

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Investigating the Dynamics of Non-Equilibrium Behavior in Eukaryotic Transcriptional Regulation

Permalink

<https://escholarship.org/uc/item/9x90519z>

Author

Liu, Jonathan

Publication Date

2021

Peer reviewed|Thesis/dissertation

Investigating the Dynamics of Non-Equilibrium Behavior in Eukaryotic Transcriptional
Regulation

by

Jonathan Liu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Physics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Hernan Garcia, Chair

Professor Ahmet Yildiz

Professor Yun S. Song

Summer 2021

Investigating the Dynamics of Non-Equilibrium Behavior in Eukaryotic Transcriptional
Regulation

Copyright 2021
by
Jonathan Liu

Abstract

Investigating the Dynamics of Non-Equilibrium Behavior in Eukaryotic Transcriptional Regulation

by

Jonathan Liu

Doctor of Philosophy in Physics

University of California, Berkeley

Professor Hernan Garcia, Chair

Biological systems provide a rich environment for studying non-equilibrium phenomena, an exciting new frontier in physics. Of particular interest is the study of transcriptional regulation—the process by which transcription factors regulate the activity of mRNA production. Recent evidence suggests that in eukaryotes, organisms whose cells possess nuclei, transcription involves non-equilibrium effects such as energy expenditure. In this work, we undergo a systematic dissection of transcriptional regulation in eukaryotes and demonstrate the necessity for non-equilibrium models in the case of the developing fruit fly. We then investigate some theoretical ramifications of non-equilibrium models of chromatin accessibility that incorporate transient dynamics. Finally, we describe a novel experimental and computational framework for studying the entire transcription cycle—consisting of mRNA initiation, elongation, and cleavage—in order to investigate transcriptional regulation beyond that of simple regulation of initiation.

To my parents, who made all of this possible.

Contents

Contents	ii
List of Figures	vi
List of Tables	viii
1 Introduction	1
1.1 The equilibrium paradigm of transcriptional regulation	1
1.2 The non-equilibrium regime	2
1.3 Experimental and computational tools	4
1.4 Overview of dissertation	6
2 Dissecting equilibrium and non-equilibrium models of chromatin accessibility in development	8
2.1 Introduction	9
2.2 Results	13
2.2.1 A thermodynamic MWC model of activation and chromatin accessibility by Bicoid and Zelda	13
2.2.2 Dynamical prediction and measurement of input-output functions in development	15
2.2.3 The thermodynamic MWC model fails to predict activation of <i>hunchback</i> in the absence of Zelda	17
2.2.4 No thermodynamic model can recapitulate the activation of <i>hunchback</i> by Bicoid alone	21
2.2.5 A non-equilibrium MWC model also fails to describe the <i>zelda</i> ⁻ data	24
2.2.6 Transcription factor-driven chromatin accessibility can capture all aspects of the data	24
2.3 Discussion	29
2.4 Acknowledgments	33
2.5 Methods and Materials	33
2.5.1 Predicting Zelda binding sites	33
2.5.2 Fly Strains	33

2.5.3	Zelda germline clones	34
2.5.4	Sample preparation and data collection	34
2.5.5	Image analysis	35
2.5.6	Data Analysis	35
3	The role of transient transcription factor inputs in models of chromatin accessibility	36
3.1	Introduction	37
3.2	Results	38
3.2.1	A simple steady-state Markov chain model of chromatin accessibility	38
3.2.2	Extending the steady-state model to account for backwards transitions	39
3.2.3	Transient transcription factor input dynamics can decrease noise in transcription onset time distributions	41
3.3	Discussion	43
3.4	Acknowledgements	45
3.5	Methods and Materials	45
4	Dynamic single-cell characterization of the eukaryotic transcription cycle	46
4.1	Introduction	47
4.2	Results	49
4.2.1	Dual-color reporter for dissecting the transcription cycle	49
4.2.2	Transcription cycle parameter inference using Markov Chain Monte Carlo	52
4.2.3	MCMC successfully infers calibration between eGFP and mCherry intensities	53
4.2.4	Inference of single-cell initiation rates recapitulates and improves on previous measurements	55
4.2.5	Elongation rate inference reveals single-molecule variability in RNAP stepping rates	58
4.2.6	Inference reveals functional dependencies of cleavage times	61
4.2.7	Uncovering single-cell mechanistic correlations between transcription cycle parameters	61
4.3	Discussion	63
4.3.1	Dissecting the transcription cycle at the single-cell level	65
4.3.2	Comparison to existing analysis techniques	66
4.3.3	Future improvements	66
4.3.4	Outlook	67
4.4	Acknowledgements	67
4.5	Methods and Materials	68
4.5.1	DNA constructs	68
4.5.2	Fly strains	68
4.5.3	Sample preparation and data collection	68

4.5.4	Image analysis	69
4.5.5	Data Analysis	69
Appendices		70
A Supplementary Information for Chapter 2		70
A.1	Equilibrium Models of Transcription	70
A.1.1	An overview of equilibrium thermodynamics models of transcription .	70
A.1.2	Thermodynamic MWC model	72
A.1.3	Constraining model parameters	74
A.2	Input-Output measurements, predictions, and characterization	76
A.2.1	Input measurement methodology	76
A.2.2	MS2 fluorescence simulation protocol	77
A.2.3	Extracting initial RNAP loading rate and transcriptional onset time .	79
A.3	Mitotic repression is necessary to recapitulate Bicoid- and Zelda-mediated regulation of <i>hunchback</i> using the thermodynamic MWC model	81
A.4	The effect of the <i>zelda</i> ⁻ background on the Bicoid concentration spatiotemporal profile	83
A.5	State-space exploration of theoretical models	86
A.5.1	General methodology of state-space exploration	86
A.5.2	State space exploration with the thermodynamic MWC model	88
A.6	Failures and assumptions of thermodynamic models of transcription	88
A.6.1	Generalized thermodynamic model	88
A.6.2	Generalized thermodynamic model state space exploration	92
A.6.3	Extended generalized thermodynamic model with transcription factor binding in the inaccessible state	93
A.6.4	Investigation of the failure of thermodynamic models	94
A.6.5	Re-examining thermodynamic models of transcriptional regulation . .	94
A.7	Non-equilibrium MWC model	97
A.7.1	Non-equilibrium MWC model	97
A.7.2	Non-equilibrium MWC model state space exploration	99
A.7.3	Alternative non-equilibrium MWC model with strong mitotic repression	102
A.8	Transcription factor-driven model of chromatin accessibility	103
A.8.1	Transcription factor-driven model of chromatin accessibility	103
A.8.2	Exploring alternatives to the additive transcription factor-driven transition rate	105
A.8.3	Transcription factor-driven model of chromatin accessibility state space exploration	107
A.9	Supplementary Videos	108
B Supplementary Information for Chapter 4		110
B.1	Full Model	110

B.2	Characterization of photobleaching in experimental setup	113
B.3	Justification for approximating transcript cleavage as instantaneous	115
B.4	MCMC inference procedure	115
B.4.1	Overview and application of MCMC	115
B.4.2	Justification of scaled observation model due to fluorescence noise behavior	118
B.4.3	Curation of inference results	119
B.4.4	Validation of inference results	121
B.5	Validation of the RNAP processivity assumption	125
B.6	Comparing intra- and inter-embryo variability	126
B.7	Full distributions of transcriptional parameters as a function of embryo position	126
B.8	Comparison of variability in mean initiation rate reported by our inference with static measurements	127
B.9	Comparison of distribution of elongation rates with other works	129
B.10	Theoretical investigation of single-cell distribution of elongation rates	129
B.11	Single-cell correlation analysis using full posterior distributions	132
B.12	Supplementary Videos	133
	Bibliography	144

List of Figures

1.1	A simple activation model of transcriptional regulation.	4
1.2	Brief overview of nascent RNA labeling technology.	6
2.1	Three models of chromatin accessibility and transcriptional regulation.	12
2.2	Thermodynamic MWC model of transcriptional regulation by Bicoid and Zelda.	14
2.3	Prediction and measurement of dynamical input-output functions.	16
2.4	The thermodynamic MWC model can explain <i>hunchback</i> transcriptional dynamics in wild-type, but not <i>zelda</i> ⁻ , embryos.	19
2.5	Failure of thermodynamic models to describe Bicoid-dependent activation of <i>hunchback</i>	23
2.6	Non-equilibrium MWC model of transcriptional regulation cannot predict the observed t_{on} delay.	26
2.7	A model of transcription factor-driven chromatin accessibility is sufficient to recapitulate <i>hunchback</i> transcriptional dynamics.	29
3.1	Results of steady state input model with equal, irreversible transitions.	40
3.2	Results of steady state input model with equal, reversible transitions.	41
3.3	Results of transient input model.	44
4.1	Theoretical model of the transcription cycle and experimental setup.	51
4.2	MCMC inference procedure.	55
4.3	Calibration of MS2 and PP7 fluorescence signals.	56
4.4	Inferred transcription-cycle parameters.	60
4.5	Single-cell correlations between transcription cycle parameters.	63
A.1	Equilibrium thermodynamic model of simple activation.	71
A.2	States, weights, and rate of RNAP loading diagram for the thermodynamic MWC model.	73
A.3	Measurements of input transcription-factor concentration dynamics.	78
A.4	MS2 fluorescence calculation protocol.	80
A.5	Outline of fitting to the trapezoidal model of transcription.	83
A.6	A thermodynamic MWC model including mitotic repression can recapitulate <i>hunchback</i> regulation by Bicoid and Zelda.	84

A.7	Comparison of eGFP-Bicoid measurements in wild-type and <i>zelda</i> ⁻ embryos. . .	86
A.8	Description of state-space metrics and boundary-exploration algorithm.	90
A.9	Exploration of state space.	92
A.10	Intuition for failure of equilibrium models.	95
A.11	A simple kinetic model of transcriptional activation.	96
A.12	Example of a four-state time-dependent model with one Bicoid binding site and no closed chromatin state.	100
A.13	State space exploration for non-equilibrium MWC model with strong mitotic repression for up to five Bicoid binding sites.	102
A.14	Testing the transcription factor-driven model of chromatin accessibility.	106
A.15	Different potential schemes of Bicoid- and Zelda-mediated transition into the accessible state, for a model with $m = 1$ transcriptionally silent state.	107
B.1	Detailed description of reporter construct used in this work.	111
B.2	Investigation of photobleaching in experimental setup.	114
B.3	Scaling of fluorescence measurement noise with overall fluorescence intensity. . .	120
B.4	Automated curation of data.	134
B.5	Overview of MCMC inference validation.	136
B.6	Comparison of intra- and inter-embryo variability.	137
B.7	Single cell distributions of inferred parameters.	138
B.8	Comparison of coefficients of variation (CV) between inferred mean initiation rates and instantaneous counts of number of nascent RNA transcripts.	139
B.9	Comparison of distribution of elongation rates with previous studies.	140
B.10	Single-molecule simulations of elongation dynamics require molecular variability to describe empirical distributions.	142
B.11	Monte Carlo simulation of error in single-cell analysis.	143

List of Tables

B.1	Mean and standard deviation of model parameters used in single-cell simulations.	122
B.2	Comparison of Spearman rank correlation coefficients and p -values between experimental and simulated single-cell correlations.	125
B.3	Parameters used in single-molecule Monte Carlo simulation of elongation rates. .	131

Acknowledgments

My PhD was a five-year affair, and in that span there were a number of people who were invaluable to my whole experience.

Hernan Garcia was one of the greatest advisors I could ask for and was so immeasurably important for shaping me into the scientist I am today. From our original coffee chat during my first Berkeley visit, I thought that we had such common interests and ways of thinking, both in and out of research. The PhD experience is so dependent on the advisor-student relationship, and I couldn't have asked for a better one myself.

Ahmet Yildiz and Yun S. Song were fantastic thesis committee members, and I learned a great deal about how to do good science from both of them. In addition, I always will appreciate them responding to dissertation logistics requests on time and generally making the tedious part of the graduation process as smooth as possible. On a similar note, Mike Eisen and Oskar Hallatschek were great members of my qualifying exam committee, and contributed to making the qualifying exam a deservedly harrowing but also immensely valuable experience.

Academic departments cannot function without their administrators, and I am deeply indebted to Anne Takizawa, Joelle Miles, and Donna Sakima in the Physics department for making my life so much easier. I'm sure there were many invisible problems that they solved on my behalf that I was never aware of.

The entire Garcia lab was such an energetic group to work with, and I had so much fun doing science and making friends along the way. Special shoutouts to Liz Eck for being a five-star rotation mentor and co-author, Meghan Turner for co-starting and co-ending a side project with me (someday we'll witness FLP/FRT working with live imaging, I'm sure!) as well as being a great peer in all things Beyond Academia, and Nick Lammers and Gabriella Martini for being part of the GOAT cohort in the Garcia lab. As for the others — Yang Joon Kim, Simon Alamos, Armando Reimer, Jake Zhao, Dennis Sun, Jacques Bothma, Matty Norstad, Myron Child, Emma Luu, Paul Talledo, Brandon Schlomann, Bruno Moretti, Giana Cirolia, and Yasemin Kirişçioglu — you all made the lab an incredible environment to work in. I also had the distinct pleasure of supervising Liya Oster, Donald Hansen, and Scout Weber during their short stints in the lab, and learned so much about the challenges of effective mentoring as a result (not that any of you needed it!)

In a way, all of this stemmed from the broader PBoC family. Rob Phillips sparked my initial interest in biophysics and first put the Garcia lab on my radar while I was still an undergraduate at Caltech. My infrequent but regular interactions with him have always been inspiring and I can trace many of my formative thought processes as a biophysicist to his APH 161 course at Caltech.

In a similar vein, Stephanie Johnson, Niranjana Srinivas, Tiffany Vora, and Weslee Glenn have been incredibly helpful career mentors, especially as I struggled with navigating my path to graduating and leaving the traditional academic system during my final year at Berkeley. During said navigation, I met Greg Huber and Angela Pisco at the CZ Biohub, who have

been nothing but supportive and welcoming. I had a fantastic internship experience there and am ecstatic to continue working with them in my next role.

I greatly enjoyed living in the city of Berkeley, and will always look back fondly at some of my favorite places. Berkeley Bowl and Monterey Market had the best grocery shopping experiences, and I suspect I will never find similar grocery stores elsewhere. Nomad Cafe, Babette, Jumpin' Java, Yali's, Free Speech Movement, and Rasa Caffè made some killer coffee and provided great ambience when I needed it. On a similar note, Asha, Teance, Far Leaves, and Blue Willow had some seriously amazing tea. Truly, I was able to experience the gamut of caffeinated beverages during graduate school.

I was a part of multiple groups and organizations at Berkeley, in which I made very meaningful relationships. The Berkeley Science Review was incredibly fun to write for and helped spark my interest in scientific writing. At QB3, Mackenzie Smith honed my skills as a writer and Rosanne Lurie programmed very useful professional development events (as well as being another very helpful mentor). I also had a wonderful experience working one summer with Sean Burns, Brittany Johnson, Brittany Daws, and Emily Ramirez at the SURF office.

Beyond Academia is one of the most welcoming and uplifting (not to mention seriously productive!) communities I've ever been a part of, and I am so glad I decided to be a part of it. I very much enjoyed co-directing for a year with Angélica González-Sánchez and Elyse Kedzie. The whole BA family will always be dear to me and I hope to see them again in future conferences.

I consider myself incredibly fortunate to have participated in amazing musical opportunities at Berkeley and grew immensely as a musician. Chamber Choir was a second home for me, and I will cherish the friends and mentors I met there: Magen Solomon, Wei Cheng, Nate Ben-Horin, Zach Manlapid, Jen Liu, Jakob Dahl, Rebecca Herman, Alyssa Kim, Cole Stanford, Joe Arquette, Emily Liquin, Kimball Strong, Clay Halbert, Luke Dzwonczyk, Sarah Ancheta, Barry Fung, Winston Yin, Indu Pereira, Rosie Ward, Amy Liu, Mariah Ronningen, Arlyss Hays, and too many more to list. In addition, Chris Brandes was a fantastic voice teacher and Gabby Lochard was a phenomenal singer and recital partner.

Grad school is a challenging time, but friends help make it better. Micah Brush, Matthew Quenneville, and Sam Badman were some of the first people I met in my Physics cohort, and I had so many great experiences with them: post-work beers at Triple Rock, General Chicken at King Dong (long live the King Dong Club!), board games at Victory Point, to list a few. Eugene Vinitzky, Roger Huang, Dean Tan, Heidi Klumpe, and Sean Nachnani were all stellar housemates over my five years in Berkeley (and eventually Sunnyvale). QinQin Yu, Marie Lu, Sarah Gord, Celine Liong, and Connie Hsueh all were great friends in Berkeley and beyond, and I shared many a great moment of celebration and/or commiseration with them. As for everyone else I forgot to list, my apologies in advance!

I could never have done any of this without the support of my family. To Josh and Kat — I'm so glad we were all in the Bay Area for much of this time, and that we were able to celebrate your wedding before the pandemic and your eventual move to Minnesota. To Mom and Dad — words cannot describe how grateful I am to you for fostering my education and growth. Now I can finally say that I'm done with school!

Finally, to Helena — can you believe it's already been three and a half years since we met at that WCS social? I am always in awe at how a chance encounter can turn into love, but in our case it all felt so natural. Thanks for having the patience to wait for me to finish grad school and I cannot wait to see where the future takes us.

Chapter 1

Introduction

If the 20th century was the golden age of physics, then the 21st century is arguably the golden age of biology. After crucial discoveries such as the crystal structure of DNA ushered in the era of modern molecular biology in the mid-20th century, we now possess unprecedented capacity to visualize, manipulate, and engineer biological systems.

These technological advances, from fluorescence microscopy to DNA sequencing and synthesis, have paved the way for a sea change in the way we approach fundamental biological discovery. We now have the tools to investigate life sciences with the same quantitative rigor that has historically characterized the physical sciences.

From the physicist's perspective, biology also presents incredibly rich systems to study complex behavior. For example, processes such as kinetic proofreading (Hopfield, 1974) and chemotaxis (Berg and Brown, 1972) provide fertile natural phenomena with which to develop and investigate physical theories of dynamical systems out of equilibrium.

Of particular interest is the study of transcription, a key step of the central dogma and one of the most prominent arenas in which biology exhibits logical control and regulation. Building off of decades of deep biophysical research in investigating transcription, this work attempts to examine the role of non-equilibrium processes in eukaryotic transcriptional regulation.

1.1 The equilibrium paradigm of transcriptional regulation

Transcription is a highly complex biological process in which an RNA polymerase (RNAP) molecule binds to a particular gene's promoter region and transcribes the gene's DNA into its messenger RNA (mRNA) counterpart (Fig. 1.1A). Oftentimes this process is tightly regulated, controlled by proteins called transcription factors (TFs) that can enhance or repress transcriptional activity by binding to specific regions of DNA. The concentrations of RNAP and TFs themselves often fluctuate in time, adding temporal dynamics into the fray.

Physical models based on equilibrium statistical mechanics have been used to resounding success in describing transcription regulation in bacterial systems (Ackers et al., 1982; Buchler et al., 2003; Vilar et al., 2003; Bolouri and Davidson, 2003; Bakk et al., 2004; Bintu et al., 2005b,a; Zeng et al., 2010; He et al., 2010; Garcia and Phillips, 2011; Brewster et al., 2012; Sherman and Cohen, 2012; Cui et al., 2013; Brewster et al., 2014; Sepulveda et al., 2016; Razo-Mejia et al., 2018). In these models, the regulation of gene output by TFs is reduced to a simple physical picture of molecules binding and unbinding to and from regulatory DNA in the vicinity of genes. For example, Figure 1.1B shows a toy equilibrium model of simple activation for a system possessing one binding site each for an activator and RNAP. By expressing the system in terms of its various configurations and statistical weights, transcription output can then be calculated from the partition function of the system by computing p_{bound} , the probability of the system being in an RNAP-bound state.

These equilibrium models can be derived from the more general kinetic picture shown in the middle of Figure 1.1C using four assumptions. First, we assume the *occupancy hypothesis* (Fig. 1.1C, i), which states that the downstream rate of RNAP loading (and thus, transcription output) is proportional to the probability of the system being in an RNAP-bound microstate (Phillips et al., 2019). Second, the system must be in quasi-steady state with respect to global activator dynamics. Here, the binding and unbinding of activators must occur on a timescale much faster than that of the natural modulation of activator concentration over time, such that the system responds to changes in activator concentration by instantaneously reaching its equilibrium value (Fig. 1.1C, ii). Third, the system must also be in quasi-steady state with respect to nascent RNA transcript production (Fig. 1.1C, iii). Again, this implies that the binding and unbinding kinetics of activators must be much faster than the timescale of transcript production. Finally, the system must conserve energy—no net energy must be consumed over time as the system transitions between microstates (Fig. 1.1C, iv).

1.2 The non-equilibrium regime

Recent studies suggest that such equilibrium models might not be entirely sufficient, hinting at non-equilibrium effects at play (Garcia et al., 2012; Hammar et al., 2014). Such effects may be even more prominent in eukaryotic systems, which possess more complex regulatory structures and energy-expending mechanisms than bacteria, for example through the wrapping of DNA around nucleosomes (Polach and Widom, 1995; Levine, 2010; Schulze and Wallrath, 2007; Lam et al., 2008; Raveh-Sadka et al., 2009; Li and Gilmour, 2011; Fussner et al., 2011; Bai et al., 2011; Li et al., 2014a; Hansen and O’Shea, 2015). Furthermore, transcription often occurs out of steady state, and transient dynamics may play a dramatic role in transcription regulation that cannot be captured by equilibrium models (Wong and Gunawardena, 2020).

For example, recent work has shown that the full spatiotemporal features of certain transcription patterns in the developing fruit fly embryo appear to violate bounds imposed by equilibrium models (Kim and O’Shea, 2008; Estrada et al., 2016; Park et al., 2019; Eck et al., 2020).

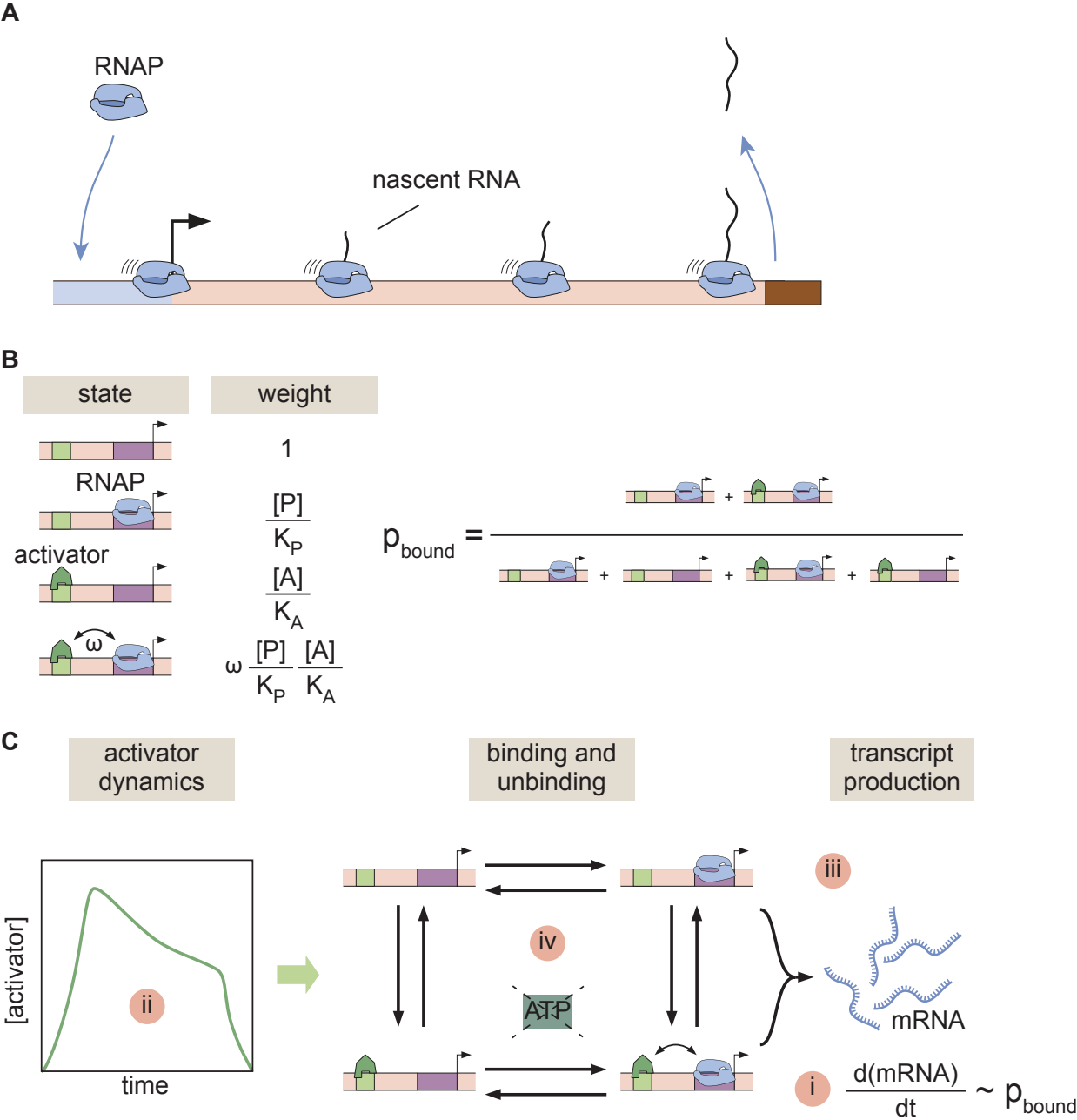


Figure 1.1: A simple activation model of transcriptional regulation. See caption on next page.

Figure 1.1: A simple activation model of transcriptional regulation. (A) Cartoon illustration of RNAP loading onto a promoter and transcribing a nascent RNA molecule. (B) Toy equilibrium model of simple activation. Here, a gene possesses one binding site each for activator (green) and RNAP (purple). Given equilibrium binding energies that lead to dissociation constants K_A and K_P , the Boltzmann weight of each state can be written using the concentrations of activator and RNAP polymerase, respectively. The system exhibits activation if the cooperativity factor ω between the activator and RNAP is greater than one. The equilibrium probability of having RNAP bound is then given by p_{bound} . (C) A more general kinetic model of transcriptional activation. An idealized picture of transcription regulation through simple activation. The system fluctuates between different states via binding and unbinding of activator and RNAP, and mRNA transcript initiation can occur when RNAP is bound. In addition, the concentration of activator itself can change over time. Under the assumptions of the occupancy hypothesis (i), quasi-steady state (ii, iii), and conservation of energy (iv), this picture reduces to the equilibrium model in (A).

However, it remains to be seen exactly which equilibrium assumptions are broken, and to what extent. One of the main difficulties in studying non-equilibrium processes lies in a detailed characterization of how far the system is from equilibrium—the space of possible non-equilibrium models is astoundingly vast. Furthermore, it is unclear if any observed non-equilibrium behavior is a universal feature of eukaryotic systems, or instead specific to certain genes. Thoroughly generalizing studies of non-equilibrium transcriptional regulation will present a significant challenge in future biological research. Ultimately, tying these physical insights to mechanistic processes will shed a deeper light into our understanding of how transcription is coupled to biophysical quantities such as energy and time.

1.3 Experimental and computational tools

To date, transcription has mostly been studied in detail using *in vitro* approaches (Bai et al., 2006; Herbert et al., 2008) or *in vivo* measurements that require the fixation of cellular material and lack the temporal resolution to uncover how transcriptional regulation unfolds in real time (Roeder, 1991; Femino et al., 1998; Raj et al., 2006; Saunders et al., 2006; Muse et al., 2007; Zenklusen et al., 2008; Core et al., 2008; Fuda et al., 2009; Wyart et al., 2010; Churchman and Weissman, 2011; Boettiger and Levine, 2013; Little et al., 2013; Xu et al., 2015; Zoller et al., 2018). Such techniques can only produce a static measurement of transcriptional behavior, often obscuring the rich dynamics of non-equilibrium processes that distinguish them from the equilibrium regime. Because systems like the fly embryo are out of steady state, time-resolved measurements are necessary to capture the full dynamics of

transcription. Furthermore, live-cell approaches are necessary to study transcription in its native, living context.

Recently it has become possible to dissect transcription in single living cells and in their full dynamical complexity using tools such as MS2 or PP7 that make it possible to fluorescently label nascent transcripts (Bertrand et al., 1998; Golding et al., 2005; Chao et al., 2008; Larson et al., 2011a). Figure 1.2A gives an overview of how these tools work. Briefly, a set of stem loop sequences are inserted in a reporter gene body. Upon transcription, the nascent RNA in these sequences folds into hairpin loops, which bind with high specificity to constitutively expressed complementary coat proteins that are fused to fluorescent proteins. This binding results in a high local concentration of fluorophores at the nascent transcriptional locus, producing a fluorescent punctum that provides a quantitative readout of the number of nascent RNA transcripts on the gene (Fig. 1.2B). The intensity of the fluorescent spot is linearly correlated with the number of active RNAP molecules on the gene (Fig. 1.2C; Garcia et al. (2013); Lucas et al. (2013)). Both MS2 and PP7 function according to this setup and differ only in their stem loop sequences, which bind to MCP and PCP coat proteins, respectively.

These technologies have yielded insights into a broad range of phenomena that dictate transcriptional dynamics. Examples include intrinsic transcriptional noise in yeast (Hocine et al., 2013), kinetic splicing effects in human cells (Coulon et al., 2014), elongation rates (Garcia et al., 2013; Fukaya et al., 2017) and chromatin accessibility (Dufourt et al., 2018; Yamada et al., 2019; Eck et al., 2020) in *Drosophila melanogaster*, and transcriptional bursting in mammalian cells (Tantale et al., 2016), *Dictyostelium* (Chubb et al., 2006; Muramoto et al., 2012; Corrigan and Chubb, 2014), fruit flies (Garcia et al., 2013; Lucas et al., 2013; Bothma et al., 2014; Fukaya et al., 2016; Falo-Sanjuan et al., 2019; Lammers et al., 2020) and *Caenorhabditis elegans* (Lee et al., 2019).

Despite these technological breakthroughs, the resulting experimental data are only as powerful as their downstream analysis permits. Although nascent RNA labeling technology such as MS2 provides a quantitative readout of the real-time activity of RNAP molecules on a gene, producing measurements that can be compared to the theoretical models discussed earlier is no simple task.

For example, the raw microscopy output from a single 2-hour experiment contains several tens of gigabytes of complex fluorescence images in four dimensions (space + time). As a result, parsing the data requires sophisticated image analysis tools that can process these datasets into single-cell time-series readouts of transcriptional activity.

Further, new computational analysis techniques are required to glean relevant biophysical insights from these noisy and complex fluorescent time series and study the regulation of processes such as transcription onset times (Dufourt et al., 2018; Fritzsche et al., 2018; Eck et al., 2020) or transcriptional initiation rates (Garcia et al., 2013). Although recent works using techniques such as Hidden Markov Models have pushed the frontiers of extracting useful signal from these noisy data (Falo-Sanjuan et al., 2019; Berrocal et al., 2020; Lammers et al., 2020), novel computational analyses of these signals remain a work in progress in parallel to the development of increasingly refined experimental tools.

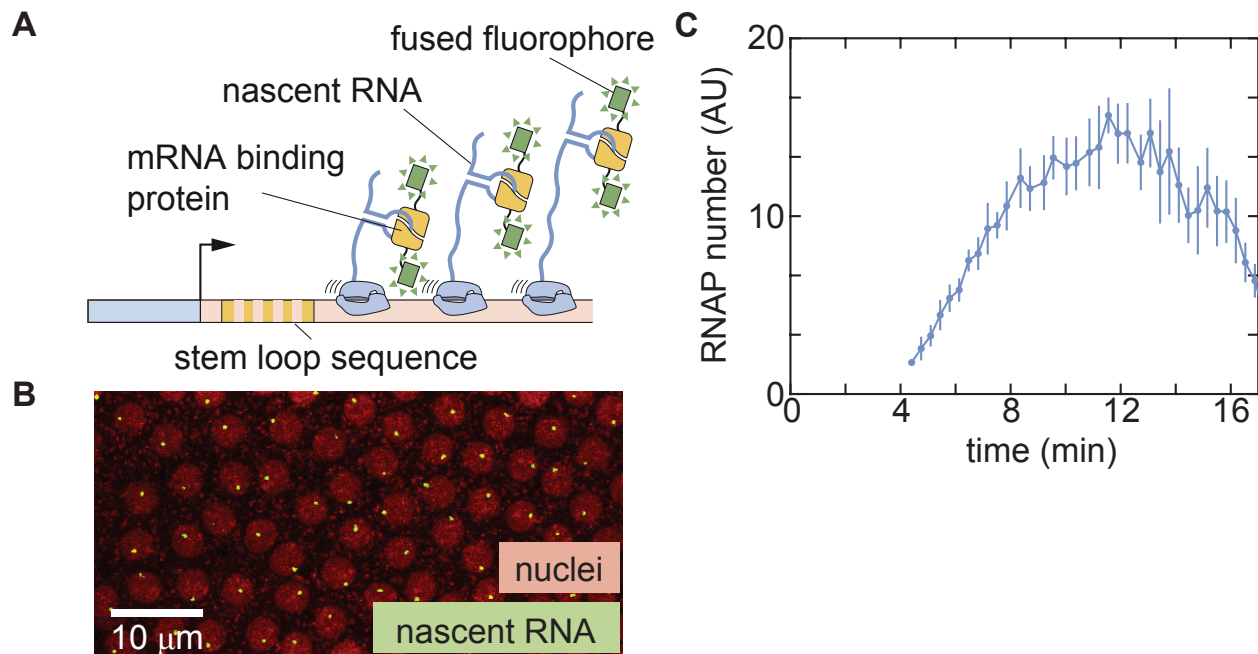


Figure 1.2: Brief overview of nascent RNA labeling technology. (A) A stem loop sequence is inserted into the body of a gene, producing hairpin loops in the nascent RNA transcript. Complementary mRNA binding proteins fused to fluorophores then bind to these hairpin loops, resulting in the localization of fluorescence at the transcriptional locus. (B) When a tagged gene is active, its output will result in bright fluorescent puncta in a microscope’s field of view (green), shown here in the case of a developing fruit fly embryo. Cell nuclei (red) are labeled to give positional context. (C) The instantaneous brightness of a fluorescent spot corresponds linearly with the number of active RNAP molecules on the gene.

The work described here implements the MS2 and PP7 technologies inside the early developing *Drosophila melanogaster* embryo to study theoretical models of eukaryotic transcriptional regulation. Since the fly’s early developmental processes are tightly coordinated in space and time, the fruit fly embryo presents an ideal system for studying eukaryotic transcriptional regulation. Subsequent computational analysis of the fluorescent signals generated from these nascent RNA-labeling technologies allow for a tight dialogue between theory, experiment, and computation.

1.4 Overview of dissertation

This dissertation comprises three main sections, derived from two published works and one work in progress.

Chapter 2 describes a comprehensive experimental and theoretical dissection of equilibrium and non-equilibrium models of transcriptional regulation in the context of eukaryotic chromatin accessibility. By examining the regulatory action of the pioneering factors Zelda and Bicoid in relation to the activity of a *hunchback* reporter gene, we discovered that the observed transcriptional dynamics could not be explained by equilibrium theories. This motivated a systematic investigation of non-equilibrium extensions to our equilibrium model of transcriptional regulation. After ruling out a non-equilibrium version of the famed Monod-Wyman-Changeux allosteric model of gene regulation, in which TFs indirectly mediate gene state transitions by modifying state energy levels, we eventually successfully proposed a different class of non-equilibrium model that involved direct TF-mediated catalysis of chromatin accessibility. This work was published in 2020 in *eLife*.

Chapter 3 follows up on the insights gained from Chapter 2 and conducts a theoretical investigation into the consequences of non-equilibrium transcriptional regulation for precision of developmental timing. Building off of the TF-mediated non-equilibrium model of chromatin accessibility introduced at the end of Chapter 1, we probe the effect of transient input TF dynamics on single-cell distributions of first passage times of transcriptional onset. We show that in the quasi-steady state regime, there is a limit to how well a population of cells can synchronize their transcriptional onset times, but that this limit can be surpassed in the transient regime, where input TFs are allowed to vary their concentrations in time.

Chapter 4 extends the scope of studying transcriptional regulation beyond that of simple transcription initiation, describing an experimental and computational framework for quantitatively studying the whole transcription cycle—initiation, elongation, and cleavage of mRNA transcripts. Previous live imaging studies tended to focus on particular aspects of transcription (such as initiation or elongation), and a principled methodology for simultaneous investigation of the whole transcription cycle at the single-cell level was lacking. By utilizing a dual-color reporter with 5' and 3' ends tagged with MS2 and PP7 stem loops, respectively, we can acquire data that informs on transcription initiation, elongation, and cleavage. We then develop a simple effective model and use the Bayesian inference technique of Markov Chain Monte Carlo to generate quantitative estimates of initiation, elongation, and cleavage rates at the single-cell level. Finally, we demonstrate how these single-cell results can be harnessed to produce a suite of insights into processes from single-molecule elongation dynamics to covariation between transcription cycle parameters. This work was published in 2021 in *PLoS Computational Biology*.

Finally, Appendices A and B provide detailed supplementary information on the work described in Chapters 2 and 4.

Chapter 2

Dissecting equilibrium and non-equilibrium models of chromatin accessibility in development

Foreword

I started my PhD during a critical juncture in the field of developmental biology. Tools such as live cell imaging were beginning to surface as viable experimental techniques for studying gene regulatory systems in organisms such as the fruit fly. At the same time, physicists were turning their attention toward complex biological systems—cells, genetic networks, and the like—in increasing numbers. The time was truly ripe for physics-based approaches that could take advantage of the extremely rich datasets produced by these imaging techniques.

My advisor, Hernan Garcia, had recently started his lab at UC Berkeley with exactly that type of goal. Using live cell imaging, his lab was beginning to generate highly quantitative microscopy datasets of transcription in developing fruit fly embryos. With the ability to visualize both input transcription factor dynamics with fluorophore fusions and output transcriptional activity with the MS2-MCP technology, we could test biophysical models of transcriptional regulation with unprecedented precision.

This project was borne out of this sudden deluge of high-quality data. My collaborator and fellow graduate student, Elizabeth (Liz) Eck, had painstakingly generated a comprehensive dataset of transcriptional activity in a *hunchback* reporter. Specifically, recent studies had indicated that chromatin accessibility, a facet of transcription unique to eukaryotes, was heavily regulated by a pioneering factor called *Zelda*. By measuring *hunchback* activity in both wild-type and a *zelda* mutant, we possessed an exquisite dataset with a perturbation targeted towards studying chromatin accessibility. When I joined the lab, the data had been collected and all that was left was the data analysis and modeling.

At first (like all scientific endeavors) the path seemed easy. We had a dataset. We had a model—in this case, a Monod-Wyman-Changeux (MWC)-based model of chromatin

accessibility recently proposed by Leonid Mirny. All we had to do was validate the model against the data and extend the equilibrium statistical mechanical picture of transcriptional regulation to account for the impact of chromatin accessibility. The *zelda* mutant provided the perturbation with which we could test the predictions of the model.

With high confidence, we thus approached the research problem, and found immediately that the model failed. No matter what parameters we changed, the MWC model of chromatin accessibility never seemed to explain the data. So, we returned to the whiteboard and brainstormed. Eventually we concluded that the equilibrium paradigm was insufficient, and that we had to invoke non-equilibrium effects. In the end, we were able to come up with a non-equilibrium model of chromatin accessibility that successfully explained our measurements.

Of course, refutation of a model is no small claim. What came next was years of careful analysis, by which we systematically investigated different classes of equilibrium and non-equilibrium models of chromatin accessibility. We finally settled on a clear narrative, in which we broke assumptions one by one until we arrived at our working model.

This project led to a publication in *eLife*, co-authored by myself and Liz. I couldn't be happier with my first graduate school project. Ultimately, this was a much-needed combination of experimental and theoretical biophysics, and an important keystone work that was pivotal for the development of my scientific style.

2.1 Introduction

Over the last decade, hopeful analogies between genetic and electronic circuits have posed the challenge of predicting the output gene expression of a DNA regulatory sequence in much the same way that the output current of an electronic circuit can be predicted from its wiring diagram (Endy, 2005). This challenge has been met with a plethora of theoretical works, including thermodynamic models, which use equilibrium statistical mechanics to calculate the probability of finding transcription factors bound to DNA and to relate this probability to the output rate of mRNA production (Ackers et al., 1982; Buchler et al., 2003; Vilar and Leibler, 2003; Bolouri and Davidson, 2003; Bintu et al., 2005b,a; Sherman and Cohen, 2012). Thermodynamic models of bacterial transcription launched a dialogue between theory and experiments that has largely confirmed their predictive power for several operons (Ackers et al., 1982; Bakk et al., 2004; Zeng et al., 2010; He et al., 2010; Garcia and Phillips, 2011; Brewster et al., 2012; Cui et al., 2013; Brewster et al., 2014; Sepulveda et al., 2016; Razo-Mejia et al., 2018) with a few potential exceptions (Garcia et al., 2012; Hammar et al., 2014).

Following these successes, thermodynamic models have been widely applied to eukaryotes to describe transcriptional regulation in yeast (Segal et al., 2006; Gertz et al., 2009; Sharon et al., 2012; Zeigler and Cohen, 2014), human cells (Giorgetti et al., 2010), and the fruit fly *Drosophila melanogaster* (Jaeger et al., 2004a; Zinzen et al., 2006; Segal et al., 2008; Fakhouri et al., 2010; Parker et al., 2011; Kanodia et al., 2012; White et al., 2012; Samee et al., 2015; Sayal et al., 2016). However, two key differences between bacteria and eukaryotes

cast doubt on the applicability of thermodynamic models to predict transcriptional regulation in the latter. First, in eukaryotes, DNA is tightly packed in nucleosomes and must become accessible in order for transcription factor binding and transcription to occur (Polach and Widom, 1995; Levine, 2010; Schulze and Wallrath, 2007; Lam et al., 2008; Raveh-Sadka et al., 2009; Li and Gilmour, 2011; Fussner et al., 2011; Bai et al., 2011; Li et al., 2014a; Hansen and O’Shea, 2015). Second, recent reports have speculated that, unlike in bacteria, the equilibrium framework may be insufficient to account for the energy-expending steps involved in eukaryotic transcriptional regulation, such as histone modifications and nucleosome remodeling, calling for non-equilibrium models of transcriptional regulation (Kim and O’Shea, 2008; Estrada et al., 2016; Li et al., 2018; Park et al., 2019).

Recently, various theoretical models have incorporated chromatin accessibility and energy expenditure in theoretical descriptions of eukaryotic transcriptional regulation. First, models by Mirny (2010), Narula and Igoshin (2010), and Marzen et al. (2013) accounted for chromatin occluding transcription-factor binding by extending thermodynamic models to incorporate the Monod-Wyman-Changeux (MWC) model of allostery (Fig. 2.1A; Monod et al., 1965). This thermodynamic MWC model assumes that chromatin rapidly transitions between accessible and inaccessible states via thermal fluctuations, and that the binding of transcription factors to accessible DNA shifts this equilibrium toward the accessible state. Like all thermodynamic models, this model relies on the “occupancy hypothesis” (Hammar et al., 2014; Garcia et al., 2012; Phillips et al., 2019): the probability p_{bound} of finding RNA polymerase (RNAP) bound to the promoter, a quantity that can be easily computed, is linearly related to the rate of mRNA production $\frac{d}{dt}(\text{mRNA})$, a quantity that can be experimentally measured, such that

$$\frac{d}{dt}(\text{mRNA}) = R p_{bound}. \quad (2.1)$$

Here, R is the rate of mRNA production when the system is in an RNAP-bound state (see Section A.1.1 for a more detailed overview). Additionally, in all thermodynamic models, the transitions between states are assumed to be much faster than both the rate of transcriptional initiation and changes in transcription factor concentrations. This separation of time scales, combined with a lack of energy dissipation in the process of regulation, makes it possible to consider the states to be in equilibrium such that the probability of each state can be computed using its Boltzmann weight (Garcia et al., 2007).

Despite the predictive power of thermodynamic models, eukaryotic transcription may not adhere to the requirements imposed by the thermodynamic framework. Indeed, Narula and Igoshin (2010), Hammar et al. (2014), Estrada et al. (2016), Scholes et al. (2017), and Li et al. (2018) have proposed theoretical treatments of transcriptional regulation that maintain the occupancy hypothesis, but make no assumptions about separation of time scales or energy expenditure in the process of regulation. When combined with the MWC mechanism of DNA allostery, these models result in a non-equilibrium MWC model (Fig. 2.1B). Here, no constraints are imposed on the relative values of the transition rates between states and energy can be dissipated over time. To our knowledge, neither the thermodynamic MWC

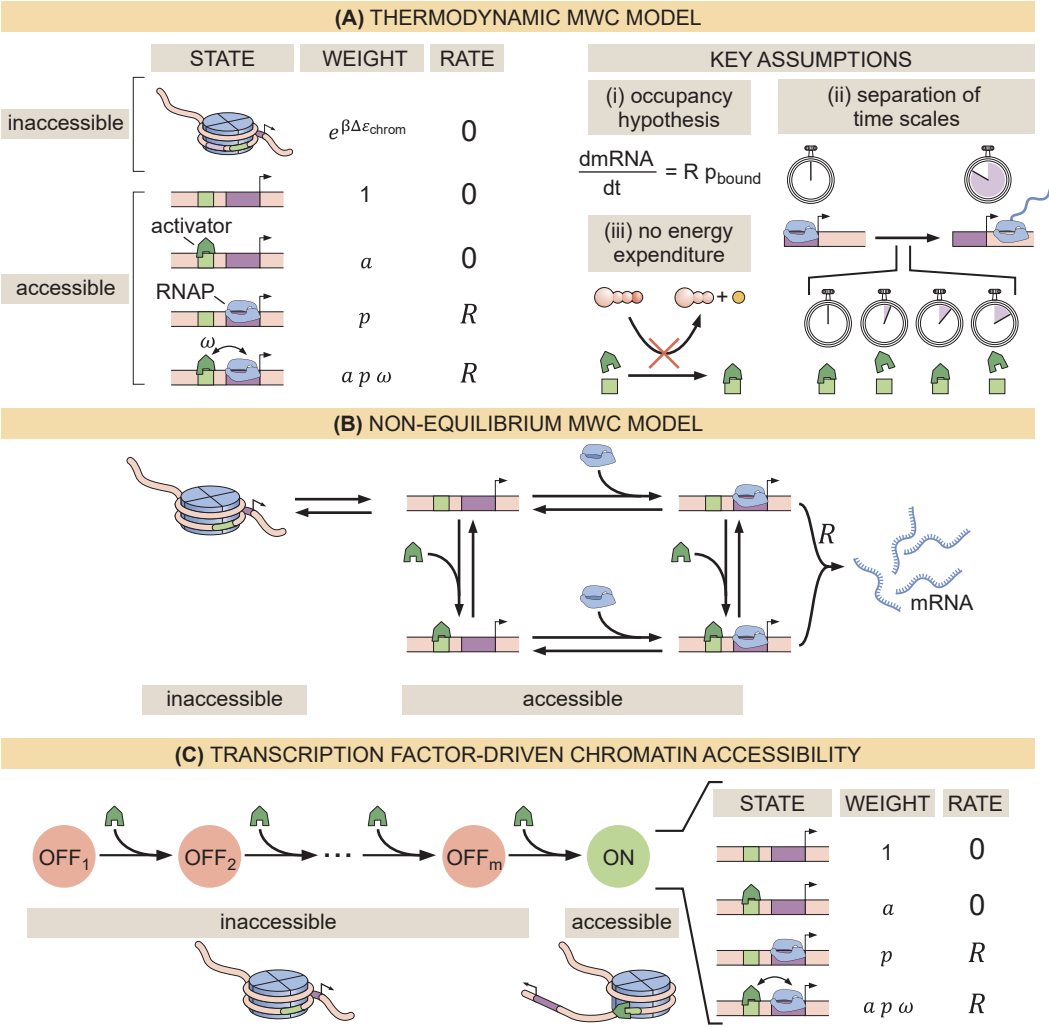


Figure 2.1: Three models of chromatin accessibility and transcriptional regulation. See caption on next page.

Figure 2.1: Three models of chromatin accessibility and transcriptional regulation. (A) Thermodynamic MWC model where chromatin can be inaccessible or accessible to transcription factor binding. Each state is associated with a statistical weight given by the Boltzmann distribution and with a rate of transcriptional initiation. $\Delta\varepsilon_{\text{chrom}}$ is the energy cost associated with making the DNA accessible and ω is an interaction energy between the activator and RNAP. $a = [\text{activator}]/K_a$ and $p = [\text{RNAP}]/K_p$ with K_a and K_p being the dissociation constants of the activator and RNAP, respectively. This model assumes the occupancy hypothesis, separation of time scales, and lack of energy expenditure described in the text. (B) Non-equilibrium MWC model where no assumptions about separation of time scales or energy expenditure are made. Transition rates that depend on the concentration of the activator or RNAP are indicated by an arrow incorporating the respective protein. (C) Transcription factor-driven chromatin accessibility model where the activator catalyzes irreversible transitions of the DNA through m silent states before it becomes accessible. Once this accessible state is reached, the system is in equilibrium.

model nor the non-equilibrium MWC model have been tested experimentally in eukaryotic transcriptional regulation.

Here, we performed a systematic dissection of the predictive power of these MWC models of DNA allostery in the embryonic development of the fruit fly *Drosophila melanogaster* in the context of the step-like activation of the *hunchback* gene by the Bicoid activator and the pioneer-like transcription factor Zelda (Driever et al., 1989; Nien et al., 2011; Xu et al., 2014). Specifically, we compared the predictions from these MWC models against dynamical measurements of input Bicoid and Zelda concentrations and output *hunchback* transcriptional activity. Using this approach, we discovered that no thermodynamic or non-equilibrium MWC model featuring the regulation of *hunchback* by Bicoid and Zelda could describe the transcriptional dynamics of this gene. Following recent reports of the regulation of *hunchback* and *snail* (Desponds et al., 2016; Dufourt et al., 2018) and inspired by discussions of non-equilibrium schemes of transcriptional regulation (Coulon et al., 2013; Wong and Gunawardena, 2020), we proposed a model in which Bicoid and Zelda, rather than passively biasing thermal fluctuations of chromatin toward the accessible state, actively assist the overcoming of an energetic barrier to make chromatin accessible through the recruitment of energy-consuming histone modifiers or chromatin remodelers. This model (Fig. 2.1C) recapitulated all of our experimental observations. This interplay between theory and experiment establishes a clear path to identify the molecular steps that make DNA accessible, to systematically test our model of transcription factor-driven chromatin accessibility, and to make progress toward a predictive understanding of transcriptional regulation in development.

2.2 Results

2.2.1 A thermodynamic MWC model of activation and chromatin accessibility by Bicoid and Zelda

During the first two hours of embryonic development, the *hunchback* P2 minimal enhancer (Margolis et al., 1995; Driever et al., 1989; Perry et al., 2012; Park et al., 2019) is believed to be devoid of significant input signals other than activation by Bicoid and regulation of chromatin accessibility by both Bicoid and Zelda (Perry et al., 2012; Xu et al., 2014; Hannon et al., 2017). As a result, the early regulation of *hunchback* provides an ideal scaffold for a stringent test of simple theoretical models of eukaryotic transcriptional regulation.

Our implementation of the thermodynamic MWC model (Fig. 2.1A) in the context of *hunchback* states that in the inaccessible state, neither Bicoid nor Zelda can bind DNA. In the accessible state, DNA is unwrapped and the binding sites become accessible to these transcription factors. Due to the energetic cost of opening the chromatin ($\Delta\varepsilon_{\text{chrom}}$), the accessible state is less likely to occur than the inaccessible one. However, the binding of Bicoid or Zelda can shift the equilibrium toward the accessible state (Adams and Workman, 1995; Miller and Widom, 2003; Mirny, 2010; Narula and Igoshin, 2010; Marzen et al., 2013).

In our model, we assume that all binding sites for a given molecular species have the same binding affinity. Relaxing this assumption does not affect any of our conclusions (as we will see below in Sections 2.2.3 and 2.2.4). Bicoid upregulates transcription by recruiting RNAP through a protein-protein interaction characterized by the parameter ω_{bp} . We allow cooperative protein-protein interactions between Bicoid molecules, described by ω_b . However, since to our knowledge there is no evidence of direct interaction between Zelda and any other proteins, we assume no interaction between Zelda and Bicoid, or between Zelda and RNAP.

In Fig. 2.2A, we illustrate the simplified case of two Bicoid binding sites and one Zelda binding site, plus the corresponding statistical weights of each state given by their Boltzmann factors. Note that the actual model utilized throughout this work accounts for at least six Bicoid binding sites and ten Zelda binding sites that have been identified within the *hunchback* P2 enhancer (Section 2.5.1; Driever and Nusslein-Volhard, 1988; Driever et al., 1989; Park et al., 2019). This general model is described in detail in Section A.1.2.

The probability of finding RNAP bound to the promoter is calculated by dividing the sum of all statistical weights featuring RNAP by the sum of the weights corresponding to all possible system states. This leads to

$$p_{\text{bound}} = \frac{\underbrace{(1+z)^{n_z}}_{\text{inaccessible state}} p \left(1 + \sum_{i=1}^{n_b} \binom{n_b}{i} b^i \omega_b^{i-1} \omega_{bp}^i \right)}{\underbrace{(1+z)^{n_z}}_{\text{Zelda binding}} \underbrace{\left(1 + p + \sum_{j=0,1} \sum_{i=1}^{n_b} \binom{n_b}{i} b^i \omega_b^{i-1} p^j \omega_{bp}^{ij} \right)}_{\text{Bicoid and RNAP binding}}}, \quad (2.2)$$

where $b = [\text{Bicoid}]/K_b$, $z = [\text{Zelda}]/K_z$, and $p = [\text{RNAP}]/K_p$, with $[\text{Bicoid}]$, $[\text{Zelda}]$, and

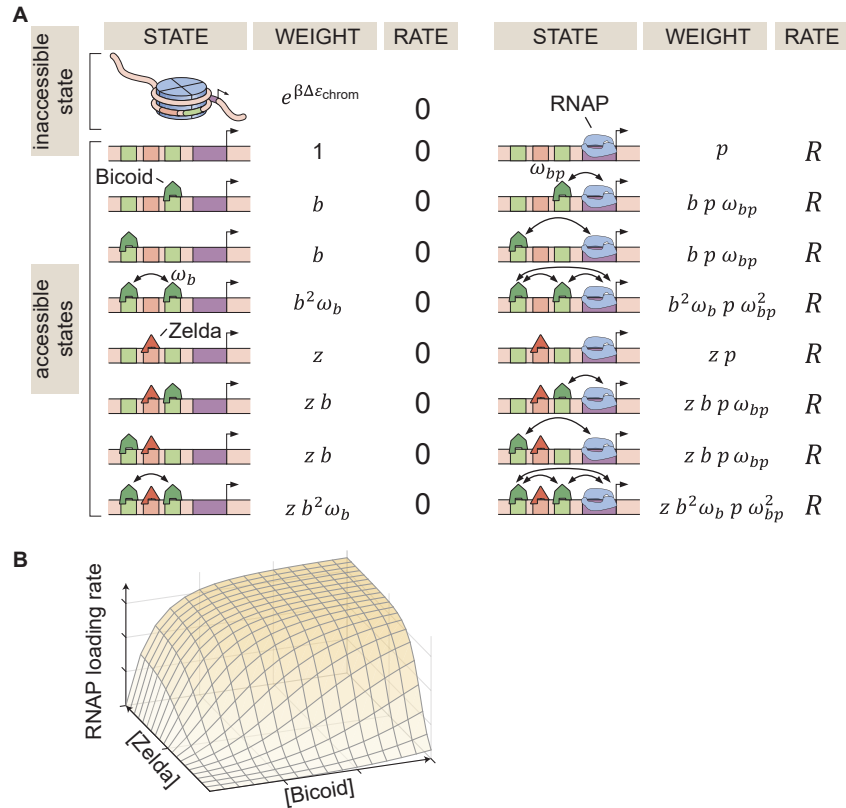


Figure 2.2: Thermodynamic MWC model of transcriptional regulation by Bicoid and Zelda. (A) States and statistical weights for a simplified version of the *hunchback* P2 enhancer. In this model, we assume that chromatin occluded by nucleosomes is not accessible to transcription factors or RNAP. Parameters are defined in the text. (B) 3D input-output function predicting the rate of RNAP loading (and of transcriptional initiation) as a function of Bicoid and Zelda concentrations for a given set of model parameters.

$[RNAP]$ being the concentrations of Bicoid, Zelda, and RNAP, respectively, and K_b , K_z , and K_p their dissociation constants (see Sections A.1.1 and A.1.2 for a detailed derivation). Given a set of model parameters, plugging p_{bound} into Equation 2.1 predicts the rate of RNAP loading as a function of Bicoid and Zelda concentrations as shown in Fig. 2.2B. Note that in this work, we treat the rate of transcriptional initiation and the rate of RNAP loading interchangeably.

2.2.2 Dynamical prediction and measurement of input-output functions in development

In order to experimentally test the theoretical model in Fig. 2.2, it is necessary to measure both the inputs – the concentrations of Bicoid and Zelda – as well as the output rate of RNAP loading. Typically, when testing models of transcriptional regulation in bacteria and eukaryotes, input transcription-factor concentrations are assumed to not be modulated in time: regulation is in steady state (Ackers et al., 1982; Bakk et al., 2004; Segal et al., 2008; Garcia and Phillips, 2011; Sherman and Cohen, 2012; Cui et al., 2013; Little et al., 2013; Raveh-Sadka et al., 2009; Sharon et al., 2012; Zeigler and Cohen, 2014; Xu et al., 2015; Sepulveda et al., 2016; Estrada et al., 2016; Razo-Mejia et al., 2018; Zoller et al., 2018; Park et al., 2019). However, embryonic development is a highly dynamic process in which the concentrations of transcription factors are constantly changing due to their nuclear import and export dynamics, and due to protein production, diffusion, and degradation (Edgar and Schubiger, 1986; Edgar et al., 1987; Jaeger et al., 2004b; Gregor et al., 2007b). As a result, it is necessary to go beyond steady-state assumptions and to predict and measure how the *instantaneous*, time-varying concentrations of Bicoid and Zelda at each point in space dictate *hunchback* output transcriptional dynamics.

In order to quantify the concentration dynamics of Bicoid, we utilized an established Bicoid-eGFP line (Sections 4.5.2, 4.5.3, and 4.5.4; Fig. 2.3A and Fig. A.3A; Video A.9.1; Gregor et al., 2007b; Liu et al., 2013). As expected, this line displayed the exponential Bicoid gradient across the length of the embryo (Section A.2.1; Fig. A.3B). We measured mean Bicoid nuclear concentration dynamics along the anterior-posterior axis of the embryo, as exemplified for two positions in Fig. 2.3A. As previously reported (Gregor et al., 2007b), after anaphase and nuclear envelope formation, the Bicoid nuclear concentration quickly increases as a result of nuclear import. These measurements were used as inputs into the theoretical model in Fig. 2.2.

Zelda concentration dynamics were measured in a Zelda-sfGFP line (Sections 4.5.2, 4.5.3, and 4.5.4; Fig. 2.3B; Video A.9.2; Hamm et al., 2017). Consistent with previous results (Staudt et al., 2006; Liang et al., 2008; Dufourt et al., 2018), the Zelda concentration was spatially uniform along the embryo (Fig. A.3). Contrasting Fig. 2.3A and B reveals that the overall concentration dynamics of both Bicoid and Zelda are qualitatively comparable. As a result of Zelda’s spatial uniformity, we used mean Zelda nuclear concentration dynamics averaged across all nuclei within the field of view to test our model (Section A.2.1; Fig. 2.3B).

Given the high reproducibility of the concentration dynamics of Bicoid and Zelda (Fig. A.3), we combined measurements from multiple embryos by synchronizing their anaphase in order to create an “averaged embryo” (Section A.2.1), an approach that has been repeatedly used to describe protein and transcriptional dynamics in the early fly embryo (Garcia et al., 2013; Bothma et al., 2014, 2015; Berrocal et al., 2020; Lammers et al., 2020).

Our model assumes that *hunchback* output depends on the instantaneous concentration of input transcription factors. As a result, at each position along the anterior-posterior axis of

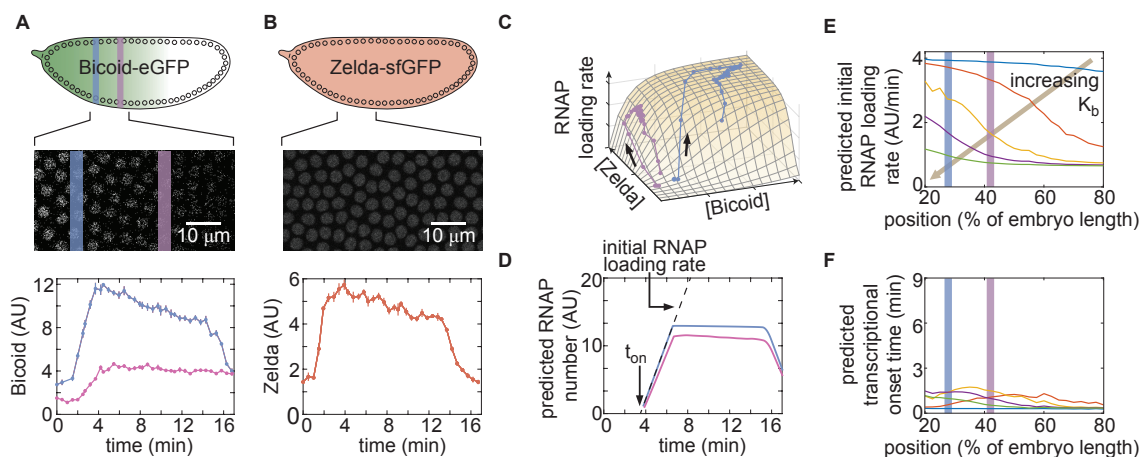


Figure 2.3: Prediction and measurement of dynamical input-output functions. (A) Measurement of Bicoid concentration dynamics in nuclear cycle 13. Color denotes different positions along the embryo and time is defined with respect to anaphase. (B) Zelda concentration dynamics. These dynamics are uniform throughout the embryo. (C) Trajectories defined by the input concentration dynamics of Bicoid and Zelda along the predicted input-output surface. Each trajectory corresponds to the RNAP loading-rate dynamics experienced by nuclei at the positions indicated in (A). (D) Predicted number of RNAP molecules actively transcribing the gene as a function of time and position along the embryo, and calculation of the corresponding initial rate of RNAP loading and the time of transcriptional onset, t_{on} . (E, F) Predicted *hunchback* (E) initial rate of RNAP loading and (F) t_{on} as a function of position along the embryo for varying values of the Bicoid dissociation constant K_b . (A, B, error bars are standard error of the mean nuclear fluorescence in an individual embryo, averaged across all nuclei at a given position; D, the standard error of the mean predicted RNAP number in a single embryo, propagated from the errors in A and B, is thinner than the curve itself; E, F, only mean predictions are shown so as to not obscure differences between them; we imaged $n=6$ Bicoid-GFP and $n=3$ Zelda-GFP embryos.)

the embryo, the combined Bicoid and Zelda concentration dynamics define a trajectory over time along the predicted input-output function surface (Fig. 2.3C). The resulting trajectory predicts the rate of RNAP loading as a function of time. However, instead of focusing on calculating RNAP loading rate, we used it to compute the number of RNAP molecules actively transcribing *hunchback* at each point in space and time, a more experimentally accessible quantity (Section 2.2.3). This quantity can be obtained by accounting for the RNAP elongation rate and the cleavage of nascent RNA upon termination (Section A.2.2; Fig. A.4; Bothma et al., 2014; Lammers et al., 2020) yielding the predictions shown in Fig. 2.3D.

Instead of examining the full time-dependent nature of our data, we analyzed two main dynamical features stemming from our prediction of the number of RNAP molecules actively transcribing *hunchback*: the initial rate of RNAP loading and the transcriptional onset time, t_{on} , defined by the slope of the initial rise in the predicted number of RNAP molecules, and the time after anaphase at which transcription starts as determined by the x-intercept of the linear fit to the initial rise, respectively (Fig. 2.3D).

Examples of the predictions generated by our theoretical model are shown in Fig. 2.3E and F, where we calculate the initial rate of RNAP loading and t_{on} for different values of the Bicoid dissociation constant K_b . This framework for quantitatively investigating dynamic input-output functions in living embryos is a necessary step toward testing the predictions of theoretical models of transcriptional regulation in development.

2.2.3 The thermodynamic MWC model fails to predict activation of *hunchback* in the absence of Zelda

In order to test the predictions of the thermodynamic MWC model (Fig. 2.3E and F), we used the MS2 system (Bertrand et al., 1998; Garcia et al., 2013; Lucas et al., 2013). Here, 24 repeats of the MS2 loop are inserted in the 5' untranslated region of the *hunchback* P2 reporter (Garcia et al., 2013), resulting in the fluorescent labeling of sites of nascent transcript formation (Fig. 2.4A; Video A.9.3). This fluorescence is proportional to the number of RNAP molecules actively transcribing the gene (Garcia et al., 2013). The experimental mean fluorescence as a function of time measured in a narrow window (2.5% of the total embryo length, averaged across nuclei in the window) along the length of the embryo (Fig. 2.4B) is in qualitative agreement with the theoretical prediction (Fig. 2.3D).

To compare theory and experiment, we next obtained the initial RNAP loading rates (Fig. 2.4C, blue points) and t_{on} (Fig. 2.4D, blue points) from the experimental data (Section A.2.3; Fig. A.5B). The step-like shape of the RNAP loading rate (Fig. 2.4C, blue points) agrees with previous measurements performed on this same reporter construct (Garcia et al., 2013). The plateaus at the extreme anterior and posterior positions were used to constrain the maximum and minimum theoretically allowed values in the model (Section A.1.3). With these constraints in place, we attempted to simultaneously fit the thermodynamic MWC model to both the initial rate of RNAP loading and t_{on} . For a given set of model parameters, the

measurements of Bicoid and Zelda concentration dynamics predicted a corresponding initial rate of RNAP loading and t_{on} (Fig. 2.3E and F). The model parameters were then iterated using standard curve-fitting techniques (Section 4.5.5) until the best fit to the experimental data was achieved (Fig. 2.4C and D, blue lines).

Although the model accounted for the initial rate of RNAP loading (Fig. 2.4C, blue line), it produced transcriptional onset times that were much lower than those that we experimentally observed (Fig. A.6B, purple line). We hypothesized that this disagreement was due to our model not accounting for mitotic repression, when the transcriptional machinery appears to be silent immediately after cell division (Shermoen and O’Farrell, 1991; Gottesfeld and Forbes, 1997; Parsons and Tg, 1997; Garcia et al., 2013). Thus, we modified the thermodynamic MWC model to include a mitotic repression window term, implemented as a time window at the start of the nuclear cycle during which no transcription could occur; the rate of mRNA production is thus given by

$$\frac{dmRNA}{dt} = \begin{cases} 0 & \text{if } t < t_{MitRep} \\ R p_{bound} & \text{if } t \geq t_{MitRep} \end{cases}, \quad (2.3)$$

where R and p_{bound} are as defined in Eqns. 2.1 and 2.2, respectively, and t_{MitRep} is the mitotic repression time window over which no transcription can take place after anaphase (Sections A.1.2 and A.3). After incorporating mitotic repression, the thermodynamic MWC model successfully fit both the rates of RNAP loading and t_{on} (Fig. 2.4C and D, blue lines, Fig. A.6A and B blue lines).

Given this success, we next challenged the model to perform the simpler task of explaining Bicoid-mediated regulation in the absence of Zelda. This scenario corresponds to setting the concentration of Zelda to zero in the models in Section A.1.2 and Fig. 2.2. In order to test this seemingly simpler model, we repeated our measurements in embryos devoid of Zelda protein (Video A.9.4). These $zelda^-$ embryos were created by inducing clones of non-functional $zelda$ mutant ($zelda^{294}$) germ cells in female adults (Sections 4.5.2, 2.5.3; Liang et al., 2008). All embryos from these mothers lack maternally deposited Zelda; female embryos still have a functional copy of $zelda$ from their father, but this copy is not transcribed until after the maternal-to-zygotic transition, during nuclear cycle 14 (Liang et al., 2008). We confirmed that the absence of Zelda did not have a substantial effect on the spatiotemporal pattern of Bicoid (Section A.4; Xu et al., 2014).

While close to 100% of nuclei in wild-type embryos exhibited transcription along the length of the embryo (Fig. 2.4E, blue; Video A.9.5), measurements in the $zelda^-$ background revealed that some nuclei never displayed any transcription during the entire nuclear cycle (Video A.9.6). Specifically, transcription occurred only in the anterior part of the embryo, with transcription disappearing completely in positions posterior to about 40% of the embryo length (Fig. 2.4E, red). We confirmed that no visible transcription spots were present in $zelda^-$ embryo posteriors by imaging in the posteriors of three $zelda^-$ embryos. These embryos are not included in our total embryo counts.

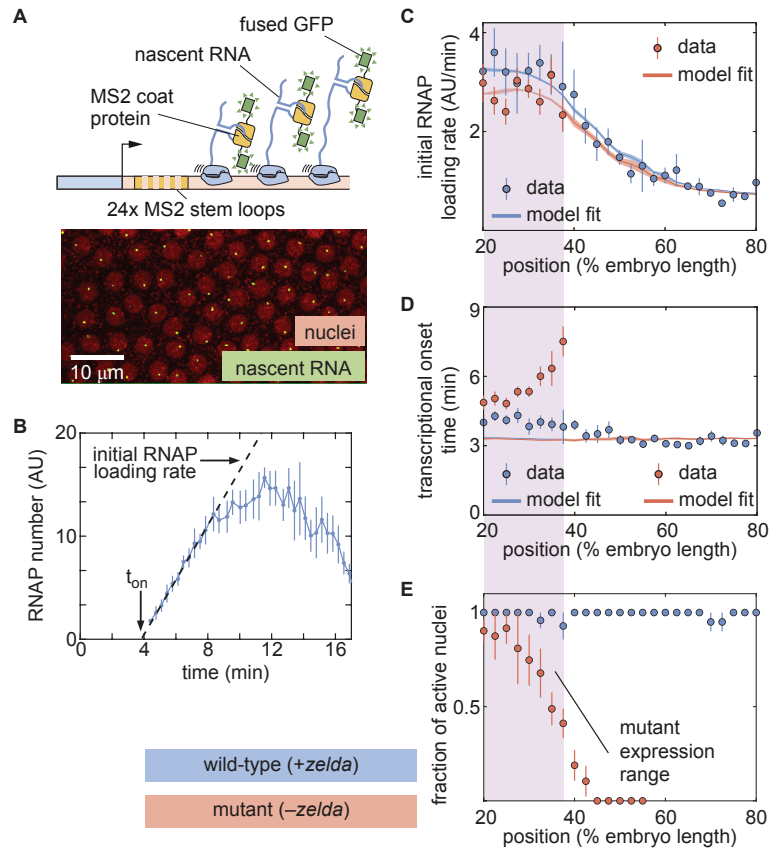


Figure 2.4: The thermodynamic MWC model can explain *hunchback* transcriptional dynamics in wild-type, but not *zelda*⁻, embryos. (A) The MS2 system measures the number of RNAP molecules actively transcribing the *hunchback* reporter gene in live embryos. (B) Representative MS2 trace featuring the quantification of the initial rate of RNAP loading and t_{on} . (C) Initial RNAP loading rate and (D) t_{on} for wild-type (blue points) and *zelda*⁻ (red points) embryos, compared with best fit to the thermodynamic MWC model (lines). The red and blue fit lines are close enough to overlap substantially. (E) Fraction of transcriptionally active nuclei for wild-type (blue) and *zelda*⁻ (red) embryos. Active nuclei are defined as nuclei that exhibited an MS2 spot at any time during the nuclear cycle. Purple shading indicates the spatial range over which at least 30% of nuclei in the *zelda*⁻ background display transcription. (B, error bars are standard error of the mean observed RNAP number, averaged across nuclei in a single embryo; C, D solid lines indicate mean predictions of the model, shading represents standard error of the mean; C, D, E, error bars in data points represent standard error of the mean over 11 wild-type embryos (blue) or 12 *zelda*⁻ embryos (red))

From those positions in the mutant embryos that did exhibit transcription in at least 30% of observed nuclei, we extracted the initial rate of RNAP loading and t_{on} as a function of position. Interestingly, these RNAP loading rates were comparable to the corresponding rates in wild-type embryos (Fig. 2.4C, red points). However, unlike in the wild-type case (Fig. 2.4D, blue points), t_{on} was not constant in the $zelda^-$ background. Instead, t_{on} became increasingly delayed in more posterior positions until transcription ceased posterior to 40% of the embryo length (Fig. 2.4D, red points). Together, these observations indicated that removing Zelda primarily results in a delay of transcription with only negligible effects on the underlying rates of RNAP loading, consistent with previous fixed-embryo experiments (Nien et al., 2011; Foo et al., 2014) and with recent live-imaging measurements in which Zelda binding was reduced at specific enhancers (Dufourt et al., 2018; Yamada et al., 2019). We speculate that the loss of transcriptionally active nuclei posterior to 40% of the embryo length is a direct result of this delay in t_{on} : by the time that onset would occur in those nuclei, the processes leading to the next mitosis have already started and repressed transcriptional activity.

Next, we attempted to simultaneously fit the model to the initial rates of RNAP loading and t_{on} in the $zelda^-$ mutant background. Although the model recapitulated the observed initial RNAP loading rates (Fig. 2.4C, red line), we noticed a discrepancy between the observed and fitted transcriptional onset times of up to ~ 5 min (Fig. 2.4D, red). While the mutant data exhibited a substantial delay in more posterior nuclei, the model did not produce significant delays (Fig. 2.4D, red line). Further, our model could not account for the lack of transcriptional activity posterior to 40% of the embryo length in the $zelda^-$ mutant (Fig. 2.4E, red).

These discrepancies suggest that the thermodynamic MWC model cannot fully describe the transcriptional regulation of the *hunchback* promoter by Bicoid and Zelda. However, the attempted fits in Fig. 2.4C and D correspond to a particular set of model parameters and therefore do not completely rule out the possibility that there exists some parameter set of the thermodynamic MWC model capable of recapitulating the $zelda^-$ data.

In order to determine whether this model is *at all* capable of accounting for the $zelda^-$ transcriptional behavior, we systematically explored how its parameters dictate its predictions. To characterize and visualize the limits of our model, we examined two relevant quantitative features of our data. First, we defined the offset in the transcriptional onset time as the value of the onset time at the position 20% along the embryo length, the most anterior position studied here (Fig. 2.5A), namely

$$\text{offset} = t_{on}(x = 20\%) \quad (2.4)$$

where x is the position along the embryo. Second, we measured the average transcriptional onset delay along the anterior-posterior axis (Fig. 2.5A). This quantity is defined as the area under the curve of t_{on} versus embryo position, from 20% to 37.5% along the embryo (the positions where the $zelda^-$ embryos display transcription in at least 30% of nuclei), divided

by the corresponding distance along the embryo

$$\langle \text{onset delay} \rangle = \frac{1}{37.5\% - 20\%} \int_{20\%}^{37.5\%} (t_{on}(x) - t_{on}(x = 20\%)) dx, \quad (2.5)$$

where the offset in the onset time was used to define the zero of this integral (Section A.5.1). While the offset in t_{on} is similar for both wild-type and $zelda^-$ backgrounds (approximately 4 min), the average t_{on} delay corresponding to the wild-type data is close to 0 min, and is different from the value of about 0.7 min obtained from measurements in the $zelda^-$ background within experimental error (Fig. 2.5C, ellipses).

Based on Estrada et al. (2016) and as detailed in Section A.5.1, we used an algorithm to efficiently sample the parameter space of the thermodynamic MWC model (dissociation constants K_b and K_z , protein-protein interaction terms ω_b and ω_{bp} , energy to make the DNA accessible $\Delta\varepsilon_{\text{chrom}}$, and length of the mitotic repression window t_{MitRep}), and to calculate the corresponding t_{on} offset and average t_{on} delay for each parameter set. Fig. 2.5B features three specific realizations of this parameter search; for each combination of parameters considered, the predicted t_{on} is calculated and the corresponding t_{on} offset and average t_{on} delay computed. Although the wild-type data overlap with the thermodynamic MWC model region, the range of the t_{on} offset and average t_{on} delay predicted by the model (Fig. 2.5C, green) did not overlap with that of the $zelda^-$ data. We concluded that our thermodynamic MWC model is not sufficient to explain the regulation of *hunchback* by Bicoid and Zelda.

2.2.4 No thermodynamic model can recapitulate the activation of *hunchback* by Bicoid alone

Since the failure of the thermodynamic MWC model to predict the $zelda^-$ data does not necessarily rule out the existence of another thermodynamic model that can account for our experimental measurements, we considered other possible thermodynamic models. Conveniently, an arbitrary thermodynamic model featuring n_b Bicoid binding sites can be generalized using the mathematical expression

$$\frac{dmRNA}{dt} = \frac{\left(\sum_{i=0}^{n_b} P_{1,i} R [Bicoid]^i \right)}{p_{inacc} + \sum_{r=0}^1 \sum_{i=0}^{n_b} P_{r,i} [Bicoid]^i}, \quad (2.6)$$

where p_{inacc} and $P_{r,i}$ are *arbitrary* weights describing the states in our generalized thermodynamic model, R is a rate constant that relates promoter occupancy to transcription rate, and the r and i summations refer to the numbers of RNAP and Bicoid molecules bound to the enhancer, respectively (Section A.6.1; Bintu et al., 2005a; Estrada et al., 2016; Scholes et al., 2017). Note, that this generalized thermodynamic model also included the possibility of Bicoid binding to the inaccessible chromatin state (Section A.6.3).

Although this generalized thermodynamic model contains many more parameters than the thermodynamic MWC model previously considered, we could still systematically explore

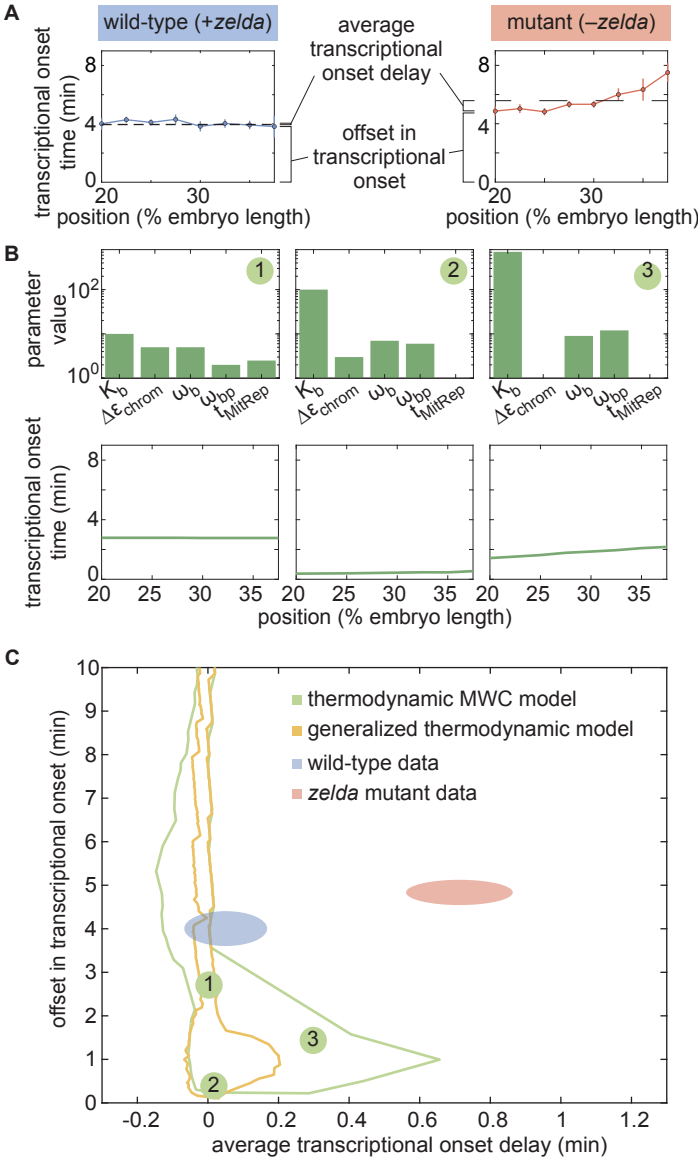


Figure 2.5: Failure of thermodynamic models to describe Bicoid-dependent activation of *hunchback*. See caption on next page.

Figure 2.5: Failure of thermodynamic models to describe Bicoid-dependent activation of *hunchback*. (A) Experimentally determined t_{on} with offset and average delay. Horizontal dashed lines indicate the average t_{on} delay with respect to the offset in t_{on} at 20% along the embryo for wild-type and *zelda*⁻ data sets. (B) Exploration of t_{on} offset and average t_{on} delay from the thermodynamic MWC model. Each choice of model parameters predicts a distinct t_{on} profile of along the embryo. (C) Predicted range of t_{on} offset and average t_{on} delay for the three cases featured in B (green points), for all possible parameter choices of the thermodynamic MWC model (green region), as well as for all thermodynamic models considering 12 Bicoid binding sites (yellow region), compared with experimental data (red and blue regions). (A, C, error bars/ellipses represent standard error of the mean over 11 and 12 embryos for the wild-type and *zelda*⁻ datasets, respectively; B, solid lines indicate mean predictions of the model)

reasonable values of these parameters and the resulting t_{on} offsets and average t_{on} delays (Section A.6.2). For added generality, and to account for recent reports suggesting the presence of more than six Bicoid binding sites in the *hunchback* minimal enhancer (Park et al., 2019), we expanded this model to include up to 12 Bicoid binding sites.

The generalized thermodynamic model also failed to explain the *zelda*⁻ data (Section A.6.2; Fig. 2.5C, yellow). Note that the region of parameter space occupied by the generalized thermodynamic model does not entirely include that of the thermodynamic MWC model due to differences in the constraints of parameter values used in the parameter exploration, as described in Sections A.1.3 and A.6.2. Nevertheless, our results strongly suggest that no thermodynamic model of Bicoid-activated *hunchback* transcription can predict transcriptional onset in the absence of Zelda, casting doubt on the general applicability of these models to transcriptional regulation in development.

Qualitatively, the reason for the failure of thermodynamic models to predict *hunchback* transcriptional is revealed by comparing Bicoid and Zelda concentration dynamics to those of the MS2 output signal (Fig. A.10). The thermodynamic models investigated in this work have assumed that the system responds *instantaneously* to any changes in input transcription factor concentration. As a result, since Bicoid and Zelda are imported into the nucleus by around 3 min into the nuclear cycle (Fig. 2.3A and B), these models always predict that transcription will ensue at approximately that time. Thus, thermodynamic models cannot accommodate delays in the t_{on} such as those revealed by the *zelda*⁻ data (see Section A.6.4 for a more detailed explanation). Rather than further complicating our thermodynamic models with additional molecular players to attempt to describe the data, we instead decided to examine the broader purview of non-equilibrium models to attempt to reach an agreement between theory and experiment.

2.2.5 A non-equilibrium MWC model also fails to describe the *zelda*⁻ data

Thermodynamic models based on equilibrium statistical mechanics can be seen as limiting cases of more general kinetic models that lie out of equilibrium (Section A.6.5; Fig. 2.1B). Following recent reports (Estrada et al., 2016; Li et al., 2018; Park et al., 2019) that the theoretical description of transcriptional regulation in eukaryotes may call for models rooted in non-equilibrium processes – where the assumptions of separation of time scales and no energy expenditure may break down – we extended our earlier models to produce a non-equilibrium MWC model (Sections A.6.5 and A.7.1; Kim and O’Shea, 2008; Narula and Igoshin, 2010). This model, shown for the case of two Bicoid binding sites in Fig. 2.6A, accounts for the dynamics of the MWC mechanism by positing transition rates between the inaccessible and accessible chromatin states, but makes no assumptions about the relative magnitudes of these rates, or about the rates of Bicoid and RNAP binding and unbinding.

Since this model can operate out of steady state, we calculate the probabilities of each state as a function of time by solving the system of coupled ordinary differential equations (ODEs) associated with the system shown in Fig. 2.6A. Consistent with prior measurements (Blythe and Wieschaus, 2016), we assume that chromatin is inaccessible at the start of the nuclear cycle. Over time, the system evolves such that the probability of it occupying each state becomes nonzero, making it possible to calculate the fraction of time RNAP is bound to the promoter and, through the occupancy hypothesis, the rate of RNAP loading. Mitotic repression is still incorporated using the term t_{MitRep} . For times $t < t_{MitRep}$, the system can evolve in time but the ensuing transcription rate is fixed at zero.

We systematically varied the magnitudes of the transition rates and solved the system of ODEs in order to calculate the corresponding t_{on} offset and average t_{on} delay. Due to the combinatorial increase of free parameters as more Bicoid binding sites are included in the model, we could only explore the parameter space for models containing up to five Bicoid binding sites (Section A.7.2; Fig. 2.6B and Fig. A.9). Regardless, none of the non-equilibrium MWC models with up to five Bicoid binding sites came close to reaching the mutant t_{on} offset and average t_{on} delay (Fig. 2.6B). Additionally, an alternative version of this non-equilibrium MWC model where the system could not evolve in time until after the mitotic repression window had elapsed yielded similar conclusions (see Section A.7.3 for details). We conjecture that the observed behavior extends to the biologically relevant case of six or more binding sites. Thus, we conclude that the more comprehensive non-equilibrium MWC model still cannot account for the experimental data, motivating an additional reexamination of our assumptions.

2.2.6 Transcription factor-driven chromatin accessibility can capture all aspects of the data

Since even non-equilibrium MWC models incorporating energy expenditure and non-steady behavior could not explain the *zelda*⁻ data, we further revised the assumptions of our model

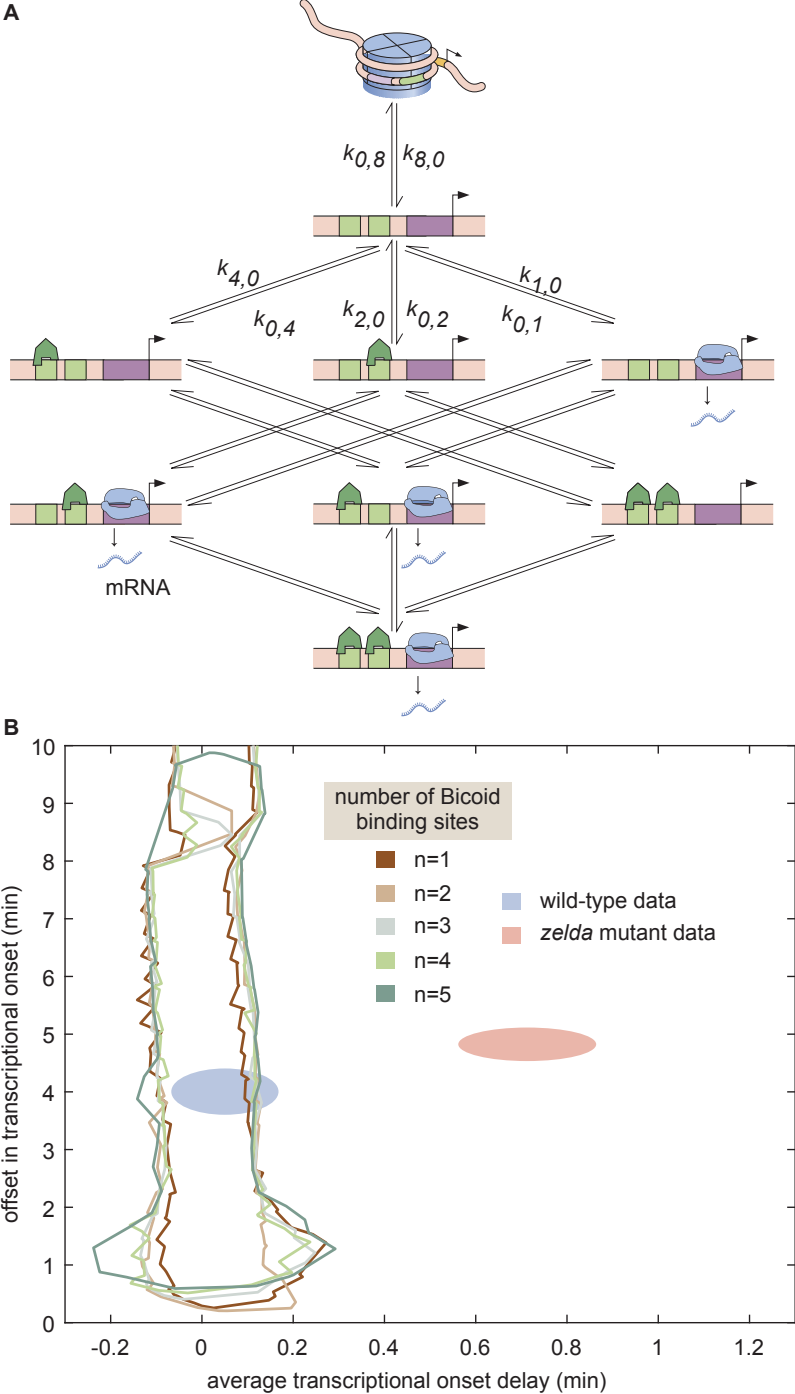


Figure 2.6: Non-equilibrium MWC model of transcriptional regulation cannot predict the observed t_{on} delay. See caption on next page.

Figure 2.6: Non-equilibrium MWC model of transcriptional regulation cannot predict the observed t_{on} delay. (A) Model that makes no assumptions about the relative transition rates between states or about energy expenditure. Each transition rate i, j represents the rate of switching from state i to state j . See Section A.7.1 for details on how the individual states are labeled. (B) Exploration of t_{on} offset and average t_{on} delay attainable by the non-equilibrium MWC models as a function of the number of Bicoid binding sites compared to the experimentally obtained values corresponding to the wild-type and $zelda^-$ mutant backgrounds. While the non-equilibrium MWC model can explain the wild-type data, the exploration reveals that it fails to explain the $zelda^-$ data, for up to five Bicoid binding sites. (B, ellipses represent standard error of the mean over 11 and 12 embryos for the wild-type and $zelda^-$ datasets, respectively)

in an effort to quantitatively predict the regulation of t_{on} along the embryo. In accordance with the MWC model of allostery, all of our theoretical treatments so far have posited that the DNA is an allosteric molecule that transitions between open and closed states as a result of thermal fluctuations (Narula and Igoshin, 2010; Mirny, 2010; Marzen et al., 2013; Phillips et al., 2013).

In the MWC models considered here, the presence of Zelda and Bicoid does not affect the microscopic rates of DNA opening and closing; rather, their binding to open DNA shifts the equilibrium of the DNA conformation toward the accessible state. However, recent biochemical work has suggested that Zelda and Bicoid play a more direct role in making chromatin accessible. Specifically, Zelda has been implicated in the acetylation of chromatin, a histone modification that renders nucleosomes unstable and increases DNA accessibility (Li et al., 2014b; Li and Eisen, 2018). Further, Bicoid has been shown to interact with the co-activator dCBP, which possesses histone acetyltransferase activity (Fu et al., 2004). Additionally, recent studies by Desponds et al. (2016) in *hunchback* and by Dufourt et al. (2018) in *snail* have proposed the existence of multiple transcriptionally silent steps that the promoter needs to transition through before transcriptional onset. These steps could correspond to, for example, the recruitment of histone modifiers, nucleosome remodelers, and the transcriptional machinery (Li et al., 2014b; Park et al., 2019), or to the step-wise unraveling of discrete histone-DNA contacts (Culkin et al., 2017). Further, Dufourt et al. (2018) proposed that Zelda plays a role in modulating the number of these steps and their transition rates.

We therefore proposed a model of transcription factor-driven chromatin accessibility in which, in order for the DNA to become accessible and transcription to ensue, the system slowly and irreversibly transitions through m transcriptionally silent states (Section A.8.1; Fig. 2.7A). We assume that the transitions between these states are all governed by the same rate constant π . Finally, in a stark deviation from the MWC framework, we posit that these

transitions can be catalyzed by the presence of Bicoid and Zelda such that

$$\pi = c_b[\textit{Bicoid}] + c_z[\textit{Zelda}]. \quad (2.7)$$

Here, π describes the rate (in units of inverse time) of each irreversible step, expressed as a sum of rates that depend separately on the concentrations of Bicoid and Zelda, and c_b and c_z are rate constants that scale the relative contribution of each transcription factor to the overall rate (see Section A.8.2 for a more detailed discussion of this choice). We emphasize that this is only one potential model, and there may exist several other non-equilibrium models capable of describing our data.

In this model of transcription factor-driven chromatin accessibility, once the DNA becomes irreversibly accessible after transitioning through the m non-productive states, we assume that, for the rest of the nuclear cycle, the system equilibrates rapidly such that the probability of it occupying any of its possible states is still described by equilibrium statistical mechanics. Like in our previous models, transcription only occurs in the RNAP-bound states, obeying the occupancy hypothesis. Further, our model assumes that if the transcriptional onset time of a given nucleus exceeds that of the next mitosis, this nucleus will not engage in transcription. Thus, this transcription factor-driven model is an extension of the non-equilibrium MWC model with two crucial differences: (i) we allow for multiple inaccessible states preceding the transcriptionally active state, and (ii) the transitions between these states are *actively* driven by Bicoid or Zelda.

Unlike the thermodynamic and non-equilibrium MWC models, this model of transcription factor-driven chromatin accessibility quantitatively recapitulated the observation that posterior nuclei in *zelda*⁻ embryos do not engage in transcription as well as the initial rate of RNAP loading, and t_{on} for both the wild-type and *zelda*⁻ mutant data (Fig. 2.7B and C). Additionally, we found that a minimum of $m = 3$ steps was required to sufficiently explain the data (Section A.8.3; Fig. A.14). Interestingly, unlike all previously considered models, the model of transcription factor-driven chromatin accessibility did not require mitotic repression to explain t_{on} (Sections A.3 and A.8.1). Instead, the timing of transcriptional output arose directly from the model's initial irreversible transitions (Fig. A.14), obviating the need for an arbitrary suppression window in the beginning of the nuclear cycle. The only substantive disagreement between our theoretical model and the experimental data was that the model predicted that no nuclei should transcribe posterior to 60% of the embryo length, whereas no transcription posterior to 40% was experimentally observed in the embryo (Fig. 2.7B and C). Finally, note that this model encompasses a much larger region of parameter space than the thermodynamic and non-equilibrium MWC models and, as expected from the agreement between model and experiment described above, contained both the wild-type and *zelda*⁻ data points within its domain (Fig. 2.7D).

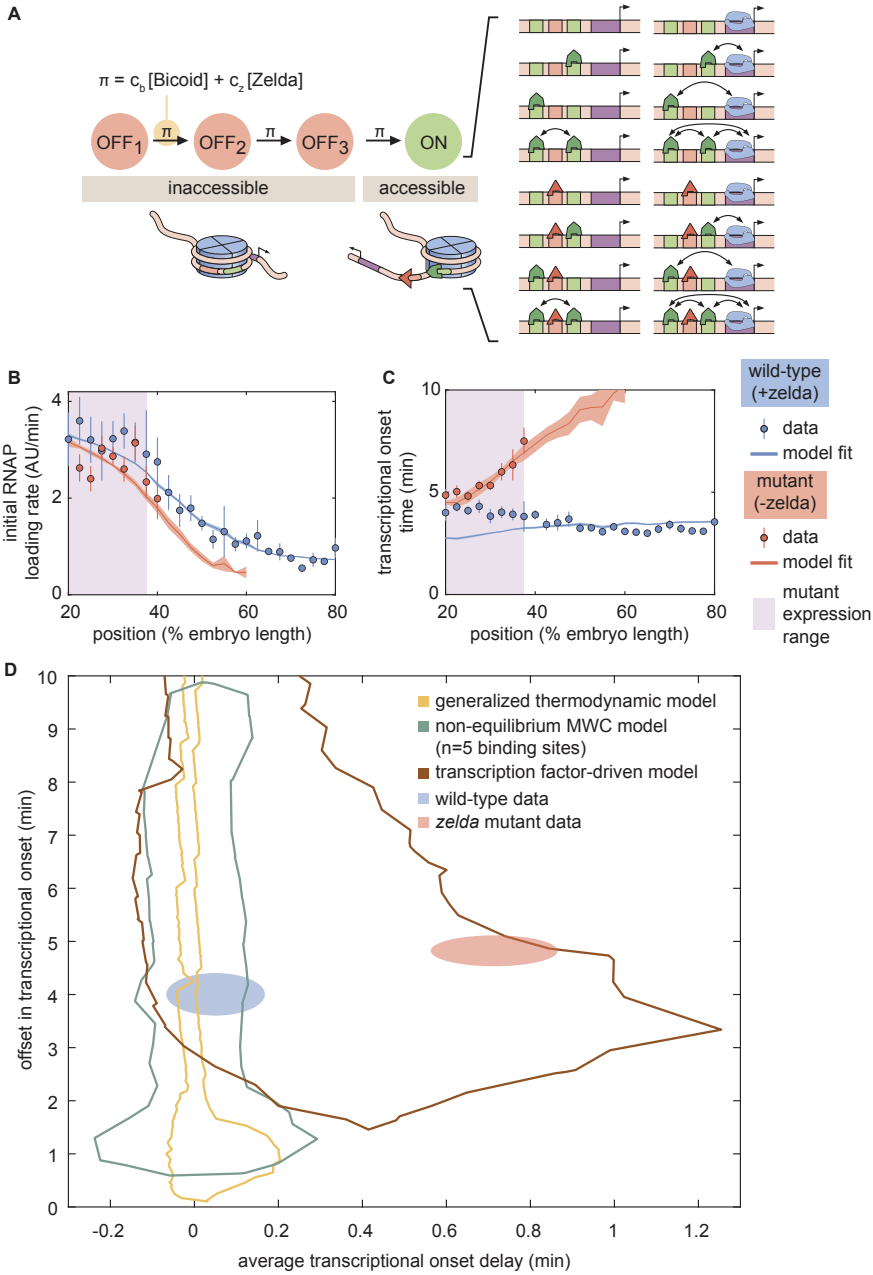


Figure 2.7: A model of transcription factor-driven chromatin accessibility is sufficient to recapitulate *hunchback* transcriptional dynamics. See caption on next page.

Figure 2.7: A model of transcription factor-driven chromatin accessibility is sufficient to recapitulate *hunchback* transcriptional dynamics. (A) Overview of the proposed model, with three ($m = 3$) effectively irreversible Zelda and/or Bicoid-mediated intermediate transitions from the inaccessible to the accessible state. (B, C) Experimentally fitted (B) initial RNAP loading rates and (C) t_{on} for wild-type and *zelda*⁻ embryos using a single set of parameters and assuming six Bicoid binding sites. (D) The domain of t_{on} offset and average t_{on} delay covered by this transcription factor-driven chromatin accessibility model (brown) is much larger than those of the generalized thermodynamic model (yellow) and the non-equilibrium MWC models (green), and easily encompasses both experimental datasets (ellipses). (B-D, error bars/ellipses represent standard error of the mean over 11 and 12 embryos for the wild-type and *zelda*⁻ datasets, respectively)

2.3 Discussion

For four decades, thermodynamic models rooted in equilibrium statistical mechanics have constituted the null theoretical model for calculating how the number, placement and affinity of transcription factor binding sites on regulatory DNA dictates gene expression (Bintu et al., 2005a,b). Further, the MWC mechanism of allostery has been proposed as an extra layer that allows thermodynamic and more general non-equilibrium models to account for the regulation of chromatin accessibility (Mirny, 2010; Narula and Igoshin, 2010; Marzen et al., 2013).

In this investigation, we tested thermodynamic and non-equilibrium MWC models of chromatin accessibility and transcriptional regulation in the context of *hunchback* activation in the early embryo of the fruit fly *D. melanogaster* (Driever et al., 1989; Nien et al., 2011; Xu et al., 2014). While chromatin state (accessibility, post-translational modifications) is highly likely to influence transcriptional dynamics of associated promoters, specifically measuring the influence of chromatin state on transcriptional dynamics is challenging because of the sequential relationship between changes in chromatin state and transcriptional regulation. However, the *hunchback* P2 minimal enhancer provides a unique opportunity to dissect the relative contribution of chromatin regulation on transcriptional dynamics because, in the early embryo, chromatin accessibility at *hunchback* is granted by both Bicoid and Zelda (Hannon et al., 2017). The degree of *hunchback* transcriptional activity, however, is regulated directly by Bicoid (Driever and Nusslein-Volhard, 1989; Driever et al., 1989; Struhl et al., 1989). Therefore, while genetic elimination of Zelda function interferes with acquisition of full chromatin accessibility, the *hunchback* locus retains a measurable degree of accessibility and transcriptional activity stemming from Bicoid function, allowing for a quantitative determination of the contribution of Zelda-dependent chromatin accessibility on the transcriptional dynamics of the locus.

With these attributes in mind, we constructed a thermodynamic MWC model which,

given a set of parameters, predicted an output rate of *hunchback* transcription as a function of the input Bicoid and Zelda concentrations (Fig. 2.2B). In order to test this model, it was necessary to acknowledge that development is not in steady-state, and that both Bicoid and Zelda concentrations change dramatically in space and time (Fig. 2.3A and B). As a result, we went beyond widespread steady-state descriptions of development and introduced a novel approach that incorporated transient dynamics of input transcription-factor concentrations in order to predict the instantaneous output transcriptional dynamics of *hunchback* (Fig. 2.3C). Given input dynamics quantified with fluorescent protein fusions to Bicoid and Zelda, we both predicted output transcriptional activity and measured it with an MS2 reporter (Figs. 2.3D and 2.4B).

This approach revealed that the thermodynamic MWC model sufficiently predicts the timing of the onset of transcription and the subsequent initial rate of RNAP loading as a function of Bicoid and Zelda concentration. However, when confronted with the much simpler case of Bicoid-only regulation in a *zelda* mutant, the thermodynamic MWC model failed to account for the observations that only a fraction of nuclei along the embryo engaged in transcription, and that the transcriptional onset time of those nuclei that do transcribe was significantly delayed with respect to the wild-type setting (Fig. 2.4D and E). Our systematic exploration of all thermodynamic models (over a reasonable parameter range) showed that no thermodynamic model featuring regulation by Bicoid alone could quantitatively recapitulate the measurements performed in the *zelda* mutant background (Fig. 2.5C, yellow).

This disagreement could be resolved by invoking an unknown transcription factor that regulates the *hunchback* reporter in addition to Bicoid. However, at the early stages of development analyzed here, such a factor would need to be both maternally provided and patterned in a spatial gradient to produce the observed position-dependent transcriptional onset times. To our knowledge, none of the known maternal genes regulate the expression of this *hunchback* reporter in such a fashion (Chen et al., 2012; Perry et al., 2012; Xu et al., 2014). We conclude that the MWC thermodynamic model cannot accurately predict *hunchback* transcriptional dynamics.

To explore non-equilibrium models, we retained the MWC mechanism of chromatin accessibility, but did not demand that the accessible and inaccessible states be in thermal equilibrium. Further, we allowed for the process of Bicoid and RNAP binding, as well as their interactions, to consume energy. For up to five Bicoid binding sites, no set of model parameters could quantitatively account for the transcriptional onset features in the *zelda* mutant background (Fig. 2.6B). While we were unable to investigate models with more than five Bicoid binding sites due to computational complexity (Estrada et al., 2016), the substantial distance in parameter space between the mutant data and the investigated models (Fig. 2.6B) suggested that a successful model with more than five Bicoid binding sites would probably operate near the limits of its explanatory power, similar to the conclusions from studies that explored *hunchback* regulation under the steady-state assumption (Park et al., 2019). Thus, despite the simplicity and success of the MWC model in predicting the effects of protein allostery in a wide range of biological contexts (Keymer et al., 2006; Swem et al., 2008; Martins and Swain, 2011; Marzen et al., 2013; Rapp and Yifrach, 2017; Razo-Mejia

et al., 2018; Chure et al., 2019; Rapp and Yifrach, 2019), the observed transcriptional onset times could not be described by any previously proposed thermodynamic MWC mechanism of chromatin accessibility, or even by a more generic non-equilibrium MWC model in which energy is continuously dissipated (Tu, 2008; Kim and O’Shea, 2008; Narula and Igoshin, 2010; Estrada et al., 2016; Wang et al., 2017).

Since Zelda is associated with histone acetylation, which is correlated with increased chromatin accessibility (Li et al., 2014b; Li and Eisen, 2018), and Bicoid interacts with the co-activator dCBP, which has histone acetyltransferase activity (Fu et al., 2004; Fu and Ma, 2005; Park et al., 2019), we suspect that both Bicoid and Zelda actively drive DNA accessibility. A molecular pathway shared by Bicoid and Zelda to render chromatin accessible is consistent with our results, and with recent genome-wide experiments showing that Bicoid can rescue the function of Zelda-dependent enhancers at high enough concentrations (Hannon et al., 2017). Thus, the binding of Bicoid and Zelda, rather than just biasing the equilibrium toward the open chromatin state as in the MWC mechanism, may trigger a set of molecular events that locks DNA into an accessible state. In addition, the promoters of *hunchback* (Desponds et al., 2016) and *snail* (Dufourt et al., 2018) may transition through a set of intermediate, non-productive states before transcription begins.

We therefore explored a model in which Bicoid and Zelda catalyze the transition of chromatin into the accessible state via a series of slow, effectively irreversible steps. These steps may be interpreted as energy barriers that are overcome through the action of Bicoid and Zelda, consistent with the coupling of these transcription factors to histone modifiers, nucleosome remodelers (Fu et al., 2004; Li et al., 2014b; Li and Eisen, 2018; Park et al., 2019), and with the step-wise breaking of discrete histone-DNA contacts to unwrap nucleosomal DNA (Culkin et al., 2017). In this model, once accessible, the chromatin remains in that state and the subsequent activation of *hunchback* by Bicoid is described by a thermodynamic model.

Crucially, this transcription factor-driven chromatin accessibility model successfully replicated all of our experimental observations. A minimum of three transcriptionally silent states were necessary to explain our data (Fig. 2.7D and Fig. A.14C). Interestingly, recent work dissecting the transcriptional onset time distribution of *snail* also suggested the existence of three such intermediate steps in the context of that gene (Dufourt et al., 2018). Given that, as in *hunchback*, the removal and addition of Zelda modulates the timing of transcriptional onset of *sog* and *snail* (Dufourt et al., 2018; Yamada et al., 2019), we speculate that transcription factor-driven chromatin accessibility may also be at play in these pathways. Thus, taken in consideration with similar works examining the dynamics of transcription onset (Desponds et al., 2016; Dufourt et al., 2018; Fritzsche et al., 2018; Li et al., 2018), our results strongly suggest that chromatin state does not fluctuate thermodynamically, but rather progresses through a series of stepwise, transcription factor-driven transitions into a final RNAP-accessible configuration (Coulon et al., 2013).

Intriguingly, accounting for these intermediate states also obviated the need for the *ad hoc* imposition of a mitotic repression window (Sections A.3 and A.8.1), which was required in the thermodynamic MWC model (Fig. A.6). Our results thus suggest a mechanistic

interpretation of the phenomenon of mitotic repression after anaphase, where the promoter must traverse through intermediary transcriptionally silent states before transcriptional onset can occur.

These clues into the molecular mechanisms of action of Bicoid, Zelda, and their associated modifications to the chromatin landscape pertain to a time scale of a few minutes, a temporal scale that is inaccessible with widespread genome-wide and fixed-tissue approaches. Here, we revealed the regulatory action of Bicoid and Zelda by utilizing the dynamic information provided by live imaging to analyze the transient nature of the transcriptional onset time, highlighting the need for descriptions of development that go beyond steady state and acknowledge the highly dynamic changes in transcription-factor concentrations that drive developmental programs.

While we showed that one model incorporating transcription factor-driven chromatin accessibility could recapitulate *hunchback* transcriptional regulation by Bicoid and Zelda, and is consistent with molecular evidence on the modes of action of these transcription factors, other models may have comparable explanatory power. In the future, a systematic exploration of different classes of models and their unique predictions will identify measurements that determine *which* specific model is the most appropriate description of transcriptional regulation in development and *how* it is implemented at the molecular level. While all the analyses in this work relied on mean levels of input concentrations and output transcription levels, detailed studies of single-cell features of transcriptional dynamics such as the distribution of transcriptional onset times (Narula and Igoshin, 2010; Dufourt et al., 2018; Fritzsche et al., 2018) could shed light on these chromatin-regulating mechanisms. Simultaneous measurement of local transcription-factor concentrations at sites of transcription and of transcriptional initiation with high spatiotemporal resolution, such as afforded by lattice light-sheet microscopy (Mir et al., 2018), could provide further information about chromatin accessibility dynamics. Finally, different theoretical models may make distinct predictions about the effect of modulating the number, placement, and affinity of Bicoid and Zelda sites (and even of nucleosomes) in the *hunchback* enhancer. These models could be tested with future experiments that implement these modulations in reporter constructs.

In sum, here we engaged in a theory-experiment dialogue to respond to the theoretical challenges of proposing a passive MWC mechanism for chromatin accessibility in eukaryotes (Mirny, 2010; Narula and Igoshin, 2010; Marzen et al., 2013); we also questioned the suitability of thermodynamic models in the context of development (Estrada et al., 2016). At least regarding the activation of *hunchback*, and likely similar developmental genes such as *snail* and *sog* (Dufourt et al., 2018; Yamada et al., 2019), we speculate that Bicoid and Zelda actively drive chromatin accessibility, possibly through histone acetylation. Once chromatin becomes accessible, thermodynamic models can predict *hunchback* transcription without the need to invoke energy expenditure and non-equilibrium models. Regardless of whether we have identified the only possible model of chromatin accessibility and regulation, we have demonstrated that this dialogue between theoretical models and the experimental testing of their predictions at high spatiotemporal resolution is a powerful tool for biological discovery. The new insights afforded by this dialogue will undoubtedly refine theoretical descriptions

of transcriptional regulation as a further step toward a predictive understanding of cellular decision-making in development.

2.4 Acknowledgments

We are grateful to Jack Bateman, Jacques Bothma, Mike Eisen, Jeremy Gunawardena, Jane Kondev, Oleg Igoshin, Rob Phillips, Christine Rushlow and Peter Whitney for their guidance and comments on our manuscript. We thank Kenneth Irvine and Yuanwang Pan for providing the *his-irfp* fly line. This work was supported by the Burroughs Wellcome Fund Career Award at the Scientific Interface, the Sloan Research Foundation, the Human Frontiers Science Program, the Searle Scholars Program, the Shurl and Kay Curci Foundation, the Hellman Foundation, the NIH Director's New Innovator Award (DP2 OD024541-01), and an NSF CAREER Award (1652236) (HGG), an NSF GRFP (DGE 1752814) and UC Berkeley Chancellor's Fellowship (EE), and the DoD NDSEG graduate fellowship (JL).

2.5 Methods and Materials

2.5.1 Predicting Zelda binding sites

Zelda binding sites in the *hunchback* promoter were identified as heptamers scoring 3 or higher using a Zelda alignment matrix (Harrison et al., 2011) and the Advanced PASTER entry form online (<http://stormo.wustl.edu/consensus/cgi-bin/Server/Interface/patser.cgi>) (Hertz et al., 1990; Hertz and Stormo, 1999). PATSER was run with setting "Seq. Alphabet and Normalization" as "a:t 3 g:c 2" to provide the approximate background frequencies as annotated in the Berkeley Drosophila Genome Project (BDGP)/Celera Release 1. Reverse complementary sequences were also scored.

2.5.2 Fly Strains

Bicoid nuclear concentration was imaged in embryos from line *yw; his2av-mrfp1;bicoidE1, egfp-bicoid* (Gregor et al., 2007b). Similarly, Zelda nuclear concentration was determined by imaging embryos from line *sfgfp-zelda;+;his-irfp*. The *sfgfp-zelda* transgene was obtained from Hamm et al. (2017) and the *his-iRFP* transgene is courtesy of Kenneth Irvine and Yuanwang Pan.

Transcription from the *hunchback* promoter was measured by imaging embryos resulting from crossing female virgins *yw;HistoneRFP;MCP-NoNLS(2)* with male *yw;P2P-MS2-LacZ/cyo;+* (Garcia et al., 2013).

In order to image transcription in embryos lacking maternally deposited Zelda protein, we crossed mother flies whose germline was *w, his2av-mrfp1,zelda(294),FRT19A;+;MCP-egfp(4F)/+* obtained through germline clones (see below) with fathers carrying the *yw;P2P-MS2-LacZ/cyo;+* reporter. The *zelda(294)* transgene is courtesy of Christine Rushlow (Liang

et al., 2008). The *MCP-egfp(4F)* transgene expresses approximately double the amount of MCP than the *MCP-egfp(2)* (Garcia et al., 2013), ensuring similar levels of MCP in the embryo in all experiments.

Imaging Bicoid nuclear concentration in embryos lacking maternally deposited Zelda protein was accomplished by replacing the *MCP-egfp(4F)* transgene described in the previous paragraph with the *bicoidE1, egfp-bicoid* transgene used for imaging nuclear Bicoid in a wildtype background. We crossed mother flies whose germline was *w, his2av-mrfp1, zelda(294), FRT19A; +; bicoidE1, egfp-bicoid/+* obtained through germline clones (see below) with *yw* fathers.

2.5.3 Zelda germline clones

In order to generate mother flies containing a germline homozygous null for *zelda*, we first crossed virgin females of *w, his2av-mrfp1, zelda(294), FRT19A/FM7, y, B; +; MCP-egfp(4F)/TM3, ser* (or *w, his2av-mrfp1, zelda(294), FRT19A; +; bicoidE1, egfp-bicoid/+* to image nuclear Bicoid) with males of *ovoD, hs-FLP, FRT19A; +; +* (Liang et al., 2008). The resulting heterozygotic offspring were heat-shocked in order to create maternal germline clones as described in Liang et al. (2008). The resulting female virgins were crossed with male *yw; P2P-MS2-LacZ/cyo; +* (Garcia et al., 2013) to image transcription or male *yw* to image nuclear Bicoid concentration.

Male offspring are null for zygotic *zelda*. Female offspring are heterozygotic for functional *zelda*, but zygotic *zelda* is not transcribed until nuclear cycle 14 (Liang et al., 2008), which occurs after the analysis in this work. All embryos lacking maternally deposited Zelda showed aberrant morphology in nuclear size and shape (data not shown), as previously reported (Liang et al., 2008; Staudt et al., 2006).

2.5.4 Sample preparation and data collection

Sample preparation followed procedures described in Bothma et al. (2014), Garcia and Gregor (2018), and Lammers et al. (2020).

Embryos were collected and mounted in halocarbon oil 27 between a semipermeable membrane (Lumox film, Starstedt, Germany) and a coverslip. Data collection was performed using a Leica SP8 scanning confocal microscope (Leica Microsystems, Biberach, Germany). Imaging settings for the MS2 experiments were the same as in Lammers et al. (2020), except the Hybrid Detector (HyD) for the His-RFP signal used a spectral window of 556-715 nm. The settings for the Bicoid-GFP measurements were the same, except for the following. The power setting for the 488 nm line was 10 μ W. The confocal stack was only 10 slices in this case, rather than 21, resulting in a spacing of 1.11 μ m between planes. The images were acquired at a time resolution of 30 s, using an image resolution of 512 x 128 pixels.

The settings for the Zelda-sfGFP measurements were the same as the Bicoid-GFP measurements, except different laser lines were used for the different fluorophores. The sf-GFP excitation line was set at 485 nm, using a power setting of 10 μ W. The His-iRFP

excitation line was set at 670 nm. The HyD for the His-iRFP signal was set at a 680-800 nm spectral window. All specimens were imaged over the duration of nuclear cycle 13.

2.5.5 Image analysis

Images were analyzed using custom-written software following the protocol in Garcia et al. (2013). Briefly, this procedure involved segmenting individual nuclei using the histone signal as a nuclear mask, segmenting each transcription spot based on its fluorescence, and calculating the intensity of each MCP-GFP transcriptional spot inside a nucleus as a function of time.

Additionally, the nuclear protein fluorescences of the Bicoid-GFP and Zelda-sfGFP fly lines were calculated as follows. Using the histone-labeled nuclear mask for each individual nucleus, the fluorescence signal within the mask was extracted in xyz, as well as through time. For each timepoint, the xy signal was averaged to give an average nuclear fluorescence as a function of z and time. This signal was then maximum projected in z, resulting in an average nuclear concentration as a function of time, per single nucleus. These single nucleus concentrations were then averaged over anterior-posterior position to create the protein concentrations reported in the main text.

2.5.6 Data Analysis

All fits in the main text were performed by minimizing the least-squares error between the data and the model predictions. Unless stated otherwise, error bars reflect standard error of the mean over multiple embryo measurements. See Section A.2.1 for more details on how this was carried out for model predictions.

Chapter 3

The role of transient transcription factor inputs in models of chromatin accessibility

Foreword

After wrapping up the project from Chapter 2, it was clear that modeling chromatin accessibility would require us to think deeply about the non-steady dynamics of transcription in development. Instead of the traditional equilibrium picture of transcriptional regulation, which abstracts away the time dimension, our working model at the end of Chapter 2 made use of a Markov chain formalism to describe the onset of transcription at the start of a nuclear cycle.

Most importantly, it appeared that the dynamics of input transcription factor concentrations—here, *Zelda* and *Bicoid*—played a huge role in determining transcription onset times. From a theoretical standpoint, this was extremely interesting, albeit difficult to work with. Indeed, while Markov chain formalisms are foundational to the modeling of waiting time distributions, the presence of transients—in this case, the coupling of time-dependent concentrations to transition rates—makes theoretical analysis difficult, if not intractable.

Nevertheless, we wondered if these transient input factor dynamics could be important, not just in terms of modulating transcription onset times. Could we investigate the role of these transients from the lens of developmental precision? In development, synchrony of transcription is of utmost importance in order to precisely choreograph patterns of gene expression in time as well as space. However, steady-state Markov chain dynamics come with clear constraints on the noise in first passage time distributions, imposing fundamental restrictions to synchrony in transcription onset times.

The core hypothesis of this project was that, compared to steady state, transient dynamics could reduce noise in first passage time distributions in a Markov chain model of chromatin accessibility. What follows is a short theoretical investigation of this premise.

3.1 Introduction

The results from Chapter 2 as well as from prior works investigating non-equilibrium transcriptional regulation in eukaryotes (Estrada et al., 2016; Li et al., 2018; Dufourt et al., 2018) indicate a growing need for models that can explain the non-steady-state behavior exhibited by systems such as the developing fly embryo. Indeed, our work (Eck et al., 2020) as well as others (Dufourt et al., 2018; Fritsch et al., 2018) suggest that metrics such as transcription onset times can provide clear phenotypes for distinguishing between equilibrium and non-equilibrium behavior. These types of analyses provide the first glimpses into the importance of *transients* in transcriptional regulation, an importance facet (Wong and Gunawardena, 2020) often neglected in earlier studies of non-equilibrium systems.

Our analysis in Chapter 2 considered the impact of pioneering transcription factor dynamics on transcription onset times in a simple model of chromatin accessibility. There, we restricted ourselves to mean-level investigations, which was sufficient to indicate some direct action of pioneering factors such as Zelda or Bicoid in the opening of chromatin for transcriptional competence.

However, in a population of cells such as in a developing fly embryo, the *distribution* of transcription onset times across cells contains much more information than the mean onset time alone. Indeed, other works investigating these distributions have been able to constrain parameters such as the number of state transitions in a model and their corresponding transition rates (Dufourt et al., 2018; Fritsch et al., 2018). By utilizing single-cell statistics, these works could fit theoretical probability distribution functions of transcription onset times to single-cell empirical distributions. Since these fits are constrained by higher-order moments of a probability distribution, they provide higher precision than simple fits to population-averaged means. Nevertheless, these works did not consider non-steady concentrations of the input transcription factors themselves (Fig. 2.3A and B), unlike our analysis in Chapter 2.

Thus, a unified analysis of transcription onset times should investigate both the impact of single-cell distributions as well as of transient input transcription factor dynamics. To our knowledge, such a study has yet to be carried out, although there is ample theoretical and experimental opportunity for this type of analysis to inform on models of transcriptional regulation in development.

In this Chapter, we begin by constructing a simple steady-state Markov chain model of pioneering factor-mediated chromatin accessibility. After an initial exploration of the model's behavior, we define a feature space to characterize the resulting distributions of single-cell transcription onset times and identify fundamental limits to the synchrony of timing across cells in the steady-state regime. Next, we show that, by considering transient input transcription factor dynamics, the resulting distributions of transcription onset times can exceed these limits.

Our theoretical insights carry two main implications. First, our results indicate that the transient nature of transcription factor dynamics not only impacts distributions of transcription onset times, but also can help improve important developmental outcomes such as transcriptional synchrony. This suggests that these dynamics are biophysical quantities

that could be tuned or even optimized by biological systems for improved evolutionary fitness. Second, our ability to probe and constrain models of chromatin accessibility increases with the addition of input transcription factor dynamics as a new tunable parameter. With the advent of technologies to control transcription factor dynamics, such as optogenetics (Johnson and Toettcher, 2018), experimental validation of the theoretical models explored in this work enter the realm of possibility, if not plausibility.

3.2 Results

3.2.1 A simple steady-state Markov chain model of chromatin accessibility

Inspired by our model of transcription factor-mediated chromatin accessibility presented at the end of Chapter 2, we begin by developing a simple steady-state model of transcription onset times based on Markov chain dynamics. Ultimately, we are interested in investigating the synchrony of transcription onset times, quantified by the spread in the distribution of onset times produced by the model.

We consider a Markov chain with $k + 1$ states labeled with indices i , with the first state labeled with index 0. The system begins in state 0 at time $t = 0$ and we will assume the final state $i = k$ is absorbing. Thus, the first k states correspond to transcriptionally OFF states, while the final ($k + 1$)th state is transcriptionally active—reaching the final state results in the onset of transcription. Prior works investigating these types of chromatin accessibility models have conjectured a minimum of $k = 3$ or so OFF states (Dufourt et al., 2018; Eck et al., 2020).

We will allow for forwards and backwards transition rates between all states, except for the final absorbing state, which will have no backwards transition out of it. Denote the transition from state i to state j with the transition rate $\beta_{i,j}$ and assume that these transition rates are constant in time. Thus, we have the reaction network



In the transcription factor-mediated picture of chromatin accessibility, the forward rates would be coupled to concentrations of an input pioneering factor such as Zelda or Bicoid in an on-rate-like fashion. We assume that the backwards rates are independent of transcription factor concentrations in an off-rate-like fashion.

We will be interested in the mean and variance of the distribution of transcription onset times — the time required to start at state 0 and reach the final state k . To make this possible, we will first consider the simple case where we only have forward transition rates β that are equal in magnitude, constant in time, and independent from each other (Fig. 3.1A). In this case, the probability density $P_k(t)$ is simply given by a Gamma distribution with shape parameter k and rate parameter β , which inform on the number of states and the

speed of transitions, respectively. $P_k(t)$ then has the form

$$P_k(t) = \frac{\beta^k}{(k-1)!} t^{k-1} e^{-\beta t}, \quad (3.2)$$

where Γ is the Gamma function. Figure 3.1B shows analytical and simulated results for this model's distribution of transcription onset times. The mean μ_k and variance σ_k^2 of these distributions have simple analytical expressions and are given by

$$\mu_k = \frac{k}{\beta} \quad (3.3)$$

and

$$\sigma_k^2 = \frac{k}{\beta^2}, \quad (3.4)$$

respectively. For this simple model, both the mean and variance are simple functions of k and β , allowing for straightforward analytical investigation.

To investigate the distribution of transcription onset times, we consider a two-dimensional feature space consisting of the mean onset time on the x-axis and the squared CV (variance divided by square mean) in the onset time on the y-axis. The squared CV is a measure of the “noise” of the system at a given mean. For this simple model, we have

$$\mu_k = \frac{k}{\beta} \quad (3.5)$$

and

$$CV_k^2 = \frac{1}{k}. \quad (3.6)$$

Thus, for this model consisting of equal, irreversible reactions, the squared CV is independent of the transition rates β and depends only on the number of steps k . Plotting the CV and mean in our feature space defined above results in a series of horizontal lines, with each line corresponding to the particular number of steps in the model (Fig. 3.1C). For a more realistic scenario with finite statistics, numerical simulations are shown in colored points. So, in this model by measuring the CV and the mean we can infer the number of states and the transition rates.

3.2.2 Extending the steady-state model to account for backwards transitions

The model presented above, while simple, is highly idealized in that it only considers equal, irreversible transitions. Next, we consider a model where we now allow for backwards transitions as well as forward transitions (Fig. 3.2A). We retain the idea of equal forward transition rates β , but now allow for equal backwards transitions of magnitude βf (except from the final absorbing state k). Here, f is constrained to be non-negative and parameterizes

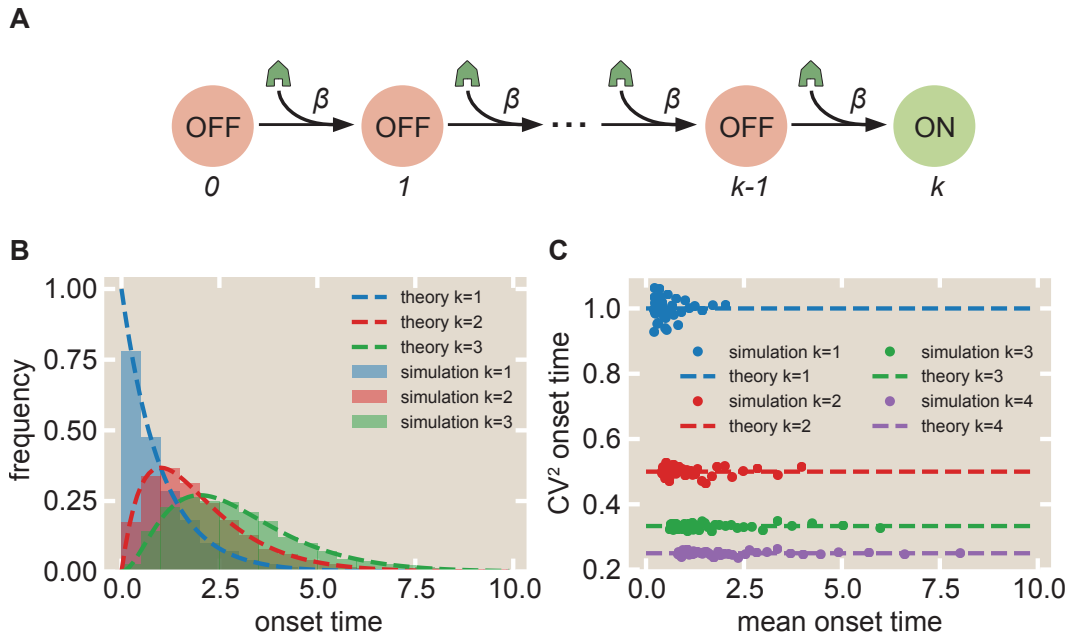


Figure 3.1: Results of steady state input model with equal, irreversible transitions. (A) Cartoon illustration of model. (B) Simulated and analytical distributions of onset times for varying numbers of steps k and transition rate magnitude $\beta = 1$. (C) Simulated and analytical results of model in feature space consisting of mean and squared CV in onset times, for varying numbers of steps k and transition rate magnitude β ranging from 0.5 to 5.

the fractional magnitude of the backwards rate compared to the forwards. The model thus consists of the following states and transitions

$$0 \xrightleftharpoons[\beta f]{\beta} 1 \xrightleftharpoons[\beta f]{\beta} \dots \xrightarrow{\beta} k. \quad (3.7)$$

Intuitively, as f increases, the relative strength of the backwards rates increases, so the distribution of onset times should broaden as the time to reach the final state becomes more stochastic. Figure 3.2B shows some distributions of transcription onset times in this model, for $k = 3$ steps and varying f . As expected, increasing f broadens the distributions.

Figure 3.2C shows these same results on the two-dimensional feature space. We see that as the backwards transition rate increases, the overall noise increases (colored points). Qualitatively, with a backwards transition rate the system is more likely to spend extra time hopping between states before reaching the final absorbing state, increasing the overall time to finish as well as the variability in finishing times.

Because actual irreversible reactions are effectively impossible to achieve in reality, the performance of the Gamma distribution model (i.e. equal, irreversible forward transitions)

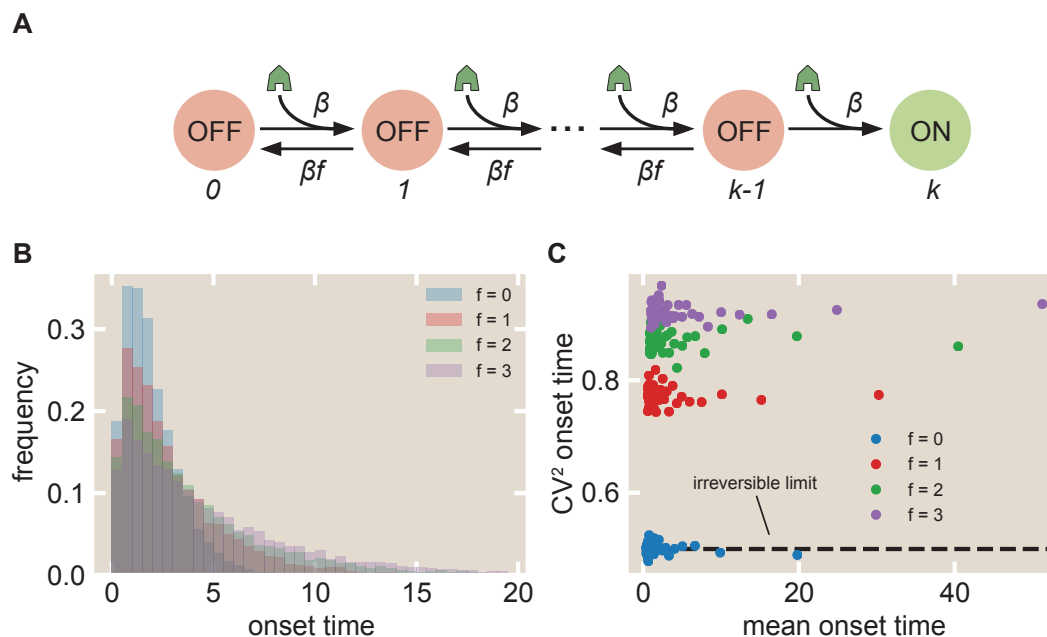


Figure 3.2: Results of steady state input model with equal, reversible transitions. (A) Cartoon illustration of model. (B) Simulated distributions of onset times for varying backwards rate fraction f , $k = 3$ steps, and transition rate magnitude $\beta = 1$. (C) Results of model with $k = 3$ steps in feature space consisting of mean and squared CV in onset times, for varying f (points) and transition rate magnitude β ranging from 0.1 to 5.1. The steady-state limit with irreversible forwards rates is shown in the black dashed line.

represents a bound to the noise performance of a real system (Fig. 3.2C, black dashed line). With this more realistic scenario of backwards transitions, the overall noise is higher. In addition, this means that the presence of backwards transitions rates can effectively introduce degeneracy. For example, a system with three states and backwards transitions may look practically indistinguishable from one with two states but irreversible forwards transitions. This complexity may make future experimental investigations into mapping these model states onto biochemical pathways quite challenging.

3.2.3 Transient transcription factor input dynamics can decrease noise in transcription onset time distributions

In the previous sections, we only considered models where the forward transition rates (the ones coupled to transcription factor concentrations) were constant in time. However, in systems like the developing fruit fly embryo, these concentrations are highly transient due to the formation and breakdown of the nuclear membrane between cell cycles. For a pioneering

factor such as Zelda, this results in rapid expulsion from the nucleus upon the onset of mitosis, and a fast re-introduction to the nucleus as the nucleus enters interphase during each cell cycle (Fig. 2.3B). Thus, after each mitotic division, there is a transient period during which the concentration of pioneering factors at a given gene locus is out of steady state. A more realistic model should account for the effect (if any) of these non-steady input dynamics on the resulting transcription onset time distributions.

Here, we investigate the effect of non-steady state transcription factor dynamics by modifying the forward transition rates β in our model to allow for transient rates $\beta(t)$ (Fig. 3.3A). For now, we will assume a reasonable form for the transition rate. Considering β to be a linearly correlated with an input transcription factor concentration, for example, we assume this transient $\beta(t)$ behaves like the concentration of a simple diffusive process with form

$$\beta(t) = \beta_0(1 - e^{-t/\tau}). \quad (3.8)$$

Here, β_0 is the asymptotic, saturating value of $\beta(t)$, and τ is the time constant governing the time-varying nature of the transition rate. The exponential term in the parantheses is motivated by the fact that any solution to the diffusion equation in free space will feature an $e^{-t/\tau}$ term. In this case, we expect τ to be highly dependent on the diffusion constant of the transcription factor in consideration. Furthermore, since we are considering the scenario in which the nuclear concentration of transcription factor immediately following anaphase begins at zero and increases to a steady-state value after a long time, the $1 - e^{-t/\tau}$ form captures this time-dependent process in a manner consistent with the diffusion equation. For comparison, the time plots of the steady state and transient input are shown in Fig. 3.3B, for $\tau = 3$ and $\beta_0 = 1$.

First, we examine the case like the first model presented in Fig. 3.1A, consisting of equal and irreversible forward transitions ($f = 0$), except now the rates $\beta(t)$ are transient. Because of the time-varying nature of $\beta(t)$, the resulting distribution $P_k(t)$ for the case of equal, irreversible forward transition rates no longer obeys a simple Gamma distribution, and an analytical solution is difficult (or even impossible) to obtain. Nevertheless, we can easily simulate the distributions numerically, which is shown in Figure 3.3C for $k = 2$ steps. We see that increasing the time constant τ results in a rightward shift of the onset time distribution, as expected since the time-varying transition rate profile will result in slower initial transition rates.

Figure 3.3D shows this irreversible transient model in the feature space holding $k = 2$ fixed while varying τ (colored points). The analytical constraint imposed in the steady-state limit is shown with the black dashed line. We see that as τ increases, the squared CV decreases. By slowing down the pioneering factor concentration dynamics, the system's overall noise performance improves.

Figure 3.3E shows the results for holding $\tau = 3$ fixed and varying the number of steps k in this irreversible transient model (colored points). Again, the constraints in the steady-state limit are shown in dashed lines. For a given number of steps k , the transient model always

exhibits a smaller squared CV than the steady-state regime. Thus, transient dynamics improve noise performance of the system.

Intuitively, having a time-dependent input profile will make earlier transitions “weaker” so that transitions that happen before the expected time are less likely, tightening the overall distribution of onset times. The relevant timescale is the dimensionless ratio $\frac{\beta_0}{\tau}$ —the faster the intrinsic transition rate β_0 is compared to the transient input timescale τ , the larger the effects of the transient input. This manifests more strongly in the feature space for low values of the mean onset time, where the discrepancy between steady-state and transient is more apparent.

So, accounting for transient pioneering factor dynamics can increase synchrony of transcription onset times. However, this analysis only considered the simple case with equal, irreversible forward transitions. Does this result generalize?

Earlier, we saw that, in the steady-state case, the presence of finite backwards transition rates decreased the overall noise performance of the model (Fig. 3.2C). The greater the backwards transition rates, the higher the noise. Now, we extend our transient model to account for these backwards rates and show how the transient dynamics can counteract this performance loss.

Similar to Figure 3.2A, we will assume a model with equal forward transition rates $\beta(t)$ and equal backward transition rates $\beta_0 f$. We will compare the steady state case with the transient input case, parameterized with timescale τ . Note that we will only consider the forward transition rates to be transient, and assume that the backward transition rates are still time-independent.

In Figure 3.3F, we explore the feature space in the steady-state vs. transient cases, with the steady-state ideal case of equal, irreversible transitions as a reference (black dashed line). We consider the case with fixed $k = 2$ and $f = 0.2$, and varying τ . As shown earlier, the steady-state case with a backwards transition rate (black points) possesses higher noise than the steady-state irreversible limit. However, using a transient rate can reduce the noise (colored points)—as τ increases, the noise decreases. Thus, by slowing down the system (increasing τ), the model counters the increase in noise caused by the presence of backwards transitions. This suggests that the diffusive timescale τ and the backwards rate fraction f have opposing effects on noise, and that pioneering factor concentration dynamics are a tunable parameter for developmental timing.

3.3 Discussion

With the growing evidence for non-equilibrium processes in transcriptional regulation (Estrada et al., 2016; Li et al., 2018; Dufourt et al., 2018; Eck et al., 2020) comes the need for precise theoretical investigations into the different ways these non-equilibrium aspects ultimately influence biological systems. Here, we studied the impact of transient pioneering factor concentration dynamics on distributions of transcription onset times in a simple model of chromatin accessibility. After developing a toy Markov chain model of transcription onset,

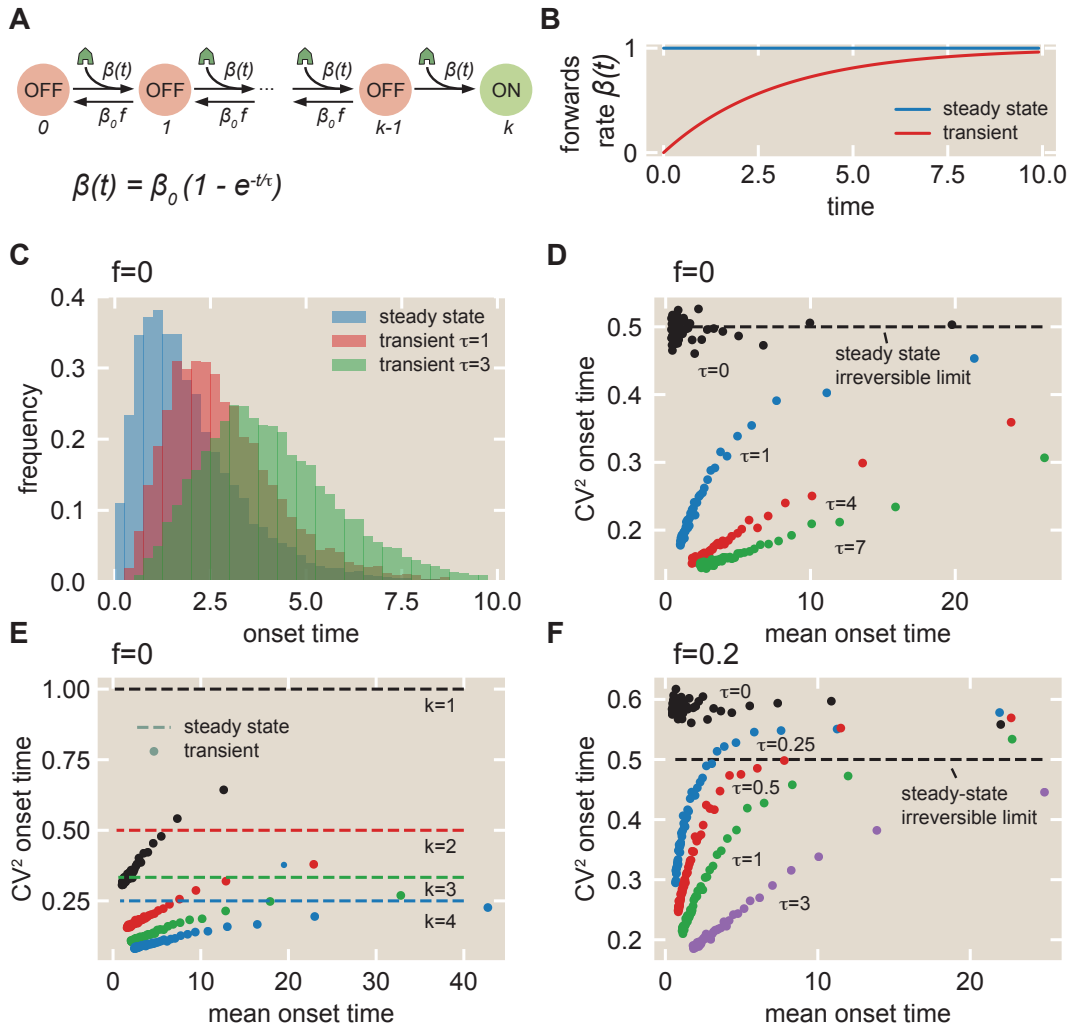


Figure 3.3: Results of transient input model. (A) Cartoon illustration of model with time-dependent forward transition rates $\beta(t)$ and time-independent backwards transition rates $\beta_0 f$. (B) Example time profiles of steady-state and transient transition rates $\beta(t)$, for $\tau = 3$ and $\beta_0 = 1$. (C) Simulated distributions of irreversible transient model ($f = 0$) in steady state and various values of τ , for $k = 2$ steps. (D) Simulated results of irreversible transient model in feature space for varying τ and $k = 2$ steps (points), along with steady-state limit (black dashed line). (E) Simulated results of irreversible transient model in feature space for varying k and $\tau = 3$ (points), along with steady-state limits (dashed lines). (F) Simulated results of reversible transient model ($f = 0.2$) in feature space for varying τ and $k = 2$ steps (points), along with steady-state limit with irreversible forward transitions (black dashed line). Values of β_0 used in (D-F) ranged from 0.1 to 5.1.

we showed that, in steady state, the spread of the model’s distribution of onset times was constrained by the number of states in the model along with the magnitudes of the transition rates between states.

In contrast, extending the model to allow for transient forward transition rates—motivated by physiologically relevant concentration dynamics of pioneering factors such as Zelda or Bicoid—allowed the resulting onset time distributions to exceed these constraints and decrease in squared CV. Thus, naturally occurring transients, due to biological processes such as nuclear import and export of transcription factors, may actually help boost precision of developmental timing in systems such as the fruit fly embryo.

Importantly, this result ties in a physical process—the rich dynamics of transient phenomena—with a biologically relevant quantity—precision of transcription onset in development. This carries two important consequences. First, our results suggest that transcription factor dynamics can serve as important parameters for downstream processes like transcription onset. As a result, the underlying determinants of these dynamics, such as diffusion coefficients, may be tunable biophysical quantities that can be evolutionarily optimized for precise development.

Second, the presence of these tunable parameters motivates future perturbation experiments to more deeply test these theoretical models. Using tools such as optogenetics (Johnson and Toettcher, 2018) to modulate transcription factor concentration dynamics, we can probe the predictions generated by different types of models to distinguish between, and ultimately validate, models of chromatin accessibility.

3.4 Acknowledgements

We thank Nick Lammers and Brandon Schlomann for fruitful discussions on the formulation and investigation of the models explored in this work.

3.5 Methods and Materials

Numerical simulations were implemented in Python with the *numpy* package. To simulate the distributions of transcription onset times, we used a Gillespie algorithm (Gillespie, 1976). In the case of the transient Markov chain model of chromatin accessibility, a modified Gillespie algorithm that accounted for time-dependent rates was used (Purtan and Udrea, 2013). For each simulation, $N = 10000$ cells were simulated.

Further details can be found in the Jupyter notebooks on the Github repository (<https://github.com/GarciaLab/OnsetTimeTransientInputs>).

Chapter 4

Dynamic single-cell characterization of the eukaryotic transcription cycle

Foreword

If there's one takeaway from studying biology, it's that life is complicated. For most of my PhD, I studied a pared-down system in an effort to reduce the scope of analysis, focusing on regulation of transcriptional initiation for a single gene in the developing fruit fly embryo, *hunchback*. But I always wondered if we were missing out on exciting parts of the bigger picture.

For example, the MS2 live imaging technology that formed the wheelhouse of our lab's experimental expertise had traditionally been used to generate signals that serve as a readout for transcriptional initiation. In principle, though, an MS2 signal reports on the number of actively transcribing RNA polymerase molecules on a gene as a function of time. That signal contains a wealth of information, yet traditional analysis had tended to use it as a proxy for promoter activity.

This next project was borne out of that motivation to make more effective use of our nascent RNA-labeling technologies. Originally, we were interested in using a dual-color reporter consisting of MS2 and PP7 stem loops on either end of a reporter gene to label nascent RNA transcripts with two separate fluorophores. This signal would let us examine the time delay between polymerase molecules reaching the 5' and 3' end of a gene, in order to estimate an average elongation velocity that could be used in our models of transcription.

This side project blossomed, though, when I realized that this setup wouldn't just let us estimate elongation speeds; we could study the whole transcription cycle—initiation, elongation, and cleavage of mRNA—with temporal resolution at the single-cell level. This was particularly exciting because, while previous works had used similar setups to examine initiation or elongation alone, no one had attempted to simultaneously measure and describe the whole transcription cycle with live imaging technologies. In particular, cleavage is a relatively understudied process compared to initiation and elongation, and I had always

wondered if there was anything interesting to be studied in that realm.

The main challenge here was to come up with a rigorous computational framework to quantify aspects of the transcription cycle from noisy, single-cell live imaging data. I ended up settling on a Bayesian framework using Markov Chain Monte Carlo, and was pleased to find that it performed well in inferring parameter values of a simple model of the transcription cycle from our dual-color reporter data. We eventually published our results in *PLoS Computational Biology*.

On a broader level, this project made me realize the importance of data processing and interpretation. With biological data quickly exploding in size and complexity, oftentimes the crucial part of a project lies in cleverly figuring out which aspects of the data are most informative, and then constructing useful statistical or mathematical metrics to concisely describe the data. Here, that process consisted of developing a simple statistical framework to convert microscopy data into simple model parameters that could then be investigated with theory. Thematically, this approach resonated strongly with me and reinforced my desire to continue working at the messy interface between experiment and theory.

4.1 Introduction

The eukaryotic transcription cycle consists of three main steps: initiation, elongation, and cleavage of the nascent RNA transcript (Fig. 4.1A; Alberts (2015)). Crucially, each of these three steps can be controlled to regulate transcriptional activity. For example, binding of transcription factors to enhancers dictates initiation rates (Spitz and Furlong, 2012), modulation of elongation rates helps determine splicing efficiency (De La Mata et al., 2003), and regulation of cleavage controls aspects of 3' processing such as alternative polyadenylation (Tian and Manley, 2016).

The steps of the transcription cycle can be coupled with each other. For example, elongation rates contribute to determining mRNA cleavage and RNA polymerase (RNAP) termination efficiency (Pinto et al., 2011; Hazelbaker et al., 2013; Fong et al., 2015; Liu et al., 2017), and functional linkages have been demonstrated between transcription initiation and termination (Moore and Proudfoot, 2009; Mapendano et al., 2010). Nonetheless, initiation, elongation, and transcript cleavage have largely been studied in isolation. In order to dissect the entire transcription cycle, it is necessary to develop a holistic approach that makes it possible to understand how the regulation of each step dictates mRNA production and to unearth potential couplings among these steps.

To date, the processes of the transcription cycle have mostly been studied in detail using *in vitro* approaches (Bai et al., 2006; Herbert et al., 2008) or genome-wide measurements that require the fixation of cellular material and lack the spatiotemporal resolution to uncover how the regulation of the transcription cycle unfolds in real time (Roeder, 1991; Saunders et al., 2006; Muse et al., 2007; Core et al., 2008; Fuda et al., 2009; Churchman and Weissman, 2011). Only recently has it become possible to dissect these processes in living cells and in their full dynamical complexity using tools such as MS2 or PP7 to fluorescently label

nascent transcripts at single-cell resolution (Bertrand et al., 1998; Golding et al., 2005; Chao et al., 2008; Larson et al., 2011a). These technological advances have yielded insights into, for example, intrinsic transcriptional noise in yeast (Hocine et al., 2013), kinetic splicing effects in human cells (Coulon et al., 2014), elongation rates in *Drosophila melanogaster* (Garcia et al., 2013; Fukaya et al., 2017), and transcriptional bursting in mammalian cells (Tantale et al., 2016), *Dictyostelium* (Chubb et al., 2006; Muramoto et al., 2012; Corrigan and Chubb, 2014), fruit flies (Garcia et al., 2013; Lucas et al., 2013; Bothma et al., 2014; Fukaya et al., 2016; Faló-Sanjuan et al., 2019; Lammers et al., 2020) and *Caenorhabditis elegans* (Lee et al., 2019).

Despite the great promise of MS2 and PP7, using these techniques to comprehensively analyze the transcription cycle is hindered by the fact that the signal from these *in vivo* RNA-labeling technologies convolves contributions from all aspects of the cycle. Specifically, the fluorescence signal from nascent RNA transcripts persists throughout the entire cycle of transcript initiation, elongation, and cleavage; further, a single gene can carry many tens of transcripts. Thus, at any given point, an MS2 or PP7 signal reports on the contributions of transcripts in various stages of the transcription cycle (Ferraro et al., 2016). Precisely interpreting an MS2 or PP7 signal therefore demands an integrated approach that accounts for this complexity.

Here, we present a method for analyzing live-imaging data from the MS2 and PP7 techniques in order to dynamically characterize the steps—initiation, elongation, and cleavage—of the full transcription cycle at single-cell resolution. While the transcription cycle is certainly more nuanced and can include additional effects such as sequence-dependent pausing (Gaertner and Zeitlinger, 2014), we view the quantification of these effective parameters as a key initial step for testing theoretical models. This method combines a dual-color MS2/PP7 fluorescent reporter (Hocine et al., 2013; Coulon et al., 2014; Fukaya et al., 2017) with Bayesian statistical inference techniques and quantitative modeling. As a proof of principle, we applied this analysis to the transcription cycle of a *hunchback* reporter gene in the developing embryo of the fruit fly *Drosophila melanogaster*. We validate our approach by comparing our inferred average initiation and elongation rates with previously reported results.

Crucially, our analysis also delivered novel single-cell statistics of the whole transcription cycle that were previously unmeasurable using genome-wide approaches, making it possible to generate distributions of parameter values necessary for investigations that go beyond simple population-averaged analyses (Raj et al., 2006; Zenklusen et al., 2008; Wyart et al., 2010; Sanchez et al., 2011; So et al., 2011; Coulon et al., 2013; Little et al., 2013; Sanchez et al., 2013; Sanchez and Golding, 2013; Jones et al., 2014; Senecal et al., 2014; Xu et al., 2015; Albayrak et al., 2016; Gomez-Schiavon et al., 2017; Shaffer et al., 2017; Serov et al., 2017; Lucas et al., 2018; Munsky et al., 2018; Zoller et al., 2018; Miura et al., 2019; Ali et al., 2020; Filatova et al., 2020). We show that, by taking advantage of time-resolved data, our inference is able to filter out uncorrelated noise, such as that originating from random measurement error, in these distributions and retain sources of correlated variability (such as biological and systematic noise). By combining these statistics with theoretical models, we revealed substantial variability in RNAP stepping rates between individual molecules, demonstrating

the utility of our approach for testing hypotheses of the molecular mechanisms underlying the transcription cycle and its regulation.

This unified analysis enabled us to investigate couplings between the various transcription cycle parameters at the single-cell level, whereby we discovered a surprising correlation of cleavage rates with nascent transcript densities. These discoveries illustrate the potential of our method to sharpen hypotheses of the molecular processes underlying the regulation of the transcription cycle and to provide a framework for testing those hypotheses.

4.2 Results

To quantitatively dissect the transcription cycle in its entirety from live imaging data, we developed a simple model (Fig. 4.1A) in which RNAP molecules are loaded at the promoter of a gene of total length L with a time-dependent loading rate $R(t)$. For simplicity, we assume that each individual RNAP molecule behaves identically and independently: there are no interactions between molecules. While this assumption is a crude simplification, it nevertheless allows us to infer effective average transcription cycle parameters.

We parameterize this $R(t)$ as the sum of a constant term $\langle R \rangle$ that represents the mean, or time-averaged, rate of initiation, and a small temporal fluctuation term given by $\delta R(t)$ such that $R(t) = \langle R \rangle + \delta R(t)$. This mean-field parameterization is motivated by the fact that many genes are well approximated by constant rates of initiation (Garcia et al., 2013; Lucas et al., 2013; Eck et al., 2020; Lammers et al., 2020). The fluctuation term $\delta R(t)$ allows for slight time-dependent deviations from the mean initiation rate. As a result, this term makes it possible to account for time-dependent behavior that can occur over the course of a cell cycle once the promoter has turned on. After initiation, each RNAP molecule traverses the gene at a constant, uniform elongation rate v_{elon} . Upon reaching the end of the gene, there follows a deterministic cleavage time, τ_{cleave} , after which the nascent transcript is cleaved.

We do not consider RNAP molecules that do not productively initiate transcription (Darzacq et al., 2007) or that are paused at the promoter (Core et al., 2008), as they will provide no experimental readout. Based on experimental evidence (Garcia et al., 2013), we assume that these RNAP molecules are processive, such that each molecule successfully completes transcription, with no loss of RNAP molecules before the end of the gene (see Section B.5 for a validation of this hypothesis).

4.2.1 Dual-color reporter for dissecting the transcription cycle

As a case study, we investigated the transcription cycle of early embryos of the fruit fly *D. melanogaster*. Specifically, we focused on the P2 minimal enhancer and promoter of the *hunchback* gene during the 14th nuclear cycle of development; the gene is transcribed in a step-like pattern along the anterior-posterior axis of the embryo with a 26-fold modulation in overall mRNA count between the anterior and posterior end (Fig. 4.1B; Driever and Nusslein-Volhard (1989); Margolis et al. (1995); Perry et al. (2012); Garcia et al. (2013)). As

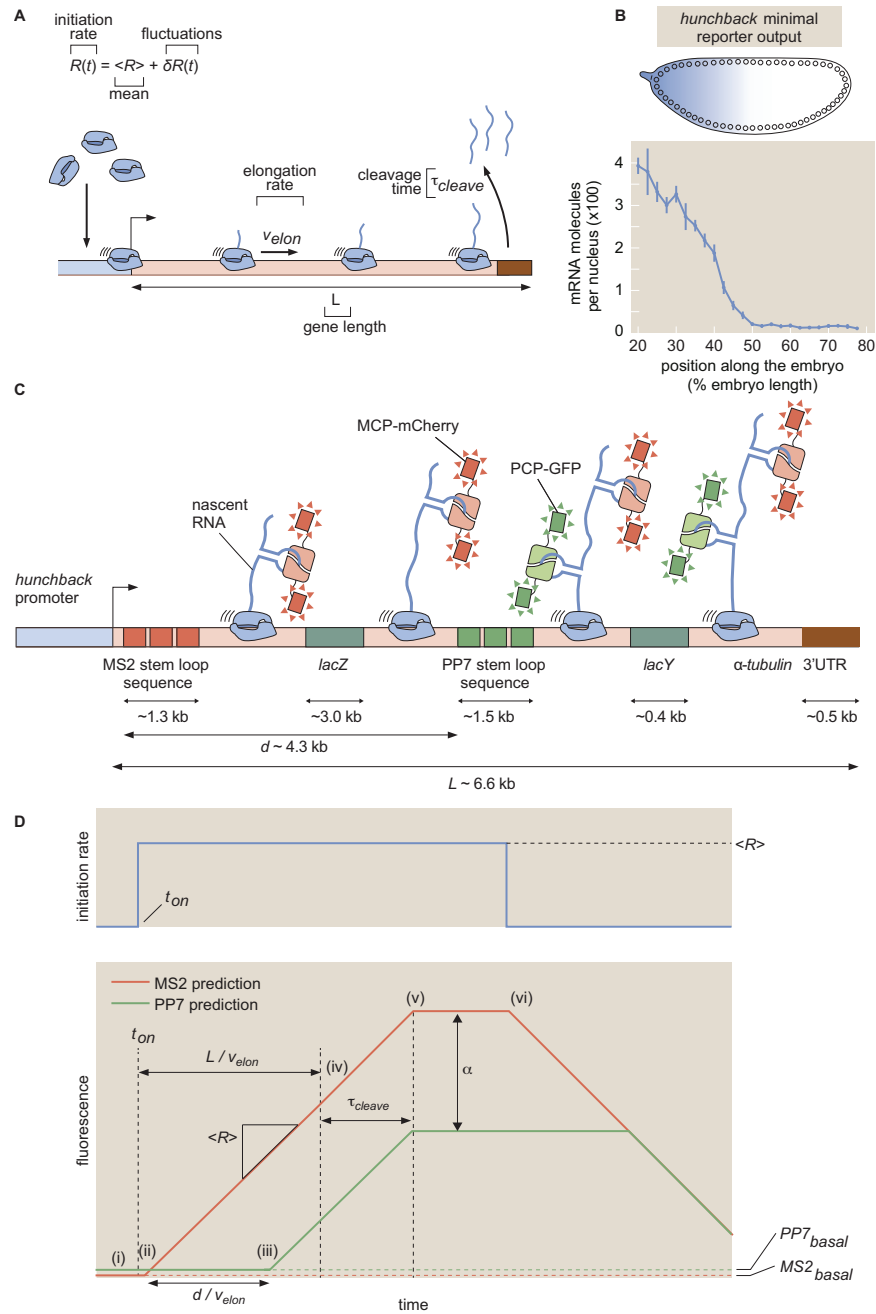


Figure 4.1: Theoretical model of the transcription cycle and experimental setup. See caption on next page.

Figure 4.1: Theoretical model of the transcription cycle and experimental setup. (A) Simple model of the transcription cycle, incorporating nascent RNA initiation, elongation, and cleavage. (B) The reporter construct, which is driven by the *hunchback* P2 minimal enhancer and promoter, is expressed in a step-like fashion along the anterior-posterior axis of the fruit fly embryo. (C) Transcription of the stem loops results in fluorescent puncta with the 5' mCherry signal appearing before the signal from 3' GFP. Only one stem loop per fluorophore is shown for clarity, but the actual construct contains 24 repeats of each stem loop. (D, top) Relationship between fluorescence trace profiles and model parameters for an initiation rate consisting of a pulse of constant magnitude $\langle R \rangle$. (D, bottom, i) At first, the zero initiation rate results in no fluorescence other than the basal levels $MS2_{basal}$ and $PP7_{basal}$ (red and green dashed lines). (ii) When initiation commences at time t_{on} , RNAP molecules load onto the promoter and elongation of nascent transcripts occurs, resulting in a constant increase in the MS2 signal (red curve). (iii) After time $\frac{d}{v_{elon}}$, the first RNAP molecules reach the PP7 stem loops and the PP7 signal also increases at a constant rate. (iv) After time $\frac{L}{v_{elon}}$, the first RNAP molecules reach the end of the gene, and (v) after the cleavage time τ_{cleave} , these first nascent transcripts are cleaved. The subsequent loss of fluorescence is balanced by the addition of new nascent transcripts, resulting in a plateauing of the signal. (vi) Once the initiation rate shuts off, no new RNAP molecules are added and both fluorescence signals will start to decrease due to cleavage of the nascent transcripts still on the gene. Because elongation continues after initiation has ceased, the 5' MS2 signal begins decreasing before the 3' PP7 signal. The MS2 and PP7 fluorescent signals are rescaled to be in the same arbitrary units with the calibration factor α . (Data in (B) adapted from Garcia et al. (2013) with the line representing the mean and error bars representing the standard error across 24 embryos.)

a result, the fly embryo provides a natural modulation in mRNA production rates, with the position along the anterior-posterior axis serving as a proxy for mRNA output.

To visualize the transcription cycle, we utilized the MS2 and PP7 systems for live imaging of nascent RNA production (Garcia et al., 2013; Lucas et al., 2013; Fukaya et al., 2016). Using a two-color reporter construct similar to that reported in Hocine et al. (2013), Coulon et al. (2014), and Fukaya et al. (2017), we placed the MS2 and PP7 stem loop sequences in the 5' and 3' ends, respectively, of a transgenic *hunchback* reporter gene (Fig. 4.1C; see Fig. B.1 for more construct details). The *lacZ* sequence and a portion of the *lacY* sequence from *Escherichia coli* were placed as a neutral spacer (Chen et al., 2012) between the MS2 and PP7 stem loops.

As an individual RNAP molecule transcribes through a set of MS2/PP7 stem loops,

constitutively expressed MCP-mCherry and PCP-GFP fusion proteins bind their respective stem loops, resulting in sites of nascent transcript formation that appear as fluorescent puncta under a laser-scanning confocal microscope (Fig. 4.2A and Video B.12.1). The fluorescent signals did not exhibit noticeable photobleaching (Section B.2 and Fig. B.2). Since *hunchback* becomes transcriptionally active at the start of the nuclear cycle before slowly decaying into a transcriptionally silent state (Garcia et al., 2013; Liu et al., 2013; Liu and Ma, 2015), we restrict our analysis to the initial 18 minute window after mitosis where the promoter remains active.

The intensity of the puncta in each color channel is linearly related to the number of actively transcribing RNAP molecules that have elongated past the location of the associated stem loop sequence (Garcia et al., 2013), albeit with different arbitrary fluorescence units. After reaching the end of the gene, which contains the 3'UTR of the α -*tubulin* gene (Chen et al., 2012), the nascent RNA transcript undergoes cleavage. Because the characteristic timescale of mRNA diffusion is about two order of magnitudes faster than the time resolution of our experiment, we approximate the cleavage of a single transcript as resulting in the instantaneous loss of its associated fluorescent signal in both channels (Section B.3). We included a few additional parameters in our model to make it compatible with this experimental data: a calibration factor α between mCherry and eGFP intensities, a time of transcription onset t_{on} after mitosis at which the promoter switches on, and basal levels of fluorescence in each channel $MS2_{basal}$ and $PP7_{basal}$ (see Section B.1 for more details). The qualitative relationship between the model parameters and the fluorescence data is described in Figure 4.1D, which considers the case of a pulse of constant initiation rate.

4.2.2 Transcription cycle parameter inference using Markov Chain Monte Carlo

We developed a statistical framework to estimate transcription-cycle parameters (Fig. 4.1A) from fluorescence signals. Time traces of mCherry and eGFP fluorescence intensity are extracted from microscopy data such as shown in Figure 4.2A and Video B.12.1 to produce a dual-signal readout of nascent RNA transcription at single-cell resolution (Fig. 4.2B, data points; see Methods and Materials for details). To extract quantitative insights from the observed fluorescence data, we used the established Bayesian inference technique of Markov Chain Monte Carlo (MCMC) (Geyer, 1992) to infer the effective parameter values in our simple model of transcription: the calibration factor between mCherry and eGFP intensities α , the time-dependent transcription initiation rate, separated into the constant term $\langle R \rangle$ and fluctuations $\delta R(t)$, the elongation rate v_{elon} , the cleavage time τ_{cleave} , the time of transcription onset t_{on} , and the basal levels of fluorescence in each channel $MS2_{basal}$ and $PP7_{basal}$.

The details of the inference procedure are described in Section B.4.1. Briefly, the inference was run separately for each single cell, yielding chains of sampled parameter values (Fig. 4.2C). These resulting chains exhibited rapid mixing and rapidly decaying auto-correlation functions (Fig. 4.2D), indicative of reliable fits. Corner plots of the fits indicated reasonable posterior

distributions (Fig. 4.2E).

From these single-cell fits, the mean value of each parameter’s chain was retained for further analysis. The final dataset was produced by filtering with an automated procedure that relied on overall fit quality (Section B.4.3 and Fig. B.4). This curation procedure did not introduce noticeable bias in the results (Fig. B.4G-I). A small minority of the rejected cells (Fig. B.4E) exhibited highly time-dependent behavior reminiscent of transcriptional bursting (Rodriguez and Larson, 2020), which lies outside the scope of our model and is explored more in the Discussion. A sample fit is shown in Figure 4.2B. To aggregate the results, we constructed a distribution from the inferred parameter from each single-cell. Intra-embryo variability between single cells was greater than inter-embryo variability (Section B.6 and Fig. B.6). As a result, unless stated otherwise, all statistics reported here were aggregated across 355 single cells combined between 7 embryos, and all shaded errors reflect the standard error of the mean.

4.2.3 MCMC successfully infers calibration between eGFP and mCherry intensities

Due to the fact that the MS2 and PP7 stem loop sequences were associated with mCherry and eGFP fluorescent proteins, respectively, the two experimental fluorescent signals possessed different arbitrary fluorescent units, related by the scaling factor α given by

$$\alpha = \frac{F_{MS2}}{F_{PP7}}, \quad (4.1)$$

where F_{MS2} and F_{PP7} are the fluorescence values generated by a fully transcribed set of MS2 and PP7 stem loops, respectively. Although α has units of AU_{MS2}/AU_{PP7} , we will express α without units in the interest of clarity of notation.

We inferred single-cell values of α using the inference methodology. As shown in the blue histogram in Figure 4.3A, our inferred values of α possessed a mean of 0.145 ± 0.004 (SEM) and a standard deviation of 0.068.

As an independent validation, we measured α by using another two-color reporter, consisting of 24 alternating, rather than sequential, MS2 and PP7 loops (Wu et al., 2014; Chen et al., 2018; Child et al., 2020) inserted at the 5’ end of our reporter construct (Fig. 4.3B). Thus, this reporter had a total of 48 stem loops, with 24 each of MS2 and PP7.

Figure 4.3C shows a representative trace of a single spot containing our calibration construct (see Video B.12.2 for full movie). For each time point, the mCherry fluorescence in all measured single-cell traces was plotted against the corresponding eGFP fluorescence (Fig. 4.3D, yellow points). The mean α was then calculated by fitting the resulting scatter plot to a line going through the origin (Fig. 4.3D, black line). The best-fit slope yielded the experimentally calculated value of $\alpha = 0.154 \pm 0.001$ (SEM). A distribution for α was also constructed by dividing the mCherry fluorescence by the corresponding eGFP fluorescence for each datapoint in Figure 4.3D, yielding the histogram in Figure 4.3A (yellow), which possessed

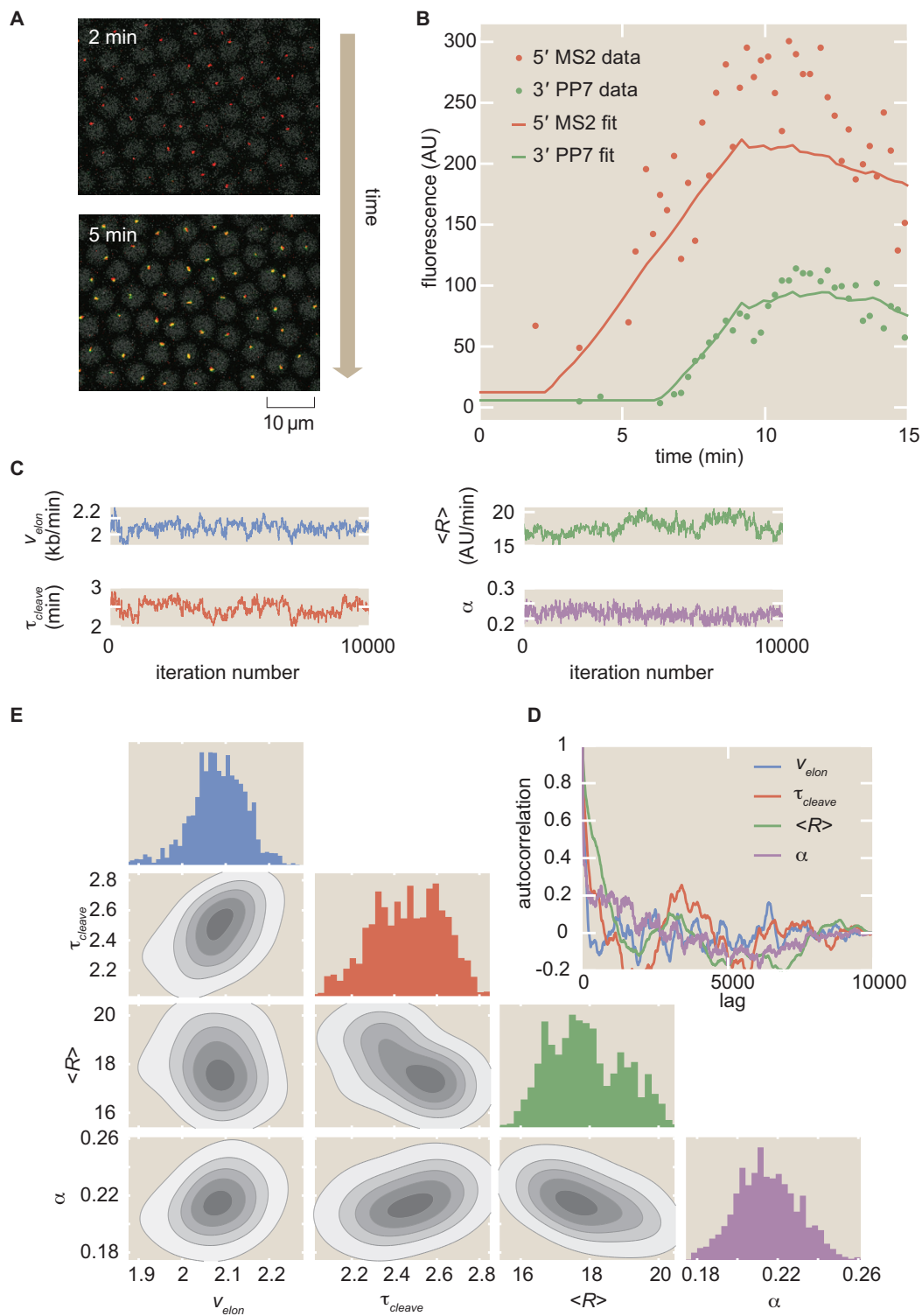


Figure 4.2: MCMC inference procedure. See caption on next page.

Figure 4.2: MCMC inference procedure. (A) Snapshots of confocal microscopy data over time, with MS2-mCherry (red) and PP7-eGFP (green) puncta reporting on transcription activity. Gray circles correspond to iRFP-labeled histones whose fluorescence is used as a fiduciary marker for cell nucleus segmentation (see Methods and Materials for details). (B) Sample single-cell MS2 and PP7 fluorescence (points) together with best-fits of the model using MCMC inference (curves). (C) Raw MCMC inference chains for the elongation rate v_{elon} , cleavage time τ_{cleave} , mean initiation rate $\langle R \rangle$, and calibration factor α for the inference results of a sample single cell. (D) Auto-correlation function for the raw chains in (A) as a function of lag (i.e. inference sample number). (E) Corner plot of the raw chains shown in (C).

a standard deviation of 0.073. Our independent calibration agreed with our inference, thus validating the inferred values of α .

Interestingly, binning the cells by position along the embryo revealed a slight position dependence in the scaling factor. As shown in Figure 4.3E, both the directly measured and inferred α displayed higher values in the anterior, about 0.15, and lower values in the posterior, about 0.1. The fact that this position dependence is observed in both in the calibration experiments and inference suggests that this spatial modulation in the value of α is not an artifact of the constructs or our analysis, but a real feature of the system. We speculate that this spatial dependence could stem from differential availability of MCP-mCherry and PCP-GFP along the embryo, leading to a modulation in the maximum occupancy of the MS2 stem loops versus the PP7 stem loops (Wu et al., 2012).

Regardless, our data demonstrate that the inferred and calibrated α can be used interchangeably, obviating the need for the control. Thus, the MS2 signals for each single cell could be rescaled to the same units as the PP7 signal (Fig. 4.3F) within a single experiment, greatly increasing the power of the inference methodology. All plots in the main text and supplementary information, unless otherwise stated, reflect these rescaled values using the overall mean value of $\alpha = 0.145$ obtained from the inference.

4.2.4 Inference of single-cell initiation rates recapitulates and improves on previous measurements

After validating the accuracy of our inference method in inferring transcription initiation, elongation, and cleavage dynamics using simulated data (Section B.4.4 and Fig. B.5), we inferred these transcriptional parameters for the *hunchback* reporter gene as a function of the position along the anterior-posterior axis of the embryo. The suite of quantitative measurements on the transcription cycle produced by the aggregated inference results is shown in Figures 4.4A, C, E, and F. Full distributions of these parameters can be found in

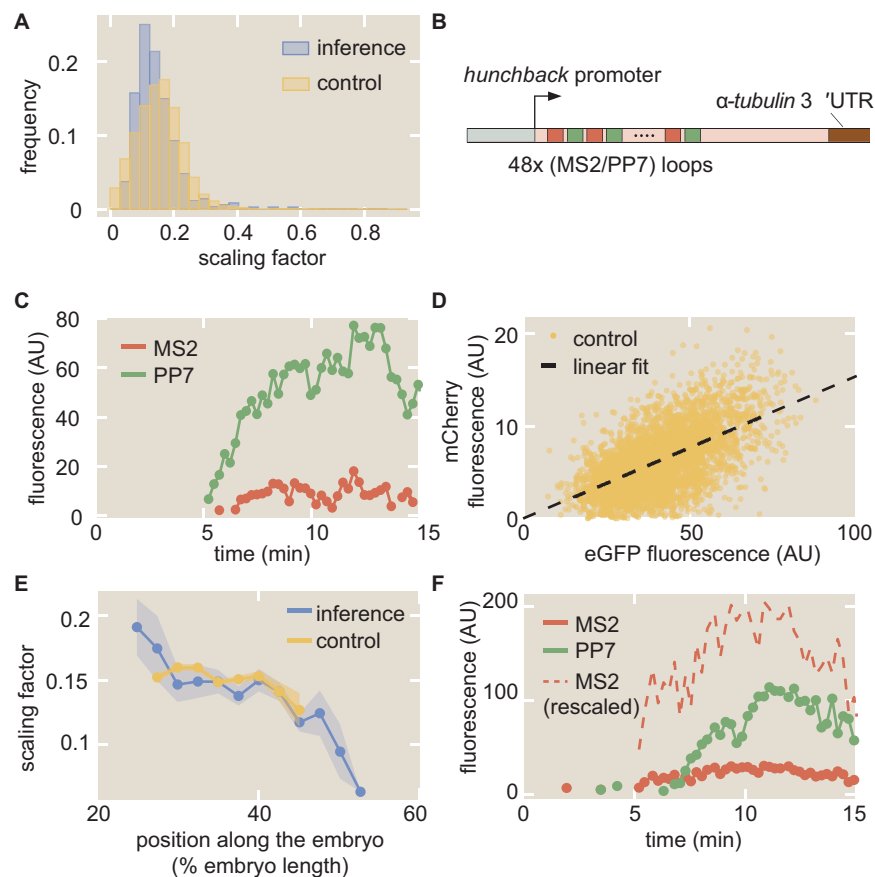


Figure 4.3: Calibration of MS2 and PP7 fluorescence signals. (A) Histogram of inferred values of α at the single-cell level from inference (blue), along with histogram of α values from the control experiment (yellow). (B) Schematic of construct used to measure the calibration factor α using 24 interlaced MS2/PP7 loops each (48 loops in total). (C) Sample single-cell MS2 (red) and PP7 (green) traces from this control experiment. (D) Scatter plot of MS2 and PP7 fluorescence values for each time point (yellow) along with linear best fit (black) resulting in $\alpha = 0.154 \pm 0.001$. (E) Position-dependent mean value of α in both the inference (blue) and the control experiment (yellow). (F) Representative raw and rescaled MS2 and PP7 traces for a sample single cell in the inference data set. (A,D,E, data were collected for 314 cells across 4 embryos for the interlaced reporter, and for 355 cells across 7 embryos for the reporter with MS2 on the 5' and PP7 on the 3' of the gene (Fig. 4.1C); shaded regions in (E) reflect standard error of the mean. Measurement conditions for both experiments are described in Methods and Materials.)

Fig. B.7.

Control of initiation rates is one of the predominant, and as a result most well-studied, strategies for gene regulation (Roeder, 1991; Spitz and Furlong, 2012; Lenstra et al., 2016). Thus, comparing our inferred initiation rates with previously established results comprised a crucial benchmark for our methodology. Our inferred values of the mean initiation rate $\langle R \rangle$ exhibited a step-like pattern along the anterior-posterior axis of the embryo, qualitatively reproducing the known *hunchback* expression profile (Fig. 4.4A, blue). As a point of comparison, we also examined the mean initiation rate measured by Garcia et al. (2013), which was obtained by manually fitting a trapezoid (Figure 4.1D) to the average MS2 signal (Fig. 4.4A, black). The quantitative agreement between these two dissimilar analysis methodologies demonstrates that our inference method can reliably extract the average rate of transcription initiation across cells.

Measurements of cell-to-cell variability in transcription initiation rate have uncovered, for example, the existence of transcriptional bursting and mechanisms underlying the establishment of precise developmental boundaries (Raj et al., 2006; Sanchez and Golding, 2013; Zenklusen et al., 2008; Little et al., 2013; Jones et al., 2014; Lucas et al., 2018; Zoller et al., 2018). Yet, to date, these studies have mostly employed techniques such as single-molecule FISH to count the number of nascent transcripts on a gene or the number of cytoplasmic mRNA molecules (Femino et al., 1998; Raj et al., 2006; Pare et al., 2009; Zenklusen et al., 2008; Wyart et al., 2010; So et al., 2011; Boettiger and Levine, 2013; Little et al., 2013; Jones et al., 2014; Senecal et al., 2014; Fei et al., 2015; Padovan-Merhar et al., 2015; Xu et al., 2015; Albayrak et al., 2016; Skinner et al., 2016; Bartman et al., 2016; Gomez-Schiavon et al., 2017; Hendy et al., 2017; Munsky et al., 2018; Zoller et al., 2018; Miura et al., 2019). In principle, these techniques do not report on the variability in transcription initiation alone; they convolve this measurement with variability in other steps of the transcription cycle (Padovan-Merhar et al., 2015; Lenstra et al., 2016).

Our inference approach isolates the transcription initiation rate from the remaining steps of the transcription cycle at the single-cell level, making it possible to calculate, for example, the coefficient of variation (CV; standard deviation divided by the mean) of the mean rate of initiation. Our results yielded values for the CV along the embryo that were fairly uniform, with a maximum value of around 40% (Fig. 4.4F, blue). This value is roughly comparable to that obtained for *hunchback* using single-molecule FISH (Little et al., 2013; Xu et al., 2015; Zoller et al., 2018).

One of the challenges in measuring CV values, however, is that informative biological variability is often convolved with undesired experimental noise, such as experimental measurement noise inherent to fluorescence microscopy. In general, this experimental noise can contain both random, uncorrelated components as well as systematic components, the latter of which combines with actual biological variability to form overall correlated noise. Although we currently cannot entirely separate biological variability from experimental noise with our data and inference method, a strategy for at least separating uncorrelated from correlated was recently implemented in the context of snapshot-based fluorescent data (Zoller et al., 2018). By utilizing a dual-color measurement of the same biological signal, one can separate

the total variability in a dataset into uncorrelated measurement noise and correlated noise, which includes components such as true biological variability and systematic measurement error.

Building on this strategy, we first took a single snapshot from our live-imaging data and calculated the total squared CV of the fluorescence of spots at a single time point (Fig. 4.4B, dark plus light purple). Compared to the squared CV from the inferred mean initiation rate (Fig. 4.4B, blue), the squared CV from the snapshot was larger by about 0.1, suggesting that the inference method reported on a somewhat lower level of overall variability.

To investigate this disparity in measured variability further, we then rewrote the squared CV from the snapshot approach as the sum of uncorrelated and correlated noise components

$$CV_{total}^2 = CV_{uncorrelated}^2 + CV_{correlated}^2. \quad (4.2)$$

The magnitudes of each noise component were estimated by using the data from the interlaced reporter introduced in Figure 4.3B. To do so, we utilized the fact that, in principle, the mCherry and GFP signals from this experiment reflected the same underlying biological process, and assumed that deviations between the two signals were a result of uncorrelated measurement noise. Thus, we could apply the two-color formalism introduced in Elowitz et al. (2002) to calculate the uncorrelated and correlated noise components from snapshots taken from the interlaced reporter construct (see Section B.8 and Figure B.8 for more details).

The bar graph shown in Figure 4.4B shows that, once the uncorrelated noise (light purple) is subtracted from the total noise of our snapshot-based measurement, the remaining correlated variability (dark purple), which includes the biological variability, is slightly lower than the variability of our inference results (blue). Thus, our inference mostly captures correlated variability and filters out the bulk of the uncorrelated noise, similarly to techniques such as single-molecule FISH (Zoller et al., 2018) but with the added advantage of also being able to resolve temporal information. Because such fixed tissue techniques ultimately provide static measurements that convolve signals from transcription initiation with those of elongation and cleavage, it is important to note that this is a qualitative comparison between the ability of fixed-tissue and live-imaging to separate correlated and uncorrelated variability. Thus, our results further validate our approach and demonstrate its capability to capture measures of cell-to-cell variability in the transcription cycle with high precision.

4.2.5 Elongation rate inference reveals single-molecule variability in RNAP stepping rates

Next, we investigated the ability of our inference approach to report on the elongation rate v_{elon} . Nascent RNA elongation plays a prominent role in gene regulation, for example, in dosage compensation in *Drosophila* embryos (Larschan et al., 2011), alternative splicing in human cells (De La Mata et al., 2003; Batsché et al., 2006), and gene expression in plants (Wu et al., 2016). Our method inferred an elongation rate v_{elon} that was relatively constant along the embryo (Fig. 4.4C), lending support to previous reports indicating a lack of regulatory

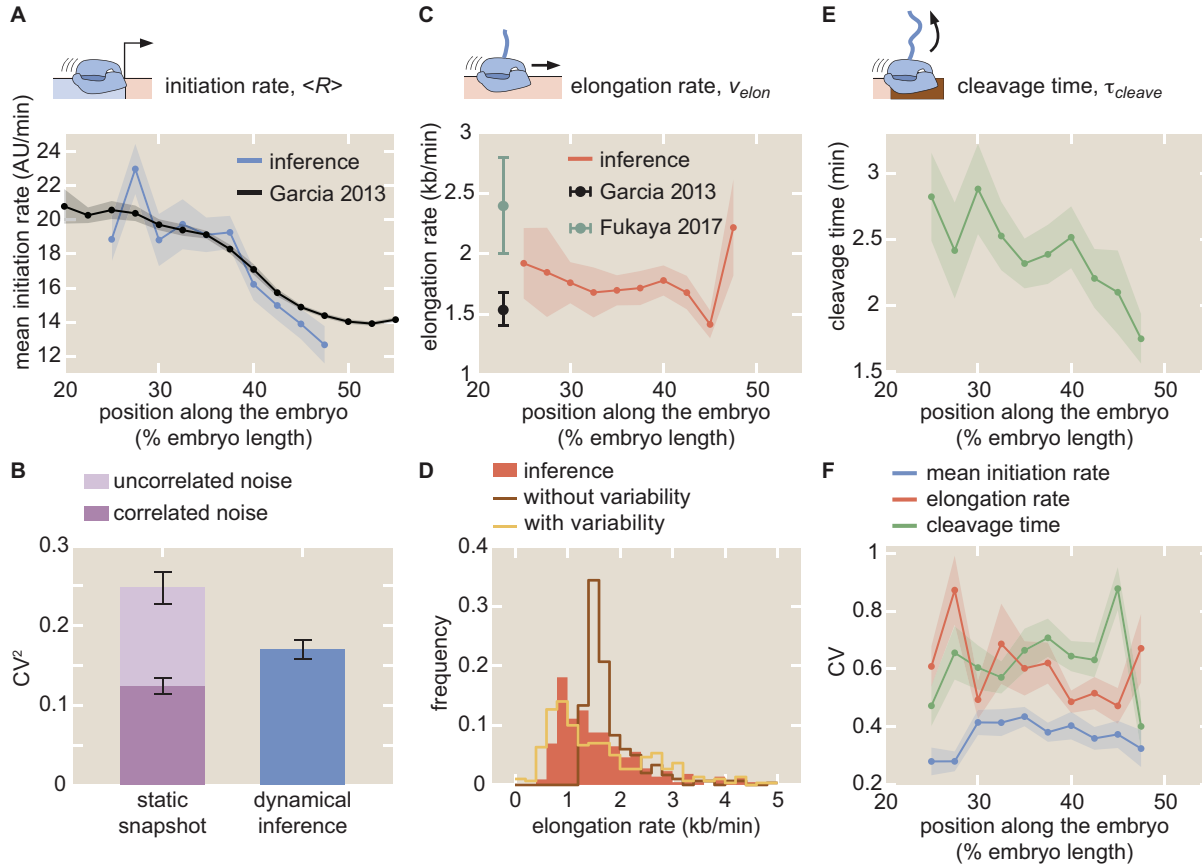


Figure 4.4: Inferred transcription-cycle parameters. See caption on next page.

control of the elongation rate in the early fly embryo (Fukaya et al., 2017). We measured a mean elongation rate of 1.72 ± 0.05 kb/min (SEM; $n = 355$), consistent with previous measurements of the fly embryo (Fig. 4.4C, black and teal; Garcia et al. (2013); Fukaya et al. (2017)), as well as with measurements from other techniques and model organisms, which range from about 1 kb/min to upwards of 4 kb/min (Femino et al., 1998; Golding et al., 2005; Darzacq et al., 2007; Boireau et al., 2007; Ardehali and Lis, 2009; Palangat and Larson, 2012; Hocine et al., 2013; Coulon et al., 2014; Fuchs et al., 2014; Tantale et al., 2016; Lenstra et al., 2016). In addition, the CV of the elongation rate was roughly uniform across embryo position (Fig. 4.4F, red).

Like cell-to-cell variability in transcription initiation, single-cell distributions of elongation rates can provide crucial insights into, for example, promoter-proximal pausing (Serov et al., 2017), traffic jams (Klumpp and Hwa, 2008; Klumpp, 2011), transcriptional bursting (Choubey et al., 2015, 2018), and noise propagation (Ali et al., 2020). While genome-wide approaches have had huge success in measuring mean properties of elongation (Core et al., 2008; Carrillo Oesterreich et al., 2010), they remain unable to resolve single-cell distributions

Figure 4.4: Inferred transcription-cycle parameters. (A) Mean inferred transcription initiation rate as a function of embryo position (blue), along with rescaled previously reported results (black, Garcia et al. (2013)). (B) Comparison of the squared CV of the mean initiation rate inferred using our approach (blue) or obtained from examining the fluorescence of transcription spots in a single snapshot (light plus dark purple). While snapshots captured a significant amount of uncorrelated noise (light purple), our inference accounts mostly for correlated noise (compare blue and dark purple). See Section B.8 and Fig. B.8 for details. (C) Inferred elongation rate as a function of embryo position (red), along with previously reported results (black, Garcia et al. (2013); teal, Fukaya et al. (2017)). (D) Distribution of inferred single-cell elongation rates in the anterior 40% of embryo (red), along with best fit to mean and standard deviation using single-molecule simulations with and without RNAP-to-RNAP variability (gold and brown, respectively, see Section B.10 for details). (E) Inferred cleavage time as a function of embryo position. (F) CV of the mean initiation rate (blue), elongation rate (red), and cleavage time (green) as a function of embryo position. (A, C, E, shaded error reflects standard error of the mean across 355 nuclei in 7 embryos, or of previously reported mean results; B, F, shaded error or black error bars represent bootstrapped standard errors of the CV or CV^2 for 100 bootstrap samples each; C, error bars reflect standard error of the mean for Garcia et al. (2013) and lower (25%) and upper (75%) quintiles of the full distribution from Fukaya et al. (2017).)

of elongation rates. We examined the statistics of single-cell elongation rates in the anterior 40% of the embryo, where the initiation rate was roughly constant, and inferred a broad distribution of elongation rates with a standard deviation of around 1 kb/min and a long tail extending to values upwards of 4 kb/min (Fig. 4.4D, red). This large spread was consistent with observations of large cell-to-cell variability in elongation rates (Palangat and Larson, 2012; Lenstra et al., 2016) using a wide range of techniques, as well as with measurements from similar two-color live imaging experiments (Hocine et al. (2013); Fukaya et al. (2017); Section B.9; Fig. B.9).

To illustrate the resolving power of examining elongation rate distributions, we performed theoretical investigations of cell-to-cell variability in this transcription cycle parameter. Following Klumpp and Hwa (2008), we considered a model where RNAP molecules stochastically step along a gene and cannot overlap or pass each other (Section B.10). The model simulated MS2 and PP7 fluorescences that were then run through the inference procedure, in order to account for the presence of inferential noise (Section B.4.4).

First, we considered a scenario where the stepping rate of each RNAP molecule is identical. In this case, the sole driver of cell-to-cell variability is the combination of stochastic stepping

behavior with traffic jamming due to steric hindrance of RNAP molecules. As shown in brown in Figure 4.4D, this model cannot account for the wide distribution of observed single-cell elongation rates.

In contrast, by allowing for substantial variability in the elongation rate of individual RNAP molecules, the model can reproduce the empirical distribution of single-cell elongation rates. As shown in gold in Figure 4.4D, the model can quantitatively approximate the inferred distribution within error (Fig. B.10D). This single-molecule variability is consistent with *in vitro* observations of substantial molecule-to-molecule variability in RNAP elongation rates (Tolić-Nørrelykke et al., 2004; Larson et al., 2011b), thus demonstrating the ability of our approach to engage in the *in vivo* dissection of the transcription cycle at the single-molecule level.

4.2.6 Inference reveals functional dependencies of cleavage times

Finally, we inferred values of the cleavage time τ_{cleave} . Through processes such as alternative polyadenylation (Tian and Manley, 2016; Jung et al., 2009) and promoter-terminator crosstalk (Moore and Proudfoot, 2009; Mapendano et al., 2010), events at the 3' end of a gene exert substantial influence over overall transcription levels (Bentley, 2014). Although many investigations of mRNA cleavage and RNAP termination have been carried out in fixed-tissue samples (Richard and Manley, 2009; Kuehner et al., 2011), live-imaging studies with single-cell resolution of this important process remain sparse; some successes have been achieved in yeast and in mammalian cells (Lenstra et al., 2016). We inferred a mean mRNA cleavage time in the range of 1.5-3 min (Fig. 4.4E), consistent with values obtained from live imaging in yeast (Larson et al., 2011a) and mammalian cells (Boireau et al., 2007; Darzacq et al., 2007; Coulon et al., 2014; Tantale et al., 2016). Interestingly, as shown in Figure 4.4E, the inferred mRNA cleavage time was dependent on anterior-posterior positioning along the embryo, with high values (~ 3 min) in the anterior end and lower values toward the posterior end (~ 1.5 min). While the reasons for this position dependence are unknown, such dependence could result from the presence of a spatial gradient of a molecular species that regulates cleavage. Importantly, such a modulation could not have been easily revealed using genome-wide approaches that, by necessity, average information across multiple cells.

The CV of the cleavage time slightly increased toward the posterior end of the embryo (Fig. 4.4F, green). Thus, although cleavage remains an understudied process compared to initiation and elongation, both theoretically and experimentally, these results provide the quantitative precision necessary to carry out such mechanistic analyses.

4.2.7 Uncovering single-cell mechanistic correlations between transcription cycle parameters

In addition to revealing trends in average quantities of the transcription cycle along the length of the embryo, the simultaneous nature of the inference afforded us the unprecedented ability to investigate single-cell correlations between transcription-cycle parameters. We used

the Spearman rank correlation coefficient (ρ) as a non-parametric measure of inter-parameter correlations. The mean initiation rate and the cleavage time exhibited a negative correlation ($\rho = -0.52$, $p\text{-val} \approx 0$; Fig. 4.5A). This negative correlation at the single-cell level should be contrasted with the positive relation between these magnitudes at the position-averaged level, where the mean initiation rate and cleavage time both increased in the anterior of the embryo (Fig. 4.4A and E). Thus, our analysis unearthed a quantitative relationship that was obscured by a naive investigation of spatially averaged quantities, an approach often used in fixed (Zoller et al., 2018) and live-imaging (Lammers et al., 2020) studies, as well as in genome-wide investigations (Combs and Eisen, 2017; Haines and Eisen, 2018). We also detected a small negative correlation ($\rho = -0.21$, $p\text{-val} = 5 \times 10^{-5}$) between elongation rates and mean initiation rates (Fig. 4.5B). Finally, we detected a small positive correlation ($\rho = 0.35$, $p\text{-val} = 2 \times 10^{-11}$) between cleavage times and elongation rates (Fig. 4.5C). These results are consistent with prior studies implicating elongation rates in 3' processes such as splicing and alternative polyadenylation: slower elongation rates increased cleavage efficiency (De La Mata et al., 2003; Pinto et al., 2011).

The observed negative correlation between cleavage time and mean initiation rate (Fig. 4.5A), in conjunction with the positive correlation between cleavage time and elongation rate (Fig. 4.5C), suggested a potential underlying biophysical control parameter: the mean nascent transcript density on the reporter gene body ρ given by

$$\rho = \frac{\langle R \rangle}{v_{elon}}. \quad (4.3)$$

Possessing units of (AU/kb), this mean transcript density estimates the average number of nascent RNA transcripts per kilobase of template DNA. Plotting the cleavage time as a function of the mean transcript density yielded a negative correlation ($\rho = -0.55$, $p\text{-val} \approx 0$) that was stronger than any of the other correlations between transcription-cycle parameters at the single-cell level (Fig. 4.5D). Mechanistically, the correlation between cleavage time and mean transcript density suggests that, on average, more closely packed nascent transcripts at the 3' end of a gene cleave faster.

Further investigations using simulations indicated that this relationship did not arise from spurious correlations in the inference procedure itself (Section B.4.4 and Fig. B.5E-H), but rather captured real correlations in the data. Furthermore, although the four inter-parameter correlations investigated here only used mean values obtained from the inference methodology, a Monte Carlo simulation involving the full Bayesian posterior distribution confirmed the significance of the results (Section B.11 and Fig. B.11).

Using an absolute calibration for a similar reporter gene (Garcia et al., 2013) led to a rough scaling of $1 \text{ AU} \approx 1 \text{ molecule}$ corresponding to a maximal RNAP density of about 20 RNAP molecules/kb in Figure 4.5D. With a DNA footprint of 40 bases per molecule (Selby et al., 1997), this calculation suggests that, in this regime, RNAP molecules are densely distributed, occupying about 80% of the reporter gene. We hypothesize that increased RNAP density could lead to increased pausing as a result of traffic jams (Klumpp and Hwa, 2008; Klumpp, 2011). Due to this pausing, transcripts would be more available for cleavage,

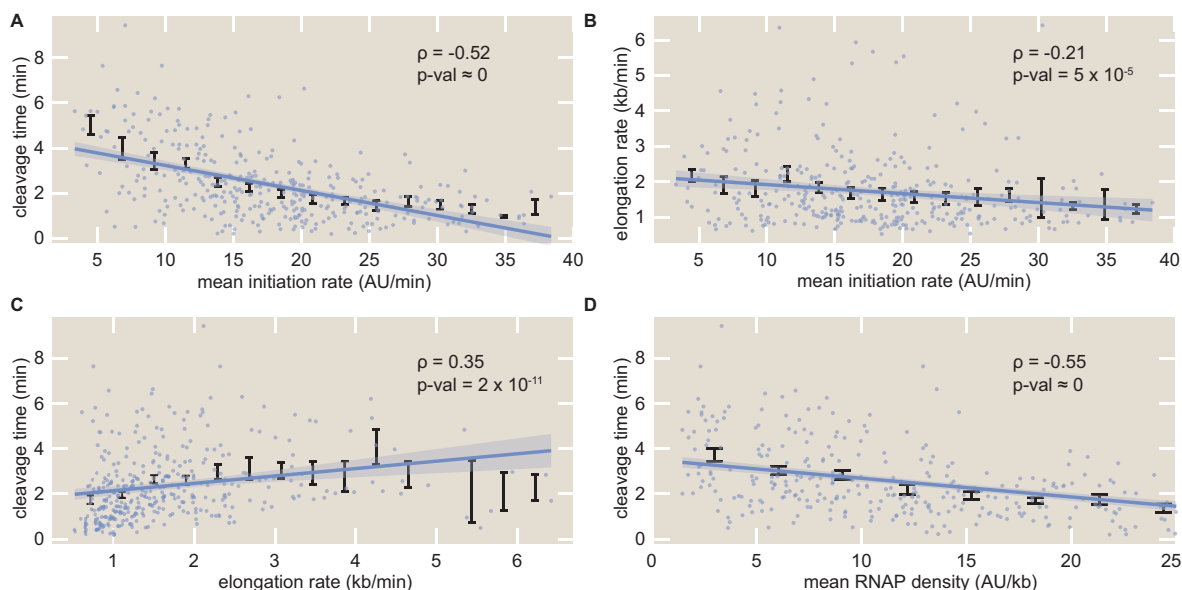


Figure 4.5: Single-cell correlations between transcription cycle parameters. Spearman rank correlation coefficients and associated p-values between (A) mean initiation rate and cleavage time, (B) mean initiation rate and elongation rate, (C) elongation rate and cleavage time, and (D) mean RNAP density and cleavage time. Blue points indicate single-cell values; black points and error bars indicate mean and SEM, respectively, binned across x-axis values. Lines and shaded regions indicate generalized linear model fit and 95% confidence interval, respectively, and are shown for ease of visualization (see Methods and Materials for details).

increasing overall cleavage efficiency. Regardless of the particular molecular mechanisms underlying our observations, we anticipate that this ability to resolve single-cell correlations between transcription parameters, combined with perturbative experiments, will provide ample future opportunities for studying the underlying biophysical mechanisms linking transcription processes.

4.3 Discussion

Over the last two decades, the genetically encoded MS2 (Bertrand et al., 1998) and PP7 (Chao et al., 2008) RNA labeling technologies have made it possible to measure nascent and cytoplasmic RNA dynamics *in vivo* in many contexts (Golding et al., 2005; Chubb et al., 2006; Darzacq et al., 2007; Larson et al., 2011a; Garcia et al., 2013; Lucas et al., 2013; Hocine et al., 2013; Coulon et al., 2014; Bothma et al., 2014; Lenstra et al., 2015, 2016; Fukaya et al., 2016; Tantale et al., 2016; Fukaya et al., 2017; Chen et al., 2018; Dufourt et al., 2018; Fritsch et al., 2018; Falo-Sanjuan et al., 2019; Li et al., 2019; Lee et al., 2019; Lammers et al.,

2020; Eck et al., 2020). However, such promising experimental techniques can only be as powerful as their underlying data-analysis infrastructure. For example, while initial studies using MS2 set the technological foundation for revealing transcriptional bursts in bacteria (Golding et al., 2005), single-celled eukaryotes (Chubb et al., 2006; Larson et al., 2009), and animals (Garcia et al., 2013; Lucas et al., 2013), only recently did analysis techniques become available to reliably obtain parameters such as transcriptional burst frequency, duration, and amplitude (Coulon et al., 2014; Desponds et al., 2016; Corrigan et al., 2016; Lammers et al., 2020; Bowles et al., 2020).

In this work, we established a novel method for inferring quantitative parameters of the entire transcription cycle—initiation, elongation and cleavage—from live imaging data of nascent RNA dynamics. Notably, this method offers high spatiotemporal resolution at the single-cell level, resolving aspects of transcriptional activity within the body of an organism and at sub-minute resolution. Furthermore, while our experimental setup utilized two fluorophores, we found that the calibration between their intensities could be inferred directly from the data (Fig. 4.3), rendering independent calibration and control experiments unnecessary.

After validating previously discovered spatial modulations in the mean initiation rate, we discovered an unreported modulation of the cleavage time with respect to embryo position that mirrored that of the mean initiation rate (Fig. 4.4E). Although such a relationship at first would suggest a positive correlation between initiation and cleavage, the presence of significant negative correlation at the single-cell level refutes this idea (Fig. 4.5A). Instead, we speculate that the spatial modulation of the cleavage time could instead underlie a coupling with a spatial gradient of some molecular factor that controls this transcription cycle parameter (El Kaderi et al., 2009), possibly due to effects such as gene looping (O’Sullivan et al., 2004; Tan-Wong et al., 2008).

These features are unattainable by widespread, but still powerful, genome-wide techniques that examine fixed samples, such as global run-on sequencing (GRO-seq) to measure elongation rates *in vivo* (Danko et al., 2013; Jonkers and Lis, 2015). Additionally, while fixed-tissue technologies such as single-molecule RNA-FISH provide superior spatial and molecular resolution to current live imaging technologies (Little et al., 2013; Zoller et al., 2018), the fixation process necessarily prevents temporal analysis of the same single cell to study these dynamic transcriptional processes. Thus, live imaging approaches offer a complementary approach to widespread RNA-FISH studies of transcriptional dynamics (Femino et al., 1998; Raj et al., 2006; Pare et al., 2009; Zenklusen et al., 2008; Wyart et al., 2010; So et al., 2011; Boettiger and Levine, 2013; Little et al., 2013; Jones et al., 2014; Senecal et al., 2014; Fei et al., 2015; Padovan-Merhar et al., 2015; Xu et al., 2015; Albayrak et al., 2016; Skinner et al., 2016; Bartman et al., 2016; Gomez-Schiavon et al., 2017; Hendy et al., 2017; Munsky et al., 2018; Zoller et al., 2018; Miura et al., 2019).

4.3.1 Dissecting the transcription cycle at the single-cell level

From elucidating the nature of mutations (Luria and Delbruck, 1943) and revealing mechanisms of transcription initiation (Zenklusen et al., 2008; Sanchez et al., 2011; So et al., 2011; Sanchez et al., 2013; Sanchez and Golding, 2013; Little et al., 2013; Hocine et al., 2013; Jones et al., 2014; Xu et al., 2015; Choubey et al., 2015; Zoller et al., 2018; Choubey et al., 2018; Filatova et al., 2020), transcription elongation (Boettiger et al., 2011; Serov et al., 2017; Ali et al., 2020), and translational control (Cai et al., 2006), to enabling the calibration of fluorescent proteins in absolute units (Rosenfeld et al., 2005, 2006; Teng et al., 2010; Brewster et al., 2014; Kim et al., 2016; Bakker and Swain, 2019), examining single-cell distributions through the lens of theoretical models has made it possible to extract molecular insights about biological function that are inaccessible through the examination of averaged quantities. The single-cell measurements afforded by our approach made it possible to infer full distributions of transcription parameters (Fig. 4.4B, D, and F). This single-cell resolution motivates a dialogue between theory and experiment for studying transcription initiation, elongation, and cleavage at the single-cell level.

We showed how our inferred distributions of initiation rates effectively filter out most uncorrelated measurement noise, which we expect to be dominated by experimental noise, while retaining information on sources of correlated noise, including underlying biological variability (Fig. 4.4B). Additionally, our theoretical model of elongation rate distributions make it possible to test mechanistic models of RNAP transit along the gene. While still preliminary and far from conclusive, our results suggest that cell-to-cell variability in elongation rates arises from single-molecule variability in stepping rates, and that processes such as stochasticity in stepping behavior and traffic jamming due to steric hindrance alone cannot account for the observed elongation rate distributions (Fig. 4.4D). Such statistics could then be harnessed to make predictions for future perturbative experiments that utilize, for example, mutated RNAP molecules with altered elongation rates (Chen et al., 1996) or reporter genes with differing spacer lengths between MS2 and PP7 stem loops sequences.

Finally, the simultaneous single-cell inference of transcription-cycle parameters granted us the novel capability to investigate couplings between transcription initiation, elongation, and cleavage, paving the way for future studies of mechanistic linkages between these processes. In particular, the observed coupling of the mRNA cleavage time with RNAP density (Fig. 4.5D) suggests future experiments utilizing, for example, orthogonal stem loops on either side of the 3'UTR as potential avenues for investigating mechanisms such as RNAP traffic jams (Klumpp and Hwa, 2008; Klumpp, 2011), inefficient or rate-limiting nascent RNA cleavage (Fong et al., 2015; Jung et al., 2009), and promoter-terminator looping (Hampsey et al., 2011). Other potential experiments could include perturbative effects, such as introducing inhibitors of transcription initiation, elongation, and/or cleavage and assessing the downstream impact on the inferred transcriptional parameters to see if the perturbed effects are separable or convolved between parameters.

4.3.2 Comparison to existing analysis techniques

Our method provides a much-needed framework for applying statistical inference for the analysis of live imaging data of nascent transcription, complementing existing Bayesian approaches (Gupta et al., 2018, 2020) as well as expanding the existing repertoire of model-driven statistical techniques to analyze single-cell protein reporter data (Heron et al., 2007; Finkenstädt et al., 2008; Suter et al., 2011; Zechner et al., 2014). In particular, compared to auto-correlation analysis of transcriptional signals (Coulon and Larson, 2016), another powerful method of analyzing live imaging transcription data, our method is quite complementary.

First, auto-correlation analysis typically requires a time-homogeneous transcript initiation process (Coulon and Larson, 2016), and benefits immensely from having experimental data acquired over long time windows to enhance the auto-correlation signal (although recent work has improved on the ability to analyze short time windows (Desponds et al., 2016)). In contrast, our model-driven inference approach can account for slight time dependence and can fit short time traces. This is of particular relevance to the fly embryo, where each cell cycle in early development is incredibly short (here, we only examined 18 minutes of data) and transcription initiation switches from OFF to ON and back to OFF within that timeframe.

Second, auto-correlation analysis depends strongly on signal-noise ratio, namely the ability to resolve single-or-few-transcript fluctuations in the number of actively transcribing polymerases on a gene (Larson et al., 2011a; Coulon et al., 2014). Our approach, however, can be applied even if the signal-noise ratio can only resolve differences in transcript number of several transcripts, rather than just one.

Third, our model-driven approach benefits from explicitly parameterizing the various steps of the transcription cycle, allowing for the separation of processes such as elongation and cleavage. In contrast, while the auto-correlation technique has the advantage of not relying on a particular specific model, it does rely on unknown parameters such as the overall transcript dwell time, which is a combination of elongation and cleavage. Thus, it becomes harder to separate contributions from these different processes. Additionally, auto-correlation approaches cannot produce absolute rates of transcriptional processes, such as the quantified rates of mean transcription initiation obtained in this work.

4.3.3 Future improvements

Future improvements to experimental or inferential resolution could sharpen precision of single-cell results, increasing confidence in the distributions obtained through this methodology. For example, technologies such as lattice light-sheet microscopy (Chen et al., 2014; Mir et al., 2017, 2018) would vastly improve spatiotemporal imaging resolution and reduce uncertainty in measurements. While this increased resolution is unlikely to dramatically change the statistics reported here, it could potentially push the analysis regime to the single-molecule level, necessitating the parallel development of increasingly refined models that can account for stochasticity and fluctuations that are not resolved with bulk measurements. In addition,

while our analysis restricted itself to consider only nascent RNA labeling technologies, this methodology could be extended to also examine mature labeled RNA in the nucleus and cytoplasm of an organism, providing a more complete picture of transcription.

One important caveat of our method is the failure to account for genes that undergo transcriptional bursting (Rodriguez and Larson, 2020). Here, the initiation rate fluctuates much more rapidly in time such that our assumption of a constant mean transcription initiation rate breaks down. We chose not to address this regime in this work because only a small minority of cells (4%) studied exhibited bursting behavior. Nevertheless, although our model does not capture bursting behavior (Section B.4.3; Fig. B.4E and F), transcriptional bursting remains a prevalent phenomenon in eukaryotic transcription and thus motivates extensions to this work to account for its behavior. For example, one possible implementation to account for transcriptional bursting could first utilize the widespread two-state model used to describe this phenomenon (Peccoud and Ycart, 1995) in order to partition a time trace into ON and OFF time windows. Then the MCMC inference method developed in this work could be used to quantify the transcription cycle during the ON and OFF windows with finer precision.

4.3.4 Outlook

To conclude, while we demonstrated this inference approach in the context of the regulation of a *hunchback* reporter in *Drosophila melanogaster*, it can be readily applied to other genes and organisms in which MS2 and PP7 have been already implemented (Golding et al., 2005; Chubb et al., 2006; Darzacq et al., 2007; Garcia et al., 2013; Lucas et al., 2013; Tantale et al., 2016; Lee et al., 2019; Sato et al., 2020), or where non-genetically encoded RNA aptamer technologies such as Spinach (Paige et al., 2011; Sato et al., 2020) are available. Thus, we envision that our analysis strategy will be of broad applicability to the quantitative and molecular *in vivo* dissection of the transcription cycle and its regulation across many distinct model systems.

4.4 Acknowledgements

We thank Sandeep Choubey, Antoine Coulon, Jane Kondev, Anders Sejr Hansen, Mustafa Mir, Rob Phillips, Manuel Razo-Mejia, and Matthew Ronshaugen for thoughtful comments on the manuscript. We also are grateful to Florian Jug, Nick Lammers, and Armando Reimer for their crucial work in developing the image analysis code used here.

This work was supported by the Burroughs Wellcome Fund Career Award at the Scientific Interface, the Sloan Research Foundation, the Human Frontiers Science Program, the Searle Scholars Program, the Shurl and Kay Curci Foundation, the Hellman Foundation, the NIH Director’s New Innovator Award (DP2 OD024541-01), and an NSF CAREER Award (1652236) (HGG), an NSF GRFP (DGE 1752814) (EE, MT), a UC Berkeley Chancellor’s Fellowship (EE), a Kfas scholarship (YJK), and an DoD NDSEG graduate fellowship (JL).

4.5 Methods and Materials

4.5.1 DNA constructs

The fly strain used to express constitutive MCP-mCherry and PCP-eGFP consisted of two transgenic constructs. The first construct, MCP-NoNLS-mCherry, was created by replacing the eGFP in MCP-NoNLS-eGFP (Garcia et al., 2013) with mCherry. The second construct, PCP-NoNLS-eGFP, was created by replacing MCP in the aforementioned MCP-NoNLS-eGFP with PCP, sourced from Larson et al. (2011a). Both constructs were driven with the *nanos* promoter to deliver protein maternally into the embryo. The constructs lacked nuclear localization sequences because the presence these sequences created spurious fluorescence puncta in the nucleus that decreased the overall signal quality. Both constructs were incorporated into fly lines using P-element transgenesis, and a single stable fly line was created by combining all three transgenes.

The reporter construct P2P-MS2-lacZ-PP7 was cloned using services from GenScript. It was incorporated into the fly genome using PhiC31-mediated Recombinase Mediated Cassette Exchange (RMCE) (Bateman et al., 2006), at the 38F1 landing site.

Full details of construct and sequence information can be found in a public Benchling folder.

4.5.2 Fly strains

Transcription of the *hunchback* reporter was measured by imaging embryos resulting from crossing *yw;MCP-NoNLS-mCherry,Histone-iRFP;MCP-NoNLS-mCherry,PCP-NoNLS-GFP* female virgins with *yw;P2P-MS2-LacZ-PP7* males. The *Histone-iRFP* transgene was provided as a courtesy from Kenneth Irvine and Yuanwang Pan.

4.5.3 Sample preparation and data collection

Sample preparation followed procedures described in Bothma et al. (2014), Garcia and Gregor (2018), and Lammers et al. (2020). To summarize, embryos were collected, dechorinated with bleach and mounted between a semipermeable membrane (Lumox film, Starstedt, Germany) and a coverslip while embedded in Halocarbon 27 oil (Sigma). Excess oil was removed with absorbent paper from the sides to flatten the embryos slightly. Data collection was performed using a Leica SP8 scanning confocal microscope (Leica Microsystems, Biberach, Germany). The MCP-mCherry, PCP-eGFP, and Histone-iRFP were excited with laser wavelengths of 488 nm, 587 nm, and 670 nm, respectively, using a White Light Laser. Average laser powers on the specimen (measured at the output of a 10x objective) were 35 μ W and 20 μ W for the eGFP and mCherry excitation lasers, respectively. Three Hybrid Detectors (HyD) were used to acquire the fluorescent signal, with spectral windows of 496-546 nm, 600-660 nm, and 700-800 nm for the eGFP, mCherry, and iRFP signals, respectively. The confocal stack consisted of 15 equidistant slices with an overall z-height of 7 μ m and an inter-slice distance

of $0.5 \mu\text{m}$. The images were acquired at a time resolution of 15 s, using an image resolution of 512×128 pixels, a pixel size of 202 nm, and a pixel dwell time of $1.2 \mu\text{s}$. The signal from each frame was accumulated over 3 repetitions. Data were taken for 355 cells over a total of 7 embryos, and each embryo was imaged over the first 25 min of nuclear cycle 14.

4.5.4 Image analysis

Images were analyzed using custom-written software following the protocols in Garcia et al. (2013) and Lammers et al. (2020). This software contains MATLAB code automating the analysis of all microscope images obtained in this work, and can be found on a public Github repository (https://github.com/GarciaLab/BcdZldHb_mRNADynamics). Briefly, this procedure involved segmenting individual nuclei using the Histone-iRFP signal as a nuclear mask, segmenting each transcription spot based on its fluorescence, and calculating the intensity of each MCP-mCherry and PCP-eGFP transcription spot inside a nucleus as a function of time. The Trainable Weka Segmentation plugin for FIJI (Arganda-Carreras et al., 2017), which uses the FastRandomForest algorithm, was used to identify and segment the transcription spots. The final intensity of each spot over time was obtained by integrating pixel intensity values in a small window around the spot and subtracting the background fluorescence measured outside of the active transcriptional locus. When no activity was detected, a value of NaN was assigned.

4.5.5 Data Analysis

Inference was done using *MCMCstat*, an adaptive MCMC algorithm (Haario et al., 2001, 2006). Figures were generated using the open-source gramm package for MATLAB, developed by Pierre Morel (Morel, 2018). Generalized linear regression used in Fig. 4.5 utilized a normally distributed error model and was performed using MATLAB's *glmfit* function. All scripts relating to the MCMC inference method developed in this work are available at the associated Github repository (<https://github.com/GarciaLab/TranscriptionCycleInference>).

Appendix A

Supplementary Information for Chapter 2

A.1 Equilibrium Models of Transcription

A.1.1 An overview of equilibrium thermodynamics models of transcription

In this section we give a brief overview of the theoretical concepts behind equilibrium thermodynamics models of transcription. For a more detailed overview, we refer the reader to Bintu et al. (2005b) and Bintu et al. (2005a). These models invoke statistical mechanics in order to calculate bulk properties of a system by enumerating the probability of each possible microstate of the system. The probability of a given microstate is proportional to its Boltzmann weight $e^{-\beta\varepsilon}$, where ε is the energy of the microstate and $\beta = (k_B T)^{-1}$ with k_B being the Boltzmann constant and T the absolute temperature of the system (Garcia et al., 2007).

Specific examples of these microstates in the context of simple activation are featured in Fig. A.1. As reviewed in Garcia et al. (2007), the Boltzmann weight of each of these microstates can also be written in a thermodynamic language that accounts for the concentration of the molecular species, their dissociation constant to DNA, and a cooperativity term ω that accounts for the protein-protein interactions between the activator and RNAP. To calculate the probability of finding RNAP bound to the promoter p_{bound} , we divide the sum of the weights of the RNAP-bound states by the sum of all possible states

$$p_{bound} = \frac{\frac{[P]}{K_p} + \omega \frac{[P][A]}{K_p K_a}}{1 + \frac{[P]}{K_p} + \frac{[A]}{K_a} + \omega \frac{[P][A]}{K_p K_a}}. \quad (\text{A.1})$$

Here, $[P]$ and $[A]$ are the concentrations of RNAP and activator, respectively. K_p and K_a are their corresponding dissociation constants, and ω indicates an interaction between

activator and RNAP: $\omega > 1$ corresponds to cooperativity, whereas $0 < \omega < 1$ corresponds to anti-cooperativity.

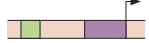
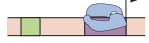
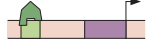
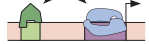
STATE	WEIGHT	RATE
	1	0
	$\frac{[P]}{K_p}$	R
	$\frac{[A]}{K_a}$	0
	$\omega \frac{[P]}{K_p} \frac{[A]}{K_a}$	R

Figure A.1: Equilibrium thermodynamic model of simple activation. A promoter region with one binding site for an activator molecule has four possible microstates, each with its corresponding statistical weight and rate of RNAP loading.

Using p_{bound} , we write the subsequent rate of mRNA production by assuming the *occupancy hypothesis*, which states that

$$\frac{dmRNA}{dt} = R p_{bound}, \quad (\text{A.2})$$

where R is an underlying rate of transcriptional initiation (usually interpreted as the rate of loading RNAP from the promoter-bound state). In the case of simple activation illustrated in Fig. A.1, the overall transcriptional initiation rate is then given by

$$\frac{dmRNA}{dt} = R \frac{\frac{[P]}{K_p} + \omega \frac{[P]}{K_p} \frac{[A]}{K_a}}{1 + \frac{[P]}{K_p} + \frac{[A]}{K_a} + \omega \frac{[P]}{K_p} \frac{[A]}{K_a}}. \quad (\text{A.3})$$

From Eq. A.1, one can derive the Hill equation that is frequently used to model biophysical binding. In the limit of high cooperativity, $\omega \frac{[P]}{K_p} \gg 1$ and $\omega \frac{[A]}{K_a} \gg 1$ such that

$$p_{bound} = \frac{\omega \frac{[P]}{K_p} \frac{[A]}{K_a}}{1 + \omega \frac{[P]}{K_p} \frac{[A]}{K_a}}. \quad (\text{A.4})$$

If we then define a new binding constant $K'_a = \frac{K_a K_p}{\omega [P]}$, we get the familiar Hill equation of order 1 with a binding constant K'_a

$$p_{bound} = \frac{\frac{[A]}{K'_a}}{1 + \frac{[A]}{K'_a}} \quad (\text{A.5})$$

In general, any Hill equation of order n can be derived from a more fundamental equilibrium thermodynamic model of simple activation possessing n activator binding sites in the appropriate limits of high cooperativity. Thus, any time a Hill equation is invoked, equilibrium thermodynamics is implicitly used, bringing with it all of the underlying assumptions described in Section A.6.5. This highlights the importance of rigorously grounding the assumptions made in any model of transcription, to better discriminate between the effects of equilibrium and non-equilibrium processes.

A.1.2 Thermodynamic MWC model

In the thermodynamic MWC model, we consider a system with six Bicoid binding sites and ten Zelda binding sites. In addition, we allow for RNAP binding to the promoter.

In our model, the DNA can be in either an accessible or an inaccessible state. The difference in free energy between the two states is given by $-\Delta\varepsilon_{\text{chrom}}$, where $\Delta\varepsilon_{\text{chrom}}$ is defined as

$$\Delta\varepsilon_{\text{chrom}} = \varepsilon_{\text{accessible}} - \varepsilon_{\text{inaccessible}}. \quad (\text{A.6})$$

Here, $\varepsilon_{\text{accessible}}$ and $\varepsilon_{\text{inaccessible}}$ are the energies of the accessible and inaccessible states, respectively. A positive $\Delta\varepsilon_{\text{chrom}}$ signifies that the inaccessible state is at a lower energy level, and therefore more probable, than the accessible state. We assume that all binding sites for a given molecular species have the same binding affinity, and that all accessible states exist at the same energy level compared to the inaccessible state. Thus, the total number of states is determined by the combinations of occupancy states of the three types of binding sites as well as the presence of the inaccessible, unbound state. We choose to not allow any transcription factor or RNAP binding when the DNA is inaccessible.

In this equilibrium model, the statistical weight of each accessible microstate is given by the thermodynamic dissociation constants K_b , K_z , and K_p of Bicoid, Zelda, and RNAP respectively. The statistical weight for the inaccessible state is $e^{\Delta\varepsilon_{\text{chrom}}}$. We allow for a protein-protein interaction term ω_b between nearest-neighbor Bicoid molecules, as well as a pairwise cooperativity ω_{bp} between Bicoid and RNAP. However, we posit that Zelda does not interact directly with either Bicoid or RNAP. For notational convenience, we express the statistical weights in terms of the non-dimensionalized concentrations of Bicoid, Zelda, and RNAP, given by b , z and p , respectively, such that, for example, $b \equiv \frac{[\text{Bicoid}]}{K_b}$. Fig. A.2 shows the states and statistical weights for this thermodynamic MWC model, with all the associated parameters.

Incorporating all the microstates, we can calculate a statistical mechanical partition function, the sum of all possible weights, which is given by

$$Z = e^{\Delta\varepsilon_{\text{chrom}}/k_B T} + \underbrace{(1+z)^{10}}_{\text{Zelda binding}} \underbrace{(1+b+b^2\omega_b+\dots+b^6\omega_b^5+p+pb\omega_{bp}+\dots+pb^6\omega_b^5\omega_{bp}^6)}_{\text{Bicoid and RNAP binding}}. \quad (\text{A.7})$$

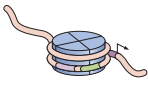
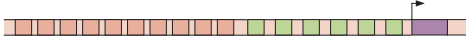
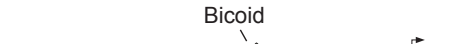










STATE	WEIGHT	RATE
	$e^{\beta \Delta \epsilon_{chrom}}$	0
	1	0
	b	0
	$b^2 \omega_b$	0
	z	0
	$z b$	0
...		
	$z^{10} b^6 \omega_b^5$	R
	p	R
	$b p \omega_{bp}$	R
	$b^2 \omega_b p \omega_{bp}^2$	R
	$z p$	R
	$z b p \omega_{bp}$	R
...		
	$z^{10} b^6 \omega_b^5 p \omega_{bp}^6$	R

Figure A.2: States, weights, and rate of RNAP loading diagram for the thermodynamic MWC model, containing six Bicoid binding sites, ten Zelda binding sites, and a promoter.

Using the binomial theorem

$$(a + b)^N = \sum_{n=0}^N \binom{N}{n} a^n b^{N-n},$$

Eq. A.7 can be expressed more compactly as

$$Z = e^{\Delta\varepsilon_{chrom}/k_B T} + (1 + z)^{10} \left(1 + p + \sum_{j=0,1} \sum_{i=1}^6 \binom{6}{i} b^i \omega_b^{i-1} p^j \omega_{bp}^{ij} \right). \quad (\text{A.8})$$

From this partition function, we can calculate p_{bound} , the probability of being in an RNAP-bound state. This term is given by the sum of the statistical weights of the RNAP-bound states divided by the partition function

$$p_{bound} = \frac{1}{Z} \left((1 + z)^{10} p \left(1 + \sum_{i=1}^6 \binom{6}{i} b^i \omega_b^{i-1} \omega_{bp}^i \right) \right). \quad (\text{A.9})$$

In this model, we once again assume that the transcription associated with each microstate is zero unless RNAP is bound, in which case the associated rate is R . Then, the overall transcriptional initiation rate is given by the product of p_{bound} and R

$$\frac{dmRNA}{dt} = R \frac{1}{Z} \left((1 + z)^{10} p \left(1 + \sum_{i=1}^6 \binom{6}{i} b^i \omega_b^{i-1} \omega_{bp}^i \right) \right). \quad (\text{A.10})$$

Note that since the MS2 technology only measures nascent transcripts, we can ignore the effects of mRNA degradation and focus on transcriptional initiation.

A.1.3 Constraining model parameters

The transcription rate R of the RNAP-bound states can be experimentally constrained by making use of the fact that the *hunchback* minimal reporter used in this work produces a step-like pattern of transcription across the length of the fly embryo (Fig. 2.4C, blue points). Since in the anterior end of the embryo, the observed transcription appears to level out to a maximum value, we assume that Bicoid binding is saturated in this anterior end of the embryo such that

$$p_{bound}(b \rightarrow \infty) \approx 1. \quad (\text{A.11})$$

In this limit, Eq. A.10 can be written as

$$\frac{dmRNA}{dt} = R_{max} \approx R, \quad (\text{A.12})$$

where R_{max} is the maximum possible transcription rate. Importantly, R_{max} is an experimentally observed quantity rather than a free parameter. As a result, the model parameter R is determined by experimentally measurable quantity R_{max} .

The value of p can also be constrained by measuring the transcription rate in the embryo's posterior, where we assume Bicoid concentration to be negligible. Here, the observed transcription bottoms out to a minimum level R_{min} (Fig. 2.4C, blue points), which we can connect with the model's theoretical minimum rate. Specifically, in this limit, b approaches zero in Eq. A.10 such that all Bicoid-dependent terms drop out, resulting in

$$\frac{dmRNA}{dt} = R_{min} \approx \frac{1}{Z} \left((1+z)^{10} p \right) R_{max}, \quad (\text{A.13})$$

where we have replaced R with R_{max} as described above. Next, we can express p in terms of the other parameters such that

$$p \approx \frac{R_{min} \left(e^{\Delta\varepsilon_{chrom}/k_B T} + (1+z)^{10} \right)}{\left(R_{max} - R_{min} \right) (1+z)^{10}}. \quad (\text{A.14})$$

Thus, p is no longer a free parameter, but is instead constrained by the experimentally observed maximum and minimum rates of transcription R_{max} and R_{min} , as well as our choices of K_z and $\Delta\varepsilon_{chrom}$. In our analysis, R_{max} and R_{min} are calculated by taking the mean RNAP loading rate across all embryos from the anterior and posterior of the embryo respectively, extrapolated using the trapezoidal fitting scheme described in Section A.2.3.

Finally, we expand this thermodynamic MWC model to also account for suppression of transcription in the beginning of the nuclear cycle via mechanisms such as mitotic repression (Section A.3). To make this possible, we include a trigger time term t_{MitRep} , before which we posit that no readout of Bicoid or Zelda by *hunchback* is possible and the rate of RNAP loading is fixed at 0. For times $t > t_{MitRep}$, the system behaves according to Eq. A.10. Thus, given the constraints stemming from direct measurements of R_{max} and R_{min} , the model has six free parameters: $\Delta\varepsilon_{chrom}$, ω_b , ω_{bp} , K_b , K_z , and t_{MitRep} . The final calculated transcription rate is then integrated in time to produce a predicted MS2 fluorescence as a function of time (Section A.2.2).

For subsequent parameter exploration of this model (Section A.5.1), constraints were placed on the parameters to ensure sensible results. Each parameter was constrained to be strictly positive such that:

- $\Delta\varepsilon_{chrom} > 0$
- $K_b > 0$
- $K_z > 0$
- $\omega_b > 0$

- $\omega_{bp} > 0$
- $0 < t_{MitRep} < 10$.

where an upper limit of 10 min was placed on the mitotic repression term to ensure efficient parameter exploration. This was justified because none of the observed transcriptional onset times in the data were larger than this value (Fig. 2.4D).

A.2 Input-Output measurements, predictions, and characterization

A.2.1 Input measurement methodology

Input transcription-factor measurements were carried out separately in individual embryos containing a eGFP-Bicoid transgene in a *bicoid* null mutant background (Gregor et al., 2007b) or a Zelda-sfGFP CRISPR-mediated homologous recombination at the endogenous *zelda* locus (Hamm et al., 2017). Over the course of nuclear cycle 13, the fluorescence inside each nucleus was extracted (details given in Section 4.5.4), resulting in a measurement of the nuclear concentration of each transcription factor over time. Six eGFP-Bicoid and three Zelda-sfGFP embryos were imaged.

Representative fluorescence traces of eGFP-Bicoid for a single embryo indicate that the magnitude of eGFP-Bicoid fluorescence decreases for nuclei located toward the posterior of the embryo (Fig. A.3A). Further, the nuclear fluorescence of eGFP-Bicoid at 8 min into nuclear cycle 13 (Fig. A.3B) exhibited the known exponential decay of Bicoid, with a mean decay length of $23.5\% \pm 0.6\%$ of the total embryo length, consistent with but slightly different than previous measurements that suggested a mean decay length of $19.1\% \pm 0.8\%$ (Liu et al., 2013). This discrepancy could stem, for example, from minor differences in acquisition from the laser-scanning two-photon microscope used in Liu et al. (2013) versus the laser-scanning confocal microscope used here, such as differences in axial resolution (due both to different choices of objectives and the inherent differences in axial resolution of one-photon and two-photon fluorescence excitation processes). Nevertheless, the difference was minute enough that we felt confident in our eGFP-Bicoid measurements.

Intra-embryo variability in eGFP-Bicoid nuclear fluorescence, defined by the standard deviation across nuclei within a single embryo divided by the mean, was in the range of 10-30%, as was the inter-embryo variability, defined by the standard deviation of the mean amongst nuclei, across different embryos (Fig. A.3C, blue and black, respectively). Six separate eGFP-Bicoid embryos were measured.

Similarly, representative fluorescence time traces of Zelda-sfGFP for a single embryo are shown in Fig. A.3D. Unlike the eGFP-Bicoid profile, the Zelda-sfGFP nuclear fluorescence was approximately uniform across embryo position (Fig. A.3E), consistent with previous fixed-tissue measurements (Staudt et al., 2006; Liang et al., 2008). Intra-embryo variability in Zelda-sfGFP nuclear fluorescence was very low (less than 10%), whereas inter-embryo

variability was relatively higher, up to 20% (Fig. A.3F, red and black, respectively). Three separate Zelda-sfGFP embryos were measured.

Due to the consistency of Zelda-sfGFP nuclear fluorescence, we assumed the Zelda profile to be spatially uniform in our analysis, and thus created a mean Zelda-sfGFP measurement for each individual embryo by averaging all mean nuclear fluorescence traces in space across the anterior-posterior axis of the embryo (Fig. A.3D, inset). This mean measurement was used as an input in the theoretical models. However, we still retained inter-embryo variability in Zelda, as described below.

To combine multiple embryo datasets as inputs to the models explored throughout this work, the fluorescence traces corresponding to each dataset were aligned at the start of nuclear cycle 13, defined as the start of anaphase. Because each embryo may have possessed slightly different nuclear cycle lengths and/or experimental sampling rates (due to the manual realignment of the z-stack to keep nuclei in focus), the individual datasets were not combined in order to create average Bicoid and Zelda profiles across embryos. Instead, a simulation and model prediction were performed for each combination of measured input Bicoid and Zelda datasets, essentially an *in silico* experiment covering a portion of the full embryo length. In all, outputs at each embryo position were predicted in at least three separate simulations. Subsequent analyses used the mean and standard error of the mean of these amalgamated simulations. With six GFP-Bicoid datasets and three Zelda-GFP datasets, there were 18 unique combinations of input embryo datasets; for a single set of parameters used in a particular model, each derived metric (e.g. t_{on}) was calculated using predicted outputs from each of the 18 possible input combinations. This procedure provided full embryo coverage and resulted in a distribution of the derived metric for that particular set of parameters. From this distribution, the mean and standard error of the mean were calculated, leading to the lines and shading in plots such as Fig. A.6.

A.2.2 MS2 fluorescence simulation protocol

To calculate a predicted MS2 fluorescence trace from measured Bicoid and Zelda inputs for a given theoretical model, we utilized a simple model of transcription initiation, elongation, and termination. First, the dynamic transcription-factor concentrations were used as inputs to each of the theoretical models outlined throughout Chapter 2. These models generated a rate of RNAP loading as a function of time and space across the embryo over the course of nuclear cycle 13.

For each position along the anterior-posterior axis, the predicted rate of RNAP loading was integrated over time to generate a predicted MS2 fluorescence trace. Given the known reporter construct length L of 5.2 kb (Garcia et al., 2013), we assume that RNAP molecules are loaded onto the start of the gene at a rate $R(t)$ predicted by the particular model under consideration (Fig. A.4; see Sections A.1.2, A.6.1, A.7.1, and A.8.1 for model details). Each RNAP molecule traverses the gene at a constant velocity v of 1.54 kb/min, as measured

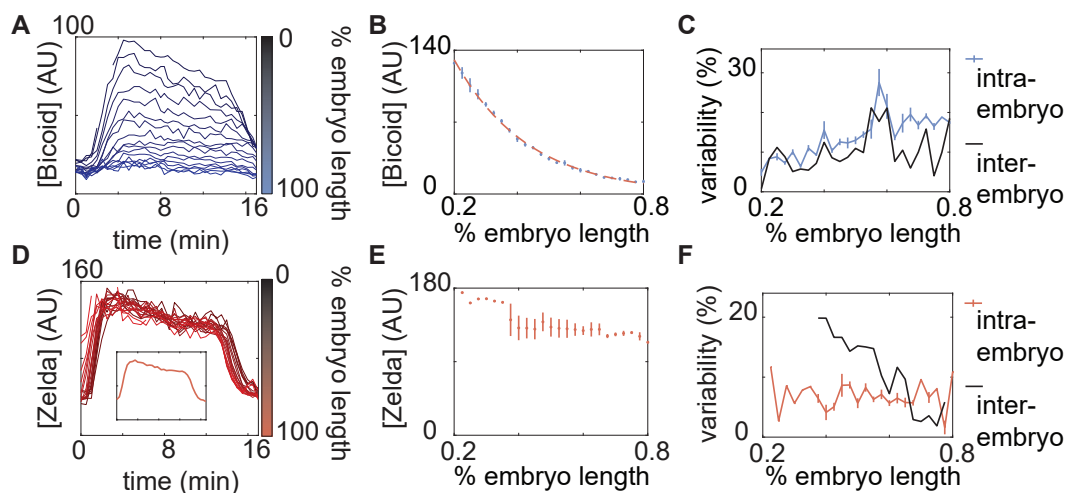


Figure A.3: Measurements of input transcription-factor concentration dynamics. (A) Nuclear eGFP-Bicoid concentration as a function of time into nuclear cycle 13 across various positions along the anterior-posterior axis of a single embryo. (B) eGFP-Bicoid concentration at 8 min into nuclear cycle 13 as a function of position along the embryo averaged over all measured embryos ($n=6$). The fit of the concentration profile to an exponential function (dashed line) results in a decay length of $23\% \pm 0.6\%$ embryo length. (C) Intra- and inter-embryo variability in eGFP-Bicoid nuclear fluorescence along the anterior-posterior axis. (D) Zelda-sfGFP concentration as a function of time into nuclear cycle 13 across various anterior-posterior positions of a single embryo. (D, inset) Zelda-sfGFP concentration averaged over the data shown in D. (E) Zelda-sfGFP concentration at 8 min into nuclear cycle 13 as a function of position along the anterior-posterior axis of the embryo averaged over all measured embryos ($n=3$). Note that anterior of 40% and posterior of 77.5% only a single embryo was measured; no error bars were calculated. (F) Intra- and inter-embryo variability in Zelda-sfGFP nuclear fluorescence along the anterior-posterior axis. (B,E, error bars represent standard error of the mean nuclear fluorescence, measured across embryos; C,F, error bars represent standard error of the mean intra-embryo variability, measured across embryos.)

experimentally by Garcia et al. (2013). With these numbers, we calculate an elongation time

$$t_{elon} = \frac{L}{v}. \quad (\text{A.15})$$

Finally, we assume that upon reaching the end of the reporter gene, the RNAP molecules terminate and disappear instantly such that they no longer contribute to spot fluorescence.

The MS2 fluorescence signal reports on the number of RNAP molecules actively occupying the gene at any given time and, under the assumptions outlined above, is given by the integral

$$F(t) = \alpha \int_0^t \left(R(t') - R(t' - t_{elon}) \right) dt', \quad (\text{A.16})$$

where $F(t)$ is the predicted fluorescence value, $R(t)$ is the RNAP loading rate predicted by each specific model, $R(t - t_{elon})$ is the time-shifted loading rate that accounts for RNAP molecules finishing transcription at the end of the gene, and α is an arbitrary scaling factor to convert from absolute numbers of RNAP molecules to arbitrary fluorescence units. The predicted value $F(t)$ was scaled by α to match the experimental data.

The final predicted MS2 signal was modified in a few additional ways. First, any RNAP molecule that had not yet reached the position of the MS2 stem loops had its fluorescence value set to zero (Fig. A.4, i), since only RNAP molecules downstream of the MS2 stem loop sequence exhibit a fluorescent signal. Second, RNAP molecules that were only partially done elongating the MS2 stem loops contributed a partial fluorescence intensity, given by the ratio of the distance traversed through the stem loops to the total length of the stem loops

$$F_{partial} = \frac{L_{partial}}{L_{loops}},$$

where $F_{partial}$ is the partial fluorescence contributed by an RNAP molecule within the stem loop sequence region, $L_{partial}$ is the distance within the stem loop sequence traversed, and L_{loops} is the length of the stem loop sequence (Fig. A.4, ii). For this reporter construct, the length of the stem loops was approximately $L_{loops} = 1.28 \text{ kb}$. RNAP molecules that had finished transcribing the MS2 stem loops contributed the full amount of fluorescence (Fig. A.4, iii). Finally, to make this simulation compatible with the trapezoidal fitting scheme in Section A.2.3, we included a falling signal at the end of the nuclear cycle, achieved by setting $R(t) = 0$ after 17 min into the nuclear cycle and thus preventing new transcription initiation events.

Given the predicted MS2 fluorescence trace, the rate of RNAP loading and t_{on} were extracted with the fitting procedure used on the experimental data (Section A.2.3).

A.2.3 Extracting initial RNAP loading rate and transcriptional onset time

To extract the initial rate of RNAP loading and the transcriptional onset time t_{on} used in the data analysis, we fit both the experimental and calculated MS2 signals to a constant loading rate model, the trapezoidal model (Garcia et al., 2013).

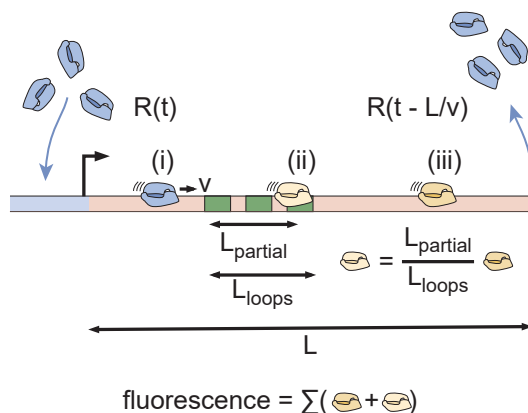


Figure A.4: MS2 fluorescence calculation protocol. RNAP molecules load onto the reporter gene at a time-dependent rate $R(t)$, after which they elongate at a constant velocity v . Upon reaching the end of the gene after a length L has been transcribed, they are assumed to terminate and disappear instantly, given by the time-shifted rate $R(t - \frac{L}{v})$. The time-dependent MS2 fluorescence is calculated by summing the contributions of RNAP molecules that are located before, within, or after the MS2 stem loop sequence (i, ii, and iii, respectively).

The trapezoidal model provides a heuristic fit of the main features of the MS2 signal by assuming that the RNAP loading rate is either zero or some constant value r (Fig. A.5A). At time t_{on} , the loading rate switches from zero to this constant value r , producing a linear rise in the MS2 signal. After the elongation time t_{elong} , the loading of new RNAP molecules onto the gene is balanced by the loss of RNAP molecules at the end of the gene, producing a plateau in the MS2 signal. Finally, at the end of the nuclear cycle, transcription ceases at t_{off} and the RNAP loading rate switches back to zero, producing the falling edge of the MS2 signal and completing the trapezoidal shape. Because we only consider the initial dynamics of transcription in the nuclear cycle in this investigation, we do not explore the behavior of t_{off} .

Fig. A.5B shows the results of fitting the mean MS2 fluorescence from a narrow window within a single embryo to the trapezoidal model. With this fit, we can extract the initial rate of RNAP loading (given by the initial slope) as well as t_{on} (given by the intercept of the fit onto the x-axis).

As a consistency check, the t_{on} values extrapolated from the trapezoidal fit of the data were compared with the experimental time points at which the first MS2 spots were observed for both the wild-type and *zelda*⁻ mutant experiments (Fig. A.5C). Due to the detection limit of the microscope, this latter method reports on the time at which a few RNAP molecules have already begun transcribing the reporter gene, rather than a “true” transcriptional onset time. Using the first frame of spot detection yields similar trends to the trapezoidal

fits, except that the measured first frame times are systematically larger, especially in the mutant data. Additionally, utilizing the first frame of detection to measure t_{on} appears to be a noisier method, likely because the actual MS2 spots cannot be observed below a finite signal-detection limit, whereas the extrapolated t_{on} from the trapezoidal fit corresponds to a “true” onset time below the signal-detection limit. For this reason, we decided to rely on the trapezoidal fit to extract t_{on} , rather than using the first frame of spot detection.

A.3 Mitotic repression is necessary to recapitulate Bicoid- and Zelda-mediated regulation of *hunchback* using the thermodynamic MWC model

As described in Section 2.2.3 of the main text, a mitotic repression window was incorporated into the thermodynamic MWC model (Section A.1.2) in order to explain the observed transcriptional onset times of *hunchback*. Here, we justify and explain this theoretical modification in greater detail.

Fig. A.6A and B depicts the experimentally observed initial rates of RNAP loading and t_{on} across the length of the embryo (blue points) for the wild-type background. After constraining the maximum and minimum theoretically allowed rates of RNAP loading (Section A.1.3), we attempted to simultaneously fit the thermodynamic MWC model to both the rate of RNAP loading and t_{on} .

The fit results demonstrate that while the thermodynamic MWC model can recapitulate the measured step-like rate of RNAP loading at *hunchback* (Fig. A.6A, purple line), it fails to predict the t_{on} throughout the embryo (Fig. A.6B, purple line; see Sections A.2.2 and A.2.3 for details about experimental and theoretical calculations). This model yields values of t_{on} that are much smaller than those experimentally observed, a trend that holds throughout the length of the embryo. This disagreement becomes more evident when comparing the output transcriptional activity reported by the measured MS2 fluorescence with the input concentrations of Bicoid and Zelda. Specifically, the Bicoid and Zelda concentration measurements at 45% along the embryo, shown for a single embryo in Fig. A.6C, are used in conjunction with the previously mentioned best-fit model parameters to predict the output MS2 signal at the same position. This prediction can then be directly compared with experimental data (Fig. A.6D, purple line vs. black points, respectively). Whereas the model predicts that transcription will commence around 1 min after anaphase due to the concurrent increase in the Bicoid and Zelda concentrations, the observed MS2 signal begins to increase around 4 min after anaphase (Fig. A.6D). As a result, the predicted transcriptional dynamics in Fig. A.6D are systematically shifted in time with respect to the observed data.

The observed disagreement in t_{on} suggests that in this model, transcription is prevented from starting at the time dictated solely by the increase of Bicoid and Zelda concentrations.

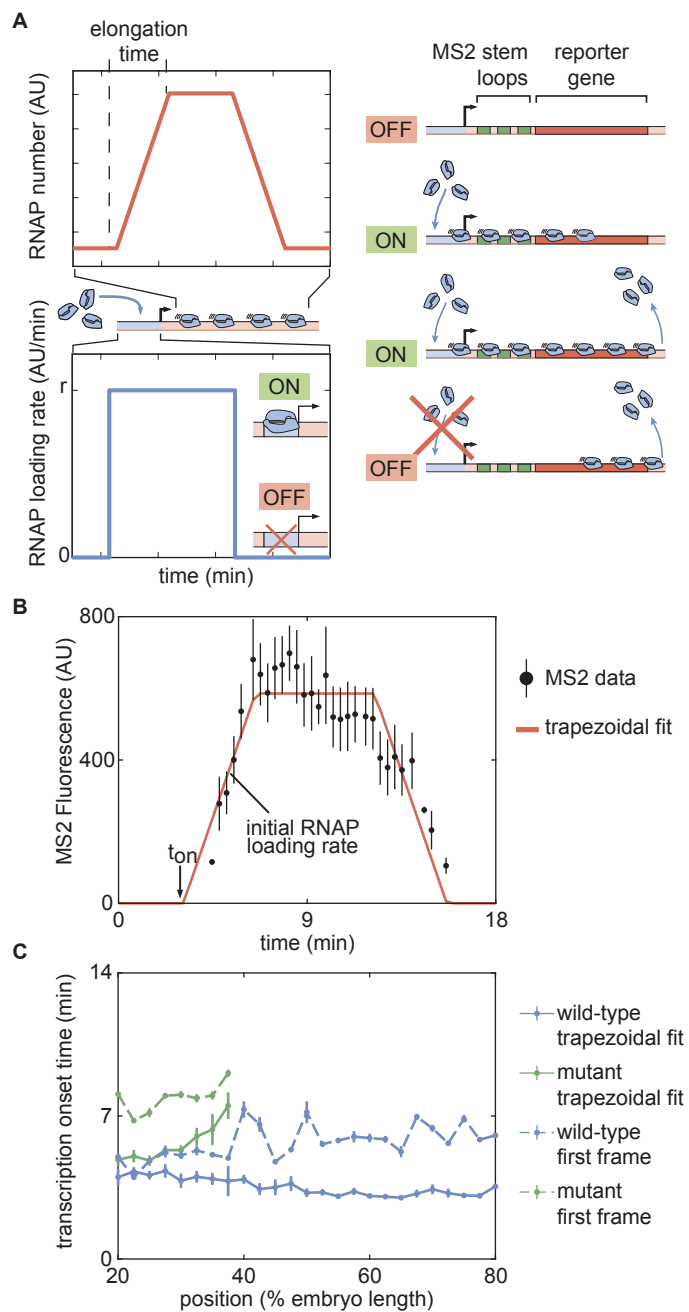


Figure A.5: Outline of fitting to the trapezoidal model of transcription. See caption on next page.

Figure A.5: Outline of fitting to the trapezoidal model of transcription. (A) The trapezoidal model of transcription, where transcription begins at an onset time t_{on} and loads RNAP molecules with a constant rate r . (B) Results of fitting the MS2 fluorescence data from a position in a single embryo to the trapezoidal model to extract t_{on} and the initial rate of RNAP loading. (C) Comparison of inferred t_{on} values between the trapezoidal model (solid lines) and using the time of first detection of signal in a fluorescence spot (dashed lines) for both wild-type and *zelda*⁻ backgrounds. (B, error bars are standard error of the mean averaged over multiple nuclei within the embryo, for data in a wild-type background at 50% along the embryo length; C, error bars are standard error of the mean, averaged across embryos).

While we speculate that this effect could stem from processes such as RNAP escape from the promoter, DNA replication at the start of the cell cycle, and post-mitotic nucleosome clearance from the promoter, we choose not to commit to a detailed molecular picture and instead ascribe this transcriptional refractory period at the beginning of the nuclear cycle to mitotic repression, the observation that the transcriptional machinery cannot operate during mitosis (Shermoen and O’Farrell, 1991; Gottesfeld and Forbes, 1997; Parsons and Tg, 1997; Garcia et al., 2013). To account for this phenomenon, we revised our thermodynamic MWC model by stating that *hunchback* can only read out the inputs and begin transcription after a specified mitotic repression time window following the previous anaphase (Section A.1.3).

Since we expect mitotic repression to operate independently of position along the length of the embryo (Shermoen and O’Farrell, 1991), we assumed that the duration of mitotic repression was uniform throughout the embryo. After incorporating a uniform 3 min mitotic repression window into the thermodynamic MWC model (Fig. A.6C and D, grey shaded region), the model successfully recapitulates t_{on} throughout the embryo (Fig. A.6B and D, blue curves), while still explaining the observed rates of RNAP loading (Fig. A.6A, blue curve). Thus, once mitotic repression is accounted for, the thermodynamic MWC model based on statistical mechanics can quantitatively recapitulate the regulation of *hunchback* transcription by Bicoid and Zelda.

A.4 The effect of the *zelda*⁻ background on the Bicoid concentration spatiotemporal profile

Our models rest on the assumption that the Bicoid gradient remains unaltered regardless of whether these measurements are made in the wild-type or *zelda*⁻ backgrounds. To confirm this assumption, we measured eGFP-Bicoid concentrations in a *zelda*⁻ background. These flies were heterozygous for eGFP-labeled Bicoid and for wild-type Bicoid, resulting in roughly

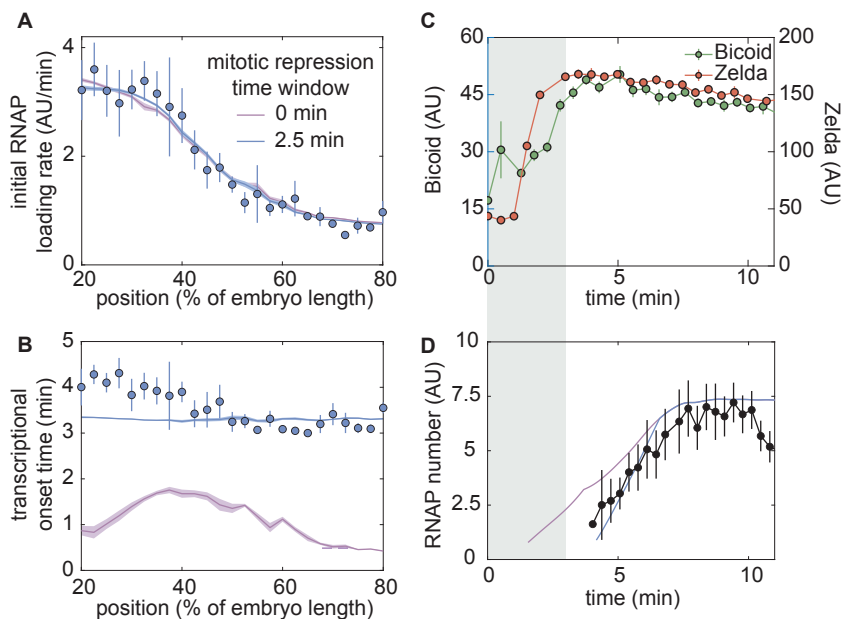


Figure A.6: A thermodynamic MWC model including mitotic repression can recapitulate *hunchback* regulation by Bicoid and Zelda. (A) Measured initial rates of RNAP loading and (B) t_{on} (blue points) across the length of the embryo, compared to fits to the thermodynamic MWC model with and without accounting for mitotic repression (blue and purple curves, respectively). (C) Nuclear concentration dynamics of Bicoid and Zelda with proposed mitotic repression window (gray shading). (D) Predicted MS2 dynamics with no mitotic repression term or a 3 min mitotic repression window compared to experimental measurements. (A,B, solid lines indicate mean predictions of the model and shading represents standard error of the mean, while points indicate data and error bars represent the standard error of the mean, across 11 embryos; C, D, data from single embryos at 45% of the embryo length with error bars representing the standard error of the mean across nuclei, errors in model predictions in D were negligible and are obscured by the prediction curve; fitted parameter values for a 3 min mitotic repression window were $\Delta\epsilon_{chrom} = 10 k_B T$, $K_b = 34 AU$, $K_z = 500 AU$, with different arbitrary fluorescent units for Bicoid and Zelda, $\omega_b = 10$, $\omega_{bp} = 0.4$, for a model assuming six Bicoid binding sites.)

50% of total Bicoid being labeled with eGFP. As shown in Fig. A.7A and B, the resultant eGFP-Bicoid nuclear fluorescence levels in nuclear cycle 13 in the *zelda*⁻ background (red) were roughly half the magnitude of the equivalent measurements in the wild-type background (blue), a trend that held both in time and along the embryo. After doubling the heterozygote eGFP-Bicoid nuclear fluorescence measurements to rescale them (Fig. A.7B, black), the two eGFP-Bicoid curves became similar, although the *zelda*⁻ eGFP-Bicoid values were systematically lower than in the wild-type background. The normalized difference, defined as the absolute value of the difference between the wild-type and *zelda*⁻ profiles at each position in the embryo divided by the value of the wild-type profile at the position, averaged across all measured positions, was $15\% \pm 2\%$. This value is within the range of the inter-embryo variability of eGFP Bicoid in wild-type background embryos (Fig. A.3C). Measuring the decay length of the eGFP-Bicoid profile in the *zelda*⁻ background also yielded a slightly different result: $21\% \pm 1\%$ of the total embryo length, as opposed to $23.5\% \pm 0.6\%$ in the wild-type background (dashed curves, see also Fig. A.3B).

Having compared the spatial profile of Bicoid in both backgrounds, we then contrasted the dynamics of nuclear Bicoid import. To quantify this analysis, we calculated the time to reach 50% and 90% of the maximum eGFP-Bicoid fluorescence signal for wild-type and *zelda*⁻ embryos, at each position along the anterior-posterior axis (Fig. A.7A, blue and red dashed lines). Because the raw fluorescence signals were noisy enough to confound this calculation, we first smoothed the signals using a moving average filter of ten datapoints (Fig. A.7A, lines).

Fig. A.7C and D show the times to reach 50% and 90% of maximum fluorescence for the anterior positions in both embryo backgrounds, where transcription was observed, respectively. In both backgrounds, the 50% and 90% times are similar to within approximately 1 min, indicating that the dynamics of nuclear eGFP-Bicoid at the start of nuclear cycle 13 are quantitatively comparable. Thus, we concluded that differences in transcription between the two embryo backgrounds do not stem from differences in Bicoid dynamics.

In summary, the dynamics of nuclear Bicoid concentration are quantitatively comparable in both wild-type and *zelda*⁻ backgrounds, whereas the overall Bicoid concentration is slightly lower in the *zelda*⁻ case. Nevertheless, these differences in concentration would have a negligible effect on our overall conclusions: in the context of our models, an overall rescaling in the magnitude of the Bicoid gradient between the wild-type and *zelda*⁻ backgrounds can be compensated by a corresponding rescaling in the dissociation constant of Bicoid, K_b . Because our systematic exploration of theoretical models considers many possible parameter values (Section A.5.1), this rescaling has no effect on our conclusion that the equilibrium models are insufficient to explain the *zelda*⁻ data. As a result, and given that our statistics for the wild-type eGFP-Bicoid data consisted of more embryos than the data for the *zelda*⁻ background, we used this wild-type data in our analyses as an input to both the wild-type and *zelda*⁻ model calculations.

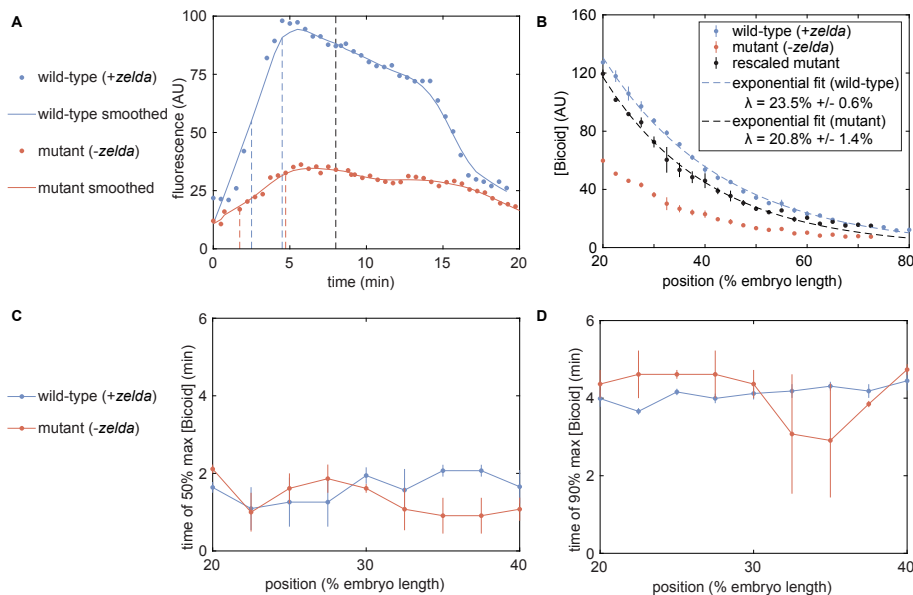


Figure A.7: Comparison of eGFP-Bicoid measurements in wild-type and *zelda*⁻ embryos. (A) Example mean nuclear eGFP-Bicoid concentrations for a single embryo at 30% along the embryo’s length, for wild-type (blue) and *zelda*⁻ (red) backgrounds. Datapoints are raw values and lines are smoothed results. The raw fluorescence at 8 min into the nuclear cycle, indicated by the black dashed line, is calculated to yield (B). The blue and red dashed lines correspond to the times to reach 50% and 90% of the maximum fluorescence for the smoothed wild-type and *zelda*⁻ signals, respectively. (B) eGFP-Bicoid measurements in wild-type (blue) and *zelda*⁻ mutant embryos (red), along with rescaled mutant profiles (black). Fits to an exponentially decaying function yield decay lengths in each background (blue and black dashed curves). (C, D) Time to reach 50% and 90% of maximum nuclear eGFP-Bicoid fluorescence for wild-type (blue) and *zelda*⁻ (red) backgrounds. A total of $n=3$ embryos were measured in the *zelda*⁻ background, compared to $n=6$ for wild-type. All error bars are standard error of mean across embryos.

A.5 State-space exploration of theoretical models

A.5.1 General methodology of state-space exploration

To help visualize the limits of our models, we collapsed our observations onto a three-dimensional state space, following a method similar to that described in Estrada et al. (2016). In this space, the x-axis was the average t_{on} delay. This magnitude was computed by integrating the t_{on} across 20% to 37.5% of the embryo length, corresponding to the range in which both wild-type and *zelda*⁻ experiments exhibited transcription in at least 30% of

observed nuclei (Figs. A.8A and 2.5A), as defined in Equation 2.5. The offset in t_{on} at 20% embryo length (Eq. 2.4) was the y-axis in the state space. The z-axis was given by the average initial rate of RNAP loading between 20% and 37.5% of the embryo length (Fig. A.8B).

Combined, the average t_{on} delay, t_{on} offset, and average initial RNAP loading rate provide a simplified description of our data as well as of our theoretical predictions. Each theoretical model inhabits a finite region in this three-dimensional state space, which we can calculate by systematically varying model parameters. Fig. A.8A and B show an example of how the three parameters are calculated using the *zelda*⁻ background data presented in Fig. 2.4C and D (red points) in the main text.

Due to the large number of parameters in each model explored, the corresponding state-space boundaries were generated by efficient sampling of the underlying high-dimensional parameter space. Although in actuality the state space contained three dimensions, we illustrate the sampling process here with a two-dimensional example, using only the offset and average delay in transcriptional onset time, for ease of visualization (Fig. A.8C). The methodology is similar to the one described in Estrada et al. (2016). Briefly, a starting set of 50 points was generated, each with a randomized set of initial parameters, the specifics of which depended on the model being tested (Fig. A.8C, i). The state space was sectioned into 100 slices along each orthogonal axis (Fig. A.8C, ii). The most extremal points in each slice were found, resulting in two extremal values each for the t_{on} offset and average t_{on} delay (Fig. A.8C, iii). For each of these points, a new set of five points was generated using random parameters within a small neighborhood of the seed points determined by the extremal points of the previous iteration (Fig. A.8C, iv). These new points were plotted; some of these points may be more extreme than the previous set of points. Steps ii-iv were iterated, resulting in a growing boundary over time (Fig. A.8C, v). This algorithm was run in the full three-dimensional state space, where 100 three-dimensional columns along the orthogonal xy-, yz-, and xz-planes were used instead of two-dimensional slices.

Constraints imposed by the data were used to filter unrealistic results and ensure rapid convergence of the algorithm. First, if the simulated average t_{on} delay was less than -0.5 min or greater than 2 min, the point was filtered out. This removal was justified experimentally, since none of the observed average t_{on} delays were outside of this range (Fig. 2.5A). Second, if the simulated average initial loading rate was smaller than 1 AU/min or greater than 4 AU/min, the point was also filtered out. This was also justified experimentally, since none of the observed initial RNAP loading rates between 20% and 37.5% embryo position lay outside this range (Fig. 2.4C). Points that fulfilled these constraints were retained for the next iteration of the algorithm. This process was repeated until the resulting space of points no longer grew appreciably, resulting in an estimate of the size and shape of the state space for each of the models presented in Sections A.1.2, A.6.1, A.7.1, and A.8.1.

To determine whether the algorithm had indeed converged, the total volume of each model's region in state space was tracked with each iteration number. If the algorithm worked well, then this volume would approach some maximum value. Fig. A.8D shows the volume of the state space corresponding to each model presented in this work (normalized by the volume at the final iteration number) as a function of the iteration number. Each model

converged to a finite value, indicating that the parameter space occupied by the models had been thoroughly explored.

A.5.2 State space exploration with the thermodynamic MWC model

Fig. A.9A and Video A.9.7 show the resulting three-dimensional state space for the thermodynamic MWC model (green), as well as all of the theoretical models considered here. We plotted the wild-type and *zelda*⁻ data on the same state space, represented as small ellipsoids of uncertainty. Any successful model must occupy a region that overlaps both the wild-type and *zelda*⁻ data.

As shown in Fig. A.9A and Video A.9.7, the state space corresponding to the thermodynamic MWC model fails to overlap with the *zelda*⁻ data. To more clearly reveal this disagreement, this three-dimensional state space was projected onto the xy-plane, the space incorporating the average t_{on} delay and t_{on} offset information. To do this projection, we noticed that both the wild-type and *zelda*⁻ data only occupied average initial loading rate values between 2.5 AU/min and 3.6 AU/min (Fig. A.9A and Video A.9.7). As a result, only points in that range of initial loading rates were retained for the projection. The resulting two-dimensional representation of our exploration is shown in Fig. A.9B. Even in this simplified representation, the failure of the thermodynamic MWC model (Fig. A.9B, green) is evident. Therefore, we utilized this representation throughout Chapter 2 and Appendix A (Figs. 2.5C, 2.6B, and 2.7D, and Figs. A.13 and A.14C).

A.6 Failures and assumptions of thermodynamic models of transcription

A.6.1 Generalized thermodynamic model

The generalized thermodynamic model is an extension of the thermodynamic MWC model presented in Section A.1.2. For extra generality, we assume the presence of twelve Bicoid binding sites and one RNAP binding site, but do not include the action of Zelda since the objective was to attempt to recapitulate the *zelda*⁻ mutant experimental data. We still allow for an inaccessible DNA state.

In this generalized model, the weight of each microstate can be arbitrary, rather than determined by underlying biophysical parameters. Since p_{bound} only depends on whether RNAP is bound, there is no need to distinguish between different microstates that have the same number of Bicoid molecules bound: the arbitrary coefficients allow separate microstates to effectively be combined together into the same weight. Thus, each microstate corresponds only to the overall number of bound molecules, regardless of binding site ordering. With twelve Bicoid sites, in addition to the inaccessible state, there are 27 total microstates and 26 free parameters describing the weights of each state (with the accessible, unbound

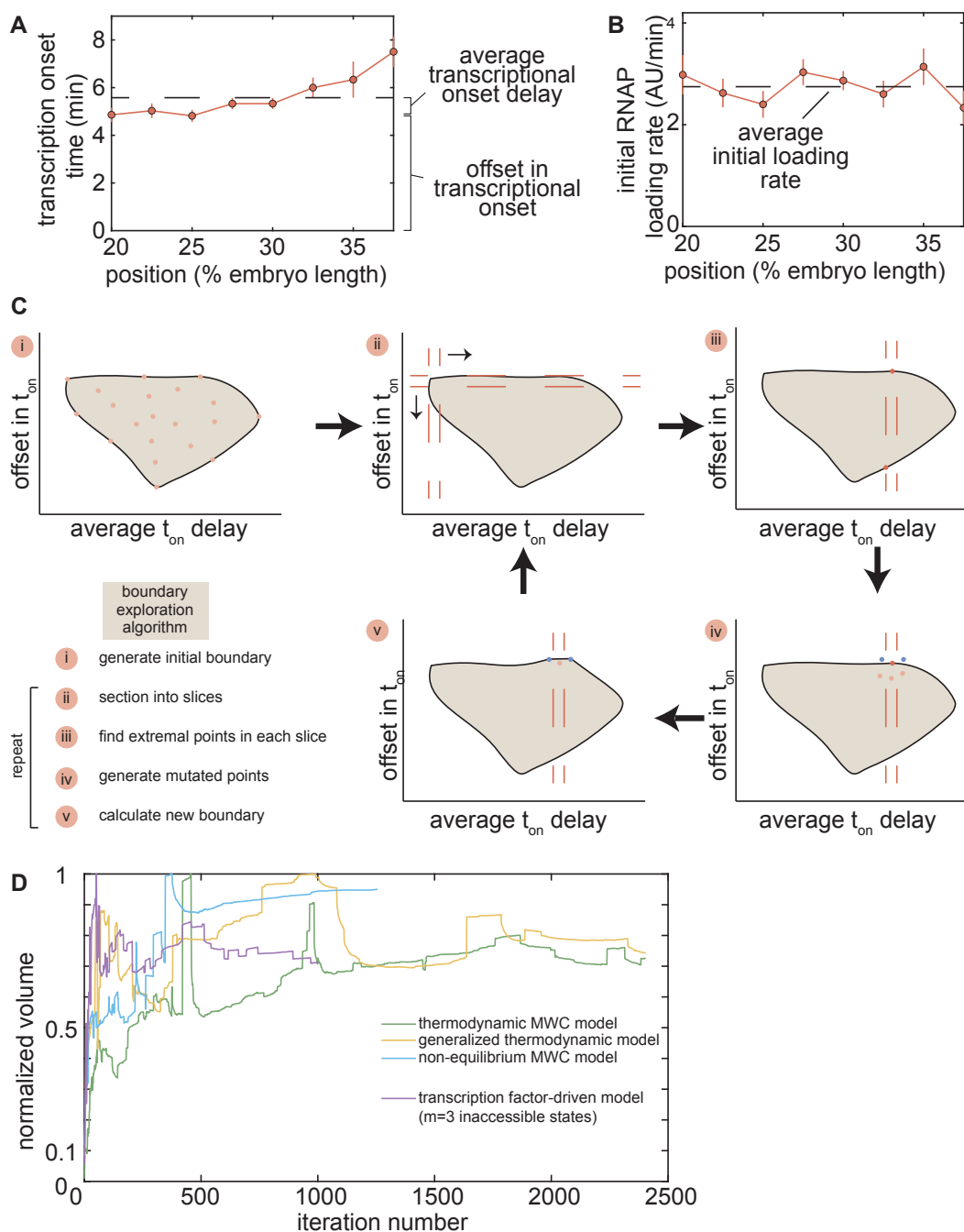


Figure A.8: Description of state-space metrics and boundary-exploration algorithm. See caption on next page.

Figure A.8: Description of state-space metrics and boundary-exploration algorithm. (A) Representative average t_{on} delay (black dashed line) and t_{on} offset for the $zelda^-$ background data in Fig. 2.4D. (B) Average initial RNAP loading rate for the $zelda^-$ background data in Fig. 2.4C. (C) Overview of the boundary-exploration algorithm for an example state space containing two dimensions. (i) A set of 50 points with random input parameters generates an initial state space of the investigated model. (ii) The space is sectioned into 10 horizontal and 10 vertical slices. (iii) The extremal points of each slice are found. (iv) For each extremal point, five new points are generated with input parameters in a small neighborhood around the parameters of this extremal point. (v) The new space is plotted with these new points, and steps (ii) - (iv) are repeated. (D) Normalized volume of state-space domain of each model investigated in this work as a function of algorithm iteration number. All volumes approach a steady value, indicating convergence.

microstate normalized to unity). Like with the thermodynamic MWC model, we assume that transcription only occurs when RNAP is bound, with the same constrained maximum rate of RNAP loading R_{max} . However, since the weights of each microstate are arbitrary, we no longer have a variable p that can be constrained by R_{min} like in Eq. A.14.

This generalized model is much more powerful than the thermodynamic MWC model due to a lack of coupling between individual microstate weights. Whereas in the previous model the underlying parameters K_b and ω_b caused similar microstates to be related mathematically, now the statistical weights for each microstate are completely independent. Physically, this scenario can arise due to, for example, higher-order cooperativities or non-identical binding energies between binding sites (Estrada et al., 2016).

The partition function in this generalized thermodynamic model is given by the polynomial

$$Z = p_{inacc} + \sum_{r=0}^1 \sum_{n=0}^{12} P_{r,n} [Bicoid]^n, \quad (\text{A.17})$$

where p_{inacc} is the weight of the inaccessible state and $P_{r,n}$ is the weight of the accessible state with r RNAP molecules bound and n Bicoid molecules bound. The overall transcriptional initiation rate is now

$$\frac{dmRNA}{dt} = \frac{1}{Z} \left(\sum_{n=0}^{12} P_{1,n} R [Bicoid]^n \right), \quad (\text{A.18})$$

where $P_{1,n}$ is the statistical weight of each RNAP-bound state and R is the corresponding rate of transcriptional initiation. Note that, as described above, R is still equal to R_{max} , the constraint described in Section A.1.3, but we no longer use the R_{min} constraint.

The resulting rate of transcriptional initiation is integrated over time to produce a simulated MS2 fluorescence trace using the same procedure as for the models presented in

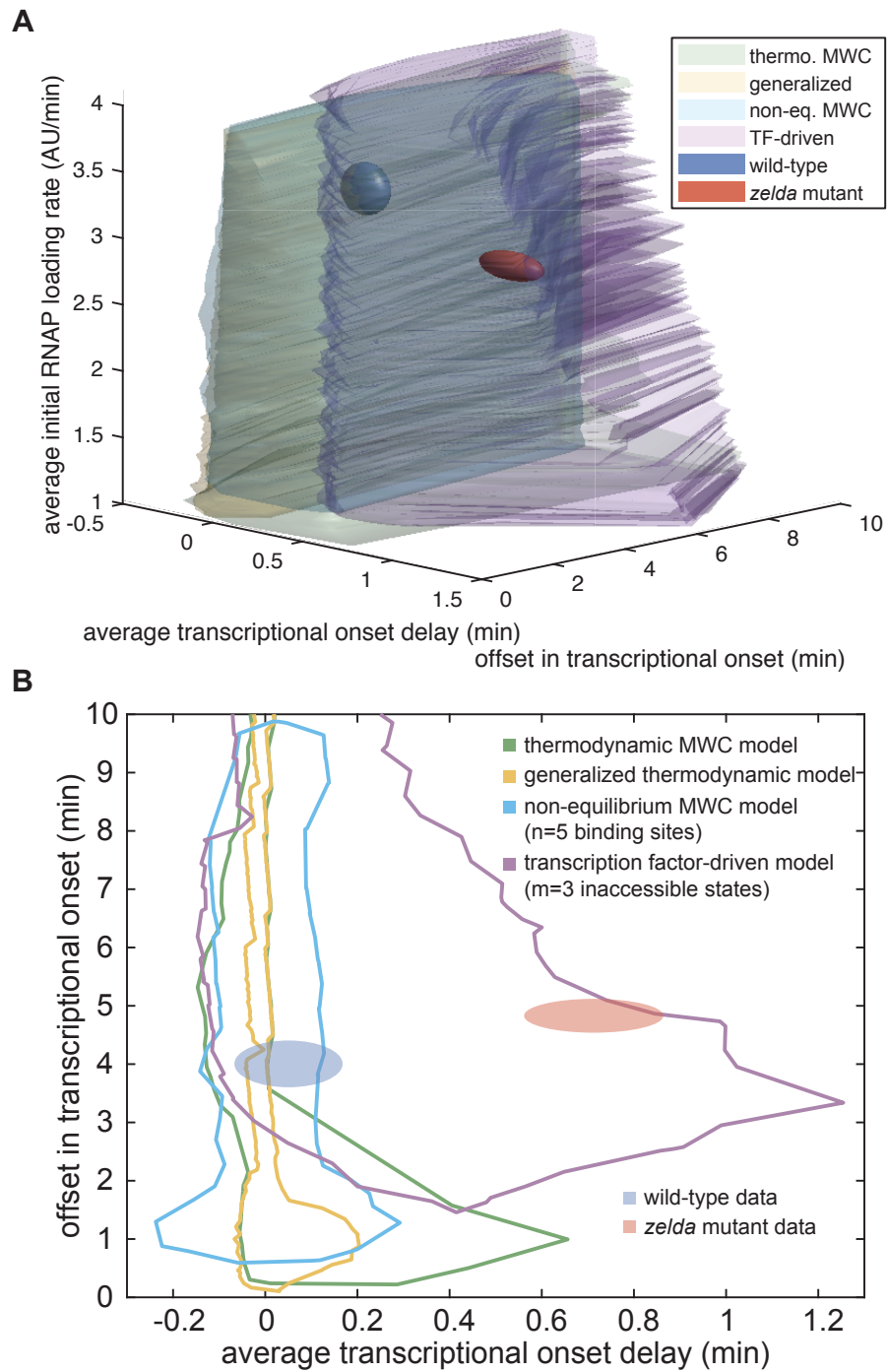


Figure A.9: Exploration of state space. See caption on next page.

Figure A.9: Exploration of state space. (A) Three-dimensional state-space exploration, showing the extents of state space of the wild-type (blue) and *zelda*⁻ (red) data as well as of various models explored in the main text. See also Video A.9.7 for an more comprehensive representation of our results. (B) Two-dimensional state-space exploration, created by projecting the three-dimensional state space in (A) for average initial loading rate values between 2.5 and 3.6 AU/min onto the xy-plane corresponding to the average t_{on} delay and t_{on} offset. Volumes (A) and areas (B) covered by the experimental data represent the standard error of the mean.

Sections A.1.2, A.7.1, and A.8.1 (see Section A.2.2 for details). As with the thermodynamic MWC model, we allow for a mitotic repression time window to account for the lack of transcription early in the nuclear cycle.

A.6.2 Generalized thermodynamic model state space exploration

Due to the high-dimensional parameter space of the generalized thermodynamic model, constraints were necessary to efficiently explore this parameter space (Section A.5.1). These constraints were placed on the values of the individual microstate weights $P_{r,n}$, based on dimensional analysis and heuristic arguments. Specifically, each weight $P_{r,n}$ is derived from a product of binding constants K_d for either Bicoid or RNAP, pairwise cooperativity parameters ω , and higher-order cooperativity terms. For the purposes of these parameter constraints, we only consider the K_d s and ω s, and ignore constraints on higher-order cooperativities. In principle, each Bicoid binding site possesses a unique K_d and protein-protein interaction terms ω with other Bicoid molecules and/or with RNAP. However, as described below, these biophysical parameters, once non-dimensionalized, can be constrained to reasonable values by scaling relations through a simple bounding scheme.

For illustrative purposes, consider the microstate with RNAP and one Bicoid molecule bound. Its weight depends on two independent binding constants p, b and a cooperativity term between RNAP and Bicoid ω_{bp} . First, we assume that the p, b terms are non-dimensionalized, i.e. they take the form $p = [RNAP]/K_p$ and $b = [Bicoid]/K_b$. Although the two individual p, b terms are in principle different since RNAP and Bicoid have can different binding energies, we can be generous about the constraints and assume that the non-dimensionalized forms are both bounded below and above by 0 and 1000, respectively. This strategy is justified by assuming that neither RNAP nor Bicoid exist in concentrations three orders of magnitude above their dissociation constants, and do not exist at negative concentrations (Estrada et al., 2016). Similarly, we can be generous about any possible cooperativities and say that ω_{bp} and ω_b have a similar bound between 0 and 1000, thus accounting for both positive and negative

cooperativities. For this state with RNAP and one Bicoid molecule bound, we can say that

$$P_{1,1} = bp\omega_{bp} \quad (\text{A.19})$$

which has bounds

$$0 < P_{1,1} < (1000)^2(1000) = 10^9 \quad (\text{A.20})$$

and thus provide a bound for the possible values that the weight $P_{1,1}$ can take.

In general, this process can be applied to enforce bounds on any microstate weight $P_{r,n}$ through constraining of the possible values of p , b , ω_{bp} , and ω_b . As a result, the weight of a microstate with more Bicoid bound (i.e. higher values of n) will have a more generous dynamic range, due to the larger powers of b and ω_b . In this way, exploration of parameter space can be made more constrained by restricting the possible values of the microstate weights $P_{r,n}$. In addition, the mitotic repression term was constrained like in the thermodynamic MWC model, where $0 < t_{MitRep} < 10$.

As a result of these constraints, the region occupied by the generalized thermodynamic model in the t_{on} offset and average t_{on} delay space does not entirely include that of the thermodynamic MWC model, whose parameters were only constrained to be positive values (Section A.1.3) Nevertheless, this model still fails to capture the delays observed in the *zelda*⁻ data (Fig. A.9B, yellow).

A.6.3 Extended generalized thermodynamic model with transcription factor binding in the inaccessible state

The generalized thermodynamic model (Section A.6.1) encompasses all possible thermodynamic models with up to twelve Bicoid binding sites that can be bound in the accessible state. However, a potentially more general class of models involves those where Bicoid can also bind to the inaccessible state. For example, Bicoid action could conceivably result in some pioneering activity by directly binding to chromatin in the inaccessible state and facilitating RNAP binding and transcription. Here, we show that these models can be reformulated into the generalized thermodynamic model presented above.

If we allow for Bicoid to bind to any of the twelve binding sites in the inaccessible states, then we introduce l new microstates with individual Boltzmann weights P_l , one for each Bicoid-bound inaccessible state, in addition to the unbound inaccessible state with weight P_{inacc} . Nevertheless, as long as the ensuing transcription rate of each Bicoid-bound inaccessible state is zero, then the net effect of these additional inaccessible states could simply be described by a single effective inaccessible state with Boltzmann weight $P'_{inacc} = P_{inacc} + \sum_l P_l$. The resulting state space exploration (Section 2.2.4 and Fig. 2.5C, yellow), which explores the whole parameter space of reasonable values of P_{inacc} , would thus also capture the behavior of this single effective inaccessible state. As a result, models that consider the binding of Bicoid to the inaccessible states are contained within our generalized thermodynamics model.

A.6.4 Investigation of the failure of thermodynamic models

Here, we provide an intuitive explanation for why thermodynamic models fail to recapitulate the delay in t_{on} for $zelda^-$ embryos. The combination of the occupancy hypothesis and the assumption of separation of times scales described in Section A.6.5 imply that the rate of transcriptional initiation at any moment in time is an *instantaneous* readout of the Bicoid concentration at that time point. Thus, any thermodynamic model is *memoryless*. Intuitively, this means that a thermodynamic model requires transcription to begin as soon as the Bicoid concentration crosses a certain “threshold” since time delays between input and output require some sense of memory. Examination of the dynamic measurements of MS2 output in $zelda^-$ embryos reveals that no matter what “threshold” concentration of Bicoid is assigned for the start of transcription, the model cannot simultaneously describe two values of t_{on} corresponding to different positions along the anterior-posterior axis (Fig. A.10A and B).

Another self-consistency check of a thermodynamic model is to examine the concentration of Bicoid at t_{on} for various positions along the embryo. Due to the memoryless nature of thermal equilibrium, a valid thermodynamic model predicts that, at different positions along the embryo, t_{on} will occur when Bicoid reaches the same threshold value. For the $zelda^-$ data, however, the level of Bicoid at each anterior-posterior position’s t_{on} value actually *decreases* with increasing t_{on} , suggesting the failure of the thermodynamic model (Fig. A.10C). Thus, the strong position-dependent delay in t_{on} for the $zelda^-$ data cannot be explained by an instantaneous Bicoid readout mechanism.

More generally, the memoryless nature of thermodynamic models implies that, given any input-output function that increases monotonically with Bicoid and Zelda concentration, the ensuing onset time of transcription cannot be later than the time at which Bicoid or Zelda reach their maximal values. This is a reflection of a generic feature of thermodynamic models, namely that only instantaneous couplings in time can exist, and that time delays are impossible (Coulon et al., 2013; Wong and Gunawardena, 2020). By inspecting the nuclear concentrations of Bicoid and Zelda in Fig. A.3, we notice that times of maximal nuclear concentration for both transcription factors all occur around 4.5 min. This time is much earlier than the delayed transcriptional onsets exhibited in the $zelda^-$ data (Fig. 2.4D, red points), providing further evidence for the unsuitability of thermodynamic models in describing the observed delay in the transcriptional onset time along the anterior-posterior axis of the embryo.

A.6.5 Re-examining thermodynamic models of transcriptional regulation

Thermodynamic models based on equilibrium statistical mechanics can be seen as limiting cases of more general kinetic models. For example, consider simple activation, where an activator whose concentration is modulated in time regulates transcription by binding to a single site (Fig. A.11). In this generic model, the presence of activator can modulate the rates of activator and RNAP binding and unbinding through the parameters α , β , γ , and δ .

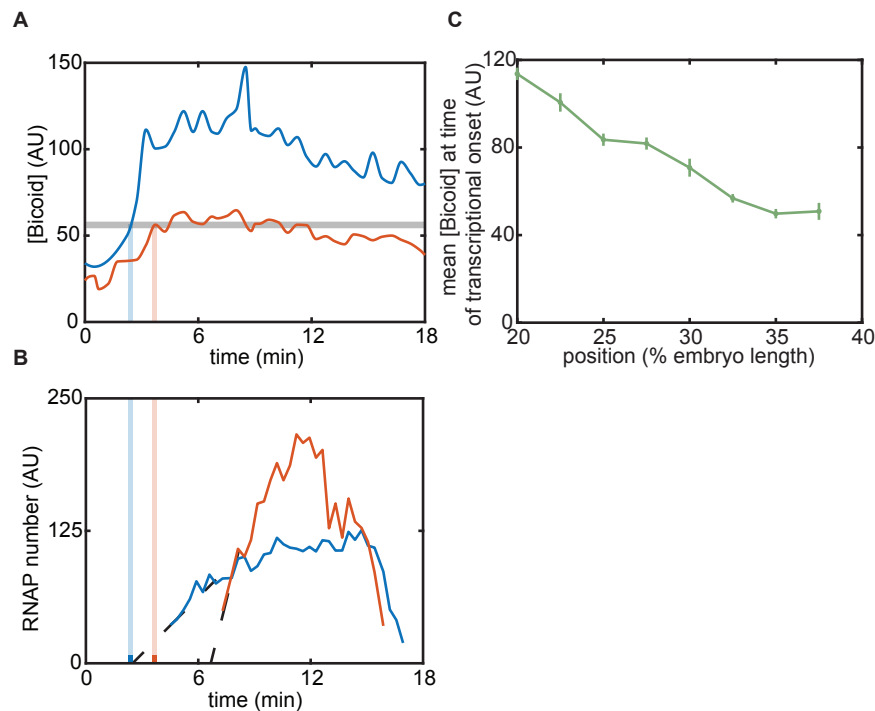


Figure A.10: Intuition for failure of equilibrium models. (A) Mean Bicoid concentrations for two positions along the embryo (blue, red), with a “threshold” chosen to attempt to match the corresponding t_{on} in (B). (B) MS2 fluorescence signal for the two positions shown in (A) for the $zelda^-$ experiment. Note that no single threshold value of Bicoid can match the timings in (A) with the transcriptional onset times in (B). (C) Mean Bicoid concentration at t_{on} as a function of position for the $zelda^-$ data.

In order to reduce kinetic models to thermodynamic models where the probabilities of each state are dictated by Boltzmann weights such as those in Fig. 2.2A, four conditions must be fulfilled. First, the rate of mRNA production must be linearly related to the probability of finding RNAP bound to the promoter (Fig. A.11i). This *occupancy hypothesis* is necessary for Eq. A.2 to hold. Second, the time scales of binding and unbinding of RNAP and transcription factors must be much faster than the time scales of the concentration dynamics of these proteins (Fig. A.11ii). Third, these time scales must also be much faster than the rate of transcriptional initiation and mRNA production (Fig. A.11iii). Under these conditions of *separation of time scales*, the binding and unbinding of proteins quickly reaches steady state while the overall concentrations of these molecular players are modulated (Segel and Slemrod, 1989). Fourth, there must be no energy input into the system (Fig. A.11iv). This condition demands “detailed balance” (Vilar and Leibler, 2003; Ahsendorf et al., 2014; Hill, 1985): the product of state transition rates in the clockwise direction over a closed loop is equal to the

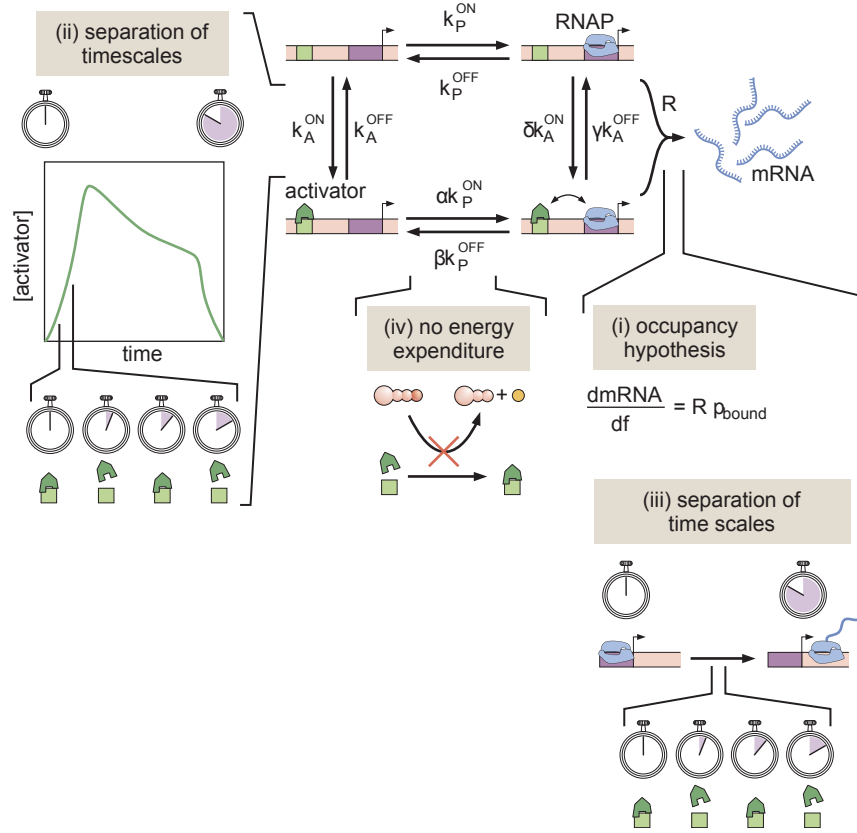


Figure A.11: A simple kinetic model of transcriptional activation in which activator molecules influence RNAP binding kinetics. The assumptions that make it possible to turn this kinetic model into a thermodynamic one are (i) the occupancy hypothesis, (ii, iii) a separation of time scales between binding and unbinding rates, and activator and mRNA production dynamics, respectively, and (iv) no energy expenditure (detailed balance).

product going in the counterclockwise direction, a constraint known as the *cycle condition* (Estrada et al., 2016). In the case of Fig. A.11, this requirement implies that

$$k_P^{ON} \delta k_A^{ON} \beta k_P^{OFF} k_A^{OFF} = k_P^{OFF} k_A^{ON} \alpha k_P^{ON} \gamma k_A^{OFF}. \quad (\text{A.21})$$

If these four conditions are met, then the system is effectively in equilibrium and the various binding states adopt probabilities that can be calculated using equilibrium statistical mechanics.

A.7 Non-equilibrium MWC model

A.7.1 Non-equilibrium MWC model

The non-equilibrium MWC model is an extension of the thermodynamic MWC model presented in Section A.1.2, where we now relax the assumption of separation of time scales (Fig. A.11 ii and iii) and make it possible to assume, for example, that the system responds instantaneously to changes in activator concentration. Here, we explicitly simulate the full system of ordinary differential equations (ODEs) that describe the dynamics of the system out of steady state. Additionally, we allow for energy to be expended and thus do not enforce detailed balance through the cycle condition (Fig. A.11iv). We still employ a mitotic repression window term, before which no transcription is allowed.

We consider a generic model with n Bicoid binding sites, and again ignore *Zelda* since we are only interested in recapitulating the *zelda*⁻ mutant data. As a result, this new model has $n + 1$ total binding sites which, together with the closed chromatin state, results in a total of $2^{n+1} + 1 = N$ microstates. In the case of six Bicoid binding sites, this results in $N = 129$ total microstates. We assign each microstate x_i a label i and describe the transition rate from state j to state i using k_{ij} , where i, j range from 0 to $N - 1$, inclusive.

In matrix notation, we write the system of ODEs as

$$\frac{d\vec{X}}{dt} = K\vec{X}, \quad (\text{A.22})$$

where \vec{X} is a vector containing the fractional occupancy of each microstate x_i and K is a matrix containing all the transition rates k_{ij} . Normalizing such that the sum of all the components in the vector \vec{X} is unity, we now have a vector representing the instantaneous probability of being in each microstate.

To relate the occupancies of the different states to the rate of transcriptional initiation, we retain the occupancy hypothesis presented earlier: that p_{bound} , the probability of being in a microstate with a bound RNAP molecule, is linearly related to the overall average transcriptional initiation rate that we determine from experimentally measurements.

For this particular system, it is helpful to define an intuitive microstate labeling system. Because the relevant physical processes are the binding and unbinding of Bicoid and RNAP molecules, we can represent any microstate in binary form, where the total number of digits is the total number of binding sites $n + 1$, and each digit represents an individual binding site. Our convention is to assign the first digit to the promoter, and the subsequent ones to the Bicoid sites. By assigning 0 to an unbound site and 1 to a bound site, we can rewrite each unique microstate's label i in binary form. For example, for a model with six Bicoid sites, the label for the microstate with no RNAP bound and the first two Bicoid sites occupied is represented with

$$i = \text{bin}(0110000) = 48. \quad (\text{A.23})$$

Here, $\text{bin}()$ indicates taking the base 2 value of the binary label in the parentheses. The closed chromatin state is added manually and assigned to the last position in our binary label, x_{N-1} .

This convention allows us to intuitively define each unique label for the system’s microstates and provides a way to map the physical contents of a microstate with its associated label i .

In general, the overall transition matrix K can be very complex. However, we benefit from the fact that the only non-zero transitions k_{ij} are the ones that correspond to physical processes: modifying the open/closed chromatin state, and binding and unbinding of Bicoid or RNAP molecules. In this binary notation, these constraints imply that the only nonzero transitions are the ones that represent individual flips between 0 and 1, as well as between the open and closed states 0 and $N - 1$. The transition matrix K is then easier to write, since it is clear from the binary representation which transitions must be nonzero. Finally, diagonal elements k_{ii} are entirely constrained because they represent probability loss from a particular state i , and must be equal to the negative of the rest of the column i , such that the sum over each column in K is zero.

Given that the Bicoid concentration changes as a function of time and that we assume first-order binding kinetics, whichever rates k_{ij} correspond to Bicoid binding rates must be multiplied by this time-dependent nuclear concentration. In contrast, all off-rates are independent of Bicoid concentration. To keep subsequent parameter exploration simple, we non-dimensionalized the Bicoid concentration by rescaling it by its approximate scale. This was achieved by dividing all Bicoid concentrations by the average Bicoid concentration, calculated by averaging the mean Bicoid nuclear fluorescence across all datasets, anterior-posterior positions, and time points, yielding approximately 35 arbitrary fluorescence units. Thus, all of the transition rates k_{ij} in the model here are expressed in units of inverse minutes.

To model transcription specifically, we assumed that at the beginning of the nuclear cycle, the system is in the closed chromatin state: $x_i(t = 0) = 0$ except for the closed chromatin state $x_{N-1}(t = 0) = 1$. We simulated the full trajectory of all the microstates x_i over time by solving the system of ODEs given in Eq. A.22. Finally, we calculated p_{bound} by summing the x_i ’s that correspond to RNAP-bound states, and then computed the subsequent transcriptional initiation rate by multiplying p_{bound} with the transcription rate R . Here, R is the same R_{max} as in Sections A.1.2 and A.6.1 but again we do not constrain the model using R_{min} , just as in Section A.6.1.

Fig. A.12A shows an example of this model for a system with only one Bicoid binding site and no closed chromatin state, for simplicity, resulting in a four-state network. The binary indexing labels (shown beneath each state in light pink) can be converted into the base-10 labels (light teal) ranging from 0 to 3. The connection matrix for this system is

$$C = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \quad (\text{A.24})$$

and the corresponding transition rate matrix K is

$$K = \begin{bmatrix} k_{00} & k_{01} & k_{02} & 0 \\ k_{10} & k_{11} & 0 & k_{13} \\ k_{20} & 0 & k_{22} & k_{23} \\ 0 & k_{31} & k_{32} & k_{33} \end{bmatrix}, \quad (\text{A.25})$$

where, in this example, k_{02} represents the transition rate from state j to state i . The diagonal elements k_{ii} are equal to the negative of the sum of the elements in the rest of the column in order to preserve conservation of probability. For example, $k_{00} = -(k_{10} + k_{20} + k_{30})$.

With all this information in hand, we solve for the occupancy of each of the four states using the matrix ODE

$$\begin{bmatrix} \frac{dx_0}{dt} \\ \frac{dx_1}{dt} \\ \frac{dx_2}{dt} \\ \frac{dx_3}{dt} \end{bmatrix} = \begin{bmatrix} k_{00} & k_{01} & k_{02} & 0 \\ k_{10} & k_{11} & 0 & k_{13} \\ k_{20} & 0 & k_{22} & k_{23} \\ 0 & k_{31} & k_{32} & k_{33} \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}. \quad (\text{A.26})$$

In this case, the occupancy hypothesis relates p_{bound} to the overall transcription rate, resulting in

$$\frac{dmRNA}{dt} = Rp_{bound} = R \frac{x_1 + x_3}{x_0 + x_1 + x_2 + x_3}. \quad (\text{A.27})$$

This model can produce time-dependent behavior not found in the thermodynamic models. Fig. A.12B contains an example of a hypothetical input Bicoid activator concentration that switches instantaneously from zero to a finite value. In the thermodynamic models, the predicted transcriptional initiation rate also responds instantaneously (Fig. A.12B, top). In contrast, for a suitable set of parameters, the non-equilibrium MWC model predicts a slow response over time (Fig. A.12B, bottom).

To produce a simulated MS2 fluorescence trace, the resulting rate of mRNA production is integrated over time using the same procedure (Section A.2.2) as the models presented in Sections A.1.2, A.6.1, and A.8.1. As with the thermodynamic MWC model, we allow for a time window of mitotic repression to account for the lack of transcription early in the nuclear cycle. Specifically, this was implemented by allowing the system to evolve over time, but fixing transcription to zero ($R = 0$) until after the mitotic repression time t_{MitRep} . An alternative formulation of the model, in which the whole system is frozen such that no transitions between states are allowed until after t_{MitRep} , is discussed below in Section A.7.3.

A.7.2 Non-equilibrium MWC model state space exploration

In the parameter exploration of this model (Section A.5.1), the transition rates k_{ij} were constrained with minimum and maximum values of $k_{min} = 1$ and $k_{max} = 10^5$ respectively, in units of inverse minutes. These bounds were conservatively chosen using the following estimates. First, we estimate the values of the possible unbinding rates k_{off} . We assume that RNAP and Bicoid obey the same unbinding kinetics. Estimates of *in vivo* single-molecule

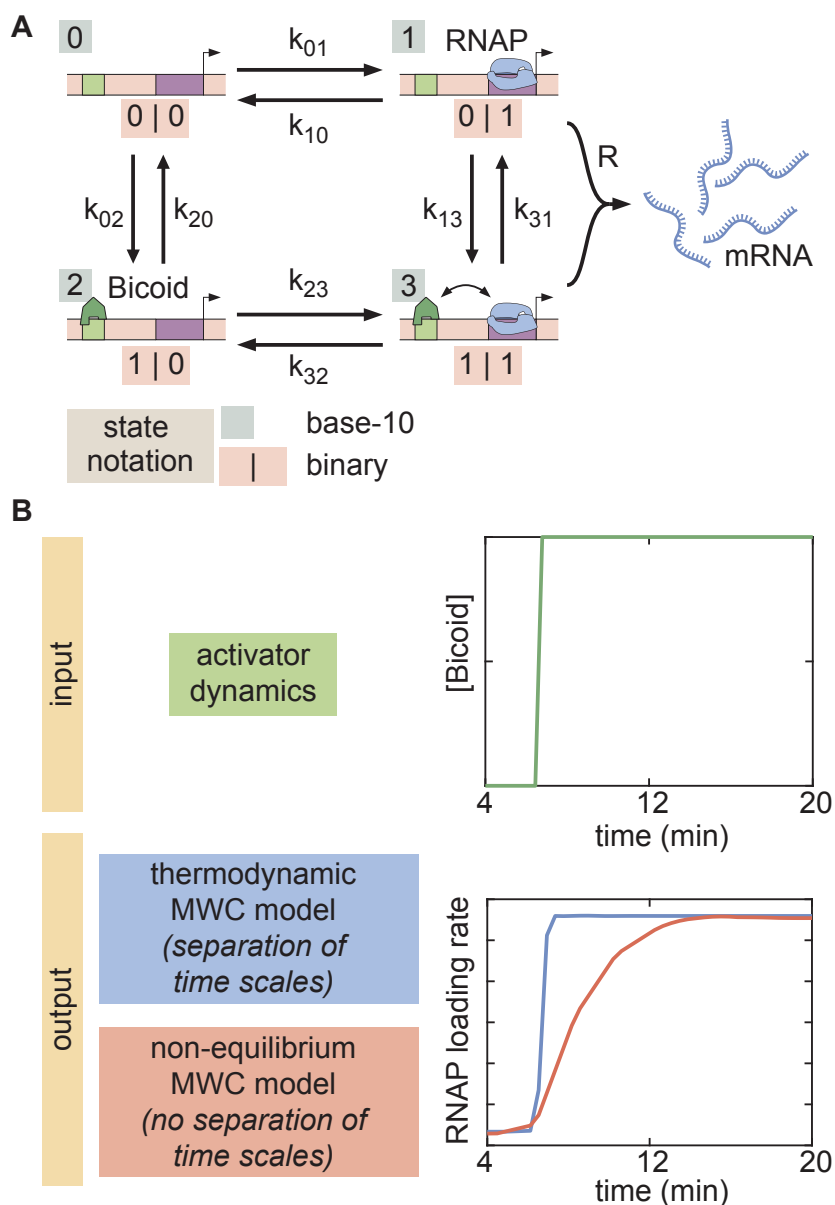


Figure A.12: Example of a four-state time-dependent model with one Bicoid binding site and no closed chromatin state. (A) The binary label for each state (light pink) can be converted into a base-10 label for each state (light teal). The transition rates k_{ij} are defined as the transition rate from state i into state j using this labeling system. (B) For an example input activator concentration temporal profile that is a step function, the time-dependent response is compared for the cases of separation of time scales and lack thereof. In the former, the transcriptional initiation rate responds instantaneously to the increase in activator input, while the response is slower in the latter.

binding kinetics inferred from Mir et al. (2018) indicate that the lifetime of Bicoid on DNA is on the order of 3 s^{-1} . Second, we estimate the values of the possible on-rates k_{on} using the classic Berg-Purcell equation for the case of a diffusion-limited binding to a perfectly absorbing spherical receptor (Berg and Purcell, 1977). In this case, the on-rate of molecule binding is given by

$$k_{on} = 4\pi D a c_0, \quad (\text{A.28})$$

where D is the diffusion coefficient of the molecule, a is the estimated size of the spherical receptor, and c_0 is the background concentration of the molecular species. Since here we are talking about transcription factor binding to a Bicoid binding site, we assume a to be on the order of 5 nm. We assume that RNAP and Bicoid obey the same diffusion characteristics, leading to a diffusion coefficient of approximately $0.3 \mu\text{m}^2\text{s}^{-1}$ (Gregor et al., 2007b). Finally, Bicoid is present at concentrations between 10 nM and 55 nM in the nucleus (Gregor et al., 2007a), and we assume that nuclear RNAP concentrations exist within the same range. Plugging these values into Eq. A.28 yields estimates for the maximum and minimum on-rates:

$$\begin{aligned} k_{on}^{max} &\sim (4\pi)(0.3 \mu\text{m}^2\text{s}^{-1})(1 \mu\text{m})(55 \text{ nM}) \\ &\sim 0.5 \text{ s}^{-1} \sim 30 \text{ min}^{-1}. \end{aligned}$$

and

$$\begin{aligned} k_{on}^{min} &\sim (4\pi)(0.3 \mu\text{m}^2\text{s}^{-1})(1 \mu\text{m})(10 \text{ nM}) \\ &\sim 0.05 \text{ s}^{-1} \sim 3 \text{ min}^{-1}. \end{aligned}$$

Thus, our maximum and minimum transition rate bounds of $k_{min} = 1 \text{ min}^{-1}$ and $k_{max} = 10^5 \text{ min}^{-1}$ lie outside these estimated binding and unbinding rates. The mitotic repression term was constrained like in the thermodynamic MWC model, where $0 < t_{MitRep} < 10$.

One caveat of the state-space exploration approach is that the high dimensionality of the non-equilibrium MWC model prevented us from calculating the full state-space boundary using six Bicoid binding sites. Due to computational costs, we were only able to accurately produce a state-space boundary for this model (Section A.7.1) using five Bicoid binding sites. Running the exploration for a model with six Bicoid binding sites took over two weeks on our own server, and the algorithm had not noticeably converged in the end.

The results of the state space exploration for the non-equilibrium MWC model using five Bicoid binding sites resulted in larger average t_{on} delays than the thermodynamic models (Sections A.1.2 and A.6.1). However, this model, like those, failed to reproduce the delays observed in the *zelda*⁻ data (Fig. A.9B, cyan).

Interestingly, the total areas covered by each non-equilibrium MWC model did not monotonically increase with Bicoid binding site number (Fig. 2.6B). This phenomenon where the state space of a model does not strictly increase with binding site number has been previously observed (Estrada et al., 2016) and the reason for this effect remains uncertain.

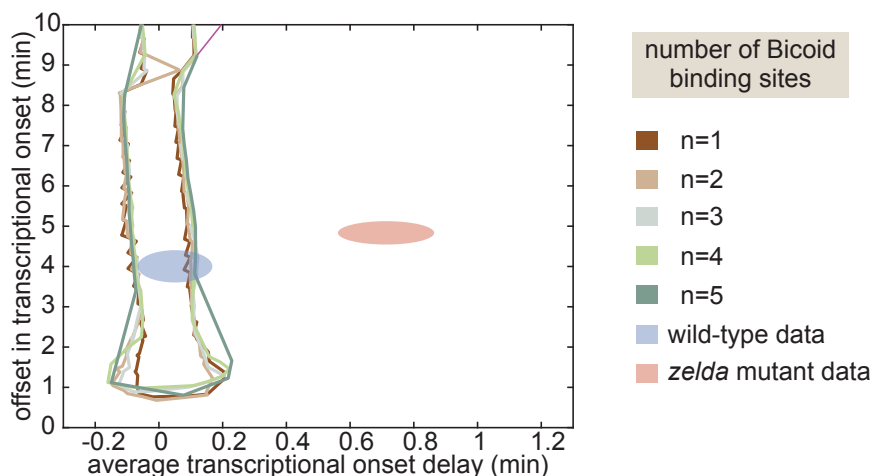


Figure A.13: State space exploration for non-equilibrium MWC model with strong mitotic repression for up to five Bicoid binding sites.

A.7.3 Alternative non-equilibrium MWC model with strong mitotic repression

In the main text, we entertained a non-equilibrium MWC model where mitotic repression blocks any productive transcription ($R = 0$) until the mitotic repression window t_{MitRep} has passed. Before this time, in this model, the system can nevertheless transition through its different states over time.

In an alternate formulation of this non-equilibrium MWC model, we consider a form of mitotic repression that we call strong mitotic repression. Here, the system itself is frozen in the initial inaccessible state and not allowed to evolve until after t_{MitRep} . After t_{MitRep} , the system evolves through time according to the same rules as the original non-equilibrium MWC model.

Repeating the state space exploration for this model, for up to five Bicoid binding sites, yielded similar conclusions. Namely, the model could not describe the average delay and offset in transcriptional onset time in the absence of Zelda (Fig. A.13). The intuition behind this is that, while this stronger form of mitotic repression could potentially achieve longer delays, the crucial feature of the *zelda*⁻ data is not merely a delayed transcription onset time, but a *position-dependent* delay that increases towards the posterior of the embryo. This stronger form of mitotic repression does not result in a mechanism capable of achieving such delay. In contrast, the final transcription factor-driven model (Section A.8.1) does provide such a mechanism by coupling the inaccessible-to-accessible transition to the position-dependent Bicoid gradient.

A.8 Transcription factor-driven model of chromatin accessibility

A.8.1 Transcription factor-driven model of chromatin accessibility

The transcription factor-driven model of chromatin accessibility is a slight modification of the thermodynamic MWC model (Section A.1.2) that replaces the MWC mechanism of chromatin transitions with a direct driving action due to Bicoid and Zelda. Here, we retain the idea of inaccessible vs. accessible states, but no longer demand that these states be in thermodynamic equilibrium. Instead, the system begins in the inaccessible state and undergoes a series of m identical, slow, and effectively irreversible transitions to the accessible state. Once these transitions into the accessible state occur, the system can rapidly and reversibly occupy all of its accessible microstates such that the probability of the system being in any of these microstates is described by thermodynamic equilibrium. The accessible states are governed by the same rules and parameters as the thermodynamic MWC model (Section A.1.2), albeit without the $\Delta\varepsilon_{chrom}$ parameter since now the transition from the inaccessible to accessible state is unidirectional.

We consider two possible contributions for these irreversible transitions: a Bicoid-dependent pathway and a Zelda-dependent pathway (Fig. A.15A, see Section A.8.2 for a discussion on this choice of parameterization). We assume the transition rates to be first-order in Bicoid and Zelda, respectively, such that

$$\pi_b = c_b[Bicoid] \quad (\text{A.29})$$

and

$$\pi_z = c_z[Zelda]. \quad (\text{A.30})$$

Here, π_b is the Bicoid-dependent contribution to the transition rates and π_z is the corresponding Zelda-dependent contribution. There are two input parameters c_b and c_z that give the relative speed of each transition rate contribution. The overall rate π of each irreversible transition is given by the sum

$$\pi = \pi_b + \pi_z = c_b[Bicoid] + c_z[Zelda]. \quad (\text{A.31})$$

Because the accessible states are in thermodynamic equilibrium with each other, we can effectively treat them as a single state and describe the entire system with $m + 1$ states, corresponding to the inaccessible, intermediate, and accessible states. We label the inaccessible state with 0, the $m - 1$ intermediate states with 1 through $m - 1$, and the final accessible state with m . Thus, we describe the probability p_i of the system being in the state i with the probability vector \vec{P}

$$\vec{P} = \begin{bmatrix} p_0 \\ p_1 \\ \dots \\ p_m \end{bmatrix}. \quad (\text{A.32})$$

Calculating the overall RNAP loading rate then simply corresponds to rescaling p_{bound} with the overall probability $p_m(t)$ of being in the accessible state:

$$\frac{dmRNA}{dt} = R p_{bound} p_m, \quad (\text{A.33})$$

where R is the same maximum rate used in Section A.1.2. Note that $p_m(t)$ is a time-dependent quantity that changes over time. To calculate $p_m(t)$, we solve the corresponding system of ODEs that describes the time evolution of \vec{P}

$$\frac{d\vec{P}}{dt} = \Pi \vec{P}, \quad (\text{A.34})$$

where Π is the transition rate matrix describing the time evolution of the system. Π , by definition, is a square matrix with dimension $m + 1$. Given the initial condition that the system begins in the inaccessible state

$$\vec{P} = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix} \quad (\text{A.35})$$

the system of ODEs can be solved to find the probability of being in the accessible state $p_m(t)$. For example, for $m = 3$ irreversible steps, Π takes the form

$$\Pi = \begin{bmatrix} -\pi & 0 & 0 & 0 \\ \pi & -\pi & 0 & 0 \\ 0 & \pi & -\pi & 0 \\ 0 & 0 & \pi & 0 \end{bmatrix}, \quad (\text{A.36})$$

where π is given by Eq. A.31.

For simplicity, the time evolution of \vec{P} was solved using MATLAB's `ode15s` solver.

With the probability $p_m(t)$ of the system being in the accessible state calculated, we now calculate the probability p_{bound} of RNAP bound to the promoter in the accessible states, which lie in thermodynamic equilibrium with each other. Because we now only have accessible states, the partition function is

$$Z = (1 + z)^{10} \left(1 + p + \sum_{j=0,1} \sum_{i=1}^6 \binom{6}{i} b^i \omega_b^{i-1} p^j \omega_{bp}^{ij} \right), \quad (\text{A.37})$$

where z , p , and b correspond to the non-dimensionalized concentrations of Zelda, RNAP, and Bicoid, respectively, and ω_b and ω_{bp} are the cooperativities between Bicoid molecules and between Bicoid and RNAP, respectively. Thus, the overall transcriptional initiation rate is

given by

$$\begin{aligned} \text{Rate} &= \frac{R}{Z} \left((1+z)^{10} p \left(1 + \sum_{i=1}^6 \binom{6}{i} b^i \omega_b^{i-1} \omega_{bp}^i \right) \right) p_m \\ &= R \frac{\left(p \left(1 + \sum_{i=1}^6 \binom{6}{i} b^i \omega_b^{i-1} \omega_{bp}^i \right) \right)}{\left(1 + p + \sum_{j=0,1} \sum_{i=1}^6 \binom{6}{i} b^i \omega_b^{i-1} p^j \omega_{bp}^{ij} \right)} p_m. \end{aligned} \quad (\text{A.38})$$

Due to the lack of the inaccessible state in the partition function and because we assume that Zelda does not directly interact with Bicoid or RNAP, now the presence of Zelda mathematically separates out so that only Bicoid influences transcription. The calculation above is a standard equilibrium statistical mechanical calculation, except that we have weighted the final result with $p_m(t)$, the probability of being in the accessible states. The resulting rate is integrated to produce a simulated MS2 fluorescence trace using the same procedure (Section A.2.2) as the models presented in Sections A.1.2, A.6.1, and A.7.1.

Interestingly, we found that a mitotic repression term was not necessary to recapitulate the data, since the presence of intermediary states produced the necessary delay to explain the experimentally observed t_{on} values in the data (Fig. 2.4D, points).

In order to sufficiently explain the data, we found that a minimum of $m = 3$ irreversible steps was necessary. Fig. A.14A and B show the results of fitting this model to the observed rates of RNAP loading and t_{on} for the wild-type and $zelda^-$ data, for increasing values of m (wild-type results not shown, since all values of m easily explained the wild-type data). We see that while lower values of m do a poor job of recapitulating the data, once we reach $m = 3$ the model sufficiently predicts the experimental data within experimental error. For values of m higher than 3, explanatory power increases marginally. Considering the parameter exploration of this model (Section A.8.3) highlights the necessity of having at least $m = 3$ steps.

A.8.2 Exploring alternatives to the additive transcription factor-driven transition rate

In Section A.8.1, we defined the transition rate between the transcriptionally silent states in our transcription factor-driven model of chromatin accessibility as

$$\pi = c_b[Bicoid] + c_z[Zelda]. \quad (\text{A.39})$$

Here, we assumed that Zelda and Bicoid operate independently and in parallel to catalyze the transitions from the inaccessible to accessible state (Fig. A.15A). Our choice in using two independent Zelda- and Bicoid-mediated transitions was primarily motivated by the fact that, to our knowledge, no direct interactions between Bicoid and Zelda have been reported to date. However, this is not the only possible choice of model formulation. Here, we discuss

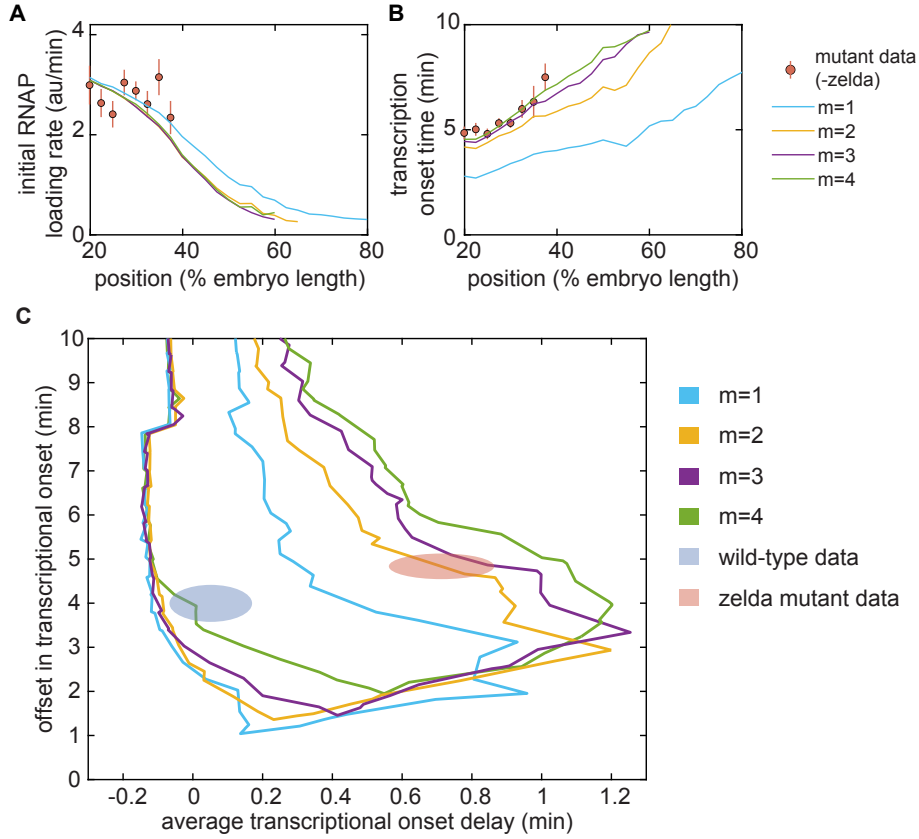


Figure A.14: Testing the transcription factor-driven model of chromatin accessibility. (A,B) Best-fit results of the transcription factor-driven model to the mutant *zelda*⁻ data. (A) initial RNAP loading rates, and (B) t_{on} , for varying numbers m of transcriptionally silent states. (C) Parameter exploration in average t_{on} delay and t_{on} offset state space for increasing values of m .

and rule out two alternative mechanisms of Zelda- and Bicoid-mediated transitions from the inaccessible to accessible state.

As a first alternative, instead of an independent and additive mechanism, we could imagine a scenario where Bicoid and Zelda act simultaneously (Fig. A.15B). Here, each stochastic transition is given by

$$\pi = c[Bicoid][Zelda] \quad (\text{A.40})$$

where c is some constant with units of $[Bicoid]^{-1}[Zelda]^{-1}min^{-1}$.

In a second alternative, Bicoid and Zelda could act sequentially. Here, each stochastic transition contains an intermediary state (Fig. A.15C). In this case, the transition rate will

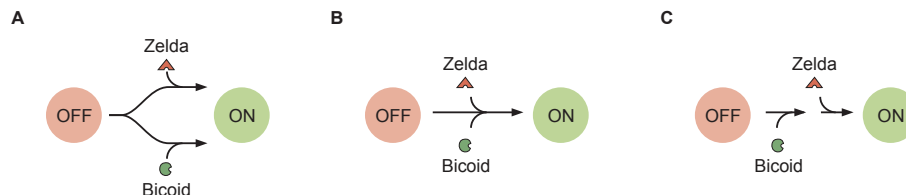


Figure A.15: Different potential schemes of Bicoid- and Zelda-mediated transition into the accessible state, for a model with $m = 1$ transcriptionally silent state. (A) The model used in the main text, where Bicoid and Zelda provide independent pathways for chromatin to transition into the accessible state. (B) A scheme where Bicoid and Zelda act simultaneously on the transition. (C) A scheme where Bicoid acts first, and then Zelda, on the same pathway.

be dependent on Bicoid and Zelda such that

$$\pi \sim \frac{c_1[Bicoid]c_2[Zelda]}{c_2[Zelda] + c_1[Bicoid]}, \quad (\text{A.41})$$

where c_1 and c_2 are constants with units of $[Bicoid]^{-1}min^{-1}$ and $[Zelda]^{-1}min^{-1}$.

One critical experimental observation is that transcription occurs even in the absence of Zelda, albeit at a delayed capacity. Since removing Zelda would set π to zero in these alternative models, transcription would not occur at all, and so both of the proposed alternative mechanisms can be ruled out. More generally, the existence of transcription in the absence of Zelda requires that there must exist some independence between Bicoid-and-Zelda-mediated transitions from the OFF to the ON state. Otherwise, no transition, and hence no transcription, could occur in the absence of Zelda.

A.8.3 Transcription factor-driven model of chromatin accessibility state space exploration

In the parameter exploration of this model (Section A.5.1), the parameters were constrained as

- $c_b > 0$
- $c_z > 0$.

The parameters shared with the thermodynamic MWC model retained the constraints described in Section A.1.3.

Fig. A.14C shows the state space explorations (see Section A.5.1) of this transcription factor-driven model for increasing numbers of intermediate steps m . Not until $m = 3$ does the model explain the both the wild-type and $zelda^-$ data, indicating that $m = 3$ is the

minimum number of irreversible steps necessary. In the state space exploration shown in Fig. 2.7D and Fig. A.9, the number of irreversible steps was fixed at $m = 3$.

Unlike the other models investigated (Sections A.1.2, A.6.1, and A.7.1), the transcription factor-driven model of chromatin accessibility occupied a region in state space that encompassed both the wild-type and *zelda*⁻ data (Fig. A.9, purple).

A.9 Supplementary Videos

- A.9.1. **Video 1.** Measurement of eGFP-Bicoid. Movie of eGFP-Bicoid fusion in an embryo in nuclear cycle 13. Time is defined with respect to the previous anaphase. (https://www.dropbox.com/s/a1073e3l468r5oa/BcdGFP_5-23-17.avi?dl=0)
- A.9.2. **Video 2.** Measurement of Zelda-sfGFP. Movie of Zelda-sfGFP fusion in an embryo in nuclear cycle 13. Time is defined with respect to the previous anaphase. (https://www.dropbox.com/s/wuhsb5crl1pb8vn/ZldGFP_2-10-17.avi?dl=0)
- A.9.3. **Video 3.** Measurement of MS2 fluorescence in a wild-type background. Movie of MS2 fluorescent spots in a wild-type background embryo in nuclear cycle 13. Time is defined with respect to the previous anaphase. (<https://www.dropbox.com/s/frcd4ywu4ki49ih/WTMS2.avi?dl=0>)
- A.9.4. **Video 4.** Measurement of MS2 fluorescence in a *zelda*⁻ background. Movie of MS2 fluorescent spots in a *zelda*⁻ background embryo in nuclear cycle 13. Time is defined with respect to the previous anaphase. (<https://www.dropbox.com/s/wmzuwxfa02bpz7h/Zld-MS2.avi?dl=0>)
- A.9.5. **Video 5.** Transcriptionally active nuclei in a wild-type background. Movie of MS2 fluorescent spots in a wild-type background embryo in nuclear cycle 13, with transcriptionally active nuclei labeled with an overlay. Time is defined with respect to the previous anaphase. (https://www.dropbox.com/s/ugp6pl6o2p3w19e/WTMS2_6-22-16_FracOn.avi?dl=0)
- A.9.6. **Video 6.** Transcriptionally active nuclei in a *zelda*⁻ background. Movie of MS2 fluorescent spots in a *zelda*⁻ background embryo in nuclear cycle 13, with transcriptionally active nuclei labeled with an overlay. Time is defined with respect to the previous anaphase. (https://www.dropbox.com/s/ndc2jlxxtg7kp/Zld-MS2_8-06-15_FracOn.avi?dl=0)
- A.9.7. **Video 7.** Exploration of three-dimensional space consisting of average initial RNAP loading rate and offset and average delay in transcriptional onset time. The models explored in the main text inhabit domains in this space, whereas the wild-type and *zelda*⁻ data inhabit ellipsoids of uncertainty. Whereas the thermodynamic MWC, generalized thermodynamic, and non-equilibrium MWC model with up to

five Bicoid binding sites cannot explain the *zelda*⁻ data, the transcription factor-driven model with three inaccessible states can adequately encompass both datasets. (<https://www.dropbox.com/s/k41kh0ibhpblzda/MetricParameters3D.avi?dl=0>)

Appendix B

Supplementary Information for Chapter 4

B.1 Full Model

To predict MS2 and PP7 fluorescence traces, we utilized a simple model of transcription initiation, elongation, and cleavage. The entire model has the following free parameters:

- $\langle R \rangle$, the mean transcription initiation rate
- $\delta R(t)$, the time-dependent fluctuations in the transcription initiation rate around the mean $\langle R \rangle$
- v_{elon} , the RNAP elongation rate
- τ_{cleave} , the mRNA cleavage time
- t_{on} , the time of transcription onset after the previous mitosis, where $t = 0$ corresponds to the start of anaphase
- $MS2_{basal}$, the basal level of MCP-mCherry fluorescence
- $PP7_{basal}$, the basal level of PCP-eGFP fluorescence
- α , the scaling factor between MCP-mCherry and PCP-eGFP arbitrary fluorescence units

Note that the fluctuations $\delta R(t)$ are independent for each time point, and exist to allow for a slight time dependence in the overall initiation rate. Thus, $\delta R(t)$ parameterizes a set of independent constant offsets in the overall loading rate at each time point.

First, the parameters $\langle R \rangle$, $\delta R(t)$, t_{on} , v_{elon} , and τ_{cleave} were used to generate a map $x_i(t)$ of the position of each actively transcribing RNAP molecule i along the body of the reporter gene, as a function of time. Although the model is represented with continuous time, the

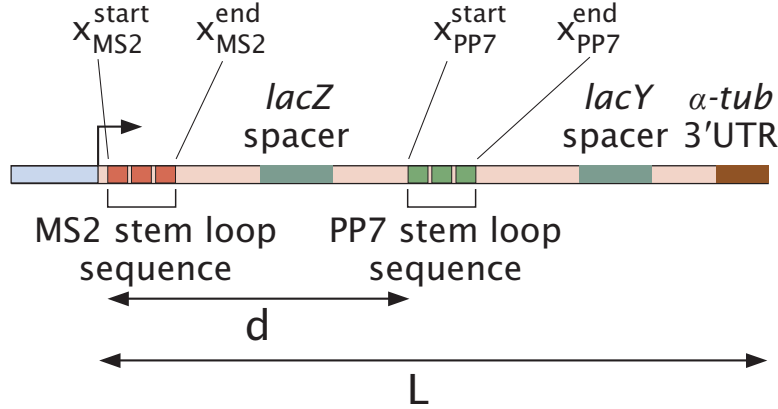


Figure B.1: Detailed description of reporter construct used in this work. Labeled positions are $x_{MS2}^{start} = 0.024$ kb, $x_{MS2}^{end} = 1.299$ kb, $x_{PP7}^{start} = 4.292$ kb, and $x_{PP7}^{end} = 5.758$ kb, where $x = 0$ corresponds to the 3' end of the promoter. Distances are $d = 4.27$ kb and $L = 6.63$ kb.

subsequent computational simulation used for the statistical inference relies on discrete timesteps. Thus, given a computational time step dt , $R(t)dt$ RNAP molecules are loaded at time point t at the promoter $x = 0$, where

$$R(t) = \begin{cases} 0 & t < t_{on} \\ \langle R \rangle + \delta R(t) & t \geq t_{on}. \end{cases} \quad (\text{B.1})$$

Note while $R(t)dt$ is a floating point number, the model utilizes discrete numbers of RNAP molecules. As a result, $R(t)dt$ is rounded down to the nearest integer since the model cannot load fractional numbers of RNAP molecules. After initiation, each RNAP molecule proceeds forward with the constant elongation rate v_{elon} . Once an RNAP molecule reaches the end of the gene, an additional cleavage time τ_{cleave} elapses after which the nascent transcript is cleaved and disappears instantly. This assumption of instantaneous disappearance following cleavage is justified in Section B.3 based on the diffusion time scale of individual mRNA molecules.

From this position map, and based on the locations of the stem loop sequences along the reporter construct (Fig. B.1), we calculate the predicted MS2 and PP7 fluorescence signals. The contribution to the MS2 signal $F_i^{MS2}(t)$ of an individual RNAP molecule i at position $x_i(t)$ is given by

$$F_i^{MS2}(t) = \begin{cases} 0 & x_i(t) < x_{MS2}^{start} \\ \frac{x_i(t) - x_{MS2}^{start}}{x_{MS2}^{end} - x_{MS2}^{start}} F_{MS2} & x_{MS2}^{start} \leq x_i(t) < x_{MS2}^{end}, \\ F_{MS2} & x_i(t) \geq x_{MS2}^{end} \end{cases}, \quad (\text{B.2})$$

where x_{MS2}^{start} and x_{MS2}^{end} are the start and end positions of the MS2 stem loop sequence, respectively, and F_{MS2} is the mCherry fluorescence produced by a single RNAP molecule that has transcribed the entire set of MS2 stem loops. Here, we also assume that RNAP molecules that have only partially transcribed the MS2 stem loops result in a fractional fluorescence given by the fractional length of the MS2 stem loop sequence transcribed. Similarly, the contribution to the PP7 signal $F_i^{PP7}(t)$ is given by

$$F_i^{PP7}(t) = \begin{cases} 0 & x_i(t) < x_{PP7}^{start} \\ \frac{x_i(t) - x_{PP7}^{start}}{x_{PP7}^{end} - x_{PP7}^{start}} F_{PP7} & x_{PP7}^{start} \leq x_i(t) < x_{PP7}^{end} \\ F_{PP7} & x_i(t) \geq x_{PP7}^{end} \end{cases}, \quad (\text{B.3})$$

where x_{PP7}^{start} and x_{PP7}^{end} are the start and end positions of the PP7 stem loop sequence, respectively, and F_{PP7} is the GFP fluorescence produced by a single RNAP molecule that has transcribed the entire set of PP7 stem loops. Note that we assume that the MCP-mCherry and PCP-GFP fluorophores effectively bind instantaneously to all their associated stem loops once they are transcribed. Due to the high numbers of nascent transcripts on the reporter gene (Fig. 4.5D), we expect that corrections to this assumption due to incomplete, stochastic, and/or non-instantaneous fluorophore binding will not introduce substantial deviations to the model.

The temporal dynamics of the total MS2 and PP7 signals $F_{MS2}(t)$ and $F_{PP7}(t)$ are then obtained by summing over all the individual RNAP molecule contributions for each timepoint

$$F_{MS2}(t) = \sum_{i=1}^N F_i^{MS2}(t) \quad (\text{B.4})$$

$$F_{PP7}(t) = \sum_{i=1}^N F_i^{PP7}(t), \quad (\text{B.5})$$

where i is the index of each individual RNAP molecule and N is the total number of loaded RNAP molecules. The final signal is then modified by accounting for the scaling factor α and the basal fluorescence values of $MS2_{basal}$ and $PP7_{basal}$. α is necessary because the two fluorescent protein signals have different arbitrary units (Fig. 4.3). Further, the two basal fluorescence values are incorporated to account for the experimentally observed low baseline fluorescence in each fluorescent channel. The final signals $F'_{MS2}(t)$ and $F'_{PP7}(t)$ are then given by

$$F'_{MS2}(t) = \begin{cases} MS2_{basal}/\alpha & F_{MS2}(t) < MS2_{basal} \\ F_{MS2}(t)/\alpha & F_{MS2}(t) \geq MS2_{basal} \end{cases} \quad (\text{B.6})$$

and

$$F'_{PP7}(t) = \begin{cases} PP7_{basal} & F_{PP7}(t) < PP7_{basal} \\ F_{PP7}(t) & F_{PP7}(t) \geq PP7_{basal} \end{cases}. \quad (\text{B.7})$$

All of the model parameters introduced in this section were used as free parameters in the fitting procedure described in Section B.4.

Note that the model does not make mechanistic claims about the nature of the cleavage process, which could potentially be convolved with processes such as transcriptional pausing. Specifically, if RNAP pausing were to happen 3' of the PP7 stem loop sequence, then it is effectively indistinguishable from cleavage at the 3' UTR.

However, we stress that our model is only an effective parameterization, and so we make no mechanistic claims as to the source of a particular cleavage time value. What our model interprets as cleavage could stem from pausing at the 3'UTR of the reporter, for example, or from continued elongation past the 3'UTR due to inefficient cleavage and termination processes. These would exhibit the same experimental signals—namely, persistence of fluorescent signal after the expected time of signal loss—and thus is a challenge of experimental resolution and not of model formulation.

B.2 Characterization of photobleaching in experimental setup

To determine whether photobleaching was present in our experimental setup, we conducted an experiment with the dual-color 5'/3' tagged reporter (Fig. 4.1C) where half of the field of view was illuminated using the experimental settings described in the Methods and Materials section (Fig. B.2A, purple), and the other half was illuminated at half the temporal sampling rate (Fig. B.2A, yellow).

Since the measurement conditions were identical except for the sampling rate for both reporter constructs used in this work, any systematic differences between the two measurement conditions could only stem from this different sampling rate. Thus, if the experimental settings were in the photobleaching regime, then the purple region would exhibit fluorescence at a systematically lower intensity compared to the yellow region. Figures B.2B and C shows the fluorescence intensities of mCherry and eGFP as a function of time at a particular anterior-posterior position of the embryo for both 0.5x and 1x sampling rates, where data points indicate fluorescence averaged within the anterior-posterior position (indicated schematically by the dashed box in Fig. B.2A) and error bars indicate standard error across cells. The plots reveal that, qualitatively, there is no obvious systematic difference between the two illumination regions.

To quantify photobleaching, we defined the average normalized difference Δ between illuminated regions. This magnitude is calculated by subtracting the fluorescence value at 1x sampling rate F_{1x} by that at 0.5x sampling rate, dividing by the fluorescence value at 0.5x sampling rate $F_{0.5x}$, and then averaging across all time points $N_{timepoints}$ and embryo positions $N_{positions}$

$$\Delta = \sum_{i=1}^{N_{timepoints}} \sum_{j=1}^{N_{positions}} \frac{1}{N_{timepoints}} \frac{1}{N_{positions}} \frac{F_{1x}^{ij} - F_{0.5x}^{ij}}{F_{0.5x}^{ij}}. \quad (\text{B.8})$$

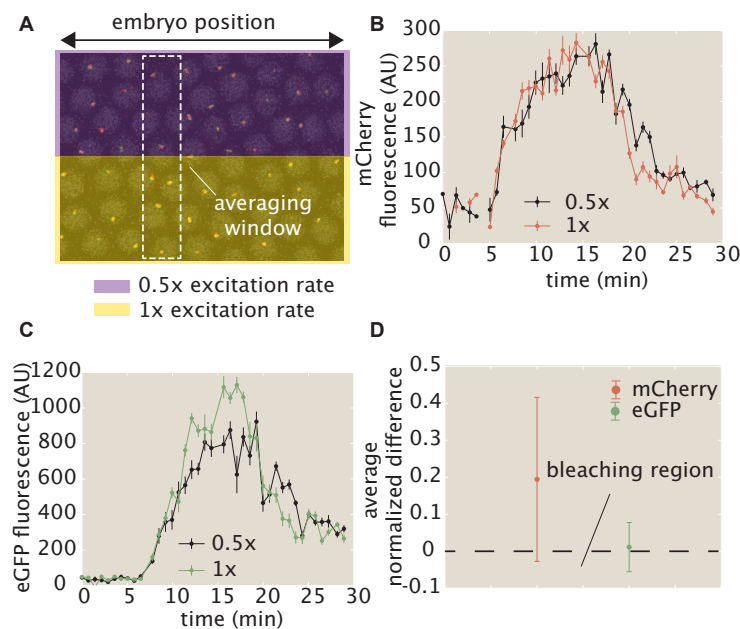


Figure B.2: Investigation of photobleaching in experimental setup. (A) Control experiment where half of the field of view is illuminated at the standard experimental settings (yellow), and the other half of the field of view is imaged at half of the illumination rate (purple). (B, C) The (B) mCherry and (C) eGFP fluorescence signals at a given anterior-posterior embryo position, averaged across cells within that position (white dashed rectangle in (A)), do not exhibit photobleaching. (D) The average normalized difference between illuminated regions, averaged across time points and anterior-posterior embryo positions, are approximately zero within error. A negative value would indicate the presence of photobleaching. (B, C, error bars indicate standard error of the mean averaged across cell nuclei in the field of view; D, error bars indicate standard error of the mean averaged across time points and embryo positions).

For example, for the curves shown in Fig. B.2B, this entails subtracting the red curve by the black curve, dividing by the black curve, and then averaging for all anterior-posterior embryo positions. An overall value of less than zero means that the 1x sampling rate produces systematically lower fluorescence intensities, indicating that our experimental settings are in the photobleaching regime.

As seen in Figure B.2D, the average normalized difference Δ is consistent with zero for both fluorophores (within standard error, measured across all time points and anterior-posterior positions). Thus, we conclude that our data are not in the photobleaching regime.

B.3 Justification for approximating transcript cleavage as instantaneous

In the model presented in Section B.1, we assumed that, when a nascent RNA transcript is cleaved at the end of the reporter gene, its MS2 and PP7 fluorescence signals disappear instantaneously. Here, we justify this assumption by demonstrating that the timescale of mRNA diffusion away from the active locus is much shorter than the experimental resolution of our system.

When a nascent RNA transcript is cleaved, it diffuses away from the gene locus. For a free particle with diffusion coefficient D , the characteristic timescale τ to diffuse a length scale L is given by

$$\tau \sim \frac{L^2}{D}. \quad (\text{B.9})$$

In the context of the experiment performed here, this can be interpreted as the timescale for a cleaved mRNA transcript to diffuse away from the diffraction-limited fluorescence punctum at the locus.

We can estimate the characteristic timescale τ by plugging in the following values. Assume that the completed transcript possesses a typical mRNA diffusion coefficient of $D \sim 0.1 \mu\text{m}^2/\text{s}$ (Gorski et al., 2006). The length scale L corresponds to the Abbe diffraction limit, which yields $L \sim 250 \text{ nm}$ for green light with a wavelength of about 500 nm and a microscope with a numerical aperture of 1. Plugging these values into the equation yields a diffusion time scale of

$$\tau \sim \frac{(250\text{nm})^2}{0.1\mu\text{m}^2/\text{s}} \sim 0.625 \text{ s}. \quad (\text{B.10})$$

As a result, a newly cleaved mRNA transcript will typically diffuse away from the locus in less than a second, meaning that its MS2 and PP7 fluorescence signal will vanish much faster than our experimental time resolution of 15 s . For this reason, we can justify approximating the cleavage process as instantaneously removing the fluorescent signals of newly cleaved transcripts.

B.4 MCMC inference procedure

B.4.1 Overview and application of MCMC

The inference procedures described in the main text were carried out using the established technique of Markov Chain Monte Carlo (MCMC). Specifically, we used the MATLAB package MCMCstat, an adaptive MCMC technique (Haario et al., 2001, 2006). For detailed descriptions, we refer the reader to the MCMCstat website (<https://mjllaine.github.io/mcmcstat/>), as well as to a technical overview of MCMC (Geyer, 1992). Briefly, MCMC allows for an estimation of the parameter values of a model that best fit the experimentally observed data along with an associated error. In this work, we use MCMC to infer the best fit

values of the transcription cycle parameters given observed fluorescence data at the single-cell level. Then, we combine these inference results across cells to construct distributions of inferred values across the ensemble of cells.

MCMC calculates a Bayesian posterior probability distribution of each free parameter given the data by stochastically sampling different parameter values. For a given set of observations D and a model with parameters θ , the so-called posterior probability distribution of θ possessing a particular set of values is given by Bayes' theorem

$$\underbrace{p(\theta|D)}_{\text{posterior}} = \frac{\overbrace{p(D|\theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(D)}_{\text{evidence}}}. \quad (\text{B.11})$$

This posterior distribution is a combination of three components: the likelihood, prior, and evidence. This latter term represents the probability of the observations possessing their particular values, and allows the overall posterior distribution to be normalized. In practice, the evidence term is often dropped since MCMC can still yield accurate results without requiring this normalization. Thus, we have

$$\underbrace{p(\theta|D)}_{\text{posterior}} \propto \overbrace{p(D|\theta)}^{\text{likelihood}} \overbrace{p(\theta)}^{\text{prior}}. \quad (\text{B.12})$$

The prior function contains *a priori* assumptions about the probability distribution of parameter values θ , and the likelihood function represents the probability of obtaining the observations, given a particular set of parameters θ . Thus, the *most likely* set of parameters θ occurs when the product of the likelihood and prior is maximized, resulting in a maximum in the posterior function. MCMC extends this by sampling different values of θ such that an approximation of the full posterior distribution is also obtained.

The prior distributions for the inferred parameters were set as follows. The prior distribution for the fluctuations in the initiation rate $\delta R(t)$ at each time point was assumed to be a Gaussian distribution centered around 0 AU/min with a standard deviation of 30 AU/min. This penalized fluctuations that strayed too far from zero, smoothing the overall initiation rate $R(t)$. For the rest of the parameters, a uniform distribution was chosen using the following uniform intervals:

- v_{elon} : [0, 10] kb/min
- t_{on} : [0, 10] min
- α : [0, 1]
- τ_{cleave} : [0, 20] min

- $MS2_{basal}$: $[0, 50]$ AU
- $PP7_{basal}$: $[0, 50]$ AU
- $\langle R \rangle$: $[0, 40]$ AU/min

These intervals were justified with the following arguments. Previous elongation rate measurements have indicated values between around 1 and 4 kb/min (Fig. B.9; (Ardehali and Lis, 2009)), so we approximately doubled this range for flexibility. Previous measurements of the transcription onset time t_{on} for *hunchback* range from about 1 to 6 min (Garcia et al., 2013), so we chose a similarly flexible interval. The calibration factor α must take on values between 0 and 1, since, under the experimental settings used, mCherry exhibits weaker absolute fluorescence than eGFP (see for example, Fig. 4.3C). Although the cleavage time is not well understood, estimates lie on the order of minutes (Lenstra et al., 2016)—we chose a large interval to be conservative. Based on our experimental data (e.g. Fig. 4.2B), basal levels of MS2 and PP7 fluorescence lie comfortably in the range $[0, 50]$ AU. Finally, as observed in our data and also reported in Garcia et al. (2013), the mean rates of initiation lie comfortably in the range $[0, 40]$ AU/min (Fig. 4.4A).

For the likelihood function, a Gaussian error function was used

$$p(D|\theta) = e^{-SS}, \quad (\text{B.13})$$

where SS is a scaled sum-of-squares residual function given by

$$SS = \sum_t \frac{(F_{data} - F_{prediction})^2}{F_{data}}. \quad (\text{B.14})$$

Here, the summation runs over individual time points, F_{data} corresponds to the MS2 or PP7 fluorescence at a given timepoint, and $F_{prediction}$ corresponds to the predicted MS2 or PP7 fluorescence according to the model, for a given set of parameter values. That is,

$$F_{data} = \{MS2_1, \dots, MS2_N, PP7_1, \dots, PP7_N\} \quad (\text{B.15})$$

where the subscripts indicate the time index over N time points. Similarly,

$$F_{prediction} = \{MS2_1^{pred}, \dots, MS2_N^{pred}, PP7_1^{pred}, \dots, PP7_N^{pred}\} \quad (\text{B.16})$$

where the superscripts indicate that these are model predictions evaluated at the experimental time points. The presence of F_{data} in the denominator scales the overall sum-of-squares residual function by the mean signal intensity and is required because the measurement noise in the fluorescence scales linearly with fluorescence intensity (Section B.4.2 and Fig. B.3).

The MCMC approach samples values of parameters θ to approximate the posterior probability distribution. There are several algorithms that achieve this—the adaptive technique used in the MCMCstat package is an efficient algorithm that updates the sampling technique to more quickly arrive at the converged distribution.

For each inference run, an initial condition of parameter values is chosen. The algorithm then stochastically updates the next set of parameter values based on the current and previous values of the posterior distribution function. After a preset number of updates (typically at least on the order of thousands), the algorithm stops, resulting in a *chain* of MCMC parameter value samples. The initial period following the initial condition, known as the *burn-in* time, is typically discarded since the results are not reliable. The remaining values of the chain comprise an approximation of the underlying posterior probability distribution, with smaller errors for longer run times.

For the purposes of this work, the MCMC procedure was run by separately inferring parameter values for the data corresponding to each single cell. For each inference, random parameter values were chosen for the initial condition of the sampling algorithm in order to prevent initial condition bias from affecting the inference results. The algorithm was run for a total of 20,000 iterations, which, after removing a burn-in window of length 10,000, resulted in a chain of length 10,000 for each of the 355 cells examined. To assess whether or not the algorithm was run for a sufficient number of iterations, the final chain was examined for *rapid mixing*, where the sampled values of a particular parameter rapidly fluctuate around a converged value. Figure 4.2C highlights this rapid mixing in the inferred transcription cycle parameters of a sample single cell. The lack of long-timescale correlations, also exemplified by the quick decay of the auto-correlation function of each chain (Fig. 4.2D), indicates that the algorithm has converged. In addition, a corner plot of the three transcription cycle parameters (Fig. 4.2E) illustrates the pairwise correlations between them, demonstrating that the inference did not encounter degenerate solutions, and that each parameter has a fairly unimodal distribution.

These diagnostics provided a check on the quality of the inference results. Afterwards, the mean value of each parameter’s final chain was then retained for each single cell for use in the further statistical analysis carried out in the main text.

B.4.2 Justification of scaled observation model due to fluorescence noise behavior

The observation model parameterized by the sum-of-squares residual in Equation B.14 is scaled by dividing by the overall fluorescence intensity. This is needed because the fluorescence noise is not constant, but rather scales linearly with overall intensity. Here, we demonstrate this behavior by examining the fluorescence noise exhibited in our system.

A priori, if we consider that the fluorescent signals in our experiment are the result of the sum of many individual fluorophores, then we would expect that, if an individual fluorophore possesses some intrinsic constant measurement error with variance σ^2 , then the associated error of N fluorophores would have a similarly scaled overall measurement error with variance $N\sigma^2$. Since N is proportional to the overall mean fluorescent signal, the observation model in Equation B.14 thus needs the mean signal in the denominator.

To validate this scaling of the variance with the mean, we examined the data from the

dual-color interlaced MS2/PP7 reporter construct from Figure 4.3B. These data constitute, in principle, a two-point measurement of the same underlying biological process, so we reasoned that we could utilize this measurement to quantify the scaling of fluorescence noise with respect to overall fluorescence intensity.

Specifically, by creating bins of eGFP fluorescence measurement from the scatterplot in Figure 4.3D, we calculated how the variance of associated mCherry fluorescence values within a bin scaled with eGFP fluorescence (here a proxy for overall fluorescence intensity). If the calculated variance increased with overall fluorescence, this would indicate that the fluorescence measurement noise is not constant, but rather scaled positively with signal strength. Figure B.3 shows this calculated variance (red), along with bootstrapped standard error, as a function of bin value (i.e. eGFP fluorescence). We see that the variance indeed increases with bin value fairly linearly, confirming our hypothesis. If we then scale the variances by dividing by the mean mCherry fluorescence within a bin, we recover a constant scaling, as expected (black).

The fluorescence intensity of each detected MS2 or PP7 spot was calculated by integrating the pixel intensities in a small circular neighborhood with a fixed radius of about 1 micron around each spot center and subtracting by the background fluorescence, calculated by fitting a Gaussian to the spatial fluorescence profile (see Methods and Materials). While the number of detected pixels does contribute to the fluorescence intensity (and thus variance across measurements), the size of a spot does correlate with overall transcriptional activity – thus, the scaling of signal variance depends on multiple factors but would be expected to increase with spot brightness, and to a lesser degree, size, both of which contribute to the overall integrated intensity within the neighborhood.

The observed behavior of fluorescence variance is intriguing because previous work using the same spot detection methodology found that the dominant contributor to fluorescence noise was background fluorescence outside of the actively transcribing locus (Garcia et al., 2013). In contrast, this work is consistent with a scenario where the noise intrinsic to the individual fluorophore molecules dominates, leading to the observed scaling of fluorescent noise with the mean intensity. We speculate that, in this work, the difference in fluorescence noise behavior stems from the usage of mCherry, whose signal is lower, and therefore noisier, than that of GFP in the context of the fruit fly embryo (Fig. 4.3F). In addition, other differences such as usage of MS2-mCherry instead of MS2-GFP and a different maternal fly line driving different levels of constitutive MCP-mCherry and PCP-GFP could change the relative strength of background fluorescence noise.

B.4.3 Curation of inference results

Individual single cell inference results were filtered automatically and then run through an automated curation procedure for final quality control. First, due to experimental and computational imaging limits, some MS2 or PP7 trajectories were too short to run a meaningful inference on. As a result, we automatically skipped over any cell with an MS2 or

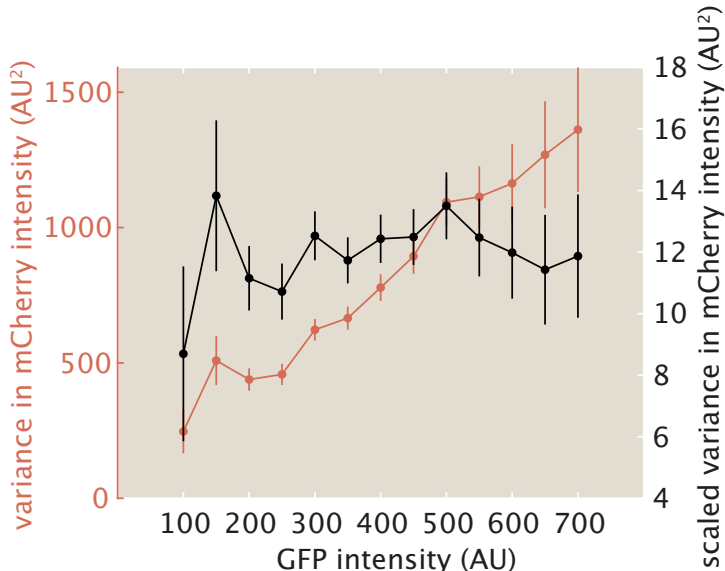


Figure B.3: Scaling of fluorescence measurement noise with overall fluorescence intensity. Variance of mCherry fluorescence at a particular GFP fluorescence (red), from the dual-color interlaced reporter construct from Figure 4.3B, along with variance scaled by dividing out the mean mCherry fluorescence (black).

PP7 signal with fewer than 30 datapoints. This amounted to 626 cells skipped out of a total of 1053, with 427 (41%) retained.

Second, the retained cells were run through an automated curation pipeline. For each single-cell fit, we calculated the average squared normalized residual δ^2 , defined as

$$\delta^2 = \sum_{\text{timepoints}} \frac{(F_{\text{data}} - F_{\text{fit}})^2}{F_{\text{data}}^2}, \quad (\text{B.17})$$

where the summation occurs over all time points and F_{data} and F_{fit} correspond to the fluorescence data and fit, respectively. Thus, δ^2 gives a measure of how good or bad, on average, each single-cell fit is. Figure B.4A and B show histograms of the average squared normalized residual δ^2 for the entire $n = 427$ dataset, with log and linear x-axes. We see that the vast majority of data possesses values of δ^2 smaller than unity, with a long tail at higher values corresponding to bad fits. We decided to implement a cutoff of $\delta_{\text{cutoff}}^2 = 1$ (red line), where any cell with a higher value of δ^2 was automatically discarded.

In sum, 355 cells of data were retained out of 427 total after this curation process. We reasoned that, since we still ended up with hundreds of single cells of data, the resultant statistical sample size was large enough to extract meaningful conclusions.

To assess the rejected fits for underlying biological causes, we did a qualitative examination for common features. There were several sources of bad fits. First, some traces possessed low

signal-to-noise ratio (Fig. B.4C) that nevertheless yielded reasonable fits that were slightly above δ_{cutoff}^2 . Still others simply had poor fits, possibly due to running into issues with the inference algorithm such as getting trapped in local minima (Fig. B.4D). We consider improvements to the algorithm to be outside the scope of this work, since the retained data still contain enough statistical size to provide interpretable results.

Finally, one potential biological source confounding the model could be the presence of substantial transcriptional bursting of the promoter. Although the majority of the traces we analyzed indicated that the *hunchback* reporter gene studied here possessed a promoter that was effectively ON during the cell cycle studied, a small fraction of traces (4% of the filtered cells) possessed substantial time dependence of the fluorescence signal, potentially resulting from rapid switching of the promoter between ON and OFF states (Fig. B.4E).

The presence of transcriptional bursts is of high biological significance, but capturing the behavior would require more specific models (e.g. two-state telegraph models like Lammers et al. (2020)). As a result, we relegate extensions of the model that can account for transcriptional bursting for future work. Thus, our work provides a self-contained framework applicable for describing the behavior of promoters that are primarily ON for the duration of the experiment and that do not experience transcriptional bursting.

Due to the variety of sources contributing to the rejected fits, we opted for a conservative approach and only analyzed the cells with high signal quality that did not exhibit the complications mentioned above. The number of retained fits were still much higher than the number of rejected fits (Fig. B.4F).

To check that the curation procedure did not incur substantial bias, we compared the average inferred mean initiation rate, elongation rate, and cleavage time as a function of embryo position between the post-filtering curated and uncurated datasets of size $n = 355$ and $n = 427$, respectively (Fig. B.4G-I). We observed no substantial difference between the two datasets, indicating that the curation procedure was not systematically altering the inference results.

B.4.4 Validation of inference results

To assess the accuracy of the inference method, we validated our MCMC approach against a simulated dataset. Using the inferred distribution of model parameters from the experimental data, we generated a simulated dataset with our theoretical model (Section B.1) and ran the MCMC inference on it.

The simulated dataset consisted of 300 cells. The model parameters used to simulate each individual cell's MS2 and PP7 fluorescences were drawn randomly from a Gaussian distribution, with mean μ and standard deviation σ calculated from the distribution of inferred model parameters from the experimental data. Table B.1 shows the parameters used in the Gaussian distributions generating each single cell's model parameters. We chose to fix the time-dependent fluctuations in the initiation rate $\delta R(t)$ at zero since these fluctuations are not well understood at the single-cell level, and the *hunchback* reporter studied here is well parameterized by a mean initiation rate (Fig. 4.2B).

	mean (μ)	standard deviation (σ)
$\langle R \rangle$	16.6 AU/min	5.1 AU/min
$\delta R(t)$	0	0
v_{elon}	1.8 kb/min	0.8 kb/min
τ_{cleave}	3.1 min	1.4 min
t_{on}	3.5 min	1.6 min
α	0.16	0.05
MS2 _{basal}	10 AU	5 AU
PP7 _{basal}	10 AU	5 AU

Table B.1: Mean and standard deviation of model parameters used in single-cell simulations.

In addition, fluorescence measurement error was generated for each single cell and at each time point by drawing a random number from a Gaussian distribution with mean 0 and standard deviation $10 \times \sqrt{F_{sim}}$ AU, where F_{sim} is the fluorescence at each time point, and adding this random number to the MS2 or PP7 fluorescence at that time point (prior to rescaling the MS2 fluorescence with the scaling factor α). Here, the $\sqrt{F_{sim}}$ factor in the magnitude of the fluorescence noise accounts for our observation that the variance of the fluorescence measurement noise scales linearly with the mean signal intensity (Fig. B.3).

Figure B.5A shows an example of the simulated MS2 and PP7 fluorescence from a single cell along with their corresponding fits. The resulting MCMC-sampled values of the mean initiation rate, elongation rate, and cleavage time are shown in the histograms in Figure B.5B (blue), along with the ground truth for that single cell (red line). As described in Section B.4.1, the mean value of each sampled distribution was retained for downstream statistical analysis.

The accuracy of the inference was investigated on three levels: 1) systematic errors affecting mean analyses, 2) random errors affecting measurements of distributions, and 3) spurious correlations between parameters affecting inter-parameter correlations.

First, the scaled error ε for each parameter was calculated on a single-cell basis as defined by

$$\varepsilon = \frac{x_{infer} - x_{truth}}{\mu_x}, \quad (\text{B.18})$$

where x represents the model parameter being investigated, the subscripts indicate whether the quantity is the inferred result or the ground truth for that single cell, and μ_x is the population mean of the parameter value from the experimental data (i.e., the values of the “mean” column in Table B.1). For example, for the mean initiation rate $\langle R \rangle$, $\mu_{\langle R \rangle}$ takes the value 16.6 AU/min. ε gives a unitless measure of the magnitude of inference error of each single cell, where a value of 1 indicates an error that is as large as the population mean itself. Because the scaled error is defined as the error due to inference for a single cell, it is an intensive quantity that is independent from the overall dataset size.

Figure B.5C shows the histogram of single-cell scaled errors $\varepsilon_{\langle R \rangle}$, $\varepsilon_{v_{\text{elong}}}$, and $\varepsilon_{\text{tau}_{\text{cleave}}}$ for the inferred mean initiation rate, elongation rate, and cleavage time, respectively. The majority of the scaled errors fall between -0.5 and 0.5 , indicating that most inferred results possess relatively small error.

The systematic error on measurements of the ensemble mean can be estimated by calculating the mean of the scaled errors shown in Figure B.5C. Doing so results in a value of -0.06 ± 0.01 , -0.01 ± 0.02 , and 0.04 ± 0.02 (mean and SEM) for the mean scaled error of the mean initiation rate, elongation rate, and cleavage time, respectively. For context, this means that, if the mean cleavage time is ~ 3 min, then the systematic error in the cleavage time is ~ 10 sec, about the time resolution of the data. Thus, the systematic error for each parameter is a couple orders of magnitude below that of the experimental mean value of each parameter, indicating that the inference provides an accurate and precise readout of the mean.

While the inference's systematic error across cells may be small, the presence of individual single-cell errors will affect measurements of distributions of parameters. To investigate the impact of these random errors, we quantified the fraction of total empirically inferred variability that consisted of inferential error. Specifically, for a parameter x , we separated the variance of single-cell measurements as

$$\sigma_{x,\text{total}}^2 = \sigma_{x,\text{empirical}}^2 + \sigma_{x,\text{inference}}^2, \quad (\text{B.19})$$

where $\sigma_{x,\text{total}}^2$ represents the overall single-cell variability observed in the data (the combination of empirical and inferential variability), $\sigma_{x,\text{inference}}^2$ represents the error inherent to our inference process, and $\sigma_{x,\text{empirical}}^2$ represents the true empirical variability after subtracting out inferential error $\sigma_{x,\text{inference}}^2$. Note that $\sigma_{x,\text{total}}^2$ is the square of the values in the standard deviation column in Table B.1.

Dividing by the square of the population means μ_x yields

$$\frac{\sigma_{x,\text{total}}^2}{\mu_x^2} = \frac{\sigma_{x,\text{empirical}}^2}{\mu_x^2} + \frac{\sigma_{x,\text{inference}}^2}{\mu_x^2}. \quad (\text{B.20})$$

Note that these are just squared CV terms, and that the last term is simply the square of the scaled error ε defined earlier

$$CV_{x,\text{total}}^2 = CV_{x,\text{empirical}}^2 + \varepsilon_x^2. \quad (\text{B.21})$$

Thus, the overall impact of the inferential error can be quantified by calculating the relative magnitudes of the contributions of $CV_{x,\text{empirical}}^2$ and ε^2 to the total variability $CV_{x,\text{total}}^2$. Figure B.5D shows this separation, where the dark bars represent the squared scaled error ε^2 , the light bars represent the true empirical variability $CV_{x,\text{empirical}}^2$, and the overall bars represent the total variability $CV_{x,\text{total}}^2$ obtained from the values of μ and σ in Table B.1.

All three model parameters—initiation, elongation, and cleavage—possess no more than approximately 25% inferential error. Nevertheless, the presence of this much error indicates

that measurements of distributions of these parameters will be somewhat confounded by the inherent error present in our inference method, highlighting the general difficulty in measuring values beyond the mean.

However, these errors in the inference of the variability of the transcription cycle parameters should not impact the results of investigating the distribution of elongation rates in Figure 4.4D, since the simulated results there were also pushed through the inference pipeline and should pick up similar inferential noise. Furthermore, the variances of the simulated distributions in the presence or absence of single-molecule elongation variability differed by essentially around a factor of two (Fig. B.10D), twice as much as the random error exhibited in the simulated results here (see Section B.10 for details).

Future improvements on increasing the accuracy of measurements of distributions could be achieved, for example, by utilizing interleaved loops such as those introduced in Figure 4.3B. Here, two orthogonal species of mRNA binding proteins fused to different fluorescent proteins would bind to interleaved loops located at the 5' end of the construct. In addition, a second pair of mRNA binding proteins would bind to an analogous set of interleaved loops located at the 3' end. The result would be a four-color experiment, with two colors reporting on transcription at the 5' end of the transcript, and two different colors reporting on transcription the 3' end. In this scenario, the data would provide independent readouts of the same underlying signal, making it possible to perform two independent inferences on the same nucleus. This would allow for the decomposition of the inference into biological variability and inferential error using techniques analogous to those presented in B.8.

Finally, we examined the inference method for spurious correlations to investigate the accuracy of the experimental single-cell correlations shown in Figure 4.5. The presence of spurious correlations would reflect inherent couplings in the inference method itself, since the simulation parameters were generated independently and stochastically.

Figure B.5E-H show the single-cell correlations using the Spearman rank correlation coefficient between model parameters for the simulated dataset, as well as between the mean RNAP density and the cleavage time, as defined in the main text. Linear regression fits are also displayed for intuitive visualization. We discovered a slight positive correlation ($\rho = 0.15$) between the elongation rate and the cleavage time (Fig. B.5G, p -val = 0.01). In contrast, there was no significant correlation between the mean initiation rate and the cleavage time, the mean initiation rate and the elongation rate, and the mean RNAP density and the cleavage time (Fig. B.5E, F, and H). Although the relationship between the elongation rate and the cleavage time possessed the same, albeit weaker, correlation as found in the data (Fig. 4.5C), the main finding in the main text of the correlation between the mean RNAP density and the cleavage time was not reproduced by the simulations (Fig. B.5H). The comparisons of Spearman rank correlation coefficients and p -values between the data and simulations are summarized in Table B.2.

Thus, our results validated the single-cell correlations discovered in the main text, indicating that the experimental results were not the product of spurious correlations.

	<u>initiation</u> <u>cleavage</u>	<u>initiation</u> <u>elongation</u>	<u>elongation</u> <u>cleavage</u>	<u>RNAP density</u> <u>cleavage</u>
data	$\rho = -0.52$ $p\text{-val} \approx 0$ negative correlation	$\rho = -0.21$ $p\text{-val} = 5 \times 10^{-5}$ negative correlation	$\rho = 0.35$ $p\text{-val} = 2 \times 10^{-11}$ positive correlation	$\rho = -0.55$ $p\text{-val} \approx 0$ negative correlation
simulation	$\rho = 0.07$ $p\text{-val} = 0.24$ insignificant correlation	$\rho = 0.01$ $p\text{-val} = 0.86$ insignificant correlation	$\rho = 0.15$ $p\text{-val} = 0.01$ positive correlation	$\rho = -0.01$ $p\text{-val} = 0.86$ insignificant correlation

Table B.2: Comparison of Spearman rank correlation coefficients and p -values between experimental and simulated single-cell correlations.

B.5 Validation of the RNAP processivity assumption

The calibration between the MS2 and PP7 signals (Fig. 4.3) provided an opportunity to test the processivity assumption presented in the main text, namely that the majority of loaded RNAP molecules transcribe to the end of the gene without falling off. To estimate the processivity quantitatively, we assume that a series of N RNAP molecules transcribes past the MS2 stem loop sequence at the 5' end of the reporter gene, and that only pN successfully transcribe past the PP7 stem loop sequence at the 3' end. Here, we define p to be the processivity factor, and require $0 < p < 1$. Thus, $p = 1$ indicates maximal processivity where every RNAP molecule that transcribes the MS2 sequence also transcribes the PP7 sequence, and $p = 0$ indicates minimal processivity, where no RNAP molecules make it to the PP7 sequence.

We assume that no RNAP molecules fall off the gene while they transcribe the interlaced MS2/PP7 loops used in the calibration experiment described in Figure 4.3B. Under this assumption, N RNAP molecules will fully transcribe both sets of stem loop sequences, allowing us to define the scaling factor as the ratio of total fluorescence values

$$\alpha_{\text{calib}} = \frac{NF_{MS2}}{NF_{PP7}} = \frac{F_{MS2}}{F_{PP7}}. \quad (\text{B.22})$$

Note that, in this simple model, RNAP molecules can still fall off the gene after they transcribe the set of MS2/PP7 loops. Now, we consider the construct with MS2 and PP7 at opposite ends of the gene used in the main text. Allowing a fraction p of RNAP molecules to fall off the gene between the MS2 and PP7 loops, we arrive at a scaling factor

$$\alpha_{\text{infer}} = \frac{NF_{MS2}}{pNF_{PP7}} = \frac{F_{MS2}}{pF_{PP7}}. \quad (\text{B.23})$$

We can thus calculate the processivity p from taking the ratio of the true and biased scaling factors

$$p = \frac{\alpha_{\text{calib}}}{\alpha_{\text{infer}}}. \quad (\text{B.24})$$

Taking the mean value of α_{calib} from our control experiment using the interlaced MS2/PP7 loops to be the true value and the mean value of α_{infer} from the inference from the main text to be the biased value, we calculate a mean processivity of $p = 0.96$, with a negligible standard error of 4.81×10^{-5} . Thus, on average, 96% of RNAP molecules that successfully transcribe the 5' MS2 stem loop sequence also successfully transcribe the 3' PP7 stem loop sequence, confirming previous results (Femino et al., 1998; Garcia et al., 2013) and lending support to the processivity assumption invoked in our model.

B.6 Comparing intra- and inter-embryo variability

In the analysis in the main text, we treated all single cell inference results equally within one statistical set. In principle, this is justified only if the variability between single cells is at least as large as the variability between individual embryos. In this section we prove this assumption.

Here, we examine two quantities: the *intra-embryo variability*, defined as the variance in a parameter across all single cells in a single embryo, and the *inter-embryo variability*, defined as the variance across embryos in the single-embryo mean of a parameter. We examined these two quantities for the three primary inferred parameters—the mean initiation rate, elongation rate, and cleavage time.

Figure B.6A-C shows the results of this comparison as a function of embryo position, where the red (blue) lines indicate the intra- (inter-) embryo variability and the red (blue) shaded regions indicate the standard error (bootstrapped standard error) in the intra- (inter-) embryo variability. For all of the parameters, the intra-embryo variability is at least as large as the inter-embryo variability, validating our treatment of all of the single-cell inference results as a single dataset, regardless of embryo.

This is seen more clearly when the data are averaged across embryo position. As shown in Fig. B.6D, the inter-embryo variability of each parameter is substantially higher than the intra-embryo variability.

B.7 Full distributions of transcriptional parameters as a function of embryo position

Figure 4.4 presents inferred values of the transcriptional parameters in the form of population means and CVs as a function of embryo position. We chose this form of presentation to focus on spatial variation of these parameters via a succinct visualization.

Figure B.7 shows the full distributions of the transcriptional parameters as a function of embryo position. For each parameter, the observed variability at a particular position in the embryo is quite broad, indicating substantial cell-to-cell variability. Nevertheless, there is no clear indication of multimodal behavior, indicating that the mean is still a reliable metric of population-averaged behavior.

B.8 Comparison of variability in mean initiation rate reported by our inference with static measurements

A widespread strategy to measure variability in transcription initiation relies on techniques such as single-molecule FISH (smFISH), which count the number of nascent transcripts at a transcribing locus in a fixed sample (Femino et al., 1998; Raj et al., 2006; Pare et al., 2009; Zenklusen et al., 2008; Wyart et al., 2010; So et al., 2011; Boettiger and Levine, 2013; Little et al., 2013; Jones et al., 2014; Senecal et al., 2014; Padovan-Merhar et al., 2015; Xu et al., 2015; Albayrak et al., 2016; Skinner et al., 2016; Bartman et al., 2016; Gomez-Schiavon et al., 2017; Hendy et al., 2017; Munsky et al., 2018; Zoller et al., 2018; Miura et al., 2019). These single time point measurements are typically interpreted as reporting on the cell-to-cell variability in transcription initiation. Further, under the right conditions, the variability reported by this method has been shown to be dominated by biological sources of variability and to have a negligible contribution from experimental sources of noise (Zoller et al., 2018).

Inspired by these measurements in fixed embryos, we sought to determine how well our approach could report on biological variability. To do so, we contrasted the inference results of the transcriptional activity of our *hunchback* reporter with a snapshot-based analysis inspired by single-molecule FISH (Zoller et al., 2018). Specifically, we calculated the CVs in the raw MS2 and PP7 fluorescence in snapshots taken at 10 minutes after the start of nuclear cycle 14, from the same post-curation cells analyzed with the inference method. We reasoned that, since this calculation does not utilize the full time-resolved nature of the data, it provides a baseline measurement of total noise that encompasses both experimental and biological variability. As a point of comparison, we also calculated the CV in the instantaneous MS2 signal from another work using a similar P2P-MS2-lacZ construct (Eck et al., 2020).

Figure B.8A shows the CV as a function of embryo position as reported by these different approaches. For the static measurements (red, green, and blue), the CV values lay around 20% to 80%. The CV of the inferred mean initiation rate (purple) exhibited similar values, although it was slightly lower in a systematic fashion. This difference was likely due to the fact that the inference relies on time-dependent measurements that can average out certain sources of error such as experimental noise, whereas such time averaging is not possible in the context of single time point measurements.

To succinctly quantify variability in the mean initiation rate, we then calculated the position-averaged squared CV for the same measurements in Figure B.8A. The resulting

squared CV values are shown in Figure B.8B. Although the static measurements possessed essentially identical squared CVs (blue, red, green), the inference method exhibited a clear reduction in the squared CV (purple).

To test whether the discrepancy in the variability between time-resolved and snapshot-based measurements arose from differences in the experimental error of each technique, we used the formalism introduced by Elowitz et al. (2002) to separate the noise in the system into uncorrelated and correlated components. Here, uncorrelated noise represents random measurement error, while correlated noise contains both systematic measurement error as well as true biological variability. To perform this separation, we utilized the alternating MS2-PP7 reporter used in the calibration calculation (Fig. 4.3B). Because the MS2 and PP7 fluorescent signals in this reporter construct should, in principle, reflect the same underlying biological signal, deviations in each signal from each other should report on the relative magnitudes of both types of noise.

First, we defined the deviations δ_{MS2} and δ_{PP7} of each instantaneous MS2 and PP7 fluorescent signal from the mean MS2 and PP7 fluorescence signals, averaged across nuclei and time

$$\delta_{MS2} = \frac{F_{MS2}}{\langle F_{MS2} \rangle} - 1 \quad (\text{B.25})$$

$$\delta_{PP7} = \frac{F_{PP7}}{\langle F_{PP7} \rangle} - 1, \quad (\text{B.26})$$

where F_{MS2} and F_{PP7} are the respective instantaneous MS2 and PP7 fluorescence values for a given nucleus and time point, and $\langle F_{MS2} \rangle$ and $\langle F_{PP7} \rangle$ are the respective mean MS2 and PP7 fluorescence values, averaged across nuclei and time points. Using these deviations, the uncorrelated and correlated noise terms are defined as

$$\eta_{uncorr}^2 = \frac{1}{2} \langle (\delta_{MS2} - \delta_{PP7})^2 \rangle \quad (\text{B.27})$$

$$\eta_{corr}^2 = \langle \delta_{MS2} \delta_{PP7} \rangle, \quad (\text{B.28})$$

where the brackets indicate an ensemble average over time points and cells (Elowitz et al., 2002). From this, the total noise η_{tot}^2 , defined as the variance σ^2 divided by the mean squared μ^2 , is simply the uncorrelated and correlated noise components added in quadrature

$$\eta_{tot}^2 = \frac{\sigma^2}{\mu^2} = \eta_{uncorr}^2 + \eta_{corr}^2. \quad (\text{B.29})$$

Note that the total noise η_{tot}^2 is simply the squared coefficient of variation. Thus, the squared coefficient of variation (CV^2) of our data is equal to η_{tot}^2 and can be separated into the uncorrelated and correlated components.

Figure B.8B shows this CV^2 (averaged across all embryo positions) for snapshots of the interlaced loop construct compared with the separated uncorrelated and correlated noise sources. Intriguingly, the uncorrelated and correlated noise (yellow) each contribute about

half to the overall noise. We posit that the relative magnitude of partitioning between correlated and uncorrelated noise also holds for the static measurements of spot fluorescence (Fig. B.8B, blue, red and green). As a result, given this assumption, we can calculate the correlated and uncorrelated variability contributions to total squared CV from these static measurements. This is shown in light and dark red in the case of the static MS2 fluorescence measurement in Figure B.8B. The figure reveals that the correlated noise component of the static measurements (dark red) is only slightly smaller than the overall noise measured by the inference (purple), suggesting that our inference method primarily reports on correlated variability.

As a result, the MCMC inference method can quantitatively capture the true biological variability in the mean initiation rate while separating out most of the uncorrelated contribution due to random experimental noise. Thus our results support the power of model-driven inference approaches in providing clean readouts of variability in transcriptional parameters.

B.9 Comparison of distribution of elongation rates with other works

As an additional validation of our inference results, we compared the distribution of single-cell inferred elongation rates with those reported in two similar works by Hocine et al. (2013) and Fukaya et al. (2017). Both of these works used a two-color live imaging reporter like the one utilized in this work, and measured the time delay between the onset of each stem loop signal to estimate a single-cell mean elongation rate. Fukaya et al. (2017) studied a similar *hunchback* reporter to the one used here, while Hocine et al. (2013) used a reporter construct in yeast.

Figure B.9 shows the comparison of distributions of elongation rates. Because the reporter constructs and analysis techniques differed between works, a quantitative comparison is not possible. Nevertheless, all three sets of results report a significant cell-to-cell variability in mean elongation rate, ranging from 1 kb/min to 3 kb/min.

B.10 Theoretical investigation of single-cell distribution of elongation rates

To investigate the molecular mechanisms underling single-cell distributions of elongation rates obtained from the inference, we developed a single-molecule theoretical model. We were interested in how the observed variability in single-cell elongation rates could constrain models of the single-molecule variability in RNAP elongation rates. To disregard effects due to position-dependent modulations in the transcription initiation rate, we only studied cells anterior of 40% along the embryo length, where the initiation rate was roughly constant.

The model was adapted from the stochastic Monte Carlo simulation used in Klumpp and Hwa (2008), which accounts for the finite size of RNAP molecules (Fig. B.10A). Here, single RNAP molecules are represented by one-dimensional objects of size $N_{footprint}$ that traverse a gene consisting of a one-dimensional lattice with a total number of sites, corresponding to single base pairs, equal to N_{sites} . The position of the active site of molecule i is given by x_i , which takes integer values—each integer corresponds to a single base pair of the gene lattice. Because RNAP molecules have a finite size, given by $N_{footprint}$, an RNAP molecule i thus occupies the lattice sites from x_i to $x_i + N_{footprint}$. In this model, we do not incorporate sequence-dependent RNAP pausing along the gene.

New RNAP molecules are loaded at the start of the gene located at $x = 0$. Due to the exclusionary interactions between molecules, simultaneously simulating the motion of all molecules is unfeasible, and a simulation rule dictating the order of events is necessary.

At each simulation timestep dt , a randomized sequence of indices is created from the following sequence

$$\mathcal{I} = \{0, 1, \dots, N\}, \quad (\text{B.30})$$

where $\{1, \dots, N\}$ correspond to any RNAP molecules $i = 1, \dots, N$ already existing on the gene, and 0 corresponds to the promoter loading site that generates new RNAP molecules.

Choosing indices i from the random sequence \mathcal{I} obtained above, the following actions are taken. If the index i indicates that an RNAP molecule was chosen ($i > 0$), then that RNAP molecule advances forward with stochastic rate ϵ . This probability is simulated by drawing a random number from a Poisson distribution with parameter ϵdt , thus giving an expected distance traveled of ϵdt per timestep (recall that, for a Poisson distribution with parameter ϵdt , the resulting random variable corresponds to the number of occurrences in a time frame dt). If this movement would cause the RNAP molecule to overlap with another RNAP molecule, then no action is taken. Otherwise, the RNAP molecule moves forward the number of steps given by the generated random variable.

If no RNAP molecule on the gene is chosen ($i = 0$), an RNAP molecule is loaded using a probability parameterized by the term βdt , only if no already existing RNAP molecules overlap with the footprint of the new RNAP molecule. If such an overlap occurs, then no action is taken. Otherwise, to calculate the probability of loading, a random number is drawn from a Poisson distribution with parameter βdt . If this number is one or higher, then the loading event is considered a success. The process is repeated until a total simulation time T has elapsed.

To simulate potential single-molecule variability, each RNAP molecule can possess a different stepping rate ϵ . For a given RNAP molecule i , its stochastic stepping rate ϵ_i is drawn from a truncated normal distribution Tr with mean μ_ϵ and standard deviation σ_ϵ and lower and upper limits 1 and infinity bp/sec, respectively

$$\epsilon_i = Tr(\epsilon, \sigma_\epsilon, 0, \infty). \quad (\text{B.31})$$

Once the position of the active site of an RNAP molecule exceeds that of the total number of sites N_{sites} , i.e. the molecule reaches the end of the gene, it is removed from the simulation after the cleavage time τ elapses..

Parameter	Description	Value
T	total simulation time	600 sec
dt	simulation timestep	0.5 sec
N_{sites}	size of lattice	6626 bp
$N_{footprint}$	RNAP footprint (Selby et al., 1997)	40 bp
μ_β	mean loading rate	0.17 sec^{-1}
σ_β	standard deviation of loading rate	0.05 sec^{-1}
μ_τ	mean cleavage time	2.5 min
σ_τ	standard deviation of cleavage time	1.6 min
μ_ϵ	mean elongation rate	free parameter
σ_ϵ	standard deviation of elongation rate	free parameter

Table B.3: Parameters used in single-molecule Monte Carlo simulation of elongation rates.

Finally, to account for single-cell variability in the transcription initiation rate, the loading rate β and cleavage time τ were allowed to vary across each simulated cell j by drawing these magnitudes from a Gaussian distribution with parameters reflecting the actual data. Since *hunchback* is known to load new nascent RNA transcripts at a rate of 1 molecule every 6 seconds in the anterior of the embryo (Garcia et al., 2013), we thus chose the mean of this distribution μ_β to be 1 molecule/6 s = 0.17 s^{-1} . The standard deviation σ_β was chosen to be this mean multiplied by the CV in the initiation rate in the anterior inferred in the main text, resulting in a value of 0.05 s^{-1} . Thus, for simulated cell j

$$\beta_j = N(\mu_\beta, \sigma_\beta), \tag{B.32}$$

where any negative value was replaced with zero.

Similarly, the cleavage time τ_j for each simulated cell was drawn from a Gaussian distribution with mean $\mu_\tau = 2.5 \text{ min}$ and standard deviation $\sigma_\tau = 1.6 \text{ min}$. These values were obtained from the distribution of inferred cleavage times in the anterior of the embryo. The values of each simulation parameter are summarized in Table B.3.

From these simulations, the positions of each RNAP molecule on the gene as a function of time were saved and then fed into the model of the reporter gene (Section B.1), producing simulated single-cell MS2 and PP7 fluorescence traces (Fig. B.10B). Simulated fluorescence noise was added using the same parameters as in the validation simulations discussed earlier (Section B.5, Table B.1, and Fig. B.5). These fluorescence traces were then run through the inference pipeline (Section B.4.1), resulting in inferred distributions of single-cell mean elongation rates from the single-molecule elongation simulation.

In order to compare these results with the empirically inferred distribution of elongation rates (Fig. 4.4D, red), we first considered a scenario where the single-molecule variability in

stepping rates σ_ϵ was fixed at zero and the mean stepping rate μ_ϵ was varied from 0.6 to 2.1 kb/min. While the combination of exclusionary interactions between RNAP molecules, stochasticity in single-molecule stepping, and inferential noise did produce some cell-to-cell variability (Fig. B.10C, top row), the resulting distributions nevertheless were unable to reproduce the large variance observed in the data. This can be seen by plotting the mean and variance of the simulated distributions (Fig. B.10D, blue), where we see that the variance in the case of $\sigma_\epsilon = 0$ is always below that of the data (Fig. B.10D, purple).

Next, we allowed σ_ϵ to vary, simulating small to moderate variability with values of $\sigma_\epsilon = 0.3$ kb/min and $\sigma_\epsilon = 0.6$ kb/min. As expected, this single-molecule variability caused the inferred single-cell elongation rate distributions to widen (Fig. B.10C, middle and bottom rows). In the presence of this variability, there existed parameter sets where the mean and variance of the simulated distributions quantitatively matched the empirical distribution within error (Fig. B.10D, red and gold).

The distributions presented in the main text correspond to the following parameter values. For the case with no molecular variability in elongation rates (Fig. 4.4D, brown), we used $\mu_\epsilon = 0.9$ kb/min and $\sigma_\epsilon = 0$ kb/min, chosen as the simulated parameter set with results closest to the inferred mean and variance of empirical elongation rates (Fig. B.10D, lower black arrow). For the case with molecular variability in elongation rates (Fig. 4.4D, gold), we used $\mu_\epsilon = 0.9$ kb/min and $\sigma_\epsilon = 0.3$ kb/min, chosen as a representative example of a simulation possessing a mean and variance in elongation rate that agreed with the inferred mean and variance of empirical elongation rates within error (Fig. B.10D, upper black arrow), as well as qualitatively agreeing with the inferred distribution (Fig. 4.4D, gold).

B.11 Single-cell correlation analysis using full posterior distributions

The single-cell inter-parameter correlations presented in the main text (Fig. 4.5) were based off of mean values from the posterior distributions obtained from the inference procedure for ease of interpretation and visualization. In principle, these correlations could possess high amounts of uncertainty due to uncertainty in the single-cell parameter estimates. Here, we conduct a correlation analysis based on the full posterior distributions from the inference and validate the mean results presented in the main text.

To do so, we used a Monte Carlo simulation to construct a distribution of Spearman correlation coefficients and investigated if the mean Spearman correlation coefficients presented in Fig. 4.5 agreed with these simulated distributions.

First, we extracted the mean and variance of the inferred posterior distribution obtained from each single cell, for each transcriptional parameter (Fig. 4.2C and E). We then simulated $N = 50,000$ new single-cell datasets comprising the mean initiation rate, elongation rate, and cleavage time, where these values were generated from Gaussian distributions parameterized by the means and variances from each parameter's posterior distribution at the single-cell

level.

Thus, each of the $N = 50,000$ simulations resulted in a simulated dataset of $n = 355$ cells with randomly generated transcriptional parameter values obtained from the information inside the single-cell inferred posterior distributions from the experimental data. We then calculated an individual Spearman correlation coefficient and associated p-value for each simulation, generating an $N = 50,000$ distribution for each correlation relationship.

Figure B.11A and B show the ensuing distribution of p-values for the Spearman correlation coefficient between the mean initiation rate and elongation rate, as well as between the elongation rate and cleavage time, respectively. The p-values for the relationships between the mean initiation rate and cleavage time and between the mean RNAP density and cleavage time were essentially zero due to floating point error. Thus, the distributions of p-values for all four inter-parameter relationships were extremely small and support the statistical significance of their associated correlations.

Figure B.11C shows the simulated distributions of Spearman correlation coefficients for all four relationships (histograms), along with the values obtained from the simpler mean analysis presented in the main text (dashed lines). We see that using the full posterior via this Monte Carlo simulation yields distributions that are in agreement with the results from the mean analysis, and that the distributions themselves are narrow, with widths of around 0.05. As a result, the correlations obtained from utilizing only mean inferred parameters quantitatively agree with the results obtained from utilizing the full Bayesian posterior obtained from the MCMC inference procedure.

Thus, our original analysis is robust, and we chose to retain its presentation in the main text for simplicity and ease of understanding.

B.12 Supplementary Videos

- B.12.1. **Video 1.** Measurement of main reporter construct. Movie of P2P-MS2-lacZ-PP7 reporter construct used in an embryo in nuclear cycle 14. Fluorescence intensities are maximum projections in the z-plane. Time is defined with respect to the previous anaphase. (https://www.dropbox.com/s/l9oiwjgl4hx3uq/Video_1.avi?dl=0)
- B.12.2. **Video 2.** Measurement of interlaced reporter construct. Movie of P2P-24x(MS2/PP7) reporter construct used in an embryo in nuclear cycle 14. Fluorescence intensities are maximum projections in the z-plane. Time is defined with respect to the previous anaphase. (https://www.dropbox.com/s/5xhtobnzjjac20g/Video_2.avi?dl=0)

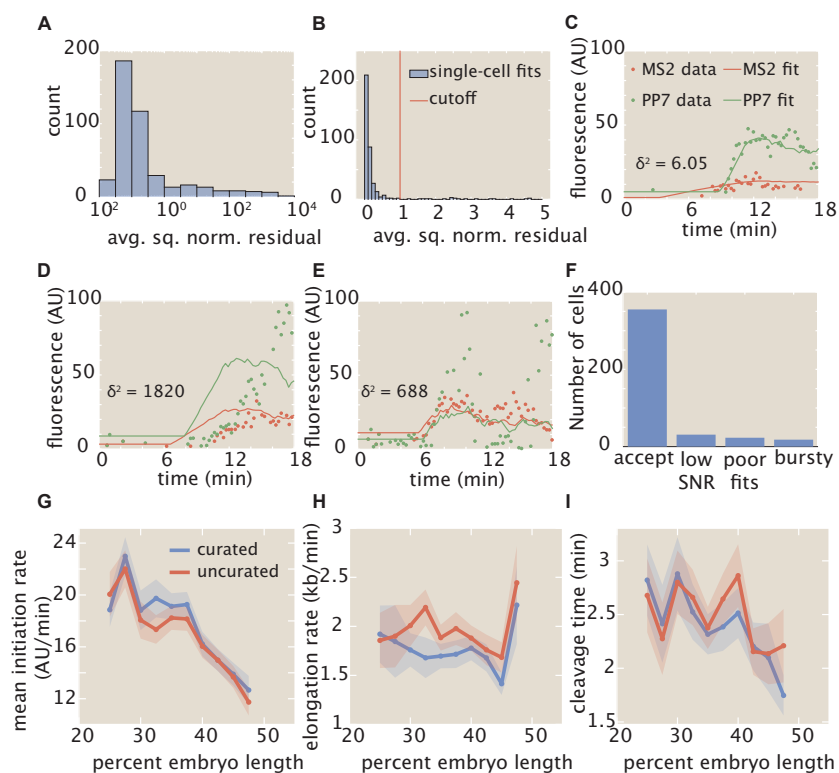


Figure B.4: Automated curation of data. (A, B) Histograms (blue) of average squared normalized residual of single-cell fits, in log (A) and linear (B) scale, with cutoff of $\delta_{cutoff}^2 = 1$ shown in red in (B). (C) Example of bad fit from poor signal-to-noise ratio (SNR). (D) Example of bad fit of otherwise reasonable data from issues in fitting algorithm, for example due to local minima. (E) Example of bad fit due to potential presence of substantial bursting of promoter. (F) Number of single cell fits in each class of rejected fit, along with number of accepted fits, after the initial filtering based on number of time points. Altogether, 84% of filtered fits were accepted. The percentages of filtered fits in the three rejected categories (low SNR, poor fits, bursting) were 7%, 5%, and 4%, respectively. The data shown in C-E are in each fluorophore's intrinsic arbitrary unit without rescaling, to present the fluorescence intensities in their raw form. (G, H, I) Comparison of average inferred (G) mean initiation rate, (H) elongation rate, and (I) cleavage time as a function of embryo position, between curated (blue) and uncurated (red) datasets. Values of δ^2 were 6.05, 1820, and 688 for the example fits shown in C-E, respectively, here given to illustrate the qualitative correspondence of δ^2 as a metric with the overall goodness-of-fit. Shading in G-I represent standard error of the mean for 355 and 427 cells across 7 embryos for curated and uncurated datasets, respectively.

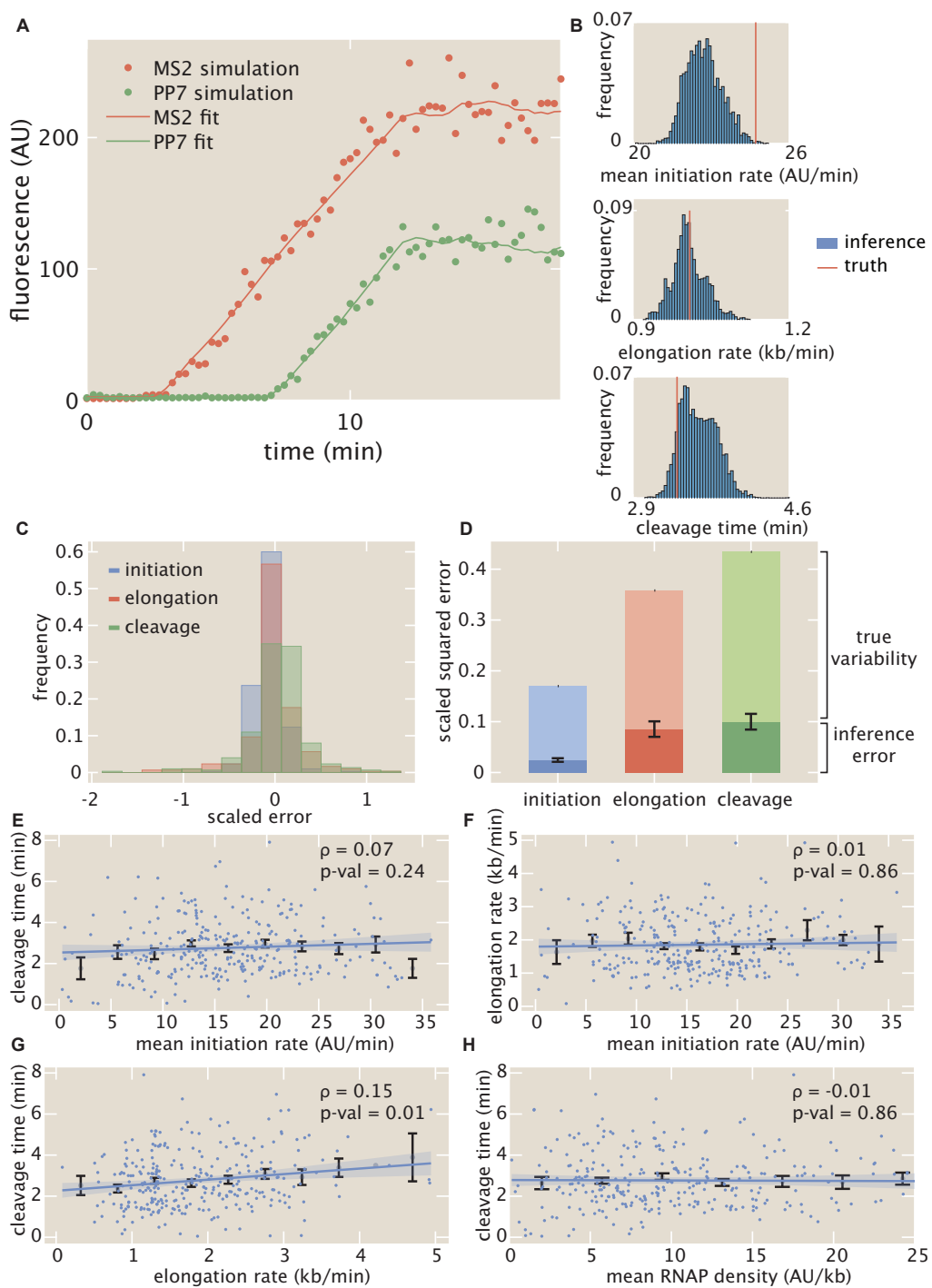


Figure B.5: Overview of MCMC inference validation. See caption on next page.

Figure B.5: Overview of MCMC inference validation. (A) Example single-cell simulated data and inferred fits. (B) MCMC inference results for the simulated data in (A) for the mean initiation rate, elongation rate, and cleavage time. The histogram represents the raw MCMC sampled values, and the red line is the ground truth for this particular cell. The mean value of each histogram is then retained for further statistical analysis. (C) Scaled error of initiation, elongation, and cleavage for each simulated cell. (D) Comparison of relative magnitudes of random inference error and true experimental variability for the initiation, elongation, and cleavage parameters. (E, F, G, H) Single-cell correlations along with Spearman correlation coefficients and p-values for simulated data between (E) mean initiation rate and cleavage time, (F) mean initiation rate and elongation rate, (G) elongation rate and cleavage time, and (H) mean RNAP density and cleavage time, respectively. Blue points indicate single-cell values; black points and error bars indicate mean and SEM, respectively, binned across x-axis values. Line and shaded region indicate generalized linear model fit and 95% confidence interval, respectively. Linear fits were calculated using a generalized linear regression model and are presented for ease of visualization (see Methods and Materials for details).

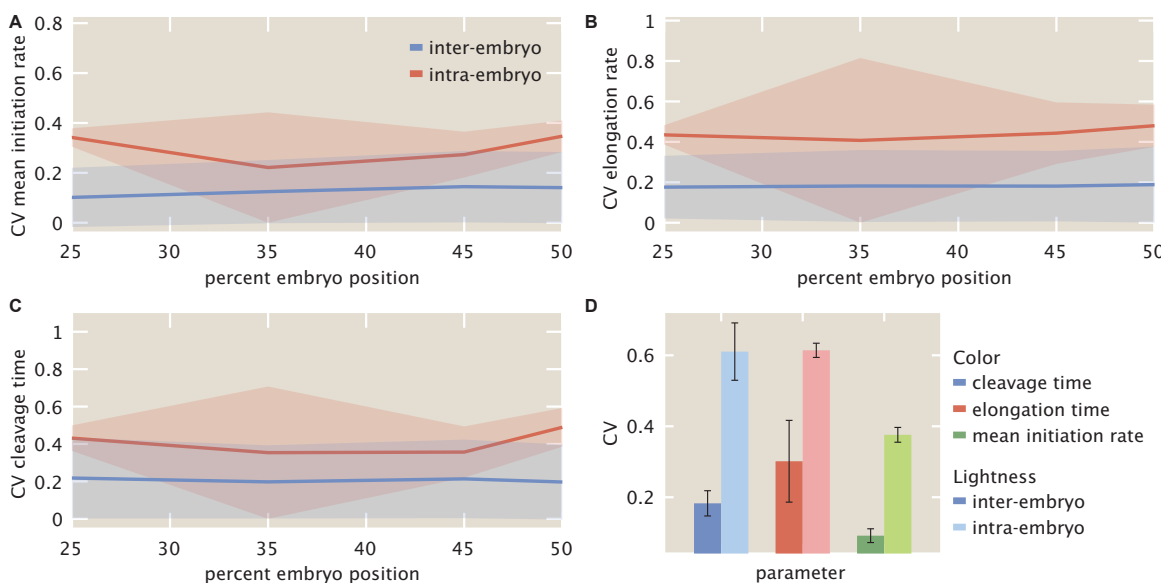


Figure B.6: Comparison of intra- and inter-embryo variability for inferred (A) mean initiation rates, (B) elongation rates, and (C) cleavage times, as a function of embryo position. (D) Intra- and inter-embryo variability for transcriptional parameters averaged across all embryo positions. (A-C, lines and shaded regions indicate mean and standard error of the mean, respectively; D, error bars indicate bootstrapped standard error error across 100 bootstrap samples. Data were taken over 355 cells across 7 embryos, with approximately 10-90 cells per embryo in the region of the embryo examined here.)

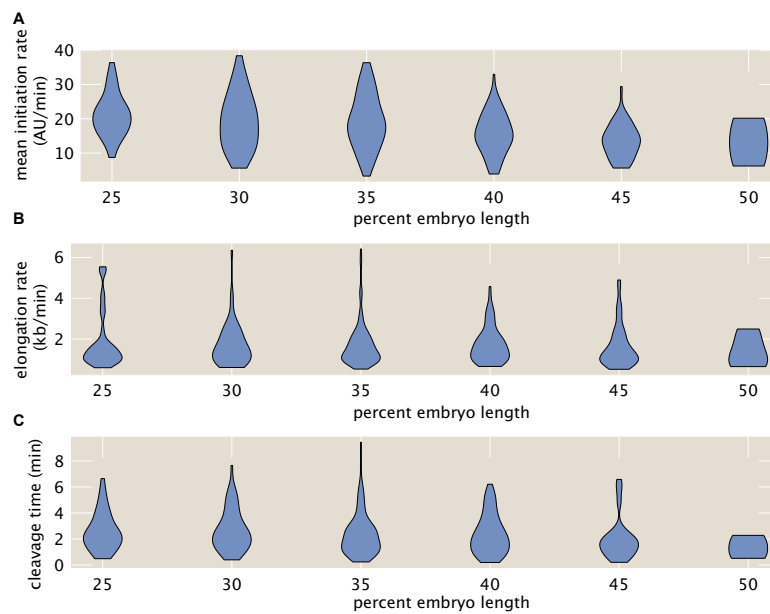


Figure B.7: Single cell distributions of inferred parameters. (A-C) Full single-cell distributions of (A) mean initiation rate, (B) elongation rate, and (C) cleavage time as a function of embryo position.

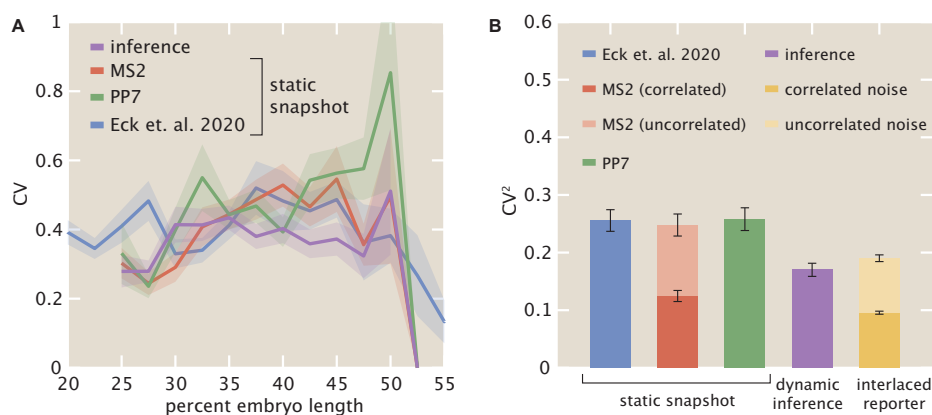


Figure B.8: Comparison of coefficients of variation (CV) between inferred mean initiation rates and instantaneous counts of number of nascent RNA transcripts. (A) Position-dependent CV of inferred mean initiation rate (purple) compared with static measurements of MS2 and PP7 raw fluorescence (red, green) from the dual-color reporter (Fig. 4.1C), as well as with static measurements of MS2 data from Eck et al. (2020) (blue). (B) Position-averaged squared CVs of the same measurements, where the entire dataset is treated as a single sample and embryo position information is disregarded. In addition, separation of uncorrelated and correlated sources of variability are shown, calculated using the reporter described in Fig. 4.3B. (A, Shaded regions indicate bootstrapped standard error of the mean; B, error bars indicate bootstrapped standard error of the mean for $n = 100$ bootstrap samples.)

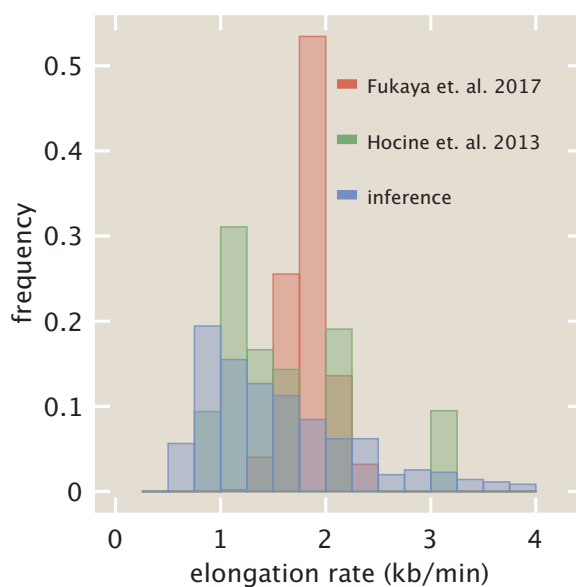


Figure B.9: Comparison of distribution of elongation rates (green) with previous studies (Hocine et al. (2013), red and Fukaya et al. (2017), blue). Distributions of previous studies were adapted from Figs. 2D and 2A of Hocine et al. (2013) and Fukaya et al. (2017), respectively.

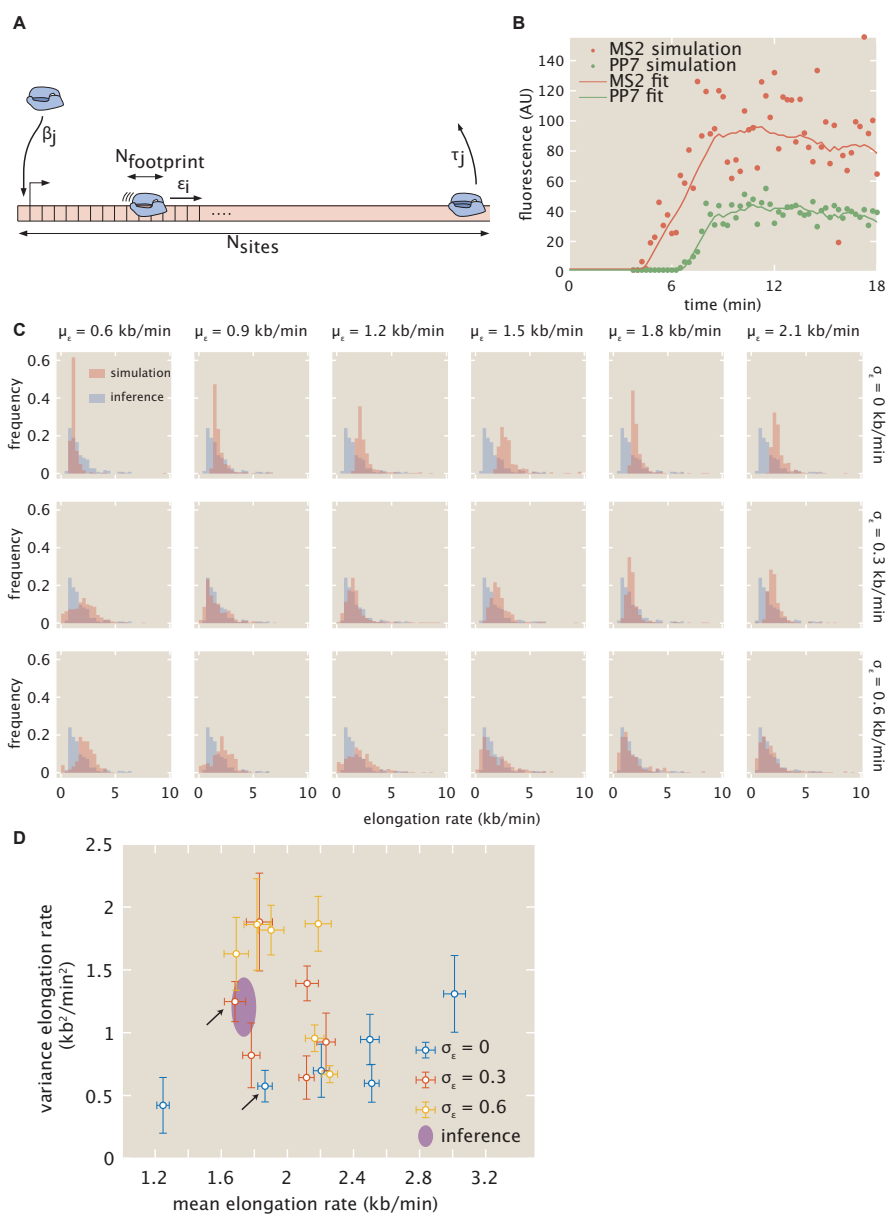


Figure B.10: Single-molecule simulations of elongation dynamics require molecular variability to describe empirical distributions. See caption on next page.

Figure B.10: Single-molecule simulations of elongation dynamics require molecular variability to describe empirical distributions. (A) Cartoon overview of simulation. RNAP molecules with footprint $N_{footprint}$ stochastically advance along a one-dimensional gene represented as a lattice with N_{sites} unique sites, with each site equivalent to a single base pair. Each RNAP molecule i possesses an intrinsic stepping rate ϵ_i , and each cell j stochastically loads new RNAP molecules at the promoter with rate β_j and cleaves finished RNAP molecules after a cleavage time τ_j . (B) Sample simulated MS2 and PP7 fluorescence traces for a single cell, using the single-molecule simulation with parameters $\mu_\epsilon = 1.8$ kb/min and $\sigma_\epsilon = 0$ kb/min, along with inferred fits. (C) Simulated distributions of elongation rates (red) for varying values of μ_ϵ and σ_ϵ , compared with inferred empirical distribution from data (blue). (D) Mean and variance of simulated and empirical distributions of elongation rates for varying values of μ_ϵ and σ_ϵ . Without enough variability in the elongation rate of individual RNAP molecules (blue), the single-molecule model cannot produce the variance observed in the data (purple). However, in the presence of enough molecular variability, the empirical distribution's mean and variance can be reproduced for certain parameter sets (red and gold). Black arrows correspond to parameter sets used for simulated distributions presented in the main text (Fig. 4.4D).

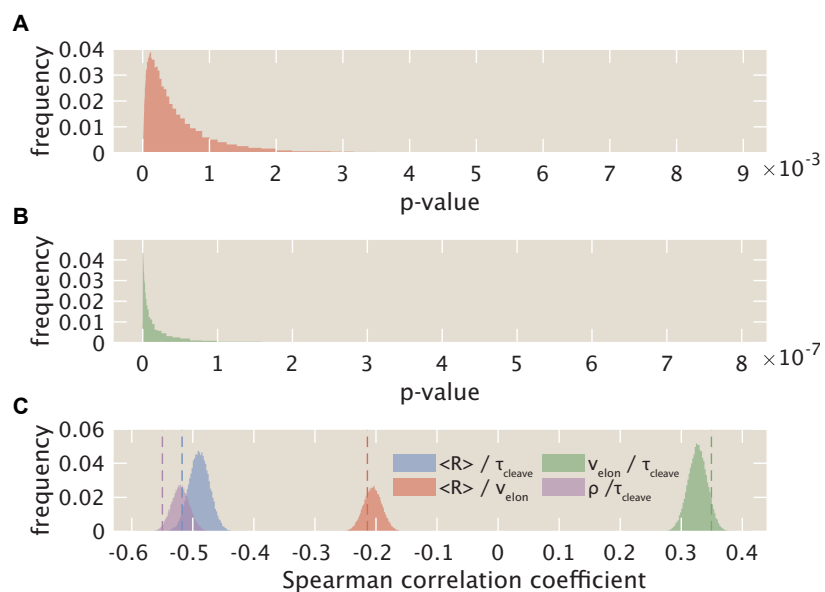


Figure B.11: Monte Carlo simulation of error in single-cell analysis. (A, B) p-values of Spearman correlation coefficient for relationships between mean initiation rate and elongation rate (A) and between elongation rate and cleavage time (B). The p-values for the relationships between mean initiation rate and cleavage time as well as between mean RNAP density and cleavage time were essentially zero due to floating point error. (C). Distributions of Spearman correlation coefficients between mean initiation rate and cleavage time (blue), mean initiation rate and elongation rate (red), elongation rate and cleavage time (green), and mean RNAP density and cleavage time (purple). Results from mean-level analysis (Fig. 4.5) are shown in dashed lines.

Bibliography

- G. K. Ackers, A. D. Johnson, and M. A. Shea. Quantitative model for gene regulation by lambda phage repressor. *Proc Natl Acad Sci U S A*, 79(4):1129–33, 1982. ISSN 0027-8424 (Print).
- C C Adams and J L Workman. Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Molecular and Cellular Biology*, 15(3):1405–1421, mar 1995. ISSN 0270-7306. doi: 10.1128/mcb.15.3.1405.
- T. Ahsendorf, F. Wong, R. Eils, and J. Gunawardena. A framework for modelling gene regulation which accommodates non-equilibrium mechanisms. *BMC Biol*, 12:102, 2014. ISSN 1741-7007 (Electronic) 1741-7007 (Linking). doi: 10.1186/s12915-014-0102-4.
- Cem Albayrak, Christian A. Jordi, Christoph Zechner, Jing Lin, Colette A. Bichsel, Mustafa Khammash, and Savaş Tay. Digital Quantification of Proteins and mRNA in Single Mammalian Cells. *Molecular Cell*, 61(6):914–924, 2016. ISSN 10974164. doi: 10.1016/j.molcel.2016.02.030.
- Bruce Alberts. *Molecular biology of the cell*. Garland Science, Taylor and Francis Group, New York, NY, sixth edition. edition, 2015. ISBN 9780815344322 (hardcover) 0815344325 (hardcover) 9780815344643 (paperback) 0815344643 (paperback) 9780815345244 (looseleaf) 0815345240 (looseleaf).
- Md Zulfikar Ali, Sandeep Choubey, Dipjyoti Das, and Robert C. Brewster. Probing Mechanisms of Transcription Elongation Through Cell-to-Cell Variability of RNA Polymerase. *Biophysical Journal*, 118(7):1769–1781, 2020. ISSN 15420086. doi: 10.1016/j.bpj.2020.02.002. URL <https://doi.org/10.1016/j.bpj.2020.02.002>.
- M. B. Ardehali and J. T. Lis. Tracking rates of transcription and splicing in vivo. *Nat Struct Mol Biol*, 16(11):1123–4, 2009. ISSN 1545-9985 (Electronic) 1545-9985 (Linking). doi: 10.1038/nsmb1109-1123.
- Ignacio Arganda-Carreras, Verena Kaynig, Curtis Rueden, Kevin W. Eliceiri, Johannes Schindelin, Albert Cardona, and H. Sebastian Seung. Trainable Weka Segmentation: A machine learning tool for microscopy pixel classification. *Bioinformatics*, 33(15):2424–2426, 2017. ISSN 14602059. doi: 10.1093/bioinformatics/btx180.

- L. Bai, A. Ondracka, and F. R. Cross. Multiple sequence-specific factors generate the nucleosome-depleted region on *cln2* promoter. *Mol Cell*, 42(4):465–76, 2011. ISSN 1097-4164 (Electronic) 1097-2765 (Linking). doi: 10.1016/j.molcel.2011.03.028.
- Lu Bai, Thomas J. Santangelo, and Michelle D. Wang. Single-molecule analysis of RNA polymerase transcription. *Annual Review of Biophysics and Biomolecular Structure*, 35: 343–360, 2006. ISSN 10568700. doi: 10.1146/annurev.biophys.35.010406.150153.
- A. Bakk, R. Metzler, and K. Sneppen. Sensitivity of σ in phage lambda. *Biophys J*, 86(1 Pt 1):58–66, 2004.
- E. Bakker and P. S. Swain. Estimating numbers of intracellular molecules through analysing fluctuations in photobleaching. *Sci Rep*, 9(1):15238, 2019. ISSN 2045-2322 (Electronic) 2045-2322 (Linking). doi: 10.1038/s41598-019-50921-7.
- C. R. Bartman, S. C. Hsu, C. C. Hsiung, A. Raj, and G. A. Blobel. Enhancer regulation of transcriptional bursting parameters revealed by forced chromatin looping. *Mol Cell*, 62(2): 237–247, 2016. ISSN 1097-4164 (Electronic) 1097-2765 (Linking). doi: 10.1016/j.molcel.2016.03.007.
- J. R. Bateman, A. M. Lee, and C. T. Wu. Site-specific transformation of drosophila via *phic31* integrase-mediated cassette exchange. *Genetics*, 173(2):769–77, 2006. ISSN 0016-6731 (Print) 0016-6731 (Linking). doi: genetics.106.056945[pii]10.1534/genetics.106.056945.
- Eric Batsché, Moshe Yaniv, and Christian Muchardt. The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nature Structural and Molecular Biology*, 13(1):22–29, 2006. ISSN 15459993. doi: 10.1038/nsmb1030.
- D. L. Bentley. Coupling mrna processing with transcription in time and space. *Nat Rev Genet*, 15(3):163–75, 2014. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi: 10.1038/nrg3662.
- H. C. Berg and D. A. Brown. Chemotaxis in escherichia coli analysed by three-dimensional tracking. *Nature*, 239(5374):500–4, 1972. ISSN 0028-0836 (Print) 0028-0836 (Linking).
- H. C. Berg and E. M. Purcell. Physics of chemoreception. *Biophys J*, 20(2):193–219, 1977. ISSN 0006-3495 (Print) 0006-3495 (Linking). doi: S0006-3495(77)85544-6[pii]10.1016/S0006-3495(77)85544-6.
- A. Berrocal, N. C. Lammers, H. G. Garcia, and M. B. Eisen. Kinetic sculpting of the seven stripes of the drosophila even-skipped gene. *Elife*, 9, 2020. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.61635.
- E. Bertrand, P. Chartrand, M. Schaefer, S. M. Shenoy, R. H. Singer, and R. M. Long. Localization of *ash1* mrna particles in living yeast. *Mol Cell*, 2(4):437–45, 1998. ISSN 1097-2765 (Print) 1097-2765 (Linking). doi: S1097-2765(00)80143-4[pii].

- L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips. Transcriptional regulation by the numbers: applications. *Curr Opin Genet Dev*, 15(2):125–35, 2005a.
- L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips. Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev*, 15(2):116–24, 2005b.
- S. A. Blythe and E. F. Wieschaus. Establishment and maintenance of heritable chromatin structure during early drosophila embryogenesis. *Elife*, 5, 2016. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.20148.
- A. N. Boettiger and M. Levine. Rapid transcription fosters coordinate snail expression in the drosophila embryo. *Cell Rep*, 3(1):8–15, 2013. ISSN 2211-1247 (Electronic). doi: 10.1016/j.celrep.2012.12.015.
- Alistair N. Boettiger, Peter L. Ralph, and Steven N. Evans. Transcriptional Regulation: Effects of Promoter Proximal Pausing on Speed, Synchrony and Reliability. *PLoS Computational Biology*, 7(5), 2011. ISSN 1553734X. doi: 10.1371/journal.pcbi.1001136.
- S. Boireau, P. Maiuri, E. Basyuk, M. de la Mata, A. Knezevich, B. Pradet-Balade, V. Backer, A. Kornblihtt, A. Marcello, and E. Bertrand. The transcriptional cycle of hiv-1 in real-time and live cells. *J Cell Biol*, 179(2):291–304, 2007. ISSN 0021-9525 (Print) 0021-9525 (Linking). doi: 10.1083/jcb.200706018.
- H. Bolouri and E. H. Davidson. Transcriptional regulatory cascades in development: initial rates, not steady state, determine network kinetics. *Proc Natl Acad Sci U S A*, 100(16): 9371–6, 2003. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi: 10.1073/pnas.1533293100.
- J. P. Bothma, H. G. Garcia, E. Esposito, G. Schlissel, T. Gregor, and M. Levine. Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living drosophila embryos. *Proc Natl Acad Sci U S A*, 111(29):10598–10603, 2014. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1410022111.
- J. P. Bothma, H. G. Garcia, S. Ng, M. W. Perry, T. Gregor, and M. Levine. Enhancer additivity and non-additivity are determined by enhancer strength in the drosophila embryo. *Elife*, 4:e07956, 2015. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.07956.
- Jonathan R. Bowles, Caroline Hoppe, Hilary L. Ashe, and Magnus Rattray. Scalable inference of transcriptional kinetic parameters from ms2 time series data. *bioRxiv*, page 2020.12.04.412049, 2020. doi: 10.1101/2020.12.04.412049.
- R. C. Brewster, F. M. Weinert, H. G. Garcia, D. Song, M. Rydenfelt, and R. Phillips. The transcription factor titration effect dictates level of gene expression. *Cell*, 156(6):1312–23, 2014. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2014.02.022.

- Robert C. Brewster, Daniel L. Jones, and Rob Phillips. Tuning Promoter Strength through RNA Polymerase Binding Site Design in *Escherichia coli*. *PLoS Computational Biology*, 8(12), dec 2012. ISSN 1553734X. doi: 10.1371/journal.pcbi.1002811.
- N. E. Buchler, U. Gerland, and T. Hwa. On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A*, 100(9):5136–41, 2003. ISSN 0027-8424 (Print).
- L. Cai, N. Friedman, and X. S. Xie. Stochastic protein expression in individual cells at the single molecule level. *Nature*, 440(7082):358–62, 2006. ISSN 1476-4687 (Electronic).
- Fernando Carrillo Oesterreich, Stephan Preibisch, and Karla M. Neugebauer. Global analysis of nascent rna reveals transcriptional pausing in terminal exons. *Molecular Cell*, 40(4): 571–581, 2010. ISSN 10972765. doi: 10.1016/j.molcel.2010.11.004. URL <http://dx.doi.org/10.1016/j.molcel.2010.11.004>.
- J. A. Chao, Y. Patskovsky, S. C. Almo, and R. H. Singer. Structural basis for the coevolution of a viral rna-protein complex. *Nat Struct Mol Biol*, 15(1):103–5, 2008. ISSN 1545-9985 (Electronic) 1545-9985 (Linking). doi: nsmb1327[pii]10.1038/nsmb1327.
- B. C. Chen, W. R. Legant, K. Wang, L. Shao, D. E. Milkie, M. W. Davidson, C. Janetopoulos, X. S. Wu, 3rd Hammer, J. A., Z. Liu, B. P. English, Y. Mimori-Kiyosue, D. P. Romero, A. T. Ritter, J. Lippincott-Schwartz, L. Fritz-Laylin, R. D. Mullins, D. M. Mitchell, J. N. Bembenek, A. C. Reymann, R. Bohme, S. W. Grill, J. T. Wang, G. Seydoux, U. S. Tulu, D. P. Kiehart, and E. Betzig. Lattice light-sheet microscopy: imaging molecules to embryos at high spatiotemporal resolution. *Science*, 346(6208):1257998, 2014. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1257998.
- H. Chen, Z. Xu, C. Mei, D. Yu, and S. Small. A system of repressor gradients spatially organizes the boundaries of bicoid-dependent target genes. *Cell*, 149(3):618–29, 2012. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2012.03.018.
- H. Chen, M. Levo, L. Barinov, M. Fujioka, J. B. Jaynes, and T. Gregor. Dynamic interplay between enhancer-promoter topology and gene activity. *Nat Genet*, 50(9):1296–1303, 2018. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi: 10.1038/s41588-018-0175-z.
- Yan Chen, David Chafin, David H. Price, and Arno L. Greenleaf. *Drosophila* RNA polymerase II mutants that affect transcription elongation. *Journal of Biological Chemistry*, 271(11): 5993–5999, 1996. ISSN 00219258. doi: 10.1074/jbc.271.11.5993.
- Myron Barber Child, Jack R. Bateman, Amir Jahangiri, Armando Reimer, Nicholas C. Lammers, Nica Sabouni, Diego Villamarin, Grace C. McKenzie-Smith, Justine E. Johnson, Daniel Jost, and Hernan G. Garcia. Live imaging and biophysical modeling support a button-based mechanism of somatic homolog pairing in *drosophila*. *bioRxiv*, page 265108, 2020. doi: 10.1101/2020.08.30.265108.

- S. Choubey, J. Kondev, and A. Sanchez. Deciphering transcriptional dynamics in vivo by counting nascent rna molecules. *PLoS Comput Biol*, 11(11):e1004345, 2015. ISSN 1553-7358 (Electronic) 1553-734X (Linking). doi: 10.1371/journal.pcbi.1004345.
- Sandeep Choubey, Jane Kondev, and Alvaro Sanchez. Distribution of Initiation Times Reveals Mechanisms of Transcriptional Regulation in Single Cells. *Biophysical Journal*, 114(9):2072–2082, 2018. ISSN 00063495. doi: 10.1016/j.bpj.2018.03.031. URL <http://linkinghub.elsevier.com/retrieve/pii/S0006349518304077>.
- J. R. Chubb, T. Trcek, S. M. Shenoy, and R. H. Singer. Transcriptional pulsing of a developmental gene. *Curr Biol*, 16(10):1018–25, 2006. ISSN 0960-9822 (Print) 0960-9822 (Linking). doi: S0960-9822(06)01426-6[pii]10.1016/j.cub.2006.03.092.
- L. S. Churchman and J. S. Weissman. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330):368–+, 2011. ISSN 0028-0836. doi: Doi10.1038/Nature09652.
- G. Chure, M. Razo-Mejia, N. M. Belliveau, T. Einav, Z. A. Kaczmarek, S. L. Barnes, M. Lewis, and R. Phillips. Predictive shifts in free energy couple mutations to their phenotypic consequences. *Proc Natl Acad Sci U S A*, 116(37):18275–18284, 2019. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1907869116.
- P. A. Combs and M. B. Eisen. Genome-wide measurement of spatial expression in patterning mutants of drosophila melanogaster. *F1000Res*, 6:41, 2017. ISSN 2046-1402 (Print) 2046-1402 (Linking). doi: 10.12688/f1000research.9720.1.
- Leighton J. Core, Joshua J. Waterfall, and John T. Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848, 2008. ISSN 00368075. doi: 10.1126/science.1162228.
- A. M. Corrigan, E. Tunnacliffe, D. Cannon, and J. R. Chubb. A continuum model of transcriptional bursting. *Elife*, 5, 2016. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.13051.
- Adam M. Corrigan and Jonathan R. Chubb. Regulation of transcriptional bursting by a naturally oscillating signal. *Current Biology*, 24(2):205–211, 2014. ISSN 09609822. doi: 10.1016/j.cub.2013.12.011. URL <http://dx.doi.org/10.1016/j.cub.2013.12.011>.
- A. Coulon and D.R. Larson. Fluctuation analysis: Dissecting transcriptional kinetics with signal theory. *Methods in Enzymology*, 2016.
- A. Coulon, C. C. Chow, R. H. Singer, and D. R. Larson. Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nat Rev Genet*, 14(8):572–84, 2013. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi: 10.1038/nrg3484.

- A. Coulon, M. L. Ferguson, V. de Turris, M. Palangat, C. C. Chow, and D. R. Larson. Kinetic competition during the transcription cycle results in stochastic rna processing. *Elife*, 3, 2014. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.03939.
- L. Cui, I. Murchland, K. E. Shearwin, and I. B. Dodd. Enhancer-like long-range transcriptional activation by lambda ci-mediated dna looping. *Proc Natl Acad Sci U S A*, 110(8):2922–7, 2013. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1221322110.
- J. Culkin, L. de Bruin, M. Tompitak, R. Phillips, and H. Schiessel. The role of dna sequence in nucleosome breathing. *Eur Phys J E Soft Matter*, 40(11):106, 2017. ISSN 1292-895X (Electronic) 1292-8941 (Linking). doi: 10.1140/epje/i2017-11596-2.
- Charles G. Danko, Nasun Hah, Xin Luo, André L. Martins, Leighton Core, John T. Lis, Adam Siepel, and W. Lee Kraus. Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells. *Molecular Cell*, 50(2):212–222, 2013. ISSN 10972765. doi: 10.1016/j.molcel.2013.02.015.
- X. Darzacq, Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair, and R. H. Singer. In vivo dynamics of rna polymerase ii transcription. *Nat Struct Mol Biol*, 14(9):796–806, 2007. ISSN 1545-9993 (Print) 1545-9985 (Linking). doi: 10.1038/nsmb1280.
- Manuel De La Mata, Claudio R. Alonso, Sebastián Kadener, Juan P. Fededa, Matías Blaustein, Federico Pelisch, Paula Cramer, David Bentley, and Alberto R. Kornblihtt. A slow RNA polymerase II affects alternative splicing in vivo. *Molecular Cell*, 12(2):525–532, 2003. ISSN 10972765. doi: 10.1016/j.molcel.2003.08.001.
- J. Desponds, H. Tran, T. Ferraro, T. Lucas, C. Perez Romero, A. Guillou, C. Fradin, M. Coppey, N. Dostatni, and A. M. Walczak. Precision of readout at the hunchback gene: Analyzing short transcription time traces in living fly embryos. *PLoS Comput Biol*, 12(12):e1005256, 2016. ISSN 1553-7358 (Electronic) 1553-734X (Linking). doi: 10.1371/journal.pcbi.1005256.
- W. Driever and C. Nusslein-Volhard. A gradient of bicoid protein in drosophila embryos. *Cell*, 54(1):83–93, 1988. ISSN 0092-8674 (Print) 0092-8674 (Linking).
- W. Driever and C. Nusslein-Volhard. The bicoid protein is a positive regulator of hunchback transcription in the early drosophila embryo. *Nature*, 337(6203):138–43, 1989. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi: 10.1038/337138a0.
- W. Driever, G. Thoma, and C. Nusslein-Volhard. Determination of spatial domains of zygotic gene expression in the drosophila embryo by the affinity of binding sites for the bicoid morphogen. *Nature*, 340(6232):363–7, 1989. ISSN 0028-0836 (Print) 0028-0836 (Linking). doi: 10.1038/340363a0.

- J. Dufourt, A. Trullo, J. Hunter, C. Fernandez, J. Lazaro, M. Dejean, L. Morales, S. Nait-Amer, K. N. Schulz, M. M. Harrison, C. Favard, O. Radulescu, and M. Lagha. Temporal control of gene expression by the pioneer factor zelda through transient interactions in hubs. *Nat Commun*, 9(1):5194, 2018. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/s41467-018-07613-z.
- E. Eck, J. Liu, M. Kazemzadeh-Atoufi, S. Ghoreishi, S. A. Blythe, and H. G. Garcia. Quantitative dissection of transcription in development yields evidence for transcription factor-driven chromatin accessibility. *Elife*, 9, 2020. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.56429.
- B. A. Edgar and G. Schubiger. Parameters controlling transcriptional activation during early drosophila development. *Cell*, 44(6):871–7, 1986. ISSN 0092-8674 (Print) 0092-8674 (Linking). doi: 0092-8674(86)90009-7[pil].
- B. A. Edgar, G. M. Odell, and G. Schubiger. Cytoarchitecture and the patterning of fushi tarazu expression in the drosophila blastoderm. *Genes Dev*, 1(10):1226–37, 1987. ISSN 0890-9369 (Print) 0890-9369 (Linking).
- Belal El Kaderi, Scott Medler, Sarita Raghunayakula, and Athar Ansari. Gene looping is conferred by activator-dependent interaction of transcription initiation and termination machineries. *Journal of Biological Chemistry*, 284(37):25015–25025, 2009. ISSN 00219258. doi: 10.1074/jbc.M109.007948.
- M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–6, 2002.
- D. Endy. Foundations for engineering biology. *Nature*, 438(7067):449–53, 2005. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: nature04342[pil]10.1038/nature04342.
- J. Estrada, F. Wong, A. DePace, and J. Gunawardena. Information integration and energy expenditure in gene regulation. *Cell*, 166(1):234–44, 2016. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2016.06.012.
- W. D. Fakhouri, A. Ay, R. Sayal, J. Dresch, E. Dayringer, and D. N. Arnosti. Deciphering a transcriptional regulatory code: modeling short-range repression in the drosophila embryo. *Mol Syst Biol*, 6:341, 2010. ISSN 1744-4292 (Electronic) 1744-4292 (Linking). doi: msb200997[pil]10.1038/msb.2009.97.
- J. Faló-Sanjuan, N. C. Lammers, H. G. Garcia, and S. J. Bray. Enhancer priming enables fast and sustained transcriptional responses to notch signaling. *Dev Cell*, 50(4):411, 2019. ISSN 1878-1551 (Electronic) 1534-5807 (Linking). doi: 10.1016/j.devcel.2019.07.002.
- Jingyi Fei, Digvijay Singh, Qiucen Zhang, Seongjin Park, Divya Balasubramanian, Ido Golding, Carin K. Vanderpool, and Taekjip Ha. Determination of in vivo target search

- kinetics of regulatory noncoding RNA. *Science*, 347(6228):1371–1374, 2015. ISSN 10959203. doi: 10.1126/science.1258849.
- A. M. Femino, F. S. Fay, K. Fogarty, and R. H. Singer. Visualization of single rna transcripts in situ. *Science*, 280(5363):585–90, 1998. ISSN 0036-8075 (Print) 0036-8075 (Linking).
- T. Ferraro, T. Lucas, M. Clemot, J. De Las Heras Chanes, J. Desponds, M. Coppey, A. M. Walczak, and N. Dostatni. New methods to image transcription in living fly embryos: the insights so far, and the prospects. *Wiley Interdiscip Rev Dev Biol*, 5(3):296–310, 2016. ISSN 1759-7692 (Electronic). doi: 10.1002/wdev.221.
- Tatiana Filatova, Nikola Popovic, and Ramon Grima. Statistics of nascent and mature RNA fluctuations in a stochastic model of transcriptional initiation , elongation , pausing , and termination. *bioRxiv*, 2020.
- Bärbel Finkenstädt, Elizabeth A. Heron, Michal Komorowski, Kieron Edwards, Sanyi Tang, Claire V. Harper, Julian R.E. Davis, Michael R.H. White, Andrew J. Millar, and David A. Rand. Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics*, 24(24):2901–2907, 2008. ISSN 13674803. doi: 10.1093/bioinformatics/btn562.
- Nova Fong, Kristopher Brannan, Benjamin Erickson, Hyunmin Kim, Michael A. Cortazar, Ryan M. Sheridan, Tram Nguyen, Shai Karp, and David L. Bentley. Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Molecular Cell*, 60(2):256–267, 2015. ISSN 10974164. doi: 10.1016/j.molcel.2015.09.026.
- S. M. Foo, Y. Sun, B. Lim, R. Ziukaite, K. O’Brien, C. Y. Nien, N. Kirov, S. Y. Shvartsman, and C. A. Rushlow. Zelda potentiates morphogen activity by increasing chromatin accessibility. *Curr Biol*, 24(12):1341–6, 2014. ISSN 1879-0445 (Electronic) 0960-9822 (Linking). doi: 10.1016/j.cub.2014.04.032.
- C. Fritzscher, S. Baumgartner, M. Kuban, D. Steinshorn, G. Reid, and S. Legewie. Estrogen-dependent control and cell-to-cell variability of transcriptional bursting. *Mol Syst Biol*, 14(2): e7678, 2018. ISSN 1744-4292 (Electronic) 1744-4292 (Linking). doi: 10.15252/msb.20177678.
- D. Fu and J. Ma. Interplay between positive and negative activities that influence the role of bicoid in transcription. *Nucleic Acids Res*, 33(13):3985–93, 2005. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gki691.
- D. Fu, Y. Wen, and J. Ma. The co-activator creb-binding protein participates in enhancer-dependent activities of bicoid. *J Biol Chem*, 279(47):48725–33, 2004. ISSN 0021-9258 (Print) 0021-9258 (Linking). doi: 10.1074/jbc.M407066200.

- Gilad Fuchs, Yoav Voichek, Sima Benjamin, Gilad Shlomit, Ido Amit, and Oren Moshe. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biology*, 15(5):R69, 2014. doi: 10.1186/gb-2014-15-5-r69.
- N. J. Fuda, M. B. Ardehali, and J. T. Lis. Defining mechanisms that regulate rna polymerase ii transcription in vivo. *Nature*, 461(7261):186–92, 2009. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature08449.
- T. Fukaya, B. Lim, and M. Levine. Enhancer control of transcriptional bursting. *Cell*, 166(2):358–368, 2016. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2016.05.025.
- T. Fukaya, B. Lim, and M. Levine. Rapid rates of pol ii elongation in the drosophila embryo. *Curr Biol*, 27(9):1387–1391, 2017. ISSN 1879-0445 (Electronic) 0960-9822 (Linking). doi: 10.1016/j.cub.2017.03.069.
- Eden Fussner, Reagan W. Ching, and David P. Bazett-Jones. Living without 30nm chromatin fibers. *Trends in Biochemical Sciences*, 36(1):1–6, jan 2011. ISSN 09680004. doi: 10.1016/j.tibs.2010.09.002.
- Bjoern Gaertner and Julia Zeitlinger. RNA polymerase II pausing during development. pages 1179–1183, 2014. doi: 10.1242/dev.088492.
- H. G. Garcia and R. Phillips. Quantitative dissection of the simple repression input-output function. *Proc Natl Acad Sci U S A*, 108(29):12173–8, 2011. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1015616108.
- H. G. Garcia, A. Sanchez, J. Q. Boedicker, M. Osborne, J. Gelles, J. Kondev, and R. Phillips. Operator sequence alters gene expression independently of transcription factor occupancy in bacteria. *Cell Rep*, 2(1):150–61, 2012. ISSN 2211-1247 (Electronic). doi: 10.1016/j.celrep.2012.06.004.
- H. G. Garcia, M. Tikhonov, A. Lin, and T. Gregor. Quantitative imaging of transcription in living drosophila embryos links polymerase activity to patterning. *Curr Biol*, 23(21):2140–5, 2013. ISSN 1879-0445 (Electronic) 0960-9822 (Linking). doi: 10.1016/j.cub.2013.08.054.
- Hernan G. Garcia and Thomas Gregor. *Live Imaging of mRNA Synthesis in Drosophila*, pages 349–357. Springer New York, New York, NY, 2018.
- Hernan G. Garcia, Jane Kondev, Nigel Orme, Julie A. Theriot, and Rob Phillips. A first exposure to statistical mechanics for life scientists. *arXiv preprint arXiv:0708.1899*, 2007.
- J. Gertz, E. D. Siggia, and B. A. Cohen. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457(7226):215–8, 2009.

- Charles J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992. doi: 10.1097/EDE.0b013e3181.
- D. T. Gillespie. General method for numerically simulating stochastic time evolution of coupled chemical-reactions. *Journal of Computational Physics*, 22(4):403–434, 1976. ISSN 0021-9991.
- L. Giorgetti, T. Siggers, G. Tiana, G. Caprara, S. Notarbartolo, T. Corona, M. Pasparakis, P. Milani, M. L. Bulyk, and G. Natoli. Noncooperative interactions between transcription factors and clustered dna binding sites enable graded transcriptional responses to environmental inputs. *Mol Cell*, 37(3):418–28, 2010. ISSN 1097-4164 (Electronic) 1097-2765 (Linking). doi: 10.1016/j.molcel.2010.01.016.
- I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, 2005. ISSN 0092-8674 (Print).
- M. Gomez-Schiavon, L. F. Chen, A. E. West, and N. E. Buchler. Bayfish: Bayesian inference of transcription dynamics from population snapshots of single-molecule rna fish in single cells. *Genome Biol*, 18(1):164, 2017. ISSN 1474-760X (Electronic) 1474-7596 (Linking). doi: 10.1186/s13059-017-1297-9.
- Stanislaw A. Gorski, Miroslav Dunder, and Tom Misteli. The road much traveled: trafficking in the cell nucleus. *Current Opinion in Cell Biology*, 18(3):284–290, 2006. ISSN 09550674. doi: 10.1016/j.ceb.2006.03.002.
- J. M. Gottesfeld and D. J. Forbes. Mitotic repression of the transcriptional machinery. *Trends Biochem Sci*, 22(6):197–202, 1997. ISSN 0968-0004 (Print) 0968-0004 (Linking).
- T. Gregor, D. W. Tank, E. F. Wieschaus, and W. Bialek. Probing the limits to positional information. *Cell*, 130(1):153–64, 2007a. ISSN 0092-8674 (Print).
- T. Gregor, E. F. Wieschaus, A. P. McGregor, W. Bialek, and D. W. Tank. Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell*, 130(1):141–52, 2007b. ISSN 0092-8674 (Print).
- Sanjana Gupta, Liam Hainsworth, Justin Hogg, Robin Lee, and James Faeder. Evaluation of Parallel Tempering to Accelerate Bayesian Parameter Estimation in Systems Biology. *Proceedings - 26th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, PDP 2018*, pages 690–697, 2018. ISSN 2377-5750. doi: 10.1109/PDP2018.2018.00114.
- Sanjana Gupta, Robin E.C. Lee, and James R. Faeder. Parallel Tempering with Lasso for model reduction in systems biology. *PLoS Computational Biology*, 16(3):1–22, 2020. ISSN 15537358. doi: 10.1371/journal.pcbi.1007669. URL <http://dx.doi.org/10.1371/journal.pcbi.1007669>.

- Heikki Haario, Eero Saksman, and Johanna Tamminen. An Adaptive Metropolis Algorithm. *Bernoulli*, 7(2):223, 2001. ISSN 13507265. doi: 10.2307/3318737. URL <http://www.jstor.org/stable/3318737?origin=crossref>.
- Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. DRAM: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, 2006. ISSN 09603174. doi: 10.1007/s11222-006-9438-0.
- J. E. Haines and M. B. Eisen. Patterns of chromatin accessibility along the anterior-posterior axis in the early drosophila embryo. *PLoS Genet*, 14(5):e1007367, 2018. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 10.1371/journal.pgen.1007367.
- Danielle C. Hamm, Elizabeth D. Larson, Markus Nevil, Kelsey E. Marshall, Eliana R. Bondra, and Melissa M. Harrison. A conserved maternal-specific repressive domain in Zelda revealed by Cas9-mediated mutagenesis in *Drosophila melanogaster*. *PLOS Genetics*, 13(12): e1007120, December 2017. ISSN 1553-7404. doi: 10.1371/journal.pgen.1007120. URL <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007120>.
- P. Hammar, M. Wallden, D. Fange, F. Persson, O. Baltekin, G. Ullman, P. Leroy, and J. Elf. Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. *Nat Genet*, 46(4):405–8, 2014. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi: 10.1038/ng.2905.
- Michael Hampsey, Badri Nath Singh, Athar Ansari, Jean Philippe Lainé, and Shankarling Krishnamurthy. Control of eukaryotic gene expression: Gene loops and transcriptional memory. *Advances in Enzyme Regulation*, 51(1):118–125, 2011. ISSN 00652571. doi: 10.1016/j.advenzreg.2010.10.001. URL <http://dx.doi.org/10.1016/j.advenzreg.2010.10.001>.
- C. E. Hannon, S. A. Blythe, and E. F. Wieschaus. Concentration dependent chromatin states induced by the bicoid morphogen gradient. *Elife*, 6, 2017. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.28275.
- A. S. Hansen and E. K. O’Shea. cis determinants of promoter threshold and activation timescale. *Cell Rep*, 12(8):1226–33, 2015. ISSN 2211-1247 (Electronic). doi: 10.1016/j.celrep.2015.07.035.
- M. M. Harrison, X. Y. Li, T. Kaplan, M. R. Botchan, and M. B. Eisen. Zelda binding in the early drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS Genet*, 7(10):e1002266, 2011. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 10.1371/journal.pgen.1002266.
- Dane Z. Hazelbaker, Sebastian Marquardt, Wiebke Wlotzka, and Stephen Buratowski. Kinetic Competition between RNA Polymerase II and Sen1-Dependent Transcription Termination. *Molecular Cell*, 2013. ISSN 10972765. doi: 10.1016/j.molcel.2012.10.014.

- H. H. He, C. A. Meyer, H. Shin, S. T. Bailey, G. Wei, Q. Wang, Y. Zhang, K. Xu, M. Ni, M. Lupien, P. Mieczkowski, J. D. Lieb, K. Zhao, M. Brown, and X. S. Liu. Nucleosome dynamics define transcriptional enhancers. *Nat Genet*, 42(4):343–7, 2010. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi: 10.1038/ng.545.
- O. Hendy, Jr. Campbell, J., J. D. Weissman, D. R. Larson, and D. S. Singer. Differential context-specific impact of individual core promoter elements on transcriptional dynamics. *Mol Biol Cell*, 28(23):3360–3370, 2017. ISSN 1939-4586 (Electronic) 1059-1524 (Linking). doi: 10.1091/mbc.E17-06-0408.
- Kristina M. Herbert, William J. Greenleaf, and Steven M. Block. Single-Molecule Studies of RNA Polymerase: Motoring Along. *Annual Review of Biochemistry*, 77(1):149–176, 2008. ISSN 0066-4154. doi: 10.1146/annurev.biochem.77.073106.100741.
- Elizabeth A. Heron, Bärbel Finkenstädt, and David A. Rand. Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study. *Bioinformatics*, 23(19): 2596–2603, 2007. ISSN 13674803. doi: 10.1093/bioinformatics/btm367.
- G. Z. Hertz and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)*, 15(7-8): 563–577, August 1999. ISSN 1367-4803.
- Gerald Z. Hertz, George W. Hartzell, and Gary D. Stormo. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Bioinformatics*, 6(2):81–92, apr 1990. ISSN 13674803. doi: 10.1093/bioinformatics/6.2.81.
- Terrell L. Hill. *Cooperativity theory in biochemistry : steady-state and equilibrium systems*. Springer-Verlag, New York, 1985. ISBN 0387961038.
- Sami Hocine, Pascal Raymond, Daniel Zenklusen, Jeffrey A. Chao, and Robert H. Singer. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nature Methods*, 10(2):119–121, 2013. ISSN 15487091. doi: 10.1038/nmeth.2305.
- J. J. Hopfield. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc Natl Acad Sci U S A*, 71(10):4135–9, 1974.
- J. Jaeger, M. Blagov, D. Kosman, K. N. Kozlov, Manu, E. Myasnikova, S. Surkova, C. E. Vanario-Alonso, M. Samsonova, D. H. Sharp, and J. Reinitz. Dynamical analysis of regulatory interactions in the gap gene system of drosophila melanogaster. *Genetics*, 167(4):1721–37, 2004a. ISSN 0016-6731 (Print) 0016-6731 (Linking). doi: 10.1534/genetics.104.027334.
- J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K. N. Kozlov, Manu, E. Myasnikova, C. E. Vanario-Alonso, M. Samsonova, D. H. Sharp, and J. Reinitz. Dynamic control of positional information in the early drosophila embryo. *Nature*, 430(6997):368–71, 2004b. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature02678.

- Heath E. Johnson and Jared E. Toettcher. Illuminating developmental biology with cellular optogenetics. *Current Opinion in Biotechnology*, 52:42–48, 2018. ISSN 18790429. doi: 10.1016/j.copbio.2018.02.003. URL <https://doi.org/10.1016/j.copbio.2018.02.003>.
- D. L. Jones, R. C. Brewster, and R. Phillips. Promoter architecture dictates cell-to-cell variability in gene expression. *Science*, 346(6216):1533–6, 2014. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1255301.
- Iris Jonkers and John T. Lis. Getting up to speed with transcription elongation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, 16(3):167–177, 2015. ISSN 14710080. doi: 10.1038/nrm3953.
- WC Jung, T. Tschaplinski, L. Wang, J. Glazebrook, and JP Greenberg. Priming in systemic plant immunity. *Science*, 324(5923):89–91, 2009. ISSN 0036-8075. doi: 10.1126/science. URL <https://www.sciencemaginedigital.org/sciencemagazine/22{ }june{ }2018/MobilePagedArticle.action?articleId=1404350{&}app=false{#}articleId1404350>.
- J. S. Kanodia, H. L. Liang, Y. Kim, B. Lim, M. Zhan, H. Lu, C. A. Rushlow, and S. Y. Shvartsman. Pattern formation by graded and uniform signals in the early drosophila embryo. *Biophys J*, 102(3):427–33, 2012. ISSN 1542-0086 (Electronic) 0006-3495 (Linking). doi: 10.1016/j.bpj.2011.12.042.
- J. E. Keymer, R. G. Endres, M. Skoge, Y. Meir, and N. S. Wingreen. Chemosensing in escherichia coli: two regimes of two-state receptors. *Proc Natl Acad Sci U S A*, 103(6): 1786–91, 2006. ISSN 0027-8424 (Print) 0027-8424 (Linking). doi: 0507438103[pii]10.1073/pnas.0507438103.
- H.D. Kim and E.K. O’Shea. A quantitative model of transcription factor-activated gene expression. *Nat Struct Mol Biol*, 15:1192–1198, 2008.
- N. H. Kim, G. Lee, N. A. Sherer, K. M. Martini, N. Goldenfeld, and T. E. Kuhlman. Real-time transposable element activity in individual live cells. *Proc Natl Acad Sci U S A*, 113(26):7278–83, 2016. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1601833113.
- Stefan Klumpp. Pausing and Backtracking in Transcription Under Dense Traffic Conditions. *Journal of Statistical Physics*, 142(6):1252–1267, 2011. ISSN 00224715. doi: 10.1007/s10955-011-0120-3.
- Stefan Klumpp and Terence Hwa. Stochasticity and traffic jams in the transcription of ribosomal RNA: Intriguing role of termination and antitermination. *Proceedings of the National Academy of Sciences of the United States of America*, 105(47):18159–18164, 2008. ISSN 00278424. doi: 10.1073/pnas.0806084105.

- Jason N. Kuehner, Erika L. Pearson, and Claire Moore. Unravelling the means to an end: RNA polymerase II transcription termination. *Nature Reviews Molecular Cell Biology*, 12(5):283–294, 2011. ISSN 14710072. doi: 10.1038/nrm3098. URL <http://dx.doi.org/10.1038/nrm3098>.
- F. H. Lam, D. J. Steger, and E. K. O’Shea. Chromatin decouples promoter threshold from dynamic range. *Nature*, 453(7192):246–50, 2008. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature06867.
- N. C. Lammers, V. Galstyan, A. Reimer, S. A. Medin, C. H. Wiggins, and H. G. Garcia. Multimodal transcriptional control of pattern formation in embryonic development. *Proc Natl Acad Sci U S A*, 117(2):836–847, 2020. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1912500117.
- E. Larschan, E. P. Bishop, P. V. Kharchenko, L. J. Core, J. T. Lis, P. J. Park, and M. I. Kuroda. X chromosome dosage compensation via enhanced transcriptional elongation in drosophila. *Nature*, 471(7336):115–8, 2011. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/nature09757.
- D. R. Larson, R. H. Singer, and D. Zenklusen. A single molecule view of gene expression. *Trends Cell Biol*, 19(11):630–7, 2009. ISSN 1879-3088 (Electronic). doi: S0962-8924(09)00188-3[pii]10.1016/j.tcb.2009.08.008.
- D. R. Larson, D. Zenklusen, B. Wu, J. A. Chao, and R. H. Singer. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science*, 332(6028):475–8, 2011a. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 332/6028/475[pii]10.1126/science.1202142.
- Matthew H. Larson, Robert Landick, and Steven M. Block. Single-Molecule Studies of RNA Polymerase: One Singular Sensation, Every Little Step It Takes. *Molecular Cell*, 41(3):249–262, 2011b. ISSN 10972765. doi: 10.1016/j.molcel.2011.01.008. URL <http://dx.doi.org/10.1016/j.molcel.2011.01.008>.
- C. Lee, H. Shin, and J. Kimble. Dynamics of notch-dependent transcriptional bursting in its native context. *Dev Cell*, 50(4):426–435 e4, 2019. ISSN 1878-1551 (Electronic) 1534-5807 (Linking). doi: 10.1016/j.devcel.2019.07.001.
- T. L. Lenstra, A. Coulon, C. C. Chow, and D. R. Larson. Single-molecule imaging reveals a switch between spurious and functional ncRNA transcription. *Mol Cell*, 60(4):597–610, 2015. ISSN 1097-4164 (Electronic) 1097-2765 (Linking). doi: 10.1016/j.molcel.2015.09.028.
- Tineke L. Lenstra, Joseph Rodriguez, Huimin Chen, and Daniel R. Larson. Transcription Dynamics in Living Cells. *Annual Review of Biophysics*, 45(1):25–47, 2016. ISSN 1936-122X. doi: 10.1146/annurev-biophys-062215-010838.

- Mike Levine. Transcriptional enhancers in animal development and evolution, sep 2010. ISSN 09609822.
- C. Li, F. Cesbron, M. Oehler, M. Brunner, and T. Hofer. Frequency modulation of transcriptional bursting enables sensitive and rapid gene regulation. *Cell Syst*, 6(4):409–423 e11, 2018. ISSN 2405-4712 (Print) 2405-4712 (Linking). doi: 10.1016/j.cels.2018.01.012.
- G. W. Li, D. Burkhardt, C. Gross, and J. S. Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–35, 2014a. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2014.02.033.
- J. Li and D. S. Gilmour. Promoter proximal pausing and the control of gene expression. *Curr Opin Genet Dev*, 21(2):231–5, 2011. ISSN 1879-0380 (Electronic) 0959-437X (Linking). doi: S0959-437X(11)00014-1[pii]10.1016/j.gde.2011.01.010.
- J. Li, A. Dong, K. Saydaminova, H. Chang, G. Wang, H. Ochiai, T. Yamamoto, and A. Pertsinidis. Single-molecule nanoscopy elucidates rna polymerase ii transcription at single genes in live cells. *Cell*, 2019. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2019.05.029.
- X. Y. Li, M. M. Harrison, J. E. Villalta, T. Kaplan, and M. B. Eisen. Establishment of regions of genomic activity during the drosophila maternal to zygotic transition. *Elife*, 3, 2014b. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.03737.
- Xiao-Yong Li and Michael B. Eisen. Zelda potentiates transcription factor binding to zygotic enhancers by increasing local chromatin accessibility during early *jem* drosophila melanogaster *jem* embryogenesis. *bioRxiv*, page 380857, 2018. doi: 10.1101/380857.
- Hsiao-Lan Liang, Chung-Yi Nien, Hsiao-Yun Liu, Mark M. Metzstein, Nikolai Kirov, and Christine Rushlow. The zinc-finger protein zelda is a key activator of the early zygotic genome in drosophila. 456(7220):400–403, 2008. ISSN 0028-0836. doi: 10.1038/nature07388. URL <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2597674/>.
- S. C. Little, M. Tikhonov, and T. Gregor. Precise developmental gene expression arises from globally stochastic transcriptional activity. *Cell*, 154(4):789–800, 2013. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2013.07.025.
- F. Liu, A. H. Morrison, and T. Gregor. Dynamic interpretation of maternal inputs by the drosophila segmentation gene network. *Proc Natl Acad Sci U S A*, 110(17):6724–9, 2013. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1220912110.
- Junbo Liu and Jun Ma. Modulation of temporal dynamics of gene transcription by activator potency in the drosophila embryo. *Development (Cambridge)*, 142(21):3781–3790, 2015. ISSN 14779129. doi: 10.1242/dev.126946.

- Xiaochuan Liu, Jaime Freitas, Dinghai Zheng, Marta S. Oliveira, Mainul Hoque, Torcato Martins, Telmo Henriques, Bin Tian, and Alexandra Moreira. Transcription elongation rate has a tissue-specific impact on alternative cleavage and polyadenylation in *Drosophila melanogaster*. *Rna*, 23(12):1807–1816, 2017. ISSN 14699001. doi: 10.1261/rna.062661.117.
- T. Lucas, T. Ferraro, B. Roelens, J. De Las Heras Chanes, A. M. Walczak, M. Coppey, and N. Dostatni. Live imaging of bicoid-dependent transcription in drosophila embryos. *Curr Biol*, 23(21):2135–9, 2013. ISSN 1879-0445 (Electronic) 0960-9822 (Linking). doi: 10.1016/j.cub.2013.08.053.
- T. Lucas, H. Tran, C. A. Perez Romero, A. Guillou, C. Fradin, M. Coppey, A. M. Walczak, and N. Dostatni. 3 minutes to precisely measure morphogen concentration. *PLoS Genet*, 14(10):e1007676, 2018. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 10.1371/journal.pgen.1007676.
- S. E. Luria and M. Delbruck. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491–511, 1943. ISSN 0016-6731 (Print).
- Christophe K. Mapendano, Søren Lykke-Andersen, Jørgen Kjems, Edouard Bertrand, and Torben Heick Jensen. Crosstalk between mRNA 3 End Processing and Transcription Initiation. *Molecular Cell*, 40(3):410–422, 2010. ISSN 10972765. doi: 10.1016/j.molcel.2010.10.012.
- J. S. Margolis, M. L. Borowsky, E. Steingrimsson, C. W. Shim, J. A. Lengyel, and J. W. Posakony. Posterior stripe expression of hunchback is driven from two promoters by a common enhancer element. *Development*, 121(9):3067–77, 1995. ISSN 0950-1991 (Print) 0950-1991 (Linking).
- B. M. Martins and P. S. Swain. Trade-offs and constraints in allosteric sensing. *PLoS Comput Biol*, 7(11):e1002261, 2011. ISSN 1553-7358 (Electronic) 1553-734X (Linking). doi: 10.1371/journal.pcbi.1002261.
- S. Marzen, H. G. Garcia, and R. Phillips. Statistical mechanics of monod-wyman-changeux (mwc) models. *J Mol Biol*, 425(9):1433–60, 2013. ISSN 1089-8638 (Electronic) 0022-2836 (Linking). doi: 10.1016/j.jmb.2013.03.013.
- J. A. Miller and J. Widom. Collaborative competition mechanism for gene activation in vivo. *Mol Cell Biol*, 23(5):1623–32, 2003. ISSN 0270-7306 (Print) 0270-7306 (Linking).
- M. Mir, A. Reimer, J. E. Haines, X. Y. Li, M. Stadler, H. Garcia, M. B. Eisen, and X. Darzacq. Dense bicoid hubs accentuate binding along the morphogen gradient. *Genes Dev*, 31(17):1784–1794, 2017. ISSN 1549-5477 (Electronic) 0890-9369 (Linking). doi: 10.1101/gad.305078.117.

- M. Mir, M. R. Stadler, S. A. Ortiz, C. E. Hannon, M. M. Harrison, X. Darzacq, and M. B. Eisen. Dynamic multifactor hubs interact transiently with sites of active transcription in drosophila embryos. *Elife*, 7:e40497, 2018. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.40497.
- L. A. Mirny. Nucleosome-mediated cooperativity between transcription factors. *Proc Natl Acad Sci U S A*, 107(52):22534–9, 2010. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 0913805107[pii]10.1073/pnas.0913805107.
- Michi Miura, Supravat Dey, Saumya Ramanayake, Abhyudai Singh, David S. Rueda, and Charles R.M. Bangham. Kinetics of HTLV-1 reactivation from latency quantified by single-molecule RNA FISH and stochastic modeling. *PLoS Pathogens*, 15(11):1–20, 2019. ISSN 15537374. doi: 10.1371/journal.ppat.1008164. URL <http://dx.doi.org/10.1371/journal.ppat.1008164>.
- J. Monod, J. Wyman, and J. P. Changeux. On the nature of allosteric transitions: A plausible model. *J Mol Biol*, 12:88–118, 1965. ISSN 0022-2836 (Print) 0022-2836 (Linking).
- Melissa J. Moore and Nick J. Proudfoot. Pre-mRNA Processing Reaches Back to Transcription and Ahead to Translation. *Cell*, 136(4):688–700, 2009. ISSN 00928674. doi: 10.1016/j.cell.2009.02.001. URL <http://dx.doi.org/10.1016/j.cell.2009.02.001>.
- Pierre Morel. Gramm: grammar of graphics plotting in Matlab. *The Journal of Open Source Software*, 3(23):568, 2018. ISSN 2475-9066. doi: 10.21105/joss.00568.
- Brian Munsky, Guoliang Li, Zachary R. Fox, Douglas P. Shepherd, and Gregor Neuert. Distribution shapes govern the discovery of predictive models for gene regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 115(29):7533–7538, 2018. ISSN 10916490. doi: 10.1073/pnas.1804060115.
- T. Muramoto, D. Cannon, M. Gierlinski, A. Corrigan, G. J. Barton, and J. R. Chubb. Live imaging of nascent rna dynamics reveals distinct types of transcriptional pulse regulation. *Proc Natl Acad Sci U S A*, 2012. ISSN 1091-6490. doi: 1117603109[pii]10.1073/pnas.1117603109.
- Ginger W. Muse, Daniel A. Gilchrist, Sergei Nechaev, Ruchir Shah, Joel S. Parker, Sherry F. Grissom, Julia Zeitlinger, and Karen Adelman. RNA polymerase is poised for activation across the genome. *Nature Genetics*, 39(12):1507–1511, 2007. ISSN 10614036. doi: 10.1038/ng.2007.21.
- J. Narula and O. A. Igoshin. Thermodynamic models of combinatorial gene regulation by distant enhancers. *IET Syst Biol*, 4(6):393, 2010. ISSN 1751-8849 (Print) 1751-8849 (Linking). doi: 10.1049/iet-syb.2010.0010.

- C. Y. Nien, H. L. Liang, S. Butcher, Y. Sun, S. Fu, T. Gocha, N. Kirov, J. R. Manak, and C. Rushlow. Temporal coordination of gene networks by zelda in the early drosophila embryo. *PLoS Genet*, 7(10):e1002339, 2011. ISSN 1553-7404 (Electronic) 1553-7390 (Linking). doi: 10.1371/journal.pgen.1002339.
- Justin M. O’Sullivan, Sue Mei Tan-Wong, Antonin Morillon, Barbara Lee, Joel Coles, Jane Mellor, and Nick J. Proudfoot. Gene loops juxtapose promoters and terminators in yeast. *Nature Genetics*, 36(9):1014–1018, 2004. ISSN 10614036. doi: 10.1038/ng1411.
- O. Padovan-Merhar, G. P. Nair, A. G. Biaesch, A. Mayer, S. Scarfone, S. W. Foley, A. R. Wu, L. S. Churchman, A. Singh, and A. Raj. Single mammalian cells compensate for differences in cellular volume and dna copy number through independent global transcriptional mechanisms. *Mol Cell*, 58(2):339–52, 2015. ISSN 1097-4164 (Electronic) 1097-2765 (Linking). doi: 10.1016/j.molcel.2015.03.005.
- Jeremy S. Paige, Karen Y. Wu, and Samie R. Jaffrey. RNA Mimics of Green Fluorescent Protein. *Science*, 333(July), 2011. doi: 10.1126/science.1207339.
- Murali Palangat and Daniel R. Larson. Complexity of RNA polymerase II elongation dynamics. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1819(7):667–672, 2012. ISSN 18749399. doi: 10.1016/j.bbagr.2012.02.024. URL <http://dx.doi.org/10.1016/j.bbagr.2012.02.024>.
- A. Pare, D. Lemons, D. Kosman, W. Beaver, Y. Freund, and W. McGinnis. Visualization of individual scr mrnas during drosophila embryogenesis yields evidence for transcriptional bursting. *Curr Biol*, 19(23):2037–42, 2009. ISSN 1879-0445 (Electronic) 0960-9822 (Linking). doi: S0960-9822(09)01848-X[pii]10.1016/j.cub.2009.10.028.
- J. Park, J. Estrada, G. Johnson, B. J. Vincent, C. Ricci-Tam, M. D. Bragdon, Y. Shulgina, A. Cha, Z. Wunderlich, J. Gunawardena, and A. H. DePace. Dissecting the sharp response of a canonical developmental enhancer reveals multiple sources of cooperativity. *Elife*, 8, 2019. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.41266.
- D. S. Parker, M. A. White, A. I. Ramos, B. A. Cohen, and S. Barolo. The cis-regulatory logic of hedgehog gradient responses: key roles for gli binding affinity, competition, and cooperativity. *Sci Signal*, 4(176):ra38, 2011. ISSN 1937-9145 (Electronic). doi: 10.1126/scisignal.2002077.
- Glenn G Parsons and Alberta Canada Tg. Mitotic Repression of RNA Polymerase II Transcription Is Accompanied by Release of Transcription Elongation Complexes. 17(10): 5791–5802, 1997.
- J Peccoud and B. Ycart. Markovian modeling of gene product synthesis. *Theor Popul Biol*, 48:222–234, 1995.

- M. W. Perry, J. P. Bothma, R. D. Luu, and M. Levine. Precision of hunchback expression in the drosophila embryo. *Curr Biol*, 22(23):2247–52, 2012. ISSN 1879-0445 (Electronic) 0960-9822 (Linking). doi: 10.1016/j.cub.2012.09.051.
- R. Phillips, N. M. Belliveau, G. Chure, H. G. Garcia, M. Razo-Mejia, and C. Scholes. Figure 1 theory meets figure 2 experiments in the study of gene expression. *Annu Rev Biophys*, 48:121–163, 2019. ISSN 1936-1238 (Electronic) 1936-122X (Linking). doi: 10.1146/annurev-biophys-052118-115525.
- Rob Phillips, Jane Kondev, Julie Theriot, and Hernan. G. Garcia. *Physical Biology of the Cell, 2nd Edition*. Garland Science, New York, 2013.
- Pedro A.B. Pinto, Telmo Henriques, Marta O. Freitas, Torcato Martins, Rita G. Domingues, Paulina S. Wyrzykowska, Paula A. Coelho, Alexandre M. Carmo, Claudio E. Sunkel, Nicholas J. Proudfoot, and Alexandra Moreira. RNA polymerase II kinetics in polo polyadenylation signal selection. *EMBO Journal*, 30(12):2431–2444, 2011. ISSN 02614189. doi: 10.1038/emboj.2011.156.
- K. J. Polach and J. Widom. Mechanism of protein access to specific DNA sequences in chromatin: A dynamic equilibrium model for gene regulation. *J Mol Biol*, 254(2):130–49, 1995. ISSN 0022-2836 (Print).
- Raluca Roxana Purnichescu Purtan and Andreea Udrea. A modified stochastic simulation algorithm for time-dependent intensity rates. *Proceedings - 19th International Conference on Control Systems and Computer Science, CSCS 2013*, pages 365–369, 2013. doi: 10.1109/CSCS.2013.101.
- A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi. Stochastic mrna synthesis in mammalian cells. *PLoS Biol*, 4(10):e309, 2006. ISSN 1545-7885 (Electronic) 1544-9173 (Linking). doi: 10.1371/journal.pbio.0040309.
- O. Rapp and O. Yifrach. Using the mwc model to describe heterotropic interactions in hemoglobin. *PLoS One*, 12(8):e0182871, 2017. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). doi: 10.1371/journal.pone.0182871.
- O. Rapp and O. Yifrach. Evolutionary and functional insights into the mechanism underlying body-size-related adaptation of mammalian hemoglobin. *Elife*, 8, 2019. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.47640.
- T. Raveh-Sadka, M. Levo, and E. Segal. Incorporating nucleosomes into thermodynamic models of transcription regulation. *Genome Res*, 19:1480–1496, 2009.
- M. Razo-Mejia, S. L. Barnes, N. M. Belliveau, G. Chure, T. Einav, M. Lewis, and R. Phillips. Tuning transcriptional regulation through signaling: A predictive theory of allosteric induction. *Cell Syst*, 6(4):456–469 e10, 2018. ISSN 2405-4712 (Print) 2405-4712 (Linking). doi: 10.1016/j.cels.2018.02.004.

- P. Richard and J. L. Manley. Transcription termination by nuclear rna polymerases. *Genes Dev*, 23(11):1247–69, 2009. ISSN 1549-5477 (Electronic) 0890-9369 (Linking). doi: 10.1101/gad.1792809.
- J. Rodriguez and D. R. Larson. Transcription in living cells: Molecular mechanisms of bursting. *Annu Rev Biochem*, 89:189–212, 2020. ISSN 1545-4509 (Electronic) 0066-4154 (Linking). doi: 10.1146/annurev-biochem-011520-105250.
- Robert G. Roeder. The complexities of eukaryotic transcription initiation: regulation of preinitiation complex assembly. *Trends in Biochemical Sciences*, 16:402–408, 1991.
- N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–5, 2005.
- N. Rosenfeld, T. J. Perkins, U. Alon, M. B. Elowitz, and P. S. Swain. A fluctuation method to quantify in vivo fluorescence data. *Biophys J*, 91(2):759–66, 2006. ISSN 0006-3495 (Print) 0006-3495 (Linking). doi: S0006-3495(06)71774-X[pii]10.1529/biophysj.105.073098.
- M. A. Samee, B. Lim, N. Samper, H. Lu, C. A. Rushlow, G. Jimenez, S. Y. Shvartsman, and S. Sinha. A systematic ensemble approach to thermodynamic modeling of gene expression from sequence data. *Cell Syst*, 1(6):396–407, 2015. ISSN 2405-4712 (Print) 2405-4712 (Linking). doi: 10.1016/j.cels.2015.12.002.
- A. Sanchez and I. Golding. Genetic determinants and cellular constraints in noisy gene expression. *Science*, 342(6163):1188–93, 2013. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1242975.
- A. Sanchez, H. G. Garcia, D. Jones, R. Phillips, and J. Kondev. Effect of promoter architecture on the cell-to-cell variability in gene expression. *PLoS Comput Biol*, 7(3):e1001100, 2011. ISSN 1553-7358 (Electronic) 1553-734X (Linking). doi: 10.1371/journal.pcbi.1001100.
- A. Sanchez, S. Choubey, and J. Kondev. Regulation of noise in gene expression. *Annu Rev Biophys*, 42:469–91, 2013. ISSN 1936-1238 (Electronic) 1936-122X (Linking). doi: 10.1146/annurev-biophys-083012-130401.
- Hanae Sato, Sulagna Das, Robert H Singer, and Maria Vera. Imaging of DNA and RNA in Living Eukaryotic Cells to Reveal Spatiotemporal Dynamics of Gene Expression. *Annual review of biochemistry*, pages 1–29, 2020. ISSN 1545-4509. doi: 10.1146/annurev-biochem-011520-104955. URL <http://www.ncbi.nlm.nih.gov/pubmed/32176523>.
- Abbie Saunders, Leighton J. Core, and John T. Lis. Breaking barriers to transcription elongation. *Nature Reviews Molecular Cell Biology*, 7(8):557–567, 2006. ISSN 14710072. doi: 10.1038/nrm1981.

- R. Sayal, J. M. Dresch, I. Pushel, B. R. Taylor, and D. N. Arnosti. Quantitative perturbation-based analysis of gene expression predicts enhancer activity in early drosophila embryo. *Elife*, 5, 2016. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.08445.
- C. Scholes, A. H. DePace, and A. Sanchez. Combinatorial gene regulation through kinetic control of the transcription cycle. *Cell Syst*, 4(1):97–108 e9, 2017. ISSN 2405-4712 (Print) 2405-4712 (Linking). doi: 10.1016/j.cels.2016.11.012.
- Sandra R. Schulze and Lori L. Wallrath. Gene Regulation by Chromatin Structure: Paradigms Established in *Drosophila melanogaster*. *Annual Review of Entomology*, 52(1):171–192, jan 2007. ISSN 0066-4170. doi: 10.1146/annurev.ento.51.110104.151007.
- E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. K. Moore, J. P. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–8, 2006. ISSN 1476-4687 (Electronic).
- E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in drosophila segmentation. *Nature*, 451(7178):535–40, 2008. ISSN 1476-4687 (Electronic). doi: nature06496[pii]10.1038/nature06496.
- Lee A. Segel and Marshall Slemrod. The quasi-steady-state assumption: a case study in perturbation. *SIAM Review*, 31(3):446–477, 1989.
- Christopher P. Selby, Ronny Drapkin, Danny Reinberg, and Aziz Sancar. RNA polymerase II stalled at a thymine dimer: Footprint and effect on excision repair. *Nucleic Acids Research*, 25(4):787–793, 1997. ISSN 03051048. doi: 10.1093/nar/25.4.787.
- A. Senecal, B. Munsky, F. Proux, N. Ly, F. E. Braye, C. Zimmer, F. Mueller, and X. Darzacq. Transcription factors modulate c-fos transcriptional bursts. *Cell Rep*, 8(1):75–83, 2014. ISSN 2211-1247 (Electronic). doi: 10.1016/j.celrep.2014.05.053.
- L. A. Sepulveda, H. Xu, J. Zhang, M. Wang, and I. Golding. Measurement of gene regulation in individual cells reveals rapid switching between promoter states. *Science*, 351(6278):1218–22, 2016. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.aad0635.
- Alexander S Serov, Alexander J. Levine, and Madhav Mani. Abortive initiation as a bottleneck for transcription in the early drosophila embryo. *ArXiv e-prints*, page 1701.06079, 2017.
- Sydney M. Shaffer, Margaret C. Dunagin, Stefan R. Torborg, Eduardo A. Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, Katrin Sproesser, Patricia A. Brafford, Min Xiao, Elliott Eggan, Ioannis N. Anastopoulos, Cesar A. Vargas-Garcia, Abhyudai Singh, Katherine L. Nathanson, Meenhard Herlyn, and Arjun Raj. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, 546(7658):431–435, 2017. ISSN 14764687. doi: 10.1038/nature22794.

- E. Sharon, Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger, and E. Segal. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol*, 30(6):521–30, 2012. ISSN 1546-1696 (Electronic) 1087-0156 (Linking). doi: 10.1038/nbt.2205.
- Marc S Sherman and Barak A Cohen. Thermodynamic State Ensemble Models of cis-Regulation. *PLoS Computational Biology*, 8(3):1–10, 2012. doi: 10.1371/journal.pcbi.1002407. URL <https://doi.org/10.1371/journal.pcbi.1002407>.
- A. W. Shermoen and P. H. O’Farrell. Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. *Cell*, 67(2):303–10, 1991. ISSN 0092-8674 (Print) 0092-8674 (Linking).
- S. O. Skinner, H. Xu, S. Nagarkar-Jaiswal, P. R. Freire, T. P. Zwaka, and I. Golding. Single-cell analysis of transcription kinetics across the cell cycle. *Elife*, 5:e12175, 2016. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.12175.
- L. H. So, A. Ghosh, C. Zong, L. A. Sepulveda, R. Segev, and I. Golding. General properties of transcriptional time series in escherichia coli. *Nat Genet*, 43(6):554–60, 2011. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi: 10.1038/ng.821.
- F. Spitz and E. E. Furlong. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*, 13(9):613–26, 2012. ISSN 1471-0064 (Electronic) 1471-0056 (Linking). doi: 10.1038/nrg3207.
- Nicole Staudt, Sonja Fellert, Ho-Ryun Chung, Herbert Jäckle, and Gerd Vorbrüggen. Mutations of the Drosophila zinc finger-encoding gene vielfältig impair mitotic cell divisions and cause improper chromosome segregation. *Molecular biology of the cell*, 17(5):2356–65, may 2006. ISSN 1059-1524. doi: 10.1091/mbc.e05-11-1056. URL <http://www.ncbi.nlm.nih.gov/pubmed/16525017><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1446075>.
- G. Struhl, K. Struhl, and P. M. Macdonald. The gradient morphogen bicoid is a concentration-dependent transcriptional activator. *Cell*, 57(7):1259–73, 1989. ISSN 0092-8674 (Print) 0092-8674 (Linking). doi: 0092-8674(89)90062-7[pii].
- D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–4, 2011. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: science.1198817[pii]10.1126/science.1198817.
- L. R. Swem, D. L. Swem, N. S. Wingreen, and B. L. Bassler. Deducing receptor signaling parameters from in vivo analysis: Luxn/ai-1 quorum sensing in vibrio harveyi. *Cell*, 134(3):461–73, 2008. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2008.06.023.

- Sue Mei Tan-Wong, Juliet D. French, Nicholas J. Proudfoot, and Melissa A. Brown. Dynamic interactions between the promoter and terminator regions of the mammalian BRCA1 gene. *Proceedings of the National Academy of Sciences of the United States of America*, 105(13): 5160–5165, 2008. ISSN 00278424. doi: 10.1073/pnas.0801048105.
- K. Tantale, F. Mueller, A. Kozulic-Pirher, A. Lesne, J. M. Victor, M. C. Robert, S. Capozzi, R. Chouaib, V. Backer, J. Mateos-Langerak, X. Darzacq, C. Zimmer, E. Basyuk, and E. Bertrand. A single-molecule view of transcription reveals convoys of rna polymerases and multi-scale bursting. *Nat Commun*, 7:12248, 2016. ISSN 2041-1723 (Electronic) 2041-1723 (Linking). doi: 10.1038/ncomms12248.
- S. W. Teng, Y. Wang, K. C. Tu, T. Long, P. Mehta, N. S. Wingreen, B. L. Bassler, and N. P. Ong. Measurement of the copy number of the master quorum-sensing regulator of a bacterial cell. *Biophys J*, 98(9):2024–31, 2010. ISSN 1542-0086 (Electronic) 0006-3495 (Linking). doi: S0006-3495(10)00175-X[pii]10.1016/j.bpj.2010.01.031.
- Bin Tian and James L. Manley. Alternative polyadenylation of mRNA precursors. *Nature Reviews Molecular Cell Biology*, 18(1):18–30, 2016. ISSN 14710080. doi: 10.1038/nrm.2016.116. URL <http://dx.doi.org/10.1038/nrm.2016.116>.
- Simon F. Tolić-Nørrelykke, Anita M. Engh, Robert Landick, and Jeff Gelles. Diversity in the Rates of Transcript Elongation by Single RNA Polymerase Molecules. *Journal of Biological Chemistry*, 279(5):3292–3299, 2004. ISSN 00219258. doi: 10.1074/jbc.M310290200.
- Y. Tu. The nonequilibrium mechanism for ultrasensitivity in a biological switch: sensing by maxwell’s demons. *Proc Natl Acad Sci U S A*, 105(33):11737–41, 2008. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.0804641105.
- J. M. Vilar and S. Leibler. Dna looping and physical constraints on transcription regulation. *J Mol Biol*, 331(5):981–9, 2003.
- J. M. Vilar, C. C. Guet, and S. Leibler. Modeling network dynamics: the lac operon, a case study. *J Cell Biol*, 161(3):471–6, 2003. ISSN 0021-9525 (Print) 0021-9525 (Linking). doi: 10.1083/jcb.200301125jcb.200301125[pii].
- Fangbin Wang, Hui Shi, Rui He, Renjie Wang, Rongjing Zhang, and Junhua Yuan. Non-equilibrium effect in the allosteric regulation of the bacterial flagellar switch. *Nature Physics*, 13(7):710–714, 2017. ISSN 1745-2481. doi: 10.1038/nphys4081.
- M. A. White, D. S. Parker, S. Barolo, and B. A. Cohen. A model of spatially restricted transcription in opposing gradients of activators and repressors. *Mol Syst Biol*, 8:614, 2012. ISSN 1744-4292 (Electronic) 1744-4292 (Linking). doi: 10.1038/msb.2012.48.
- F. Wong and J. Gunawardena. Gene regulation in and out of equilibrium. *Annu Rev Biophys*, 49:199–226, 2020. ISSN 1936-1238 (Electronic) 1936-122X (Linking). doi: 10.1146/annurev-biophys-121219-081542.

- B. Wu, J. Chen, and R. H. Singer. Background free imaging of single mrnas in live cells using split fluorescent proteins. *Sci Rep*, 4:3615, 2014. ISSN 2045-2322 (Electronic) 2045-2322 (Linking). doi: 10.1038/srep03615.
- B. Wu, C. Eliscovich, Y. J. Yoon, and R. H. Singer. Translation dynamics of single mrnas in live cells and neurons. *Science*, 352(6292):1430–5, 2016. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.aaf1084.
- Bin Wu, Jeffrey A. Chao, and Robert H. Singer. Fluorescence fluctuation spectroscopy enables quantitative imaging of single mrnas in living cells. *Biophysical Journal*, 102:2936, 2012.
- Matthieu Wyart, David Botstein, and Ned S. Wingreen. Evaluating gene expression dynamics using pairwise RNA fish data. *PLoS Computational Biology*, 6(11), 2010. ISSN 1553734X. doi: 10.1371/journal.pcbi.1000979.
- H. Xu, L. A. Sepulveda, L. Figard, A. M. Sokac, and I. Golding. Combining protein and mrna quantification to decipher transcriptional regulation. *Nat Methods*, 12(8):739–42, 2015. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi: 10.1038/nmeth.3446.
- Z. Xu, H. Chen, J. Ling, D. Yu, P. Struffi, and S. Small. Impacts of the ubiquitous factor zelda on bicoid-dependent dna binding and transcription in drosophila. *Genes Dev*, 28(6):608–21, 2014. ISSN 1549-5477 (Electronic) 0890-9369 (Linking). doi: 10.1101/gad.234534.113.
- S. Yamada, P. H. Whitney, S. K. Huang, E. C. Eck, H. G. Garcia, and C. A. Rushlow. The drosophila pioneer factor zelda modulates the nuclear microenvironment of a dorsal target enhancer to potentiate transcriptional output. *Curr Biol*, 29(8):1387–1393 e5, 2019. ISSN 1879-0445 (Electronic) 0960-9822 (Linking). doi: 10.1016/j.cub.2019.03.019.
- C. Zechner, M. Unger, S. Pelet, M. Peter, and H. Koepl. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. *Nat Methods*, 11(2):197–202, 2014. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi: 10.1038/nmeth.2794.
- R. D. Zeigler and B. A. Cohen. Discrimination between thermodynamic models of cis-regulation using transcription factor occupancy data. *Nucleic Acids Res*, 42(4):2224–34, 2014. ISSN 1362-4962 (Electronic) 0305-1048 (Linking). doi: 10.1093/nar/gkt1230.
- L. Zeng, S. O. Skinner, C. Zong, J. Sippy, M. Feiss, and I. Golding. Decision making at a subcellular level determines the outcome of bacteriophage infection. *Cell*, 141(4):682–91, 2010. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2010.03.034.
- D. Zenklusen, D. R. Larson, and R. H. Singer. Single-rna counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol*, 15(12):1263–71, 2008. ISSN 1545-9985 (Electronic) 1545-9985 (Linking). doi: nsmb.1514[pii]10.1038/nsmb.1514.

- R. P. Zinzen, K. Senger, M. Levine, and D. Papatsenko. Computational models for neurogenic gene expression in the drosophila embryo. *Curr Biol*, 16(13):1358–65, 2006. ISSN 0960-9822 (Print) 0960-9822 (Linking). doi: 10.1016/j.cub.2006.05.044.
- B. Zoller, S. C. Little, and T. Gregor. Diverse spatial expression patterns emerge from unified kinetics of transcriptional bursting. *Cell*, 175(3):835–847 e25, 2018. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi: 10.1016/j.cell.2018.09.056.