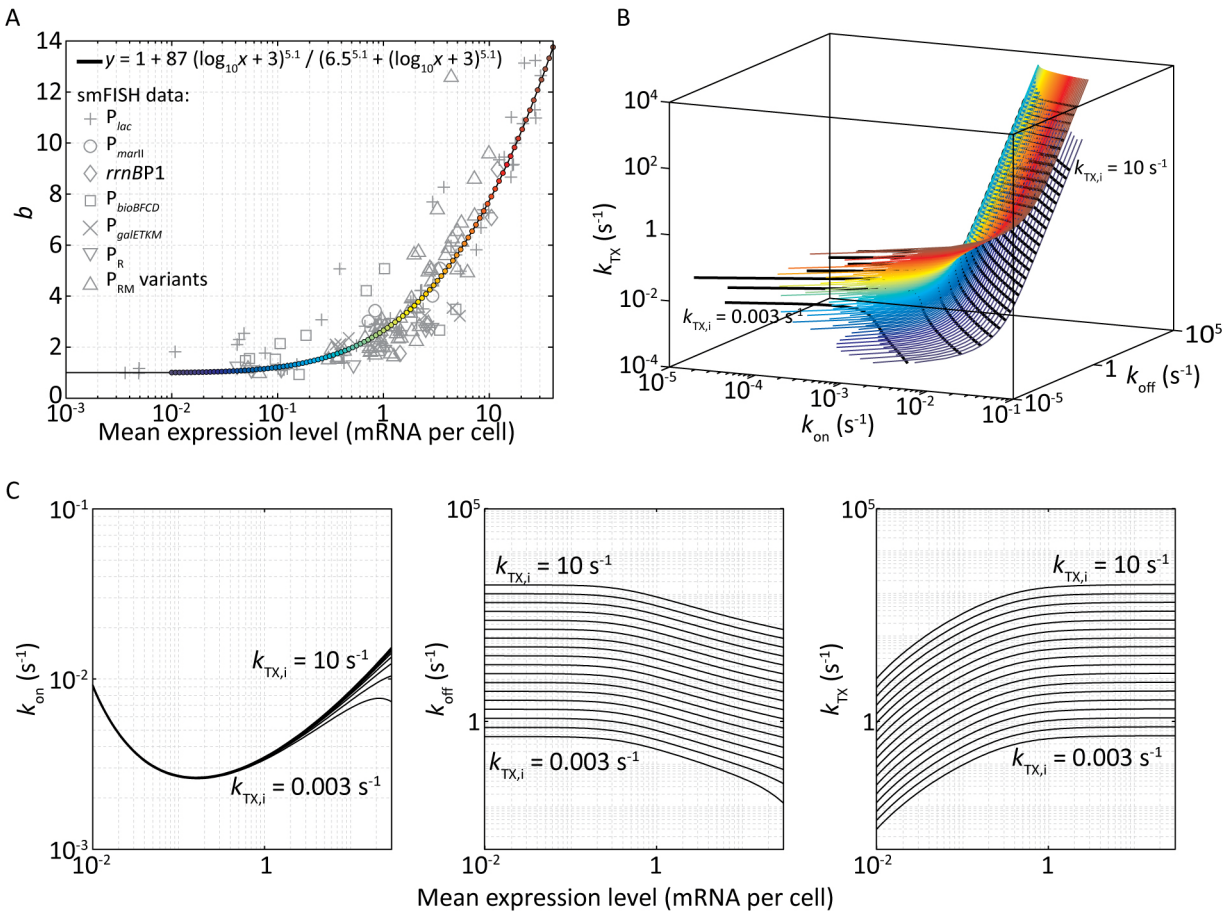


Figure D.1. Simultaneous modulation of all three kinetic parameters.

(A) b as a function of the mean expression level $\langle n \rangle$. The data (grey markers) is reproduced from **Figure 4.1B**. The data set is approximated by a Hill function (black line). Each point $(\sigma^2, \langle n \rangle)$ along the Hill curve (colored markers) defines a curve in $(k_{on}, k_{off}, k_{TX})$ space (panel B, curves of the same color).

(B) An ensemble of trajectories in k-space yielding the observed $(\sigma^2, \langle n \rangle)$ behavior. Each colored curve corresponds to a marker of the same color in panel (A). The black lines are trajectories traversing the colored lines, such that the trajectory is always perpendicular to a colored line when they intersect. Each black trajectory arises from a choice of an initial k_{TX} ranging from 0.003 s^{-1} to 10 s^{-1}.

(C) k_{on}, k_{off}, and k_{TX} as a function of $\langle n \rangle$ given by each of the black trajectories in panel (B).

REFERENCES

Abdel-Hamid, A. M. and J. E. Cronan (2007). "Coordinate expression of the acetyl coenzyme A carboxylase genes, accB and accC, is necessary for normal regulation of biotin synthesis in Escherichia coli." *J Bacteriol* **189**(2): 369-76.

Abramowitz, M. and I. Stegun (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Dover Publications, New York.

Alberts, B. (2008). *Molecular biology of the cell. Reference edition*. New York, Garland Science.

Alekshun, M. N. and S. B. Levy (1999). "The mar regulon: multiple resistance to antibiotics and other toxic chemicals." *Trends Microbiol* **7**(10): 410-3.

Alon, U. (2007). *An introduction to systems biology: design principles of biological circuits*. Boca Raton, FL, Chapman & Hall/CRC.

Arraiano, C. M., J. M. Andrade, S. Domingues, I. B. Guinote, M. Malecki, R. G. Matos, R. N. Moreira, V. Pobre, F. P. Reis, M. Saramago, I. J. Silva and S. C. Viegas (2010). "The critical role of RNA processing and degradation in the control of gene expression." *FEMS Microbiol Rev* **34**(5): 883-923.

Austin, D. W., M. S. Allen, J. M. McCollum, R. D. Dar, J. R. Wilgus, G. S. Sayler, N. F. Samatova, C. D. Cox and M. L. Simpson (2006). "Gene network shaping of inherent noise spectra." *Nature* **439**(7076): 608-11.

Ausubel, F. M. (1987). *Current protocols in molecular biology*. Brooklyn, N. Y., John Wiley & Sons.

Bar-Even, A., J. Paulsson, N. Maheshri, M. Carmi, E. O'Shea, Y. Pilpel and N. Barkai (2006). "Noise in protein expression scales with natural protein abundance." *Nat Genet* **38**(6): 636-43.

Barker, D. F. and A. M. Campbell (1980). "Use of bio-lac fusion strains to study regulation of biotin biosynthesis in Escherichia coli." *J Bacteriol* **143**(2): 789-800.

Bernstein, J. A., A. B. Khodursky, P. H. Lin, S. Lin-Chao and S. N. Cohen (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): 9697-702.

Bertrand, E., P. Chartrand, M. Schaefer, S. M. Shenoy, R. H. Singer and R. M. Long (1998). "Localization of ASH1 mRNA particles in living yeast." *Mol Cell* **2**(4): 437-45.

Blake, W. J., K. A. M, C. R. Cantor and J. J. Collins (2003). "Noise in eukaryotic gene expression." *Nature* **422**(6932): 633-7.

Blattner, F. R., G. Plunkett, 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau and Y. Shao (1997). "The complete genome sequence of Escherichia coli K-12." *Science* **277**(5331): 1453-62.

Bremer, H. and G. Churchward (1977). "An examination of the Cooper-Helmstetter theory of DNA replication in bacteria and its underlying assumptions." *J Theor Biol* **69**(4): 645-54.

Bustin, S. A. and T. Nolan (2004). "Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction." *J Biomol Tech* **15**(3): 155-66.

Cai, L., N. Friedman and X. S. Xie (2006). "Stochastic protein expression in individual cells at the single molecule level." *Nature* **440**(7082): 358-62.

Campbell, N. and J. Reece (2005). *Biology*, Pearson, Benjamin Cummings.

Campbell, R. E., O. Tour, A. E. Palmer, P. A. Steinbach, G. S. Baird, D. A. Zacharias and R. Y. Tsien (2002). "A monomeric red fluorescent protein." *Proc Natl Acad Sci U S A* **99**(12): 7877-82.

Chubb, J. R., T. Trcek, S. M. Shenoy and R. H. Singer (2006). "Transcriptional pulsing of a developmental gene." *Curr Biol* **16**(10): 1018-25.

Chubb, J. R. and T. B. Liverpool (2010). "Bursts and pulses: insights from single cell studies into transcriptional mechanisms." *Curr Opin Genet Dev* **20**(5): 478-84.

Coburn, G. A. and G. A. Mackie (1999). "Degradation of mRNA in Escherichia coli: an old problem with some new twists." *Prog Nucleic Acid Res Mol Biol* **62**: 55-108.

Cormack, B. P., R. H. Valdivia and S. Falkow (1996). "FACS-optimized mutants of the green fluorescent protein (GFP)." *Gene* **173**(1 Spec No): 33-8.

Cox, C. D., J. M. McCollum, M. S. Allen, R. D. Dar and M. L. Simpson (2008). "Using noise to probe and characterize gene circuits." *Proc Natl Acad Sci U S A* **105**(31): 10809-14.

Crick, F. (1970). "Central dogma of molecular biology." *Nature* **227**(5258): 561-3.

Datsenko, K. A. and B. L. Wanner (2000). "One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products." *Proc Natl Acad Sci U S A* **97**(12): 6640-5.
Dobrzynski, M. and F. J. Bruggeman (2009). "Elongation dynamics shape bursty transcription and translation." *Proc Natl Acad Sci U S A*.
Elowitz, M. B., A. J. Levine, E. D. Siggia and P. S. Swain (2002). "Stochastic gene expression in a single cell." *Science* **297**(5584): 1183-6.
Fang, F. C. (2005). "Sigma cascades in prokaryotic regulatory networks." *Proc Natl Acad Sci U S A* **102**(14): 4933-4.
Femino, A. M., F. S. Fay, K. Fogarty and R. H. Singer (1998). "Visualization of single RNA transcripts in situ." *Science* **280**(5363): 585-90.
Fusco, D., N. Accornero, B. Lavoie, S. M. Shenoy, J. M. Blanchard, R. H. Singer and E. Bertrand (2003). "Single mRNA molecules demonstrate probabilistic movement in living mammalian cells." *Curr Biol* **13**(2): 161-7.
Garcia, H. G., A. Sanchez, T. Kuhlman, J. Kondev and R. Phillips (2010). "Transcription by the numbers redux: experiments and calculations that surprise." *Trends Cell Biol*.
Geva-Zatorsky, N., E. Dekel, A. A. Cohen, T. Danon, L. Cohen and U. Alon (2010). "Protein dynamics in drug combinations: a linear superposition of individual-drug responses." *Cell* **140**(5): 643-51.
Gillespie, D. T. (1977). "Exact Stochastic Simulation of Coupled Chemical-Reactions." *Journal of Physical Chemistry* **81**(25): 2340-2361.
Goh, K. I. and A. Barabasi (2008). "Burstiness and memory in complex systems." *Epl* **81**(4): -.
Golding, I. and E. C. Cox (2004). "RNA dynamics in live Escherichia coli cells." *Proc Natl Acad Sci U S A* **101**(31): 11310-5.
Golding, I., J. Paulsson, S. M. Zawilski and E. C. Cox (2005). "Real-time kinetics of gene activity in individual bacteria." *Cell* **123**(6): 1025-36.
Golding, I. and E. C. Cox (2006). "Eukaryotic transcription: what does it mean for a gene to be 'on'?" *Curr Biol* **16**(10): R371-3.
Golding, I. and E. C. Cox (2008). "Chapter 8: Spatiotemporal dynamics in bacterial cells: real-time studies with single-event resolution." *Methods Cell Biol* **89**: 223-51.
Hernandez, V. J. and H. Bremer (1990). "Guanosine tetraphosphate (ppGpp) dependence of the growth rate control of rrnB P1 promoter activity in Escherichia coli." *J Biol Chem* **265**(20): 11605-14.
N. P. Higgins, Ed. (2005). *The bacterial chromosome*. Washington, D.C., ASM Press.
Hobman, J. L., C. W. Penn and M. J. Pallen (2007). "Laboratory strains of Escherichia coli: model citizens or deceitful delinquents growing old disgracefully?" *Mol Microbiol* **64**(4): 881-5.
Iost, I. and M. Dreyfus (1995). "The stability of Escherichia coli lacZ mRNA depends upon the simultaneity of its synthesis and translation." *EMBO J* **14**(13): 3252-61.
Kaern, M., T. C. Elston, W. J. Blake and J. J. Collins (2005). "Stochasticity in gene expression: from theories to phenotypes." *Nat Rev Genet* **6**(6): 451-64.
Kennell, D. and H. Riezman (1977). "Transcription and translation initiation frequencies of the Escherichia coli *lac* operon." *J Mol Biol* **114**(1): 1-21.
Kepler, T. B. and T. C. Elston (2001). "Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations." *Biophys J* **81**(6): 3116-36.
Kittisopikul, M. and G. M. Suel (2010). "Biological role of noise encoded in a genetic network motif." *Proc Natl Acad Sci U S A* **107**(30): 13300-5.
Kuhlman, T., Z. Zhang, M. H. Saier, Jr. and T. Hwa (2007). "Combinatorial transcriptional control of the lactose operon of Escherichia coli." *Proc Natl Acad Sci U S A* **104**(14): 6043-8.
Liang, S., M. Bipatnath, Y. Xu, S. Chen, P. Dennis, M. Ehrenberg and H. Bremer (1999). "Activities of constitutive promoters in Escherichia coli." *J Mol Biol* **292**(1): 19-37.
Lim, W. A. and R. T. Sauer (1989). "Alternative packing arrangements in the hydrophobic core of lambda repressor." *Nature* **339**(6219): 31-6.
Little, J. W., D. P. Shepley and D. W. Wert (1999). "Robustness of a gene regulatory circuit." *EMBO J* **18**(15): 4299-307.
Lutz, R. and H. Bujard (1997). "Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements." *Nucleic Acids Res* **25**(6): 1203-10.
Mao, C., C. R. Brown, E. Falkovskaia, S. Dong, E. Hrabeta-Robinson, L. Wenger and H. Boeger (2010). "Quantitative analysis of the transcription control mechanism." *Mol Syst Biol* **6**: 431.

Michalowski, C. B. and J. W. Little (2005). "Positive autoregulation of cI is a dispensable feature of the phage lambda gene regulatory circuitry." *J Bacteriol* **187**(18): 6430-42.
Miller, J. H. (1992). *A short course in bacterial genetics : a laboratory manual and handbook for Escherichia coli and related bacteria*. Plainview, N.Y., Cold Spring Harbor Laboratory Press.
Mitarai, N., I. B. Dodd, M. T. Crooks and K. Sneppen (2008). "The generation of promoter-mediated transcriptional noise in bacteria." *PLoS Comput Biol* **4**(7): e1000109.
Mooney, R. A., S. E. Davis, J. M. Peters, J. L. Rowland, A. Z. Ansari and R. Landick (2009). "Regulator trafficking on bacterial transcription units in vivo." *Mol Cell* **33**(1): 97-108.
Müller-Hill, B. (1996). *The lac Operon : a short history of a genetic paradigm*. Berlin ; New York, Walter de Gruyter.
Neidhardt, F., J. Ingraham and M. Schaechter (1990). *Physiology of the bacterial cell: a molecular approach*, Sinauer Associates.
Neidhardt, F. C. (1987). *Escherichia coli and Salmonella typhimurium : cellular and molecular biology*. Washington, D.C., American Society for Microbiology.
Newman, J. R., S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi and J. S. Weissman (2006). "Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise." *Nature* **441**(7095): 840-6.
Nolan, T., R. E. Hands and S. A. Bustin (2006). "Quantification of mRNA using real-time RT-PCR." *Nat Protoc* **1**(3): 1559-82.
Pare, A., D. Lemons, D. Kosman, W. Beaver, Y. Freund and W. McGinnis (2009). "Visualization of individual Scr mRNAs during Drosophila embryogenesis yields evidence for transcriptional bursting." *Curr Biol* **19**(23): 2037-42.
Paulsson, J. and M. Ehrenberg (2000). "Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks." *Phys Rev Lett* **84**(23): 5447-50.
Paulsson, J. (2004). "Summing up the noise in gene networks." *Nature* **427**(6973): 415-8.
Peabody, D. S. and F. Lim (1996). "Complementation of RNA binding site mutations in MS2 coat protein heterodimers." *Nucleic Acids Res* **24**(12): 2352-9.
Peabody, D. S. (1997). "Subunit fusion confers tolerance to peptide insertions in a virus coat protein." *Arch Biochem Biophys* **347**(1): 85-92.
Peccoud, J. and B. Ycart (1995). "Markovian Modeling of Gene-Product Synthesis." *Theoretical Population Biology* **48**(2): 222-234.
Pedersen, S. and S. Reeh (1978). "Functional mRNA half lives in E. coli." *Mol Gen Genet* **166**(3): 329-36.
Pedraza, J. M. and A. van Oudenaarden (2005). "Noise propagation in gene networks." *Science* **307**(5717): 1965-9.
Potrykus, K., D. Vinella, H. Murphy, A. Szalewska-Palasz, R. D'Ari and M. Cashel (2006). "Antagonistic regulation of Escherichia coli ribosomal RNA rrnB P1 promoter activity by GreA and DksA." *J Biol Chem* **281**(22): 15238-48.
Proux, F. and M. Dreyfus (2008). "Construction and characterization of E. coli K12 strains in which the transcription of selected genes is desynchronized from translation." *Methods Enzymol* **447**: 243-58.
Ptashne, M. (2004). *A genetic switch : phage lambda revisited*. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.
Raj, A., C. S. Peskin, D. Tranchina, D. Y. Vargas and S. Tyagi (2006). "Stochastic mRNA synthesis in mammalian cells." *PLoS Biol* **4**(10): e309.
Raj, A., P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden and S. Tyagi (2008). "Imaging individual mRNA molecules using multiple singly labeled probes." *Nat Methods* **5**(10): 877-9.
Reppas, N. B., J. T. Wade, G. M. Church and K. Struhl (2006). "The transition between transcriptional initiation and elongation in E. coli is highly variable and often rate limiting." *Mol Cell* **24**(5): 747-57.
Rosenfeld, N., J. W. Young, U. Alon, P. S. Swain and M. B. Elowitz (2005). "Gene regulation at the single-cell level." *Science* **307**(5717): 1962-5.
Rosenfeld, N., T. J. Perkins, U. Alon, M. B. Elowitz and P. S. Swain (2006). "A fluctuation method to quantify in vivo fluorescence data." *Biophys J* **91**(2): 759-66.
Sambrook, J. and D. W. Russell (2001). *Molecular cloning : a laboratory manual*. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.
Sanchez, A. and J. Kondev (2008). "Transcriptional control of noise in gene expression." *Proc Natl Acad Sci U S A* **105**(13): 5081-6.

Sanchez, A., H. G. Garcia, D. Jones, R. Phillips and J. Kondev (2011). "Effect of Promoter Architecture on the Cell-to-Cell Variability in Gene Expression." *PLoS Comput Biol* **7**(3): e1001100.
Sauer, R. T., S. R. Jordan and C. O. Pabo (1990). "Lambda repressor: a model system for understanding protein-DNA interactions and protein stability." *Adv Protein Chem* **40**: 1-61.
Schmittgen, T. D. and K. J. Livak (2008). "Analyzing real-time PCR data by the comparative C(T) method." *Nat Protoc* **3**(6): 1101-8.
Selinger, D. W., R. M. Saxena, K. J. Cheung, G. M. Church and C. Rosenow (2003). "Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation." *Genome Res* **13**(2): 216-23.
Shahrezaei, V. and P. S. Swain (2008). "Analytical distributions for stochastic gene expression." *Proc Natl Acad Sci U S A* **105**(45): 17256-61.
Shen-Orr, S. S., R. Milo, S. Mangan and U. Alon (2002). "Network motifs in the transcriptional regulation network of Escherichia coli." *Nat Genet* **31**(1): 64-8.
Simpson, M. L., C. D. Cox and G. S. Sayler (2004). "Frequency domain chemical Langevin analysis of stochasticity in gene transcriptional regulation." *J Theor Biol* **229**(3): 383-94.
Singh, D., S. J. Chang, P. H. Lin, O. V. Averina, V. R. Kaberdin and S. Lin-Chao (2009). "Regulation of ribonuclease E activity by the L4 ribosomal protein of Escherichia coli." *Proc Natl Acad Sci U S A* **106**(3): 864-9.
Svenningsen, S. L., N. Costantino, D. L. Court and S. Adhya (2005). "On the role of Cro in lambda prophage induction." *Proc Natl Acad Sci U S A* **102**(12): 4465-9.
Swain, P. S., M. B. Elowitz and E. D. Siggia (2002). "Intrinsic and extrinsic contributions to stochasticity in gene expression." *Proc Natl Acad Sci U S A* **99**(20): 12795-800.
Tan, R. Z. and A. van Oudenaarden "Transcript counting in single cells reveals dynamics of rDNA transcription." *Mol Syst Biol* **6**: 358.
Taniguchi, Y., P. J. Choi, G. W. Li, H. Chen, M. Babu, J. Hearn, A. Emili and X. S. Xie (2010). "Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells." *Science* **329**(5991): 533-8.
Thattai, M. and A. van Oudenaarden (2001). "Intrinsic noise in gene regulatory networks." *Proc Natl Acad Sci U S A* **98**(15): 8614-9.
Tokeson, J. P., S. Garges and S. Adhya (1991). "Further inducibility of a constitutive system: ultrainduction of the gal operon." *J Bacteriol* **173**(7): 2319-27.
van Zon, J. S., M. J. Morelli, S. Tanase-Nicola and P. R. ten Wolde (2006). "Diffusion of transcription factors can drastically enhance the noise in gene expression." *Biophys J* **91**(12): 4350-67.
Weickert, M. J. and S. Adhya (1993). "The galactose regulon of Escherichia coli." *Mol Microbiol* **10**(2): 245-51.
Whitelaw, N. C., S. Chong and E. Whitelaw (2010). "Tuning in to noise: epigenetics and intangible variation." *Dev Cell* **19**(5): 649-50.
Wilfinger, W. W., K. Mackey and P. Chomczynski (1997). "Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity." *Biotechniques* **22**(3): 474-6, 478-81.
Yakovleva, G. M., S. K. Kim and B. L. Wanner (1998). "Phosphate-independent expression of the carbon-phosphorus lyase activity of Escherichia coli." *Appl Microbiol Biotechnol* **49**(5): 573-8.
Yu, J., J. Xiao, X. Ren, K. Lao and X. S. Xie (2006). "Probing gene expression in live cells, one protein molecule at a time." *Science* **311**(5767): 1600-3.
Yunger, S., L. Rosenfeld, Y. Garini and Y. Shav-Tal (2010). "Single-allele analysis of transcription kinetics in living mammalian cells." *Nat Methods* **7**(8): 631-3.
Zenklusen, D., D. R. Larson and R. H. Singer (2008). "Single-RNA counting reveals alternative modes of gene expression in yeast." *Nat Struct Mol Biol* **15**(12): 1263-71.
Zong, C., L.-h. So, L. A. Sepulveda, S. O. Skinner and I. Golding (2010). "Lysogen stability is determined by the frequency of activity bursts from the fate-determining gene." *Mol Syst Biol* **6**: 440.