UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Dynamics of Gene Regulation and Chromatin Accessibility

Permalink
https://escholarship.org/uc/item/0nw2s3xj

Author
Eck, Elizabeth

Publication Date
2021

Peer reviewed|Thesis/dissertation

Dynamics of Gene Regulation and Chromatin Accessibility

by

Elizabeth Eck

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biophysics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Hernan Garcia, Chair
Professor Richard Harland
Professor Robert Phillips
Professor Michael Eisen

Fall 2021

Dynamics of Gene Regulation and Chromatin Accessibility

Abstract

Dynamics of Gene Regulation and Chromatin Accessibility

by

Elizabeth Eck

Doctor of Philosophy in Biophysics

University of California, Berkeley

Professor Hernan Garcia, Chair

Recent advances in live imaging technologies as well as non-equilibrium models are revealing new insights into the dynamic process of transcriptional regulation. The works described here all utilize MS2 imaging of transcription in live, developing embryo of fruit flies and zebrafish to investigate transcriptional dynamics in these model organisms. We utilize the resulting high resolution data to develop and test physics-based models of transcription.

We investigate two different regulatory enhancers in fruit fly embryos in three different projects. Through the first of these projects, we illuminate new information about the timing of transcriptional onset and the role of the pioneer-like factor Zelda. We connect newly discovered Zelda-dependent "hubs" of an activating transcription factor to transcriptional output for the first time. Next, we undertake a systematic dissection of transcriptional regulation in eukaryotes. Through this project, we challenge the standard equilibrium model of transcriptional regulation and demonstrate the need for non-equilibrium models. We propose a model of transcription that incorporates transient transcription factor dynamics to drive DNA accessibility. The third project extends the study of transcriptional dynamics beyond initiation. We devise a novel experimental and theoretical foundation for the study of the interconnected regulation of the entire transcription cycle. In the last project we describe here, we move beyond the *Drosophila* field and expand our MS2 technology to zebrafish and established a transgenic MS2 reporter line of fish for the first time.

The imaging technology and modeling used in these projects have already proven to be incredibly useful for studying the dynamic process of transcription, revealing new mechanisms of regulation as well as enabling the development and testing of new theoretical models of transcription. This is just the beginning and, in the coming years, further improvements in imaging technologies in combination with continued efforts on the theoretical modeling front will surely continue to bring about new discoveries in the field of transcriptional regulation.

To my parents, Robert and Carolyn, my husband, Michael, and my darling son, Elliott.

Contents

List of Figures

List of Tables

Acknowledgments

My husband and I were recently reflecting on our careers, and he asked me if the version of myself from seven years ago, at graduate school interviews, would think of current day me as successful. I thought back to what I considered to be signs of success in senior graduate students and realized that I have since checked all the boxes: I am getting my PhD from a world class university, I got a prestigious graduate fellowship, I have contributed to several successful publications, I'm moving on to an exciting post-doctoral position, and, most importantly, I thoroughly enjoyed the experience, learned a ton, made great connections with other researchers, and I am incredibly happy with where I am today. None of this would have been possible without the many people who helped me along the way.

Of course, the person who shaped my graduate career more than anyone else was my advisor, Hernan Garcia. He is, without doubt, the greatest research mentor I've ever had. His infectious enthusiasm and excitement for science is almost unparalleled. He encouraged me to build my skills in all realms related to scientific research: experimental design, data analysis, theoretical modeling, illustrating, writing, and speaking, to name a few. He is kind and considerate of his graduate students and does everything in his power to help them achieve success. One day, I hope to be a fraction of the scientist and advisor that he is. His mentorship has truly shaped my work and helped me develop my scientific style. Hernan has a beautiful family and he and his wife, Julia, are some of the nicest and most supportive people that I know. I want them both to know how much I appreciate their care and support.

Much appreciation is due to the Biophysics Graduate Group. Our Head Graduate Advisor, Susan Marqusee, our two Department Chairs over the last six years, James Hurley and Ahmet Yildiz have all been wonderful program leaders. Of course, a big thank you is due to our Graduate Student Affairs Officer, Kate Chase, who really keeps the program running day after day and is always quick to respond to any questions. Thank you and a fond farewell to the other Biophysics graduate students, especially my cohort who have recently or are soon moving on to their next phase. I'm excited to see what everyone's next steps will be.

One thing I will miss about UC Berkeley are the many outreach opportunities. I've participated in several rewarding programs. In particular, I will miss my friends at the Physical sciences Opportunities for Women in Education and Research (POWER) Bay Area program. It was a privilege to work with such an enthusiastic group of women to empower students from marginalized genders, especially nontraditional students. I learned so much about starting and running an outreach program from these talented scientists.

I want to thank my thesis committee, Richard Harland, Mike Eisen, and Rob Phillips for their support and feedback. I also send thanks to my former committee member and lab neighbor, Nipam Patel, who has moved on to be the director of the Marine Biological Laboratory (MBL) in Woods Hole, Massachusetts.

Speaking of the MBL, I've spent many summers traveling to Woods Hole. In 2017, I enrolled in the Physiology course. It was a grueling and yet incredibly fun experience where I made life-long friends and gained a new excitement for the breadth of incredible research occurring around the world. I've gone back to the MBL several times as a teaching assistant

xii

and always enjoy the exciting and collaborative environment found there. Of course, the directors of the physiology course, Rob Phillips, Jennifer Lippincott Schwartz, and Wallace Marshall all shaped the incredible atmosphere of course. Rob Phillips, in particular influenced my research and my time at the MBL. My rotation with him, exploring kinetic models of transcription heavily influenced my work in the Garcia Lab. Of course, I loved our daily (and sometime twice a day) swims to the dock of truth which built up my swimming endurance and were a welcome physical break from the lab. Rob is a deeply caring person and I know I can always turn to him when I need a keen eye to review my work or a swimming buddy.

I have truly enjoyed and learned from my teaching and mentoring opportunities in graduate school. In addition to teaching at Woods Hole, I assisted with Hernan's Physical Biology of the Cell course here at Berkeley. Hernan is an extremely engaging teacher and I learned a lot from teaching alongside him. I also had the privilege of mentoring some truly talented undergraduate students. Maryam Kazemzadeh-Atoufi was a data acquisition machine and digested the literature with amazing aptitude. Sydney Ghoreishi, who worked with me for three years is probably the most enthusiastic student I've ever met. Without her help and hard work, I never would have made as much progress through the research as we did together. Both Maryam and Sydney are now superstars at their own graduate schools and I know we'll be hearing about their work in the future.

Thank you to the other labs that I've had the privilege to collaborate and brainstorm with including, but not limited to, the Harrison lab at UW-Madison, the DePace Lab at Harvard, the Bintu Lab at Stanford, the Chica Lab at the University of Ottawa, the Oates Lab at The EPFL, and the Phillips Lab at CalTech. Our interactions, retreats, and collaborations helped me develop into the scientist that I am today. Extra special shout-outs to Matthew Eason for his collaboration on the protein sensors, Daniele Soroldoni for being my expert guide to the world of zebrafish research, and Phillips Lab graduate students Griffin Chure, Manuel Razo, Muir Morrison, and Tal Einav for who have each probably forgotten more biophysics than I will ever know and for being awesome teachers!

Any research lab environment is shaped by its lab members. Thank you so much to my lab-mates for all the laughs, support, troubleshooting, mentoring, and generosity. I have learned and at one time or another sought help from each and every one of them. The lab's founding members Armando Reimer and Simon Alamos along with Matthew (Matty) Norstad and Jacques Bothma welcomed me to the lab and showed me the ropes in the early days. Jonathan Liu was a fantastic research partner and coauthor. Yang Joon Kim was part of my biophysics cohort as well as my lab-mate. Studying for quals and navigating the program was so much easier and enjoyable together than it would have been alone. Brandon Schlomann and Bruno Moretti, thank you for taking over the zebrafish work and helping me navigate applying for postdoc positions and fellowships. Gabriella Martini was an awesome co-GSI for my semester teaching. Thanks to Pranjal Srivastava and Yovan Badal for taking over the Zelda project. Meghan Turner has been a great swimming partner and MBL roomie and co-TA. Nicholas Lammers and Jake Zhao thank you for helping me with the Zeiss scope, modeling, and general research questions. Our fantastic technicians Myron Child, Emma Luu, and Paul Talledo kept the lab run smoothly along with Laura Fantone who was a huge

help in navigating the red tape and paperwork over the past few years. If anyone wants to relive some of our glory days just look up our retreat introduction video. It sums up the general atmosphere and camaraderie we have in the lab.

I did not have the most traditional path to science or graduate school, but along the way, I have always had the support from my friends from afar. Although we don't talk everyday, I know that they will always be there to help when I need advice, to listen when I need to vent, or to cheer me on when I share triumphs. A special thanks to the Winnipeg crew, Amar, Mat, Natalie, Willow, Tiffany, and Darcie who were my family away from home. Of course, thank you to Stephanie, my childhood best friend (of 30+ years) who has literally climbed mountains, weathered lightning storms, and traveled halfway around the world with me.

Thank you to my family. My brother has always supported me and wanted to be a part of my life and have me in his. My parents have been my biggest cheerleaders and sacrificed so much for me to achieve my dreams. They have supported me through thick and thin, no matter what. They took care of me when I developed epilepsy and couldn't function for an entire year. They wouldn't let me give up looking for medication or therapies to stop my seizures. They supported my ballet career 110% putting in long hours driving me around to classes and rehearsals and even supported my decision to complete high school through independent study and correspondence in order to dance full-time. They let me move back home when I transitioned out of ballet and had to figure out what I wanted to do with the rest of my life. From childhood, they encouraged my curiosity in science and gave me the confidence that I was smart enough to find success as a scientist. I say with complete confidence that I would not be where I am or who I am today without their sacrifice, love, and support.

Lastly, my graduate school journey would not have been as joyful without the love and support of my husband, Michael. He has been so proud of my journey and success, by my side every step of the way. He takes a genuine interest in my work and has been a huge help in proofreading and suggesting edits for my writing. Our lives have changed so much since we moved to Berkeley together. We have gone on adventures around the world and built a home at UC Village. We planned and had a beautiful wedding with most of our friends and family in attendance. We've weathered a global pandemic together. Our biggest and most special change has been the arrival of our son, Elliott, our goofy little guy. Michael is uprooting his life again so we can move across the country for my new post-doctoral position. I am going to miss our Berkeley home and all of the delicious food we've discovered in the area, but I'm so excited to start the next stage of our lives together. Michael's intelligence, humor, thoughtfulness, and big heart challenge me to be a better person every day and Elliott, in his short time with us, has challenged me to write a dissertation on as little sleep as humanly possible. In all seriousness, I cannot imagine spending my life with anyone else and can't wait to raise Elliott together. I hope he grows up to be as wonderful and caring a person as his father.

Chapter 1

Introduction

The cells in an animal embryo all start out identically and contain identical genetic information, but over the course of development adopt different fates. These fates are acquired, in large part, due to the transcription of certain genes being regulated—turned on or off, or tuned to higher or lower levels—at particular times by the DNA binding of activators and repressors in the vicinity of the gene. One of the holy grails of developmental biology is to be able to predict this transcription from regulatory DNA sequences. This process is highly dynamic, changing quickly over space and time, but many of the techniques used to investigate it to date use fixed, dead tissue. Therefore, we obtain snapshots of development, losing most of the vital dynamic information.

Recent advances in live imaging technology have revealed previously unknown characteristics and dynamics of transcription. These new insights have, in some cases, shown standard theoretical models to be invalid or unsuitable when dynamics are taken into account. For example, two genes expressed in the early fruit fly embryo, *hunchback* (Desponds et al., 2016; Eck et al., 2020) and *snail* (Dufourt et al., 2018), were characterized by models based on equilibrium statistical mechanics. However, recent live expression data of these genes have shown that these models fail to predict the full dynamics of their transcription. Thus, advances in experimental techniques which allow us to investigate transcriptional dynamics in detail, present exciting opportunities to develop and test new models. This dialogue between theory and experiment will undoubtedly reveal new information about the mechanisms of transcriptional regulation. The works presented here stand on the shoulders of giants who carefully studied and characterized developmental gene regulatory networks, enabling us to tag genes involved in these networks for live imaging and to develop theoretical models of their regulation. Through the works presented here, these networks are reexamined through the lens of their transcriptional dynamics.

1.1 The dynamic regulation of transcription

Transcription, one of the first steps involved in gene expression, is the process of copying the DNA encoding a gene into messenger RNA (mRNA). This entire process occurs in a repeating cycle that starts with initiation followed by elongation and then termination. Regulation may occur at each step in the cycle.

Initiation is an especially important component of the cycle because it comprises the main juncture at which much of the regulation occurs (Alberts, 2015). This regulation is accomplished through a variety of mechanisms. This work will focus on two major modes of initiation regulation: transcription factor activity and chromatin state.

Transcription factors are proteins that can activate (turn on or speed up) or repress (turn off or slow down) transcription by binding to specific binding sites, short nucleotide motifs, in regions of regulatory DNA called enhancers. Exactly how transcription factors accomplish their roles remains a mystery in many cases. They may, for example, enhance the binding of other transcription factors or the loading of RNA polymerase (RNAP). Alternatively, they may influence their local microenvironment increasing the concentration of other transcription factors. Another possibility is that they recruit chromatin remodelers that can influence the accessibility of the DNA.

Although, great strides have been made toward decoding how the arrangement and affinity of transcription factor binding sites in enhancers direct gene expression and developmental patterns *in vivo*, we still cannot quantitatively predict transcriptional outputs simply by investigating this regulatory architecture. Part of the reason for this is the dynamic nature of transcription factors *in vivo*. Their concentrations change quickly over both time and space. In order to develop accurate and predictive models of transcription, we cannot ignore these dynamics.

Many transcriptional models attempt to predict transcriptional output from the probability that a transcription factor is occupying its binding site(s) which is dependent on its concentration as well as its affinity for the binding sequence(s) in question (Fig. 1.1). In addition to concentration changes over time and space, we are learning that transcription factors may form "hubs" or small microenvironments within which the transcription factor concentration is higher than that in the surrounding regions (Mir et al., 2017, 2018; Dufourt et al., 2018; Tsai et al., 2017; Yamada et al., 2019). Transcription sites within these microenvironments will display outputs and dynamics that are different from what would be predicted based on the spatially averaged transcription factor concentration. These local microenvironments may be especially important in regions where the transcription factor would otherwise be at concentrations that are too low for it to efficiently bind the DNA and accomplish its role. The dynamics of and mechanisms behind these hubs are areas of active research.

Another possibility for the lack of a predictive understanding of enhancer-driven transcriptional output is that the regulatory sequence alone does not provide information about chromatin state, which can also change dynamically. While the organization of eukaryotic DNA into chromatin was originally thought to be a packaging mechanism, it was revealed to

play a role in regulating the access of transcription factors to their binding sites (Spitz and Furlong, 2012; Baylin and Jones, 2011; Johnson and Dent, 2013; Hardison and Taylor, 2012). In order for transcription factors to access the DNA and transcription to begin, the DNA must unwind from nucleosomes and become accessible (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).

This regulation is not a one-way street. While the chromatin state can affect the transcription factors access to the DNA, transcription factors themselves can modulate chromatin state. Pioneer transcription factors, for example, have been suggested to be able to bind DNA that is wrapped around nucleosomes and facilitate increased accessibility for other transcription factors (Zaret and Carroll, 2011).

Fixed-cell techniques that can probe entire populations of cells allow access to the chromatin state at single time points, providing snapshots of the chromatin landscape. However, the dynamics of the changing structure as well as the molecular processes governing it (such as nucleosome remodeling, post-translational modification, and protein-protein interactions), occurs on multiple time scales and remain challenging to examine (Larson, 2011; Mellor, 2006). Probes have been developed to track chromatin modifications in live cells (Sato et al., 2021) and superresolution microscopy has been used to image chromatin structure dynamics (Nozaki et al., 2017; Barth et al., 2020; Xiang et al., 2018), but simultaneously imaging chromatin structure and transcriptional output at a single locus remains elusive. Modeling may be an avenue to bridge the gap and infer chromatin state while experimental techniques continue to be developed.

While much transcriptional regulation occurs during initiation, regulation can occur at each step in the transcription cycle. For example, splicing efficiency can be influenced by the modulation of elongation rates (Spitz and Furlong, 2012), and the cleavage of mRNA, part of the termination process, can control aspects of 3' processing (Tian and Manley, 2016). While each step of the transcription cycle is dynamic, the details of the cycle have mostly been studied via *in vitro* experiments (Bai et al., 2006; Herbert et al., 2008) and by genome wide measurements. These measurements are only possible following cell fixation (Roeder, 1991; Saunders et al., 2006; Muse et al., 2007; Core et al., 2008; Fuda et al., 2009; Churchman and Weissman, 2011). Additionally, the steps of the transcriptional cycle have mostly been studied in isolation from one another, thus the real-time information encoded in how the dynamics of these processes vary and relate to one another has remained mostly unexamined. Here too, we lack a predictive understanding of the flow of information, and its regulation, throughout the transcription cycle.

A few studies have found that the steps of the transcriptional cycle can influence one another. For instance, elongation rates play a role in determining termination efficiency, including mRNA cleavage (Pinto et al., 2011; Hazelbaker et al., 2013; Fong et al., 2015; Liu et al., 2017). Additionally, transcription initiation and termination have been functionally linked (Moore and Proudfoot, 2009; Mapendano et al., 2010). More work investigating the transcription cycle in its entirety is needed in order to understand the dynamic flow of information: how the regulation of each step modulates the other steps and how each of the

steps might be coupled to the others.

The transcription cycle is highly complex and involves many convoluted regulatory processes. Spatiotemporal dynamics are an intrinsic component of the regulation of the cycle from the modulation of transcriptional initiation, involving interacting transcription factors and a changing chromatin environment, to the interconnected regulation of initiation, elongation, and termination. Focusing on dynamics in experimental and theoretical investigations will be essential to the development of a comprehensive and predictive understanding of these vital processes.

1.2   transcriptional models

The dominant class of models of transcription over the past several decades is based in equilibrium statistical mechanics. These models reduce transcriptional regulation to the simple physical concept of regulatory molecules and transcriptional machinery binding and unbinding their binding sites in regulatory DNA. The various binding configurations in which a system can be found can be described by their statistical weight as seen in the simple example of a regulatory region containing a single activator binding site and single RNAP binding site (Fig. 1.1). These statistical weights, given by their Boltzmann weights, are based on physical parameters and characteristics of the system such as molecule concentrations, dissociation constants, and cooperativity factors. Transcriptional output, which ultimately results from RNAP loading at the promoter, is assumed to be proportional to the probability of finding the system in an RNAP-bound microstate. This probability can easily be calculated by dividing the sum of the weights of the RNAP-bound microstates by the partition function (Fig. 1.1). This assumption is called the *occupancy hypothesis* (Hammar et al., 2014; Garcia et al., 2012; Phillips et al., 2019).

The occupancy hypothesis is a key characteristic of equilibrium models, but it is not the only assumption being made. In fact, equilibrium models of transcription are a subclass of the more general kinetic models. Kinetic models are considered equilibrium models when four assumptions hold, as shown in Figure 1.2. In addition to the occupancy hypothesis (Fig. 1.2,i), the system is assumed to be in quasi-steady state with respect to transcription factor dynamics. This means that the transitions between different binding configurations (microstates) are much quicker than changes in relevant transcription factor concentrations (Fig. 1.2, ii) as well as transcriptional initiation rates (Fig. 1.2, iii). These two assumptions both have to do with separation of time scales and imply that the microstate probabilities respond instantaneously to changes in transcription factor concentrations. The last assumption required by equilibrium models is that no energy is consumed in order for the system to realize the different microstates. In other words, energy is conserved (Fig. 1.2, iv).

As discussed in Section 1.1, transcription is highly dynamic and equilibrium models do not explicitly account for time dependent processes. So, while equilibrium models have found success in some cases, especially in the field of bacterial transcriptional regulation, (Ackers et al., 1982; Buchler et al., 2003; Vilar et al., 2003; Bakk et al., 2004; Bintu et al., 2005b,a;
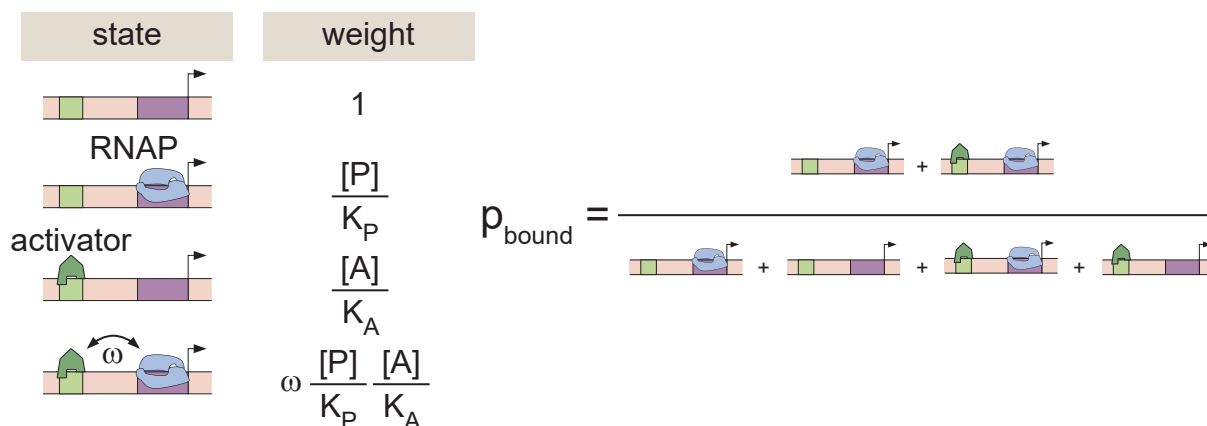
Figure 1.1: A simple equilibrium model transcriptional activation. An example gene regulatory region contains one binding site for an activator (green) and one binding site for RNAP (purple). The Boltzmann weight of each microstate (binding configuration) can be determined from dissociation constants, K_A and K_P (determined from given equilibrium binding energies), concentrations of the activator and RNAP, and the cooperativity factor between the activator and RNAP, ω. If the cooperativity factor is greater than one, the system will exhibit activation by the activator. The probability of finding RNAP bound is given by p_{bound}.

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), their applicability to the more complex process of eukaryotic transcription is less certain. There are many reasons for why equilibrium models might be insufficient in eukaryotic transcriptional systems. First, transient dynamics may be indispensable for proper regulation and these dynamics cannot be captured by any equilibrium model because that would violate the separation of timescales, a requirement of equilibrium models (Wong and Gunawardena, 2020). Also, equilibrium models cannot account for the energy-expending regulatory steps, such as histone remodeling, that can be necessary to enable access to DNA rendered inaccessible by chromatin structure (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).

Indeed, recent studies performed in the live developing *Drosophila* embryo show that the spatiotemporal patterns of transcription violate limits set by equilibrium models (Kim and O'Shea, 2008; Estrada et al., 2016; Park et al., 2019; Eck et al., 2020). This implies that at least one of the four assumptions of equilibrium models (Fig. 1.2) is invalid in these cases. However, which assumptions are incorrect and to what extent is still unknown. A challenging aspect of investigations into non-equilibrium regulation is characterizing how the systems are
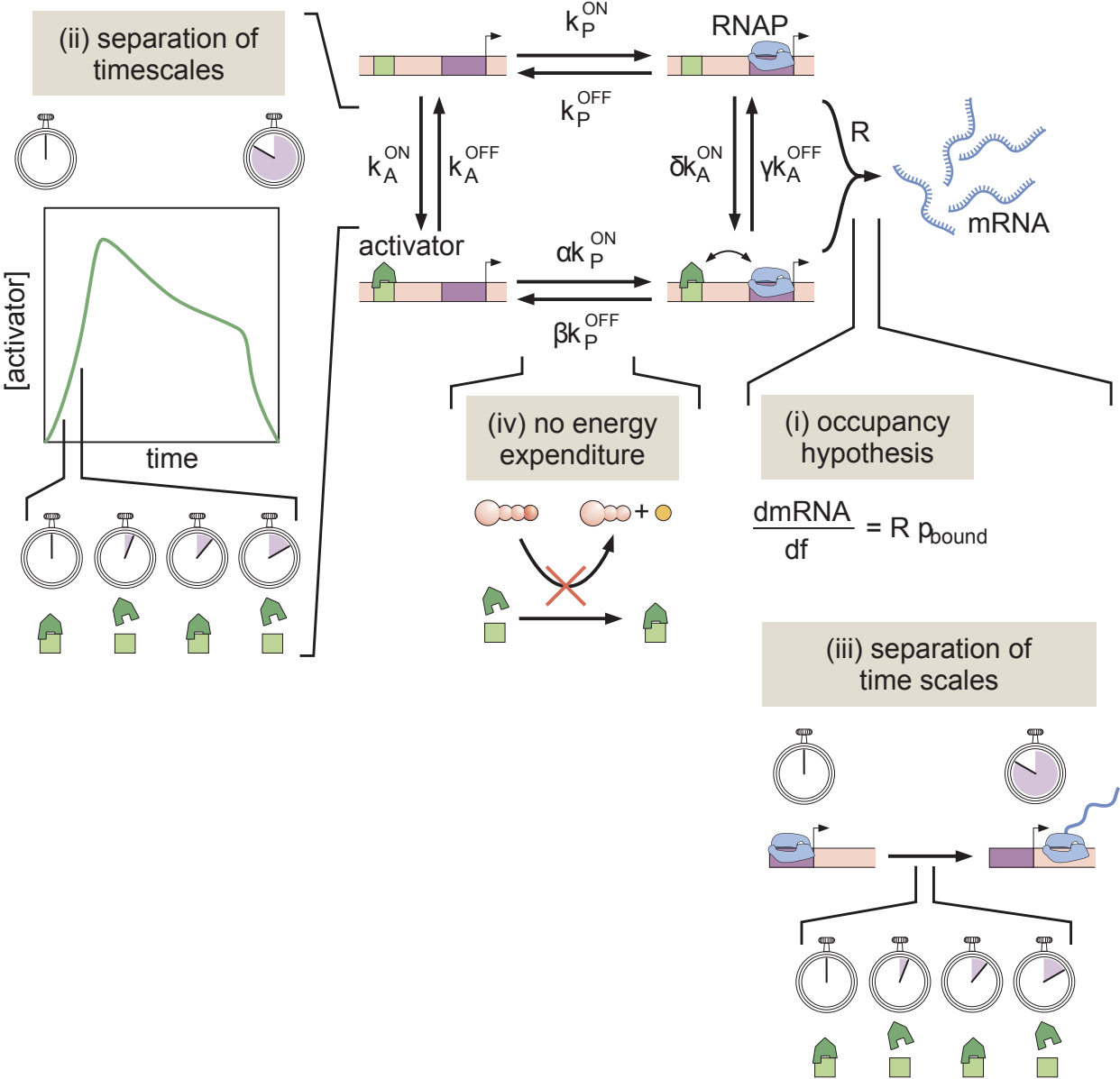
Figure 1.2: A simple kinetic model of transcriptional activation in which activator molecules influence RNAP binding kinetics. The assumptions that make it possible to turn this kinetic model into a thermodynamic one are (i) the occupancy hypothesis, (ii, iii) a separation of time scales between binding and unbinding rates, and activator and mRNA production dynamics, respectively, and (iv) no energy expenditure.

out of equilibrium (i.e. what assumptions of equilibrium models are being violated) as well as how far from equilibrium they lie. There are a vast number of potential non-equilibrium models, and distinguishing the unique consequences of each model is challenging, to say the least. We are also currently unsure whether lessons learned in these particular examples where equilibrium assumptions have been challenged—for example their non-equilibrium features—can be more broadly applied to eukaryotes. They may be specific to the organisms or even more specific to the particular genes being studied. In order to investigate this further, we need to use experimental tools that enable the study of live transcriptional dynamics are necessary.

1.3 Live imaging of transcription

Live imaging of mRNA *in vivo* relies on two main technological approaches. The first is probe hybridization to endogenous transcripts (Tutucci et al., 2018b). An example of this approach uses probes called molecular beacons (Fig. 1.3A; Tyagi and Kramer (1996)). Molecular beacons are composed of single-stranded DNA that folds into stem-loop structures. At one end of the DNA strand is a fluorescent molecule and at the other is a quencher. While in a stem-loop shape the two are near one another and fluorescence is quenched. However, the DNA sequence of the stem-loop is complimentary to the RNA target sequence and, when it binds its target, the quencher is separated from the fluorescent molecule and fluorescence can be detected. Molecular beacons have been used in a variety of contexts including human cells (Medley et al., 2005; Peng et al., 2005; Yang et al., 2011; Jha et al., 2015; de Oliveira et al., 2020), other mammalian cells (Vargas et al., 2005; Santangelo and Bao, 2007; Wang et al., 2008; Yeh et al., 2008; Kang et al., 2011), *Drosophila* oocytes (Bratu et al., 2003), chicken embryonic fibroblasts (Tyagi and Alsmadi, 2004), and *Xenopus* retinal ganglion cells (Turner-Bridger et al., 2018; Cioni et al., 2019). The downside to this technology is that the probes must be micro-injected or otherwise introduced into the cells. Recently, progress has been made through a number of approaches to visualize single transcripts *in vivo* using molecular beacons (Mao et al., 2020).

Molecular beacons only fluoresce upon binding to target RNA. The same principle is behind genetically encoded RNA aptamers (Fig. 1.3B). Aptamers bind fluorogenic dyes such as Spinach2 (Strack et al., 2013), Broccoli (Filonov et al., 2014), and RNA-Mango (Dolgosheina et al., 2014). The RNA aptamers form a secondary structure which binds the fluorogenic dyes which fluoresce upon binding (Tutucci et al., 2018b). Like the case with molecular beacons, recent innovations have allowed for the imaging of single molecules using RNA aptamers (Cawte et al., 2020).

The second approach to imaging mRNA *in vivo* is the use of repeated tags which bind fluorescently tagged molecules (Tutucci et al., 2018b). This approach relies on a pool of unbound, fluorescently tagged molecules that result in a background signal. Strategies have emerged to reduce this background including targeting the molecules to specific cellular compartments (Bertrand et al., 1998) and fine tuning their expression (Fusco et al., 2003;
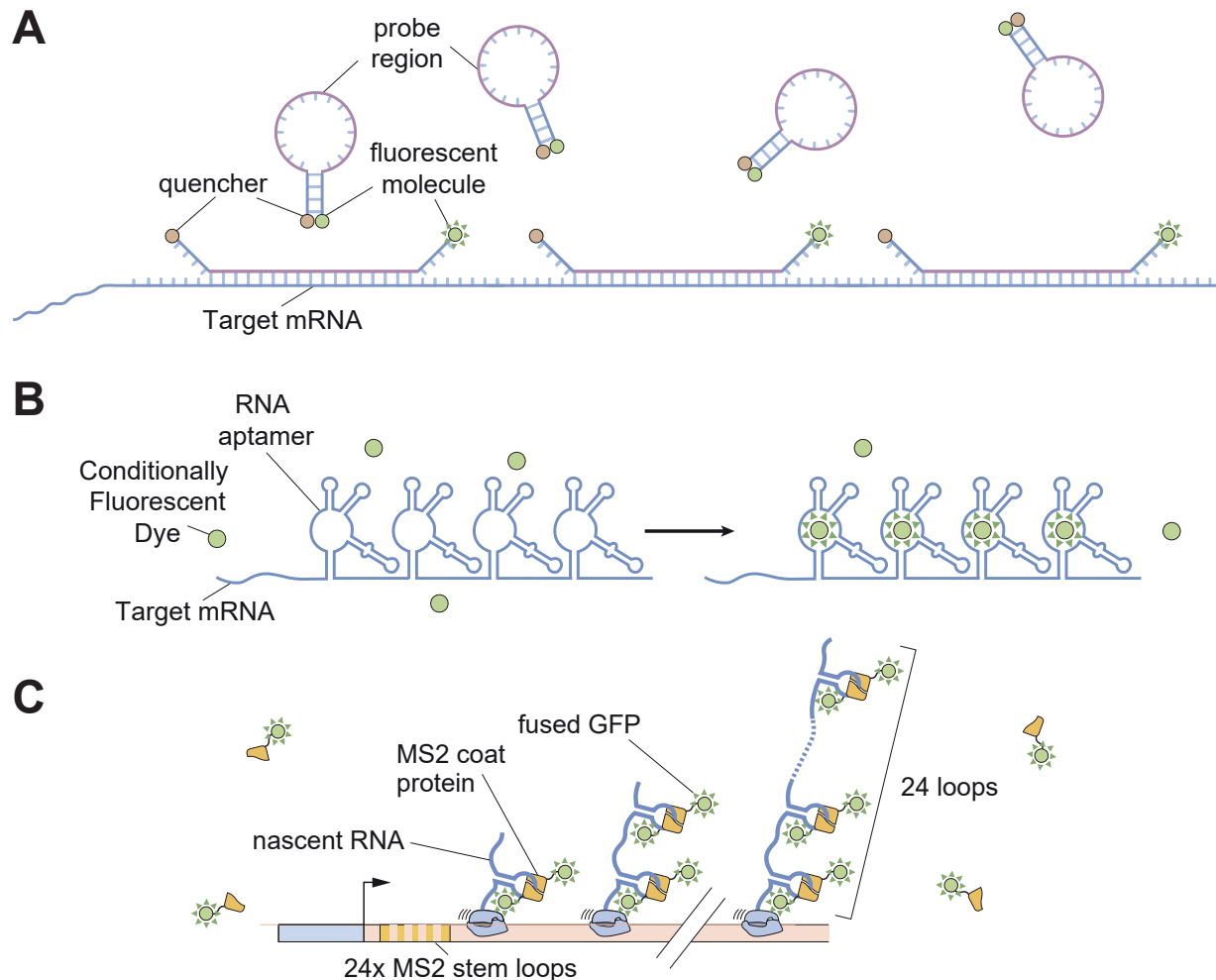
Figure 1.3: Live imaging methods. (A) Molecular beacons rely on a single stranded DNA stem-loop probe with a fluorescent molecule at one end and a quencher at the other. When the probe is folded in a stem-loop, fluorescence is quenched. When the probe binds complimentary target RNA, fluorescence can be detected. (B) RNA aptamers encoded in the target RNA bind conditionally fluorescent dyes which fluorescence upon binding. (C) MS2 stem-loop repeats are encoded in the target RNA. MS2 coat protein fused with a fluorescent protein dimerize and bind the MS2 stem-loops upon transcription, localizing the fluorescent signal and resulting in bright puncta that appear above the background fluorescence cause by unbound coat protein-fluorescent protein fusions.

Wu et al., 2012).

The works described here use the repeated tags of the MS2 and the similar PP7 imaging technologies to fluorescently label nascent transcripts (Fig. 1.3C; Bertrand et al. (1998); Golding et al. (2005); Chao et al. (2008); Larson et al. (2011a)). The MS2 system has become a standard technique in the imaging of single mRNA molecules and transcription *in vivo* (Bertrand et al., 1998; Yunger et al., 2010). In this system, the gene of interest is labeled with repeats of the bacteriophage-derived MS2 sequence that, once transcribed, will form stem-loop shapes. Each stem-loop is bound by a dimer of the MS2 coat protein-fluorescent protein fusion (MCP-FP), which is co-expressed in the same cell. This results in fluorescent puncta at the site of transcription whose intensity reports on the number of transcripts being formed (Garcia et al., 2013). The PP7-PCP system is an orthogonal, similar bacteriophage-derived stem-loop system (Chao et al., 2008; Hocine et al., 2013; Lim et al., 2001). The dual use of MS2 and PP7 allows for two color labeling. Although they are not used in the works presented here, there are other similar stem-loop labeling systems such as U1A (Brodsky and Silver, 2002; Takizawa and Vale, 2000), λ boxB (Lange et al., 2008), and Bgl (Chen et al., 2009).

The MS2 technology has been used to image transcription and transcript localization in a number of model systems including bacteria (Golding et al., 2005), yeast (Bertrand et al., 1998; Hocine et al., 2013; Tutucci et al., 2018a), amoebas (Chubb et al., 2006), *Drosophila* (Jaramillo et al., 2008; Garcia et al., 2013; Bothma et al., 2015), mammalian cells (Fusco et al., 2003; Shav-Tal et al., 2004), human cells (Janicki et al., 2004; Darzacq et al., 2007; Mor et al., 2010; Ben-Ari et al., 2010; Wu et al., 2016), and mice (Lionnet et al., 2011; Park et al., 2014; Yoon et al., 2016). This work has revealed numerous insights into features of transcription that were previously inaccessible. For example, mRNA was visualized localizing to the bud tip of living yeast cells (Bertrand et al., 1998) and moving on microtubules in mammalian (Cos) cells (Fusco et al., 2003). Transcriptional bursting has been visualized using MS2 in *Drosophila* (Lucas et al., 2013; Bothma et al., 2014; Fukaya et al., 2016; Falo-Sanjuan et al., 2019; Lammers et al., 2020a), *Caenorhabditis elegans* (Lee et al., 2019), *Dictyostelium* (Chubb et al., 2006; Muramoto et al., 2012; Corrigan and Chubb, 2014), and human cells (Tantale et al., 2016). Two-color imaging using MS2 and PP7 has enabled the investigation of transcriptional noise in yeast (Hocine et al., 2013), elongation rates in *Drosophila* (Fukaya et al., 2017), and kinetic splicing effects in human cells (Coulon et al., 2014).

To be sure that the repeated tags are accurately reporting on the gene of interest without altering its expression in any way, the tagged and untagged mRNA can be compared via single-molecule fluorescent *in situ* hybridization (smFISH; Pichon et al. (2018)). This is a fixed cell technique one version of which consists of several fluorescent oligonucleotide probes, designed to be complimentary to the target mRNA, hybridize to their target, enabling the detection of single transcripts (Femino et al., 1998). The use of multiple probes increases signal-to-noise ratios because the combined signal from multiple probes is specific while unspecific signals results from single oligonucleotides. MS2 measurements have been validated by smFISH in many studies including those performed in yeast (Tutucci et al., 2018a,b), *Drosophila* embryos (Garcia et al., 2013; Hoppe et al., 2020), human cells (Sheinberger et al.,

2017; Tantale et al., 2016), and mice cells (Yoon et al., 2016).

The syncytial *Drosophila* embryo is an excellent candidate for the use of the MS2 imaging system due to its ease of imaging and well-characterized gene regulatory network. Indeed, the use of MS2 in *Drosophila* has developed into a rich subfield of its own (Garcia and Gregor, 2018). Much of the research into vertebrate transcription, with its relevance for biomedical research, however, remains mostly constrained by studies of fixed tissue. While MS2 has been used in mammalian and human cells, its application *in vivo* in vertebrate animals remains much more limited. Zebrafish embryos present a promising new system in which to utilize MS2 to investigate the dynamics of transcriptional regulation. Like *Drosophila* embryos, zebrafish embryos are ideal for imaging studies. They present some new challenges, however, as their gene regulatory networks are less well characterized than those in *Drosophila* and their larger size can be challenging from both an imaging standpoint—the need to image a large field of view, quickly, over a long period of time—and a data handling standpoint—more than a terabyte of data can be generated from a single imaging session. The works presented here utilize *Drosophila* embryos as an invertebrate model system and extend the MS2 technology to zebrafish embryos to begin to explore the dynamics of transcriptional regulation in vertebrate development.

1.4   Overview of dissertation

Following this introduction which comprises Chapter 1 are four chapters. Chapter 2 through Chapter 4 are derived from published works while Chapter 5 is a work in progress. Finally, appendices A, B, and C provide detailed supplementary information on the work described in Chapters 2, 3, and 4 respectively.

Chapter 2 focuses on a collaboration with the Rushlow Lab (New York University) which was published in *Current Biology* in 2019. This study utilized the MS2 imaging system to investigate the dynamics of transcriptional onset. We found that, following nuclear divisions in the early fruit fly embryo, the pioneer-like factor Zelda plays a role in determining the probability of transcriptional activation, the timing of transcriptional onset, and the transcriptional initiation rates. A combination of immunostaining for the activating transcription factor, Dorsal, and smFISH for the reporter gene mRNA revealed that transcription rates are influenced by the Zelda-dependent accumulation of Dorsal ("hubs") at transcription sites. This indicates that Zelda's ability to increase transcription rates is dependent on enhancing the local concentration of activators rather than an alternative pathway, specific to Zelda. These results imply that these Zelda-dependent hubs play a functional role in regulating transcription.

Chapter 3 describes a study published in *eLife* in 2020. This work uses Zelda-regulated early expressed genes in the fruit fly embryo to carefully dissect equilibrium and non-equilibrium models of transcription focusing on the dynamics of chromatin accessibility. Through a combination of experimental and theoretical approaches, we reveal that the transcriptional dynamics observed experimentally cannot be described by any equilibrium

models including a Monod-Wyman-Changeux (MWC) allosteric model of DNA accessibility and transcription. After a systematic investigation, we ruled out the validity of most non-equilibrium alternatives, including a non-equilibrium version of the MWC model. We proposed an alternative class of models in which transcription factors directly catalyze chromatin accessibility in an energy dependent manner.

Chapter 4 is based on a work published in *PLoS Computational Biology* in 2021. Previous live imaging studies lacked tools with high enough spatiotemporal resolution as well as the methodology to simultaneously investigate the full transcriptional cycle—initiation, elongation, and termination—at the single cell level and to separate the contributions from each of the transcription cycle steps. Thus, these studies focused solely on a single step of the cycle. Our study utilizes the MS2 and PP7 imaging systems as well as the Bayesian inference technique of Markov Chain Monte Carlo to investigate the dynamics of the full transcriptional cycle from initiation to elongation and mRNA cleavage. In this case study in the developing fruit fly embryo, the dual use of MS2 and PP7 loops allows for a two-color reporter in which both the 5' and 3' ends of the transcript are tagged with a unique color. Quantitative estimates of the initiation, elongation, and nascent mRNA cleavage rates were generated by Bayesian inference. The single-cell estimates reveal insights into various transcriptional properties such as the co-variation of transcription cycle parameters, cell-to-cell variability of these parameters, and single mRNA polymerase elongation dynamics.

As evident in Chapters 2 through 4, the use of live measurements of gene expression dynamics has revolutionized the field of *Drosophila* research, revealing new insights into the fundamental process of transcription including unknown mechanisms of regulation as well as disproving some commonly accepted models. Chapter 5 expands on this revolution by extending the tools developed in *Drosophila* to the study of the dynamics of transcriptional regulation in vertebrates. We adapted the MS2 imaging system for use in live, developing zebrafish embryos. With this work we aim to understand how the embryo integrates dynamic inputs to regulate and coordinate transcription across cells and tissues. As a case study and opportunity to hone these new investigative tools, we focus on one of the better characterized and more fundamental gene networks in zebrafish development: the network that lays out the vertebrate body segments during somitogenesis. We successfully established several transgenic MCP zebrafish lines as well as several lines expressing an MS2 reporter for a gene involved in regulating somitogenesis. Preliminary measurements of transcriptional dynamics in zebrafish embryos challenge our expectations. Specifically, our measurements show that single-cell transcriptional dynamics are characterized by punctuated, periodic bursts of activity, rather than matching the smooth oscillations seen in the fluorescently tagged protein product of the gene.

Chapter 2

The *Drosophila* pioneer factor Zelda modulates the nuclear microenvironment of a Dorsal target enhancer to potentiate transcriptional output

Foreword

The common theme of my research in graduate school is the role dynamics play in transcriptional regulation. My advisor, Hernan Garcia, who started his lab just a few months before I started graduate school, was using live imaging to experimentally investigate transcription in developing fruit fly embryos. This live imaging allows us access to the dynamics of both transcription factors regulating gene expression as well as the resulting transcriptional output. This rich data, in turn, allows us to test dynamic biophysical models of transcriptional regulation.

Some of my most fruitful research experiences in graduate school resulted from collaboration both within my own lab and with other labs. I learn more from these collaboration than I do working alone. One of my most successful collaborations with another lab was with Christine Rushlow's Lab at NYU.

When I joined the Garcia Lab, most of my work focused on a transcription factor called Zelda. Christine Rushlow pioneered much of the work on this vital *Drosophila* regulatory protein. She was interested in using the imaging system that is the Garcia Lab's speciality (the MS2 system) to further investigate Zelda's role in regulating the expression dynamics of *sog*, a dorsal-ventral patterning factor of the fruit fly embryo. So, some of the first data that I collected after joining the Garcia lab involved using a *sog* reporter construct that her lab had designed to remove Zelda binding sites (and therefore Zelda input). However, preliminary

data from this transgene, as well as others that had Zelda sites removed, did not look like data from embryos that lacked Zelda protein. Somehow, it seemed, Zelda was still affecting the expression of these transgenes.

In the Garcia Lab, I learned to question our investigative tools, reporters, and experimental design for potential flaws or side-effects. So I began to investigate our reporter constructs. Around this time Natalie Dostatni's lab reported finding 36 Zelda binding sites in the original MS2 DNA sequence that affected reporter transgene expression (Lucas et al., 2018). They redesigned the MS2 sequence to remove the Zelda sites in order to remove potential Zelda influence that occurs outside of the enhancer under investigation.

When it came to the unexpected results from our reporter transgenes, I realized that our nonrepetitive MS2v5 DNA sequence, similarly, inadvertently contained many Zelda binding sites. This meant that, although Christine's Lab had carefully removed Zelda as a factor in the regulatory DNA, it may have been affecting the transgene's expression because of its binding in the MS2 component of the transgene. I needed to redesign the MS2v5 sequence to remove these Zelda sites to be sure that we had completely removed Zelda's potential effects on the transgene's expression.

I followed the design pattern for the MS2v5 sequence (Wu et al., 2015) and removed the Zelda binding sites using an alignment matrix provided by Melissa Harrison at the University of Wisconsin-Madison. Christine's lab then used this sequence to create a new *sog* reporter transgene and took over the data acquisition and analysis with some guidance from Hernan and myself. They reveal new information about Zelda's role in transcriptional dynamics and then took the project even further by examining "hubs" of high local concentration of the Dorsal activator, which turn out to be Zelda dependent. For the first time, these hubs are connected to transcriptional output.

Ultimately, this project, which was published in *Current Biology*, allowed us to investigate the expression dynamics of *sog* and the role Zelda plays in them at high spatiotemporal resolution. The next chapter jumps off from this point, investigating another fruit fly gene that is dependent on Zelda for proper expression and delving into the details of theoretical models of transcription.

2.1 Summary

Connecting the developmental patterning of tissues to the mechanistic control of RNA polymerase II remains a long-term goal of developmental biology. Many key elements have been identified in the establishment of spatial-temporal control of transcription in the early *Drosophila* embryo, a model system for transcriptional regulation. The dorsal-ventral axis of the *Drosophila* embryo is determined by the graded distribution of Dorsal (Dl), a homolog of the nuclear factor κB (NF-κB) family of transcriptional activators found in humans (Steward et al., 1984; Stathopoulos and Levine, 2002). A second maternally deposited factor, Zelda (Zld), is uniformly distributed in the embryo and is thought to act as a pioneer factor, increasing enhancer accessibility for transcription factors, such as Dl (Liang et al., 2008;

Schulz et al., 2015; Nien et al., 2011; Harrison and Eisen, 2015; Foo et al., 2014; Sun et al., 2015; Harrison et al., 2011). Here, we utilized the MS2 live imaging system to evaluate the expression of the Dl target gene *short gastrulation* (*sog*) to better understand how a pioneer factor affects the kinetic parameters of transcription. Our experiments indicate that Zld modifies probability of activation, the timing of this activation, and the rate at which transcription occurs. Our results further show that this effective rate increase is due to an increased accumulation of Dl at the site of transcription, suggesting that transcription factor "hubs" induced by Zld (Mir et al., 2017) functionally regulate transcription.

2.2 Results

Our study focused on the Dorsal (Dl) target gene *sog*, as its expression domain spans a large dynamic range of the Dl gradient, allowing us to examine how Zelda (Zld) potentiates Dl activity across the dorsal-ventral axis. Previous experiments have demonstrated that the lateral stripe of *sog* expression narrows dramatically in *zld*-null embryos (Nien et al., 2011; Kanodia et al., 2012) (Figs. 2.1A and 2.1B)and that progressively removing Zld DNA binding sites from the *sog* shadow (distal) enhancer shrinks the domain of activation of reporter genes in a linear manner (Foo et al., 2014). In order to understand how Zld influences transcription at different points along the Dl gradient, we revisited these constructs with the aim of visualizing transcription in real time by adding 24 MS2 loops to the 5' end of the *lacZ* reporter. Because previously utilized MS2 loops (Larson et al., 2011a; Forrest and Gavis, 2003; Garcia et al., 2013; Lucas et al., 2013) contained potential Zld binding sites (Lucas et al., 2018), we revised the MS2v5 (Wu et al., 2015) sequence to make a Zld binding-site-free non-repetitive version, referred to as MS2v5(-TAG) (see Section 2.4). Constructs also contained either the *sog* shadow (distal) enhancer (Hong et al., 2008; Ozdemir et al., 2014) with its three native canonical Zld binding sites, CAGGTAG (hereafter referred to as "3TAG"), or without these sites (hereafter referred to as "0TAG"; Fig. 2.1C; see Section 2.4 for enhancer sequences Foo et al. (2014)). The narrowing effect of removing Zld binding sites was confirmed by *in situ* hybridization (Figs. 2.1D and 2.1E).

By crossing these transgenic reporter lines to females expressing the MCP (MS2 coat protein)-GFP fusion gene during oogenesis (Garcia et al., 2013), we visualized the transcriptional activation of each reporter as fluorescent foci (see Fig. 2.1C and Section 2.4). These embryos also express H2Av(histone 2A variant)-RFP (Saint and Clarkson, 2000), allowing us to track nuclear cycles and record transcriptional activation events in space and time. We performed confocal live imaging over the course of nuclear cycles 10–14 (NC10–NC14), tracking the activation of the 3TAG and 0TAG reporter genes (Videos A.2.1 and A.2.2). To validate that the MS2 transgenes behaved as expected, we examined transcriptional activation events in space and time and compared those to expression as assessed by conventional *in situ* analysis. We find that the 3TAG construct is activated as early as NC10, although activation of the 0TAG construct is delayed until NC11 or NC12 (Figs. 2.2A and 2.2B; Videos A.2.1and A.2.2; also see additional Videos A.2.3, A.2.4, A.2.5, and A.2.6), in agreement with previously
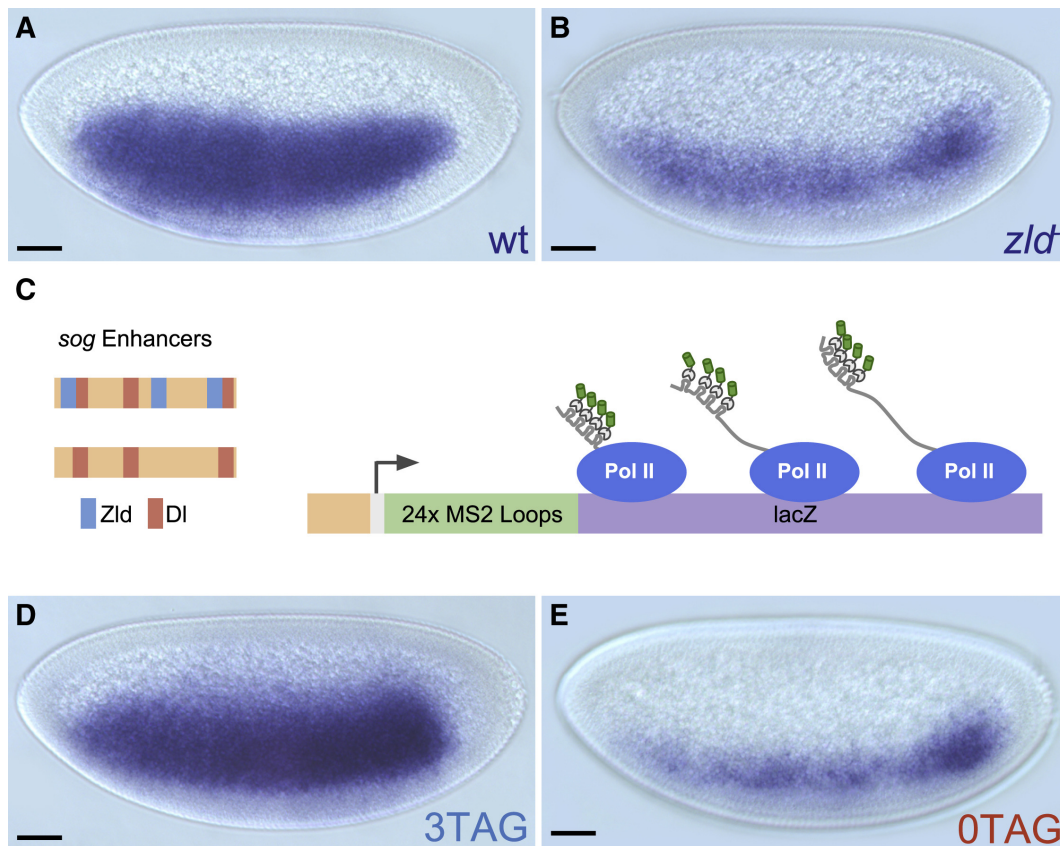
Figure 2.1: Zld potentiates Dl activity at the *sog* enhancer. (A and B) Conventional enzymatic *in situ* hybridization staining of *sog* in wild-type (A) and *zld* mutant (B) NC14 embryos. (C) Schematic representation of transgenes. MS2 loops have been incorporated into the 5' end of the transcript upstream of a *lacZ* reporter sequence. (D and E) *In situ* hybridization staining of the engineered MS2v5(-TAG) *lacZ* transgenic embryos, showing that 3TAG (D) and 0TAG (E) expression is similar to the expression of *sog* in wild-type and *zld* mutants, respectively. Scale bar shown in bottom left corner of each embryo represents 50 μm.

published results of *sog* activation in *zld* mutants (Nien et al., 2011).

To compare the spatial differences in activation, we divided the expression domain of *sog* into five discrete zones, with zone 1 comprising the mesoderm and all subsequent zones defined by 20-μm-width bands moving sequentially toward the dorsal midline of the embryo, diagrammed in Figure 2.2C. The *in situ* experiments predict that the most dorsal zones imaged would show few active nuclei in 0TAG embryos, and this was the case. Although 3TAG embryos showed similar numbers of active nuclei in each zone across all cycles (NC12–NC14), with the exception of zone 1 in NC14 due to ventral repression by Snail (Fig. 2.2D), in 0TAG embryos, the more dorsal the zone, the fewer the number of active nuclei (Fig. 2.2E). Collectively, these qualitative observations are in accordance with what is currently known about how Zld participates in transcriptional activation and provide evidence that our transgenes are faithfully reporting on the transcriptional activity of *sog* in the presence or absence of Zld.

In addition to allowing qualitative assessment of transcriptional activation, MS2 reporters continually output information on the state of transcription over time, enabling an analysis of the timing of each activation event within a nuclear cycle (Garcia et al., 2013). This was performed by measuring the time between anaphase of NC12 and the appearance of fluorescent foci in NC13 and plotting the results as cumulative distribution curves (Figs. 2.2F–2.2H). This analysis showed that nuclei in 3TAG embryos express simultaneously across the domain of expression (Fig. 2.2G; Video A.2.1). In stark contrast, we observed a significant position-dependent delay of activation in 0TAG embryos, where the ventral nuclei activate transcription well before lateral nuclei (Fig. 2.2H; Video A.2.2). This is presumably due to the highly dynamic nature of the Dl gradient, whereby Dl levels increase within and across nuclear cycles (Kanodia et al., 2009; Liberman et al., 2009; Reeves et al., 2012). Here, the 0TAG reporter is effectively acting as a readout for nuclear Dl concentration, suggesting that, in the absence of Zld binding sites, the *sog* enhancer responds to Dl levels in a concentration-dependent manner, rather than the binary switch-like response seen in the presence of Zld.

Knowing that activation is altered in 0TAG embryos, we next examined the internal kinetic features of transcription. We focused principally on two phases of transcription, which are described in Figures 2.3A and 2.3B using representative nuclei from each genotype at NC13, with the signal over time quantified in Figures 2.3C and 2.3D. The first was "ramp up," an early phase where polymerase molecules first begin to elongate as transcription begins. Here the rise in MS2 signal is attributed to polymerases accumulating over the gene body as they transcribe the MS2 loops and continue to elongate. The length of the ramp-up phase is commonly thought of as the time in which a single polymerase molecule has traversed the entire gene body (Garcia et al., 2013). The transition to the next phase, "steady state" transcription, is reached when the rate of polymerase loading is matched by rate of polymerase unloading, diagrammed in Figure 2.3E. Here, the MS2 signal levels off and fluctuates within a narrow range, as there is no net gain of nascent transcripts. We have included an equation demonstrating that the signal strength at steady-state transcription can be understood as the average gap between polymerase molecules on the gene body (Fig. 2.3E).

Using the duration of the ramp-up phase, which can be referred to as the "time to steady
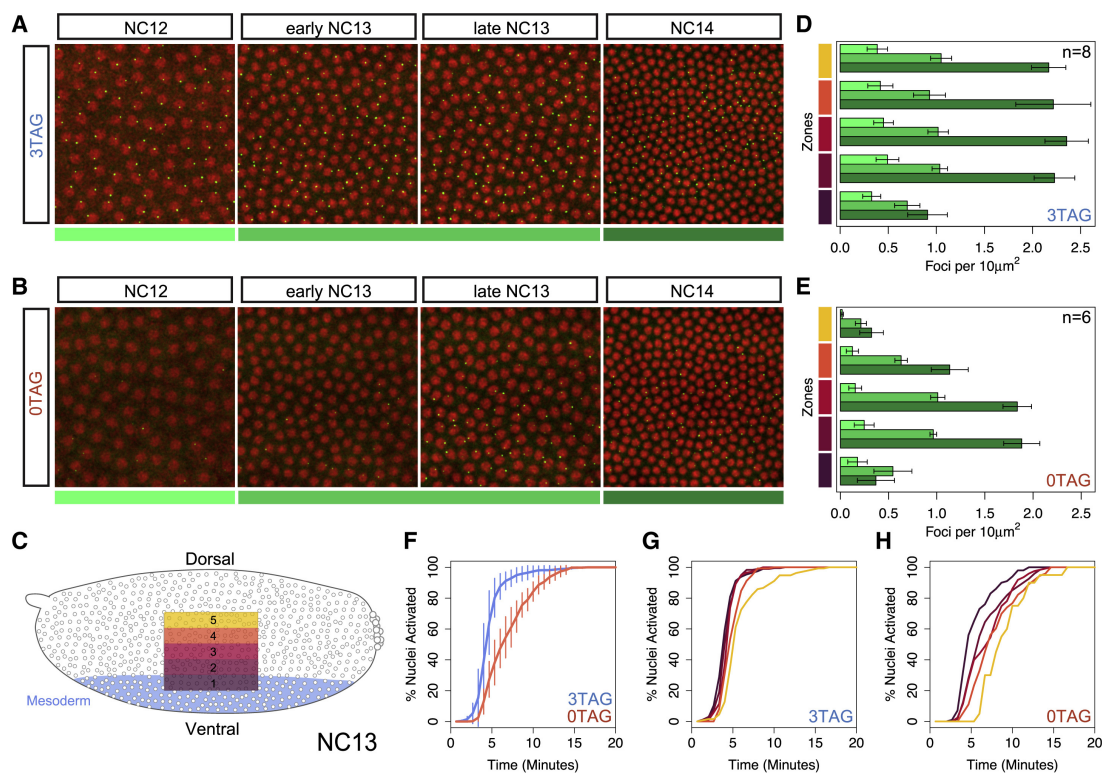
Figure 2.2: MS2 imaging reveals a position-dependent transcriptional delay in the absence of Zld binding sites. (A and B) Frames taken from live imaging Videos A.2.1 (A) and A.2.2 (B) that track transcription (green spots) from NC12 to NC14 as indicated and color coded below, NC12 (light green), NC13 (medium green), and NC14 (dark green). Nuclei (red) have been labeled using maternally loaded H2Av-RFP. (C) Bars on right side follow five zones along the dorsal-ventral axis with ventral mesoderm on bottom (zone 1) as diagrammed in the embryo schematic with blue shading defining the presumptive mesoderm of the embryo. (D and E) Quantification of the number of expressing nuclei in 3TAG (D) and 0TAG (E) NC12–NC14 embryos (color coded as in A and B) agrees with conventional *in situ* analysis, showing markedly fewer active nuclei in 0TAG embryos across consecutive nuclear cycles, especially in zones 4 and 5. In total, 8 3TAG embryos and 6 0TAG embryos were analyzed as indicated in the bar plots and plotted with error bars representing one SD of all values collected for each cycle and bin. For additional videos, see Videos A.2.3, A.2.4, A.2.5, and A.2.6.(F–H) Cumulative distribution curves of nuclei that activate transcription in NC13, excluding nuclei that never activate in NC13. Time 0 on the x axis is the start of anaphase of the previous cycle, NC12. All zones concatenated with delay values across genotypes in (F) with variance across biological replicates indicated with vertical lines showing one SD of all embryos measured. 3TAG embryos activate transcription simultaneously across the expression domain (G), and 0TAG embryos show a delay dependent on the nucleus' position in the Dl gradient (H).
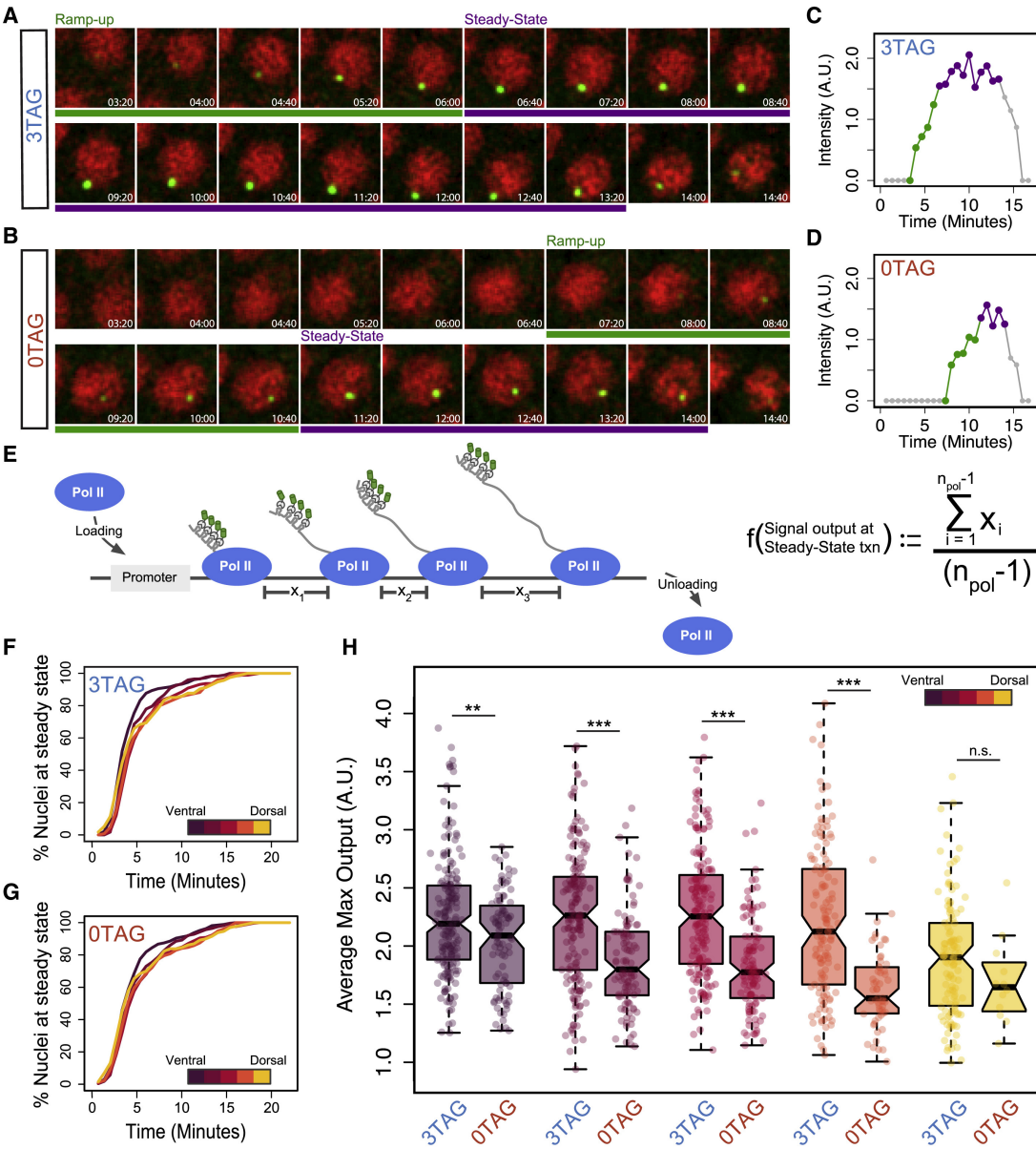
Figure 2.3: Zld promotes full saturation of polymerase on the gene body during transcriptional elongation. (A and B) Representative single nuclei tracked over NC13 from Videos A.2.1 (A) and A.2.2 (B). Time stamp (min) is shown in bottom right corner of each frame (time 0 is defined as the start of NC12 anaphase). Ramp-up and steady-state phases of transcription are highlighted with green and purple bars, respectively. Caption continues on next page.

Figure 2.3: Caption continued from previous page. (C and D) Quantification of signal intensity over time in 3TAG (C) and 0TAG (D) from representative nuclei shown in (A) and (B), respectively. Phases of transcription are highlighted with corresponding colors as in (A) and (B). Ramp up is calculated as the length of time between detection above background of the MS2 focus and max output (averaged; see Section 2.4). (E) Schematic representation of steady-state transcription, where the gene body is decorated with elongating RNA polymerases and the rate of loading is roughly matched by the rate of unloading. X values show the spacing between polymerase molecules. Spacing of polymerase molecules can be inferred from the signal output at steady state using the equation shown. (F and G) Cumulative distribution curves of the percentage of nuclei that have reached steady state in different zones of 3TAG (F) and 0TAG (G) embryos. (H) Average intensity at steady state (NC13) plotted as boxplot distributions over all five zones of the *sog* expression domain. In total for all zones, 855 and 460 nuclei were analyzed for 3TAG and 0TAG, respectively, from 8 3TAG and 6 0TAG embryos (see additional Videos A.2.1, A.2.2, A.2.3, A.2.4, A.2.5, A.2.6). Significant differences between all zones except zone 5 were found using a Welch's t test between the genotypes. 3TAG embryos show little difference over the first four zones, and 0TAG embryos show progressive loss in signal intensity over the dorsal-ventral axis.

state," we can calculate the number of nuclei that have reached steady-state transcription as cumulative distribution curves, with the percentage of all active nuclei at steady state plotted over time (Figs. 2.3F and 2.3G). There is a striking similarity between the two genotypes, indicating that Zld does not act on the speed of polymerase. In addition, the time to steady state is similar in each of the different zones, suggesting that nuclear Dl concentration has little influence on polymerase elongation rate. In contrast, when signal intensity values of steady-state transcription are averaged for each nucleus (Fig. 2.3H), it appears that both Zld and Dl are modulating the strength of transcription. Similar to our observations regarding the onset of transcriptional activation, the 3TAG reporter shows comparable max output across multiple zones until the most extreme end of the Dl gradient (zone 5), whereas the 0TAG reporter shows a progressive loss of max output across the entire gradient (Fig. 2.3F), indicating that transcriptional output rate has become a function of nuclear Dl concentration. These results suggest Zld acts upstream of elongation, for example, to either increase RNA polymerase II loading or decrease the length of pausing experienced by a given polymerase molecule. Either of these regulatory steps would affect the mean spacing of polymerase molecules at max output.

This behavior of Zld inducing uniform transcriptional activation and output across a transcriptional activator gradient could be explained by Zld's reported ability to promote the formation of transcription factor "hubs" (Mir et al., 2017, 2018; Dufourt et al., 2018).

By raising the local concentration of Dl at the site of transcription, Zld may effectively flatten the gradient of Dl experienced by the enhancer and therefore unify the levels of transcriptional output in regions of low level Dl. To test this hypothesis, we used a previously described method to examine transcription factor enrichment at sites of nascent transcript formation in *Drosophila* embryos (He et al., 2011; Xu et al., 2015). By costaining fixed embryos with an anti-Dl antibody and a single molecule (sm) fluorescent *in situ* hybridization (FISH) probe targeting the *lacZ* reporter transcript (Little et al., 2013), we could quantify the concentration of Dl protein adjacent to foci of transcription. Figure 2.4A shows the Dl gradient at comparable positions in 3TAG and 0TAG embryos. Signal overlap between puncta of Dl staining and *lacZ* staining, the presumed site of transcription, can be seen in 3D contour maps, where the surface represents the level of Dl antibody signal and the site of transcription is mapped onto the texture of the contour. We classified nuclei as either having a high, mid, or low level of Dl based on binning all nuclei imaged according to their average Dl signal intensity, which correspond spatially to zones 1, 2, and 3 in Figures 2.2 and 2.3.

Figure 2.4C uses a modified approach demonstrated by Tsai et al. (2017), where the radial intensity of the Dl antibody stain is plotted to visualize the nuclear microenvironment that surrounds a site of active transcription (*lacZ* staining). Because the nuclear concentration of Dl changes across the gradient, we divided voxel intensity by the average voxel intensity found within a nucleus. In this way, we could normalize across nuclei by defining our measurement as a unitless index describing the relative enrichment of signal at a given site of transcription, where a value of 1 indicates no enrichment. Additionally, we included a set of random points within nuclei as a control. For a full breakdown of individual enrichment curves, see Figure A.1. As predicted, we see a progressive loss in enrichment over the gradient in 0TAG embryos and a measurable gain in enrichment in 3TAG embryos, indicating that Zld's ability to drive higher transcriptional output is based on enhancing the local concentration of existing transcriptional activators rather than utilizing an additional Zld-specific activation pathway. Importantly, these results strongly suggest a functional link between Zld's reported ability to induce transcription factor aggregates (Mir et al., 2017) and transcriptional output, an important first step toward a complete understanding of Zld's ability to control gene expression.

2.3 Discussion

The precise logic governing *cis*-regulatory elements is still an evolving field after decades of research. The role of pioneer factors, such as Zld, in modifying chromatin has increased our understanding of how patterning transcription factors, such as Dl and Bicoid (Bcd), access their target enhancers (Schulz et al., 2015; Foo et al., 2014; Sun et al., 2015; Li and Eisen, 2018; Xu et al., 2014); however, questions persist as to the events that occur at the site of transcription. Several recent reports have suggested that the accumulation and stable association of transcription factor aggregates, or hubs, is important for proper transcriptional output (Mir et al., 2017, 2018; Dufourt et al., 2018; Tsai et al., 2017). Additionally, the
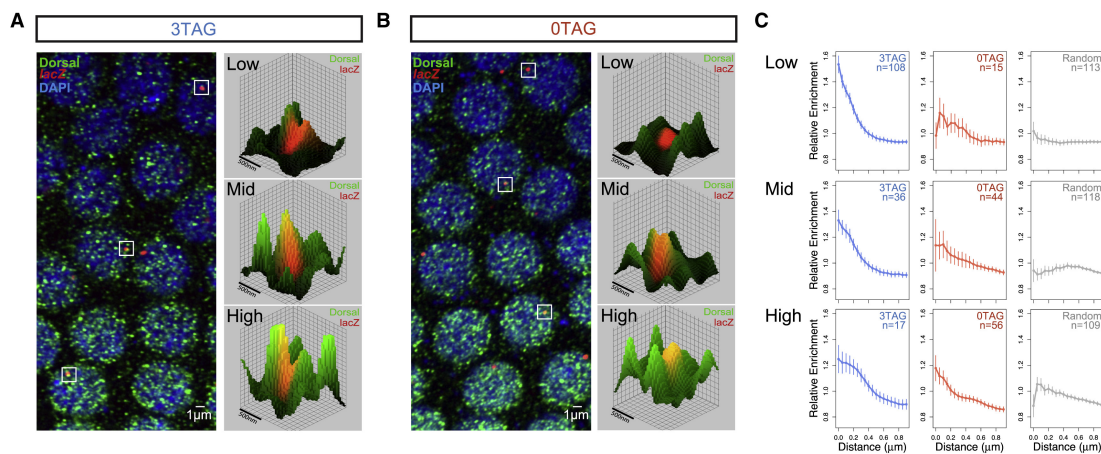
Figure 2.4: Zld increases the local concentration of Dl at the site of transcription. (A and B) Confocal images of NC13 embryos stained with anti-Dl antibodies and smFISH probes for the *lacZ* reporter genes 3TAG (A) and 0TAG (B). Dl staining appears highly punctate, indicating the possible presence of high-Dl nuclear microenvironments. Sites of active transcription are visualized as red nuclear foci that can be localized in 3D space. Select foci were isolated and visualized in 3D contour maps, where the height of surface represents the intensity of the Dl staining. A high incidence of FISH signal overlapping with Dl microdomains was observed, suggesting the concentration of Dl may have an impact on transcription. (C) The distributions of Dl signal within the microdomain of transcribing foci (see Fig. A.1 for individual enrichment curves). In regions of high nuclear Dl, both genotypes show similar distributions, but a difference is detected in regions where nuclear Dl begins to drop. Control distributions were prepared using random places in the nucleus. The numbers of nuclei (n) used for the analysis are indicated. Three embryos for each genotype were used. Error bars indicate the SEM.

Drosophila transcription factor Bcd is enriched in Zld hubs, particularly in nuclei with lower overall Bcd, suggesting that Zld interacts with transcription factors to raise their local concentration (Mir et al., 2018). Our results manipulating Zld binding at the enhancer (site of transcription) agree nicely with these recent findings and for the first time suggest a direct impact of these transcription factor hubs on transcription itself.

Our experiments identify two key parameters where Zld modifies the activity of a Dl-responsive enhancer. The first parameter is the onset of transcription across the domain of *sog*, where a position-dependent delay in transcriptional activation of the reporter was observed in the 0TAG embryos. We believe that the uniformity of this response is the result of Zld's pioneering activity to ubiquitously lower the nucleosome barrier from regions of DNA in close proximity to its DNA binding motif. Freeing up enhancers may then allow Dl to be bound more quickly at low concentrations, which may in turn lead to local enrichment of

Dl (Fig. 2.4C). In the absence of Zld, Dl must compete directly with nucleosomes to access its DNA binding sites. This competition could be more effective at high concentrations of Dl, thus leading to the concentration-dependent effects observed in 0TAG. The second parameter controlled by Zld is the uniformity of the transcriptional output over the course of a nuclear cycle. Our MS2 data of 3TAG embryos showed remarkably similar levels of total transcription in all measured positions save for the most extreme dorsally located nuclei. Our results of higher Dl enrichment in 3TAG embryos in nuclei with low Dl tracks well with the measurements of transcription. However, it remains to be seen whether these two transcriptional parameters (timing and output) are connected by a single mechanistic step mediated by Zld binding to an enhancer.

More broadly, these experiments demonstrate the influence that Zld has on global dynamics of tissue patterning. Morphogen gradients supply positional information through an analog signal: a contiguous modulation in morphogen concentration directly encodes location along a developmental axis. When creating a broad domain of expression from this signal, a conversion from analog to digital information must occur; a varying amount of transcriptional activator, in this case Dl, must be reliably converted into a uniform transcriptional response. Our work suggests that Dl's interaction with Zld allows it to produce near identical transcriptional output over a large range of nuclear Dl concentrations, and in the absence of that interaction, output is purely reflective of the underlying gradient. As uniform transcriptional domains that span large sections of morphogen gradients are a common motif across developmental systems, we believe our work helps shed light on a common principle of cellular fate decisions.

2.4 Methods

Table 2.1: Key resources table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Sheep anti-DIG-AP antibody	Roche	Cat# 11093274910, RRID:AB_2734716
Mouse anti-DL antibody (7A4)	Developmental Studies Hybridoma Bank	Cat# anti-Dorsal 7A4, RRID:AB_528204
Alexa fluor 488 goat anti mouse secondary antibody	ThermoFisher Scientific	Cat# A-11001, RRID:AB_2534069
Chemicals, Peptides, and Recombinant Proteins		
5-Bromo-4-chloro-3-indolyl phosphate p-toluidine salt(BCIP)	Roche	Cat# 10760994001
4-Nitro blue tetrazolium chloride(NBT)	Roche	Cat# 11585029001
DIG RNA labeling mix	Roche	Cat# 11277073910
RNA FISH Hybridization Buffer	Stellaris (LGC Biosearch Technologies)	Cat# SMF-HB1-10
DAPI	Sigma-Aldrich	Cat# D9542
Aqua-Poly/Mount	Polysciences	Cat# 18606-20
Number 1.5 glass coverslips	Fisher Scientific	Cat# 22266858
Gibson Assembly Master Mix	New England Biolabs, Inc	Cat# E2611S
lacZ Atto633 smFISH Probe	Shawn Little's lab	Little et al. (2013)
lacZ ISH DIG RNA Probe	Rushlow Lab	
sog ISH DIG RNA Probe	Rushlow Lab	

Table 2.1: Key resources table

Experimental Models: *Drosophila* Strains		
y[1] w[1118]	Bloomington Drosophila Stock Center	Cat# 6598
y[1] w[*]; P{*w*[+mC] = His2Av-mRFP1}II.2; P{*w*[+mC] = *nos*-MCP.EGFP}2	Bloomington Drosophila Stock Center	Cat# 60340, RRID:BDSC_60340
Maternal Triple Driver(MTD)-Gal4: P{*COG*-GAL4:VP16}; P{Gal4-*nos*.NGT}40; P{*nos*-Gal4-VP16}	Bloomington Drosophila Stock Center	Cat# 31777, RRID:BDSC_31777
UAS-shRNA-*zld*	Rushlow Lab	Sun et al. (2015)
sog 3TAG-MS2-*lacZ*	Yamada et al. (2019)	N/A
sog 0TAG-MS2-*lacZ*	Yamada et al. (2019)	N/A
Recombinant DNA		
pib-evepr-ms2v5(-TAG)-*lacZ* plasmid	Yamada et al. (2019)]	N/A
Software and Algorithms		
FIJI (ImageJ)	NIH	http://fiji.sc
MATLAB	The Mathworks Inc.	https://www.mathworks.com
Imaris	Bitplane	http://www.bitplane.org
R	The R Foundation	https://www.r-project.org
LAS X	Leica Microsystems Inc	https://www.cellularimaging.nl/leica-las-x/

Table 2.1: Key resources table

N2012	CarlZE Zeiss Inc	https://www.zeiss.com/corporate/int/home.html
Other		
Confocal microscope	Leica	SP8
Confocal microscope	Zeiss	LSM 880
Microscope	Zeiss	Axioskop
Digital camera for microscopy	Zeiss	AxioCam MRc
Power meter(X-cite)	Lumen Dynamics Group Inc, Canada	Model # XR2100
Breathable membrane (Lumox Film)	Sarstedt AG & Co.; Nümbrecht, Germany	Cat# 94.6077.317
Plastic microscope slide (3D printed)	Sculpteo; Créteil, France	N/A

2.4.1 Contact for reagent and resource sharing

Requests for any information and requests for resources or reagents should be directed to the Lead Contact, Christine Rushlow (car2@nyu.edu).

2.4.2 Experimental model and subject details

All flies were grown on standard fly (*Drosophila melanogaster*) cornmeal-molasses-yeast media. *y[1]w[1118]*(used as wild-type flies), *zld* shmir (*zld-*) (see Section 2.4.3 below) (Sun et al., 2015), and transgenic embryos (3TAG and 0TAG) were collected on yeasted grape juice agar plates. Flies of the genotype *y[1] w*; P{His2Av-mRFP1}II.2; P{nos-MCP.EGFP}2* (Bloomington Stock Number 60340) carried two transgenes, one on chromosome 3, *P{nos-MCP.EGFP}2*, which expresses the MS2 coat protein (MCP) fused to EGFP under the control of the *nanos* promoter active in oogenesis, and the other on chromosome 2, *P{His2Av-mRFP1}II.2*, which expresses RFP-tagged His2Av in all cells under the control of *His2Av*. MS2 transgenes were constructed in the following manner: MS2 loop sequences were revised since previously used MS2 loops (Larson et al., 2011a; Forrest and Gavis, 2003; Garcia et al., 2013; Lucas et al., 2018; Wu et al., 2015) contained potential Zld binding sites (Nien et al., 2011; Garcia et al., 2013; Lucas et al., 2018). The new MS2 loops sequence, MS2v5(-TAG) (see Section 2.4.3

for DNA sequence) was placed in between the eve minimal promoter and a *lacZ* reporter gene (pib-evepr-ms2v5(-TAG)-*lacZ* plasmid), then subcloned into an attB vector (pBPhi) containing *sog* enhancers with (3TAG) or without (0TAG) Zld binding sites (Foo et al., 2014) (Section 2.4.3). Transgenic lines carrying these constructs were generated by phiC31 integration in the 53B2 landing site (VK00018), Bloomington stock number 9736 (Groth et al., 2004; Bischof et al., 2007) by BestGene.

2.4.3 Method details

Depletion of maternal *zld*

Embryos were collected from females depleted of *zld* RNAs by RNAi prepared in two crosses (Sun et al., 2015): 1st cross ♀ +;+;UAS-shRNA-*zld* × ♂ P{*COG*-GAL4:VP16};P{Gal4-*nos*.NGT}40;P{*nos*-Gal4-VP16}; 2nd cross G1♀ P{*COG*-GAL4:VP16}/+;P{Gal4-*nos*.NGT}40/+;P{*nos*-Gal4-VP16}/UAS-shRNA-*zld* × ♂ yw

Sequence of the *sog 3TAG* and *0TAG* enhancers

sog 3TAG

426 bp enhancer sequence (Zld binding sites underlined):

GTTTCAGCGGAAC<u>CAGGTAG</u>GCTGGTCGATCGGAAATTCCCACCATACACA
TGTGGCTATAATGCCAACGGCATCGAGGTGCGAAAACAGATGCAGCCTCATA
AAAGGGGCGCAGATAAGGTCGCGGTTGCGTGGGAAAAGCCCATCCGACCAG
GACCAGGACGAAGCAGTGCGGTTGGCGCATCATTGCCGCCATATCTGCTATT
C<u>CTACCTG</u>CGTGGCCATGGCGATATCCTTGTGCAAGGATAAGGAGCGGGGATC
ATAAAACGCTGTCGCTTTTGTTTATGCTGCTTATTTAAATTGGCTTCTTGGCG
GGCGTTGCAACCTGGTGCTAGTCCCAATCCCAATCCCAATTCCAATCCCAATCC
ATATACCATATCCAATGCATT<u>CTACCTG</u>TCCTGGGAATTTCCGATCTGGCCGCA
CCCATAT

sog 0TAG

426 bp enhancer sequence (mutated Zld binding sites underlined):

GTTTCAGCGGAAC<u>CCAACAA</u>GCTGGTCGATCGGAAATTCCCACCATACACA
TGTGGCTATAATGCCAACGGCATCGAGGTGCGAAAACAGATGCAGCCTCATA
AAAGGGGCGCAGATAAGGTCGCGGTTGCGTGGGAAAAGCCCATCCGACCAG
GACCAGGACGAAGCAGTGCGGTTGGCGCATCATTGCCGCCATATCTGCTATT
CT<u>TGTTGG</u>GCCATGGCGATATCCTTGTGCAAGGATAAGGAGCGGGGATCATAA
AACGCTGTCGCTTTTGTTTATGCTGCTTATTTAAATTGGCTTCTTGGCGGGCG

TTGCAACCTGGTGCTAGTCCCAATCCCAATCCCAATTCCAATCCCAATCCATAT
ACCATATCCAATGCATT<u>TTGTTGG</u>GGAATTTCCGATCTGGCCGCACCCATAT

in situ hybridization

Embryos were collected and aged to be 1-3 hours old at room temperature and dechorionated in Clorox for two minutes. They were then fixed in 4% formaldehyde (1X PBS) and an equal volume of heptane for 25 minutes while shaking vigorously. Devitellinization was performed by pipetting off the bottom fixative phase and adding 4 mL of methanol and shaking vigorously for 30 s. Embryos were rinsed in methanol and transferred to ethanol for storage at -20°C. Hybridization of fixed embryos used a standard *in situ* hybridization (ISH) protocol and DIG-labeled *sog* cDNA or *lacZ* RNA antisense probes (Foo et al., 2014); hybridized at 55°C overnight). Visualization of the labeled probe was done using anti-DIG-AP (alkaline phosphatase) antibodies (Roche Biochemicals) followed by histochemical enzymatic staining reagents (Roche Biochemicals). For smFISH, Atto-633 conjugated probe sets complementary to *lacZ* (gift from Shawn Little) (Little et al., 2013) were used in hybridization experiments using Stellaris (LGC Biosearch Technologies) reagents and protocols.

Antibody staining

Antibody staining was performed at 4°C for 16 hours followed by three 20 minute washes in PBS + 0.1% Tris pH 7.0. Anti-Dl antibody (Dl_7A4) was obtained from the Developmental Studies Hybridoma Bank and used at 1:50 dilution. Embryos were then stained with Alexa Fluor 488 anti-mouse secondary antibody (Invitrogen, ThermoFisher Scientific) for 1.5 hours and washed in the same manner. After DAPI (D9542, Sigma-Aldrich) staining for 20 minutes, embryos were mounted on microscope slides using Aqua-Poly/Mount (Polysciences) and Number 1.5 glass coverslips (Fisher Scientific). Embryos were imaged with Zeiss Axiophot DIC optics and a Zeiss Cam and ZEN2012 software.

Construction of MS2v5(-TAG) vector

In order to identify potential Zld binding sites in the DNA sequence encoding MS2v5 (Wu et al., 2015), the sequence was analyzed with a Zld alignment matrix (courtesy of Melissa Harrison (Harrison et al., 2011)) using the Advanced PATSER entry form online (http://stormo.wustl.edu/consensus/cgi-bin/Server/Interface/patser.cgi) (Hertz and Stormo, 1999). PATSER was run with settings Seq. Alphabet and Normalization "a:t 3 g:c 2" to provide the approximate background frequencies as annotated in Berkeley Drosophila Genome Project (BDGP)/Celera Release 1. Complementary sequences were also scored. When PATSER identified a site scoring 3 or higher, one to three bases were modified to reduce the score of the site. After modifying the sequence, it was run through PATSER again to check that no new binding sites were inadvertently created. The process was repeated

until all sites scored 3 or higher were abolished. Sites that occurred on sequences encoding MS2 loops were carefully modified to maintain the pattern set forth in Wu et al. (Wu et al., 2015). Potential binding sites for GAGA Factor were simultaneously abolished during this process using the same methods. The entire MS2v5(-TAG) sequence was constructed as a G-block by GenScript, confirmed by sequencing, and incorporated into our reporter construct by Gibson Assembly (New England Biolabs, Inc.).

MS2v7 sequence (24 new loops in italics)

aacctacaaagg*cgtggaggatcaccccacg*ccacacttcacaatcaaggggtacaatacacaag*ggtggaggaacaccccacc*ctccagacacattacacagaaatccaatcaaaca*gaagcaccatcagggcttc*gaaatttatctcaaaaaactacaaacag*taatcaggatcaccgatta*cgtctgcaatatacgtcaaacgacgcccacgacg*ggaggacgatcacgcctcc*cgaatatcggcattcgtggctttcgaattcaatc*cgtggagcatcagcccacg*gacccaatcagagtcgaatagaactcgactttcg*cgaagagcatcagccttcg*cgccattcttacacaaaccatagtctccccttgt*cgaacagcatcagcgttcg*agcccagtacccaactcaagaaaattttactccc*gaagcagcatcagcgcttc*ggccccaagaatacatccccaacaaaatcacatc*cgagcaccaacagggctcg*gttgtttcttgtggatagtcgacaaaccgac*caaggaccatcaggccttg*tcaccaacaagacaaaaactactcttctc*gaagcagcatcagcgcttc*ctcgagcatacattgtgcctatttcttg*ggtggacgatcacgccacc*catcgcctgacgaatttcaaaacacggacaagga*cgagcacgagcacggctcg*ccacgtccaatacgattacttaggtttcggga*tcacgatcacggatcc*cgcaccttcatcacttccactcaggacatt*caagcaagcacgatca cggcttg*ctccacaagtctcaaccacagaaactaggaaatc*ggttcagcaccagcgaacc*agtcgaaggtcaaacctcttcccacaaaactgc*gaagcaggatcaccgcttc*gccattccaacataccaaatcaaaaacaattagt*cgtacagcatcagcgtacg*accacgcatcagtgactactatcaaaaaccaaac*cgttcagcaacagcgaacg*gtacacacggaaaaatcaactggtttacaaatac*gaaagacgagcacgctttc*aactattacgaaaaacatccgagg*cgatcagcaacagcgatcg*cccggcggaaaacctcacaaaaacacgacaaacg*gaagcacgaacacggcttc*cgccgacaacccacaaacttacaacgacgcaaac*ggtgcaggatcaccgcacc*gtacatcaaacacctcagatctcatt

MS2v5 sequence (24 old loops in italics; Zld sites underlined)

ggatcctacggtacttattgccaagaaa*gcacgagcatcagccgtgc*ctcc<u>aggt</u>cgaatcttcaaa*cgacgacgatcacgcgtcg*ctccagtattccagggttcatcagatcctacggtacttattgccaagaaa*gcacgagcatcagccgtgc*ctcc<u>aggt</u>cgaatcttcaaa*cgacgacgatcacgcgtcg*ctccagtattccagggttcatcagatcctacggtacttattgccaagaaa*gcacgagcatcagccgtgc*ctcc<u>aggt</u>cgaatcttcaaa*cgacgacgatcacgcgtcg*ctccagtattccagggttcatcagatcctacggtacttattgccaagaaa*gcacgag*textitcatcagccgtgcctcc<u>aggt</u>cgaatcttcaaa*cgacgacgatcacgcgtcg*ctccagtattccagggttcatcagatcctacggtacttattgccaagaaa*gcacgagcatcagccgtgc*ctcc<u>aggt</u>cgaatcttcaaa*cgacgacgatcacgcgtcg*ctccagtattccagggttcatcagatcctacggtacttattgccaagaaa*gcacgagcatcagccgtgc*ctcc<u>aggt</u>cgaatcttcaaa*cgacgacgatcac gcgtcg*ctccagtattccagggttcatcagatcctacggtacttattgccaagaaa*gcacgagcatcagccgtgc*ctcc<u>aggt</u>cgaatcttcaaa*cgacgacgatcacgcgtcg*ctccagtattccagggttcatcagatcctacggtacttattgccaagaaa*gcacgagcatcagccgtgc*ctcc<u>aggt</u>cgaatcttcaaa*cgacgacgatcacgcgtcg*ctccagtattccagggttcatcagatcctacggtacttattgcca agaaa*gcacgagcatcagccgtgc*ctcc<u>aggt</u>cgaatcttcaaa*cgacgacgatcacgcgtcg*ctccagtattccagggttcatcagatcctacggtacttattgccaagaaa*gcacgagcatcagccgtgc*ctcc<u>aggt</u>cgaatcttcaaa*cgacgacgatcacgcgtcg*ctccagtattccagggttcatcagatcctacggtacttattgccaagaaa*gcacgagcatcagccgtgc*ctcc<u>aggt</u>cgaatcttcaaa*cgacgacgatcacgcgtcg*ctccagtattccagggttcatcggatct

Live imaging

Virgin females maternally expressing MCP-GFP and H2Av-RFP were crossed with males of
the MS2 reporter lines. 0-1 hour embryos were collected, dechorionated, and transferred onto
a breathable membrane (Lumox Film, Sarstedt AG & Co.; Nümbrecht, Germany) in the
middle of a plastic microscope slide (3D printed by Sculpteo; Créteil, France). Live imaging
was performed using a Leica SP8 63X objective lens with the following settings: optical
sections: 512x512 pixels, 30 z stacks 0.69 μm apart, 12bit; zoom: 1.7; time resolution: 40 s per
frame. Laser power was measured using the X-Cite power meter, model No.XR2100) and set
at 70% (main), 30% (488 nm), and 10% (554 nm). Embryos were imaged for approximately
two hours, typically from nc 10 to early nc 14, as sog refines rapidly during mid-late nc 14
due to dynamic regulation by other factors (François et al., 1994).

High resolution imaging

Antibody and smFISH stained embryos with either 3TAG or 0TAG MS2 reporters were
imaged using a LSM Zeiss 880 confocal microscope with 100X objective using the following
settings: 1132x1132 pixels with 0.14 μm z stack step size, 16bit, 1.8 zoom. Laser power
was set at 1% (405 nm), 5% (488 nm), 15% (633 nm). All images were captured using the
the Airyscan detector array. Post-processing was carried out using the ZEN2012 software
"Airyscan Processing" feature.

Detailed genotypes

Figure 2.1

A: wt = ♀ y[1]w[∗] × ♂ y[1]w[∗]

B: *zld-* prepared by *zld* RNAi in two crosses: 1st cross ♀ +;+;UAS-shRNA-*zld* × ♂
P{*COG*-GAL4:VP16};P{Gal4-*nos*.NGT}40;P{*nos*-Gal4-VP16}; 2nd cross G1♀ P{*COG*-
GAL4:VP16}/+;P{Gal4-*nos*.NGT}40/+;P{*nos*-Gal4-VP16}/UAS-shRNA-*zld* × ♂ yw

D: 3TAG = *y[1] w[∗];sog 3TAG*-MS2-*lacZ*;+

E: 0TAG = *y[1] w[∗];sog 0TAG*-MS2-*lacZ*;+

Figure 2.2

A, C, and E: 3TAG ♀ *y[1] w[∗]*;Pw[+mC] = His2Av-mRFP1}II.2;P{w[+mC] = *nos*-
MCP.EGFP2 × ♂ *y[1] w[∗];sog 3TAG*-MS2-*lacZ*;+

B, D, and F: 0TAG ♂ *y[1] w[*]*;P{w[+mC] = His2Av-mRFP1}II.2;P{w[+mC] = *nos-*MCP.EGFP}2 × ♂ *y[1] w[*]*;*sog 0TAG*-MS2-*lacZ*;+

Figure 2.3

B, D, and F: 3TAG ♀ *y[1] w[*]*;P{w[+mC] = His2Av-mRFP1}II.2;P{w[+mC] = *nos-*MCP.EGFP}2 × ♂ *y[1] w[*]*;*sog 3TAG*-MS2-*lacZ*;+

C, E, and F: 0TAG ♀ *y[1] w[*]*;P{w[+mC] = His2Av-mRFP1}II.2;P{w[+mC] = *nos-*MCP.EGFP}2 × ♂ *y[1] w[*]*;*sog 0TAG*-MS2-*lacZ*;+

Figure 2.4

A and C: 3TAG *y[1] w[*]*;*sog 3TAG*-MS2-*lacZ*;+

B and C: 0TAG *y[1] w[*]*;*sog 0TAG*-MS2-*lacZ*;+

Figure A.1 (Related to Fig. 2.4)

A, C, and D: 3TAG *y[1] w[*]*;*sog 3TAG*-MS2-*lacZ*;+

B, C, and D: 0TAG *y[1]w[*]*;*sog 0TAG*-MS2-*lacZ*;+

Videos (Related to Figs. 2.2 and 2.3)

Videos A.2.1, A.2.3, and A.2.5: 3TAG *y[1]w[*]*; *sog 3TAG*-MS2-*lacZ*;+

Videos A.2.2, A.2.4, and A.2.6: 0TAG *y[1]w[*]*;*sog 0TAG*-MS2-*lacZ*;+

Quantification and statistical analysis

Live imaging Videos (Videos A.2.1, A.2.2, A.2.3, A.2.4, A.2.5, and A.2.6) were analyzed using the Imaris (Bitplane, Oxford Instruments, Concord MA) "spots" function over and track using retrograde motion with a max frame gap of 3. MS2 foci were assumed to be 1μm across with a z axis point spread function estimation of 2μm. After tracking, both intensity sum and position csv files were exported and analyzed using a series of custom R scripts. Tracks are assigned a nuclear cycle and zone position by referencing a manually generated annotation file containing all frames where anaphase was reached for each Video and a y axis position of ventral repression at nuclear cycle 14. Transcriptional delay values for each tracked object were generated by subtracting the current frame number by the preceding anaphase frame number. Transcriptional dynamics at different dorsal-ventral positions was analyzed by subdividing each image into five zones along the DV axis. Zone 1 comprises the

mesoderm, as determined by the Snail repression border that becomes obvious in early NC14. The remaining zones are defined by 20μm spatial bins that proceed dorsally, approximately 4 rows of nuclei per zone (schematized in Figure 2.2C).

To measure transcriptional kinetic parameters, we used individual foci and performed a linear fit on the first 25% of the intensity values over time. Time to steady-state values were calculated by intersecting the linear fit with a horizontal line generated by the averaging the top 20% of intensity values for foci signals. Statistical tests were performed using Welch's t test that assumes independent underlying variance. P values shown in Figure 2.3H are visually represented as one asterisk indicating a $p < 0.05$, two indicating $p < 0.01$, and three indicating $p < 0.001$.

The smFISH nascent transcript values shown in Figure 2.4 were obtained by extracting the total fluorescence of large nuclear localized foci assumed to be the point of active transcription. This value was then divided by intensity values of single transcripts by assuming an average 0.3μm diffraction limited point again using the Imaris "spots." These values formed a normal distribution from which the median value was selected as the fluorescence intensity value of a single transcript within a single frame. Dl intensity values for each nucleus were found by extracting the mean fluorescence of antibody stain signal within volumes defined by nuclear DAPI signal. This normalizes differences in Dl concentrations along the gradient between genotypes. Radial scans were performed using a custom R script that utilized the position values extracted from Imaris to interrogate .tif files of the Dl antibody stain. Error bars on enrichment plots in Figure 2.4C are standard error of the mean of individual enrichment curves in each positional bin. All plotting was performed with base R functions and the ggplot2 library.

2.5 Acknowledgments

2.6 Author contributions

C.A.R., H.G.G., and P.H.W. designed the overall study. E.C.E. designed and prepared the MS2v5(-TAG) loops vector. S.Y. made the *sog* enhancer-MS2v5(-TAG) reporter constructs and carried out the live imaging experiments. C.A.R., P.H.W., and S.-K.H. carried out the assays in fixed embryos. P.H.W. conceived the ideas and wrote code for the computational image analysis. P.H.W. wrote the draft manuscript, and all authors contributed to revisions.

2.7 Declaration of interests

The authors declare no competing interests.

Chapter 3

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

Foreword

When I began graduate school, advances in imaging technologies, including microscopes as well as *in vivo* approaches to visualizing molecular processes, made it possible to investigate the molecular dynamics of cellular and developmental biology at high resolution. In addition to these experimental advances, theoretical modeling is able to act as its own type of microscope into these processes (Cohen, 2004). I wanted to combine the two approaches—experimental imaging and theoretical modeling—to investigate a compelling biological question.

Hernan Garcia's work investigating the fundamental process of gene expression, perfectly fit my interests. Much of the work in the Garcia Lab uses microscopy to image live, developing *Drosophila* embryos to generate quantitative, high temporal resolution data sets of transcription. This data enables us to test and challenge long-held models of transcriptional regulation as well as develop and test new models.

My first year rotation project in the Garcia Lab involved studying the role that a protein called Zelda played in gene expression in the early fruit fly embryo. The data that I acquired during my rotation showed that the loss of Zelda caused delays in the onset of transcription that could not be explained by a simple equilibrium model that at that time was the standard, accepted model describing the system. Studies over the previous few years indicated that Zelda played the role of a "pioneer factor," somehow making chromatin accessible for other transcription factors. We possessed a large, high-quality data set that we could utilize to investigate and test theoretical models of Zelda's activity.

A physics graduate student, Jonathan Liu, rotated with the lab and worked on this project near the end of my second year of graduate school. He brought a wealth of skills and insight to the project, particularly to the data analysis and modeling components. He quickly joined

the lab and we became partners on this project that, at the time, seemed near completion.

Hernan encouraged us to find a model which described the data so that we could publish quickly. This seemed like a relatively easy task, at first. We identified the Monod-Wyman-Changeux (MWC) model of chromatin accessibility as the model most suited to describe our data. This model was first proposed by Jonathan Widom (Miller and Widom, 2003) and explored further by Leonid Mirny (Mirny, 2010). We simply needed to use our data to validate this extension of the equilibrium statistical mechanical model of transcriptional regulation. Jonathan and I both examined this model and tried everything we could to make it "fit" our data. After a week, we went back to Hernan to tell him that the model just didn't seem to recapitulate what we saw experimentally. To paraphrase, Hernan's response was, "Interesting...Now prove that this model fails."

So we set out to prove that this equilibrium, and ultimately, all equilibrium models failed to describe our data. Of course proving the failure of a model is no easy task. Over the next couple of years, we systematically explored all potential predicted outputs of several classes of model of transcriptional regulation. When all possible equilibrium models failed, we revisited the assumption made in order to use those models and then broke those assumptions, one by one, and similarly exploring outputs of different classes of non-equilibrium models. We eventually ruled out all classes of models that we explored except one of them.

The only class of models that has the potential to describe our experimental data is a class of non-equilibrium models in which transcription factors actively drive chromatin accessibility. Exactly which of these models describes the actual regulation occurring in our system remains to be determined by future experiments and modeling.

Over the course of graduate school, this project took more time than any of my other projects. It culminated in a publication in *elife* that I coauthored with Jonathan. This project was made possible through the use of transgenic flies created by Shelby Blythe at Northwestern University, who also provided guidance. Two talented undergraduate students, Maryam Kazemzadeh-Atoufi and Sydney Ghoreishi, contributed to data acquisition and analysis. Through Hernan's guidance, support, and pushing us to find a solution to our modeling quandary, I gained numerous skills as well as an appreciation for a puzzle-solving approach to modeling.

Hernan received funding to continue pursuing follow up projects to dig further into the transcription factor driven accessibility model as well as Zelda's regulatory role and potential cofactors. I have set up the materials and much of the plans for these follow up projects and the next generation of graduate and undergraduate students have taken over and will undoubtedly be successful with them.

3.1 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. 3.1A; 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

$$\frac{d}{dt}(\text{mRNA}) = R\,p_{bound}. \tag{3.1}$$

Here, R is the rate of mRNA production when the system is in an RNAP-bound state (see

(A) THERMODYNAMIC MWC MODEL

	STATE	WEIGHT	RATE
inaccessible		$e^{\beta \Delta \varepsilon_{chrom}}$	0
		1	0
accessible	activator	a	0
	RNAP	p	R
	ω	$a\,p\,\omega$	R

KEY ASSUMPTIONS

(i) occupancy hypothesis

$$\frac{d mRNA}{dt} = R\, p_{bound}$$

(ii) separation of time scales

(iii) no energy expenditure

(B) NON-EQUILIBRIUM MWC MODEL

R

mRNA

inaccessible   accessible

(C) TRANSCRIPTION FACTOR-DRIVEN CHROMATIN ACCESSIBILITY

OFF$_1$ → OFF$_2$ → … → OFF$_m$ → ON

inaccessible   accessible

STATE	WEIGHT	RATE
	1	0
	a	0
	p	R
	$a\,p\,\omega$	R
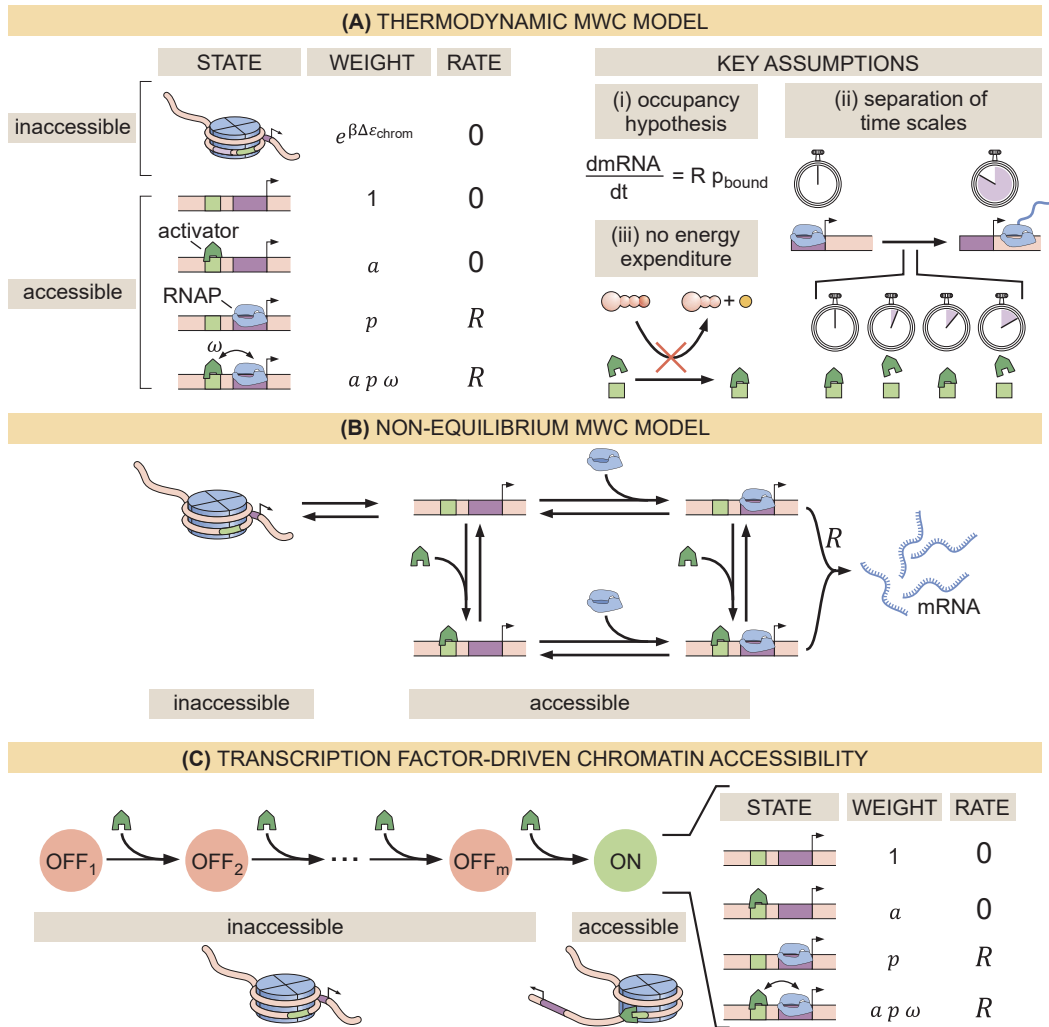
Figure 3.1: Three models of chromatin accessibility and transcriptional regulation. See caption on next page.

Section B.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.

Figure 3.1: 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.

(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. 3.1B). 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 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. 3.1C)

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.

3.2 Results

3.2.1 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. 3.1A) 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 ($\Delta\varepsilon_{\mathrm{chrom}}$), 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 3.2.3 and 3.2.4). 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. 3.2A, 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 3.5.1; Driever and Nusslein-Volhard, 1988; Driever et al., 1989; Park et al., 2019). This general model is described in detail in Section B.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

$$p_{bound} = \frac{\left(1+z\right)^{n_z} p\left(1 + \sum_{i=1}^{n_b} \binom{n_b}{i} b^i \omega_b^{i-1} \omega_{bp}^i\right)}{\underbrace{e^{\Delta\varepsilon_{chrom}/k_B T}}_{\substack{\text{inaccessible} \\ \text{state}}} + \underbrace{\left(1+z\right)^{n_z}}_{\text{Zelda binding}} \underbrace{\left(1 + p + \sum_{j=0,1} \sum_{i=1}^{n_b} \binom{n_b}{i} b^i \omega_b^{i-1} p^j \omega_{bp}^{ij}\right)}_{\text{Bicoid and RNAP binding}}},$$ (3.2)

where $b = [Bicoid]/K_b$, $z = [Zelda]/K_z$, and $p = [RNAP]/K_p$, with $[Bicoid]$, $[Zelda]$, and $[RNAP]$ being the concentrations of Bicoid, Zelda, and RNAP, respectively, and K_b, K_z, and K_p their dissociation constants (see Sections B.1.1 and B.1.2 for a detailed derivation). Given a set of model parameters, plugging p_{bound} into Equation 3.1 predicts the rate of RNAP loading as a function of Bicoid and Zelda concentrations as shown in Fig. 3.2B. Note that in this work, we treat the rate of transcriptional initiation and the rate of RNAP loading interchangeably.

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

In order to experimentally test the theoretical model in Fig. 3.2, 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. 3.3A and Fig. B.3A; Video B.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 B.2.1; Fig. B.3B). We measured mean Bicoid nuclear concentration dynamics along the anterior-posterior axis of the embryo, as exemplified for two positions in Fig. 3.3A. As previously reported (Gregor et al., 2007b), after anaphase and nuclear envelope formation, the Bicoid nuclear concentration quickly increases
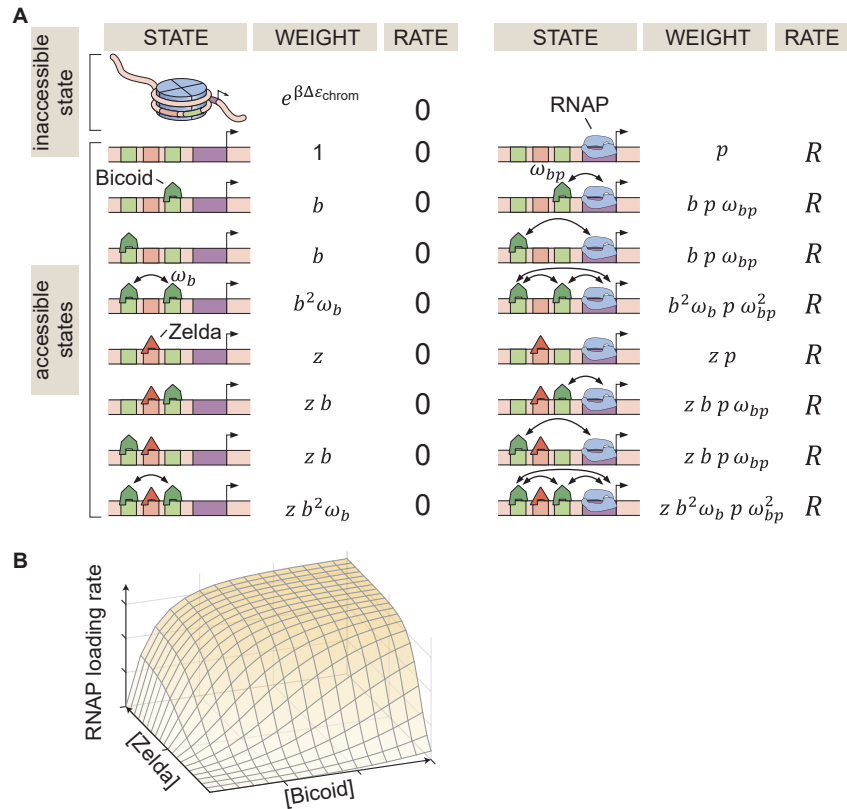
Figure 3.2: Thermodynamic MWC model of transcriptional regulation by Bicoid and Zelda. (A) States and statistical weights for a simplified version of the *hunchback* P2 enhancer. In this model, we assume that chromatin occluded by nucleosomes is not accessible to transcription factors or RNAP. Parameters 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.

as a result of nuclear import. These measurements were used as inputs into the theoretical model in Fig. 3.2.

Zelda concentration dynamics were measured in a Zelda-sfGFP line (Sections 4.5.2, 4.5.3, and 4.5.4; Fig. 3.3B; Video B.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. B.3). Contrasting Fig. 3.3A 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 B.2.1; Fig. 3.3B).

Given the high reproducibility of the concentration dynamics of Bicoid and Zelda (Fig. B.3), we combined measurements from multiple embryos by synchronizing their anaphase in order to create an "averaged embryo" (Section B.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., 2020a).

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 the embryo, the combined Bicoid and Zelda concentration dynamics define a trajectory over time along the predicted input-output function surface (Fig. 3.3C). 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 3.2.3). This quantity can be obtained by accounting for the RNAP elongation rate and the cleavage of nascent RNA upon termination (Section B.2.2; Fig. B.4; Bothma et al., 2014; Lammers et al., 2020a) yielding the predictions shown in Fig. 3.3D.

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. 3.3D).

Examples of the predictions generated by our theoretical model are shown in Fig. 3.3E 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.
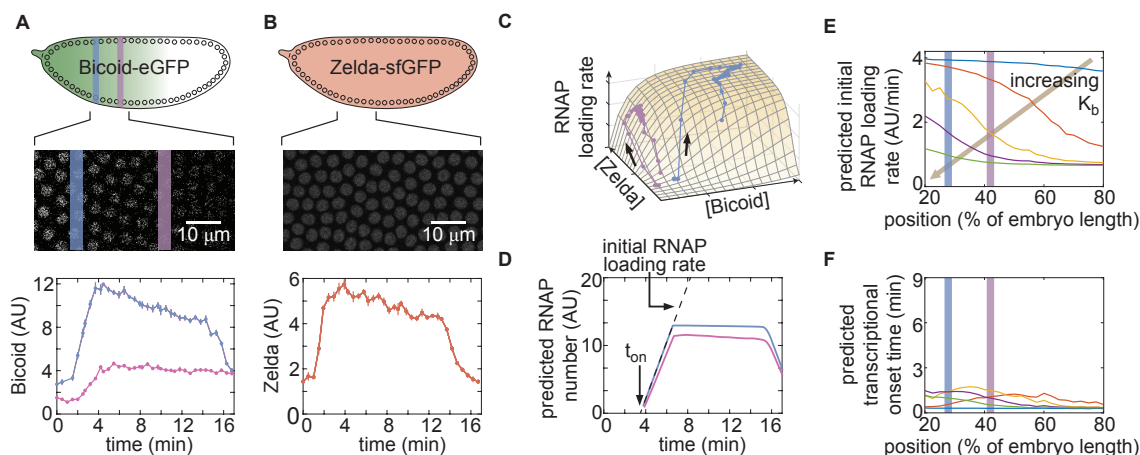
Figure 3.3: Prediction and measurement of dynamical input-output functions. (A) Measurement of Bicoid concentration dynamics in nuclear cycle 13. Color denotes different positions along the embryo and time is defined with respect to anaphase. (B) Zelda concentration dynamics. These dynamics are uniform 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.)

3.2.3 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. 3.3E 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. 3.4A; Video B.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. 3.4B) is in qualitative agreement with the theoretical prediction (Fig. 3.3D).

To compare theory and experiment, we next obtained the initial RNAP loading rates (Fig. 3.4C, blue points) and t_{on} (Fig. 3.4D, blue points) from the experimental data (Section B.2.3; Fig. B.5B). The step-like shape of the RNAP loading rate (Fig. 3.4C, 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 B.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. 3.3E 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. 3.4C and D, blue lines).

Although the model accounted for the initial rate of RNAP loading (Fig. 3.4C, blue line), it produced transcriptional onset times that were much lower than those that we experimentally observed (Fig. B.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

$$\frac{\mathrm{dmRNA}}{\mathrm{d}t} = \begin{cases} 0 & \text{if } t < t_{MitRep} \\ R\,p_{bound} & \text{if } t \geq t_{MitRep} \end{cases}, \tag{3.3}$$

where R and p_{bound} are as defined in Eqns. 3.1 and 3.2, respectively, and t_{MitRep} is the mitotic repression time window over which no transcription can take place after anaphase (Sections B.1.2 and B.3). After incorporating mitotic repression, the thermodynamic MWC model successfully fit both the rates of RNAP loading and t_{on} (Fig. 3.4C and D, blue lines, Fig. B.6A and B blue lines).
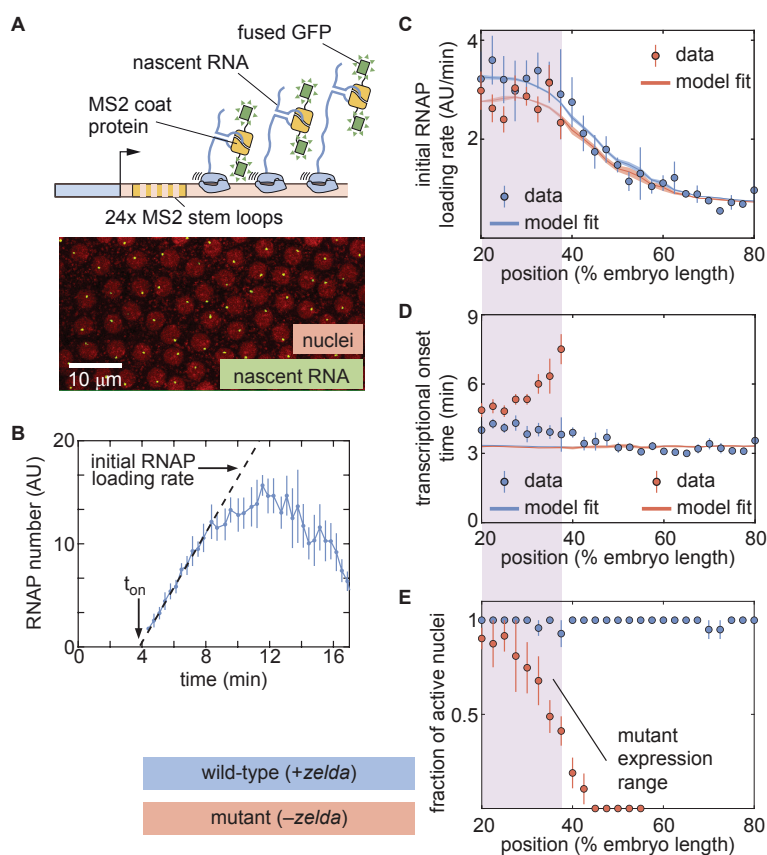
Figure 3.4: The thermodynamic MWC model can explain *hunchback* transcriptional dynamics in wild-type, but not *zelda⁻*, embryos. (A) The MS2 system measures the number of RNAP molecules actively transcribing the *hunchback* reporter gene in live embryos. (B) Representative MS2 trace featuring the quantification 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))

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 B.1.2 and Fig. 3.2. In order to test this seemingly simpler model, we repeated our measurements in embryos devoid of Zelda protein (Video B.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, 3.5.3; 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 B.4; Xu et al., 2014).

While close to 100% of nuclei in wild-type embryos exhibited transcription along the length of the embryo (Fig. 3.4E, blue; Video B.9.5), measurements in the $zelda^-$ background revealed that some nuclei never displayed any transcription during the entire nuclear cycle (Video B.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. 3.4E, 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.

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. 3.4C, red points). However, unlike in the wild-type case (Fig. 3.4D, 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. 3.4D, 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. 3.4C, red line), we noticed a discrepancy between the observed and fitted transcriptional onset times of up to ∼5 min (Fig. 3.4D, red). While the mutant data exhibited a substantial delay in more posterior nuclei, the model did not produce significant delays (Fig. 3.4D, 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. 3.4E, 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. 3.4C 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. A.1A), namely

$$\text{offset} = t_{on}(x = 20\%) \tag{3.4}$$

where x is the position along the embryo. Second, we measured the average transcriptional onset delay along the anterior-posterior axis (Fig. A.1A). 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

$$\langle \text{onset delay} \rangle = \frac{1}{37.5\% - 20\%} \int_{20\%}^{37.5\%} (t_{on}(x) - t_{on}(x = 20\%)) \, dx, \tag{3.5}$$

where the offset in the onset time was used to define the zero of this integral (Section B.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. A.1C, ellipses).

Based on Estrada et al. (2016) and as detailed in Section B.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. A.1B 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. A.1C, 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.
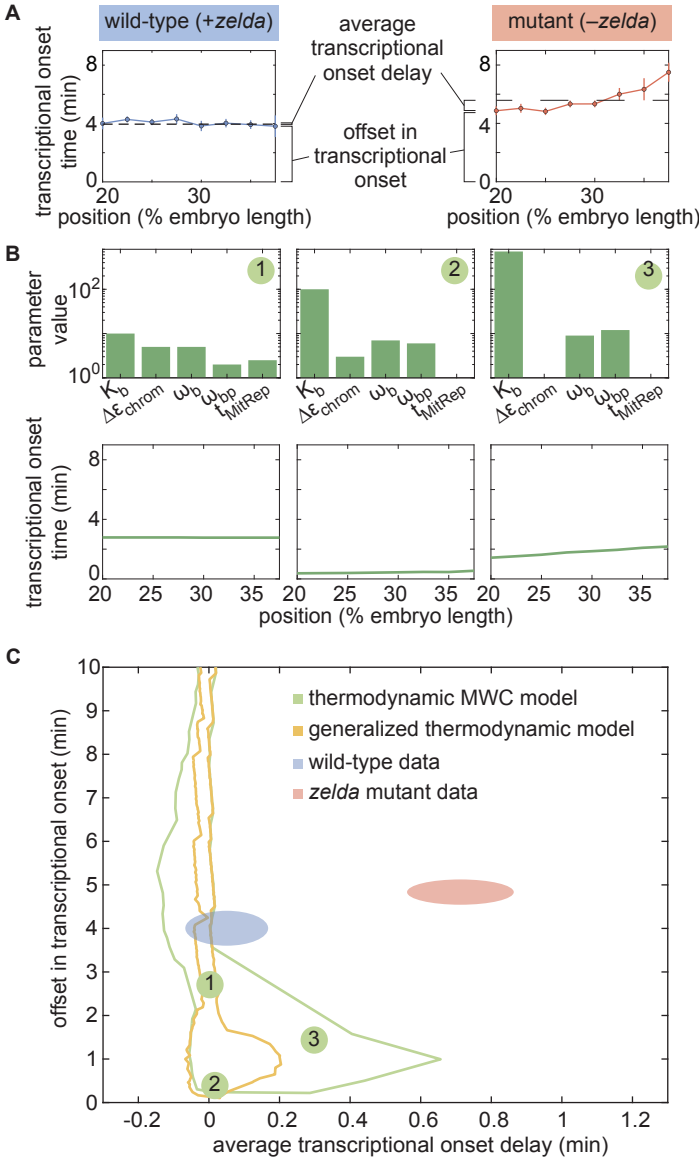
Figure 3.5: Failure of thermodynamic models to describe Bicoid-dependent activation of *hunchback*. See caption on next page.

Figure 3.5: 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)

3.2.4 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

$$\frac{d\text{mRNA}}{dt} = \frac{\left(\sum_{i=0}^{n_b} P_{1,i} R [Bicoid]^i \right)}{p_{inacc} + \sum_{r=0}^{1} \sum_{i=0}^{n_b} P_{r,i} [Bicoid]^i}, \tag{3.6}$$

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 B.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 B.6.3).

Although this generalized thermodynamic model contains many more parameters than the thermodynamic MWC model previously considered, we could still systematically explore reasonable values of these parameters and the resulting t_{on} offsets and average t_{on} delays (Section B.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 B.6.2; Fig. A.1C, 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 B.1.3 and B.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. B.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. 3.3A 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 B.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.

3.2.5 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 B.6.5; Fig. 3.1B). 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 B.6.5 and B.7.1; Kim and O'Shea, 2008; Narula and Igoshin, 2010). This model, shown for the case of two Bicoid binding sites in Fig. 3.6A, 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. 3.6A. 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 B.7.2; Fig. 3.6B and Fig. B.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. 3.6B). 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 B.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.

3.2.6   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 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.
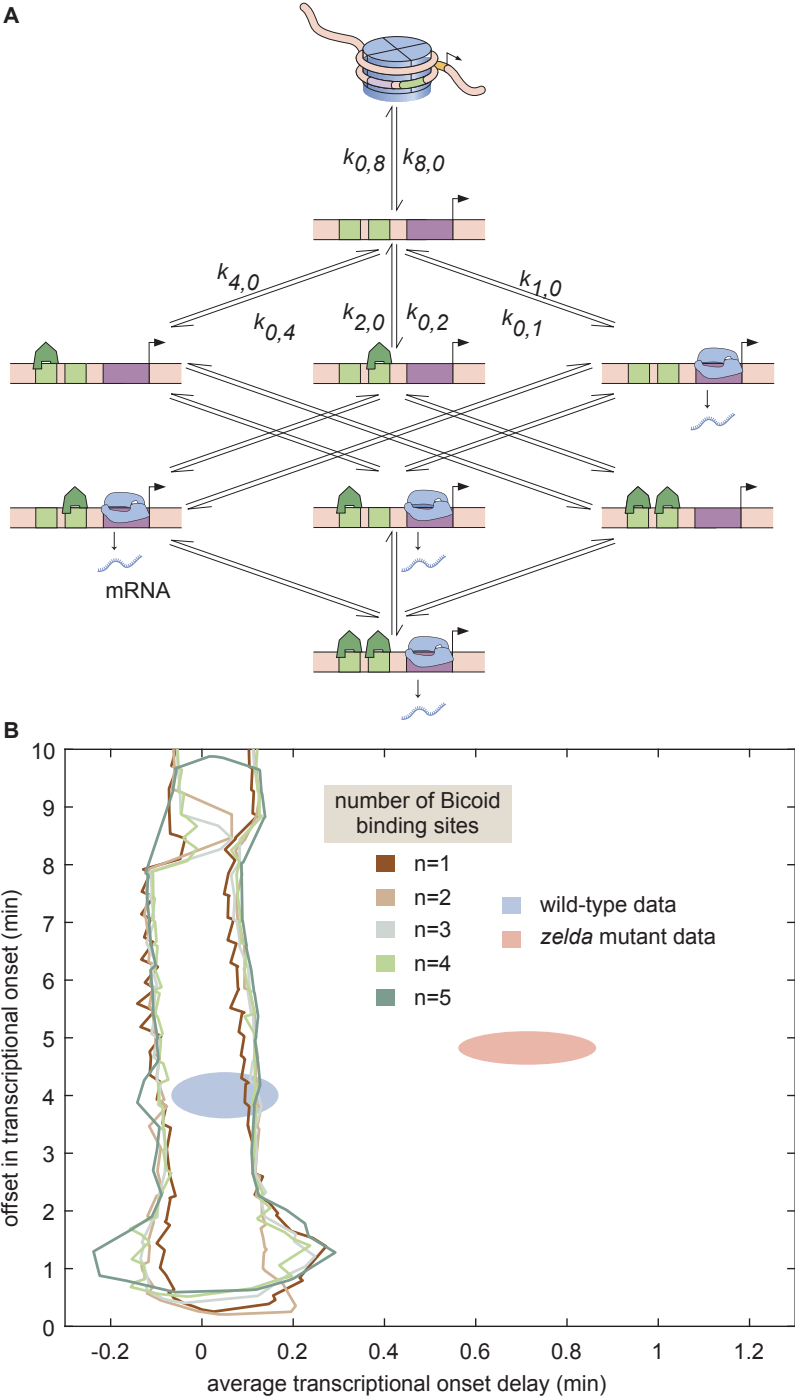
Figure 3.6: Non-equilibrium MWC model of transcriptional regulation cannot predict the observed t_{on} delay. See caption on next page.

Figure 3.6: 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 B.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)

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 B.8.1; Fig. 3.7A). 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

$$\pi = c_b[Bicoid] + c_z[Zelda]. \tag{3.7}$$

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 B.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. 3.7B and C). Additionally, we found that a minimum of $m = 3$ steps was required to sufficiently explain the data (Section B.8.3; Fig. B.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 B.3 and B.8.1). Instead, the timing of transcriptional output arose directly from the model's initial irreversible transitions (Fig. B.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. 3.7B 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. 3.7D).

3.3 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
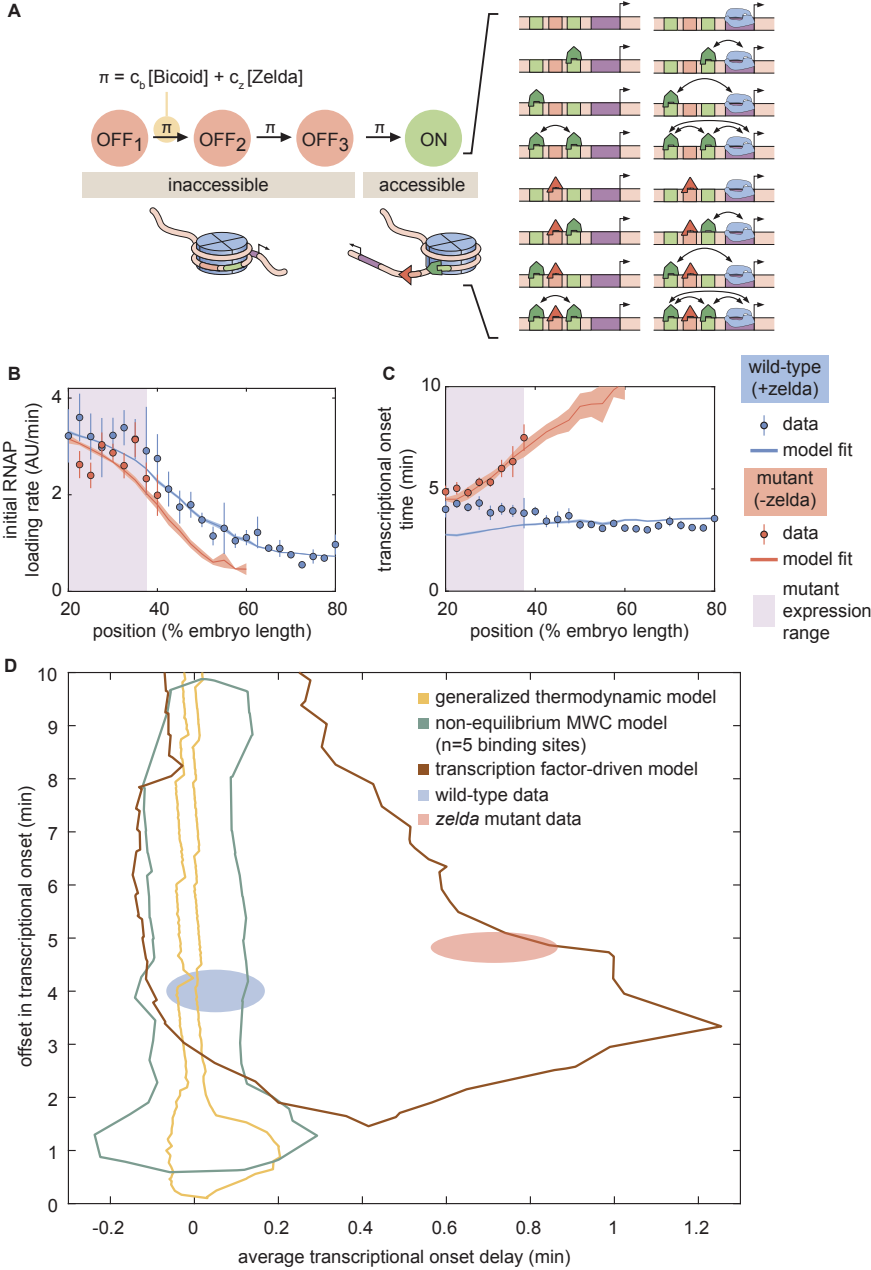
Figure 3.7: A model of transcription factor-driven chromatin accessibility is sufficient to recapitulate *hunchback* transcriptional dynamics. See caption on next page.

Figure 3.7: 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)

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. 3.2B). 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. 3.3A 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. 3.3C). 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. 3.3D and 3.4B).

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. 3.4D and E). Our systematic exploration of all thermodynamic models (over a reasonable parameter range) showed that that no thermodynamic model featuring regulation by Bicoid alone could quantitatively recapitulate the measurements performed in the *zelda* mutant background (Fig. A.1C, 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. 3.6B). 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. 3.6B) 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. 3.7D and Fig. B.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; Fritzsch 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 B.3 and B.8.1), which was required in the thermodynamic MWC model (Fig. B.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; **?**) 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.

3.4 Acknowledgments

3.5 Methods and Materials

3.5.1 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.

3.5.2 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.

3.5.3 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).

3.5.4   Sample preparation and data collection

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

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. (2020a), 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 μW. The confocal stack was only 10 slices in this case, rather than 21, resulting in a spacing of 1.11 μm 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 μW. 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.

3.5.5   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.

3.5.6   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 B.2.1 for more details on how this was carried out for model predictions.

Chapter 4

Dynamic single-cell characterization of the eukaryotic transcription cycle

Foreword

In the previous two chapters, we focused on the timing and rate of transcriptional initiation. Indeed, initiation is a very important component of the transcription cycle from a regulatory perspective. However, this approach ignores other potentially important regulatory steps. By overlooking the rest of the transcription cycle—elongation and termination, for instance—we might be missing out on a more complete understanding of the dynamics of transcriptional regulation.

Much of the work in the Garcia Lab uses MS2 as a readout of transcriptional initiation, but we are continually devising new methods to use it alone or in conjunction with other technology to explore other processes. The intensity of the fluorescent spot produced by MCP binding to the MS2 stem-loops is a measure of the number of RNA polymerase molecules actively transcribing the tagged gene. This means that the signal contains much more information than simply the transcription initiation rate and onset time. If we could devise a way to separate the signals from initiation, elongation, and mRNA cleavage, we could harness more of the potential of the MS2 technology.

The project described in this chapter arose from Jonathan and I carefully examining the full MS2 signal from the data reported in Chapter 3. We noticed that in the anterior positions of the fruit fly embryo, the transcription spot intensity continued to rise toward the end of the nuclear cycle despite falling activator concentrations. In the posterior positions, however, the spot intensities fell toward the end of the nuclear cycle. We were unable to come up with a single set of parameters that could explain the signal patterns for every position in the embryo. Our first idea, was that the high transcription rates in the anterior were causing a traffic jam at the end of the gene where the termination process was taking place, causing the fluorescent signal to continue to rise toward the end of the nuclear cycle. However, we did not have the experimental tools to test this hypothesis.

In order to investigate the nature of the puzzling fluorescent signal, we conceived an experiment using a dual-color reporter. This involved placing the MS2 and PP7 stem loops on either end of our reporter gene, and tagging them with two different fluorophores. The result is a two color signal that would let us examine the time delay between polymerase molecules reaching the 5' and 3' end of the gene. This time delay can be used to estimate an average elongation velocity that we could use in our transcription models.

This idea motivated a group effort in the lab to create the necessary flies for this dual-color reporter. We needed flies expressing coat proteins for both systems, MCP and PCP, as well as a histone tag to visualize the nuclei. Jonathan, myself, Yang Joon Kim, Meghan Turner, and Simon Alamos all worked together to create flies which expressed MCP-mCherry, PCP-GFP, and histone-iRFP. A master's degree student, Donald Hansen, later joined the project, contributing to data acquisition, analysis, and modeling.

Jonathan took the lead on this project and ran with it. He realized that the dual-color reporters would not only allow us to examine elongation velocity, but also mRNA cleavage, part of the termination process. So with these embryos, he set out to investigate the dynamics of the entire transcription cycle. This was exciting because while others had used similar experimental systems to examine initiation or elongation in isolation, no one had simultaneously investigated the entire transcription cycle with live imaging. mRNA cleavage, in particular, was understudied.

Jonathan also came up with the rigorous computational framework to quantify the steps of the transcription cycle from the noisy, single cell experimental data. We ended up with a Bayesian framework using Markov Chain Monte Carlo which did well in inferring parameters of a simple model of the transcription cycle from the dual-color data.

This project was published in *PLoS Computational Biology*. It underscores the need not only for innovative imaging technology, but also the importance of data analysis and interpretation. We are witnessing an explosion in the number and size of biological data sets and we need to address the challenge of deciphering which data is most informative and then utilize or develop the most appropriate theoretical, computational, or statistical analysis to interpret the data. The need for innovative data analysis methods is something I encountered and am still addressing in the project described in Chapter 5 which is still in progress.

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; Falo-Sanjuan et al., 2019; Lammers et al., 2020a) 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., 2020a). 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 C.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 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. C.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 C.12.1). The fluorescent signals did not exhibit noticeable photobleaching (Section C.2 and Fig. C.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
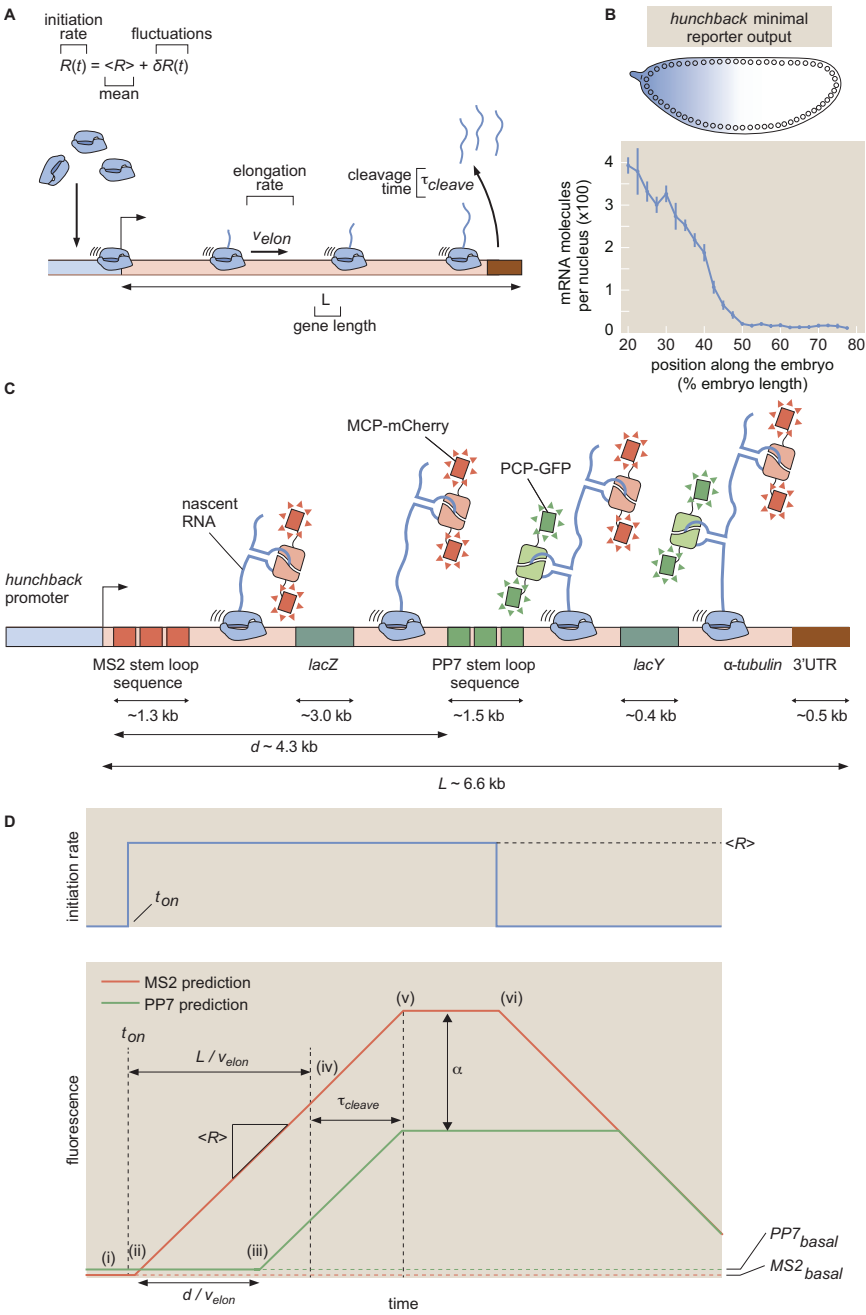
Figure 4.1: Theoretical model of the transcription cycle and experimental setup. See caption on next page.

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.)

experiment, we approximate the cleavage of a single transcript as resulting in the instantaneous loss of its associated fluorescent signal in both channels (Section C.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 C.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 C.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 C.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 C.4.3 and Fig. C.4). This curation procedure did not introduce noticeable bias in the results (Fig. C.4G-I). A small minority of the rejected cells (Fig. C.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 C.6 and Fig. C.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

$$\alpha = \frac{F_{MS2}}{F_{PP7}}, \tag{4.1}$$
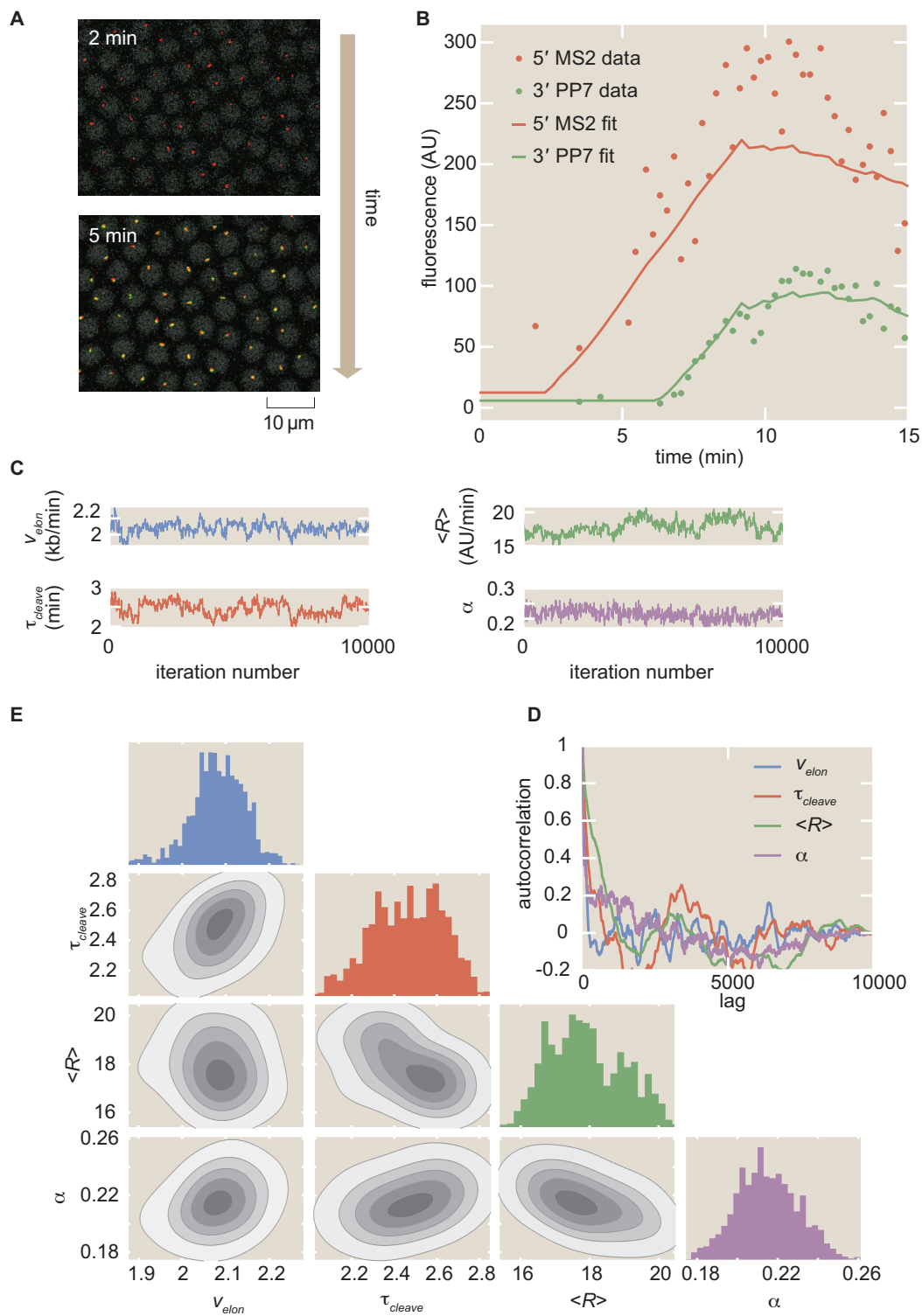
Figure 4.2: MCMC inference procedure. See caption on next page.

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).

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 $AU_{\mathrm{MS2}}/AU_{\mathrm{PP7}}$, 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 C.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 data point in Figure 4.3D, yielding the histogram in Figure 4.3A (yellow), which possessed 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
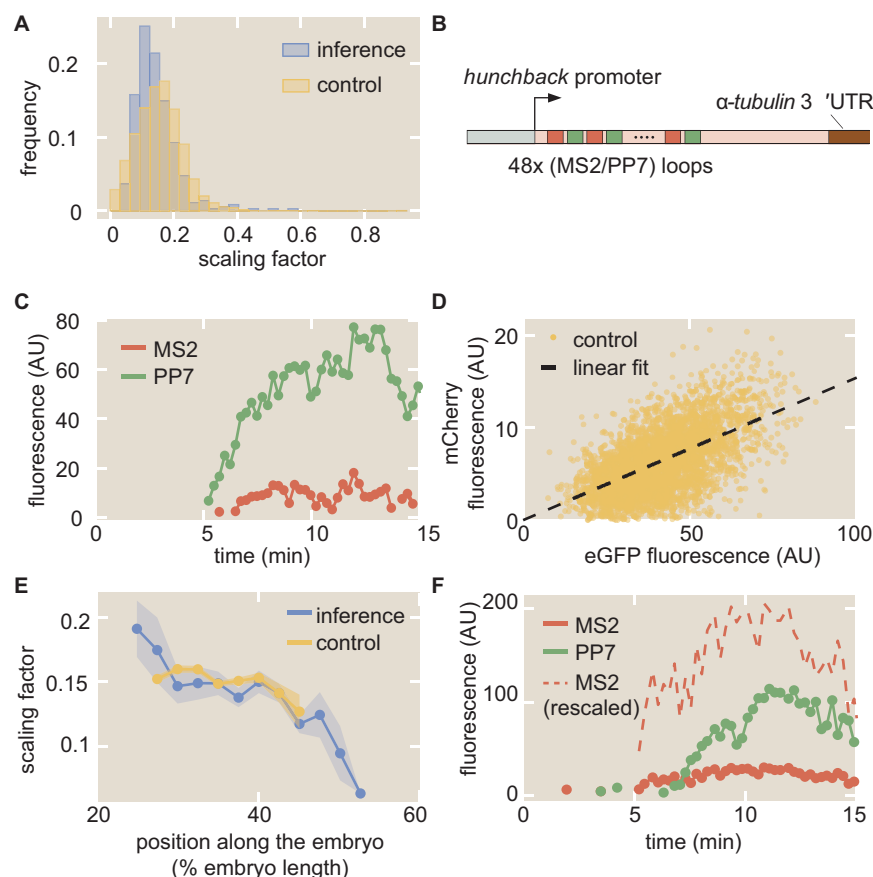
Figure 4.3: Calibration of MS2 and PP7 fluorescence signals. (A) Histogram of inferred values of α at the single-cell level from inference (blue), along with histogram of α values from the control experiment (yellow). (B) Schematic of construct used to measure the calibration factor α using 24 interlaced 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.)

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 C.4.4 and Fig. C.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 Fig. C.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

$$CV_{total}^2 = CV_{uncorrelated}^2 + CV_{correlated}^2. \tag{4.2}$$

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 C.8 and Figure C.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
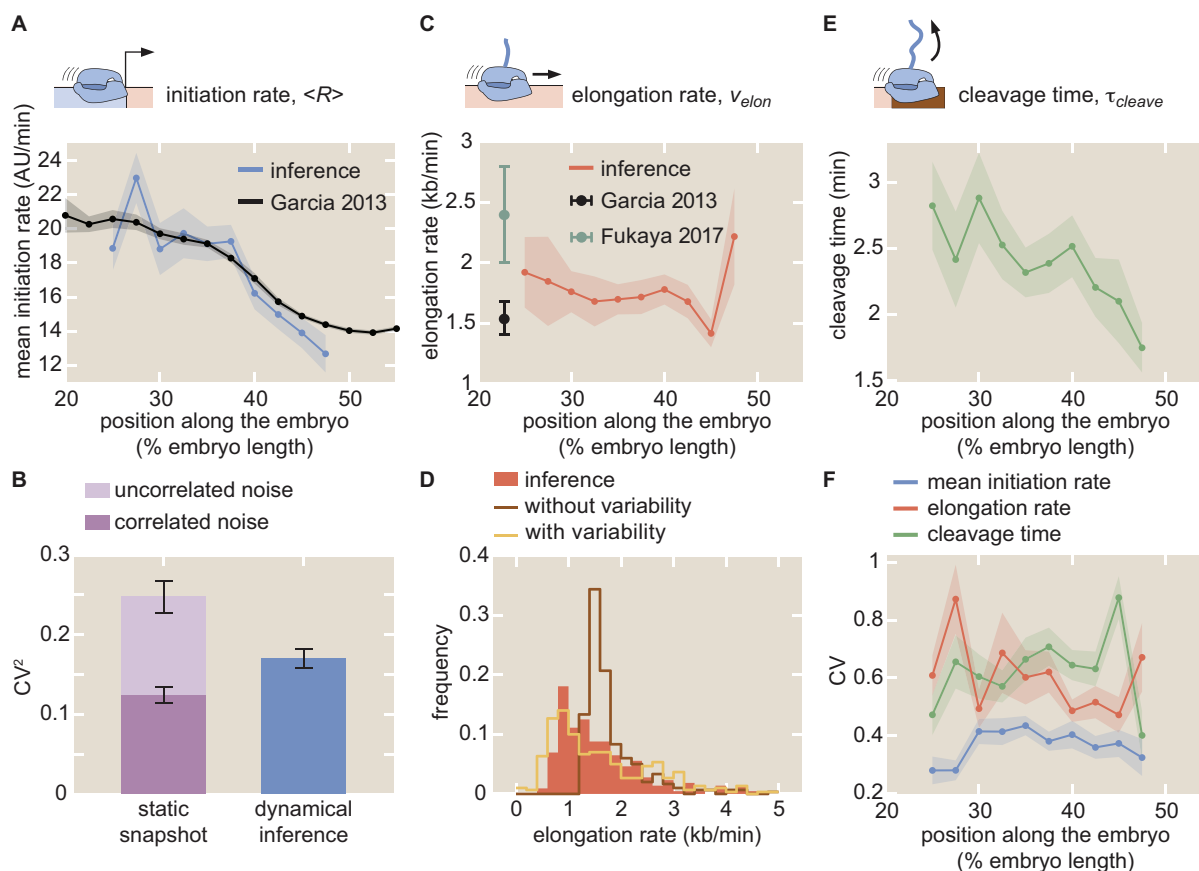
Figure 4.4: Inferred transcription-cycle parameters. See caption on next page.

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

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 C.8 and Fig. C.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 C.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).)

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 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; **?**),

they remain unable to resolve single-cell distributions 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-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 C.9; Fig. C.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 C.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 C.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. C.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 ($\sim$3 min) in the anterior end and lower values toward the posterior end ($\sim$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-val ≈ 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., 2020a) 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-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-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

$$\rho = \frac{\langle R \rangle}{v_{elon}}. \tag{4.3}$$

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-val ≈ 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 C.4.4 and Fig. C.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 C.11 and Fig. C.11).

Using an absolute calibration for a similar reporter gene (Garcia et al., 2013) led to a rough scaling of 1 AU ≈ 1 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, 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; **?**; Falo-Sanjuan et al., 2019; Li et al., 2019; Lee et al., 2019; Lammers et al., 2020a; 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., 2020a; Bowles et al., 2020).

In this work, we established a novel method for inferring quantitative parameters of
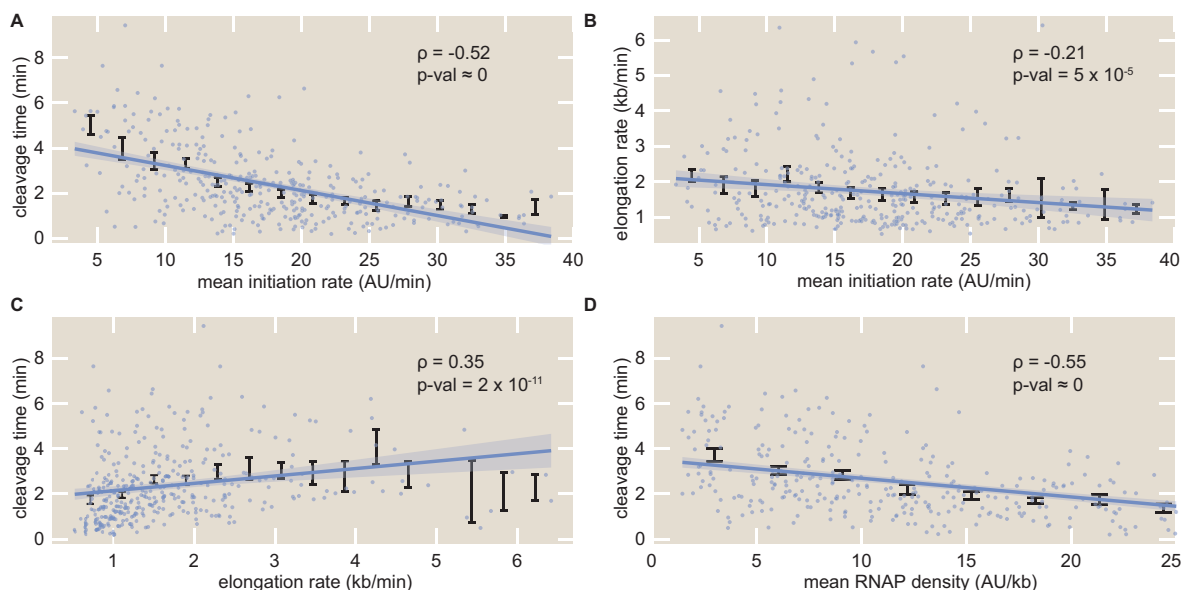
Figure 4.5: Single-cell correlations between transcription cycle parameters. Spearman rank correlation coefficients and associated p-values between (A) mean initiation rate and cleavage time, (B) mean initiation rate and elongation rate, (C) elongation rate and cleavage time, and (D) mean RNAP density 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).

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 C.4.3; Fig. C.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

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. (2020a). To summarize, embryos were collected, dechorionated 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 μW and 20 μW 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 μm and an inter-slice distance of 0.5 μm. The images were acquired at a time resolution of 15 s, using an image resolution of 512 x 128 pixels, a pixel size of 202 nm, and a pixel dwell time of 1.2 μs. 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. (2020a). 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).

Chapter 5

Quantifying the transcriptional dynamics of somitogenesis in live zebrafish embryos

Foreword

The projects described in the previous three chapters all dealt with genetic regulation in the developing fruit fly embryo. I wanted to expand the study of transcriptional dynamics at high spatiotemporal resolution to other organisms in which that data was not yet available. I was especially interested in extending our technologies to vertebrates with their relevance to biomedical research.

When I joined the Garcia Lab, Hernan had just received funding for a collaboration Andy Oates at the The École Polytechnique Fédérale de Lausanne. The main goal of this ongoing project is to use MS2 imaging and theoretical modeling developed in the Garcia Lab to study the process of somitogenesis in zebrafish, the expertise of the Oates Lab.

After a long process testing our MS2 and MCP constructs through injection of one cell stage embryos, generating transgenic fish, and identifying founders, we successfully established three lines of MS2 fish and five lines of MCP fish expressing one of three MCP-fluorescent protein designs. Preliminary data shows promise and we are currently solving the challenge of how to image these fish and how to process the large datasets we generate with them. Some datasets are over a terabyte! I am confident that we will end up learning a great deal about transcriptional dynamics with these new transgenic lines.

This project involved me traveling back and forth to The Max Planck Institute of Molecular Cell Biology and Genetics in Dresden, Germany. It was an amazing and unique opportunity to meet with scientists in Europe and to learn about the work occurring there. In Dresden, I met and was trained in zebrafish handling, transgenesis, and imaging by Daniele Soroldoni from the Oates Lab. Without his expertise I would have been lost in exploring this new model organism. I also visited Andy Oates' Lab in Lausanne and learned a great deal from

the rest of his lab.

As I prepare to graduate and move on from the Garcia Lab, I am incredibly thankful to have had this unique and rewarding experience in international collaboration. I am handing off this project to two talented and competent post-docs that joined the Garcia Lab last year, Brandon Schlomann and Bruno Moretti. I am excited to see the future results of their hard work on this project.

5.1   Introduction

The strikingly dynamic and complex process of vertebrate somitogenesis combines biochemical signaling (Hubaud and Pourquié, 2014) and tissue mechanics (Mongera et al., 2018) to achieve the robust development of morphological segments—somites—that prefigure bones and muscles in adults. This process is accomplished by a highly conserved biological clock consisting of an oscillatory network of genes collectively called clock genes (Venzin and Oates, 2020). Somites are formed rhythmically, and sequentially from the anterior to the posterior of the elongating body axis with their spacing determined by the period of the genetic oscillations. These oscillations are maintained through a negative feedback loop of the clock genes (Oates et al., 2012) and coordinated among neighboring cells through contact-dependent Notch signaling and fluid flows that rearrange neighbor cells (Fig. 5.1; Venzin and Oates (2020)). This results in tissue level coherent waves of gene expression moving from the tail toward the head of the embryo. In humans, disruption of any of these components leads to diseases such as congenital scoliosis (Hubaud and Pourquié, 2014; Bulman et al., 2000; Bray, 2016; Eckalbar et al., 2012; Sparrow et al., 2008).

Despite years of research identifying components of this system in zebrafish, we still have a poor understanding of the mechanisms behind how these components come together to maintain oscillations in single cells, to synchronize these oscillations among neighboring cells, and how mutations disrupt oscillations and cause developmental defects. This is due, at least in part, to the lack of real-time dynamic information about the system (Zinani et al., 2020). This lack of real-time information means that potentially impactful defects in the timing of transcription are difficult to infer. Indeed, live transcription data in *Drosophila* have revealed detailed temporal information which resulted in the development of new models of transcriptional regulation (Dufourt et al., 2018; Eck et al., 2020). From snapshots of gene expression profiles of somitogenesis in fixed zebrafish embryos, we can see the waves of gene expression, but we don't know which aspect of these waves are being regulated, their frequency, duration, or amplitude, let alone how they're being regulated.

Fluorescent protein reporters have recently shed light on some aspects of the dynamics of this system (Delaune et al., 2012; Shih et al., 2015), but these fluorescent proteins take a relatively long time to mature and fluoresce, making inferences about the underlying gene expression dynamics difficult. In fact, the exact same sinusoidal temporal profile of a fluorescent reporter of a protein product of a clock gene, could arise from two different potential profiles of that gene's transcription. Figure 5.2 shows how the same sinusoidal wave
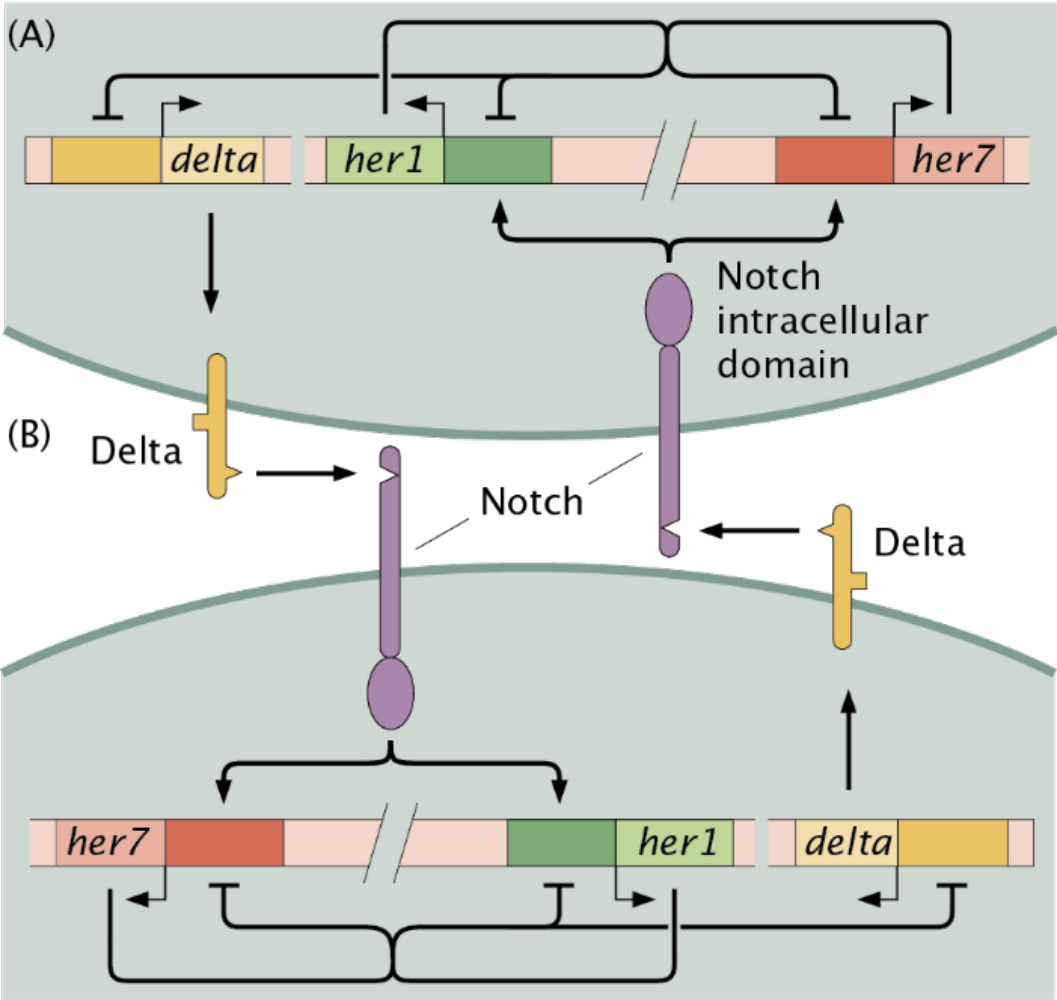
Figure 5.1: Gene regulatory network underlying somitogenesis in zebrafish. (A) A negative feedback loop of the clock genes, *her1* and *her7*, leads to oscillations within single cells. (B) Cell-to-cell coupling via the Notch signaling pathway synchronizes neighboring cells. Adapted from Oates (2020).
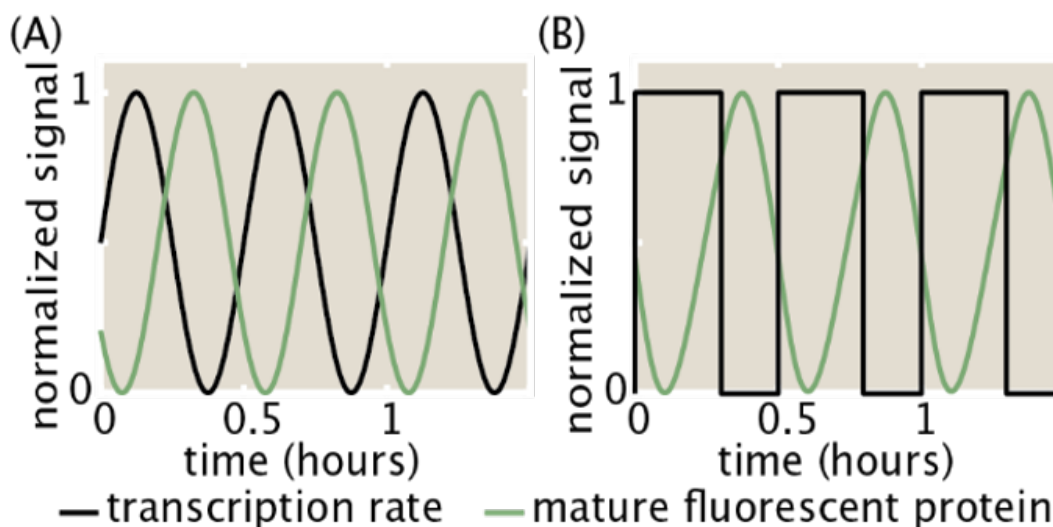
Figure 5.2: Fluorescent protein reporters cannot unequivocally reveal underlying oscillatory transcriptional dynamics. Illustrative transcriptional and protein dynamics for (A) a continuous transcriptional oscillation and (B) rhythmic gene expression bursts. The protein dynamics that result from these different transcriptional dynamics are identical. Transcription rate was prescribed to be a (A) sine function or (B) a step with period 0.5 h. Protein was simulated by assuming an mRNA decay rate of 10/h, a protein decay rate of 12/h, and a fluorescent protein maturation time of 0.3 h.

pattern of protein concentration could arise from either a transcription rate that changes following a similar sinusoidal wave pattern or a transcription rate that follows a step-function.

The MS2-MCP imaging system has been implemented to track transcriptional dynamics in a number of systems including bacteria (Golding et al., 2005), yeast (Hocine et al., 2013; Tutucci et al., 2018a), amoebas (Chubb et al., 2006), *Drosophila* (Garcia et al., 2013; Bothma et al., 2015), and human cells (Janicki et al., 2004; Darzacq et al., 2007; Ben-Ari et al., 2010; Brody et al., 2011). It has also been transgenically established in mice in order to investigate transcriptional dynamics in mouse cells (Lionnet et al., 2011).

Given zebrafish embryos' amenability for imaging, the MS2-MCP system is well suited to investigate their gene expression dynamics *in vivo*. However, the implementation of MS2 in zebrafish has been limited. A transgenic line of $\beta actin$ promoter driven NLS-td-MCP-eGFP zebrafish has been established (Campbell et al., 2015). However, this line exhibits MCP aggregates in nuclei, perhaps due to the use of tandem dimer MCP (td-MCP) (Wu et al., 2012). This line was used to visualized zygotic genome activation by imaging $\beta actin$ promoter driven MS2. DNA encoding this MS2 was injected into one cell stage embryos prior to imaging. This MCP line was also used to visualize transcript localization by imaging MS2-*nanos3* 3'UTR localizing to primordial germ cells (Campbell et al., 2015).

The same MCP design was used in another study examining how 3'UTRs promote mRNA localization to myelin sheaths (Yergert et al., 2020). In this study, the *sox10* promoter was used to drive tdMCP-eGFP expression. Both MCP and the MS2 reporter plasmid were transiently expressed in this study through their co-injection along with *tol2* transposase mRNA into one-cell stage zebrafish embryos. The same group also performed transient expression experiments to visualize a translational regulator and mRNA encoding a myelin-translated protein colocalizing in myelin sheaths (Fedder-Semmes and Appel, 2021).

An independent line of NLS-MCP-eGFP zebrafish was established in order to study how mRNA polarization orients motile cell polarity and directs tissue movement at sites of filapodia formation (Costa et al., 2019). In this study, a chimeric endothelial enhancer/promoter, *fli1ep*, was used to drive nls-MCP-eGFP expression. MS2 reporter plasmids were co-injected with *tol2* mRNA into one-cell stage embryos for mosaic expression analysis.

The MS2-MCP system was used in another zebrafish embryo study as a control to establish a new endogenous mRNA imaging system based on inactive mini-III RNase, an endonuclease that can bind and cleave double-stranded RNAs (Zhang et al., 2020). In this case, an mRNA probe was coinjected into embryos along with with MCP-GFP and the inactive mini-III RNase proteins.

To our knowledge, no stable transgenic lines of MS2 reporters have been established. Additionally, the MCP component has not been optimized to visualize and quantify transcriptional dynamics in live developing embryos.

We have successfully engineered transgenic zebrafish expressing an MS2 reporter for one of the clock genes involved in somitogenesis, *her1*, as well as several transgenic lines expressing fluorescently tagged MCP (Section 5.2.1). Preliminary data (Section 5.2.2) shows this reporter successfully tagging sites of *her1* transcription. Fluorescent puncta move as a wave from the posterior toward the anterior of the developing embryo (Fig. 5.9B; Video D.1.1). This is consistent with previous measurements of fluorescently tagged Her1 protein (Delaune et al., 2012; Shih et al., 2015; Soroldoni et al., 2014). Surprisingly, we found that these preliminary measurements of transcriptional dynamics in a developing vertebrate embryo challenge our expectations. Measurements show that the single-cell transcriptional dynamics do not match the smooth oscillations seen in the fluorescently tagged protein. Instead, the transcription occurs in punctuated, periodic bursts of activity. Transcriptional bursting has been repeatedly measured in other eukaryotes and has been found to be stochastic. However, simple simulations of bursting in zebrafish indicate that stochasticity would destroy the robust protein oscillations seen experimentally. This implies that some of the bursting parameters are being regulated but which ones and how?
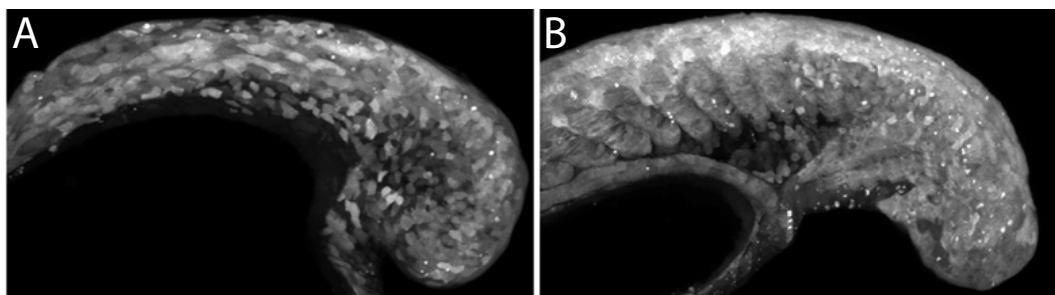
Figure 5.3: Transient expression experiment. (A and B) These sample images from the transient expression experiments show embryos that were injected with 100 pg *MCP-mNeonGreen* RNA into the cell of the one cell stage embryo. The embryo in (A) was co-injected with 50 pg of *MS2v5* RNA. mNeonGreen fluorescence is apparent in both embryos. Bright puncta are also present in both embryos regardless of the presence of an MS2 reporter. Imaging took place at approximately the 12-16 somite stage. Embryos are seen from a lateral perspective, looking at the side of the developing tail.

5.2 Results

5.2.1 Transient expression experiments and generation of transgenic zebrafish

As a first test of our MS2 and MCP constructs, we performed transient expression experiments using *MCP-mNeonGreen* RNA and the non-repetitive *MS2v5* DNA (Wu et al., 2015), both driven by the ubiquitous *Sp6* promoter (Section 5.5.1). These constructs were injected into the cell of one cell stage wild type zebrafish embryos and imaged on a Zeiss Axio Examiner.Z1 multiphoton laser scanning microscope. We tested injecting different quantities (between 0.5 pg and 100 pg) of *MCP-mNeonGreen* RNA. Fluorescence was evident via confocal imaging (Fig. 5.3; Section 5.5.3), indicating that our fluorophore was working. However, small bright puncta were seen whether *MS2v5* DNA was coinjected or not, as seen in the sample images in Figure 5.3. We hypothesized that these puncta are the result of aggregates due to high protein concentration, or some other artifact stemming from our transient injections.

In order to avoid this potential pitfall of transient expression experiments using injections, we generated transgenic lines stably expressing MCP-mNeonGreen using *I-SceI* meganuclease-mediated transgenesis (See Section 5.5.2; Soroldoni et al. (2009)). The MCP constructs (Fig. 5.4A) used to create the transgenic lines utilize a *ubiquitin* promoter (Mosimann et al., 2011) and encode either a single amino acid linker or a ten amino acid linker between the coat protein and fluorescent protein (either mNeonGreen or eGFP; Section 5.2.1). Founders were identified for each construct and tanks of their offspring were raised to adulthood for experiments. In the case of the MCP-10 amino acid linker-mNeonGreen (MCP-10AA-

mNeonGreen) construct, three founders with different expression levels (weak, medium, and strong) were identified and their offspring raised. For both of the other constructs, only a single founder was identified with weak expression levels, similar to the MCP-10AA-mNeonGreen weak line. These lines were not further utilized. In all of the MCP lines, mosaic expression (different concentrations in each cell) is evident (Figs. 5.5, 5.6, and 5.9B; Video D.1.1). Additionally, strong expression is seen in the skin (Fig. 5.5; Video D.1.1) and notochord (Figs. 5.5A, 5.6, 5.9B; Video D.1.1). No nuclear or cytoplasmic aggregates were apparent via confocal (Fig. 5.5A) or light-sheet imaging (Fig. 5.6A).

Once the transgenic MCP fish were established, transient injection of MS2 was used to check if the imaging system had the potential to be successful once the MS2 construct was also established transgenically. To do this, *Sp6* promoter-driven *MS2v5* DNA (Section 5.2.1) was injected into one-cell stage embryos expressing one of the MCP constructs. The resulting embryos were imaged on a Zeiss Axio Examiner.Z1 multiphoton laser scanning microscope at about the 12 somite stage to check for MS2 spots (Section 5.5.3). Figure 5.5A shows an MCP-mNeonGreen expressing embryo which was not injected with *MS2v5* DNA. Figures 5.5B and 5.5B' show two confocal planes of an MCP-mNeonGreen embryo that was injected with 52 pg of *MS2v5* DNA. Bright spots (red arrows) are only apparent in the injected embryo. These transient expression experiments resulted in very few, very bright spots. This may be due to regions with high concentrations of *MS2v5* DNA, which resulted from an uneven distribution of DNA after its injection. The presence of the spots, however, implied that the injected MS2 loops were being bound by MCP-mNeonGreen. This promising result encouraged us to move forward by making transgenic MS2 fish.

The MS2 reporter transgene (Fig. 5.4B) uses the shared regulatory region between *her1* and *her7* (Section 5.2.1). The *her7* coding sequence was replaced with nuclear localized mKate2, a red fluorescent protein, which serves both as a marker of transgenesis (and the presence of MS2) during embryo screening as well as a nuclear marker during imaging. The *her1* promoter drives the expression of the *her1* coding sequence followed by *MS2v5* which is inserted just before the *her1* 3'UTR. Like the MCP fish, the transgenic MS2 fish were made using *I-SceI* meganuclease-mediated transgenesis (See Section 5.5.2; Soroldoni et al. (2009)). Three founders were identified and their offspring were raised.

The MS2 reporter fish were crossed with MCP fish and the resulting embryos were imaged on a Viventis LS1 Live light sheet microscope system (Section 5.5.4). Control embryos which did not have the MS2 transgene did not show evidence of transcription spots (Fig. 5.6A). In the embryos which did receive the MS2 transgene (known by the presence of the mKate2 signal in the nuclei) transcription spots and sometimes cytoplasmic speckles were present (Fig. 5.6B). The cytoplasmic speckles may be due to MS2 signal from mRNA exported to the cytoplasm.

5.2.2 Preliminary data

Preliminary data was taken on a Viventis LS1 Live light sheet microscope system at 28°C. Waves of *her1* driven gene expression were evident. Bright, synchronized transcription spots
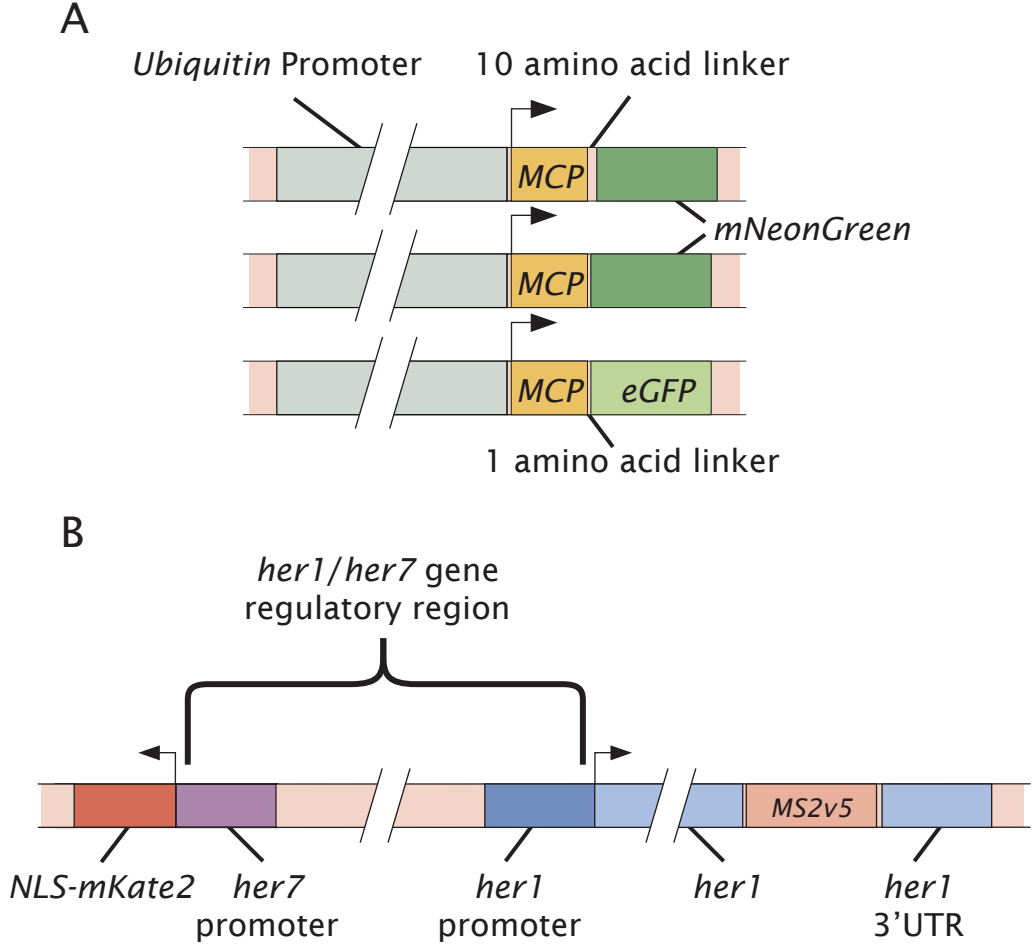
Figure 5.4: Constructs used in generating transgenic zebrafish. (A) Three different MCP constructs were made. They all contained *MCP* and either *mNeonGreen* or *eGFP*. Either one amino acid or ten amino acids linked MCP with the fluorescent protein. (B) The MS2 construct utilized the regulatory region shared between *her1* and *her7*. The *her7* coding sequence was replaced with *mKate2* containing a nuclear localization signal (NLS; from Campbell et al. (2015)). The *her1* coding sequence remained present and *MS2v5* was inserted before the *her1* 3'UTR.
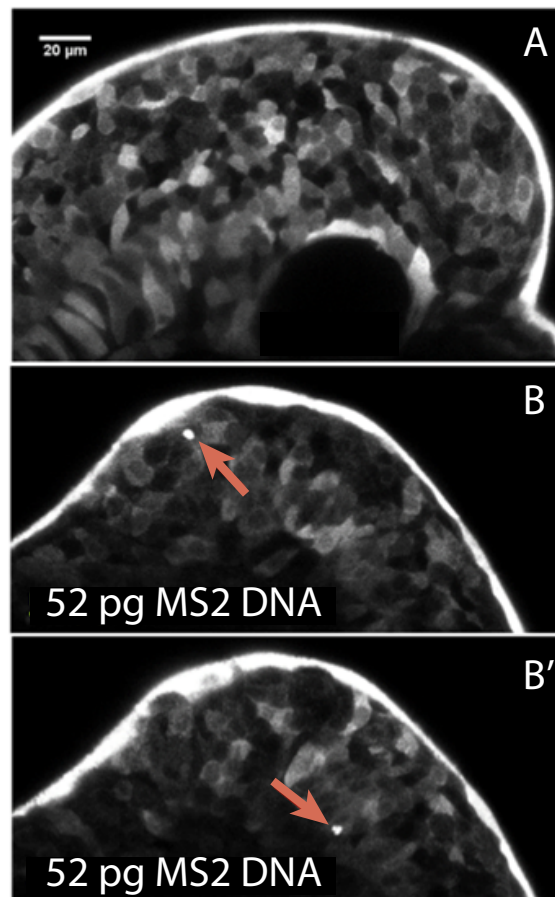
Figure 5.5: MS2 injection experiment. MCP-mNeonGreen expressing embryos were imaged (A) without injection or (B,B') after injecting the cell of the one-cell stage embryo with 52 pg *MS2v5* DNA. Bright puncta are present only in the injected embryo. Imaging took place at approximately the 12 somite stage. Embryos are seen from a lateral perspective looking at the side of the developing tail bud. Bright MCP expression can be seen in the skin of the embryos. Scale bar is 20 μm
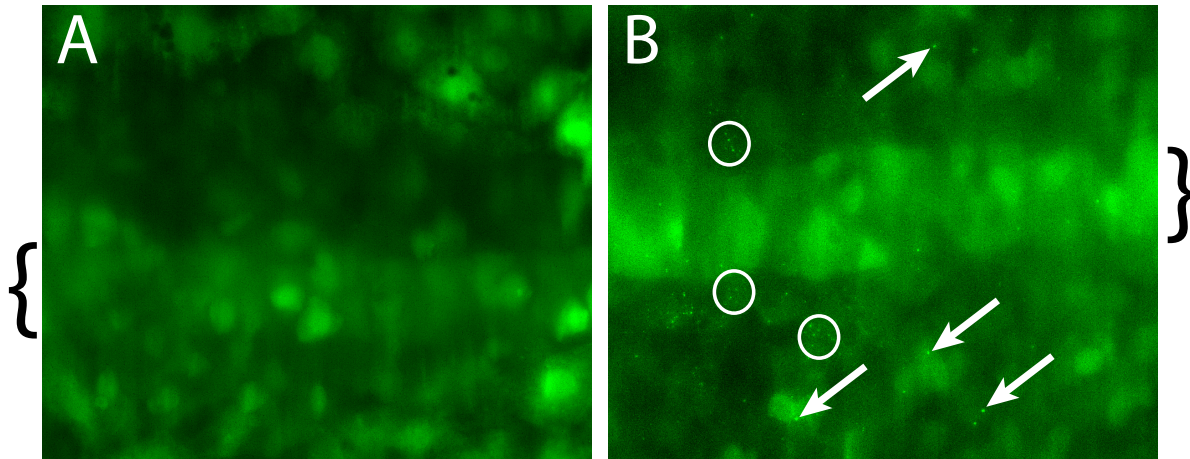
Figure 5.6: Preliminary light sheet images. Light Sheet Images of embryos expressing MCP-mNeonGreen (medium line) (A) alone or (B) along with the MS2 reporter transgene. Both embryos are seen from a dorsal perspective looking down at the future back of the animal. Notochords can be seen running horizontally across both images. Their approximately widths and positions are marked by brackets. These embryos were approximately at the four-somite stage. Mosaic expression of MCP-mNeonGreen is evident. (A) No transcription spots are evident in the fish without the MS2 transgene or mKate2 nuclear signal. (B) Transcription spots (arrows) and cytoplasmic speckles (circled) are evident only in the fish with the MS2 transgene (known by the presence of the mKate2 signal in nuclei, not shown). Note that cytoplasmic speckles were not evident in all embryos and may be due to the copy number of the MS2 transgene in the particular line imaged.

appeared in lateral rows of cells. The spots turned on and off in oscillations. This caused the spots to appear to traverse the field of cells, moving as a traveling wave from the posterior toward the anterior (Fig. 5.9B; Video D.1.1).

Transcription spots were analyzed using custom-written software following the protocol in Garcia et al. (2013) (Section 5.5.5). Briefly, this procedure involves segmenting each MCP-mNeonGreen transcription spot based on its fluorescence, and calculating its intensity as a function of time. Nuclei were tracked manually. Automated nuclear tracking was not possible due to the dim signal and quick bleaching of the mKate2 signal. Recently, the MCP fish were crossed with fish carrying an *h2b-mScarlet* transgene. This transgene results in a much brighter and robust nuclear signal (Fig. 5.7). Transcription oscillations in single cells (example shown in Figure 5.9D) show bursts of expression rather than the smooth oscillations seen in previous measurements of fluorescently tagged Her1 protein (Delaune et al., 2012; Shih et al., 2015; Soroldoni et al., 2014) which can be seen in Figure 5.8.
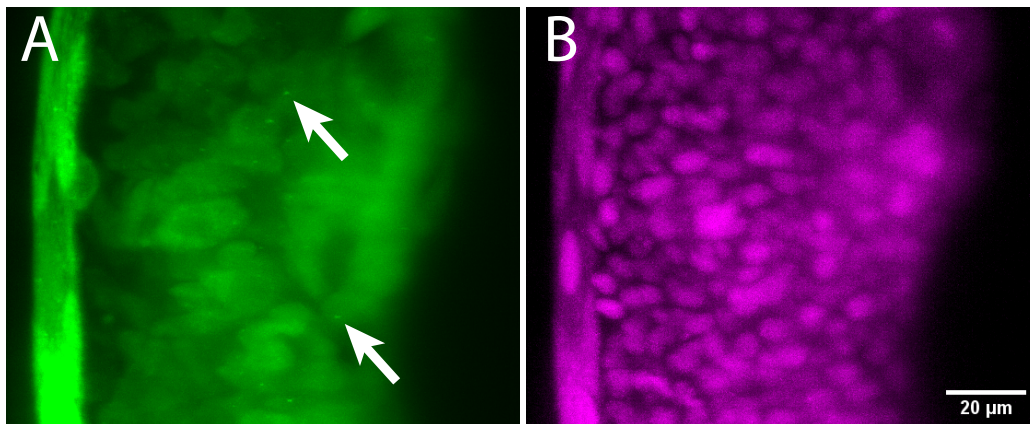
Figure 5.7: Sample confocal image of an embryo with tagged histone. This embryo is expressing MCP-10AA-mNeonGreen (medium line), the *her1-MS2v5* reporter, and h2b-mScarlet. (A) Sample transcription spots, seen in the mNeonGreen channel, are marked with arrows. (B) Nuclei marked with both nuclear localized mKate2 and h2b-mScarlet can be seen in the red channel. These images are maximum projections of a 30 μm substack. Scale bar is 20 μm.
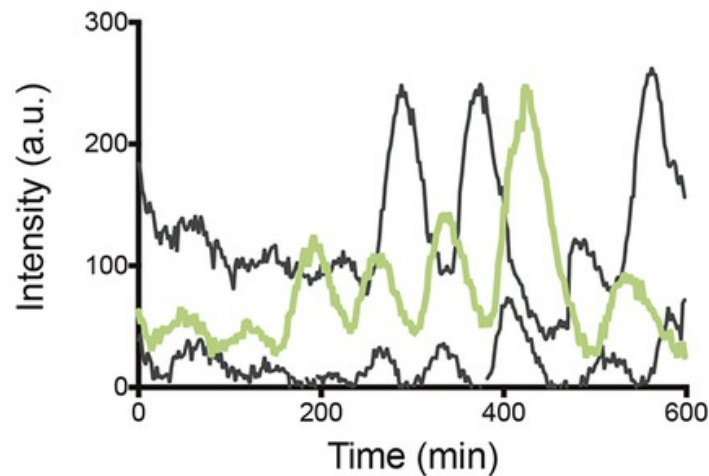
Figure 5.8: Her1-YFP oscillations. YFP signal intensity (arbitrary units) measured by tracking a regions of interest over 3 single tailbud cells (green trace). The gray traces are two additional single cells in cell culture. Plotted in 2-min intervals. Reprinted from Webb et al. (2016).
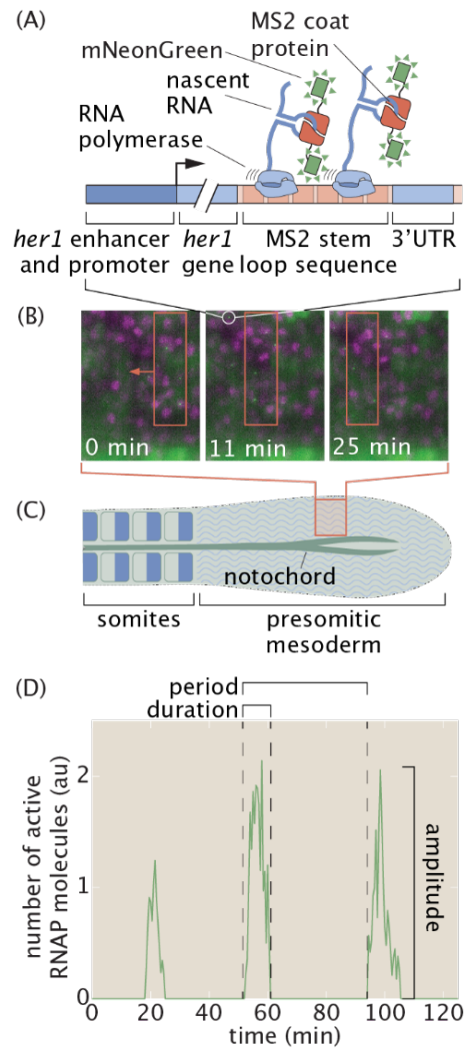
Figure 5.9: Revealing transcriptional dynamics in somitogenesis. (A) The MS2 reporter for the cyclic *her1* gene, fluorescence is proportional to the number of nascent RNA molecules. (B) Typical field of view showing sites of transcription in single nuclei. Note the wave of transcriptional activity propagating along the field of view. The bright notochord is evident at the bottom of the frame. (C) Region imaged along the presomitic mesoderm. (D) Preliminary data show that oscillations are created by bursts of transcriptional activity defined by their period, duration and amplitude.

5.3 Discussion

While designing the MCP and MS2 constructs for use in zebrafish, several considerations were made. Firstly, we wanted MCP to be at an appropriate nuclear concentration. There needs to be enough MCP so that all of the MS2 loops become bound upon transcription, a phenomenon termed "saturation." At the same time, we want to have the MCP concentration as low as possible to reduce the background fluorescence and to avoid any potential protein aggregation effects. Reducing background enables higher sensitivity of the system——fewer active polymerases are required to be transcribing the gene in order for the signal to be visible above the background. If the MCP concentration is too low, the signal would not be reliably quantifiable because the loops would not be saturated. If the MCP concentration is too high, the signal-to-noise ration would decrease and aggregates might be more likely to form. We chose to omit a nuclear localization signal from our MCP design to keep the nuclear background fluorescence lower. We also chose to not use td-MCP, which has been found to aggregate in fruit fly embryos (Armando Reimer and Myron Barber Child, personal communication, 2021). This strategy seems to have been successful in preventing aggregation of MCP in nuclei. Further test are needed to determine whether the MS2 loops are being fully saturated. The weak, medium, and strong expression of MCP-mNeonGreen in different transgenic lines will enable us to tune the concentration of MCP to achieve our goals.

While we have successfully established transgenic zebrafish lines for several MCP constructs, some optimization remains to be done. The *ubiquitin* promoter was reported to drive constitutive transgene expression throughout all developmental stages (Mosimann et al., 2011). However, the mosaic expression of MCP is unexpected and may make characterization of the various transgenic lines more challenging.

Injection of *MS2v5* DNA (Fig. 5.5) indicated that the MS2 system would likely work in our MCP fish, but that injections were not a reliable way to investigate *her1* expression dynamics. The MS2 puncta which resulted from DNA injections were much larger than expected and present in very few cells. This was likely due to an uneven distribution of injected DNA that wasn't distributed evenly to all daughter cells following cell division. We determined that the only way to reliably measure *her1* expression and neighbor cell coordination would be through stable transgenesis. Indeed, early imaging of the *her1* reporter transgene seems to show that it is replicating *her1* expression patterns and MS2 puncta seem to be present in most presomitic mesoderm cells.

While all of the transgenic fish seem to be healthy, *In situ* hybridization will be performed to check for possible somitogenesis defects or alteration of *her1* expression patterns. *MS2v5* and *her1* expression patterns will also be compared via *in situ* hybridization to ensure that the reporter transgene is faithfully replicated *her1* expression patterns.

5.3.1 Future directions

Thus far, we have worked to establish the MS2-MCP imaging system in zebrafish specifically to examine the complex gene regulatory and signaling network governing somitogenesis.
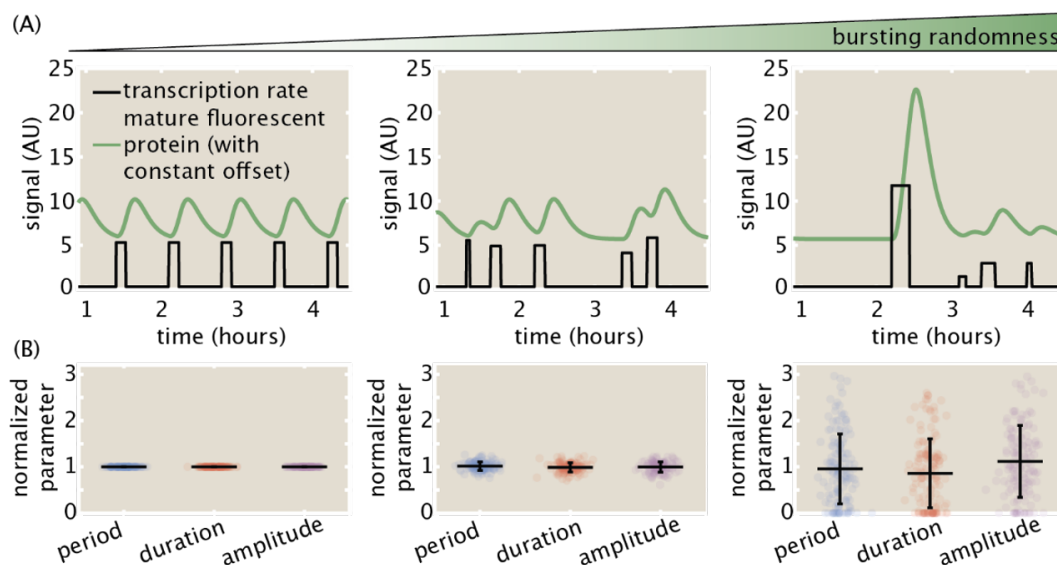
Figure 5.10: Consequences of deterministic and stochastic burst dynamics. (A) Simulations of protein dynamics resulting from increasingly stochastic values of the transcriptional busting parameters. (B) Parameter distributions used for simulations.

Preliminary data has been taken and we are confident that the system will enable us to answer some open questions about this regulation. Below are some of the first questions we will tackle through imaging and analyzing our new *her1* reporters.

How does the period, duration, and amplitude of transcriptional bursts result in robust and synchronized tissue level protein oscillations?

Transcriptional bursting is widespread in eukaryotes (Lammers et al., 2020b) and these bursts have been found to be stochastic. This means that their parameters, frequency, duration, and amplitude, can only be predicted probabilistically, not deterministically (Lammers et al., 2020a; Berrocal et al., 2020; Rodriguez et al., 2019; Halpern et al., 2015; Tunnacliffe et al., 2018; Zoller et al., 2018; Corrigan et al., 2016; Suter et al., 2011; Molina et al., 2013; Zenklusen et al., 2008; Padovan-Merhar et al., 2015; Bartman et al., 2016; Larson et al., 2011a). However, we found that in preliminary simulations of clock genes in zebrafish, stochasticity in bursting parameters can destroy the robust protein oscillations seen experimentally (Fig. 5.8). Figure 5.10 shows how the protein oscillations become more disrupted as stochasticity of the bursting parameters increases. This implies that at least some of the bursting parameters must be regulated during somitogenesis. This regulation is likely accomplished by the repressive activity of the clock genes, *her1* and *her7*, along with activation through the Notch signaling pathway. Which parameters are being regulated and to what degree, however, remains
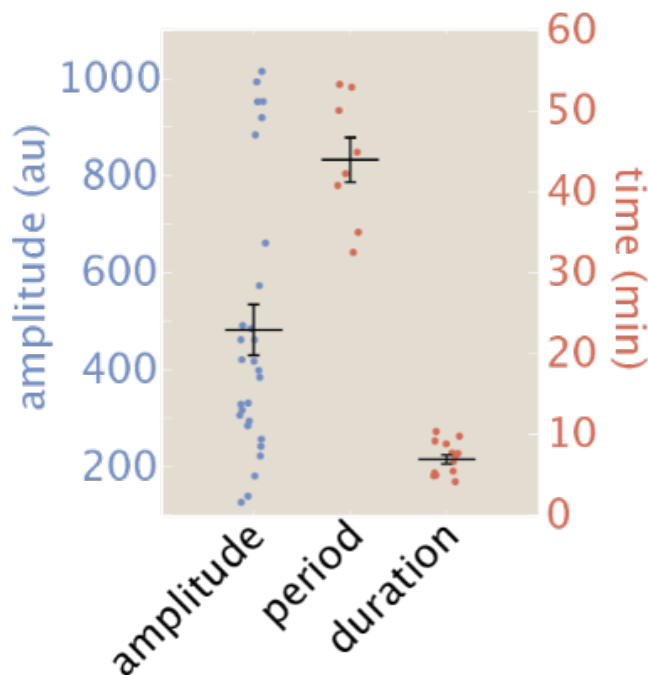
Figure 5.11: Preliminary measurements of bursting parameters. Amplitude, period, and duration of transcriptional bursts of the *her1-MS2* reporter. Data is from approximately ten cells tracked in two different embryos for two-to-three oscillations, each.

unknown.

We plan to extract the bursting parameters using novel theoretical and computation analysis that were developed for bursting observed in *Drosophila* (Bothma et al., 2014; Lammers et al., 2020a; Eck et al., 2020; Alamos et al., 2020; Lammers et al., 2020b; Berrocal et al., 2020). We will calculate the spread in the distributions of each parameter, measured in individual cells (preliminary data shown in Figure 5.11), and use simulations to determine whether the variability is small enough to result in sustained protein oscillations.

Beyond this single cell analysis, we will examine how these parameters are coordinated among neighboring cells to synchronize their transcriptional activity and give rise to tissue-level oscillations. Notch activation in fruit flies was found to increase burst duration and amplitude (Falo-Sanjuan et al., 2019). We hypothesize that, in somitogenesis, these two parameters are coordinated among neighboring cells, but burst period is not. We will test this hypothesis by measuring spatial correlations in bursting parameters using approaches recently established for studying protein level correlations in somitogenesis (Delaune et al., 2012; Bhavna, 2020).

How do clock gene and signaling mutants disrupt single-cell transcriptional dynamics and

neighbor cell synchronization?

Beyond examining the bursting parameters of the healthy somitogenesis gene regulatory network, we will examine which molecules modulate bursting parameters and how this regulation becomes disrupted in mutants. There are two main, but not mutually exclusive, mechanisms that could result in defective somite formation. The first is mutations altering the magnitudes of bursting parameters. This would still result in oscillations, but would affect downstream differentiation. The second mechanism is mutations increasing the variability in bursting parameters while leaving the average magnitudes unaffected. Similarly, cell signaling through the Notch pathway may be playing a role in ensuring that neighbor cells burst with similar magnitudes in their bursting parameters and might also reduce the intercellular variability of those parameters.

To investigate these possibilities, we will use well-established alterations of this system's molecular components while measuring live gene expression dynamics and then use theoretical and computational analysis to extract bursting parameters. We plan on initially studying two of the molecular components of the system: the clock gene *her7* and the Notch signaling pathway. Mutations in *her7* or *deltaC* (a component of the Notch pathway) result in effectively identical developmental phenotypes—disruption of proper somite formation. This is interesting because Her7 is a repressor of the clock genes that drives oscillations in individual cells while DeltaC synchronizes oscillations in neighboring cells through the activation of clock genes (Venzin and Oates, 2020). To solve the mystery of how mutations of these two components, which play different roles within this system, result in identical phenotypes, we will use two methods. First, we will use DAPT, a small molecule inhibitor of the Notch pathway, to replicate *deltaC* mutations (Venzin and Oates, 2020). Second, we will remove Her7 from the system using established morpholinos (Oates and Ho, 2002). Through these two approaches, we will uncover not only how the loss of Her7 and the Notch pathway affect bursting parameter averages and variability, but whether these two molecular components affect the same parameters or act through distinct molecular mechanisms.

Another interesting aspect of this investigation is the dichotomy of the of double mutant's effects. Combined *her7* and *deltaC* mutants lead to *her1* expression that is intermediate to the two individual mutants (Oates et al., 2005). The loss of Her7 repression leads to higher *her1* levels while the loss of the Notch pathway activation leads to lower *her1* levels. The intermediate *her1* levels in the double mutant imply that the two mutations work additively and that Her7 and Notch act at independent sites in the *her1* regulatory region. However, the phenotype of the double mutant is more extreme than the individual mutants with the disruption of somite boundaries occurring sooner (more anterior) during somitogenesis, reminiscent of the complete elimination of the oscillatory circuit (Oates et al., 2005). This implies synergy in the double mutant and suggests that Her7 and the Notch pathway interact at the *her1* regulatory region. To determine whether there truly is synergy between these two components, we will combine DAPT and her7 morpholinos and compare bursting parameters and cell synchronization to those of the individual treatments.

While preliminary data from our transgenic MS2 reporter and MCP zebrafish show promise, we will continue to optimize imaging conditions as well as our data analysis pipeline. We are confident that, in the near future, through the investigations enabled by these transgenic fish, we will gain valuable insights into the dynamic mechanisms of regulation of the zebrafish somitogenesis regulatory network as well as its misregulation due to mutations of network components. Additionally, this work may aid us in understanding how stochastic bursts of transcription can give rise to reproducible developmental outcomes.

5.4 Acknowledgements

5.5 Methods and materials

5.5.1 Genetic constructs

Codon optimization was performed using Integrated DNA Technology's (IDT's) Codon Optimization Tool (https://www.idtdna.com/pages/tools/codon-optimization-tool).

Genetic constructs can be found at the Benchling links listed below.

The following plasmids were used in transient expression (injection) experiments:

PCS2+ *Sp6-MCP-mNeonGreen* (https://benchling.com/eeck/f/lib_YSpQhHNw-dissertation-zebrafish-ms2-constructs/seq_ACd7VZer-copy-of-pcs2-mcp-ng/edit)

PCS2+ *Sp6-Ms2v5-LacZ* (https://benchling.com/eeck/f/lib_YSpQhHNw-dissertation-zebrafi sh-ms2-constructs/seq_ACd7VZer-copy-of-pcs2-mcp-ng/edit)

The following constructs were used in the generation of transgenic zebrafish:

pBSII *ISceI-ubi-MCP-1AA-EGFP* (https://benchling.com/eeck/f/lib_YSpQhHNw-dissertati on-zebrafish-ms2-constructs/seq_oVa8nxKM-copy-of-piscei_ubidrmcp-1aa-egfp/edit)

pBSII *ISceI-ubi-MCP-1AA-mNeonGreen* (https://benchling.com/garcialab/f/lib_sJIgWad5-e mail-for-robb/seq_XGxmVyZq-piscei_ubidrmcp-1aa-drng/edit)

pBSII *ISceI-ubi-MCP-10AA-mNeonGreen* (https://benchling.com/eeck/f/lib_YSpQhHNw-di ssertation-zebrafish-ms2-constructs/seq_TPjUJwRO-piscei_ubidrmcp-10aa-drng/edit)

pShave *ISceI-Δher7-NLS-mKate2-her1-10AA-MS2v5* (https://benchling.com/eeck/f/lib_YS pQhHNw-dissertation-zebrafish-ms2-constructs/seq_BszOU4Oi-copy-of-pshave_iscei_her7nls drk2_her1-10aa-ms2v5/edit)

5.5.2 Transgenesis

Stable transgenic zebrafish lines were generated using *I-SceI* meganuclease-mediated transgenesis (Soroldoni et al., 2009). Briefly, DNA was coinjected with *I-sceI* meganuclease into the cell of one-cell stage WT embryos (AB or TL).

Post-injection, *I-SceI* enzyme activity was assayed by electrophoresis. One day later, embryos were screened briefly to confirm fluorescence of mKate2, mNeonGreen, or eGFP, whichever was appropriate to the injected DNA. Injected embryos were raised to adulthood and outcrossed to WT (AB or TL) zebrafish. If offspring exhibited fluorescence, they were selected to be raised for experiments and their parent (which had been injected as an embryo) was selected as a founder.

5.5.3 Confocal imaging

A Zeiss Axio Examiner.Z1 multiphoton laser scanning microscope (Zeiss, Germany) was used for imaging during transient expression experiments and *MS2* DNA injections into MCP expressing fish. The following settings were used: Frame size 1024x1024 pixels, Speed 8 (pixel dwell time 1.27 μs), Bidirectional scanning: On, 12 bit depth, sum 4 line scans, zoom 2.1 (pixel size 0.2 μm. The Alexa 488 channel was imaged with laser power 5.5 and pinhole 90, and Gain 850. The detector was set to a window spanning 491 nm to 563 nm. A z-stack spanning 20 μm was taken with a step size of 0.5 μm (41 steps total).

5.5.4   Light sheet imaging

Embryos were imaged with a Viventis LS1 Live Light Sheet Microscope System (Viventis Microscopy, Lasaunne, Switzerland). The embryos were dechorionated and imaged dorsally in a 3D-printed imaging chamber filled with 600 μl E3 fish medium. The imaging chamber had troughs of width of 650 μm. The excitation light sheet was composed of a dual illumination scanned Gaussian beam of approximately 3.3 μm thickness (full width at half maximum) using 488 and 561 nm lasers. The signal was collected with a Nikon CFI75 Apo LWD 25x/1.1 NA objective and through 525/50-25 and 561/25 nm bandpass filters respectively onto an Andor Zyla 4.2 Plus sCMOS camera. An incubation chamber maintained the embryos at 28°C. 200 z-slices with a spacing of 0.5 μm and a frame interval of 30 seconds were acquired for up to 4 hrs. The pixel size under these settings was 0.17 μm. Embryos were imaged from approximately the two to six somite stage.

5.5.5   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 mKate2 signal as a nuclear mask, segmenting each transcription spot based on its fluorescence, and calculating the intensity of each MCP-mNeonGreen transcriptional spot as a function of time. Nuclei and spot segmentation as well as nuclear tracking was curated manually.

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.

Simon Alamos, Armando Reimer, Krishna K. Niyogi, and Hernan G Garcia. Quantitative imaging of rna polymerase ii activity in plants reveals the single-cell basis of tissue-wide transcriptional dynamics. *bioRxiv*, page 274621, 2020. doi: 10.1101/2020.08.30.274621.

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 or 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.

R. Barth, K. Bystricky, and H. A. Shaban. Coupling chromatin structure and dynamics by live super-resolution imaging. *Science Advances*, 6, 7 2020. ISSN 23752548. doi: 10.1126/ SCIADV.AAZ2196. URL `/pmc/articles/PMC7458449//pmc/articles/PMC7458449/ ?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7458449/`.

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.

Stephen B Baylin and Peter A Jones. A decade of exploring the cancer epigenome — biological and translational implications. 2011. doi: 10.1038/nrc3130. URL `www.nature. com/reviews/cancer`.

Y. Ben-Ari, Y. Brody, N. Kinor, A. Mor, T. Tsukamoto, D. L. Spector, R. H. Singer, and Y. Shav-Tal. The life of an mrna in space and time. *J Cell Sci*, 123(Pt 10):1761–74, 2010. ISSN 1477-9137 (Electronic) 0021-9533 (Linking). doi: 10.1242/jcs.062638.

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 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].

Rajasekaran Bhavna. Segmentation clock dynamics is strongly synchronized in the forming somite. *Developmental Biology*, 460:55–69, 4 2020. ISSN 1095564X. doi: 10.1016/j.ydbio.2019.03.008.

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.

Johannes Bischof, Robert K. Maeda, Monika Hediger, François Karch, and Konrad Basler. An optimized transgenesis system for drosophila using germ-line-specific c31 integrases. *Proceedings of the National Academy of Sciences of the United States of America*, 104:3312–3317, 2 2007. ISSN 00278424. doi: 10.1073/PNAS.0611511104.

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.

Diana P Bratu, Byeong-Jik Cha, Musa M Mhlanga, Fred Russell Kramer, and Sanjay Tyagi. Visualizing the distribution and transport of mrnas in living cells. 2003. URL `www.pnas.orgcgidoi10.1073pnas.2233244100`.

Sarah J. Bray. Notch signalling in context. *Nature Reviews Molecular Cell Biology 2016 17:11*, 17:722–735, 8 2016. ISSN 1471-0080. doi: 10.1038/nrm.2016.94. URL `https://www.nature.com/articles/nrm.2016.94`.

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.

Alexander S. Brodsky and Pamela A. Silver. Identifying proteins that affect mrna localization in living cells. *Methods*, 26:151–155, 2 2002. ISSN 1046-2023. doi: 10.1016/S1046-2023(02) 00017-8.

Yehuda Brody, Noa Neufeld, Nicole Bieberstein, Sebastien Z. Causse, Eva Maria Böhnlein, Karla M. Neugebauer, Xavier Darzacq, and Yaron Shav-Tal. The in vivo kinetics of rna

polymerase ii elongation during co-transcriptional splicing. *PLOS Biology*, 9:e1000573, 2011. ISSN 1545-7885. doi: 10.1371/JOURNAL.PBIO.1000573. URL `https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000573`.

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).

Michael P. Bulman, Kenro Kusumi, Timothy M. Frayling, Carole McKeown, Christine Garrett, Eric S. Lander, Robb Krumlauf, Andrew T. Hattersley, Sian Ellard, and Peter D. Turnpenny. Mutations in the human delta homologue, dll3, cause axial skeletal defects in spondylocostal dysostosis. *Nature Genetics 2000 24:4*, 24:438–441, 4 2000. ISSN 1546-1718. doi: 10.1038/74307. URL `https://www.nature.com/articles/ng0400_438`.

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).

P. D. Campbell, J. A. Chao, R. H. Singer, and F. L. Marlow. Dynamic visualization of transcription and rna subcellular localization in zebrafish. *Development*, 2015. ISSN 1477-9129 (Electronic) 0950-1991 (Linking). doi: 10.1242/dev.118968.

Adam D Cawte, Peter J Unrau, and David S Rueda. Live cell imaging of single rna molecules with fluorogenic mango ii arrays. *Nature Communications*, 11, 2020. doi: 10.1038/s41467-020-14932-7. URL `https://doi.org/10.1038/s41467-020-14932-7`.

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.

Jianbo Chen, Olga Nikolaitchik, Jatinder Singh, Andrew Wright, Craig E. Bencsics, John M. Coffin, Na Ni, Stephen Lockett, Vinay K. Pathak, and Wei Shau Hu. High efficiency of hiv-1 genomic rna packaging and heterozygote formation revealed by single virion analysis. *Proceedings of the National Academy of Sciences*, 106:13535–13540, 8 2009. ISSN 0027-8424. doi: 10.1073/PNAS.0906822106. URL `https://www.pnas.org/content/106/32/13535https://www.pnas.org/content/106/32/13535.abstract`.

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.

Jean Michel Cioni, Julie Qiaojin Lin, Anne V. Holtermann, Max Koppers, Maximilian A.H. Jakobs, Afnan Azizi, Benita Turner-Bridger, Toshiaki Shigeoka, Kristian Franze, William A. Harris, and Christine E. Holt. Late endosomes act as mrna translation platforms and sustain mitochondria in axons. *Cell*, 176:56–72.e15, 1 2019. ISSN 0092-8674. doi: 10.1016/J.CELL.2018.11.030.

J. E. Cohen. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biol*, 2(12):e439, 2004.

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.

Guilherme Costa, Joshua Bradbury, Nawseen Tarannum, and Shane P. Herbert. mrna compartmentalisation spatially orients tissue morphogenesis. *bioRxiv*, page 374850, 8 2019. doi: 10.1101/374850. URL https://www.biorxiv.org/content/10.1101/374850v3https://www.biorxiv.org/content/10.1101/374850v3.abstract.

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.

Getulio Pereira de Oliveira, Eric Zigon, Gaenna Rogers, Danny Davodian, Shulin Lu, Tijana Jovanovic-Talisman, Jennifer Jones, John Tigges, Sanjay Tyagi, and Ionita Calin Ghiran. Detection of extracellular vesicle rna using molecular beacons. *iScience*, 23:100782, 1 2020. ISSN 2589-0042. doi: 10.1016/J.ISCI.2019.100782.

E. A. Delaune, P. Francois, N. P. Shih, and S. L. Amacher. Single-cell-resolution imaging of the impact of notch signaling and mitosis on segmentation clock dynamics. *Dev Cell*, 23(5):995–1005, 2012. ISSN 1878-1551 (Electronic) 1534-5807 (Linking). doi: 10.1016/j.devcel.2012.09.009.

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.

Elena V Dolgosheina, Sunny C Y Jeng, Shanker Shyam, S Panchapakesan, Razvan Cojocaru, Patrick S K Chen, Peter D Wilson, Nancy Hawkins, Paul A Wiggins, and Peter J Unrau. Rna mango aptamer-fluorophore: A bright, high-affinity complex for rna labeling and tracking. 2014. doi: 10.1021/cb500499x. URL `https://pubs.acs.org/sharingguidelines`.

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.

Walter L. Eckalbar, Rebecca E. Fisher, Alan Rawls, and Kenro Kusumi. Scoliosis and segmentation defects of the vertebrae. *Wiley interdisciplinary reviews. Developmental biology*, 1:401–423, 5 2012. ISSN 1759-7692. doi: 10.1002/WDEV.34. URL `https://pubmed.ncbi.nlm.nih.gov/23801490/`.

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[pii].

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[pii]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[pii]10.1038/msb.2009.97.

J. Falo-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.

Karlie N. Fedder-Semmes and Bruce Appel. The akt-mtor pathway drives myelin sheath growth by regulating cap-dependent translation. *Journal of Neuroscience*, 41:8532–8544, 10 2021. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0783-21.2021. URL https://www.jneurosci.org/content/41/41/8532https://www.jneurosci.org/content/41/41/8532.abstract.

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.

Grigory S. Filonov, Jared D. Moon, Nina Svensen, and Samie R. Jaffrey. Broccoli: Rapid selection of an rna mimic of green fluorescent protein by fluorescence-based selection and directed evolution. *Journal of the American Chemical Society*, 136:16299–16308, 11 2014. ISSN 15205126. doi: 10.1021/JA508478X/SUPPL_FILE/JA508478X_SI_001.PDF. URL https://pubs.acs.org/doi/abs/10.1021/ja508478x.

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.

K. M. Forrest and E. R. Gavis. Live imaging of endogenous rna reveals a diffusion and entrapment mechanism for nanos mrna localization in drosophila. *Curr Biol*, 13(14): 1159–68, 2003. ISSN 0960-9822 (Print) 0960-9822 (Linking). doi: S0960982203004512[pii].

Vincent François, Mark Solloway, Jason W. O'Neill, John Emery, and Ethan Bier. Dorsal-ventral patterning of the drosophila embryo depends on a putative negative growth factor encoded by the short gastrulation gene. *Genes Development*, 8:2602–2616, 11 1994. ISSN 0890-9369. doi: 10.1101/GAD.8.21.2602. URL http://genesdev.cshlp.org/content/8/21/2602http://genesdev.cshlp.org/content/8/21/2602.abstract.

C. Fritzsch, 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.

D. Fusco, N. Accornero, B. Lavoie, S. M. Shenoy, J. M. Blanchard, R. H. Singer, and E. Bertrand. Single mrna molecules demonstrate probabilistic movement in living mammalian cells. *Curr Biol*, 13(2):161–7, 2003. ISSN 0960-9822 (Print) 0960-9822 (Linking).

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: 1015616108[pii]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.

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 Dundr, 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).

A. C. Groth, M. Fish, R. Nusse, and M. P. Calos. Construction of transgenic drosophila by using the site-specific integrase from phage phic31. *Genetics*, 166(4):1775–82, 2004. ISSN 0016-6731 (Print) 0016-6731 (Linking). doi: 166/4/1775[pii].

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.

Keren Bahar Halpern, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. Bursty gene expression in the intact mammalian liver. *Molecular Cell*, 58:147–156, 4 2015. ISSN 10974164. doi: 10.1016/J.MOLCEL.2015.01.027/ATTACHMENT/B4FA4D3D-BFF7-4540-83ED-478CFC3CBDB8/MMC3.XLSX.

URL `http://www.cell.com/article/S1097276515000507/fulltexthttp://www.cell.com/article/S1097276515000507/abstracthttps://www.cell.com/molecular-cell/abstract/S1097-2765(15)00050-7`.

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.

Ross C. Hardison and James Taylor. Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics 2012 13:7*, 13:469–483, 6 2012. ISSN 1471-0064. doi: 10.1038/nrg3242. URL `https://www.nature.com/articles/nrg3242`.

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.

Melissa M. Harrison and Michael B. Eisen. Transcriptional activation of the zygotic genome in drosophila. *Current Topics in Developmental Biology*, 113:85–112, 2015. ISSN 00702153. doi: 10.1016/bs.ctdb.2015.07.028.

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.

F. He, J. Ren, W. Wang, and J. Ma. A multiscale investigation of bicoid-dependent transcriptional events in drosophila embryos. *PLoS One*, 6(4):e19122, 2011. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). doi: 10.1371/journal.pone.0019122.

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*, 15(7):563–577, jul 1999. ISSN 1367-4803. doi: 10.1093/bioinformatics/15.7.563. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/15.7.563.

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. W. Hong, D. A. Hendrix, and M. S. Levine. Shadow enhancers as a source of evolutionary novelty. *Science*, 321(5894):1314, 2008. ISSN 1095-9203 (Electronic). doi: 321/5894/1314[pii]10.1126/science.1160631.

Caroline Hoppe, Jonathan R. Bowles, Thomas G. Minchington, Catherine Sutcliffe, Priyanka Upadhyai, Magnus Rattray, and Hilary L. Ashe. Modulation of the promoter activation rate dictates the transcriptional response to graded bmp signaling levels in the drosophila

embryo. *Developmental Cell*, 54:727–741.e7, 9 2020. ISSN 1534-5807. doi: 10.1016/J. DEVCEL.2020.07.007.

Alexis Hubaud and Olivier Pourquié. Signalling dynamics in vertebrate segmentation. *Nature Reviews Molecular Cell Biology 2014 15:11*, 15:709–721, 10 2014. ISSN 1471-0080. doi: 10.1038/nrm3891. URL https://www.nature.com/articles/nrm3891.

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.

S. M. Janicki, T. Tsukamoto, S. E. Salghetti, W. P. Tansey, R. Sachidanandam, K. V. Prasanth, T. Ried, Y. Shav-Tal, E. Bertrand, R. H. Singer, and D. L. Spector. From silencing to gene expression: real-time analysis in single cells. *Cell*, 116(5):683–98, 2004. ISSN 0092-8674 (Print) 0092-8674 (Linking).

A. M. Jaramillo, T. T. Weil, J. Goodhouse, E. R. Gavis, and T. Schupbach. The dynamics of fluorescently labeled endogenous gurken mrna in drosophila. *J Cell Sci*, 121(Pt 6):887–94, 2008. ISSN 0021-9533 (Print) 0021-9533 (Linking). doi: jcs.019091[pii]10.1242/jcs.019091.

Rajneesh Jha, Brian Wile, Qingling Wu, Aaron H. Morris, Kevin O. Maher, Mary B. Wagner, Gang Bao, and Chunhui Xu. Molecular beacon-based detection and isolation of working-type cardiomyocytes derived from human pluripotent stem cells. *Biomaterials*, 50:176–185, 5 2015. ISSN 0142-9612. doi: 10.1016/J.BIOMATERIALS.2015.01.043.

David G. Johnson and Sharon Y.R. Dent. Chromatin: Receiver and quarterback for cellular signals. *Cell*, 152:685–689, 2 2013. ISSN 0092-8674. doi: 10.1016/J.CELL.2013.01.017. URL http://www.cell.com/article/S0092867413000688/fulltexthttp://www.cell. com/article/S0092867413000688/abstracthttps://www.cell.com/cell/abstract/ S0092-8674(13)00068-8.

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.sciencemagazinedigital.org/sciencemagazine/22{_}june{_}2018/MobilePagedArticle.action?articleId=1404350{&}app=false{#}articleId1404350.

Won Jun Kang, Ye Lim Cho, Ju Ri Chae, Jong Doo Lee, Kyung Ju Choi, and Soonhag Kim. Molecular beacon-based bioimaging of multiple micrornas during myogenesis. *Biomaterials*, 32:1915–1922, 3 2011. ISSN 0142-9612. doi: 10.1016/J.BIOMATERIALS.2010.11.007.

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.

Jitendra S. Kanodia, Richa Rikhy, Yoosik Kim, Viktor K. Lund, Robert DeLotto, Jennifer Lippincott-Schwartz, and Stanislav Y. Shvartsman. Dynamics of the dorsal morphogen gradient. *Proceedings of the National Academy of Sciences of the United States of America*, 106:21707–21712, 12 2009. ISSN 00278424. doi: 10.1073/PNAS.0912395106.

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, 2020a. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1912500117.

N. C. Lammers, Y. J. Kim, J. Zhao, and H. G. Garcia. A matter of time: Using dynamics and theory to uncover mechanisms of transcriptional bursting. *Curr Opin Cell Biol*, 67:147–157, 2020b. ISSN 1879-0410 (Electronic) 0955-0674 (Linking). doi: 10.1016/j.ceb.2020.08.001.

S. Lange, Y. Katayama, M. Schmid, O. Burkacky, C. Brauchle, D. C. Lamb, and R. P. Jansen. Simultaneous transport of different localized mrna species revealed by live-cell imaging. *Traffic*, 9(8):1256–67, 2008. ISSN 1600-0854 (Electronic) 1398-9219 (Linking). doi: TRA763[pii]10.1111/j.1600-0854.2008.00763.x.

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.

Daniel R. Larson. What do expression dynamics tell us about the mechanism of transcription? *Current opinion in genetics development*, 21:591–599, 10 2011. ISSN 1879-0380. doi: 10. 1016/J.GDE.2011.07.010. URL `https://pubmed.ncbi.nlm.nih.gov/21862317/https://pubmed.ncbi.nlm.nih.gov/21862317/?dopt=Abstract`.

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 ¡em¿drosophila melanogaster¡/em¿ 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/.

Louisa M. Liberman, Gregory T. Reeves, and Angelike Stathopoulos. Quantitative imaging of the dorsal nuclear gradient reveals limitations to threshold-dependent patterning in drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 106:22317–22322, 12 2009. ISSN 00278424. doi: 10.1073/PNAS.0906227106.

Francis Lim, Thomas P. Downey, and David S. Peabody. Translational repression and specific rna binding by the coat protein of the pseudomonas phage pp7 *. *Journal of Biological Chemistry*, 276:22507–22513, 6 2001. ISSN 0021-9258. doi: 10.1074/JBC.M102411200. URL `http://www.jbc.org/article/S0021925820785339/fulltexthttp://www.jbc.org/article/S0021925820785339/abstracthttps://www.jbc.org/article/S0021-9258(20)78533-9/abstract`.

T. Lionnet, K. Czaplinski, X. Darzacq, Y. Shav-Tal, A. L. Wells, J. A. Chao, H. Y. Park, V. de Turris, M. Lopez-Jones, and R. H. Singer. A transgenic mouse for in vivo detection of endogenous labeled mrna. *Nat Methods*, 2011. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi: nmeth.1551[pii]10.1038/nmeth.1551.

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).

Shiqi Mao, Yachen Ying, Ruonan Wu, and Antony K. Chen. Recent advances in the molecular beacon technology for live-cell single-molecule imaging. *iScience*, 23:101801, 12 2020. ISSN 2589-0042. doi: 10.1016/J.ISCI.2020.101801.

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.

Colin D Medley, Timothy J Drake, Jeffrey M Tomasini, Richard J Rogers, and Weihong Tan. Simultaneous monitoring of the expression of multiple genes inside of single breast carcinoma cells. 2005. doi: 10.1021/ac050881y. URL `https://pubs.acs.org/sharingguidelines`.

Jane Mellor. Dynamic nucleosomes and gene transcription. *Trends in genetics : TIG*, 22:320–329, 6 2006. ISSN 0168-9525. doi: 10.1016/J.TIG.2006.03. 008. URL `https://pubmed.ncbi.nlm.nih.gov/16631276/https://pubmed.ncbi.nlm. nih.gov/16631276/?dopt=Abstract`.

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`.

N. Molina, D. M. Suter, R. Cannavo, B. Zoller, I. Gotic, and F. Naef. Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. *Proc Natl Acad Sci U S A*, 110(51):20563–8, 2013. ISSN 1091-6490 (Electronic) 0027-8424 (Linking). doi: 10.1073/pnas.1312310110.

A. Mongera, P. Rowghanian, H. J. Gustafson, E. Shelton, D. A. Kealhofer, E. K. Carn, F. Serwane, A. A. Lucio, J. Giammona, and O. Campas. A fluid-to-solid jamming transition underlies vertebrate body axis elongation. *Nature*, 561(7723):401–405, 2018. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/s41586-018-0479-2.

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 toTranscription 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`.

A. Mor, S. Suliman, R. Ben-Yishay, S. Yunger, Y. Brody, and Y. Shav-Tal. Dynamics of single mrnp nucleocytoplasmic transport and export through the nuclear pore in living cells. *Nat Cell Biol*, 12(6):543–52, 2010. ISSN 1476-4679 (Electronic) 1465-7392 (Linking). doi: ncb2056[pii]10.1038/ncb2056.

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.

Christian Mosimann, Charles K. Kaufman, Pulin Li, Emily K. Pugach, Owen J. Tamplin, and Leonard I. Zon. Ubiquitous transgene expression and cre-based recombination driven by the ubiquitin promoter in zebrafish. *Development*, 138:169–177, 1 2011. ISSN 0950-1991. doi: 10.1242/DEV.059345.

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.

Tadasu Nozaki, Ryosuke Imai, Mai Tanbo, Ryosuke Nagashima, Sachiko Tamura, Tomomi Tani, Yasumasa Joti, Masaru Tomita, Kayo Hibino, Masato T. Kanemaki, Kerstin S. Wendt, Yasushi Okada, Takeharu Nagai, and Kazuhiro Maeshima. Dynamic organization of chromatin domains revealed by super-resolution live-cell imaging. *Molecular cell*, 67: 282–293.e7, 7 2017. ISSN 1097-4164. doi: 10.1016/J.MOLCEL.2017.06.018. URL `https://pubmed.ncbi.nlm.nih.gov/28712725/`.

A. C. Oates, C. Mueller, and R. K. Ho. Cooperative function of deltac and her7 in anterior segment formation. *Dev Biol*, 280(1):133–49, 2005. ISSN 0012-1606 (Print) 0012-1606 (Linking). doi: 10.1016/j.ydbio.2005.01.010.

A. C. Oates, L. G. Morelli, and S. Ares. Patterning embryos with oscillations: structure, function and dynamics of the vertebrate segmentation clock. *Development*, 139(4):625–39, 2012. ISSN 1477-9129 (Electronic) 0950-1991 (Linking). doi: 10.1242/dev.063735.

Andrew C. Oates. Waiting on the fringe: cell autonomy and signaling delays in segmentation clocks. *Current Opinion in Genetics Development*, 63:61–70, 8 2020. ISSN 0959-437X. doi: 10.1016/J.GDE.2020.04.008.

Andrew C. Oates and Robert K. Ho. Hairye/(spl)-related (her) genes are central components of the segmentation oscillator and display redundancy with the delta/notch signaling pathway in the formation of anterior segmental boundaries in the zebrafish. *Development*, 129:2929–2946, 6 2002. ISSN 09501991. doi: 10.1242/dev.129.12.2929.

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.

Anil Ozdemir, Lijia Ma, Kevin P. White, and Angelike Stathopoulos. Su(h)-mediated repression positions gene boundaries along the dorsal-ventral axis of drosophila embryos. *Developmental Cell*, 31:100–113, 10 2014. ISSN 18781551. doi: 10.1016/j.devcel.2014.08.005.

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.bbagrm.2012.02.024. URL `http://dx.doi.org/10.1016/j.bbagrm.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.

H. Y. Park, H. Lim, Y. J. Yoon, A. Follenzi, C. Nwokafor, M. Lopez-Jones, X. Meng, and R. H. Singer. Visualization of dynamics of single endogenous mrna labeled in live mouse. *Science*, 343(6169):422–4, 2014. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science.1239200.

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.

Xiang-Hong Peng, Ze-Hong Cao, Jin-Tang Xia, Grant W Carlson, Melinda M Lewis, William C Wood, and Lily Yang. Real-time detection of gene expression in cancer cells using molecular beacon imaging: New strategies for cancer research. 2005. URL `www.aacrjournals.org`.

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.

Xavier Pichon, Mounia Lagha, Florian Mueller, and Edouard Bertrand. Molecular cell technology review a growing toolbox to image gene expression in single cells: Sensitive approaches for demanding challenges. *Molecular Cell*, 71:468–480, 2018. doi: 10.1016/j. molcel.2018.07.022. URL https://doi.org/10.1016/j.molcel.2018.07.022.

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).

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: 06-PLBI-RA-0625R2[pii]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.

G. T. Reeves, N. Trisnadi, T. V. Truong, M. Nahmad, S. Katz, and A. Stathopoulos. Dorsal-ventral gene expression in the drosophila embryo reflects the dynamics and precision of the dorsal nuclear gradient. *Dev Cell*, 2012. ISSN 1878-1551 (Electronic) 1534-5807 (Linking). doi: 10.1016/j.devcel.2011.12.007.

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.

Joseph Rodriguez, Gang Ren, Christopher R. Day, Keji Zhao, Carson C. Chow, and Daniel R. Larson. Intrinsic dynamics of a human gene reveal the basis of expression heterogeneity. *Cell*, 176:213–226.e18, 1 2019. ISSN 10974172. doi: 10.1016/J.CELL.2018.11.026/ATTACHMENT/621806DE-6B19-4774-9A2C-81C1CAE4118F/MMC9.PDF. URL `http://www.cell.com/article/S0092867418315186/fulltexthttp://www.cell.com/article/S0092867418315186/abstracthttps://www.cell.com/cell/abstract/S0092-8674(18)31518-6`.

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.

Robert Saint and Michael Clarkson. Pictures in cell biology: A functional marker for drosophila chromosomes in vivo. *Trends in Cell Biology*, 10:553, 12 2000. ISSN 0962-8924. doi: 10.1016/S0962-8924(00)01875-4.

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.

Philip J Santangelo and Gang Bao. Dynamics of filamentous viral rnps prior to egress. *Nucleic Acids Research*, 35:3602–3611, 2007. doi: 10.1093/nar/gkm246. URL `https://academic.oup.com/nar/article/35/11/3602/2402278`.

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`.

Yuko Sato, Masaru Nakao, and Hiroshi Kimura. Live-cell imaging probes to track chromatin modification dynamics. *Microscopy*, 70(5):415–422, 08 2021. ISSN 2050-5701. doi: 10.1093/jmicro/dfab030. URL `https://doi.org/10.1093/jmicro/dfab030`.

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.

K. N. Schulz, E. R. Bondra, A. Moshe, J. E. Villalta, J. D. Lieb, T. Kaplan, D. J. McKay, and M. M. Harrison. Zelda is differentially required for chromatin accessibility, transcription factor binding, and gene expression in the early drosophila embryo. *Genome Res*, 25(11): 1715–26, 2015. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi: 10.1101/gr.192682. 115.

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.

Yaron Shav-Tal, Xavier Darzacq, Shailesh M. Shenoy, Dahlene Fusco, Susan M. Janicki, David L. Spector, and Robert H. Singer. Dynamics of single mrnps in nuclei of living cells. *Science*, 304:1797–1800, 6 2004. ISSN 00368075. doi: 10.1126/SCIENCE.1099754/ FORMAT/PDF. URL `www.sciencemag.org`.

Jonathan Sheinberger, Hodaya Hochberg, Erez Lavi, Itamar Kanter, Shira Avivi, Gita Reinitz, Avital Schwed, Yuval Aizler, Eli Varon, Noa Kinor, and Yaron Shav-Tal. Cd-tagging-ms2: detecting allelic expression of endogenous mrnas and their protein products

in single cells. *Biology Methods and Protocols*, 2, 1 2017. ISSN 23968923. doi: 10.1093/ BIOMETHODS/BPX004. URL `https://academic.oup.com/biomethods/article/2/ 1/bpx004/3821150`.

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).

N. P. Shih, P. Francois, E. A. Delaune, and S. L. Amacher. Dynamics of the slowing segmentation clock reveal alternating two-segment periodicity. *Development*, 142(10): 1785–93, 2015. ISSN 1477-9129 (Electronic) 0950-1991 (Linking). doi: 10.1242/dev.119057.

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.

D. Soroldoni, D. J. Jorg, L. G. Morelli, D. L. Richmond, J. Schindelin, F. Julicher, and A. C. Oates. Genetic oscillations. a doppler effect in embryonic pattern formation. *Science*, 345 (6193):222–5, 2014. ISSN 1095-9203 (Electronic) 0036-8075 (Linking). doi: 10.1126/science. 1253089.

Daniele Soroldoni, Benjamin M. Hogan, and Andrew C. Oates. Simple and efficient transgenesis with meganuclease constructs in zebrafish. *Methods in molecular biology (Clifton, N.J.)*, 546:117–130, 2009. ISSN 10643745. doi: 10.1007/978-1-60327-977-2_8. URL `https://link.springer.com/protocol/10.1007/978-1-60327-977-2_8`.

Duncan B. Sparrow, Encarna Guillén-Navarro, Diane Fatkin, and Sally L. Dunwoodie. Mutation of hairy-and-enhancer-of-split-7 in humans causes spondylocostal dysostosis. *Human molecular genetics*, 17:3761–3766, 2008. ISSN 1460-2083. doi: 10.1093/HMG/ DDN272. URL `https://pubmed.ncbi.nlm.nih.gov/18775957/`.

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.

A. Stathopoulos and M. Levine. Linear signaling in the toll-dorsal pathway of drosophila: activated pelle kinase specifies all threshold outputs of gene expression while the bhlh

protein twist specifies a subset. *Development*, 129(14):3411–9, 2002. ISSN 0950-1991 (Print) 0950-1991 (Linking).

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/16525017http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1446075`.

Ruth Steward, Francis J. McNally, and Paul Schedl. Isolation of the dorsal locus of drosophila. *Nature*, 311:262–265, 1984. ISSN 00280836. doi: 10.1038/311262A0.

Rita L. Strack, Matthew D. Disney, and Samie R. Jaffrey. A superfolding spinach2 reveals the dynamic nature of trinucleotide repeat–containing rna. *Nature Methods 2013 10:12*, 10:1219–1224, 10 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2701. URL `https://www.nature.com/articles/nmeth.2701`.

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].

Y. Sun, C. Y. Nien, K. Chen, H. Y. Liu, J. Johnston, J. Zeitlinger, and C. Rushlow. Zelda overcomes the high intrinsic nucleosome barrier at enhancers during drosophila zygotic genome activation. *Genome Res*, 25(11):1703–14, 2015. ISSN 1549-5469 (Electronic) 1088-9051 (Linking). doi: 10.1101/gr.192542.115.

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.

Peter A. Takizawa and Ronald D. Vale. The myosin motor, myo4p, binds ash1 mrna via the adapter protein, she3p. *Proceedings of the National Academy of Sciences*, 97:5273–5278, 5 2000. ISSN 0027-8424. doi: 10.1073/PNAS.080585897. URL `https://www.pnas.org/content/97/10/5273https://www.pnas.org/content/97/10/5273.abstract`.

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. Capozi, 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.

A. Tsai, A. K. Muthusamy, M. R. Alves, L. D. Lavis, R. H. Singer, D. L. Stern, and J. Crocker. Nuclear microenvironments modulate transcription from low-affinity enhancers. *Elife*, 6: e28975, 2017. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.28975.

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.

Edward Tunnacliffe, Adam M. Corrigan, and Jonathan R. Chubb. Promoter-mediated diversification of transcriptional bursting dynamics following gene duplication. *Proceedings of the National Academy of Sciences of the United States of America*, 115:8364–8369, 8 2018. ISSN 10916490. doi: 10.1073/PNAS.1800943115/-/DCSUPPLEMENTAL. URL https://www.pnas.org/content/115/33/8364https://www.pnas.org/content/115/33/8364.abstract.

Benita Turner-Bridger, Maximillian Jakobs, Leila Muresan, Hovy Ho Wai Wong, Kristian Franze, William A. Harris, and Christine E. Holt. Single-molecule analysis of endogenous -actin mrna trafficking reveals a mechanism for compartmentalized mrna localization in axons. *Proceedings of the National Academy of Sciences of the United States of America*, 115:E9697–E9706, 10 2018. ISSN 10916490. doi: 10.1073/PNAS.1806189115/-/DCSUPPLEMENTAL. URL https://www.pnas.org/content/115/41/E9697https://www.pnas.org/content/115/41/E9697.abstract.

E. Tutucci, M. Vera, J. Biswas, J. Garcia, R. Parker, and R. H. Singer. An improved ms2 system for accurate reporting of the mrna life cycle. *Nat Methods*, 15(1):81–89, 2018a. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi: 10.1038/nmeth.4502.

Evelina Tutucci, Nathan M. Livingston, Robert H. Singer, and Bin Wu. Imaging mRNA In Vivo, from Birth to Death. *Annual Review of Biophysics*, 47(1):85–106, 2018b. ISSN 1936-122X. doi: 10.1146/annurev-biophys-070317-033037.

Sanjay Tyagi and Osama Alsmadi. Imaging native -actin mrna in motile fibroblasts. *Biophysical Journal*, 87:4153–4162, 12 2004. ISSN 0006-3495. doi: 10.1529/BIOPHYSJ.104.045153.

Sanjay Tyagi and Fred Russell Kramer. Molecular beacons: Probes that fluoresce upon hybridization. *Nature Biotechnology 1996 14:3*, 14:303–308, 1996. ISSN 1546-1696. doi: 10.1038/nbt0396-303. URL https://www.nature.com/articles/nbt0396-303.

Diana Y Vargas, Arjun Raj, Salvatore A E Marras, Fred Russell Kramer, and Sanjay Tyagi. Mechanism of mrna transport in the nucleus. 2005. URL www.pnas.orgcgidoi10.1073pnas.0505580102.

O. F. Venzin and A. C. Oates. What are you synching about? emerging complexity of notch signaling in the segmentation clock. *Dev Biol*, 460(1):40–54, 2020. ISSN 1095-564X (Electronic) 0012-1606 (Linking). doi: 10.1016/j.ydbio.2019.06.024.

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.

Wei Wang, Zong-Qiang Cui, Han Han, Zhi-Ping Zhang, Hong-Ping Wei, Ya-Feng Zhou, Ze Chen, and Xian-En Zhang. Imaging and characterizing influenza a virus mrna transport in living cells. *Nucleic Acids Research*, 36:4913–4928, 2008. doi: 10.1093/nar/gkn475. URL https://academic.oup.com/nar/article/36/15/4913/1075956.

A. B. Webb, I. M. Lengyel, D. J. Jorg, G. Valentin, F. Julicher, L. G. Morelli, and A. C. Oates. Persistence, period and precision of autonomous cellular oscillators from the zebrafish segmentation clock. *Elife*, 5, 2016. ISSN 2050-084X (Electronic) 2050-084X (Linking). doi: 10.7554/eLife.08438.

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, V. Miskolci, H. Sato, E. Tutucci, C. A. Kenworthy, S. K. Donnelly, Y. J. Yoon, D. Cox, R. H. Singer, and L. Hodgson. Synonymous modification results in high-fidelity gene expression of repetitive protein and nucleotide sequences. *Genes Dev*, 29(8):876–86, 2015. ISSN 1549-5477 (Electronic) 0890-9369 (Linking). doi: 10.1101/gad.259358.115.

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.

Wanqing Xiang, M. Julia Roberti, Jean Karim Hériché, Sébastien Huet, Stephanie Alexander, and Jan Ellenberg. Correlative live and super-resolution imaging reveals the dynamic structure of replication domains. *The Journal of Cell Biology*, 217:1973, 6 2018. ISSN 15408140. doi: 10.1083/JCB. 201709074. URL `/pmc/articles/PMC5987722//pmc/articles/PMC5987722/?report= abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5987722/`.

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.

J. Yang, T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso, J. M. Cunningham, M. de Andrade, B. Feenstra, E. Feingold, M. G. Hayes, W. G. Hill, M. T. Landi, A. Alonso, G. Lettre, P. Lin, H. Ling, W. Lowe, R. A. Mathias, M. Melbye, E. Pugh, M. C. Cornelis, B. S. Weir, M. E. Goddard, and P. M. Visscher. Genome partitioning of genetic variation for complex traits using common snps. *Nat Genet*, 43(6):519–25, 2011. ISSN 1546-1718 (Electronic) 1061-4036 (Linking). doi: 10.1038/ng.823.

Hsiao-Yun Yeh, Marylynn V Yates, Ashok Mulchandani, and Wilfred Chen. Visualizing the dynamics of viral replication in living cells via tat peptide delivery of nuclease-resistant molecular beacons. 2008. URL `www.pnas.org/cgi/content/full/`.

Katie M. Yergert, Rebecca O'Rouke, Jacob H. Hines, and Bruce Appel. Identification of 3' utr motifs required for mrna localization to myelin sheaths in vivo. *bioRxiv*, page 654616, 8 2020. ISSN 2692-8205. doi: 10.1101/654616. URL `https://www.biorxiv.org/content/10.1101/654616v2https://www.biorxiv.org/content/10.1101/654616v2.abstract`.

Young J. Yoon, Bin Wu, Adina R. Buxbaum, Sulagna Das, Albert Tsai, Brian P. English, Jonathan B. Grimm, Luke D. Lavis, and Robert H. Singer. Glutamate-induced rna localization and translation in neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 113:E6877–E6886, 11 2016. ISSN 10916490. doi: 10.1073/PNAS.1614267113/-/DCSUPPLEMENTAL. URL `https://www.pnas.org/content/113/44/E6877https://www.pnas.org/content/113/44/E6877.abstract`.

S. Yunger, L. Rosenfeld, Y. Garini, and Y. Shav-Tal. Single-allele analysis of transcription kinetics in living mammalian cells. *Nat Methods*, 7(8):631–3, 2010. ISSN 1548-7105 (Electronic) 1548-7091 (Linking). doi: nmeth.1482[pii]10.1038/nmeth.1482.

Kenneth S Zaret and Jason S Carroll. Pioneer transcription factors : establishing competence for gene expression Parameters affecting transcription factor access to target sites in chromatin Initiating events in chromatin : pioneer factors bind first. *Genes and Development*, pages 2227–2241, 2011. ISSN 08909369. doi: 10.1101/gad.176826.111.GENES.

C. Zechner, M. Unger, S. Pelet, M. Peter, and H. Koeppl. 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.

Lin Zhang, Luxi Chen, Jing Chen, Weimin Shen, and Anming Meng. Mini-iii rnase-based dual-color system for in vivo mrna tracking. *Development (Cambridge)*, 147, 11 2020. ISSN 14779129. doi: 10.1242/DEV.190728/VIDEO-6.

O. Q. H. Zinani, K. Keseroglu, A. Ay, and E. M. Ozbudak. Pairing of segmentation clock genes drives robust pattern formation. *Nature*, 2020. ISSN 1476-4687 (Electronic) 0028-0836 (Linking). doi: 10.1038/s41586-020-03055-0.

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.

Appendix A

Supplementary Information for Chapter 2

A.1 Dorsal enrichment
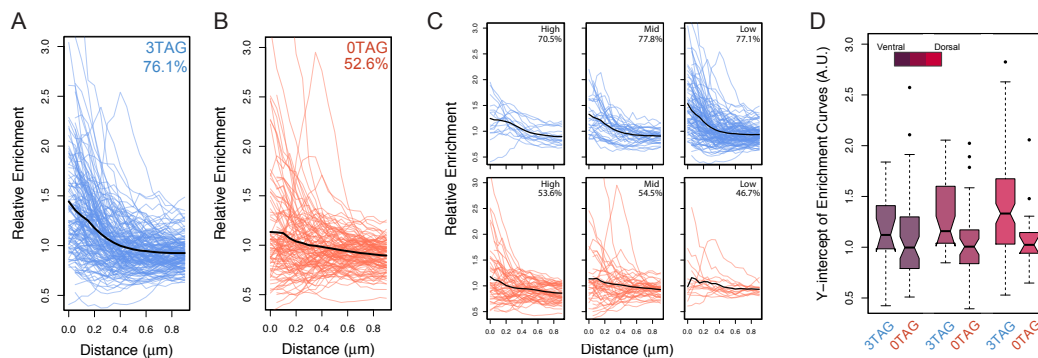
Figure A.1: Dl enrichment at the 3TAG enhancer increases across the D/V axis. Related to Figure 2.4. See caption on next page. See caption on next page.

Figure A.1: Figure S1. Dl enrichment at the 3TAG enhancer increases across the D/V axis. Related to Figure 4. See caption on next page. (A-B) Individual enrichment curves plotted together for each genotype, as indicated. Average enrichment profiles of all curves are plotted in black. Percentages show the fraction of curves that have a y-intercept greater than 1, indicating the proportion of nuclei that show net enrichment. (C) Individual enrichment curves from each bin plotted in the same manner; 3TAG on top (blue), 0TAG on bottom (red). Note that the y-intercept of the 3TAG foci increases as the nuclear concentration of Dl falls, which explains how we can observe uniform transcriptional output across the gradient in 3TAG, i.e., Dl is more enriched in regions where there is less nuclear Dl, thereby maintaining uniform output. However, the median value of 0TAG foci stays relatively flat at 1.0 across the gradient as expected since there is no enrichment without Zld. Note also that the percentages of enriched lines, which are the lines with a y-intercept greater than 1 (indicated in the upper right corner of each panel) do not appreciably change over the Dl gradient for either genotype, therefore the effect of enrichment is restricted to the amplitude of enrichment rather than the percentage of cells that are enriched. (D) Boxplots showing the distribution of y-intercepts from each spatial bin.

A.2 Supplementary videos

A.2.1. **Video S1.** Time-Lapse Video of *sog* 3TAG1, Related to **Figures 2 and 3**. Time lapse video of *sog* 3TAG1 embryo in NC10-NC14. Embryos were collected from females carrying MCP-GFP (green) and H2Av-RFP (red) mated to males homozygous for the MS2 transgene reporter lines and prepared for live imaging (see Section 2.4) on a Leica SP8 with a 63X objective lens and the following settings: optical sections: 512x512 pixels, 30 z stacks 0.69 μm apart, 12bit; zoom: 1.7; time resolution: 40 s per frame (see Section 2.4). Scale bar = 10 μm. *(https://www.dropbox.com/s/4hdejomdouyodsq/Video%20S1.mp4?dl=0)*

A.2.2. **Video S2.** Time-Lapse Video of *sog* 0TAG1, Related to **Figures 2 and 3**. Time lapse videos of *sog* 0TAG1 embryo in NC10-NC14. Embryos were collected from females carrying MCP-GFP (green) and H2Av-RFP (red) mated to males homozygous for the MS2 transgene reporter lines and prepared for live imaging (see Section 2.4) on a Leica SP8 with a 63X objective lens and the following settings: optical sections: 512x512 pixels, 30 z stacks 0.69 μm apart, 12bit; zoom: 1.7; time resolution: 40 s per frame (see Section 2.4). Scale bar = 10μm. *(https://www.dropbox.com/s/9051014umxito4o/Video%20S2.mp4?dl=0)*

A.2.3. **Video S3.** Time-Lapse Video of *sog* 3TAG2, Related to Figures 2 and 3. Time lapse videos of *sog* 3TAG2 embryo in NC10-NC14. Embryos were collected from females carrying MCP-GFP (green) and H2Av-RFP (red) mated to males homozygous for the MS2 transgene reporter lines and prepared for live imaging (see Section 2.4) on a Leica SP8 with a 63X objective lens and the following settings: optical sections: 512x512 pixels, 30 z stacks 0.69 μm apart, 12bit; zoom: 1.7; time resolution: 40 s per frame (see Section 2.4). Scale bar = 10μm. *(https://www.dropbox.com/s/srivz2qb7h9aj2n/Video%20S3.mp4?dl=0)*

A.2.4. **Video S4.** Time-Lapse Video of *sog* 0TAG2, Related to Figures 2 and 3. Time lapse videos of *sog* 0TAG2 embryo in NC10-NC14. Embryos were collected from females carrying MCP-GFP (green) and H2Av-RFP (red) mated to males homozygous for the MS2 transgene reporter lines and prepared for live imaging (see Section 2.4) on a Leica SP8 with a 63X objective lens and the following settings: optical sections: 512x512 pixels, 30 z stacks 0.69 μm apart, 12bit; zoom: 1.7; time resolution: 40 s per frame (see Section 2.4). Scale bar = 10μm. *(https://www.dropbox.com/s/3nb1e406fkrtvpm/Video%20S4.mp4?dl=0)*

Video S5.

A.2.5. **Video S5.** Time-Lapse Video of *sog* 3TAG3, Related to Figures 2 and 3. Time lapse videos of *sog* 3TAG3 embryo in NC10-NC14. Embryos were collected from females carrying MCP-GFP (green) and H2Av-RFP (red) mated to males homozygous for the MS2 transgene reporter lines and prepared for live imaging (see

Section 2.4) on a Leica SP8 with a 63X objective lens and the following settings: optical sections: 512x512 pixels, 30 z stacks 0.69 μm apart, 12bit; zoom: 1.7; time resolution: 40 s per frame (see Section 2.4). Scale bar = 10μm. *(https://www.dropbox.com/s/enh6ps2vay6sl7k/Video%20S5.mp4?dl=0)*

A.2.6. **Video S6.** Time-Lapse Video of *sog* 0TAG3, Related to Figures 2 and 3. Time lapse videos of *sog* 0TAG3 embryo in NC10-NC14. Embryos were collected from females carrying MCP-GFP (green) and H2Av-RFP (red) mated to males homozygous for the MS2 transgene reporter lines and prepared for live imaging (see Section 2.4) on a Leica SP8 with a 63X objective lens and the following settings: optical sections: 512x512 pixels, 30 z stacks 0.69 μm apart, 12bit; zoom: 1.7; time resolution: 40 s per frame (see Section 2.4). Scale bar = 10μm. *(https://www.dropbox.com/s/k3s66kqhrleh79s/Video%20S6.mp4?dl=0)*

Appendix B

Supplementary Information for Chapter 3

B.1 Equilibrium models of transcription

B.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 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. B.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

$$p_{bound} = \frac{\frac{[P]}{K_p} + \omega \frac{[P]}{K_p}\frac{[A]}{K_a}}{1 + \frac{[P]}{K_p} + \frac{[A]}{K_a} + \omega \frac{[P]}{K_p}\frac{[A]}{K_a}}. \tag{B.1}$$

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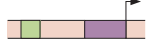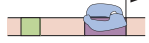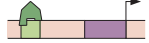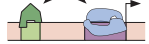
STATE	WEIGHT	RATE
	1	0
	$\dfrac{[P]}{K_p}$	R
	$\dfrac{[A]}{K_a}$	0
	$\omega \dfrac{[P]}{K_p} \dfrac{[A]}{K_a}$	R

Figure B.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

$$\frac{\mathrm{dmRNA}}{\mathrm{d}t} = R\, p_{bound},\tag{B.2}$$

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. B.1, the overall transcriptional initiation rate is then given by

$$\frac{\mathrm{dmRNA}}{\mathrm{d}t} = R\frac{\frac{[P]}{K_p} + \omega\frac{[P]}{K_p}\frac{[A]}{K_a}}{1 + \frac{[P]}{K_p} + \frac{[A]}{K_a} + \omega\frac{[P]}{K_p}\frac{[A]}{K_a}}.\tag{B.3}$$

From Eq. B.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

$$p_{bound} = \frac{\omega\frac{[P]}{K_p}\frac{[A]}{K_a}}{1 + \omega\frac{[P]}{K_p}\frac{[A]}{K_a}}.\tag{B.4}$$

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

$$p_{bound} = \frac{\frac{[A]}{K'_a}}{1 + \frac{[A]}{K'_a}}\tag{B.5}$$

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 B.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.

B.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

$$\Delta\varepsilon_{\text{chrom}} = \varepsilon_{\text{accessible}} - \varepsilon_{\text{inaccessible}}. \tag{B.6}$$

Here, $\varepsilon_{\text{accessible}}$ and $\varepsilon_{\text{accessible}}$ 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 $e^{\Delta\varepsilon}$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, $b \equiv \frac{[Bicoid]}{K_b}$. Fig. B.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

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

Figure B.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. B.7 can be expressed more compactly as

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

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

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

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

$$\frac{d\mathrm{mRNA}}{dt} = R\frac{1}{Z}\left(\left(1+z\right)^{10}p\left(1+\sum_{i=1}^{6}\binom{6}{i}b^i\omega_b^{i-1}\omega_{bp}^{i}\right)\right). \tag{B.10}$$

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

B.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. 3.4C, 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

$$p_{bound}(b \to \infty) \approx 1. \tag{B.11}$$

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

$$\frac{d\mathrm{mRNA}}{dt} = R_{max} \approx R, \tag{B.12}$$

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. 3.4C, blue points), which we can connect with the model's theoretical minimum rate. Specifically, in this limit, b approaches zero in Eq. B.10 such that all Bicoid-dependent terms drop out, resulting in

$$\frac{d\text{mRNA}}{dt} = R_{min} \approx \frac{1}{Z}\left(\left(1+z\right)^{10} p\right) R_{max}, \tag{B.13}$$

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

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

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 B.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 B.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. B.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 B.2.2).

For subsequent parameter exploration of this model (Section B.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. 3.4D).

B.2 Input-output measurements, predictions, and characterization

B.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. B.3A). Further, the nuclear fluorescence of eGFP-Bicoid at 8 min into nuclear cycle 13 (Fig. B.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. B.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. B.3D. Unlike the eGFP-Bicoid profile, the Zelda-sfGFP nuclear fluorescence was approximately uniform across embryo position (Fig. B.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. B.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. B.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. B.6.

B.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 3. 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. B.4; see Sections B.1.2, B.6.1, B.7.1, and B.8.1 for model details). Each RNAP molecule traverses the gene at a constant velocity v of 1.54 kb/min, as measured
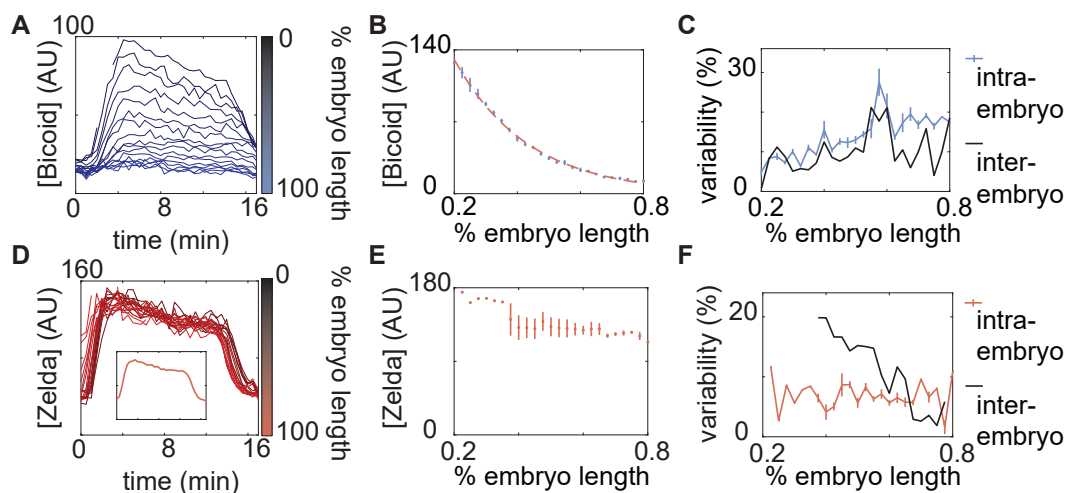
Figure B.3: Measurements of input transcription-factor concentration dynamics. (A) Nuclear eGFP-Bicoid concentration as a function of time into nuclear cycle 13 across various positions along the anterior-posterior axis of a single embryo. (B) eGFP-Bicoid concentration at 8 min into nuclear cycle 13 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% ± 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

$$t_{elon} = \frac{L}{v}. \tag{B.15}$$

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

$$F(t) = \alpha \int_0^t \Big(R(t') - R(t' - t_{elon}) \Big) dt', \tag{B.16}$$

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. B.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. B.4, ii). For this reporter construct, the length of the stem loops was approximately $L_{loops} = 1.28$ kb. RNAP molecules that had finished transcribing the MS2 stem loops contributed the full amount of fluorescence (Fig. B.4, iii). Finally, to make this simulation compatible with the trapezoidal fitting scheme in Section B.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 B.2.3).

B.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).
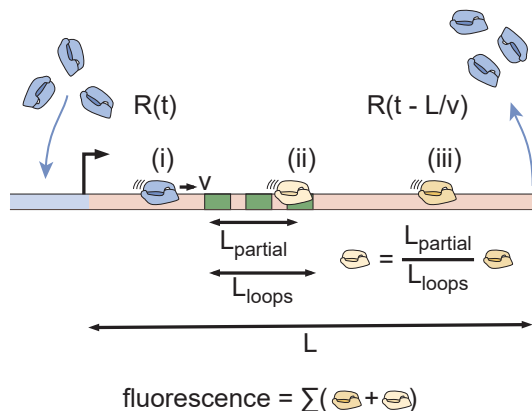
Figure B.4: MS2 fluorescence calculation protocol. RNAP molecules load onto the reporter gene at a time-dependent rate $R(t)$, after which they elongate at a constant velocity v. Upon reaching the end of the gene after a length L has been transcribed, they are assumed to terminate and disappear instantly, 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. B.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. B.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. B.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.

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

As described in Section 3.2.3 of the main text, a mitotic repression window was incorporated into the thermodynamic MWC model (Section B.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. B.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 B.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. B.6A, purple line), it fails to predict the t_{on} throughout the embryo (Fig. B.6B, purple line; see Sections B.2.2 and B.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. B.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. B.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. B.6D). As a result, the predicted transcriptional dynamics in Fig. B.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. 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
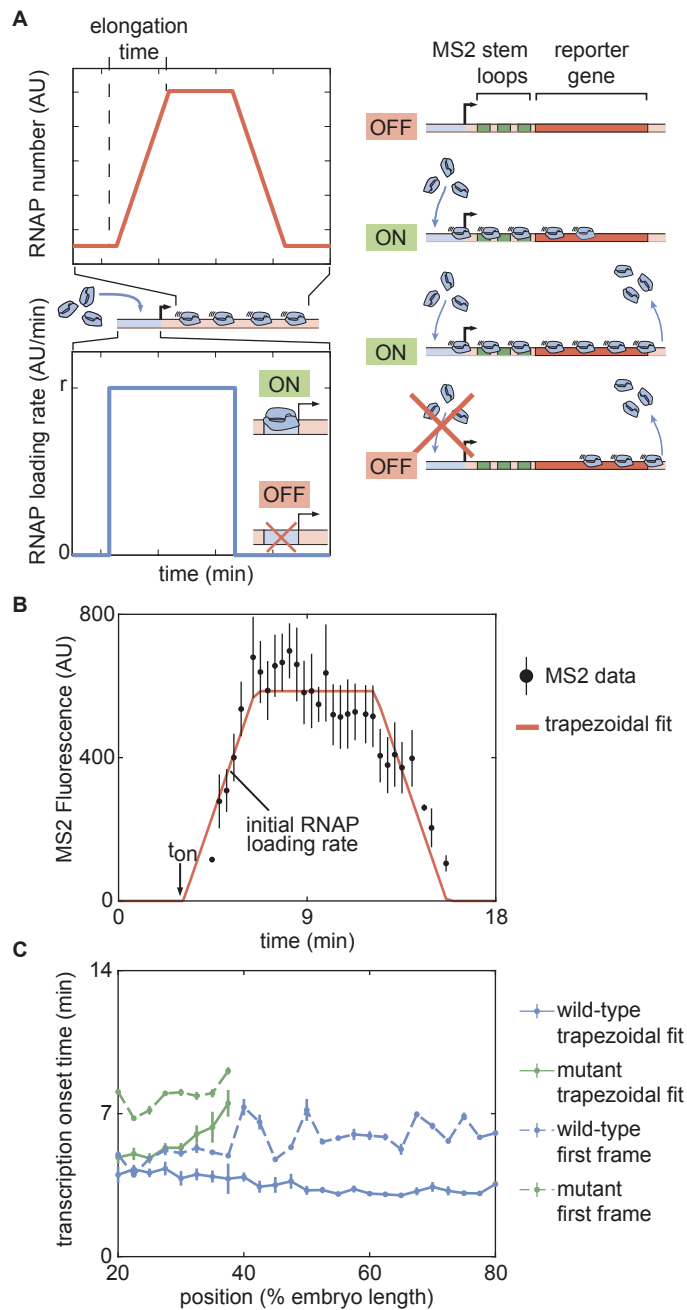
Figure B.5: Outline of fitting to the trapezoidal model of transcription. See caption on next page.

Figure B.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).

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 B.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. B.6C and D, grey shaded region), the model successfully recapitulates t_{on} throughout the embryo (Fig. B.6B and D, blue curves), while still explaining the observed rates of RNAP loading (Fig. B.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.

B.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 50% of total Bicoid being labeled with eGFP. As shown in Fig. B.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
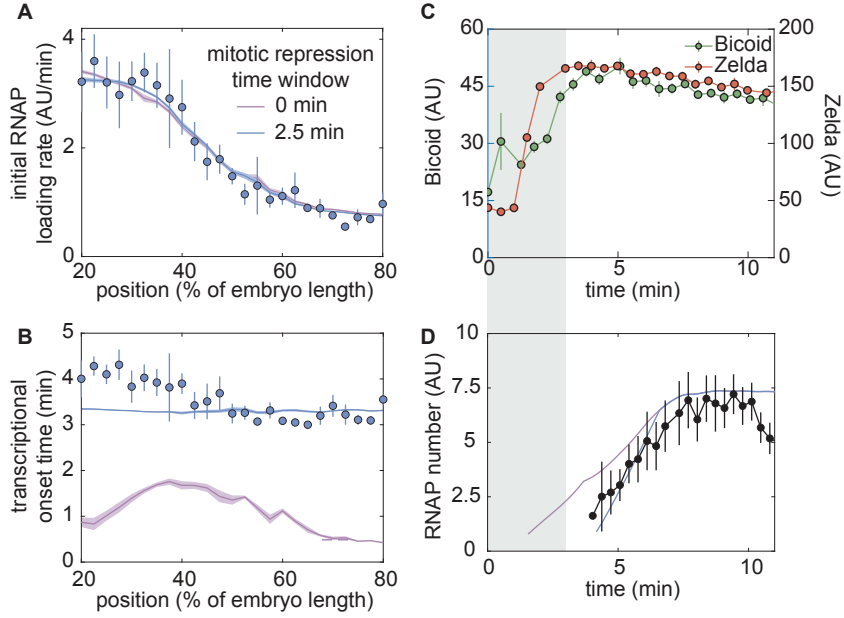
Figure B.6: A thermodynamic MWC model including mitotic repression can recapitulate *hunchback* regulation by Bicoid and Zelda. (A) Measured initial rates of RNAP loading and (B) t_{on} (blue points) across the length of the embryo, compared to fits to the thermodynamic MWC model with and without accounting 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 $\Delta\epsilon_{chrom} = 10\ k_BT$, $K_b = 34\ AU$, $K_z = 500\ AU$, with different arbitrary fluorescent units for Bicoid and Zelda, $\omega_b = 10$, $\omega_{bp} = 0.4$, for a model assuming six Bicoid binding sites.)

(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. B.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. B.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. B.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. B.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. B.7A, lines).

Fig. B.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 B.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.
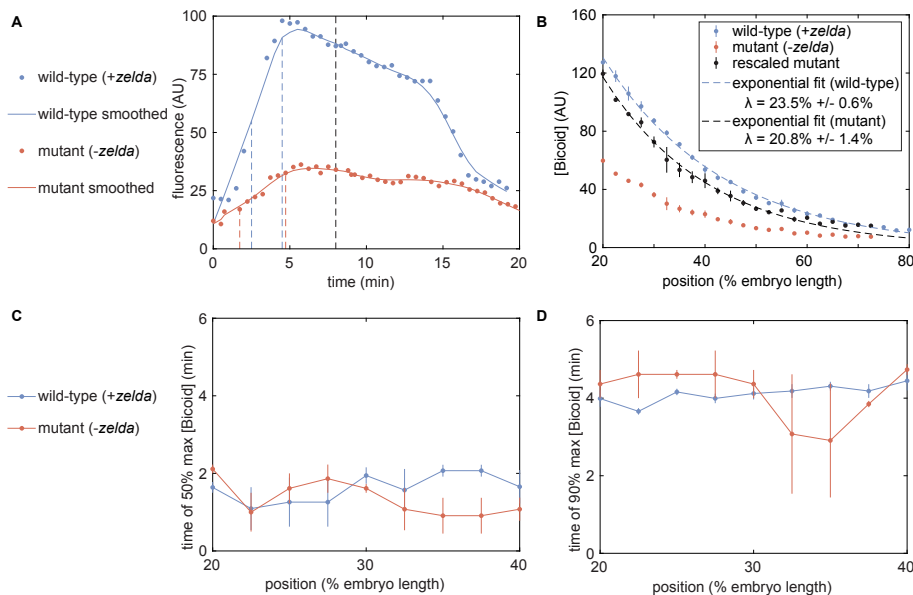
Figure B.7: Comparison of eGFP-Bicoid measurements in wild-type and *zelda⁻* embryos. (A) Example mean nuclear eGFP-Bicoid concentrations for a single embryo at 30% along the embryo's length, for wild-type (blue) and *zelda⁻* (red) backgrounds. Datapoints are raw values and lines are smoothed results. 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.

B.5 State-space exploration of theoretical models

B.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. B.8A and A.1A), as defined in Equation 3.5. The offset in t_{on} at 20% embryo length (Eq. 3.4) 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. B.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. B.8A and B show an example of how the three parameters are calculated using the $zelda^-$ background data presented in Fig. 3.4C 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. B.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. B.8C, i). The state space was sectioned into 100 slices along each orthogonal axis (Fig. B.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. B.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. B.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. B.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. A.1A). 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. 3.4C). 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 B.1.2, B.6.1, B.7.1, and B.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. B.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.

B.5.2 State space exploration with the thermodynamic MWC model

Fig. B.9A and Video B.9.7 show the resulting three-dimensional state space for the thermo-dynamic 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. B.9A and Video B.9.7, the state space corresponding to the thermo-dynamic 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. B.9A and Video B.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. B.9B. Even in this sim-plified representation, the failure of the thermodynamic MWC model (Fig. B.9B, green) is evident. Therefore, we utilized this representation throughout Chapter 3 and Appendix B (Figs. A.1C, 3.6B, and 3.7D, and Figs. B.13 and B.14C).

B.6 Failures and assumptions of thermodynamic models of transcription

B.6.1 Generalized thermodynamic model

The generalized thermodynamic model is an extension of the thermodynamic MWC model presented in Section B.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
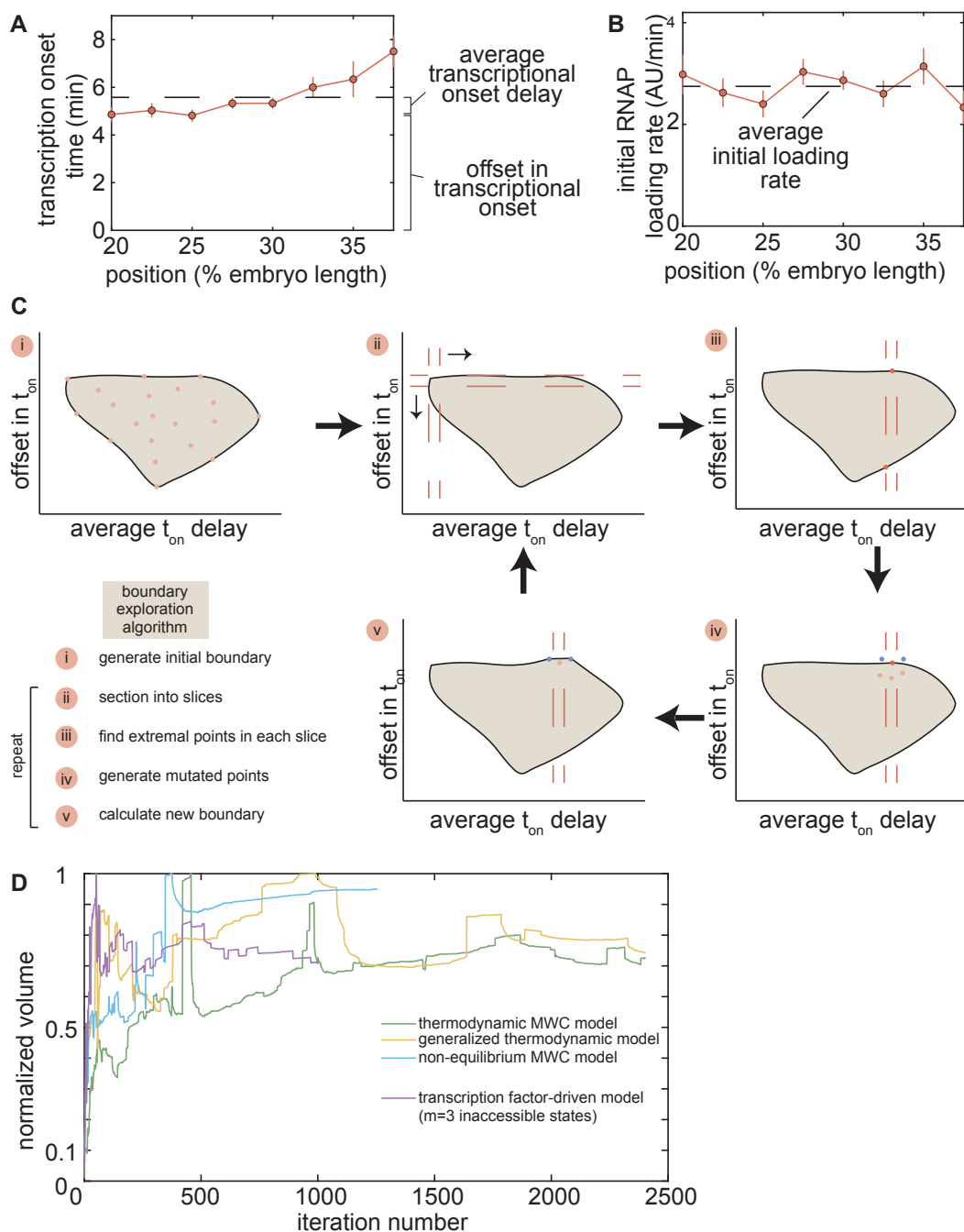
Figure B.8: Description of state-space metrics and boundary-exploration algorithm. See caption on next page.

Figure B.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. 3.4D. (B) Average initial RNAP loading rate for the *zelda$^-$* background data in Fig. 3.4C. (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. B.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

$$Z = p_{inacc} + \sum_{r=0}^{1} \sum_{n=0}^{12} P_{r,n}[Bicoid]^n, \tag{B.17}$$

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

$$\frac{d\text{mRNA}}{dt} = \frac{1}{Z}\Big(\sum_{n=0}^{12} P_{1,n}R[Bicoid]^n \Big), \tag{B.18}$$

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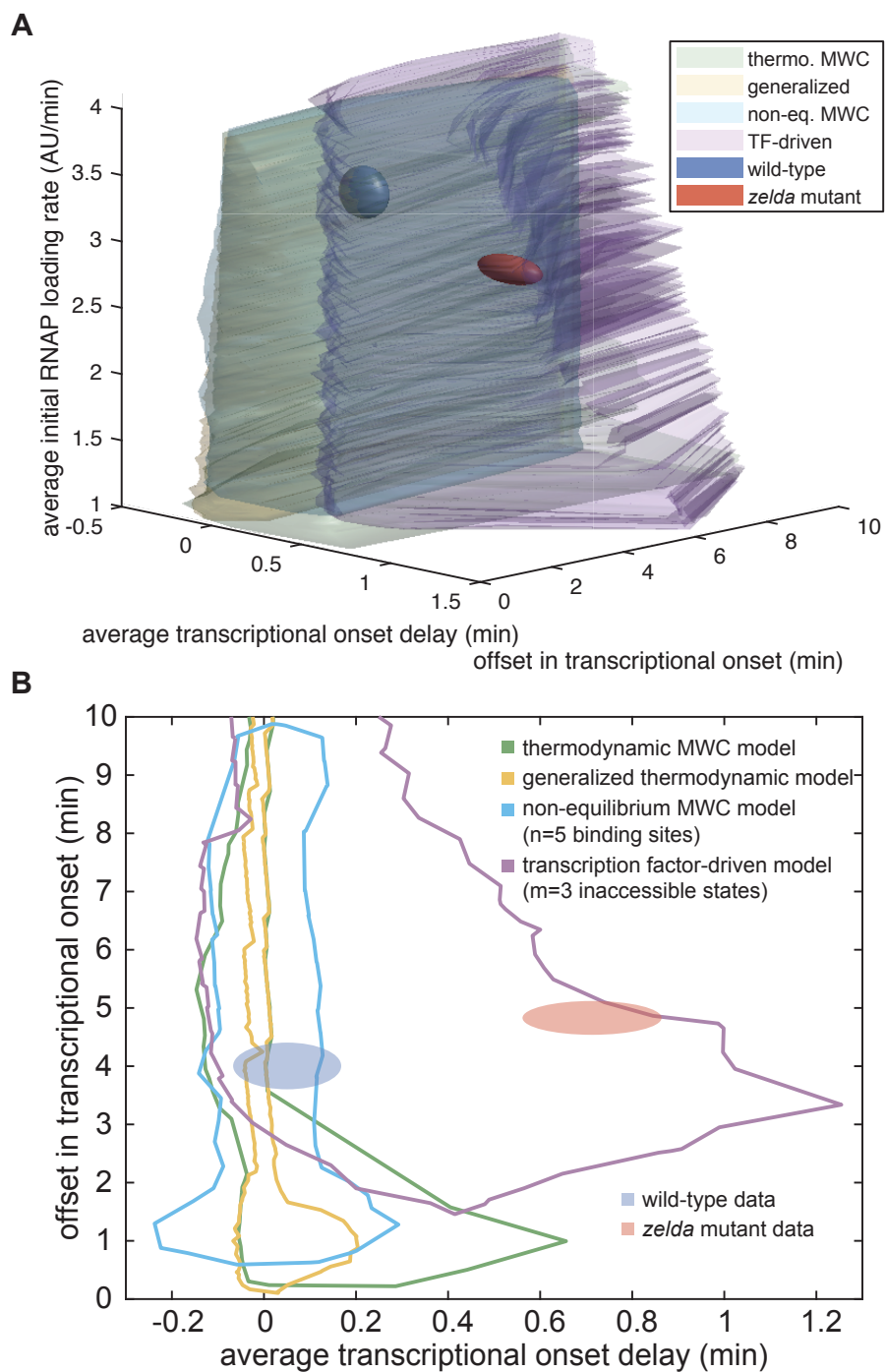
Figure B.9: Exploration of state space. See caption on next page.

Figure B.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 B.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 B.1.2, B.7.1, and B.8.1 (see Section B.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.

B.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 B.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_ds and ωs, 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

$$P_{1,1} = bp\omega_{bp} \tag{B.19}$$

which has bounds

$$0 < P_{1,1} < (1000)^2(1000) = 10^9 \tag{B.20}$$

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 B.1.3) Nevertheless, this model still fails to capture the delays observed in the $zelda^-$ data (Fig. B.9B, yellow).

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

The generalized thermodynamic model (Section B.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 3.2.4 and Fig. A.1C, 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.

B.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 B.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. B.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. B.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. B.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. 3.4D, 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.

B.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. B.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.
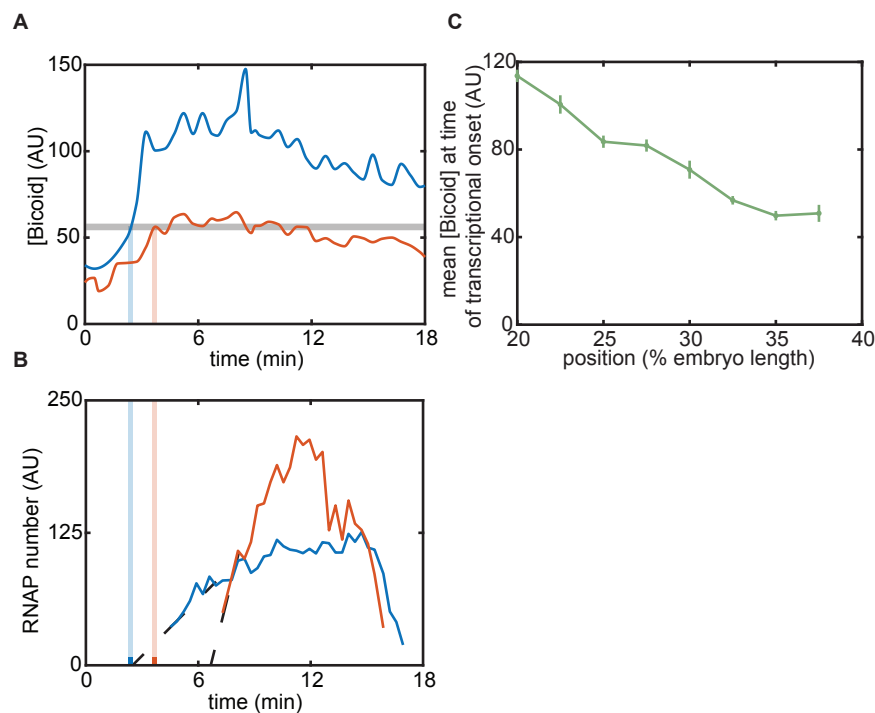
Figure B.10: Intuition for failure of equilibrium models. (A) Mean Bicoid concentrations for two positions along the embryo (blue, red), with a "threshold" chosen to to attempt to match the corresponding t_{on} in (B). (B) MS2 fluorescence signal for the two positions shown in (A) for the $zelda^-$ experiment. 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. 3.2A, 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. B.11i). This *occupancy hypothesis* is necessary for Eq. B.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. B.11ii). Third, these time scales must also be much faster than the rate of transcriptional initiation and mRNA production (Fig. B.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. B.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
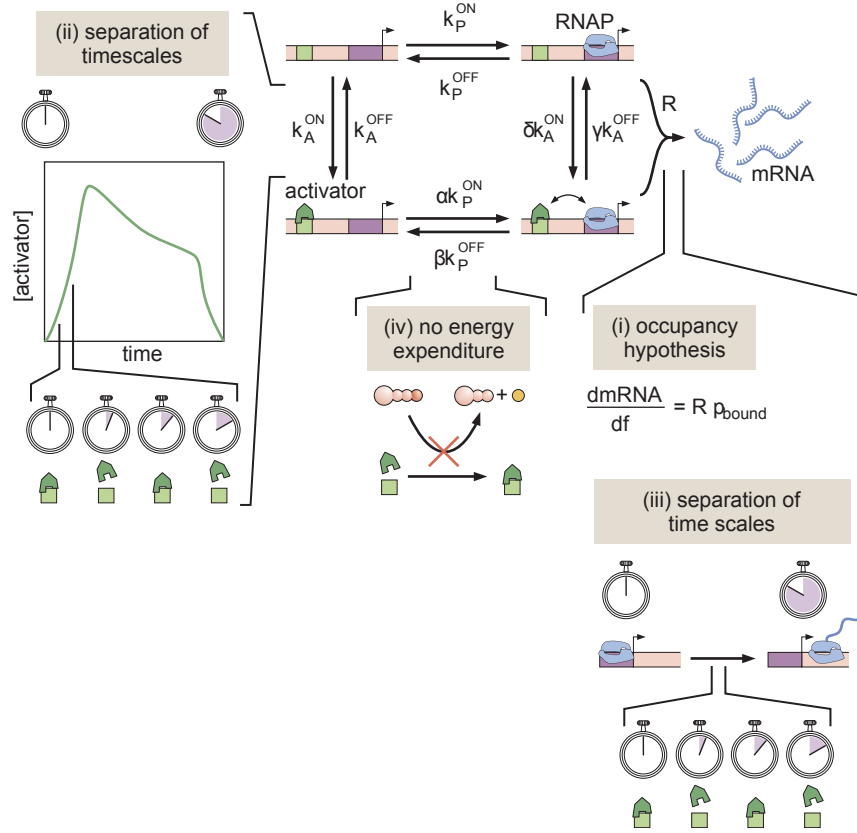
Figure B.11: A simple kinetic model of transcriptional activation in which activator molecules influence RNAP binding kinetics. The assumptions that make it possible to turn this kinetic model into a thermodynamic one are (i) the occupancy hypothesis, (ii, iii) a separation of time scales between binding 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. B.11, this requirement implies that

$$k_P^{ON} \delta k_A^{ON} \beta k_P^{OFF} k_A^{OFF} = k_P^{OFF} k_A^{ON} \alpha k_P^{ON} \gamma k_A^{OFF}. \tag{B.21}$$

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.

B.7 Non-equilibrium MWC model

B.7.1 Non-equilibrium MWC model

The non-equilibrium MWC model is an extension of the thermodynamic MWC model presented in Section B.1.2, where we now relax the assumption of separation of time scales (Fig. B.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. B.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

$$\frac{d\vec{X}}{dt} = K\vec{X}, \tag{B.22}$$

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

$$i = \text{bin}(0110000) = 48. \tag{B.23}$$

Here, 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. B.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 B.1.2 and B.6.1 but again we do not constrain the model using R_{min}, just as in Section B.6.1.

Fig. B.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

$$C = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \tag{B.24}$$

and the corresponding transition rate matrix K is

$$K = \begin{bmatrix} k_{00} & k_{01} & k_{02} & 0 \\ k_{10} & k_{11} & 0 & k_{13} \\ k_{20} & 0 & k_{22} & k_{23} \\ 0 & k_{31} & k_{32} & k_{33} \end{bmatrix}, \tag{B.25}$$

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

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

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

$$\frac{d\text{mRNA}}{dt} = Rp_{bound} = R\frac{x_1 + x_3}{x_0 + x_1 + x_2 + x_3}. \tag{B.27}$$

This model can produce time-dependent behavior not found in the thermodynamic models. Fig. B.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. B.12B, top). In contrast, for a suitable set of parameters, the non-equilibrium MWC model predicts a slow response over time (Fig. B.12B, bottom).

To produce a simulated MS2 fluorescence trace, the resulting rate of mRNA production is integrated over time using the same procedure (Section B.2.2) as the models presented in Sections B.1.2, B.6.1, and B.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 B.7.3.

B.7.2 Non-equilibrium MWC model state space exploration

In the parameter exploration of this model (Section B.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
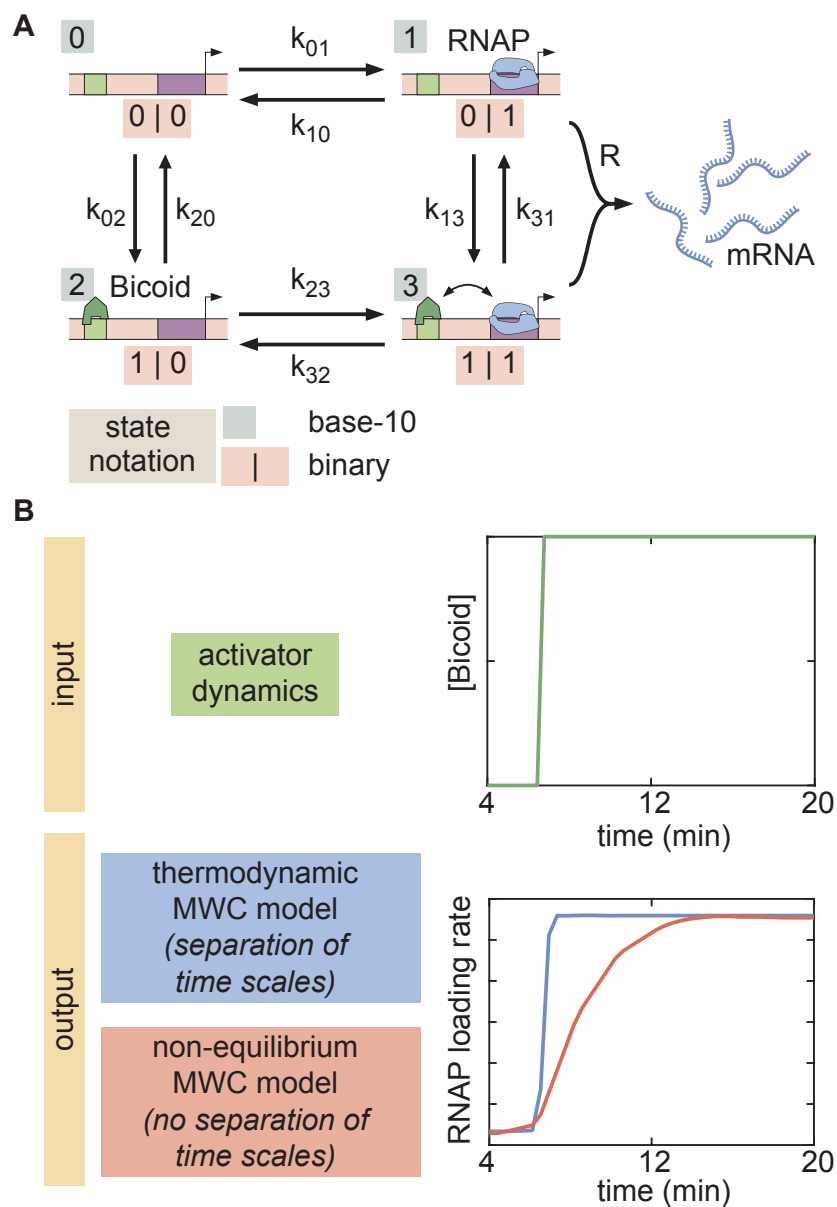
Figure B.12: Example of a four-state time-dependent model with one Bicoid binding site and no closed chromatin state. (A) The binary label for each state (light pink) can be converted into a base-10 label for each state (light teal). The transition rates k_{ij} are defined as the transition rate from 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

$$k_{on} = 4\pi D a c_o, \tag{B.28}$$

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 $\mu m^2 s^{-1}$ (Gregor et al., 2007b). Finally, Bicoid is 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. B.28 yields estimates for the maximum and minimum on-rates:

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

and

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

Thus, our maximum and minimum transition rate bounds of $k_{min} = 1~min^{-1}$ and $k_{max} = 10^5~min^{-1}$ 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 B.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 B.1.2 and B.6.1). However, this model, like those, failed to reproduce the delays observed in the *zelda*$^-$ data (Fig. B.9B, cyan).

Interestingly, the total areas covered by each non-equilibrium MWC model did not monotonically increase with Bicoid binding site number (Fig. 3.6B). 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.
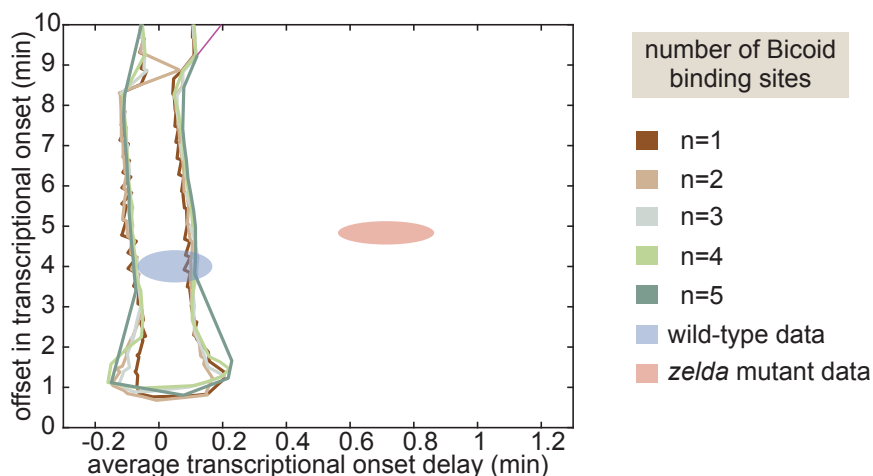
Figure B.13: State space exploration for non-equilibrium MWC model with strong mitotic repression for up to five Bicoid binding sites.

B.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. B.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 B.8.1) does provide such a mechanism by coupling the inaccessible-to-accessible transition to the position-dependent Bicoid gradient.

B.8 Transcription factor-driven model of chromatin accessibility

B.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 B.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 B.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. B.15A, see Section B.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

$$\pi_b = c_b[Bicoid] \tag{B.29}$$

and

$$\pi_z = c_z[Zelda]. \tag{B.30}$$

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

$$\pi = \pi_b + \pi_z = c_b[Bicoid] + c_z[Zelda]. \tag{B.31}$$

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 $\vec{P}$

$$\vec{P} = \begin{bmatrix} p_0 \\ p_1 \\ ... \\ p_m \end{bmatrix}. \tag{B.32}$$

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:

$$\frac{\text{dmRNA}}{\text{d}t} = R\, p_{bound}\, p_m, \tag{B.33}$$

where R is the same maximum rate used in Section B.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 $\vec{P}$

$$\frac{d\vec{P}}{dt} = \Pi \vec{P}, \tag{B.34}$$

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

$$\vec{P} = \begin{bmatrix} 1 \\ 0 \\ \dots \\ 0 \end{bmatrix} \tag{B.35}$$

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

$$\Pi = \begin{bmatrix} -\pi & 0 & 0 & 0 \\ \pi & -\pi & 0 & 0 \\ 0 & \pi & -\pi & 0 \\ 0 & 0 & \pi & 0 \end{bmatrix}, \tag{B.36}$$

where π is given by Eq. B.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

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

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

$$Rate = \frac{R}{Z}\left(\left(1+z\right)^{10}p\left(1+\sum_{i=1}^{6}\binom{6}{i}b^{i}\omega_{b}^{i-1}\omega_{bp}^{i}\right)\right)p_{m} \tag{B.38}$$

$$= R\frac{\left(p\left(1+\sum_{i=1}^{6}\binom{6}{i}b^{i}\omega_{b}^{i-1}\omega_{bp}^{i}\right)\right)}{\left(1+p+\sum_{j=0,1}\sum_{i=1}^{6}\binom{6}{i}b^{i}\omega_{b}^{i-1}p^{j}\omega_{bp}^{ij}\right)}p_{m}.$$

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 B.2.2) as the models presented in Sections B.1.2, B.6.1, and B.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. 3.4D, points).

In order to sufficiently explain the data, we found that a minimum of $m = 3$ irreversible steps was necessary. Fig. B.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 B.8.3) highlights the necessity of having at least $m = 3$ steps.

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

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

$$\pi = c_b[Bicoid] + c_z[Zelda]. \tag{B.39}$$

Here, we assumed that Zelda and Bicoid operate independently and in parallel to catalyze the transitions from the inaccessible to accessible state (Fig. B.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
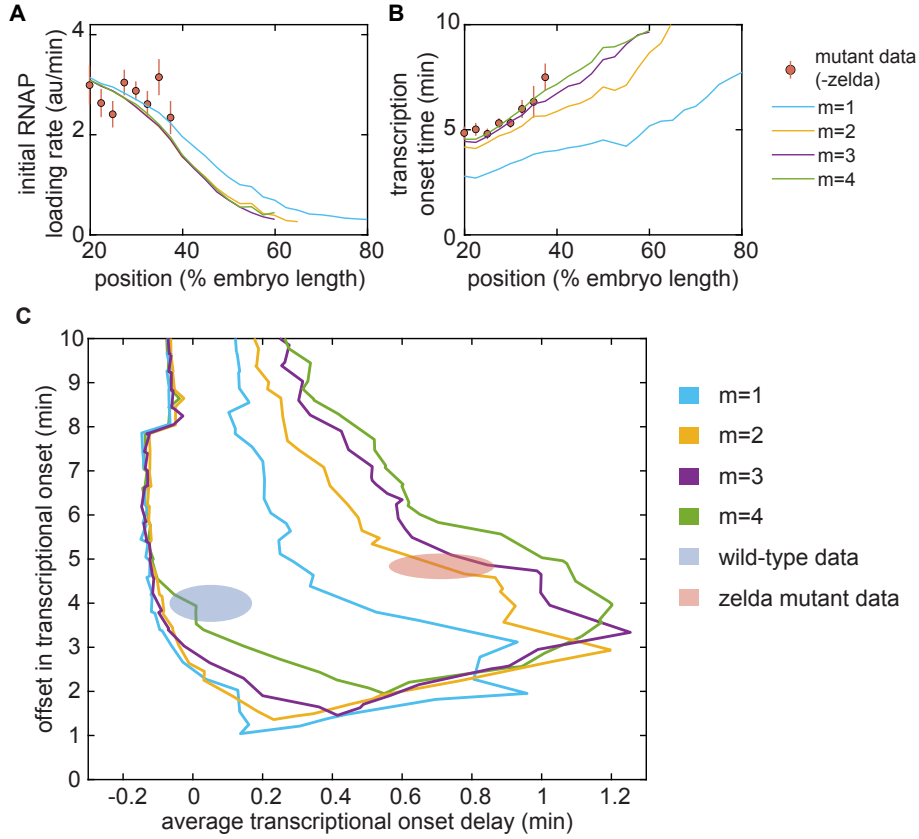
Figure B.14: Testing the transcription factor-driven model of chromatin accessibility. (A,B) Best-fit results of the transcription factor-driven model to the mutant $zelda^-$ data. (A) initial RNAP loading rates, and (B) t_{on}, for varying numbers m of transcriptionally silent states. (C) Parameter 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. B.15B). Here, each stochastic transition is given by

$$\pi = c[Bicoid][Zelda] \tag{B.40}$$

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. B.15C). In this case, the transition rate will be dependent on Bicoid and Zelda such that

$$\pi \sim \frac{c_1[Bicoid]c_2[Zelda]}{c_2[Zelda] + c_1[Bicoid]}, \tag{B.41}$$
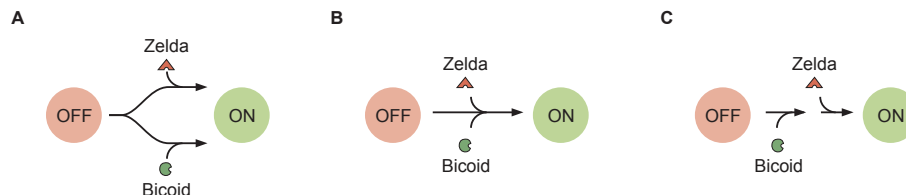
Figure B.15: Different potential schemes of Bicoid- and Zelda-mediated transition into the accessible state, for a model with $m = 1$ transcriptionally silent state. (A) The model used in the main text, where Bicoid and Zelda provide independent pathways for chromatin to transition into the accessible 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.

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.

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

In the parameter exploration of this model (Section B.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 B.1.3.

Fig. B.14C shows the state space explorations (see Section B.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. 3.7D and Fig. B.9, the number of irreversible steps was fixed at $m = 3$.

Unlike the other models investigated (Sections B.1.2, B.6.1, and B.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. B.9, purple).

B.9 Supplementary videos

B.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)*

B.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/wwhsb5crl1pb8vn/ZldGFP_2-10-17.avi?dl=0)*

B.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/frcd4yuu4ki49ih/WTMS2.avi?dl=0)*

B.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/wmzuwxfao2bpz7h/Zld-MS2.avi?dl=0)*

B.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)*

B.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/ndc2jlqxxgte7kp/Zld-MS2_8-06-15_FracOn.avi?dl=0)*

B.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 C

Supplementary Information for Chapter 4

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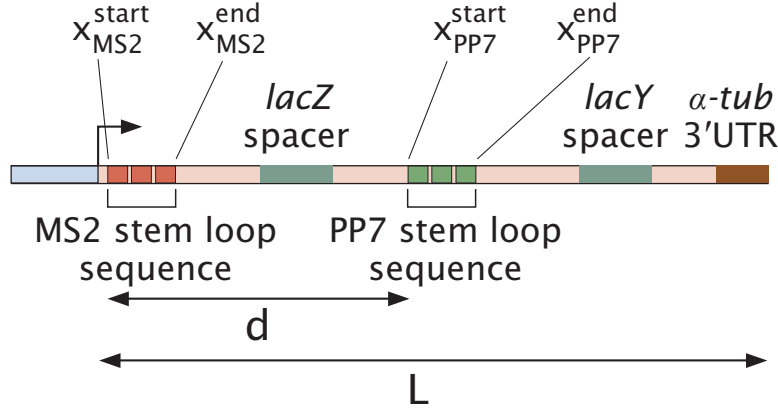
Figure C.1: Detailed description of reporter construct used in this work. Labeled positions are $x_{MS2}^{start} = 0.024$ kb, $x_{MS2}^{end} = 1.299$ kb, $x_{PP7}^{start} = 4.292$ kb, and $x_{PP7}^{end} = 5.758$ kb, where $x = 0$ corresponds to the 3' end of the promoter. Distances are $d = 4.27$ kb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

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

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

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 C.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. C.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

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

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

$$F_i^{PP7}(t) = \begin{cases} 0 & x_i(t) < x_{PP7}^{start} \\ \frac{x_i(t)-x_{PP7}^{start}}{x_{PP7}^{end}-x_{PP7}^{start}} F_{PP7} & x_{PP7}^{start} \le x_i(t) < x_{PP7}^{end} \\ F_{PP7} & x_i(t) \ge x_{PP7}^{end} \end{cases}, \tag{C.3}$$

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

$$F_{MS2}(t) = \sum_{i=1}^{N} F_i^{MS2}(t) \tag{C.4}$$

$$F_{PP7}(t) = \sum_{i=1}^{N} F_i^{PP7}(t), \tag{C.5}$$

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

$$F'_{MS2}(t) = \begin{cases} MS2_{basal}/\alpha & F_{MS2}(t) < MS2_{basal} \\ F_{MS2}(t)/\alpha & F_{MS2}(t) \ge MS2_{basal} \end{cases} \tag{C.6}$$

and

$$F'_{PP7}(t) = \begin{cases} PP7_{basal} & F_{PP7}(t) < PP7_{basal} \\ F_{PP7}(t) & F_{PP7}(t) \ge PP7_{basal}. \end{cases} \tag{C.7}$$

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

C.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. C.2A, purple), and the other half was illuminated at half the temporal sampling rate (Fig. C.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 C.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. C.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 $N_{positions}$

$$\Delta = \sum_{i=1}^{N_{timepoints}} \sum_{j=1}^{N_{positions}} \frac{1}{N_{timepoints}} \frac{1}{N_{positions}} \frac{F_{1x}^{ij} - F_{0.5x}^{ij}}{F_{0.5x}^{ij}}. \tag{C.8}$$
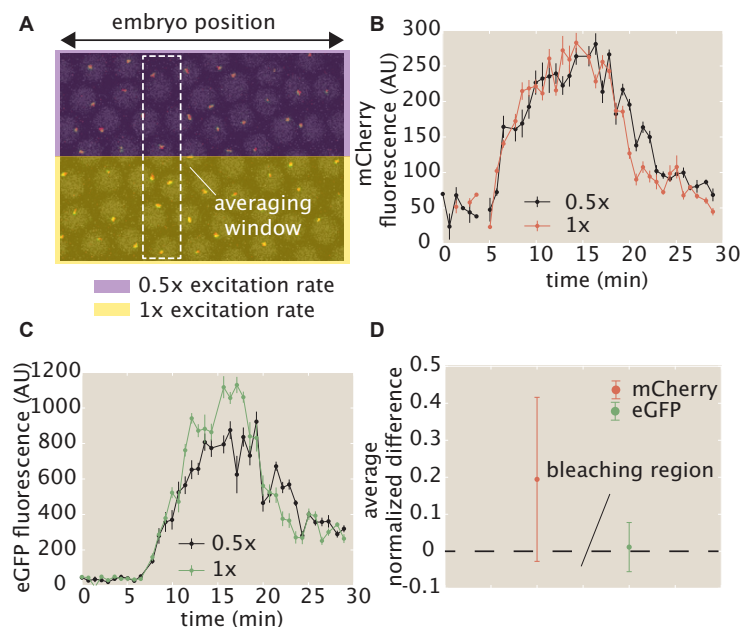
Figure C.2: Investigation of photobleaching in experimental setup. (A) Control experiment where half of the field of view is illuminated at the standard experimental settings (yellow), and the other half of the field of view is imaged at half of the illumination rate (purple). (B, C) The (B) mCherry 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. C.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 C.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.

C.3 Justification for approximating transcript cleavage as instantaneous

In the model presented in Section C.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

$$\tau \sim \frac{L^2}{D}. \tag{C.9}$$

In the context of the experiment performed here, this can 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 $D \sim 0.1 \; \mu m^2/s$ (Gorski et al., 2006). The length scale L corresponds to the Abbe diffraction limit, which yields $L \sim 250 \; nm$ 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

$$\tau \sim \frac{(250nm)^2}{0.1\mu m^2/s} \sim 0.625 \; s. \tag{C.10}$$

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.

C.4 MCMC inference procedure

C.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 the MCMCstat website (`https://mjlaine.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

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

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

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

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. C.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

$$p(D|\theta) = e^{-SS}, \tag{C.13}$$

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

$$SS = \sum_t \frac{(F_{data} - F_{prediction})^2}{F_{data}}. \tag{C.14}$$

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,

$$F_{data} = \left\{ MS2_1, \dots MS2_N, PP7_1, \dots PP7_N \right\} \tag{C.15}$$

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

$$F_{prediction} = \left\{ MS2_1^{pred}, \dots MS2_N^{pred}, PP7_1^{pred}, \dots PP7_N^{pred} \right\} \tag{C.16}$$

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 C.4.2 and Fig. C.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.

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

The observation model parameterized by the sum-of-squares residual in Equation C.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 C.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 C.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 contributes 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.

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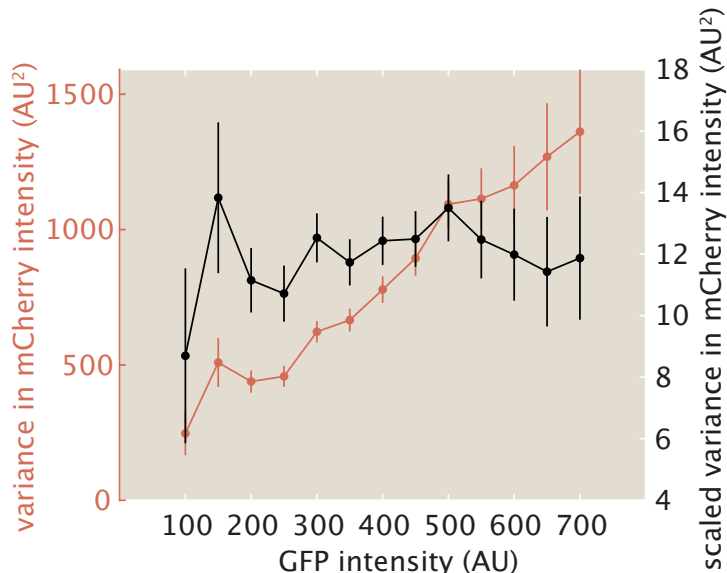
Figure C.3: Scaling of fluorescence measurement noise with overall fluorescence intensity. Variance of mCherry fluorescence at a particular GFP fluorescence (red), from the dual-color interlaced reporter construct from Figure 4.3B, along with variance scaled by dividing out the mean mCherry fluorescence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

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

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 C.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 $\delta_{cutoff}^2 = 1$ (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. C.4C) that nevertheless yielded reasonable fits that were slightly above δ^2_{cutoff}. Still others simply had poor fits, possibly due to running into issues with the inference algorithm such as getting trapped in local minima (Fig. C.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. C.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. (2020a)). 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. C.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. C.4G-I). We observed no substantial difference between the two datasets, indicating that the curation procedure was not systematically altering the inference results.

C.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 C.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 C.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 C.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. C.3).

Figure C.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 C.5B (blue), along with the ground truth for that single cell (red line). As described in Section C.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

$$\varepsilon = \frac{x_{infer} - x_{truth}}{\mu_x}, \tag{C.18}$$

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 C.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 C.5C shows the histogram of single-cell scaled errors $\varepsilon_{\langle R \rangle}$, $\varepsilon_{v_{elon}}$, and $\varepsilon_{tau_{cleave}}$ 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 C.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

$$\sigma_{x,total}^2 = \sigma_{x,empirical}^2 + \sigma_{x,inference}^2, \tag{C.19}$$

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

Dividing by the square of the population means μ_x yields

$$\frac{\sigma_{x,total}^2}{\mu_x^2} = \frac{\sigma_{x,empirical}^2}{\mu_x^2} + \frac{\sigma_{x,inference}^2}{\mu_x^2}. \tag{C.20}$$

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

$$CV_{x,total}^2 = CV_{x,empirical}^2 + \varepsilon_x^2. \tag{C.21}$$

Thus, the overall impact of the inferential error can be quantified by calculating the relative magnitudes of the contributions of $CV_{x,empirical}^2$ and ε^2 to the total variability $CV_{x,total}^2$. Figure C.5D shows this separation, where the dark bars represent the squared scaled error ε^2, the light bars represent the true empirical variability $CV_{x,empirical}^2$, and the overall bars represent the total variability $CV_{x,total}^2$ obtained from the values of μ and σ in Table C.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. C.10D), twice as much as the random error exhibited in the simulated results here (see Section C.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 C.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 C.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. C.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. C.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. C.5H). The comparisons of Spearman rank correlation coefficients and p-values between the data and simulations are summarized in Table C.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.

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

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

C.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

$$\alpha_{\text{calib}} = \frac{NF_{MS2}}{NF_{PP7}} = \frac{F_{MS2}}{F_{PP7}}. \tag{C.22}$$

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

$$\alpha_{\text{infer}} = \frac{NF_{MS2}}{pNF_{PP7}} = \frac{F_{MS2}}{pF_{PP7}}. \tag{C.23}$$

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

$$p = \frac{\alpha_{\text{calib}}}{\alpha_{\text{infer}}}. \tag{C.24}$$

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.

C.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 C.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. C.6D, the inter-embryo variability of each parameter is substantially higher than the intra-embryo variability.

C.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 C.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.

C.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 C.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 C.8A. The resulting

squared CV values are shown in Figure C.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

$$\delta_{MS2} = \frac{F_{MS2}}{\langle F_{MS2} \rangle} - 1 \tag{C.25}$$

$$\delta_{PP7} = \frac{F_{PP7}}{\langle F_{PP7} \rangle} - 1, \tag{C.26}$$

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

$$\eta_{uncorr}^2 = \frac{1}{2} \langle \left(\delta_{MS2} - \delta_{PP7} \right)^2 \rangle \tag{C.27}$$

$$\eta_{corr}^2 = \langle \delta_{MS2} \delta_{PP7} \rangle, \tag{C.28}$$

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

$$\eta_{tot}^2 = \frac{\sigma^2}{\mu^2} = \eta_{uncorr}^2 + \eta_{corr}^2. \tag{C.29}$$

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 C.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. C.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 C.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.

C.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 C.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.

C.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. C.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 time step dt, a randomized sequence of indices is created from the following sequence

$$\mathcal{I} = \{0, 1, \ldots, N\}, \tag{C.30}$$

where $\{1, \ldots, N\}$ correspond to any RNAP molecules $i = 1, \ldots, N$ 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 $\epsilon\, dt$, thus giving an expected distance traveled of $\epsilon\, dt$ per time step (recall that, for a Poisson distribution with parameter $\epsilon\, dt$, 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 $\beta\, dt$, 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 $\beta\, dt$. 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

$$\epsilon_i = Tr(\epsilon, \sigma_\epsilon, 0, \infty). \tag{C.31}$$

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 time step	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 C.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

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

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 $\mu_\tau = 2.5$ min and standard deviation $\sigma_\tau = 1.6$ 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 C.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 C.1), producing simulated single-cell MS2 and PP7 fluorescence traces (Fig. C.10B). Simulated fluorescence noise was added using the same parameters as in the validation simulations discussed earlier (Section C.5, Table C.1, and Fig. C.5). These fluorescence traces were then run through the inference pipeline (Section C.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. C.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. C.10D, blue), where we see that the variance in the case of $\sigma_\epsilon = 0$ is always below that of the data (Fig. C.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. C.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. C.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. C.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. C.10D, upper black arrow), as well as qualitatively agreeing with the inferred distribution (Fig. 4.4D, gold).

C.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 C.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 C.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.

C.12   Supplementary videos

C.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/l9oiwjjgl4hx3uq/Video_1.avi?dl=0)*

C.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)*
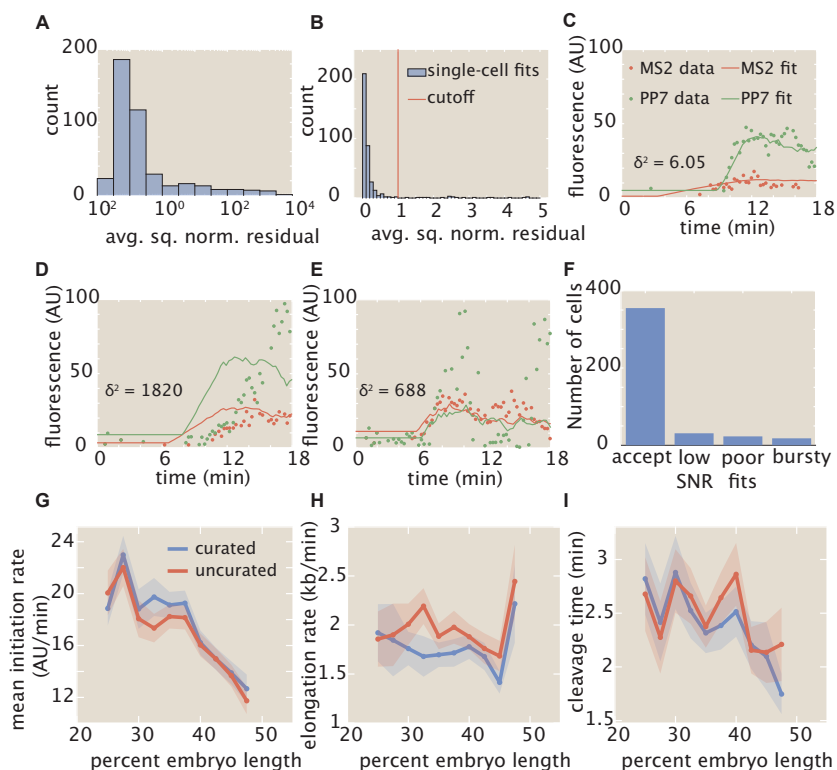
Figure C.4: Automated curation of data. (A, B) Histograms (blue) of average squared normalized residual of single-cell fits, in log (A) and linear (B) scale, with cutoff of $\delta^2_{cutoff} = 1$ shown in red in (B). (C) Example of bad fit from poor signal-to-noise ratio (SNR). (D) Example of bad fit 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.
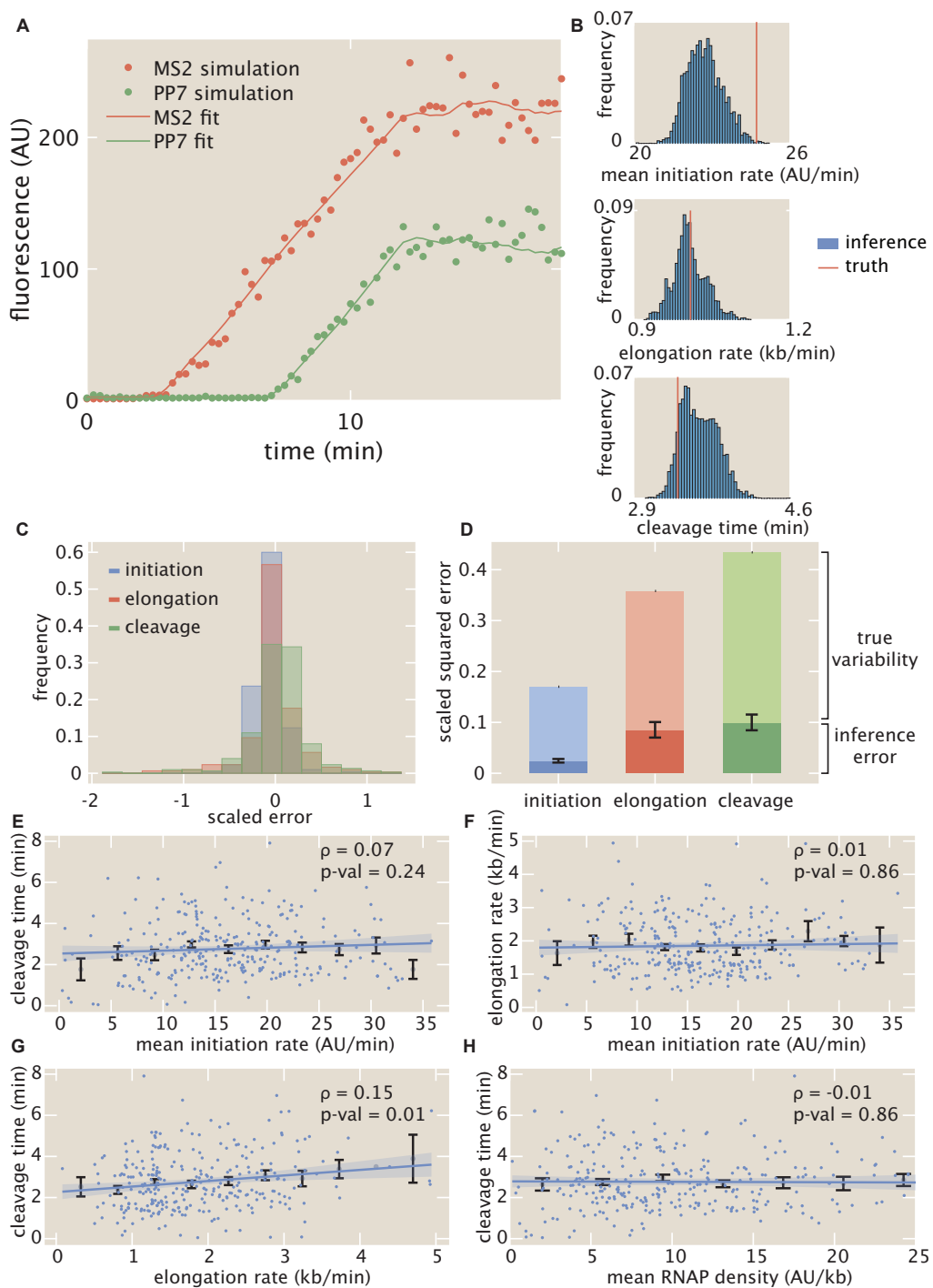
Figure C.5: Overview of MCMC inference validation. See caption on next page.

Figure C.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).
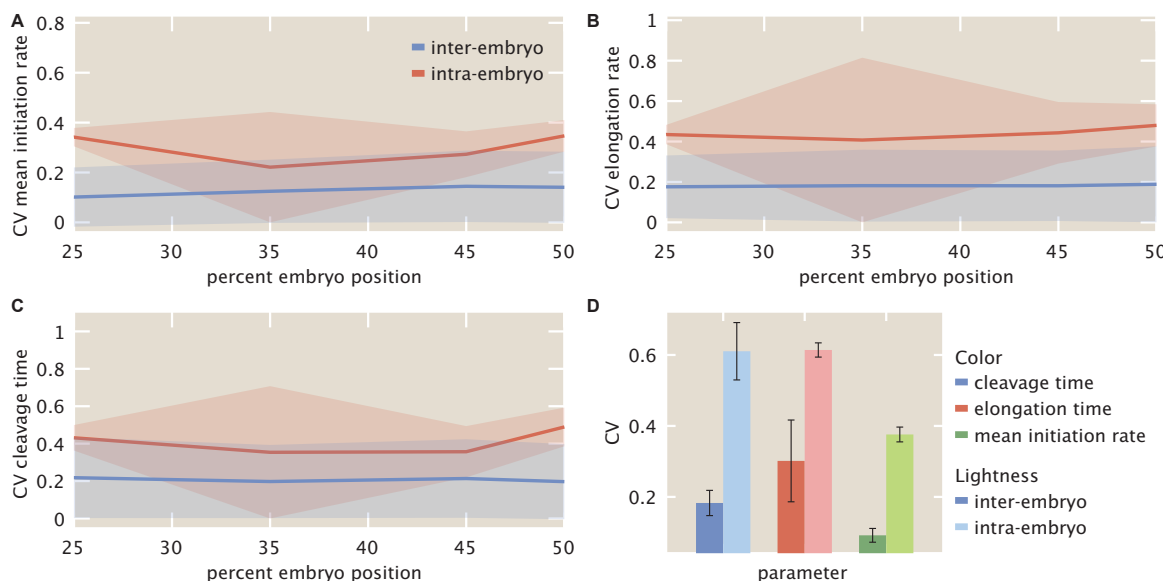
Figure C.6: Comparison of intra- and inter-embryo variability for inferred (A) mean initiation rates, (B) elongation rates, and (C) cleavage times, as a function of embryo position. (D) Intra- and inter-embryo variability for transcriptional parameters averaged across all embryo positions. (A-C, lines 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.)
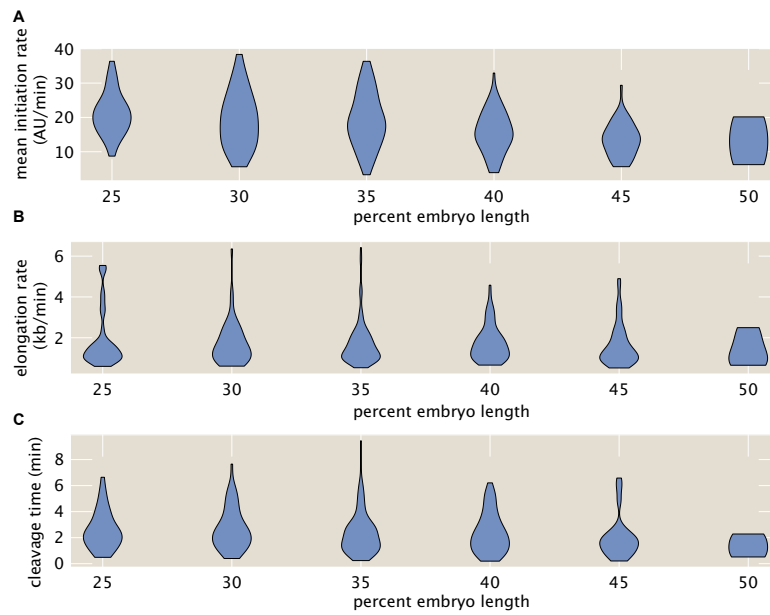
Figure C.7: Single cell distributions of inferred parameters. (A-C) Full single-cell distributions of (A) mean initiation rate, (B) elongation rate, and (C) cleavage time as a function of embryo position.
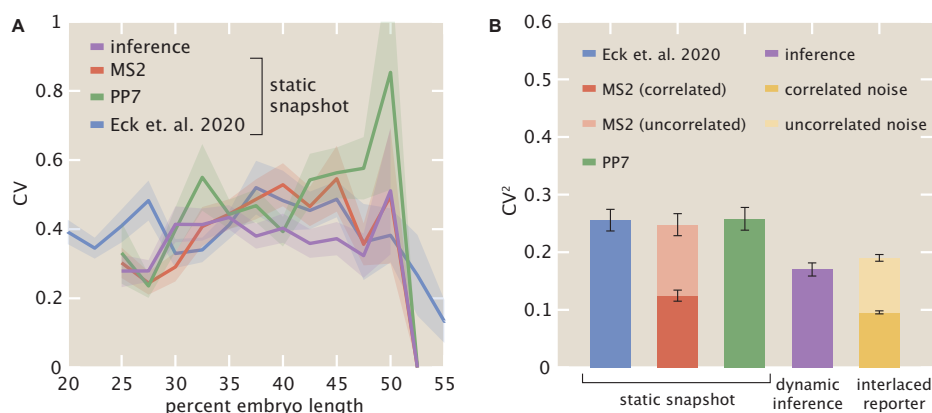
Figure C.8: Comparison of coefficients of variation (CV) between inferred mean initiation rates and instantaneous counts of number of nascent RNA transcripts. (A) Position-dependent CV of inferred mean initiation rate (purple) compared with static measurements of MS2 and PP7 raw fluorescence (red, green) 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.)
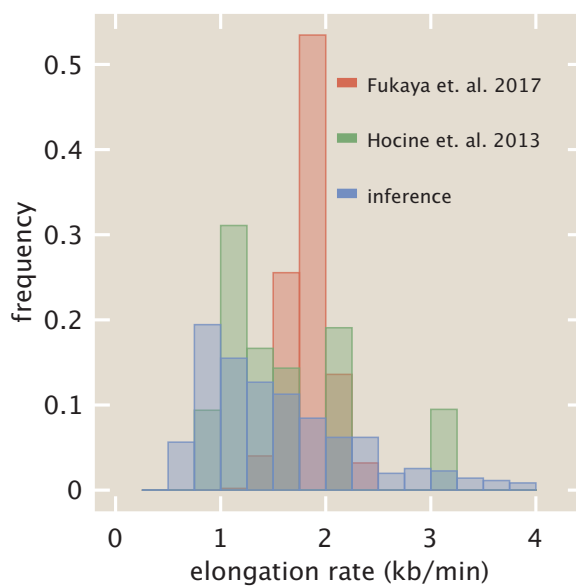
Figure C.9: Comparison of distribution of elongation rates (green) with previous studies (Hocine et al. (2013), red and Fukaya et al. (2017), blue). Distributions of previous studies were adapted from Figs. 2D and 2A of Hocine et al. (2013) and Fukaya et al. (2017), respectively.
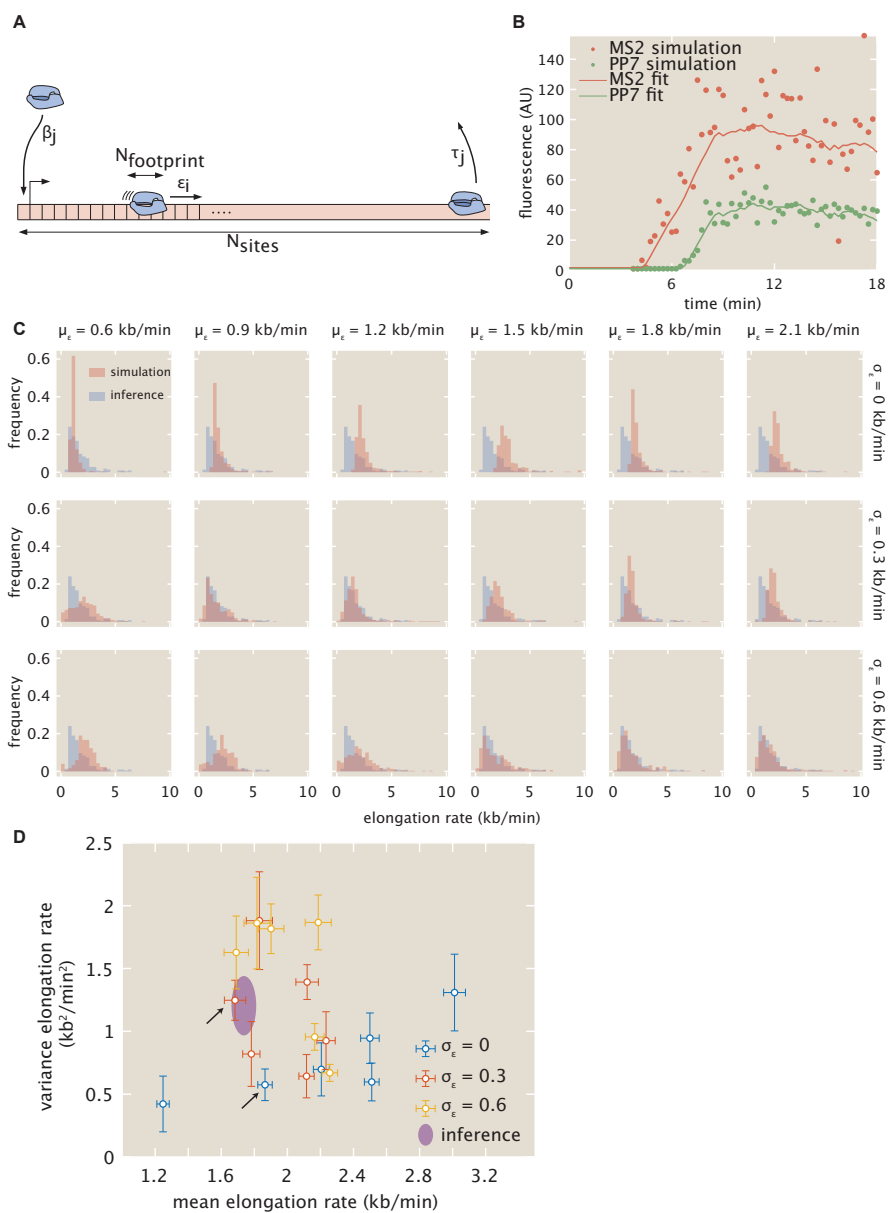
Figure C.10: Single-molecule simulations of elongation dynamics require molecular variability to describe empirical distributions. See caption on next page.

Figure C.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).
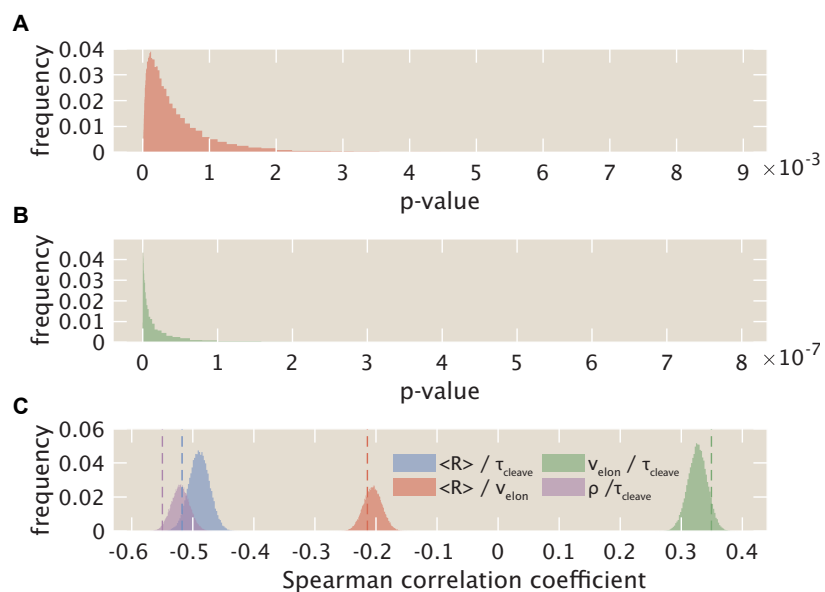
Figure C.11: Monte Carlo simulation of error in single-cell analysis. (A, B) p-values of Spearman correlation coefficient for relationships between mean initiation rate and elongation rate (A) and between elongation rate and cleavage time (B). The p-values for the relationships between mean initiation 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.

Appendix D

Supplementary Information for Chapter 5

D.1 Supplementary video

D.1.1. **Video S1.** Video of light sheet microscopy images of *her1* driven transcriptional dynamics. This video shows a zebrafish embryo expressing the *her1*-driven MS2v5 reporter transgene and *ubiquitin*-driven MCP-mNeonGreen with a 10 amino acid linker between the coat protein and fluorescent protein. The upper video shows the MCP channel. Cell migration from right to left (posterior to anterior) is evident. A vertical row of transcription spots can be see passing through the field of cells, traveling from the posterior to the anterior of the embryo. The bright notochord can be seen at the top of the frame. Bright skin cells are evident in the lower right corner of the frame. Mosaic expression of MCP is evident. Nuclear localized mKate2 can be seen in the lower video. Image planes in z were spaced 0.5 μm apart. This video shows a maximum projection of a 50 plane substack. Frames were taken approximately every 30 seconds. The full video takes place over approximately 29 minutes. Scale bar = 10 μm. *(https://www.dropbox.com/s/gn3aciuwum6qw3u/Spots%2BNuclei.avi?dl=0)*