Title

Molecular Mechanisms of Precise and Robust Gene Regulation in Drosophila

Author

Boettiger, Alistair Nicol

Publication Date

2011

Molecular Mechanisms of Precise and Robust Gene Regulation in Drosophila

by

Alistair Nicol Boettiger

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Biophysics
and the Designated Emphasis
in
Computational and Genomic Biology
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor Michael Levine, Co-chair
Professor Daniel Rokhsar, Co-chair
Professor George Oster
Professor Steven Evans

Spring 2011

Abstract

Molecular Mechanisms of Precise and Robust Gene Regulation in Drosophila

by

Alistair Nicol Boettiger

Doctor of Philosophy in Biophysics

and

the Designated Emphasis in Computational and Genomic Biology

Professor Michael Levine, Co-chair

Professor Daniel Rokhsar, Co-chair

The ornate arrangement of diverse cells into specialized tissues, organs, and higher structures characteristic of multicellular organisms is all encoded from the same genome sequence. Despite their differences, morphologically distinct cells (e.g. muscle cells and neurons) must transcribe many of the same genes. Morphological indistinguishable cells must often transcribe distinct sets of genes (e.g. different odorant receptor cells). The ensemble of genes expressed in a given cell – and the relative frequency they are expressed at, give each cell its characteristic identity more so than the presence of individual genes. Therefore understanding the genetic control of development and differentiation is a question not so much of the understanding the gene sequences themselves, but the regulatory structure of the genome which determines how they are deployed.

In order for development to unravel in such a manner that each embryo makes it through the process with all the correct parts in the correct positions at the end, this process must be exceedingly precise. Though often taken for granted, this precision becomes particular impressive if one considers the frequency with which mistakes are made in intelligently designed human built assembly processes. The developing animal must position components correctly on scales of microns (e.g. tissue boundaries) and nanometers (e.g. neuron-junctions), has no external direction of assembly, and requires

thermal noise to position many of its components (including essentially all transcription factors - proteins which regulate read access to the genome).

It is not sufficient for the process to be precise. It must also be robust to changes in the conditions in which it operates, such as different thermal environments, nutrient conditions, and chemical environments. This robustness enables a certain degree of plasticity, such that some components of the system can change and evolve new functions, without causing catastrophic failure of the rest of the system.

In my thesis research I have tried to explore some of the molecular mechanisms of gene regulation which support the precise and robust expression of multicellular genomes. Rapid advances in post-genomic technologies have exposed a broad range of fundamental differences in the organization and regulation of multicellular genomes such as Drosophila. I have worked primarily on two phenomena, the use of promoter proximal pausing as a regulatory strategy, and the use of multiple apparently redundant regulatory sequences to drive expression of the same gene. Discovery of both of these phenomena emerged from analysis of whole genome polymerase and transcription factor binding data. Using quantitative high resolution in situ and semi-automated computational image processing I have studied the detailed differences in the transcriptional activation and transcription frequency of genes regulated by these mechanisms. Through this analysis I have shown a strong correlation through more rapid and synchronous gene expression and regulation through release of promoter proximal paused polymerase. Theoretical modeling demonstrates that such an effect can be expected from regulating release of stable downstream state in a general assembly process (such as construction of the RNA Pol II pre-initiation complex).

Analysis of gene expression driven by multiple enhancers with overlapping activity compared to constructs with only a single active enhancer revealed that the process by which an enhancer binds its target transcription factors and activates expression is often limiting enough that having a second independent copy can produce detectable changes in the frequency of transcription. This reduction of natural variation in gene activation is especially important under stress conditions, such as thermal stress or reduced levels of some of the activating factors. Robustness to this sort of variation may be important both for adaptation within a species and the flexibility to allow modification of interacting pathways in the course of evolutionary modification. These investigations also revealed a corrective propensity whereby the simultaneous activity of multiple enhancers, responding to repressors as well as activators, can give rise to correctly restricted gene expression even when the elements taken in isolation drive some degree of ectopic expression.

So far both of these mechanisms have only been reliably documented in multicellular systems, suggesting that the precision and robustness they confer may be an innovation of metazoans in response to increased levels of coordination required to keep many cells functioning in the tight cooperation of a multicellular organism. Doubtless this is but scratching the surface of the mechanisms which ensure such precision and control. However the rapid improvements to both genomic tools and imaging technology make it like to be a promising field for further exploration for years to come.

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

Weak electrostatic interactions between randomly diffusing proteins and small, individual gene loci are responsible for all of the highly organized complexity we see in the biological world. When organized together in carefully constructed networks, these individually variable, uncontrolled interactions of the molecular scale generate precise, reliable control of behavior. Though the component interactions are highly dependent on the concentrations of the molecular players and the detailed environmental conditions, the input output behaviors of the signaling pathways are often very robust to such perturbations. I am interested in investigating what properties of the interaction network allow for this emergent reliability instead of compounded chaos.

1.2 Experimental System

The dorsal-ventral patterning system of the early *Drosophila* embryo provides an ideal system in which to investigate robust and reliable control of gene expression and its role in patterning. First, as a developmental tissue, all of the signals function in their naturally evolved context (unlike induction of signaling in a cell-culture system). The gene regulatory network controlling this system is well mapped [91]. We know who the critical genes are and which genes regulate the expression of which other genes. More so than in almost any other system, we have identified the specific sequence elements that convert signals (concentrations of different transcription factors) into expression patterns for particular genes [69, 70, 73, 74, 76, 75, 174, 173]. These distal control elements we call *enhancers*. For a brief review of the DV system see Hong *et al* 2008 [63].

Not only are the components of the system well characterized, but recent improvements, particularly in imaging and genomics, bring a wealth of tools to inform and test models of signal processing and gene regulation. The system is well suited for quantitative confocal imaging [25, 26]. We can measure the spatial profiles of the concentration gradients of transcription factors [26, 49], make quantitative comparisons across different genetic backgrounds [26] and even estimate molecules per nucleus [49]. With complete sequence information, we can design probes to trace the spatial and temporal expression patterns of any of the involved genes. And increasing amount of genome wide data on transcription factor binding and polymerase binding is also available [171]. Thanks to genetic mutants in DV system, some of this binding data is available separated by tissue type [170]. Meanwhile with other genetic tools we can selectively reduce the gene dosage of individual components of the system or remove the gene entirely. With available transgenic tools, we can create ectopic expression of select genes and test if these additional signals are processed as predicted by our models. We can also test predictions of how given sequences determine expression pattern by observing the profiles of reporter genes driven by these synthetic constructs. Of particular importance to the study of reliability in patterning, the small size and easy collection of embryos allows hundreds of animals to be labeled and imaged simultaneously. This allows for good statistics and quantification of the population *variability* and not just individual examples.

1.3 Research Goals

Despite the extensive knowledge of components and diverse collection of tools both to make measurements and perturbations we have a very incomplete understanding of how these components function as a system to create precise and reliable patterning. For example, there is no mechanistic explanation how the mesodermal boundary is both switch like in response and robust to perturbations in concentration, though it is one of the very first and best studied readouts of the dorsal system [69, 66, 152].

The Drosophila embryo offers a well adapted model system to understand some of the properties of the more complex, multilayered regulatory networks

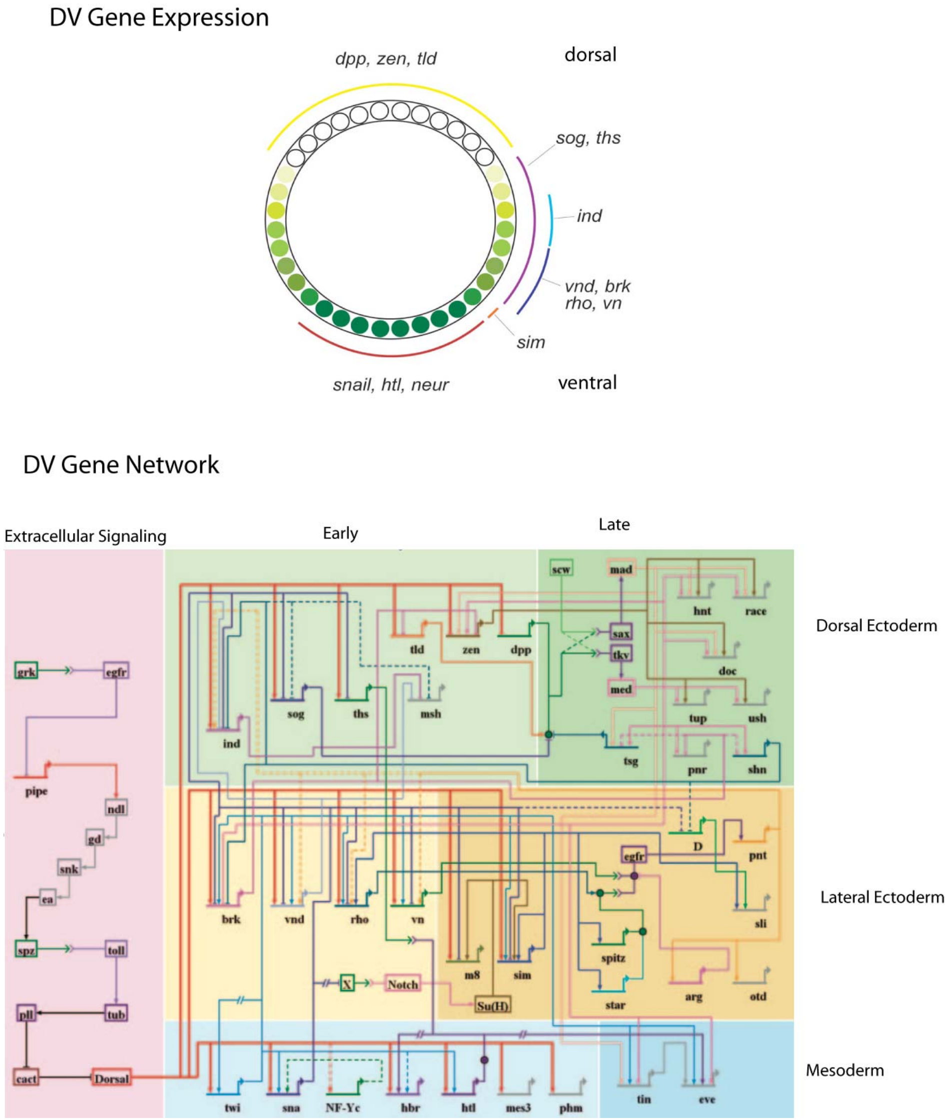

Figure 1.1: The Dorsal Ventral Patterning System. **A** Schematic cross section of an (reproduced from Hong *et al* 2008). **B** The dorsal ventral patterning network in *Drosophila* Assymetric patterning of pipe protein in the egg chamber leads to ventral activation of the Dorsal protein inside the embryo (red region). Dorsal regulates expression of genes in all three presumptive DV tissues, the mesoderm (blue), lateral ectoderm (yellow) and dorsal ectoderm (green). Some of the temporal separation in the activation of these genes is indicated by the darker shading of the later expressed genes (Reproduced from Levine and Davidson 2005).

controlling transcription of metazoan genomes. Why are some genes regulated through repression or induced through repressor of repressors, while others controlled through direct activation? Why do some genes regulate the initiation of gene expression and others regulate the mRNA elongation half way through the expression process? What are the consequences of these differences on the ability of the network to process signals in a precise and reliable manner?

In my thesis work I have focused on two particular features of early Drosophila gene regulation whose origins and effects on transcription has not previously been investigated. The first feature is the observation that the transcription of a considerable number of developmental control genes is regulated not by controlling in which cells the promoter binds pol II, but rather in which cells bound pol II is permitted to proceed to productive elongation. The latter part of my research has focused on understanding why many regulatory elements appear to have overlapping, apparently redundant roles in transcriptional regulation. Using high resolution microscopy techniques to probe the behaviors of individual cells which use or lack one of these modes of regulation, I present evidence that both regulatory strategies function in reducing cell-cell variation in expression. In response to especial requirements for precise and robust gene expression of a rapidly developing multicellular organism like the Drosophila, these regulatory schemes provide for more reliable control through individually stochastic components.

Chapter 2

Paused Polymerase and Synchronous Expression

2.1 Background: Active promoters on inactive genes

In 2006, a collaboration between the Levine, Young, and Kellis labs combined the powerful genetic mutations available in Drosophila with recently developed whole genome pol II binding assay to determine the complete tissue specific gene expression patterns in the early Drosophila embryo's three primary tissues, the dorsal ectoderm, lateral ectoderm and presumptive mesoderm. These assays provided an unexpected additional result. Many of the genes which are regulated in a tissue specific expression patterns show substantial pol II binding at the promoter of cells from all tissues – even though release of polymerase into the body of the gene is restricted to a single tissue. This result was observed both for genes which are actively repressed in the off tissue (such as the gene *sog*, which is repressed in the mesodermal tissue by snail, and shows nonetheless strong promoter pol-II binding in both the mesodermal and ecotdermal tissues) and for genes which are not activated (for example, *snail* is only activated by the high concentrations of the transcription factor Dorsal present in the mesoderm, and is not induced in ectodermal tissues. Nonetheless the promoter of this gene is strongly bound in ectodermal tissues as well as mesodermal tissues) [170]. This distinction is shown in figure 2.1.

These results challenged the common belief that access of polymerase to

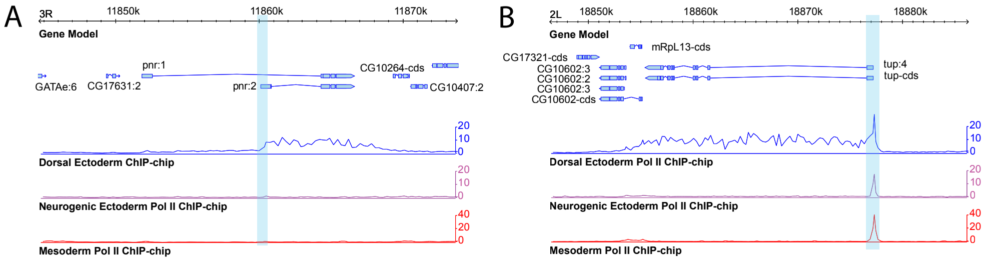

Figure 2.1: **Identification of Paused Polymerase in Drosophila by Chip-chip:** **(A)** Gene models are shown top, aligned to Pol II chromatin immunoprecipitation signal measurements from whole genome tiling array, showing locations in the genome where Pol II is bound in each of three specific tissues – the dorsal ectoderm, the neurogenic ectoderm, and the mesoderm, from Zeitlinger 2007 [170]. *pnr* is expressed only in the dorsal ectoderm – the promoter (highlighted region) is silent in the other tissues. **(B)** Genome data as in (A) for the region around the gene *tup*. In this case the promoter region is bound in all three tissue types, even though the rest of the gene is only transcribed in the dorsal ectoderm.

the promoter of a gene is the primary mode of regulating gene expression [79, 122], and raised new questions as to what the possible selective advantages, if any, of this alternate mode of gene regulation could be.

These results also illustrated several of the peculiar strengths of the Drosophila embryo as a model system for investigations in details of gene regulation on a genome wide scale. This tissue specific assay was accomplished by exploiting the robust mutant collection available for Drosophila research. A genetic trick using different mutant lines was employed to generate populations of embryos containing only one of each tissue type. Females mutants for the maternal DV signaling factor Pipe were used to produce eggs lacking any Dorsal activation and hence giving rise to purely dorsal ectodermal tissue. Females carrying a constiuitively active allele of the Toll receptor (*toll[10b]*), produced constiuitively high levels of Dorsal activation and turn most of the embryo into presumptive mesoderm type cells, expressing a complete set of genes normally characteristic of the presumptive mesodermal tissue [141, 89, 83, 44, 152, 12]. Females with two different partial loss of function alleles, (*toll[rm9]/toll[rm10]*) lay embryos which have intermediate levels of Dorsal activation throughout the embryo. This condition induces the subsequent activation of lateral ectoderm target genes, at the expense of both Dorsal ectodermal and presumptive mesodermal tissues (which require higher levels of Dorsal activation for induction) [141, 152, 12, 66].

2.2 Preliminary Studies: Pausing and Immunity

I began my investigation of potential expression differences using a cell culture assay to compare the behavior of paused and non-paused immune-response genes. Cell culture provided an easy start up and immune signaling a simple inducible system to compare the nature of activation. Based on some preliminary modeling of the effect of regulating downstream transcriptional events, I had a prediction that paused genes should be expressed faster upon induction and possibly in a more synchronous fashion as well. The development of this model and a detailed mathematical analysis is provided in the next chapter.

Here I discuss my methods and findings from these preliminary studies in cell culture. These experiments were designed in consultation with Kate Senger, a postdoctoral student already conducting research on paused and non-paused immunity genes.

2.2.1 Materials and Methods

engineering constructs

The YFP sequence was amplified by PCR from YFP containing plasmids (gift from J. Cande). The sequence was purified by gel extraction, and used to transform S2 cells using the TopoII cloning kit (Invitrogen). Successfully cloned colonies were grown up overnight in LB media with ampacilin, and plasmids were extracted using Miniprep kit (Invitrogen). Constructs were checked by sequencing. The YFP sequence was cut from the plasmid using NotI and XbaI (an enzyme co). PGL3-plasmid transformation vectors containing the upstream regulatory elements of select immunity genes fused to Luciferase were cut with NotI and XbaI and treated with CIP to prevent self-ligation. Restriction fragments were separated by gel extraction and ligated using T7 ligase for two hours before being transformed into competent cells. Sequences from the resulting colonies were checked by sequencing. Colonies with intact regulatory regions and YFP genes were selected for amplification and DNA extraction using a Maxiprep kit (Invitrogen).

cell culture and transfection

S2 cells cultured in Schnider's Insect Media at 25C on glass slides were transfected using calcium chloride shock. Briefly, 4 micrograms of DNA were mixed with 8 microliters of calcium chloride, dilute to 50 microliters with water, and added to 50 microliters of BBS. Precipitate was allowed to form for thirty minutes, and then cells were transfected with 100 microliters per well. Cells were allowed 48 hours to recover before induction and live imaging.

live imaging

Time lapse movies of live S2 cells were taken on a Nikon TE2000 inverted fluorescent microscope with an automated mechanical stage. Twenty locations for each strain were imaged using a 10x objective every two minutes. The precise focus for each location was kept the same throughout the imaging experiment. Images of cells prior to induction served as a control for leaky expression. A separate population of transfected but never induced cells provided another control for possible induction during the extra hours of incubation. This latter population exhibited no detectable gene activation events during the time of observation. Fluorescent lamp exposure was kept at less than 500 ms per image, to reduce cell damage and photobleaching.

image processing and analysis

Time lapse videos of S2 cultures were converted into image sequences. As soon as a given cell produced enough YFP to be detected above a low scale threshold, the corresponding time of the activation event was recorded. This approach generates a distribution of individual waiting time to activation for each population.

2.2.2 Results

The paused immunity gene *CecA1* exhibits faster and more synchronous response than the unpaused *mtk* gene

To investigate stochastic variability in gene induction we studied activation of the immune response of *Drosophila* cells to the bacterial cell wall component LPS. We transfected *Drosophila* S2 cells with constructs containing the regulatory region of select immunity genes from the Toll pathway, fused to YFP reporter genes. Since the gene is identical in both constructs, stochastic variability in transcription, splicing and translation should have identical noise signatures. Any difference in the overall noise signatures of the pathway we observe must result from differences in the induction mechanisms for the different control elements. We induced these cultures with LPS and recorded a time lapse film over the following four hours (see materials and methods).

From these films, we determine the time at which individual cells successfully induced and synthesized enough YFP to exceed our detection threshold. Sample time lapse images are shown in figure 2.2. Some cells start producing YFP almost immediately following induction, and the signal is detectable after half an hour latency. Other cells do not start transcribing YFP until a few hours after induction. These general characteristics are observed for all transfects we examined, however, the details of the distributions vary significantly.

Cells transfected with plasmid where YFP is downstream of the *Cecropin A1* (*CecA1*) control region exhibit on average a faster and more synchronous activation than those under control of the *Metchnikowin* (*Mtk*) regulatory element. Around one hour after induction, the number of cells expressing YFP under control of *CecA1* climbs steeply. After an hour and half, the rate at which new cells are induced drops notably, with comparatively few cells achieving above-threshold YFP induction after two hours gap since induction. In contrast, the rate of *Mtk* induction increases steadily and slowly over two hours. There is no clear peak time at which the majority of cells induce. The mean induction time is later than observed for *CecA1*. As a substantial number of cells only passed the threshold during the fourth hour of observation, the distribution suggests more cells will likely induce even after four hours, shifting the mean activation time even later. This broad distribution indicates a considerable contribution of stochastic effects to the induction process.

Both *CecA1* and *Mtk* function in the bacterial induced immune response of *Drosophila*. The control elements we used from these genes have been shown to reproduce the endogenous expression patterns when the transcribed gene portion is replaced with a reporter gene like Luciferase (Senger 2007, unpublished data). These control elements differ functionally, since *CecA1* supports a paused polymerase prior to induction, while there is no evidence of pre-attached, paused polymerase at the *Mtk* locus (Senger 2007, unpublished data).

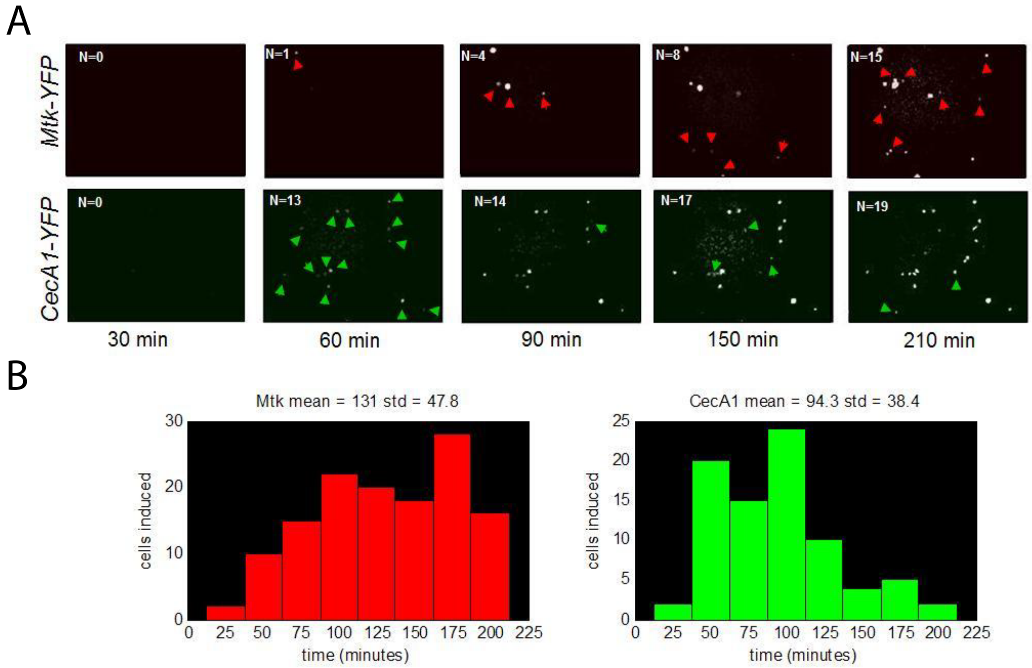

Figure 2.2: **Variable Induction Times: A** The induction response of two immunity genes, Cecropin A1 (CecA1) and Metchnikowin (Mtk) activated by LPS. Select time lapse images illustrate the variable delay time until immune response is activated in different cells. Note the steady increase in the number of active cells among the *Mtk* images (red arrows indicate newly activated cells). In contrast, most of the cells active in the final panel for *CecA1* were already active in the second panel (green arrows mark newly activated cells). **B** The distribution of the delay times is plotted for the induction under control of the *Mtk* and *CecA1* regulatory regions.

LPS activates immune response through the Toll signaling pathway [159]. This pathway activates expression of several antimicrobial peptide including the Cecropins and Metchnikowin. The molecular function of the protein product of *CecA1* is unknown. Metchnikowin is primarily an anti-fungal protein, and is also induced through the immunity deficiency (IMD) pathway.

It was also observed in the course of this study that YFP expression in *CecA1* transfects is noticeably more leaky than the expression of *Mtk*. It is concievable this results for failure to arrest and pause the polymerase which is loaded in the absence of inducing signal. However this result was not followed up on in greater detail.

These data emphasize the clear stochastic nature of gene expression in the immune response pathway. The waiting time to achieve reliable protein production is not a deterministic quantity, but a random variable dependent on the stochastic interactions of molecular signaling. Previous studies of gene induction concentrated on how transcription and translation contribute to this noise (out-put noise). Our results highlight the contribution of input noise. The differences observed between induction of *CecA1* and *Mtk* is possibly a consequence of the differences in their regulatory mechanism – one controlling release of polymerase from a paused state and the other controlling pol-II promoter binding. If this is indeed a dominant cause for the differences seen in the YFP expression in these experiments, then the data indicate that noise at this input level is not negligible relative to noise at the output level and that different control mechanisms greatly affect its modulation.

2.2.3 Follow-up studies

Though promising, the cell culture experiments suffer from a variety of restrictions, not least of which was the lack of a substantial number of inducible paused and unpaused genes to test for a more general trend. This restriction could be overcome by turning to the blastoderm embryo, where the various signals of development themselves provide precisely timed and reproducible induction of gene expression (since the embryo itself is coordinating the gene induction events and providing the appropriate signals, rather than relying on the experimentalists pipetting accuracy to provide reproducible signals and without the corresponding uncertainty for the physiological relevance of the signal level used in induction.

In addition to turning from immune signaling in cell-culture to developmental signaling the embryo, I switched from live protein imaging using YFP to mRNA detection by in situ. Though this would make it substantially more challenging to infer dynamics, it allowed for measurements of a signal closer to the upstream regulatory difference – removing the worry about differential translation properties of the genes under study. Nonetheless, using large populations of embryos sampled uniformly from a relatively short time window of development (30-40 minute window, identifiable by morphological changes) some of the dynamics of the process could still be inferred. It would also allow for higher sensitivity of detection.

2.3 Synchronous and Stochastic Activation *in vivo*

2.3.1 Introduction

In this section I present the primary findings from the *in vivo* analysis of kinetics for elongation regulated paused genes compared with non-paused, initiation regulated genes. This discussion borrows from the text and figures that we published in *Science* 2009.

We report the use of a high throughput, partially automated quantitative in situ hybridization method to examine the initial activation of gene expression in the early Drosophila embryo. The analysis of 14 different gene expression profiles in hundreds of embryos reveals two distinct patterns of induction: synchronous and stochastic. There is a tight correlation between the occurrence of stalled Pol II and synchronous patterns of gene induction. In contrast, genes lacking stalled Pol II exhibit stochastic patterns of activation. The analysis of mutant embryos with reduced Dorsal activator concentrations revealed a further distinction between stochastic and synchronous modes of transcriptional induction. Stalled genes with multiple enhancers (shadow enhancers) maintain synchronous expression patterns, whereas stalled genes with single enhancer elements exhibit stochastic activation. We propose that transcriptional synchrony helps ensure the orderly deployment of the complex gene regulatory networks that control embryogenesis.

Pol II ChIP-chip assays identified 1000 stalled genes in the early Drosophila embryo [165, 170, 108], including at least 100 developmental control genes, such as Hox genes and components of the Notch, Hedgehog, FGF, and Wnt signaling pathways. Classical studies on the regulation of the heat shock gene hsp70 suggested that stalled Pol II renders genes poised for rapid activation [61, 94]. Here we present evidence that stalled genes are activated in a synchronous fashion in the early embryo, while genes lacking stalled Pol II display stochastic patterns of induction.

Eukaryotic transcription is an intrinsically stochastic process due to variability in the recruitment and subsequent assembly of the Pol II complex and associated coactivator complexes such as Mediator, TFIID, and TFIIH [123, 125, 129, 130]. Consequently, not all cells that receive the same inducing signal would be expected to respond at precisely the same time. In principle, cell-to-cell variation in the onset of transcription might be diminished for genes containing stalled Pol II, whereby polymerase loading is uncoupled from gene activation. Indeed, this prediction can be demonstrated mathematically through analysis of the corresponding Markov Process [18].

The early Drosophila embryo is an ideal system to examine variability in the onset of de novo transcription within a developmental field of coordinately induced cells. However, most previous studies examined relatively few embryos [9, 26, 50, 49, 51]. To distinguish subtle differences in the patterns of transcriptional induction we used computational methods to process hundreds of staged embryos in a quantitative and unbiased fashion. This procedure employs in situ hybridization with a battery of fluorescent probes against large intronic regions of the target genes, high-resolution confocal microscopy, and semi-automated image segmentation algorithms (see Methods). Using this method we can identify most nascent transcripts in transcriptionally active nuclei (as in Figure 2.3A-C).

2.3.2 Results

Natural Patterns of Activation

We began the analysis with six genes that are activated at approximately the same stage (early nuclear cleavage cycle 14) in the presumptive mesoderm. Three of the genes (Mes4 [152] or NF-Yc [100], Mef2 [6], and *hbr* [106] or dof [172] or stumps [68]) lack stalled Pol II in tissues where the genes are inactive, but display Pol II binding across the length of the transcription unit in the mesoderm when expressed (see diagrams above embryos in Fig. 2.3D-F). At the time of induction, all three genes display asynchronous patterns of expression, as judged by the nuclear hybridization dots representing de novo transcripts. However, within 30 minutes after induction, hybridization dots are detected in most mesodermal nuclei of most embryos (see 2.4).

Very different results were obtained with *htl* [90] (FGF receptor), Mes2 [152] (a SANT-domain transcription factor), and Mes5 [152] (also called *Mdr49*, a membrane ATPase [135]), which contain stalled Pol II in tissues where the genes are inactive (top diagrams, blue and red Pol II traces, Fig. 2.3A-C). These genes display synchronous patterns of activation, with clear hybridization signals detected in most nuclei of most embryos (Fig. 2.3A-C).

Distinct patterns of gene activation, synchronous and stochastic, are also seen for genes expressed in the other primary embryonic tissues, the neurogenic ectoderm and dorsal ectoderm (Fig. 2.5). *thisbe* [53, 151] (FGF8)

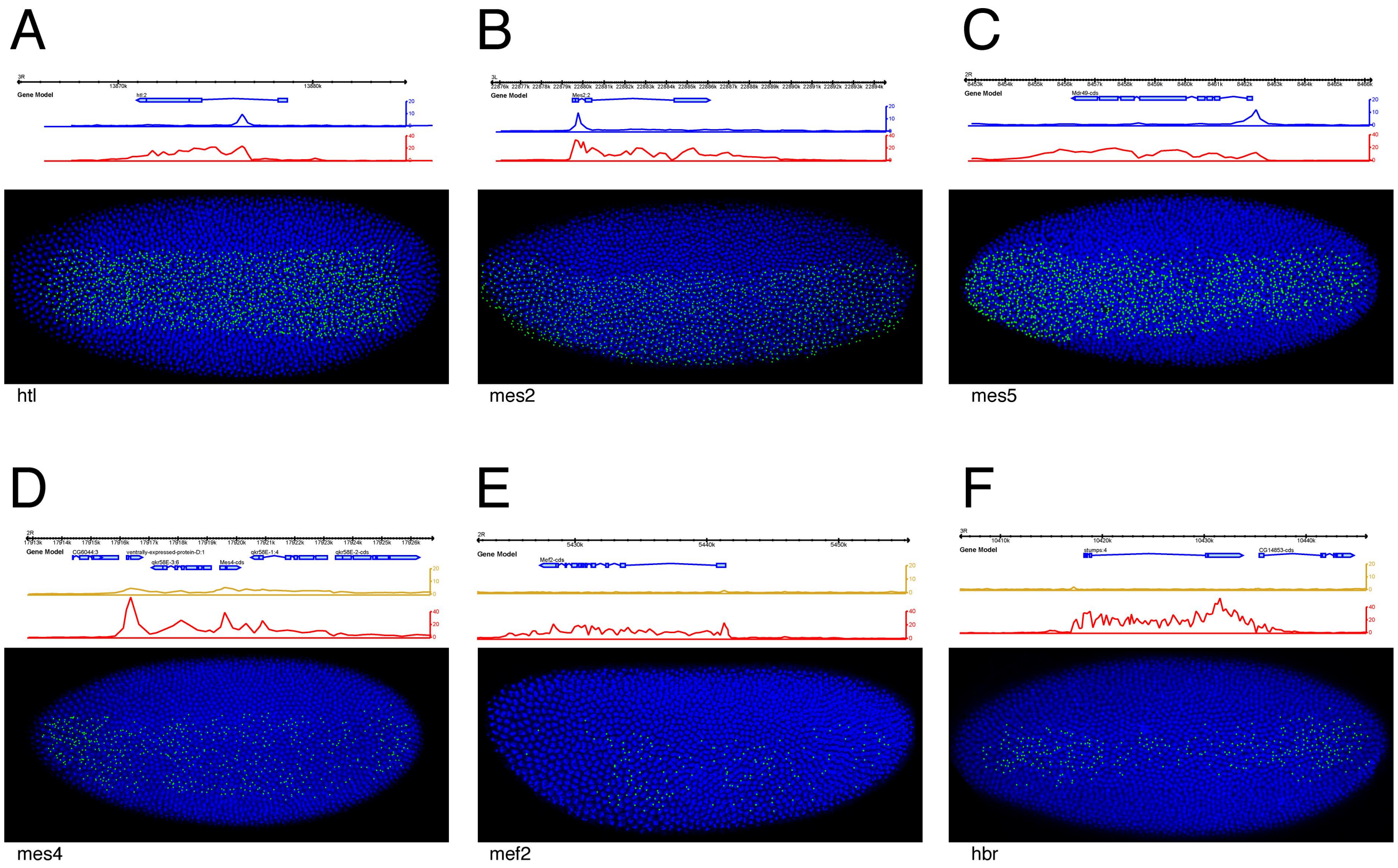

Figure 2.3: **Assaying asynchrony of gene expression in the mesoderm**. **A-F**, Representative confocal images (median activation state) of whole mount embryos in situ hybridizations against nascent transcripts (green), with counterstained nuclei (blue) are shown for each gene. The inset above each image shows the gene prediction models and PolII ChIP-chip binding data in dorsal (blue) lateral (yellow) or mesodermal (red) tissue. Genes which show Pol II peaks at the promoter in all tissues regulate transcriptional elongation i.e. stalled / paused genes (A,B,C) rather than promoter-polymerase recruitment non-stalled(D,E,F).

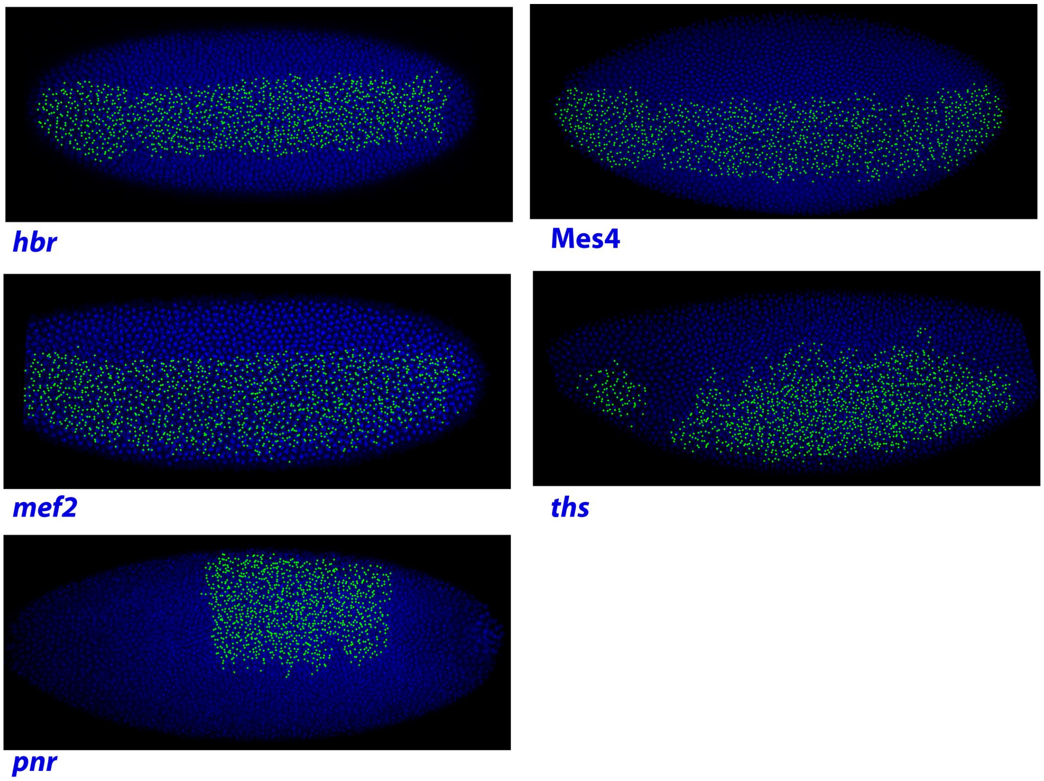

Figure 2.4: **Late expression of non-stalled genes is uniform**. When the restriction around the cephalic furrow is seen all of the non-stalled genes shown in this study have reached uniform patterns of expression.

lacks stalled Pol II and exhibits stochastic activation within the neurogenic ectoderm (Fig. 2.5A), whereas Neu3 [152] contains stalled Pol II and exhibits nearly uniform induction profiles in the same region (Fig. 2.5B). pnr [128] and tup [40] display stochastic and synchronous patterns of induction, respectively, in the dorsal ectoderm (Fig. 2.5C,D).

The degree of synchrony in expression for a given gene is inferred from the frequency of early cleavage cycle 14 embryos showing partial expression (summarized in Fig. 2.5E). The population statistics includes the analysis of an average of 50 embryos for each gene. Embryos containing a subset of nuclei with hybridization signals are represented by white or blue blocks, while uniformly activated embryos are represented by black blocks. The genes are sorted from least synchronous to most synchronous. Following this analysis we see that all of the genes lacking stalled Pol II (Mes4, *hbr*, *pnr*, *Mef2* and *ths*) clearly segregate from the synchronous genes, which contain stalled Pol II (*htl*, *sog*, Neu3, *tup*, Mes5, *vnd ush*, *rho*, and Mes2) (see Supplement for embryo stains and ChIP-chip results not shown here). All of the nonstalled genes show a higher degree of asynchronous activation than any of the stalled genes. The window of time between the detection of the first nuclei with hybridization signals until the detection of the last-activated nuclei is very short for stalled genes (2-3 min or less). In contrast, the variation in activation times for genes lacking stalled Pol II is much larger (15-20 min or more), resulting in the observed stochastic expression profiles.

Ectopic induction of Synchronous or Stochastic Expression

To determine whether the synchronous and stochastic modes of transcriptional induction are an intrinsic property of the promoters rather than enhancers, we examined the activation of 7 different genes in embryos containing an ectopic anterior-posterior Dorsal nuclear gradient [66] (Fig. 2.6). This gradient arises from the localized expression of an activated Toll receptor at the anterior pole of transgenic embryos using the Bicoid 3' UTR [99]. The resulting gradient induces novel patterns of gene expression, including the activation of mesodermal genes in the presumptive head. Nonetheless, the trend observed for stalled and nonstalled genes in wild-type embryos is also seen in these highly abnormal mutants.

The stalled Mes2 gene exhibits uniform ectopic activation in head regions (Fig. 2.6B), while the nonstalled Mes4 gene displays stochastic induction (Fig. 2.6A). Similarly, Neu3 (stalled) is uniformly activated in middle body

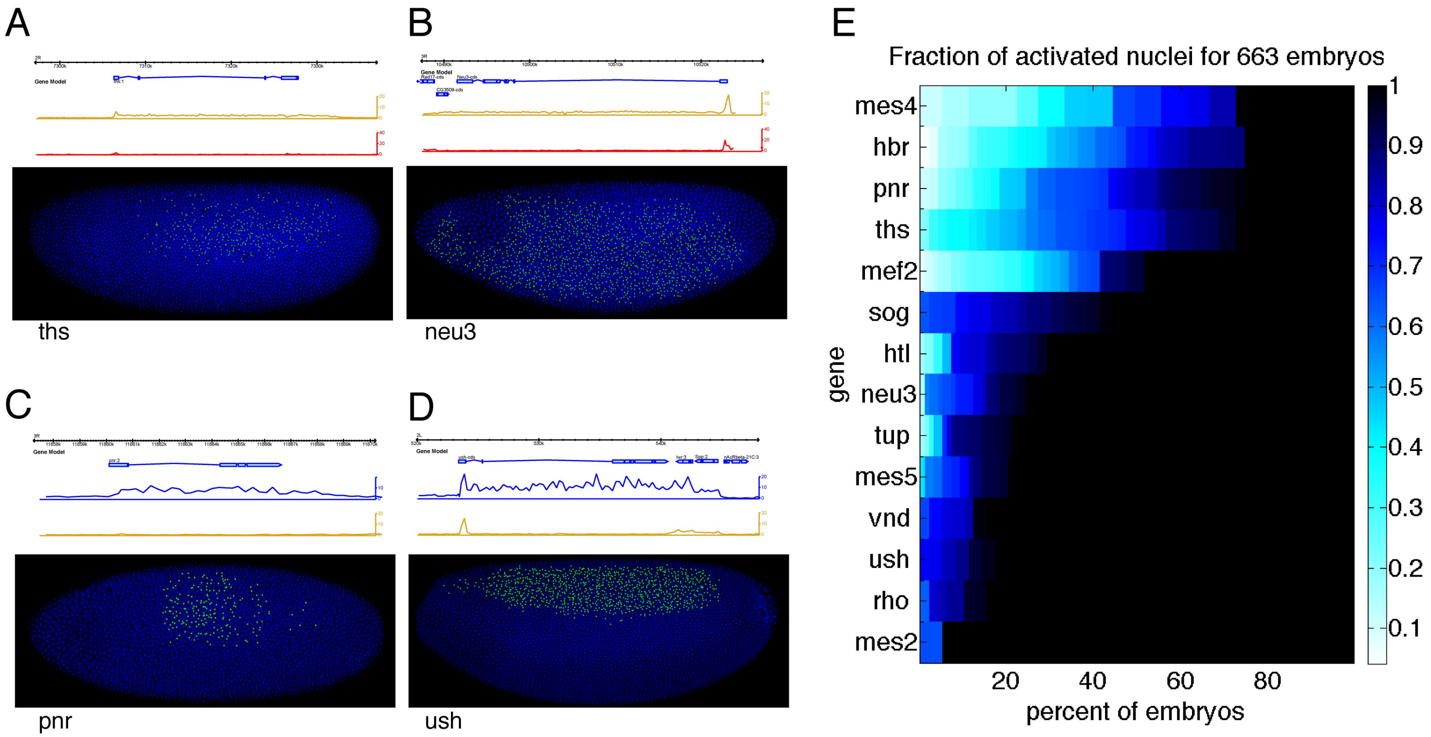

Figure 2.5: **Asynchronous expression among paused and unpaused ectodermal genes. A-D** Confocal images of nascent gene expression and corresponding Pol II binding-data for the lateral ectoderm (A,B) and dorsal ectoderm (C,D). **E** Heat map summarizing synchrony results for each gene. The color indicates the fraction of activated nuclei in the region of expression. The x-axis shows the proportion of similarly staged embryos with a fractional activation of at least n, [0,1]. The results are sorted top to bottom by the sum along the row, putting the most asynchronous genes at the top and the most synchronous at the bottom.

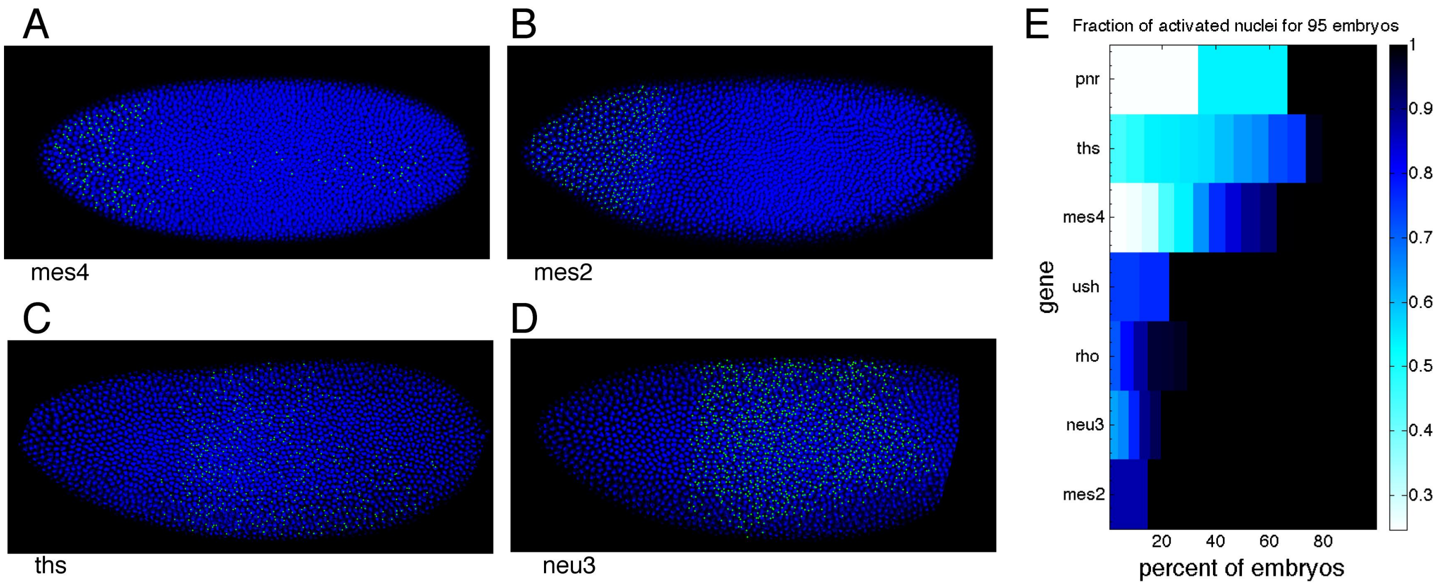

Figure 2.6: **Cell location and TF gradient do not dramatically affect synchrony.** **A-D**. Embryos containing an AP gradient of Dorsal activity due to toll[10b]-bcd-3′-UTR mRNA being expressed in the anterior pole under control of the maternal hsp83 driver [151] were assayed for asynchronies in gene expression. The ectopic activation under control of a differently shaped dorsal gradient still recapitulates the wildtype observation that stalled genes are activated in a more synchronous fashion. **E**. The results are summarized in the heat map as in Figure 2.5.

regions (Fig. 2.6D), while thisbe (nonstalled) exhibits stochastic activation profiles in the same region (Fig. 2.6C). The quantitative analysis of 95 different embryos suggests that all three nonstalled genes, pnr, thisbe, and Mes4, display stochastic activation by the ectopic Dorsal gradient. In contrast, the four stalled genes, *ush*, Neu3, *rho*, and Mes2, all display synchronous activation patterns (Fig. 2.6E). These results suggest that the two modes of induction correlate with the disposition of the promoter (stalled or nonstalled) rather than the activities of the enhancers, which are regulated quite differently in the mutant embryos as compared with wild-type [66].

There is no obvious correlation between transcriptional synchrony and threshold readouts of the Dorsal gradient in wild-type embryos. For example, both *htl* and *sog* display synchronous expression (Fig. 2.5E) even though *sog* is activated by 10-100 -fold fewer Dorsal molecules in the lateral ectoderm as compared with the regulation of *htl* in the mesoderm [63]. Perturbations in the Dorsal gradient would be expected to cause greater stochastic variation in the activation of *sog* than htl. To investigate this issue we examined embryos from *dl*/+ females which express a reduced Dorsal gradient [69].

Effects of Activator Concentration

A total of 7 stalled genes were examined in *dl*/+ embryos. Three of the genes (*htl*, Mes2, and Mes5/Mdr49) are activated by high levels of the Dorsal gradient, while the remaining 4 genes (*rho*, *vnd*, *sog*, and Neu3) are activated by low levels. Two of the three genes activated by high levels of Dorsal, *htl* and Mes2, display normal, synchronous patterns of activation in *dl*/+ embryos (Fig 2.7E,G). The third gene, *Mdr49*, shows a partial loss in synchrony (Fig 2.7F), particularly in posterior regions lacking the Bicoid activator gradient. *Mdr49* appears to be activated by Bicoid and Dorsal [107], whereas most mesodermal genes are activated by Dorsal and Twist. The latter genes are buffered to reductions in the Dorsal gradient since Twist expression is normal in *dl*/+ embryos (data not shown).

As expected, two of the Dorsal target genes that respond to low levels of the Dorsal gradient, rho and Neu3, show a marked loss of synchrony in *dl*/+ embryos (Fig 2.7A, C). In contrast, the other two genes, *vnd* and *sog*, display essentially normal patterns of synchronous activation in these embryos (Fig. 2.7B, D). Both *vnd* and *sog* contain shadow enhancers, secondary enhancers for a single Dorsal-dependent pattern of expression, whereas rho and Neu3 do not appear to contain such enhancers [64]. Shadow enhancers might compen-

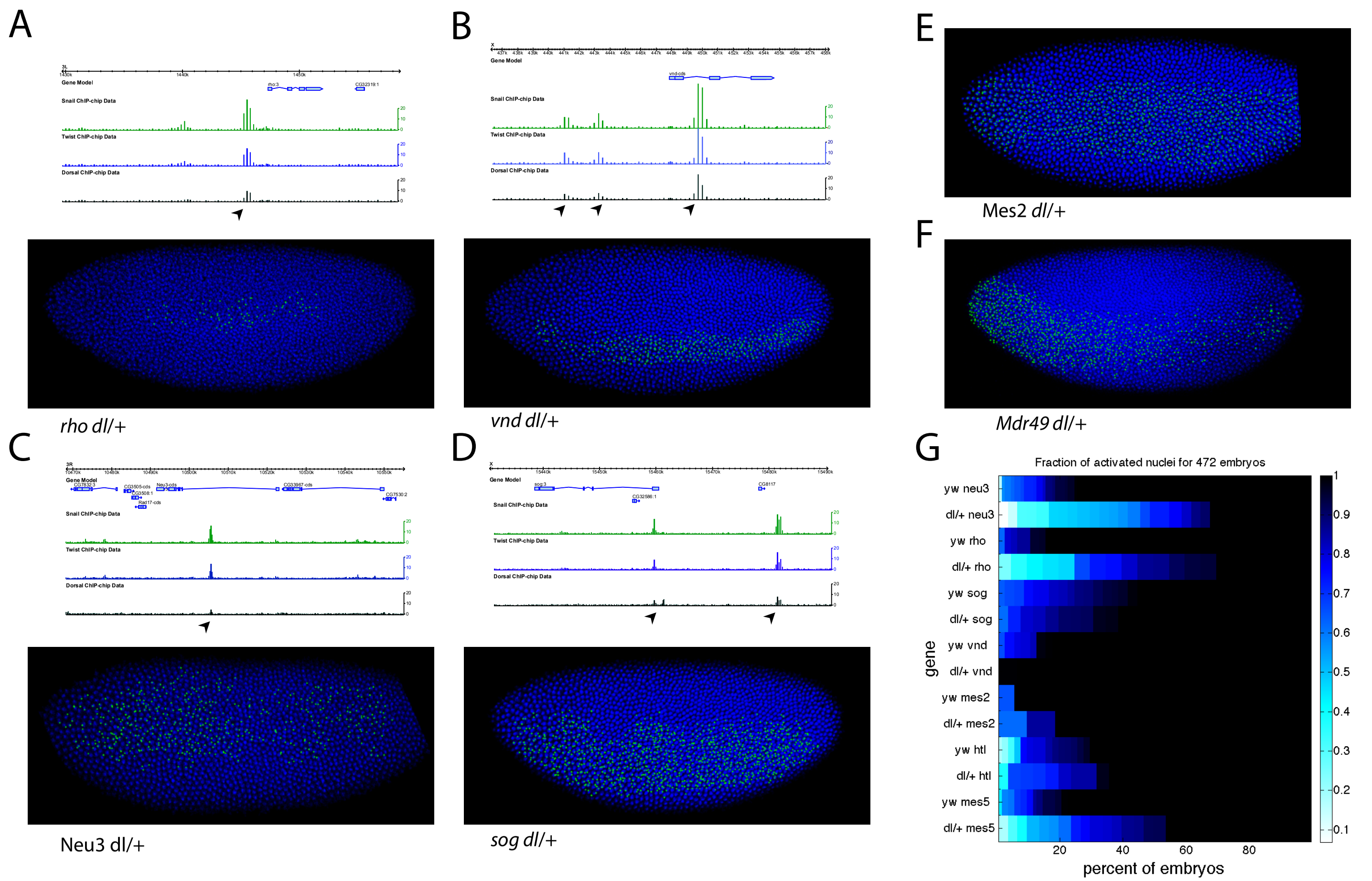

Figure 2.7: **Effect of reduced activator concentration A** The neurogenic ectoderm gene rho activates in a more stochastic fashion in Dorsal heterozygotes (*dl*/+) than in wildtype embryos. B. The neurogenic gene *vnd* has shows no detectable difference in the synchrony of expression in the *dl*/+ background. Multiple 5' enhancers of *vnd* (arrows) show similar binding of Dorsal Twist and Snail and reproduce the endogenous expression pattern [64]. *rho* has only one neurogenic ectoderm enhancer. **C**. Neu3 also shows increased evidence of stochastic activation in the *dl*/+ background. Neu3 is located in a gene rich region and regulated by a single intronic enhancer. **D**. *sog* expression shows no change in the *dl*/+ background, *sog* also has confirmed shadow enhancers which both drive the lateral ectoderm expression (29). **E-F** Effect on stalled mesodermal genes, Mes2 and *Mdr49*. **G** Summary of quantification of asynchronous expression for lateral ectoderm genes in wildtype (*yw*) and *dl*/+ embryos.

sate for the variability caused by local fluctuations in Dorsal concentrations by increasing the probability of occupancy of critical Dorsal binding sites [18, 51, 8, 11, 158].

2.3.3 Discussion

Despite 25 years of visualizing gene expression in the Drosophila embryo, previous studies failed to distinguish synchronous and stochastic modes of gene activation e.g., ([152, 100, 6, 106, 172, 68, 90, 53, 151, 40, 13, 39, 144, 171, 174]. This finding was made possible by the use of a quantitative method that examines gene expression across large populations of identical embryos rather than just a few individual embryos. Most of the developmental control genes that are active in the early embryo contain stalled Pol II and exhibit synchronous patterns of induction. Stochastic activation patterns are seen for relatively few genes engaged in dorsal-ventral patterning.

The lack of a synchronous, coordinated response of cells to patterning signals might present a challenge to proper cell fate decisions during development, particularly in the Drosophila embryo where these decisions are made quite rapidly (ie. minutes). Several fate decisions, such as the establishment of the ventral midline in Drosophila, involve distinct sender cells and receiver cells (Notch signaling). The sender population expresses a particular gene (Snail/Delta), which causes them to induce a different gene in the receivers (e.g., Sim). However, if some sender cells start expressing signaling genes before others, the late-activating senders might be converted to receivers, resulting in patterning defects (summarized in Fig. 2.8). In addition, synchronous activation might be used when the exact timing of expression is important, as in the case of pairs of genes that must maintain a balanced stoichiometry for proper function.

We propose that Pol II stalling and transcriptional synchrony helps ensure the orderly unfolding of the complex genetic programs that control development. It is likely that any given gene, or even small sets of genes, can be activated in a stochastic fashion without causing severe patterning defects. However, the reproducible and reliable development of large populations of embryos might be incrementally augmented by the acquisition of stalled Pol II on critical developmental control genes. Thus, synchrony might be a measure of population fitness. It might be difficult to demonstrate the importance of synchrony for any given gene, but we believe that if several key developmental control genes were activated in a stochastic fashion then there would

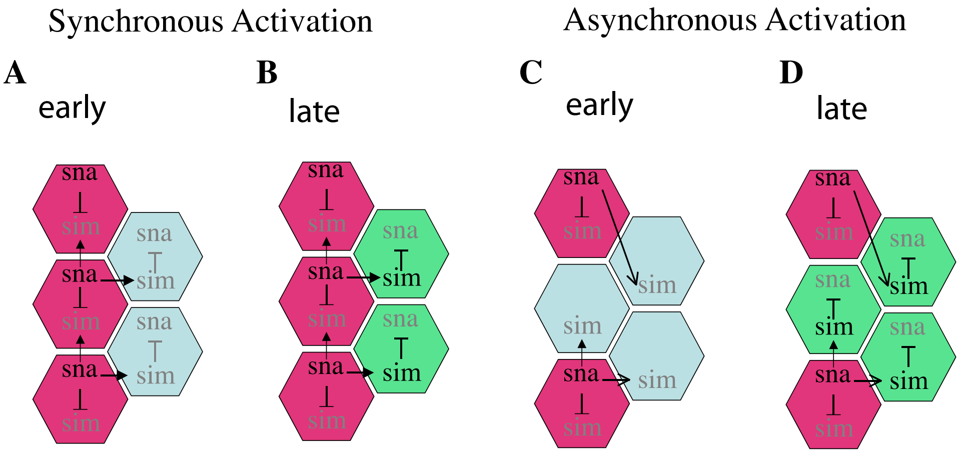

Figure 2.8: Effect of asynchronous activation among sender receiver cell populations. **A**. A sender population of mesoderm cells expressing *sna* signal through the Delta-Notch pathway to induce sim expression in neighboring cells. *sna* inhibits sim expression, preventing Delta presenting sender cells from becoming receiver cells. **B.** This leads to a single uniform line of sim expressing cells presumptive neurogenic cells adjacent to the *sna* mesoderm cells (7). **C.** Asynchronous induction of *sna* may lead to sim activation in late-responding mesoderm cells. **D.** Activation of sim leads to differentiation that prevents a normal mesoderm fate from being realized. This produces patterning defects.

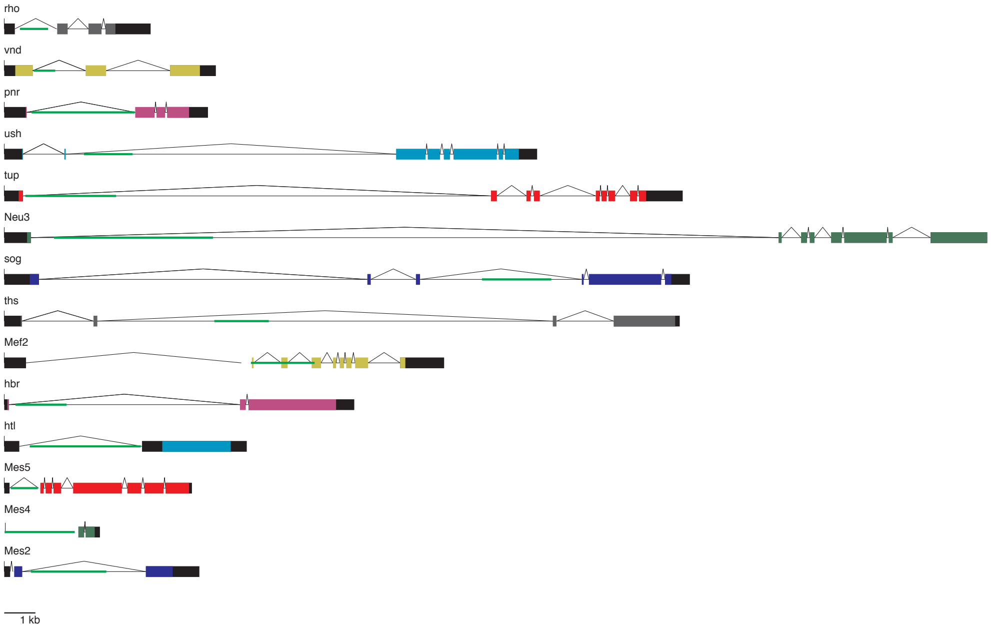

Figure 2.9: Probes for studying gene activation. Green bars represent region tiled with short antisense probes. Two to five unique probes were created for each gene and the brightest most reliable used for quantitative experiments.

be a diminishment in fitness due to stochastic errors in cell fate specification.

2.3.4 Methods

Construction of probes and embryo staining

Probes between 1 and 5 kb in length were carbonate-treated to fractionate into small probes and hybridized to 1-4 hour old fixed yellow-white embryos, as described in Kosman et al 2004 [84]. Two to five unique probes were made for each gene to ensure only clear, bright, and reproducible staining. All probes were made with digoxygen-haptens, stained with a sheep anti-dig primary antibody and Alexa Fluor 555-donkey-anti-sheep secondary antibody (Molecular Probes).

Quantitative confocal imaging and automated image analysis

Embryos were imaged on a Leica Scanning Confocal microscope with a 20x objective. Confocal photographs were taken at 1024x1024 resolution with approximately 500 nm/pixel. 10 independent z-sections from each embryo were photographed in 1 ?m intervals to find any active transcripts in the nuclei. Nuclei counter-stained with Draq5 (Biostatus, Leicestershire UK) were simultaneously imaged. All embryos were imaged near the beginning of nuclear cycle 14 when their expression first becomes significant. One exception is *sog* whose expression affects development by nuclear cycle 12, and was imaged in this cycle. Embryos from mid to late cycle 14 were also photographed to demonstrate that the probe staining itself is not stochastic and does reliably label all nuclei if the embryo was old enough to have activated them all. Approximately 50 embryos were photographed for each gene at the earliest detected moment of gene expression during the desired cycle.

We wrote an automated image segmentation program in MatlabR2008b (Mathworks) to identify and count the stained nuclei and detected probes. The script applies a space-filling algorithm to the expression staining data to determine the region of expression. Essentially, this algorithm uses the observed staining pattern to determine the boundaries of the expression region and counts all nuclei within the region. The ratio of detected transcripts to total nuclei counted in the expression region determines the measured parameter n the fraction of activated nuclei. If nascent transcripts are detected within every nuclei within the region, n=1. The script is implemented in a custom designed graphical user interface to allow user supervision and ensure appropriate classification at each step.

Quantification of Experimental Errors

Analysis of mid stage cleavage-cycle 14 embryos, (where we expect all nuclei in the expression domain are successfully induced) allows us to quantify our uncertainty in detecting activated nuclei. We find a mean uncertainty of 10% for each of our probes. Due to complexities of the dorsal boundary of the pattern, the uncertainty is slightly greater for *sog* and Neu3, with mean errors around 15%.

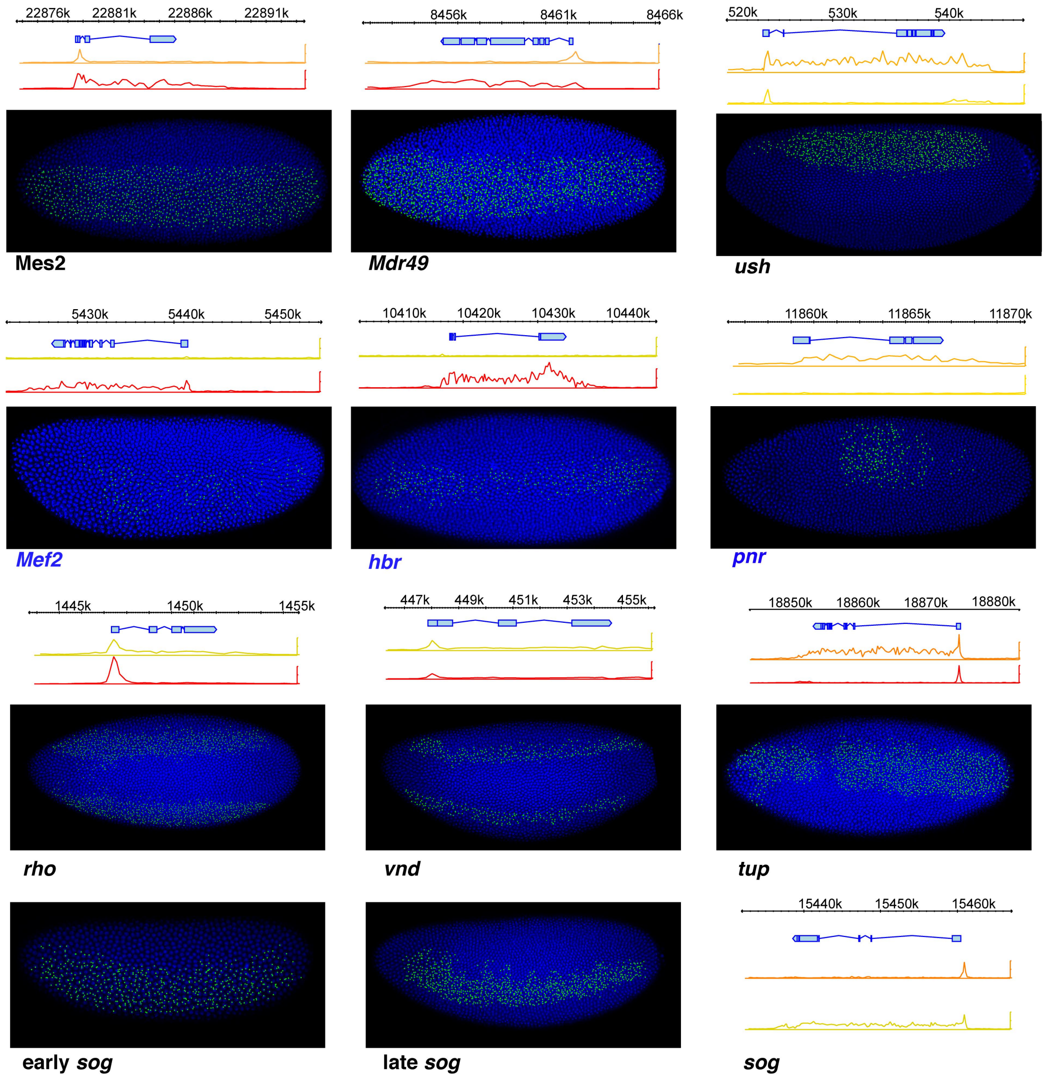

Figure 2.10: Representative in situ images for genes not shown in above images. Above each embryo, tissue specific ChIP-chip polymerase binding data and gene models are shown. Red traces are from presumptive mesoderm tissue (Mes2, Mdr49, Mef2, hbr), yellow are from presumptive lateral ectoderm tissue (*rho*

, vnd and sog*) and orange traces are from presumptive dorsal ectoderm tissue (*ush*, pnr, tup). Stalled gene names are in black, non-stalled are in blue.*

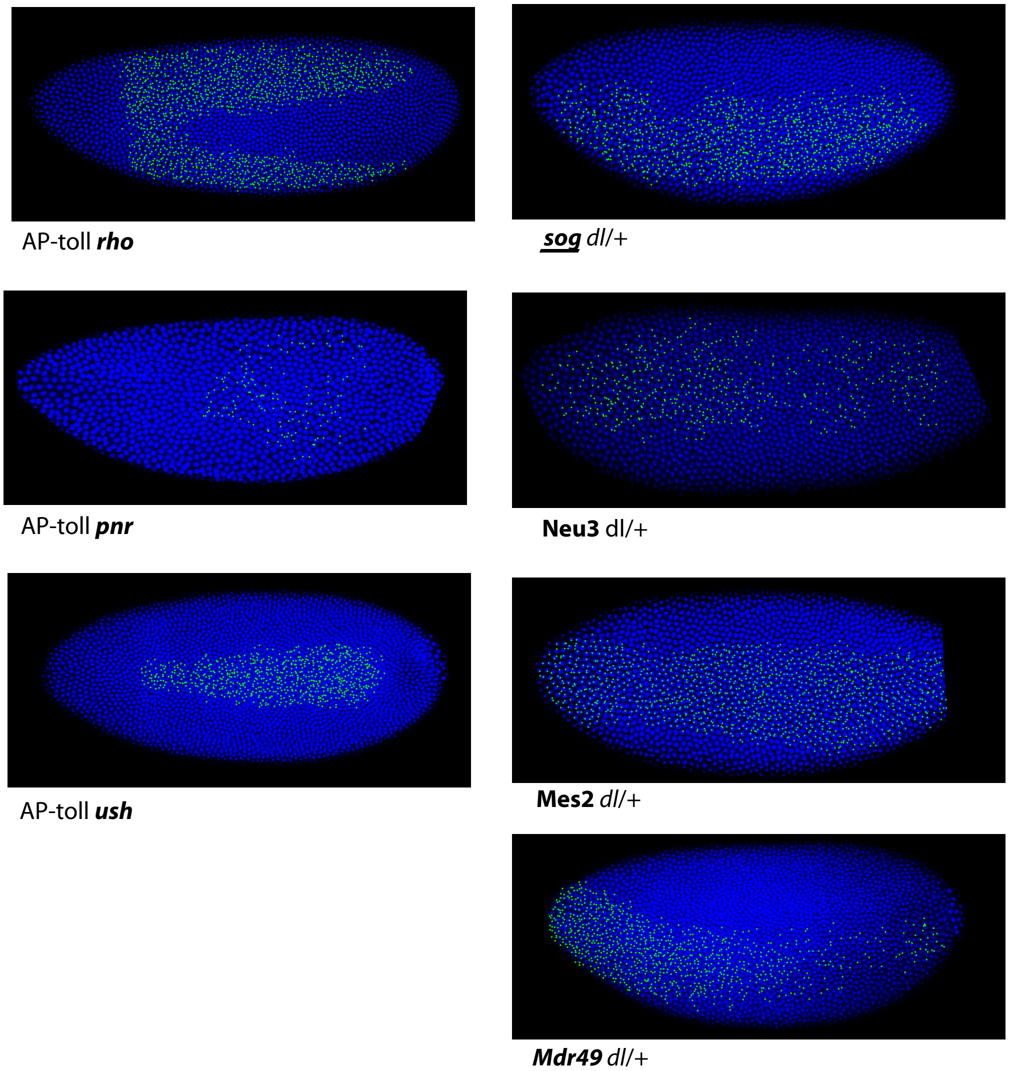

Figure 2.11: Representative in situ images for genes in mutant backgrounds not shown in above images.

2.4 Role of GAGA-binding factor (Trl) in Pausing and Synchrony

It has been proposed that GAGA binding protein may play an important role in promoter-proximal pausing. Paused genes are enriched for GAGA motifs [61]. The Lis lab has shown that removing GAGA sites in the hsp70 promoter is sufficient to convert it to a non-paused gene [88]. Whole-genome CHiP-chip assays for GAGA factor confirmed that GAGA factor is preferentially associated with genes previously annotated as paused [86]. Despite this correlation, several clearly paused genes lack GAGA binding and several clearly non-paused genes have considerable amounts of promoter proximal GAGA binding.

To further test for possible functional dependency of promoter proximal pausing and the correlated synchronous expression of these genes, I asked if the activation profile of any of the genes studied above would change in response to reduced levels of GAGA factor (which in Drosophila is well known as *trithorax-like*, Trl). Analysing embryos from mothers heterozygous for a null allele of Trl (gift from Vivek Chopra), I repeated the analysis of the synchrony of gene activation (see figure 2.12).

Of the paused genes in our study, Mes2, *rho*, *ush* and *pnr* all have promoter proximal peaks of GAGA binding, and *sog* has a weak cluster. Activation of *rho* is considerably more stochastic in this background than among wildtype embryos, suggesting that Trl may play an important role in either the correct establishment or release of paused polymerase (see fig 2.12). A more subtle change was observed Mes2. No change was observed for *sog* or *ush* however, suggesting that they are less sensitive to Trl concentration or that the Trl binding does not play a substantial role in the synchronous response of these genes. None of the genes which lack Trl ChIP-chip peaks showed any significant change in synchrony of gene expression in the Trl mutant background, suggesting that for those where we do observe a change the effect may be direct. Given that GAGA factor binds over 1500 genes [86], this is a reasonably useful control.

Taken together, we see a similar correlation of GAGA factor finding and synchrony as has been previously reported for GAGA factor binding and pausing. GAGA factor is not necessary for synchronous expression, genes both with and without GAGA that are expressed in a synchronous fashion in the wildtype embryo are still expressed in a synchronous fashion in the mutant background with reduced concentration. Nor is GAGA sufficient for synchronous expression, as genes which lack synchronous expression nonetheless have GAGA factor binding. Nonetheless those genes for which manipulations of GAGA does effect the synchrony of expression are both paused and have binding.

2.5 Further validations

From the original ChIP-chip data, it is impossible to distinguish whether a polymerase at the promoter is actually promoter proximally paused or in some transient, unstable loading state during preinitation complex (PIC) formation. (Though the tissue specific Pol-II ChIP do allow distinction of elongation regulated and initiation regulated.

The paused state of several of the genes in this study was confirmed by two post-docs in our lab. Joung-woo Hong tested several genes by "bubble assay," a technique which uses $KMnO_4$ to react exposed thymidine residues in the open transcriptional bubble of transcriptionally engaged polymerase, showing that genes like *rho* are indeed paused (Hong unpublished data). A whole genome approach was pioneered by Vivek Chopra in collaboration with Leighton Core from the Lis lab. Vivek conducted tissue specific global-run-on assays (GRO-seq) [27]. This technique detects only transcriptionally engaged polymerases (transiently bound ones are prevented from initiation through sarkosyl treatment). Transcription is reinitialized upon adding more nucleotides including Br-UTP, which facilitates isolation of newly synthesized short sequences which are sequenced and mapped back to the genome with conventional genome profiling ChIP-seq techniques. With this approach most of the genes we had classified as stalled appear to promoter-proximal paused and those we classified as non-stalled to be non-paused in *gd[7]* mutant background (dorsal ectoderm). Two exceptions should be noted: *vnd* does not appear to be paused in *gd[7]*, though it does have a confirmed shadow enhancer, and *hbr* does appear be paused. The enhancers isolated for this gene however drive especially patchy expression, one of the early indications that enhancers also play a substantial role in synchrony and uniformity of gene expression [140].

Fig 2.13 shows the GRO-seq results for *sog* and *ths*, showing the large number of transcriptionally engaged polymerases only very near the promoter of the *sog* gene in embryos from *gd*[7] mothers that produce purely dorsalized

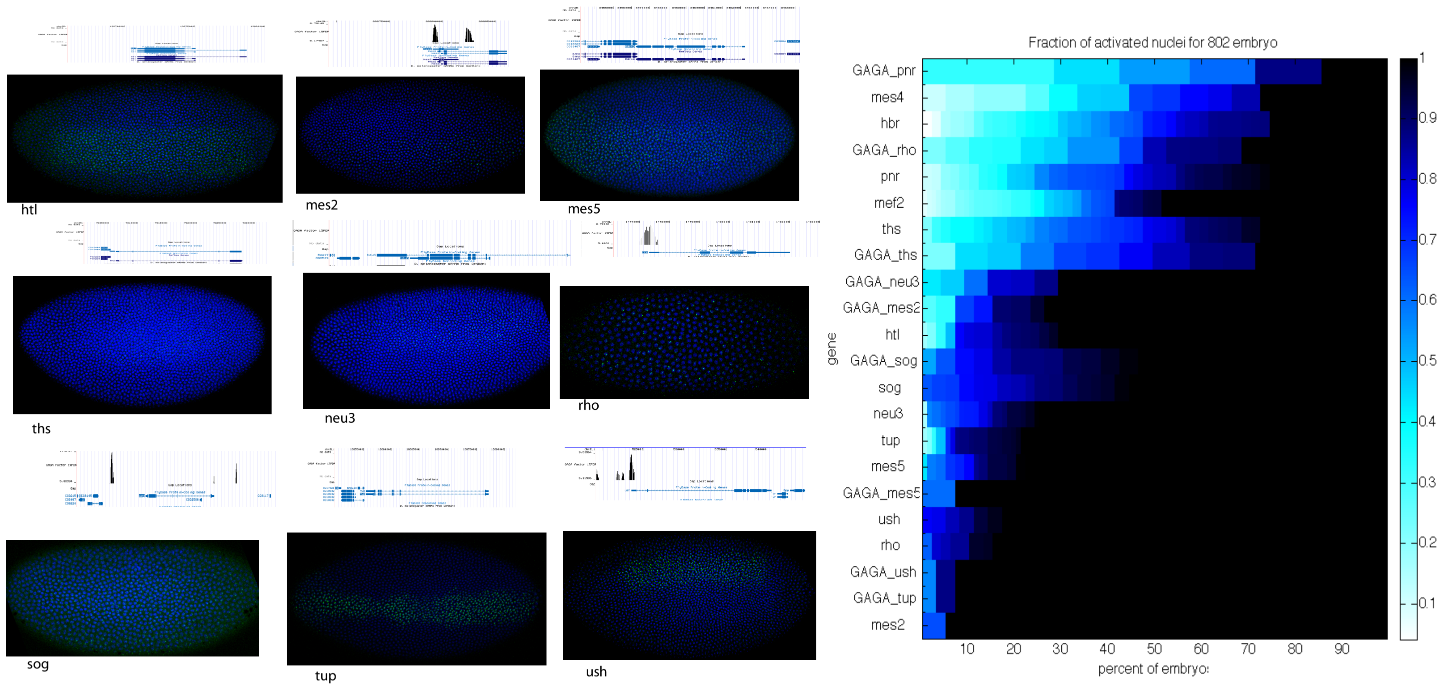

Figure 2.12: **Role of GAGA factor in synchronous expression A-H**, patterns of gene activation in embryos with only half the usual maternal contribution of GAGA factor. Above each panel is shown CHiP-chip binding data from Lee *et al* 2008 [86]. **I** Heatmap of the population of embryos activating each of the indicated genes in a GAGA heterozygote or wildtype background.

embryos in which *sog* is not expressed. In contrast the non-paused *ths* locus is completely unbound. Representative in situs for *sog* and *ths* are shown below, using an image rendering tenchnique of coloring the whole nuclei if it contains a transcription foci, originally developed by Jacques Bothma [119].

With this technique we can also create clearer images of the transcriptional activation of other paused an unpaused genes, as shown in figure 2.14. This enhanced rendering would prove especially useful for visualizing subtle differences created by modifying regulatory sequences in my subsequent imaging studies. By this point I had also managed to get reliable double stains working (primarily by increasing the washing steps to reduce background in the green channel, and by tiling longer intronic fragments with probes to improve the signal to noise). This allowed for a more direct comparison of the timing of synchrony of of activations of two genes in the same embryo. For the stains shown in Fig 2.13 and 2.14 are taken in the same embryo comparing paused and non-paused genes in the same tissue and same embryo.

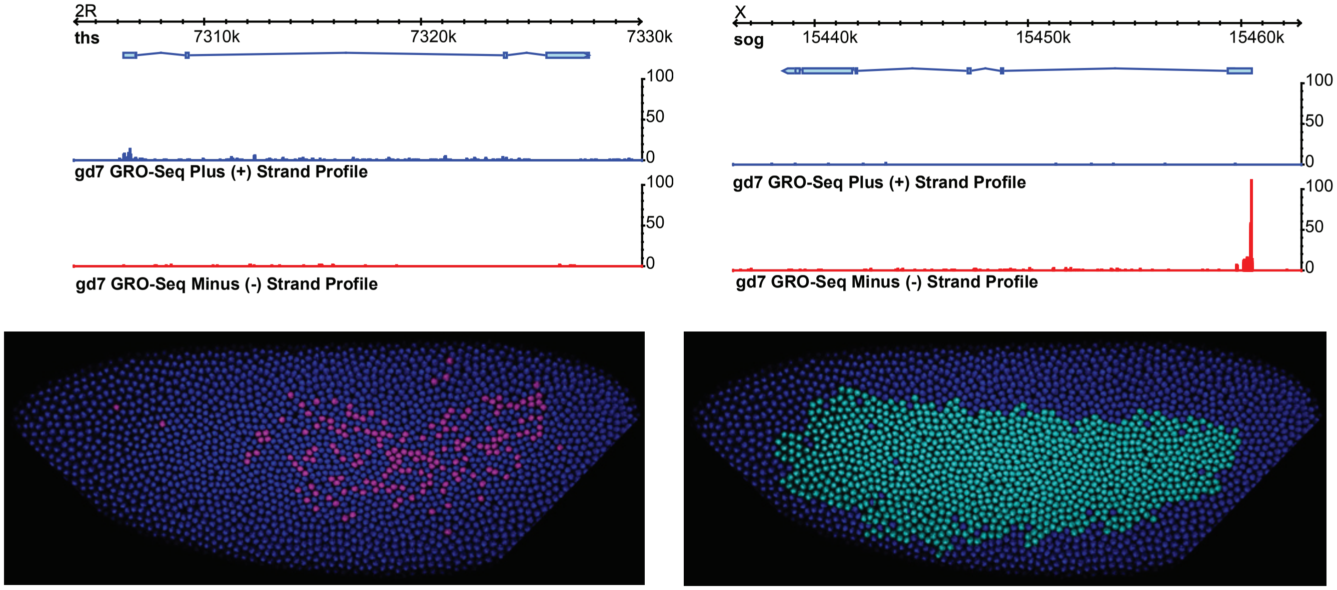

Figure 2.13: **Upper panels:** genome browser view of GROseq Pol-II competence data from the plus-strand (forward reading here) and minus-strand (backward) for dorsal ectoderm tissue for the genes *thisbe* (left) and *sog* (right). Neither gene is expressed at substantial levels in the dorsal ectoderm tissue, as shown in the upper panels by the very low read frequencies around the gene. Note however that large amounts of competent Pol-II are found at the promoter of *sog*. This is clear evidence of elongation-regulated paused polymerase state in a tissue where the gene is un-induced. **Lower panels:** Quantitative *in situ* hybridization to detect activated nuclei. Embryos are oriented anterior left, dorsal up. Note that appreciable gene activity is only observed in the lateral ectoderm, not in the mesodermal tissue (bottom of embryo) or dorsal tissue (top of embryo). Note as well the uniform pattern of early expression for the gene *sog* contrasted with the more heterogeneous expression of *thisbe*. The vertebrate names of each of these famous genes is indicated in parentheses.

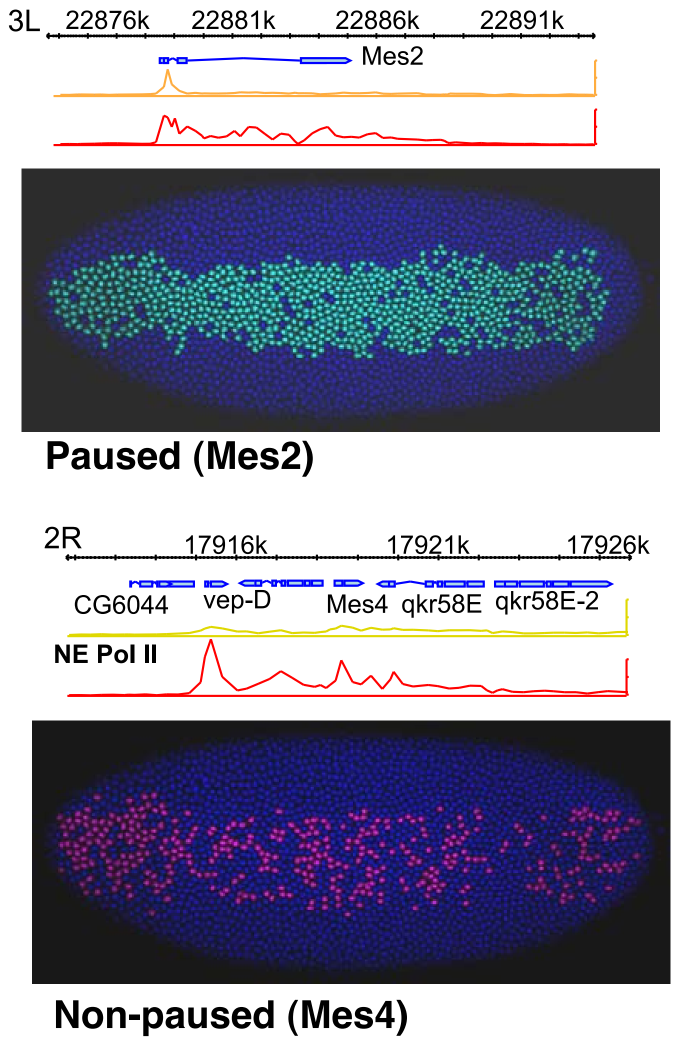

Figure 2.14: Computational labeling of whole nuclei corresponding to transcriptional state provides a clearer picture of the fraction of transcriptionally silent nuclei. Shown above are gene models with corresponding Pol II ChIP-chip binding data in mesodermal tissue (red) and non-mesodermal tissues (yellow).

Chapter 3

Understanding Synchrony through Mechanistic Modeling

Having observed a rather striking correlation between genes which are regulated by release of a promoter proximally paused polymerase and a more rapid, and synchronous activation of gene expression, I asked if this change could be causal. The most salient feature of this change is really order of the regulation – controlling the polymerase and allowing transcription of the gene to proceed only after it has engaged the transcript, rather than the controlling access to promoter from the start. One can make intuitive arguments why this mode of regulation should be faster (polymerase binding at the promoter can get a head start if it doesn't have to wait for enhancer activation first). However some of these intuitive arguments turn out to be wrong in general (as we shall see in this chapter).

To make a more rigorous study of this subject and began with some derivations models and simulations of my own. Finite state models of transcription seemed like they should be analytical solvable rather than just simulate-able. It turns out they are, though this required considerable help and though from two wonderful probability theorists with whom I first discussed the problem in Fall 2007 and which we continued to work on together for the following four years. At the end of this investigation we had not only a deeper intuition about the effect of changing the regulated step, but also a general analytic framework and toolbox of code for analyzing macromolecular assembly processes that meet certain conditions commonly found in biological assembly processes. In this chapter I relate those findings. This presentation adapted from the manuscript "Transcriptional regulation: Effects of promoter proximal pausing on speed, synchrony and reliability," written in collaboration with Dr. Peter L Ralph and Dr. Steven N Evans as equal collaborators and author.

3.1 Introduction

Investigations in yeast [79, 122] led to the hypothesis that in most organisms the recruitment of polymerase to the promoter is the primary regulated step in the activation of gene expression [77, 103, 46, 22]. However, recent studies of multicellular organisms have revealed a diverse array of other regulatory strategies, including several types of post-initiation regulation [170, 108, 58]. Zeitlinger et al. [170] generated tissue-specific whole-genome polymerase binding data in *Drosophila* and showed that regulation of polymerase release from the promoter is widespread during development. Their data shows that some 15% of tissue-specific genes bind polymerase to their promoters in *all* tissues, even though each gene only allows polymerase to proceed through the coding sequence in a specific tissue. Differential expression of these genes is made possible by a *paused state* wherein a polymerase remains stably bound but precisely stopped a short distance from the promoter and awaits a regulated release that is only triggered in the appropriate tissue [170]. Finally, many metazoa have been shown to have, genome-wide, disproportionate amounts of polymerase bound at promoter regions as compared to coding regions [27, 54, 108, 170].

This mechanism has been called *promoter proximal pausing.* It should not be confused with the stochastic stalling of a polymerase as it transcribes, a phenomenon which has also been termed "polymerase pausing". Furthermore, there are distinctions to be made between: **stalled polymerase**, a polymerase which associates in a transient, unstable manner with the promoter but does not proceed into productive transcription; **poised polymerase**, a polymerase for which the association is stable but has not escaped from the promoter to begin transcription; and **promoter proximal paused polymerase**, a polymerase that completely escapes from the promoter but "pauses" in a stable, inducible state just downstream of the promoter. It is believed that most genes which have polymerase bound to their promoters in all tissues but expressed in only some tissues fall in the last category; this promoter proximal accumulation of pol II may indicate that regulation of pausing transitions is a general feature of metazoan transcriptional control.

There are promoter proximal paused genes that are not regulated by cell-specific signals [45]; however, we stress that we are concerned solely with transcription regulation and thus we do not consider this latter class of "unregulated" genes.

Other whole genome assays for polymerase binding have demonstrated that metazoans, from Drosophila to humans, have disproportionally large fractions of polymerase bound at promoter regions compared to coding sequences (5-30% of all genes in primary human lung fibroblasts) [27, 54, 108, 170]. This observation has been used to argue that a large fraction of these genomes are "paused". Without seeing polymerase occupancy for an uninduced and state of the gene, one can not be certain that these genes are indeed elongation regulated. This ambiguous state of polymerase enrichment at promoters has been variously termed "stalled", ""lingering", or "poised", though the evidence from *Drosophila* suggests that many of are in fact elongation regulated. We focus our analysis the effects of controlling gene expression at the pausing release state compared to the initial polymerase binding event.

It remains an open question why expression of some genes is controlled further downstream than others. Several groups have postulated that pausing may ready a polymerase for rapid induction [27, 108, 61]. (Here *induction* refers to the first time at which all the components required for expression of a particular gene become available, and *expression* is when transcription of the first nascent mRNA transcript begins.) To motivate this idea, the preloaded, paused polymerase is described as a "loaded gun" ready to shoot off a single transcript as soon as it is induced. Experiments with heat shock genes – the first class of genes for which paused promoters were identified – show evidence of rapid induction consistent with this idea [168, 131]. However, pre-loading only provides an argument for why the *first* transcript would be produced more quickly. Surprisingly then, it was also observed by Yao et al. [168] that *subsequent* polymerases are recruited rapidly to promoters of induced, elongation-regulated genes as well as the first, preloaded Pol II – a phenomenon not accounted for by the loaded gun metaphor. Since most genes must be transcribed several times in order to produce functional levels of mRNA, changes in speed of induction as a whole are likely to be of more physiological consequence than changes in the time at which the first, pre-paused transcript releases.

When whole-genome studies extended the observation of pausing to cover many key developmental regulatory genes [170], further questions arose. While the selective advantage of rapid induction is reasonably apparent for stress response genes, it is harder to explain why rapid induction would be selected for in so many developmental transcription factors and signaling pathway components. An additional hypothesis, suggested by Boettiger and Levine [17], is that regulation of transcriptional elongation (for instance, by promoter proximal polymerase pausing) may have evolved to ensure more coordinated expression across populations of cells. This hypothesis was motivated by the striking correspondence between genes shown experimentally to activate in a synchronous fashion and genes shown to bind polymerase at the promoter independent of activator state but not continue elongation until activator arrival.

Recent work by Darzacq and colleagues [30] provides insight into why a regulatory interaction downstream of transcriptional pre-initiation complex (PIC) assembly may lead to more coordinated gene expression than does regulation upstream of PIC assembly. Using fluorescently tagged transcription components, they demonstrated that transcriptional initiation is a highly variable process, with only about one in ninety Pol II–gene interactions leading all the way to productive mRNA elongation [30]. Nonproductive interactions each lasted between several seconds and a minute, suggesting that abortion of transcriptional initiation can occur at different stages in assembly of the complex. Regulatory interactions that occur after this noisy assembly process would act only on transcriptionally competent polymerases, and so this mechanism might result in more synchronous expression – a hypothesis we test here.

The idea that gene expression itself is intrinsically variable (rather than variable as a result of extrinsic fluctuations in upstream quantities) is well established and is a recent focus of theoretical and experimental interest – see [125] and [126] for reviews. Stochasticity can arise at many stages of the process, including from the diffusion of molecules in the cell [160], noisy gene regulation [116], chromatin and other conformal rearrangements [34], random events during elongation [127, 134], and random dynamics of translation and degradation of mRNA and proteins [133].

Populations of single-celled organisms have been shown to take advantage of noisy gene expression to achieve clonal yet phenotypically heterogeneous populations [98]. In metazoan development, however, proper growth and development generally relies on coordination and synchrony rather than stochastic switching. For example, certain cells in the Drosophila embryo are induced to become neurons if they are next to a mesoderm cell but not mesoderm themselves [33], so uneven activation of mesoderm fate could produce early patches of mesoderm, thereby improperly inducing neuronal development in neighboring tissue. Although synchronous behavior is important for metazoa, particularly in development, it is not a universal property of all metazoan genes. For instance, genes with both synchronous and very stochastic patterns of induction have been observed in the Drosophila embryo [17]. The unique challenges of coordinating the behavior of a large number of independent cells may explain why elongation regulation aimed at release from a paused state appears to be much more dominant among metazoa like *D. melanogaster* and humans than *E. coli* or *S. cerevisiae*.

Here we investigate mathematically whether the significant change in the coordination of expression observed in experiment [17] can be explained by a change in the regulation network topology which only effects whether regulation occurs before or after PIC assembly, while keeping other details (reactions and rates) of the PIC assembly process the same. We also seek to determine which interactions in the transcriptional pathway are most important for determining the coordination of expression, and what effect different topologies have on the speed of induction and variability between sister cells in total number of mRNA synthesized.

We do this by constructing continuous-time Markov chain models of PIC assembly with states that correspond to joint configurations of the promoter and the enhancer. The (random) time taken for the chain to pass from a "start" state to an "end" state corresponds to the elapsed time between successive transcription events. The models we construct for the two different modes of regulation have a common set of transition rates, but the particular mode of regulation dictates that certain transitions are disallowed, resulting in two chains with different sets of states accessible from the "start" state. We describe this situation by saying that each model is a *topological rearrangement* of the other. Because the same set of transition rates completely parametrize both chains, (see figure 3.1) we can make meaningful comparisons between the two models. Once the Markov chains are constructed, we use the Feynman–Kac formula [38], model-specific decomposition techniques and computer algebra to find symbolic expressions for features of these first passage times that correspond to the delay between induction and transcription.

Although there has recently been much work modeling different sources of stochasticity in gene expression, most models refrain from a detailed representation of the different protein–DNA complexes involved in favor of more abstract approximations [11, 117, 118, 155, 156, 158, 98]. Two–state "on–off"

Markov chains have been used many times to model stochasticity in transcription (e.g. [116, 7]), and provide analytic solutions. Such models have been used to explain, for instance, the observation that mRNA copy number does not in general follow Poisson statistics, implying that there are “bursts” of transcription in some sense. This bursting behavior can occur if the gene transitions between an *active* state (in which transcription can occur), and an *inactive* state (in which it does not), as shown by Raj et al. [123]. Although more complicated Markov chain models have appeared, often presented via a stochastic chemical master equation [139], they are usually simulated rather than studied analytically (see [132] for a review of methods and software). A notable recent exception is Coulon et al. [28], who use matrix diagonalization to study the power spectrum and other properties of several models of regulation. A complementary set of techniques takes a broader view, using the fluctuation–dissipation theorem to work on the scale of *small* stochastic deviations from the differential equations that capture the average behaviors at equilibrium [11, 117, 118, 156, 158].

We model the intrinsic noise of regulation and polymerase recruitment using biologically-derived Markov chain models. We focus on this particular piece of the larger process of expression in greater detail than has been done previously in order to provide a detailed mathematical investigation of the role of promoter proximal pausing. Unlike simulation methods, our approach provides a tractable way to compute analytic expressions for which interpretation is direct and reliable. Moreover, it does not depend on small-noise or equilibrium assumptions, or require the passage to a continuum limit. Furthermore, the structure of the models we use is determined by biological realism rather than being constrained by mathematical tractability. Our approach is most similar to that of [28], although our methods are less computationally intensive and produce symbolic expressions which allow us to investigate phenomena in greater depth. In particular, we compare alternate modes of gene regulation and readily evaluate analytically the sensitivity of system properties to changes in rate parameters over a large proportion of parameter space.

3.2 Methods

3.2.1 Framework for modeling regulatory interactions

As a prelude to describing the actual Markov chain model of transcriptional regulation we analyze, we describe a general approach to modeling promoters, enhancers and their interactions, and illustrate this approach with a toy model of transcription that is not too cumbersome to draw – see figure 3.1.

We begin with two separate Markov chains, a *promoter chain* and an *enhancer chain* (figure 3.1A). The states of the promoter chain are the possible configurations of the components involved in polymerase loading onto the promoter (e.g. "naked DNA" or "DNA–polymerase complex") and the allowable transitions correspond to the arrivals of these components, in whichever order is permissible by the underlying biochemistry. The states of the enhancer chain are the the components involved in enhancer activation (e.g. the binding of regulatory transcription factors to the appropriate cis-control sequence for that promoter).

Next, to model the regulatory interaction between enhancer and promoter, we designate a particular configuration of the enhancer as the *permissive configuration*, and specify a particular transition of the promoter chain as the *regulated step*. We require the enhancer chain to be in the permissive configuration for the promoter chain to make the transition through the regulated step and we assume that the enhancer remains in the permissive configuration as long as the promoter chain is downstream of that step. (The specification that the enhancer remains in the bound/permissive state while the process is downstream of the regulated step is not the only possible choice, but it is perhaps the most realistic.) We choose the regulated step according to the regulation mechanism that we are modeling.

The composite stochastic process that records the states of both the promoter and enhancer chains is our resulting Markov chain model of transcription. Varying the regulated step leads to alternative topologies for this chain. We stress that, as we change the choice of regulated step, the underlying promoter and enhancer chains remain the same. In particular, the same set of rate parameters are used in both schemes and they have the same meaning. This permits meaningful comparison of different methods of regulation. Two possible regulated steps, labeled "IR gated" and "ER gated", are shown along with the corresponding Markov chains in figure 3.1. Each possible configuration of the components of the transcription complex and associated enhancer

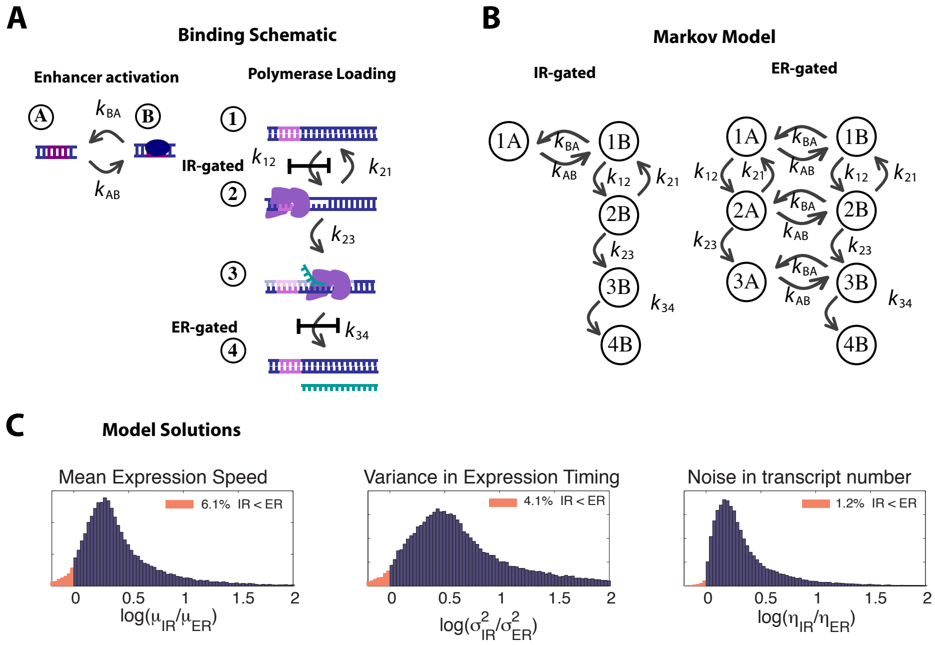

Figure 3.1: **From regulatory mechanism to Markov Chain: (A)** Schematics of two simplified schemes for initiation regulation (IR) and elongation regulation (ER). Transcription is represented in 4 steps: (1) naked DNA, (2) DNA-polymerase complex, (3) actively transcribing polymerase, and (4) completed mRNA. The enhancer is either (A) open or (B) bound. The enhancer must be bound (the permissive configuration) for the transcription chain to pass the gated step (*gate*), whose identity depends on the scheme (IR or ER). **(B)** The corresponding Markov chains for each regulation scheme. Colors of arrows denote the transition rates from (A). Note that one set of rate parameters determines all the numerical values for both chains, allowing for a direct test of the effects of topological change. **(C)** Distributions of log ratios of speed (μ), variance of expression time (σ), and transcript count variability (η) across 10,000 randomly chosen parameter vectors (as described in the text), showing that ER is faster, less variable, and produces less variability in transcript numbers over most possible combinations of rate parameters for this simple model.

elements is represented by a state of the composite chain, and the composite chain jumps from one state to another when a single molecular binding or unbinding event converts one configuration of complexes into another. For simplicity, we assume that each arrival in the end state allows one transcript to be made. After transcription occurs, the transcription complex may dissociate entirely, returning the chain to its initial state, or it may leave behind a partial *scaffold*, returning the composite chain to an intermediate state (and possibly leading to successive rounds of reinitiation and thus a "burst" of transcription products – i.e. multiple mRNA molecules being transcribed per promoter opening event).

Formally, the general composite Markov chain model is constructed as follows. Consider two promoter configurations, say, x_i and x_j, such that a direct transition from the first to the second is possible. Write $r_P(x_i, x_j)$ for the rate at which this transition occurs. For any two promoter configurations for which a direct transition is not possible, we set this rate equal to zero. Similarly, we write $r_E(y_i, y_j)$ for the transition rate from enhancer configuration y_i to enhancer configuration y_j. Denote the permissive enhancer configuration by y_*. Suppose that the regulated step of the promoter chain is the step from state x_a to state x_b. Let X^* be the set of states downstream from x_b, i.e. those states that can only be reached from the unbound state by passing through x_b. Then, the composite Markov chain takes values in a set of pairs of configurations (x, y), and it jumps from (x_i, y_i) to (x_j, y_j) at rate $q((x_i, y_i), (x_j, y_j))$, defined as follows:

$$\begin{aligned}
q((x_i, y_i), (x_j, y_i)) &= r_P(x_i, x_j), && \text{if } (x_i, x_j) \neq (x_a, x_b),\\
q((x_i, y_i), (x_i, y_j)) &= r_E(y_i, y_j), && \text{if } x_i \notin X^*,\\
q((x_a, y_i), (x_b, y_i)) &= 0, && \text{if } y_i \neq y_*,\\
q((x_a, y_*), (x_b, y_*)) &= r_P(x_a, x_b), &&
\end{aligned}$$

and $q((x_i, y_i), (x_j, y_j)) = 0$, otherwise. Denote by x_e the expressing promoter configuration with productively elongating mRNA. We are interested in the passage of the composite Markov chain from certain starting states – either the state in which both promoter and enhancer are unbound or the state to which the system returns after elongation begins – to the final, expressing state (x_e, y_*). Depending on which transition is regulated, some pairs of promoter and enhancer configurations will be unreachable from the relevant starting states; these pairs are biochemically inaccessible and are never visited, and so need not appear in our depictions or in our generator matrices (e.g. state 2A in the IR-gated scheme of figure 3.1).

Because there are generally only two promoters per gene active at the same time in a given nucleus, binding of a general transcription factor (TF) at one locus does not decrease the total concentration of the TF in the nucleus sufficiently to affect the rate of binding at the homologous locus. Furthermore, since the observed timescales of variability in induction are shorter than the expected timescale for protein translation and folding, we neglect any feedback from mRNA synthesis which might modify the transition rates. This allows us, in particular, to assume that the jump rates of the Markov chain are homogeneous in time.

3.2.2 Detailed model of transcription

We now apply this framework to examine a model of transcription that is more interesting and detailed than the toy model used above for illustrative purposes.

Many general transcription factors (TFs), such as the protein complexes TFIIA, TFIIB, etc., function together in a coordinated fashion to form the pre-initiation complex (PIC) necessary for the proper activation of transcription [56, 82, 157]. Experiments with fluorescently labeled TFs *in vivo* indicate that the components of this complex assemble on the promoter DNA [30, 148] rather than float freely in the nucleoplasm, as had been previously argued [115].

The steps of PIC assembly are not fully understood [56], although some important details are known. We analyze the assembly scheme depicted in figure 3.2, which is largely consistent with available data. The promoter is recognized by TFIID, the binding of which allows TFIIA and TFIIB to join the complex [157]. We choose this complex as the first state in our promoter model (state 1 of figure 3.2), since it is only just after this step that the regulation method may differ. TFIIB facilitates the recruitment of RNA polymerase II (Pol II) [157] (state 2). For many non-paused genes, polymerase is only detected in cells that have an activated enhancer (the cis regulatory sequence which controls expression) [170]. We call these genes *initiation regulated* and require that the enhancer reach its permissive state (B) before this association can occur. Since Mediator is important for many promoter–enhancer interactions [56, 42] it has likely also joined the complex prior to polymerase arrival. TFIIE, (state 3), and TFIIF (state 4), bind next, possibly in either order. Once both are bound (state 5), TFIIH must also bind (state 6) before Pol II starts synthesizing RNA and clears the promoter [82, 56]. TFIIH is displaced upon promoter escape [82], and if Ser 2 of the Pol II tail is not phosphorylated by CDK9 (pTEFb), transcription pauses 40–50 base pairs downstream of the promoter [131, 42, 143] (state 7). For elongation regulated genes, it is the release from this paused state that is possible only in the presence of an activated enhancer (permissive state) – which is generally believed to recruit the necessary CDK9 (and possibly other factors). Phosphorylation of Ser 2 allows the fully competent polymerase to proceed through the gene and produce a complete mRNA (state 8). The transition rates between configurations depend on the energy of association of the bond created and the concentration of the reacting components.

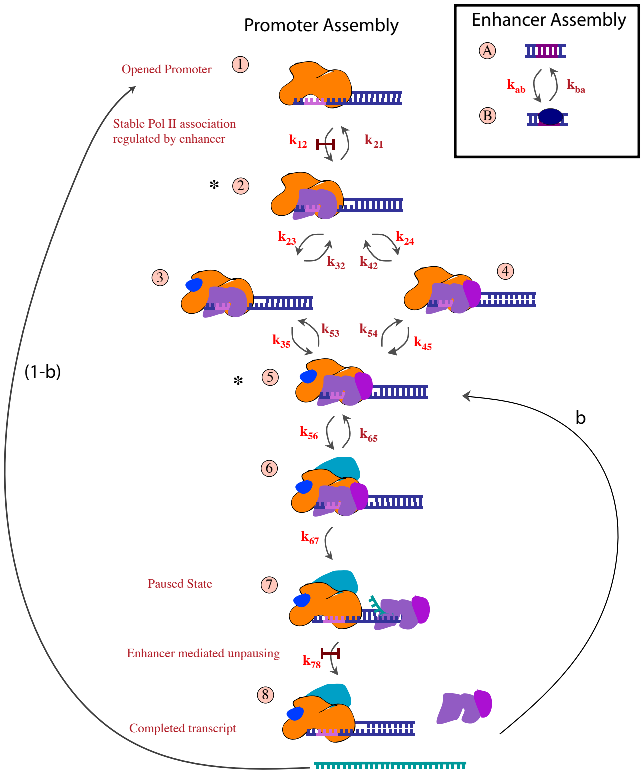

Figure 3.2: **Model of PIC assembly.** Each possible complex in the process is enumerated as a state of the promoter Markov chain. (see text for description of each complex) The promoter chain (states 1–8) is combined with the enhancer chain (states A and B) to make the full 16 state model of transcription. Transitions that in some scheme require an activated enhancer (state B) are indicated by a gate, *gate*. Forward rate transitions are in light font and backward transitions in dark font. The 1 → 2 transition is regulated in the IR scheme, and the 7 → 8 transition is regulated in the ER scheme.

Since we are interested in exploring the differences in which step of PIC assembly is regulated and not the different possible modes of enhancer activation, we use a simple abstracted two-state model of enhancer activation. A single transition switches the enhancer from the inactive state to the permissive state. For instance, a transition to the permissive state could represent the binding of a TF to the enhancer. This is not likely to be completely realistic, but if a particular step in the actual dynamics of transcription factor assembly and enhancer-promoter interaction is rate-limiting (e.g. the looping rate between a bound enhancer and its target promoter), then its behavior will be well approximated by our minimal model, with the transition from active to inactive corresponding to the rate for this limiting step.

For many paused genes, it is the phosphorylation event which is believed to be regulated [170, 42]. However, accumulating data suggests the molecular identity of the release factors may vary between paused genes. For example, some also require the recruitment of TFIIS in order to escape a "backtracked" paused state [2]. We consider any such regulation by release from pausing after PIC assembly to be *elongation regulation* (ER), and any regulation acting upstream of PIC assembly *initiation regulation* (IR).

Finally, the scaffold of transcriptional machinery that facilitates polymerase binding does not necessarily dissociate when transcription begins. Thus, reinitiation may occur by binding new polymerases (at step 5) which must still reload TFIIH which was evicted during promoter escape in order to proceed to step 6 and so on back to step 8. Repeated cycles of reinitiation may lead to a burst of mRNAs synthesized from a single promoter opening event. We denote by b the probability that the scaffold survives to cycle in a new polymerase (see figure 3.2). The scaffold breaks down before the next polymerase arrives with probability $1 - b$, in which case transcription activation must start again from state 1. We analyze both the time until the first transcript begins (for which such bursting is irrelevant) and the effect of this partial stability of the scaffold on cell–to–cell variation in total mRNA.

Our aim is not to present a definitive model of PIC assembly itself. Rather, we seek to understand the impact of different modes of regulation on a reasonable model that incorporates sufficient detail and to develop tools that can analyze effectively models of this complexity.

3.2.3 Statistical Methods

We are interested in the speed and variability of the transcription process, as measured, respectively, by the mean, μ_τ, and variance, σ_τ^2, of the delay τ between induction of the gene and expression of the first functional mRNA transcript. (Recall that by *induction* we mean the first time at which all the components required for expression of a particular gene become available, and by *expression* we mean the time when transcription of the first nascent mRNA transcript begins.) We use the mean delay to explore the hypothesis that the mechanism of elongation regulation is faster than that of initiation regulation, even when there is no polymerase initially bound (as reported in [168]). The variance of the delay is related to the degree of synchrony of expression of the first transcripts in a population of identically induced cells (studied in [17]) – allowing us to test if synchrony is a functional consequence of elongation regulation. We are also interested in the variation between activated cells of the total amount of mRNA produced in each. If we denote by $N(t)$ the random number of transcripts produced up until time t, then it follows from elementary renewal theory (see e.g. Section XI.5 in [36]) that $N(t)$ has mean approximately $\mu_{N(t)} \approx t/\mu_\tau$ and variance approximately $\sigma_{N(t)}^2 \approx \sigma_\tau^2 t/\mu_\tau^3$. A natural measure of relative variability of $N(t)$ is the squared coefficient of variation of $N(t)$, $\sigma_{N(t)}^2/\mu_{N(t)}^2$ (i.e. the variance of $N(t)$ divided by the squared mean of $N(t)$), which is thus approximately $\sigma_\tau^2/(\mu_\tau t)$. We denote the coefficient σ_τ^2/μ_τ by η, and refer to it as *transcript count variability.* The transcript count variability provides a measure of the variation in total number of rounds of transcription initiated by identical cells that have been induced for the same amount of time. Note that η has units of time:

$$\eta = \frac{\sigma_\tau^2}{\mu_\tau} \approx \frac{\sigma_{N(t)}^2}{\mu_{N(t)}^2} t.$$

However, the ratio of this quantity for the IR scheme to its counterpart for ER scheme does not depend on our choice of time scale. For any time t, this ratio is approximately the ratio of the squared coefficients of variation of $N(t)$ for the two schemes, and thus the ratio provides a way of comparing the relative variability in transcript counts between the two schemes across all times. Such a comparison is of interest because many of the known pausing regulated genes are transcription factors or cell signaling components that act in concentration dependent manners, and hence the precision of the total number of transcripts made directly affects the precision of functions downstream [17]. (Rather than the coefficient of variation, some authors consider the Fano factor of $N(t)$, defined to be $\sigma^2_{N(t)}/\mu_{N(t)}$ [155]. If $N(t)$ has a Poisson distribution, then its Fano factor is 1, and hence a Fano factor that differs from 1 indicates some form of "non-Poisson-ness". As such, the Fano factor capture a feature of the *character* of the stochasticity inherent in the number of transcripts made up to some time, whereas the squared coefficient of variation indicates the (relative) magnitude of the stochastic effects.)

We use our model to examine how these three important system properties – speed, synchrony, and transcript count variability – depend on the jump rates and how they differ between an IR and an ER regulation scheme. In both cases, the delay τ between induction and transcription corresponds to the (random) time it takes for the corresponding Markov chain to go from an initial state s to a final state f. For the chains corresponding to the models shown in figures 3.1 and 3.2, the moments of τ, the Laplace transforms of τ, and hence the probability distributions themselves, can be found analytically as we describe briefly here (for detailed discussion, see the next section).

Denote by Q the *infinitesimal generator* matrix that has off-diagonal entries q_{ij} given by the jump rate from state i to state j, and diagonal entries q_{ii} given by the negative of the sum of the jump rates out of state i. The infinitesimal generator of the chain *stopped when it hits state* f is the matrix $\widetilde{Q}$ obtained by replacing the entries in the row of Q corresponding to f with zeros. Writing $p(\cdot)$ for the probability density function of τ, the Laplace transform of p is

$$\phi(\lambda) = \int_0^\infty e^{-\lambda t} p(t)\, dt = (\lambda I - \widetilde{Q})^{-1}_{sf}. \tag{3.1}$$

In principle, the transform ϕ can be inverted to find p, as we do in figure 3.4D. Also, the n^{th} moment of τ can be found from the n^{th} derivative of ϕ:

$$\int_0^\infty t^n p(t)\, dt = (-1)^n \frac{d^n}{d\lambda^n} \phi(\lambda)\Big|_{\lambda=0}. \tag{3.2}$$

In particular, the mean and variance of τ can be computed from the first and second derivatives of $\phi(\lambda)$.

It is not necessary to carry out the differentiation in equation (3.2) explicitly, since (3.2) becomes

$$\int_0^\infty t^n p(t)\, dt = n! \sum_y (-Q_{-f})^{-(n+1)}_{sy} \widetilde{Q}_{yf} \tag{3.3}$$

after some matrix algebra, as derived in the section 3.5. Here, Q_{-f} is the submatrix of Q obtained by removing the final row and column. As shown in the Section 3.5, these expressions can be computed much more efficiently than (3.1) or (3.2).

Equation (3.1) is known as the Feynman–Kac formula [38], and it reduces our problem in principle to inverting the matrix $(\lambda I - \widetilde{Q})$. This is easy to do numerically for particular rate parameter values, but in order to make detailed general predictions about the consequences of changing the step at which the enhancer regulates transcription we require symbolic expressions for the system properties with the rates as free parameters. However, for even moderately complex chains like that described in figure 3.2, symbolic inversion of the matrix is prohibitively difficult for commonly available software.

To overcome this obstacle, we develop new analytic techniques that take advantage of the special structure of these matrices. First, we note that chains modeling transcription often have a block structure, in that we can decompose the state space according to the subset of states that must be passed through by any path of positive probability leading from the initial to the final state (we call such states *pinch points*) (see figure 3.2). (See section on analytical methods for more details) The models of initiation regulation we consider are amenable to this approach. In order for the ER scheme to be amenable to this approach, we assume that by the time the PIC assembly has reached the regulated step, the enhancer chain is in (stochastic) chemical equilibrium. Concretely, if π is the stationary probability that the enhancer is in the permissive state, then at each time the promoter chain jumps to state 7 (of figure 3.2) we suppose it jumps to state 7B with probability π and to state 7A with probability $(1 - \pi)$. (To evaluate the effect of this approximation, we investigate how our results change after removing the parameter vectors in which the enhancer chain is slow to equilibrate and hence when this approximation is the worst.) A similar decomposition for elongation regulated genes is possible using spectral theory, but the computational savings are not as great as for the pinch point decomposition. We provide a detailed description of these techniques and the accompanying proofs (plus implementations coded in MATLAB (see alistairboettiger.info, Software) in the Section 3.5.

Our approach has several advantages. Firstly, once we have derived symbolic expressions for features of interest, it is straightforward to substitute in a large number of possibilities for the transition rate vector in order to understand how those features vary with respect to the values of the transition rates. This would be computationally impossible using simulation and at best very expensive using a numerical version of the naive Feynman–Kac approach. Secondly, we are able to differentiate the symbolic expressions with respect to the transition rate parameters to determine the sensitivity with respect to the values of the parameters. It would be even more infeasible to use simulation or a numerical Feynman–Kac approach to perform such a sensitivity analysis.

3.3 Results

3.3.1 Predictions for representative parameter values

To get an initial sense of the differences between these two schemes of regulation, we first compared the transcriptional behaviors for a best-guess set of parameters, guided by measurements of promoter binding and escape rates by Darzacq et al. [30] and Degenhardt et al. [34] *in vivo* and observations in embryonic Drosophila transcription. These data do not allow us to uniquely estimate all 14 binding reaction rates in our model of PIC assembly, but they do constrain key properties, including the time scale of the rate-limiting reactions and the ratio of forward to backward reaction rates for both early binding events and later promoter engagement events. We chose parameters to be consistent with these measurements, and chose enhancer activation and deactivation rates to be consistent with induction times estimated in Drosophila [17] (which are also in the range recently reported in human cell lines [34]).

We used the following rate parameters for the model of figure 3.2:

$$\begin{aligned}&[k_{12}, k_{21}, k_{23}, k_{32}, k_{24}, k_{42}, k_{35}, k_{53}, k_{45}, k_{54}, k_{56}, k_{65}, k_{67}, k_{78}, k_{ab}, k_{ba}]\\ &\quad = [.108, .725, 10, 10, 10, 10, 10, 10, 10, .008, .005, 10, 10, 10, .01, 1]\text{sec}^{-1}.\end{aligned}$$

We found the probability density of the amount of time it takes the system to go from induced to actively transcribing, shown in figure 3.3A, by numerical inversion of the Laplace transform (equation 3.1). With these rate parameters, the mean time between induction and the start of transcription for an elongation regulated scheme is around 5 minutes, with a standard deviation of about 4 minutes, whereas an initiation regulated scheme with the same rate parameters has a mean of 16 minutes and a standard deviation of 12 minutes, consistent with experimentally estimated initiation times in Drosophila [17].

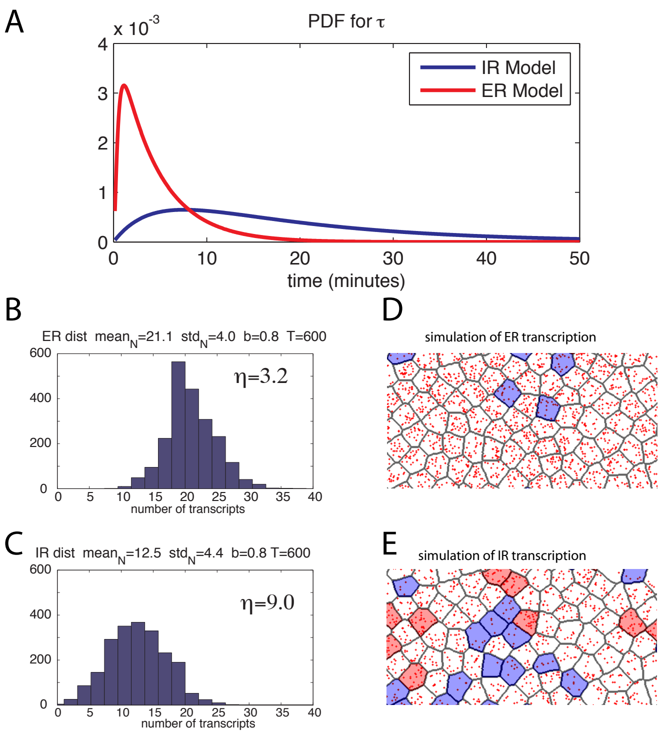

Figure 3.3: **Model Predictions: (A)** Probability distributions for first passage times: Probability density functions of the time to first transcription, obtained by inversion of symbolically calculated Laplace transforms, using rate parameters computed in experimental studies of particular transcription systems. Rates inferred from Darzacq et al. [30] measurements of promoter binding and promoter escape rates (see text). **(B)** Distribution of total transcripts among a population of simulated cells during 600 minutes of transcription under the ER scheme with parameters as in (A) and a reinitiation probability of 0.8. **(C)** as in (B) but for the IR scheme. **(D)** Individual cell simulation (see text) showing of the expected results for an mRNA counting assay on the population of cells plotted in (B). Each mRNA transcript is represented by a red dot randomly positioned within the cell. Cells with less than two-thirds of mean mRNA concentration are shaded blue, cells with more than three-halves of mean mRNA concentration are shaded red. **(E)** as in (D) but for the IR scheme.

We also described the number of mRNA produced over a given period of time at one choice of b (the probability the GTF scaffold dissociates before the return of the next polymerase). Setting $b = 0.8$, we found the distribution of the time delay between the beginning of the production of subsequent transcripts under each model. Using this distribution, we simulated the number of mRNA produced during a 600 minute period in 2000 independent cells, under both the IR and the ER scheme (for the common vector of rate parameters listed above). The resulting distributions of mRNA numbers are shown in figure 3.3B and C. To depict the amount of variability this represents, figures 3.3D and E show a cartoon of the results – for each cell pictured, we sampled a random number of mRNA as above, which are shown red dots randomly scattered within the cell. To emphasize the variability, we then colored cells blue that have less than two-thirds the mean mRNA number and colored cells red that have more than three halves the mean mRNA number.

In this example, η is 2.8 times larger in the ER scheme than in the IR scheme, so these simulations also give a sense of how a given ratio of transcript count variabilities η for the two schemes corresponds to a difference in cell-to-cell variability of transcript counts, a topic we explore in more detail below.

3.3.2 Effects of regulation scheme on expression timing

Our predictions for the time of expression and the number of transcripts in the previous subsection depended on the chosen parameter values such as the association rate of different GTFs and the average burst size of the gene expression. The values of such parameters can, for the most part, be only very approximately estimated. Moreover, they may be expected to vary considerably between different genes and different species.

Since a single vector of parameters simultaneously specifies our models for the two regulation mechanisms, we can systematically explore all possible combinations of promoter strength and enhancer activation rates and ask in each of these cases how the two mechanisms compare in terms of speed, synchrony and variability in transcript counts.

To compare the two kinds of regulation of the model in figure 3.2, we sampled 10,000 random vectors of transition rates and substituted them into our analytic expressions for μ_τ, σ_τ^2, and η, with each rate chosen independently and uniformly between 0 and 1 (we could also have used a regular grid of parameter vectors). Since we will use ratios of the relevant quantities to compare models, and these ratios are all invariant under a common linear rescaling of time, the fact that all rates are bounded by 1 is no restriction – we are effectively sampling over *all* of parameter space. (For instance, the ratio of mean expression times of the two models does not change after multiplying every rate parameter by 100.) Furthermore, independent draws of new sets of 10,000 parameter vectors and substitutions give nearly identical results, confirming that our results are not sensitive to the specifics of the sample. Additionally, discarding parameter vectors for which the enhancer dynamics are significantly slower than for the promoter chain (i.e. k_{ab} or k_{ba} is smallest) does not qualitatively change any of the results, validating our treatment of the enhancer chain when analyzing the ER scheme.

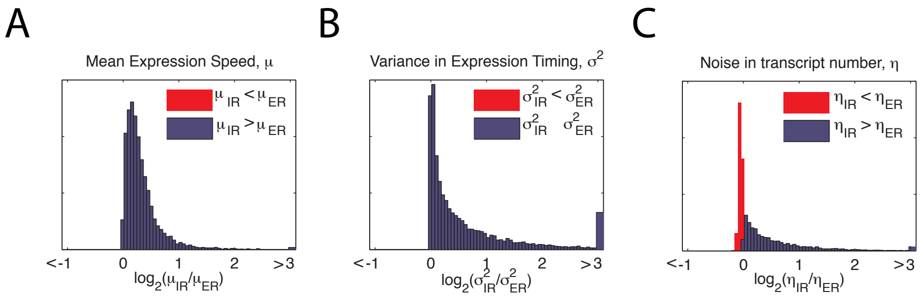

Figure 3.4: **Model Results: (A)** Comparison of log ratios of mean expression speed for the IR/ER schemes for 10,000 uniformly sampled rates. For all jump rates, the log ratio is positive (red line), indicating the ER scheme is always faster. Extreme values that would be off the edge of the graph are collected into the outermost bins. **(B)** Variance in timing of expression. **(C)** $\log_2$ ratio of noise in transcript number, measured by the squared coefficient of variation between cells of total mRNA counts $N(t)$ up to time t: $\sigma^2_{N(t)}/\mu^2_{N(t)}$ – the ratio is approximately independent of t.

In figure 3.4A–C we plot the histogram of $\log_2$ ratios for the mean delay, variance in delay, and transcript count variability for the 10,000 randomly selected parameter combinations sampled uniformly across parameter space. We found that at all sampled choices of rate parameter, and therefore in the vast majority of parameter space, the time to the first transcription event after induction is smaller and less variable (i.e. more synchronous) for elongation regulation than for initiation regulation in the realistic model of figure 3.2. Thus, both the experimentally reported speed [168] and synchrony

[17] for elongation regulated genes can be expected purely from effects of regulation topology without invoking changes in promoter strength or in the composition of the PIC.

We emphasize that this conclusion is still consistent with the possibility that a particular initiation regulated gene is expressed in a more synchronous pattern or with more rapid kinetics than some other elongation regulated gene: it is only necessary that the rate parameters are also sufficiently different. However, for the fixed set of rates associated with a given gene, the network topology of the ER scheme always improved synchrony and speed in our model of transcription relative to the corresponding IR scheme for the parameter vectors we sampled.

There is a plausible intuitive explanation for why elongation regulation is almost always faster than initiation regulation (figure 3.4A). When the regulation acts downstream, there are multiple paths which the system can take to before it reaches the regulated step – (i.e. either the enhancer can reach the permissive state first or the polymerase can load), as illustrated for the simple model in figures 3.1A and B. The system moves closer to the endpoint with whichever happens first, whereas the IR regulated scheme must wait for enhancer activation before proceeding. The combination of this intuition and our strong numerical evidence suggests a provable global inequality. However, recall that for the toy model IR is faster over about 6% of parameter space, and one can reduce the realistic model to the toy model by making appropriate transitions very fast. For example, for the toy model the choice of parameters

$$[k_{ab}, k_{ba}, k_{12}, k_{21}, k_{23}, k_{34}] = [1, 1, .1, .1, .1, .0001]$$

leads to a 5 fold increase in speed of the IR scheme relative to the ER scheme. This allows us to find parameter vectors where IR is faster than ER for the realistic model, for instance,

$$\begin{aligned}&[k_{ab}, k_{ba}, k_{12}, k_{21}, k_{23}, k_{32}, k_{24}, k_{42}, k_{35}, k_{53}, k_{45}, k_{54}, k_{56}, k_{65}, k_{67}, k_{78}]\\ &\quad = [.1, 1, .01, .01, .01, .01, .01, .01, .01, .01, .01, .01, .01, .01, .01, .0001]\end{aligned}$$

produces in the realistic model a 10 fold increase in speed for the IR scheme relative to the ER scheme. However, such reversals of the typical ordering must occur over less than one ten-thousandth of parameter space. The fact that the typical ordering is not universal and hence not the consequence of some analytically provable domination of one model by the other demonstrates the necessity of our numerical exploration of parameter space.

3.3.3 Effect of regulation scheme on mRNA concentration

The effect of the regulatory scheme on the variation in the total amount of expression among cells is perhaps the most interesting and also experimentally untested consequence of regulating release from the paused state. As discussed above, we compute a factor $\eta \approx (\sigma^2_{N(t)}/\mu^2_{N(t)})t$ for each scheme and compare the schemes by examining the ratio of the resulting quantities. If the ratio η_{IR}/η_{ER} is larger than one at a particular set of parameter values, a population of cells using the IR scheme with those rate parameters will show more variability in mRNA concentrations between cells (relative to the average over all cells) than if they were using the ER scheme with the same rate parameters. In this case, we say that the ER scheme is more *consistent* than the IR scheme.

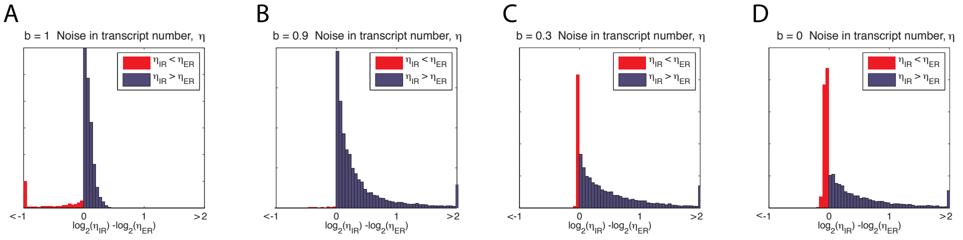

Figure 3.5: **Effect of scaffold stability** for variation in transcript number. **(A)** $\log_2$ ratio of transcript variability, η, between the IR and ER scheme when all subsequent polymerases engage an assembled scaffold $b = 1$. Extreme values that would be off the edge of the graph are collected into the outermost bins. **(B)** As in (A) when $b = 0.9$, note the ER scheme is more often substantially more coordinated, though a few parameters still make the IR scheme the more coordinated by a smaller margin. **(C)** $b = 0.3$. **(D)** $b = 0$.

We explored the logarithm of this ratio (equivalently, the difference of the logarithms of the respective η quantities) at four different values of b (the probability the scaffold does not disassemble; see figure 3.2); several of the resulting distributions are shown in figure 3.5.

When the complex is very stable, so that all polymerases find a pre-assembled scaffold to return to ($b = 1$, figure 3.5A), the ER scheme is more consistent for most rate parameters, but the differences are small. In fact, in nearly all cases at which η differs by a factor of at least 2, the IR scheme is the more consistent.

When the scaffold is still stable but less so ($b = 0.9$, figure 3.5A; mean burst size 10), the ER scheme still almost always produces more consistent numbers of transcripts among cells than the IR scheme, and the differences are much larger. If the scaffold is less stable ($b = 0.3$, figure 3.5C; mean burst size 1.4), the ER scheme is still more often more consistent than the IR scheme.

When we consider the simplest case with no bursting ($b = 0$, figure 3.5D), the ER scheme produces less variation in total transcript (smaller η) for most of parameter space. Moreover, the distribution is strongly skewed to the right, to the extent that for the 20% of parameter space where there is more than a 1.5 fold difference between the two regulatory mechanisms the ER scheme is always less variable.

We have found that, regardless of the value of b, the ER scheme is more consistent over most of parameter space. However, for that difference in consistency to be substantial, b must not be too close to 1. This is at first surprising, because if the scaffold remains assembled, so that the chain returns to state 5 of figure 3.2, an IR scheme seems to have a clear "advantage" – it does not have to wait for the enhancer to arrive, whereas the ER scheme does, and one might expect that this added stochastic event would only increase variability.

Consideration of how each chain depends on its starting state suggests an intuitive explanation for this difference. The IR scheme differs more in the amount of time it takes to reach the synthesis state when started with or without a scaffold (state 5 or state 1) than does the ER scheme. Intermediate values of b allow the possibility of some cells making many bursts by reverting to state 5 after each synthesis while other cells make dramatically less by reverting to state 1 after each synthesis. In contrast, under the ER regulation scheme, cells that start again from state 1 or from state 5 have relatively more similar synthesis times, and thus relatively less variation. The similar synthesis times result from the fact that ER is faster starting from state 1, for the reasons discussed above, and slower than IR when starting from state 5, because of the extra regulatory step before synthesis. Consequently, an ER scheme reduces the noise associated with very stable transcription scaffolds (see [117, 155, 158] for a discussion of this noise).

3.3.4 Pertinent properties of elongation regulation

To further understand why elongation regulation results in faster, more synchronous, and more consistent gene expression over a wide range of parameters we investigated alternative post-initiation regulatory schemes. This allows us to explore how changing certain properties of the model of PIC assembly (the promoter chain) will affect the results: Is the difference large because there are many steps between the IR step and the ER step, or is it because there is no allowed transition leading backward out of the state immediately before the regulated step? To explore these questions, we made modifications to the toy model of figure 3.1 which we are able to analyze without the assumption of enhancer equilibrium.

First note that, as is shown in figure 3.1C, the ER scheme is still faster, less variable, and more reliable (smaller μ, σ, and η) than the IR scheme over approximately 95% of parameter space. (It is also reassuring that the results are so similar to those for the more realistic model.)

We performed the same analysis after adding a reverse transition from state 3 back to state 2 (see figure 3.6A-B). The results are shown in figure 3.6C, and demonstrate that there is strikingly little difference between the two schemes of regulation. This suggests that the absence of a backwards transition from the state immediately preceding the regulated transition is an important factor in producing the differences between the schemes we observed above. In the ER scheme of figure 3.1, PIC assembly becomes "caught" in state 3, awaiting arrival of the enhancer. (Similarly, the ER scheme of figure 3.2 gets "caught" in state 7). After adding a transition $3 \rightarrow 2$, PIC assembly may run up and down the chain many times before it is in state 3 at the same time the enhancer is in the permissive configuration, and this counteracts any benefits in speed or reliability that may have been gained otherwise. (It is not obvious that this will happen: the ER scheme of figure 3.6B still has "more routes" from state 1A to state 4 than the IR scheme, so it may run counter to intuition that the IR scheme could be so often faster.) This furthermore suggests that regulating after a state in which PIC assembly is "caught" reduces variation – some polymerases may run from state 1 to 8 smoothly and fire very quickly, while others may go up and down the assembly process many times before they actually escape the promoter and make a transcript (as is suggested by the data of Darzacq et al. [30]), and this will substantially spread out the times at which the first transcript is created.

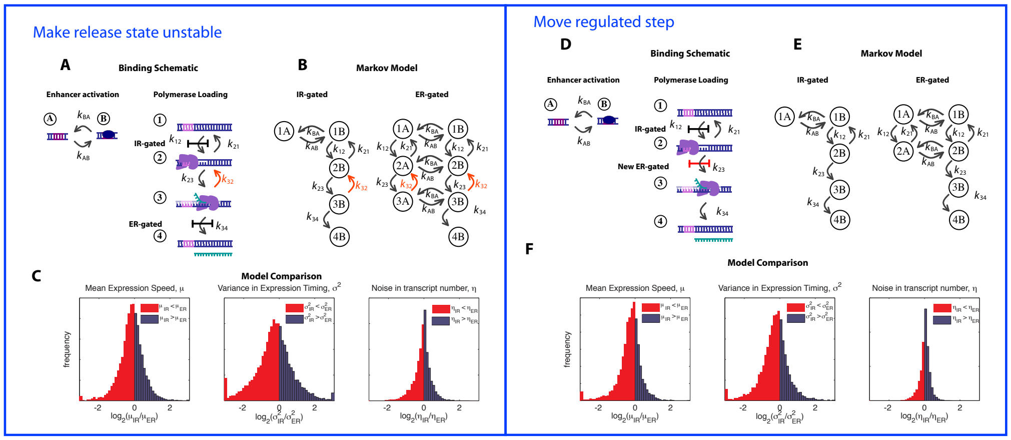

Figure 3.6: **Effect of regulated step.** **(A)** Adding a transition k_{32} which enables polymerase to exit the paused state and return to a pre-initiated state. **(B)** Effect of the added transition on the structure of the composite Markov chains. **(C)** Comparison between the models over all of parameter space when the transition k_{32} is added. **(D)** Schematic of changing the regulated step to control promoter escape rather than release from pausing. **(E)** Resulting composite Markov chains for regulating promoter escape. **(F)** Comparison between the models over all of parameter space when promoter escape is the regulated step.

We also investigated the case in which the $2 \rightarrow 3$ transition is regulated and observed a similar pattern – see figure 3.6D-F. This investigation supports the intuition that it is the stability of the paused state, not simply the parallel assembly of enhancer complex and promoter complex, that is most important in understanding the different behavior of the two regulatory schemes. It also suggests that these differences should be specific to genes that are regulated through paused (as opposed to poised or stalled) polymerase.

3.3.5 Sensitivity analysis

Small variations in rate parameters between cells will occur if the number of TF or Pol II molecules is small, so it is of interest to investigate how robust the properties of each regulation scheme are to such variation and which jump rates affect each scheme the most. To measure this sensitivity, we compute the gradient of a quantity of interest (e.g. the mean induction speed) with respect to the vector of jump rates, square the entries, and normalize so that the entries sum to one, giving a quantity we refer to as *relative sensitivity* that is analogous to the “percent variation explained” in classical analysis of variance. Our analytic solutions for the quantities of interest make this computation possible. For example, let $m(\mathbf{r})$ denote the mean transcription time of the chain when the vector of transition rates is $\mathbf{r}$. Then, the relative sensitivity of m to each rate r_i is $(\partial_{r_i} m(\mathbf{r}))^2 / \sum_j (\partial_{r_j} m(\mathbf{r}))^2$. The larger this quantity is, the larger is the relative effect a small change in r_i has on m.

To explore the sensitivity across parameter space, we computed relative sensitivities for each of the three system properties to all 16 parameters at each of the 10,000 random vectors of transition rates described above. Each of the system properties showed surprisingly similar sensitivity profiles, so we only discuss the results for the mean time to transcription. Marginal distributions of sensitivity of mean time to transcription to each parameter are shown in figure 3.7. Corresponding plots for the variance of transcription time and for transcript count variability are shown in figures S4 and S5.

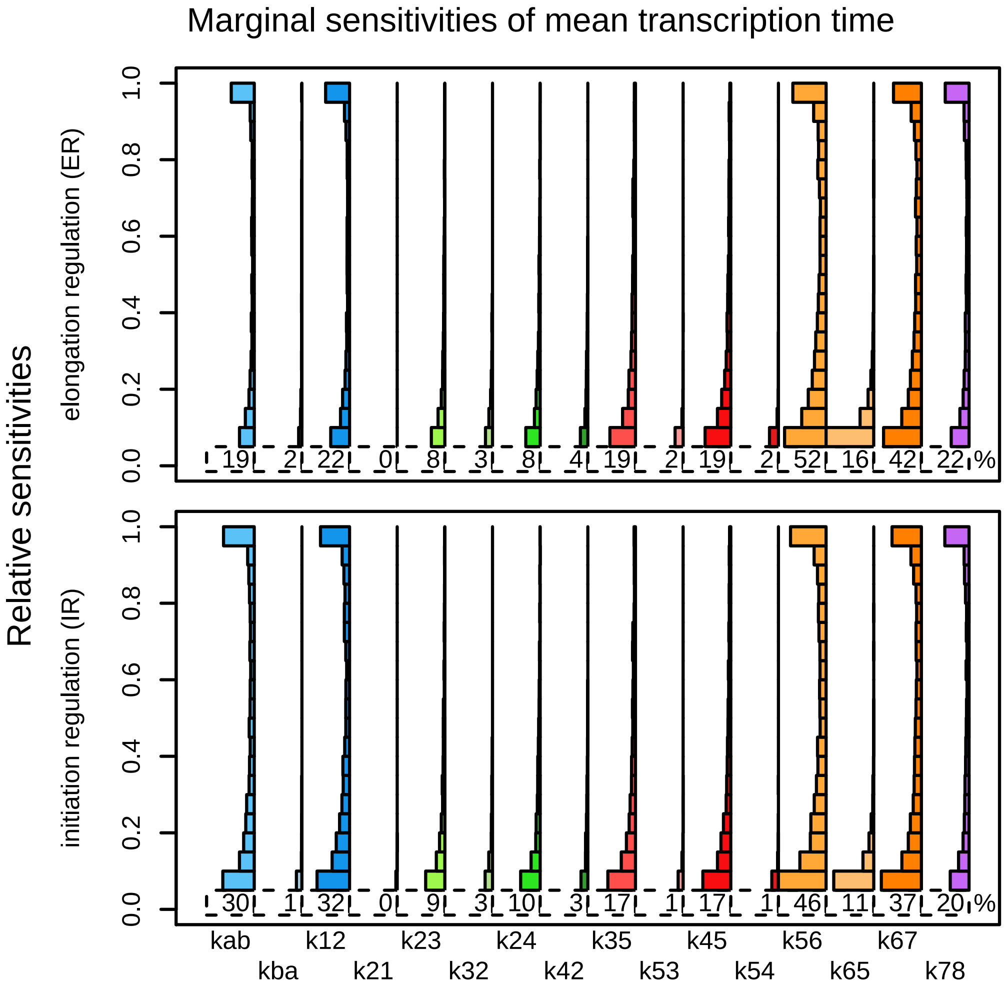

Figure 3.7: **Sensitivity Analysis** for mean expression time. Histograms of the marginal distributions of relative sensitivities for both the ER and IR schemes, across uniform random samples from parameter space, as described in the text. The smallest bin of the histogram (values below .05) is disproportionately large, and so is omitted; shown instead is the percent of parameter space on which the relative sensitivity is at least .05. Note that often only a single parameter dominates (many sensitivities are near 1), that many parameters are almost never influential, and that ER and IR are similar except for the addition of sensitivity to k_{ab}.

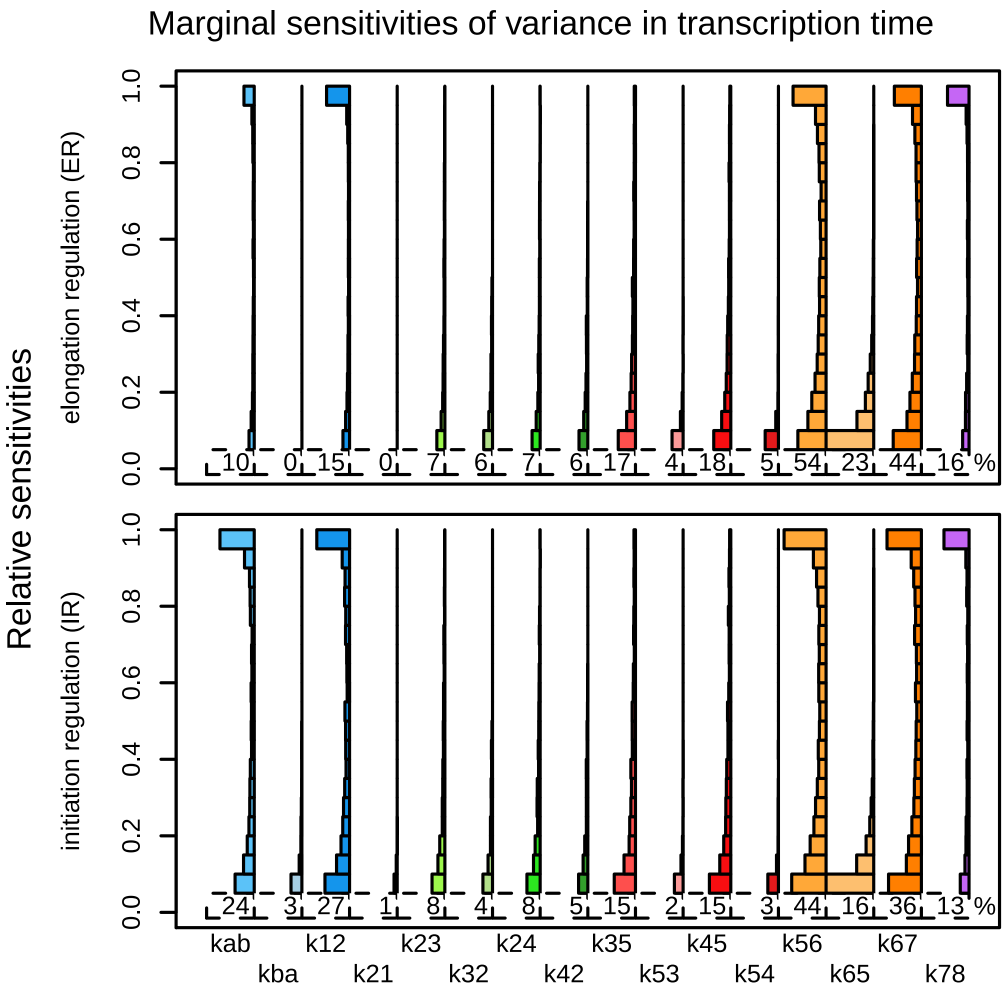

Figure 3.8: **Sensitivity analysis for variance in transcription time.** The details are the same as for Figure 3.7, except that the variance in transcription time is analyzed, rather than the mean transcription time.

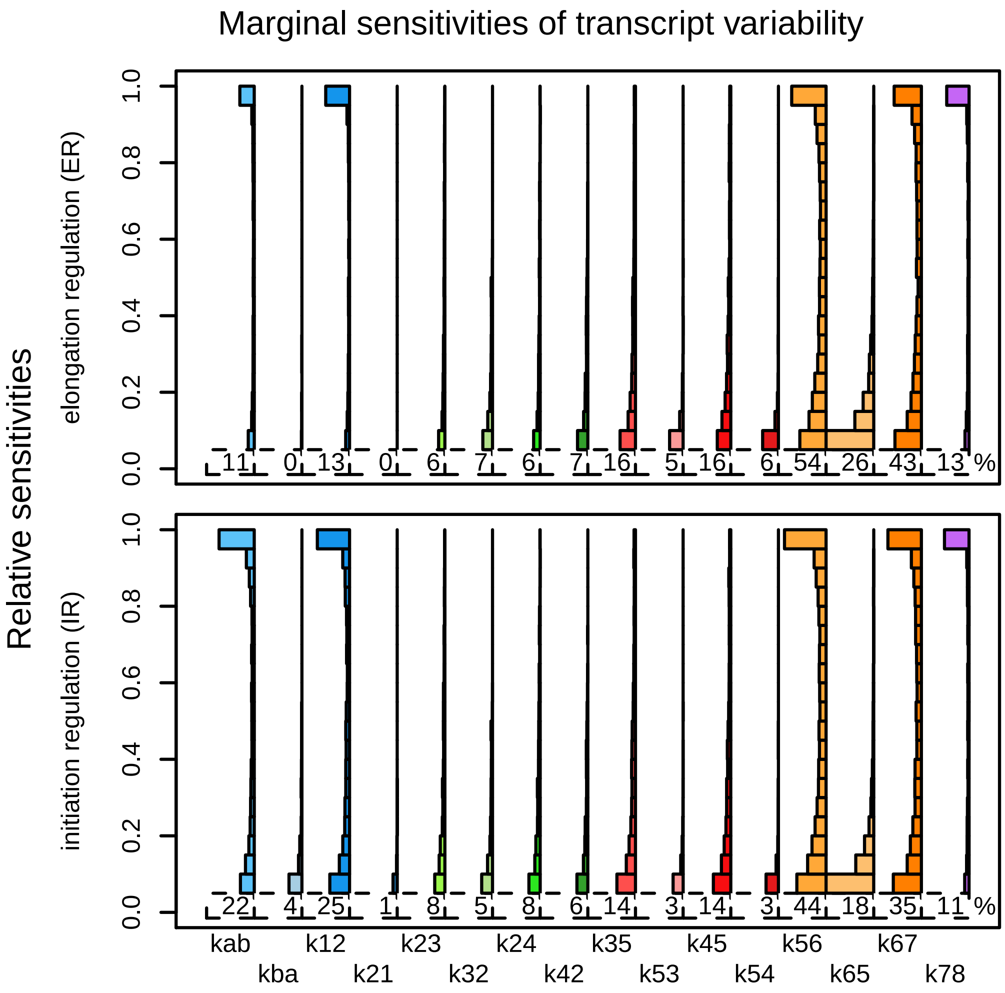

Figure 3.9: **Sensitivity analysis for transcript count variability.** The details are the same as for Figure 3.7, except that the transcript count variability is analyzed, rather than the mean transcription time.

As one might expect, for a given parameter vector the parameters to which the behavior of the models are most sensitive are generally those that happen to take the smallest value (and are thus rate-limiting): for each parameter vector, we recorded the sizes of the two parameters with the highest and second highest sensitivity values and found that their sample means were 0.147 and 0.296, respectively (whereas the sample mean of a typical parameter value will be very close to 0.5). However, just how small a given transition rate must be before it controls the system properties depends on where the corresponding edge lies in the topology of the network. As shown in figure 3.7, some parameters are relatively important throughout a large region of parameter space in both the ER and IR schemes, while others only dominate the response of the system in a small portion and some never appear.

Two further observations are evident from this analysis. First, we see which transitions in the process of activating the gene are most sensitive to small fluctuations (due to small number of TF molecules or changes in binding strength). As is apparent from figure 3.7, just 4 of the 16 promoter chain jump rates dominate the sensitivity, and these are the same for both IR and ER schemes (k_{12}, k_{56}, k_{67}, and k_{78}). The relative importance among those 4 jump rates depends on the position in parameter space, primarily through their relative sizes. Furthermore, although the ER and IR schemes have otherwise similar sensitivity profiles, the IR scheme is additionally sensitive to variation in the rate of enhancer–promoter interactions, k_{ab}. As this interaction between potentially distant DNA loci is likely rate-limiting for gene expression, the robustness of the elongation regulated scheme to fluctuations of this rate may provide a further explanation for why elongation regulated genes appear to exhibit considerably more synchronous activation. It suggests additionally that the rate of enhancer–promoter interactions is under more selective pressure for IR genes, where it has a large effect on their expression properties, than it is for ER genes, which may exhibit very similar expression properties despite having different enhancer interaction rates.

Second, we also observe that the complex assembly steps which may occur in arbitrary arrival order, namely the recruitment of TFIIE or TFIIF (governed by the jump rates k_{23}, k_{24}, k_{35}, and k_{45}) are considerably more tolerant to stochastic variation than sequential assembly steps such as the initial recruitment of the polymerase (k_{12}), the arrival of the last component of the complex, TFIIH (k_{56}), or promoter escape (k_{67}). Although between–cell variation in the total concentration of these intermediate, non-sequential binding factors will affect their binding rate parameters, it will not greatly change properties of the time to expression, thus suggesting an additional benefit of ER. This observation leads to the conclusion that the regulatory processes controlling the concentration of factors arriving in arbitrary order and the binding affinities of such factors may be under less evolutionary pressure than the corresponding quantities for factors associated with other transitions.

3.4 Discussion

Speed, synchrony, degree of cell–to–cell variability, and robustness to environmental fluctuations are important features of transcription. They are properties of the system rather than of a particular gene, DNA regulatory sequence, or gene product taken in isolation, and optimizing them can, for instance, reduce the frequency of mis-patterning events that arise due to the inherently stochastic nature of gene expression. Understanding how these properties emerge, the mechanism by which they change, and the tradeoffs involved in optimizing them all require tractable models of transcription.

Through a study of stochastic models of transcriptional activation, we demonstrated that the increased speed and synchrony of paused genes, reported by Yao et al. [168] and Boettiger et al. [17] respectively, are expected consequences of the elongation regulation shown by such genes. We also predicted that ER genes produce more consistent numbers of total transcripts than IR genes. This hypothesis can be tested directly using recently developed methods (see [5, 126] for reviews and the Section 3.5 for more details).

We furthermore explored what aspects of ER make this possible. From an examination of the effect of scaffold stability we proposed that elongation regulation should reduce the noise-amplifying nature of bursty expression. By investigating alternative models of post-initiation regulation, we also determined that our predictions depend critically on the stability of the transcriptionally engaged, paused polymerase, and would not be expected from polymerases cycling rapidly on and off the promoter (i.e. polymerase stalling).

Our investigation required us to introduce a general probabilistic framework for analyzing system properties of protein–DNA interactions. Stochastic effects, resulting from molecular fluctuations, are increasingly understood to play important roles in gene control and expression (see [125] for a review). We can now determine quantitatively how an element's location in a network affects the general properties of that network, even when the rate constants and concentrations of the network components are unknown. In particular, we quantified the extent to which system properties are sensitive to each rate parameter, something which might predict the evolutionary constraint on that component. Most previous approaches to the analysis of protein–DNA interactions have relied either on simulations, which require some knowledge of numerical rate values, or on the fluctuation–dissipation theorem, which requires the system to be near equilibrium and the noise to be small. Our methods avoid the limitations of those approaches and also make analysis of realistic models, as done in [28], significantly more feasible.

Finally, our approach is not restricted to investigating the assembly of transcriptional machinery, but may also prove useful in studying stochastic properties of a variety of regulatory DNA sequences (such as enhancers). Different assembly topologies, such as sequential versus arbitrary association mechanisms for the component TFs [56], may account for some of the observed differences in sensitivities and kinetics between otherwise similar regulatory elements. As new technologies allow better experimental determinations of these mechanisms, a theoretical framework within which one can explore their potential consequences will become increasingly important.

Taken together, our results provide theoretical and computational support for the hypothesis that the widespread appearance of promoter proximal polymerase pausing is due to its ability to mitigate noise and improve the efficiency of gene expression. The methodology we present provides a constructive framework with which to examine the properties of other macromolecular assembly processes and achieve further insights into how assembly topologies and molecular noise interact.

3.5 Analytic Methods

This section contains a more detailed discussion of the approach we use to analyze our model, including a derivation of the various analytical expressions for the moment generating functions through which we evaluate the properties of each of our models. These derivations and proofs I owe to Peter Ralph and Steve Evans, and were included in as Supporting Material in our PLoS Computational Biology paper on the subject [18]. As they form an important foundation of the conclusions discussed above, I include them here. Though the derivations are somewhat technical, we have striven to provide enough context and explanation that the causal reader should be able to follow the logic of the approach and that the reader with the basic foundation in the elements of probability and tools of calculus and linear algebra used therein should be able to follow the details of the approach.

The organization of this discussion is as follows. In Section 3.5.1 we introduce the terminology and structure that specifies the types of models we analyze. In Section 3.5.2 we review some facts about continuous-time Markov chains, including the Feynman-Kac formula. Then, in Section 3.5.3, we present an example with simplified promoter and enhancer chains to illustrate both how we combine these components to construct the two models of transcription regulation and how we may apply the Feynman-Kac formula naively to compute the Laplace transform and moments of the transcription time. Finally, in Sections 3.5.4 and 3.5.5, we describe the decomposition and approximation methods that allow us to analyze larger and more detailed models.

3.5.1 Overview of Model Formulation

As described in the main text, we use continuous-time Markov chains to model the process of *polymerase-initiation complex (PIC) assembly.* In our formulation, each state of the Markov chain corresponds to a configuration of the PIC (that is, a possible intermediary form of the complex). Once the collection of states has been determined and conveniently labeled, it only remains to specify which transitions between states are possible and to assign rates to those transitions. One may think of the states of the Markov chain as the set of vertices of a directed graph and the possible transitions as the directed edges (that is, arrows) connecting pairs of vertices in the graph. Every directed edge has an associated transition rate that does not change as time progresses. Therefore, if the label of the (random) intermediary form present at time t is X_t, then the stochastic process $X = (X_t)_{t\geq 0}$ is a time-homogeneous Markov process.

There is a distinguished state, denoted s, that corresponds to empty DNA, and another distinguished state, denoted f, that corresponds to successful transcription. The random time the chain takes to reach the final state f is the *first passage time to f*. If the chain starts in the empty state s, then this corresponds to the *transcription time* – the delay between induction and expression. For the purposes of deriving the probability distribution of the transcription time, it will also be useful to consider the probability distribution of other first passage times as well.

Properties of the transcription time have natural interpretations. For instance, consider a population of duplicate systems (cells) that are induced simultaneously. The mean (that is, expected) transcription time corresponds to the average delay between induction and expression for the population. The variance of the transcription time quantifies the variability between cells due to stochastic effects acting independently on each individual and so it indicates the degree of asynchrony in the first expression event. Similarly, if the Markov chain returns to state s each time it reaches state f, then attributes such as the mean and variance of the probability distribution of the total number of visits to state f during a finite window of time correspond to features of the collection of numbers of mRNA molecules made by members of the population during that time period.

It was our goal in the paper to compare the properties of two Markov chain models of transcription that differed only by a "topological" rearrangement in the sense that there was a correspondence between the directed edges in the two chains such that for corresponding edges the associated transition rates are equal. More specifically, we first constructed separate *promoter* and *enhancer* chains that were common to the two models and then combined them in two different ways to produce the chains that modeled transcription. Roughly speaking, the promoter and enhancer chains interacted by requiring that the enhancer chain be in its *permissive* state for the promoter to pass a certain *regulated* transition and then varying the identity of the regulated transition resulted in the two transcription regulation models.

An analytic expression for the the Laplace transform of the probability distribution of the transcription time may, in principle, be obtained from the Feynman-Kac formula, as described in Section 3.5.2. However, a naive application of this approach, with its attendant symbolic matrix inversion, quickly becomes infeasible for realistic examples with even a moderate number of states. Luckily, it is often possible to take advantage of the special structure of the transcription chains to obtain a symbolic expression for the Laplace transform and hence for the moments, or at least to provide formulas that give good approximations upon substitution of numerical values for the transition rates. In Section 3.5.4, we describe a general method for computing Laplace transforms of first passage times that relies on simplifications induced by a decomposition of the state space according to the subset of states that must be passed through by any path of positive probability leading from the initial to the final state – we call such states *pinch points*. The models of initiation regulation we consider are amenable to this approach. Unfortunately, our models of elongation regulation do not have pinch points, and a similar decomposition is not feasible. A similar decomposition described in Section 3.5.5 for chains in "parallel" is possible using spectral theory; however, it computational savings are not as great as in the case of pinch points, so we also describe a simple approximation for this case.

Our approach has several advantages. Firstly, once we have derived symbolic expressions for features of interest, it is straightforward to substitute in a large number of possibilities for the transition rate vector in order to understand how those features vary with respect to the values of the transition rates. This would be computationally impossible using simulation and at best very expensive using a numerical version of the naive Feynman-Kac formula. Secondly, the symbolic expressions can be differentiated with respect to the transition rate parameters to indicate sensitivity with respect to the values of the parameters. It would be even more infeasible to use simulation or a numerical Feynman-Kac approach to perform such a sensitivity analysis.

3.5.2 Computing first-passage times of continuous-time Markov chains

The dynamics of a time-homogeneous, continuous-time Markov chain X are fully specified by giving the state in which the chain starts and listing for each pair of distinct states $i \neq j$ the rate q_{ij} at which the chain makes a transition from i to j (if a transition from i to j is not possible, then $q_{ij} = 0$). The random time it takes the stochastic process X to leave state i has an exponential distribution with rate r_i, where $r_i = \sum_j q_{ij}$. Upon leaving state i, the probability the process jumps to state j is q_{ij}/r_i.

The quantities q_{ij} and r_i are collected into the *generator matrix* Q with elements given by $Q_{ij} = q_{ij}$ for $i \neq j$ and $Q_{ii} = -r_i$. The probability that the chain, X, is in state j at time t, given that it started in state i at time 0, is then

$$\mathbb{P}\{X_t = j \,|\, X_0 = i\} = (e^{tQ})_{ij} = \sum_{k=0}^{\infty} \frac{t^k (Q^k)_{ij}}{k!}.$$

Suppose in the representation of the Markov chain as a directed graph with arrows between states corresponding to possible transitions that if it is possible to follow a series of arrows from the state s to some state i, then it is possible to follow another series of arrows from the state i to the state f. Suppose, moreover, that there is at least one series of arrows leading from the state s to the state f. In this case, if the chain starts in state s, then with probability 1 it will eventually visit the state f.

Let τ denote the time that X first visits the state f. The Laplace transform of the random variable τ when the starting state of the chain is s, is defined as

$$\mathbb{E}[e^{-\lambda\tau} \,|\, X_0 = s] = \int_0^{\infty} e^{-\lambda t} \mathbb{P}\{\tau \in dt \,|\, X_0 = s\},$$

where λ is the transform variable.

The Laplace transform and hence, in principle, the probability distribution of τ may be computed using the modified transition matrix,

$$\widetilde{Q}_{ij} = \begin{cases} Q_{ij}, & \text{if } i \neq f, \\ 0, & \text{if } i = f. \end{cases}$$

This is the generator matrix for the *stopped* Markov chain $\widetilde{X}$, defined as $\widetilde{X}_t = X_{\min(t,\tau)}$. That is, $\widetilde{X}$ follows X up until it encounters state f, at which time it stops. Because $\widetilde{X}$ stops when it hits state f,

$$\mathbb{P}\{\tau \leq t \,|\, X_0 = s\} = \mathbb{P}\{\widetilde{X}_t = f \,|\, X_0 = s\}.$$

Integration by parts gives

$$\begin{aligned} \mathbb{E}[e^{-\lambda\tau} \,|\, X_0 = s] &= \mathbb{P}\{\tau \leq t \,|\, X_0 = s\} e^{-\lambda t}|_0^\infty + \lambda \int_0^\infty \mathbb{P}\{\tau \leq t \,|\, X_0 = s\} e^{-\lambda t} dt \\ &= \lambda \int_0^\infty \mathbb{P}\{\tau \leq t \,|\, X_0 = s\} e^{-\lambda t} dt \\ &= \int_0^\infty \lambda e^{-\lambda t} \mathbb{P}\{\widetilde{X}_t = f \,|\, X_0 = s\} dt \\ &= \int_0^\infty \lambda e^{-\lambda t} (e^{t\widetilde{Q}})_{s,f} dt \\ &= \lambda[(\lambda - \widetilde{Q})^{-1}]_{s,f}. \end{aligned}$$

The matrix $(\lambda - \widetilde{Q})$ is invertible for $\lambda > 0$; this is equation (1) in the text.

The submatrix Q_{-f} obtained by removing both the row and column indexed by f from Q (or, equivalently, $\widetilde{Q}$) is invertible and, as we observe below in Lemma 3.5.4 below, the n^{th} moment of τ is given analytically by

$$\mathbb{E}[\tau^n \,|\, X_0 = s] = (-1)^n \frac{d^n}{d\lambda^n} \lambda[(\lambda - \widetilde{Q})^{-1}]_{s,f}\Big|_{\lambda=0} = n! \sum_y (-Q_{-f})_{sy}^{-(n+1)} \widetilde{Q}_{yf}. \tag{3.4}$$

In addition to computing its moments, the probability density function of τ may be computed numerically using the inverse Laplace transform.

3.5.3 A simple example

Here is the "toy model" from the paper, described in more detail and shown in Figure 1 of the main text. The promoter assembly process is a Markov chain with four states:

- a *closed promoter* unassociated with any transcription factors;
- an *open promoter* with a loaded polymerase ready to transcribe;
- an *engaged polymerase* where the polymerase has successfully escaped the promoter; item a *completed mRNA transcript.*

The assembly process may switch back and forth between the closed and open state, depending on the arrival and stable binding of the appropriate transcription factors. Once in the actively transcribing state, the system can only leave by successful completion of transcription (i.e. entering state 4), at which time it returns to the closed promoter state and polymerase loading can occur again. The delay between induction and mRNA synthesis is represented by the time it takes the chain to get from state 1 to state 4, as depicted in Figure 3.1.

Regulation of this gene expression cascade depends on the state of a second Markov chain that describes and *enhancer.* The latter chain has only two states, A and B. The enhancer modifies the behavior of the promoter chain by the requirement that the enhancer chain must be in state B for the promoter chain to make a certain transition step. We vary the identity of this *gated* or *regulated* step and compare the resulting transcription time distributions.

We say the process is *initiation regulated* if the step from *closed* to *open* (transition $1 \to 2$ in Figure 3.1A) is regulated by the enhancer chain. That is, the enhancer chain must be in state B for the promoter chain to leave the closed state and the enhancer chain cannot leave state B while the promoter chain is in the open state.

On the other hand, we say the process is *elongation regulated* if the step from *engaged polymerase* to *completed mRNA transcript* (transition $3 \to 4$ in the Figure 3.1) is regulated by the enhancer chain. That is, the enhancer chain must be in state B for the promoter chain to move from the engaged state to the completed state. In both cases, the enhancer chain is unconstrained by the promoter chain.

These two couplings of the enhancer and promoter chains define the two new Markov chains shown in Figure 3.1B.

Having defined the system we can now compute the distribution of first passage times from a state with naked DNA to a state where the first mRNA is transcribed. The generator matrix for the initiation regulated model is (refer to Figure 3.1C, “IR composite” chain)

$$\widetilde{Q}_I = \begin{array}{c} \\ 1A \\ 1B \\ 2B \\ 3B \\ 4 \end{array} \!\!\! \begin{array}{c} \begin{array}{ccccc} 1A & 1B & 2B & 3B & 4 \end{array} \\ \begin{pmatrix} * & k_{ab} & 0 & 0 & 0 \\ k_{ba} & * & k_{12} & 0 & 0 \\ 0 & k_{21} & * & k_{23} & 0 \\ 0 & 0 & 0 & * & k_{34} \\ 0 & 0 & 0 & 0 & * \end{pmatrix} \end{array},$$

where $*$ denotes the appropriate quantity so that the rows sum to zero. The elongation regulated model has generator matrix (refer to Figure 1C, “ER composite” chain)

$$\widetilde{Q}_E = \begin{array}{c} \\ 1A \\ 2A \\ 3A \\ 1B \\ 2B \\ 3B \\ 4 \end{array} \!\!\! \begin{array}{c} \begin{array}{ccccccc} 1A & 2A & 3A & 1B & 2B & 3B & 4 \end{array} \\ \begin{pmatrix} * & k_{12} & 0 & k_{ab} & 0 & 0 & 0 \\ k_{21} & * & k_{23} & 0 & k_{ab} & 0 & 0 \\ 0 & 0 & * & 0 & 0 & k_{ab} & 0 \\ k_{ba} & 0 & 0 & * & k_{12} & 0 & 0 \\ 0 & k_{ba} & 0 & k_{21} & * & k_{23} & 0 \\ 0 & 0 & k_{ba} & 0 & 0 & * & k_{34} \\ 0 & 0 & 0 & 0 & 0 & 0 & * \end{pmatrix} \end{array}.$$

In both cases the distinguished states s and f are, respectively, state $1A$ (enhancer in state A and promoter in state 1) and state 4 (the gene is actively transcribing). With the help of a symbolic package such as *Sage* or *Mathematica*, we can apply (3.4) to find analytic expressions for the moments of the transcription time in each model. Doing so results in lengthy expressions (from which we spare the reader) and no obvious consistent ordering between the two schemes; but numerical evaluation shows that over the vast majority of parameter space, the ER scheme is faster than the IR scheme (the mean transcription time is smaller), but also more noisy (the variance of the transcription time and the transcript count variability are both larger). The distribution of the log ratios for the speed, degree of synchrony, and variation in total transcripts made are plotted in Figure 3.1 in the main text.

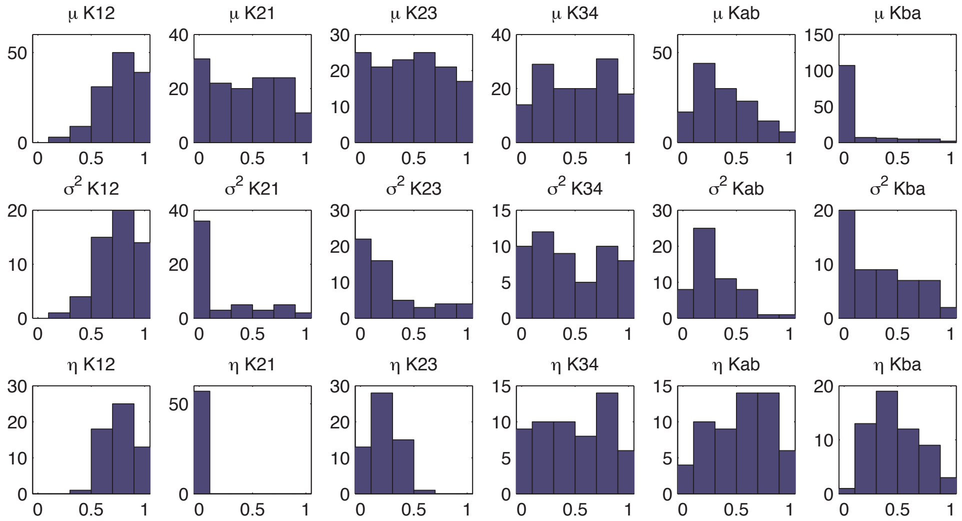

Figure 3.10: Histograms of the distributions of those parameter values where the IR scheme is faster than the ER scheme (top row), more synchronous the ER scheme (middle row) or less noisy in terms of total transcripts than the ER scheme (bottom row).

Examining the parameter combinations at which the IR model is faster (histograms are show in Figure 3.10) reveals that for this to be true, k_{12} must be fast, while k_{ba} must be slow, and k_{ab} must be even slower. This seems to be allowing both chains to reach state 3B at about the same time, since the transition $1 \to 2$ is fast, at which point the ER chain has a chance of falling back to state 3A, a possibility that the IR chain avoids.

3.5.4 Decomposition into sequential modules

In this section we present and prove analytical tools for the decomposition of a detailed transcription model into modules connected in a sequential manner. We proceed somewhat abstractly at first, but the connection with models of transcription will soon become clear.

Set-up and notation

Suppose we have a sequence of continuous-time Markov chains X^k on a sequence of state spaces $\mathcal{X}^k$ for $k \in \{1, 2, \ldots, n\}$. Suppose that each state space $\mathcal{X}^k$ has two distinguished (and distinct) states s_k and f_k. Each Markov chain

X^k represents a single "stage" of the transcription factor assembly. We assume that f_k is accessible from any state in $\mathcal{X}^k$, for each k. The entire process of transcription is modeled by a Markov chain X that is constructed by stringing the state spaces together sequentially, identifying s_k with f_{k-1} for $2 \leq k \leq n$, and leaving the transition rates the same. We call the state $f_k = s_{k+1}$ the k^{th} *pinch point*, and denote it by p_k.

For some state $b \in \mathcal{X}^k$ and a Markov chain Y on $\mathcal{X}^k$, define

$$\tau_b(Y) = \inf\{t > 0 : Y(t) = b\},$$

the time it takes the chain Y to first arrive at b.

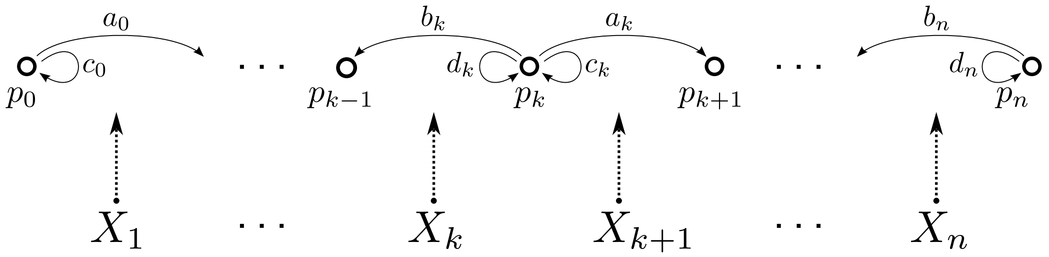

Figure 3.11: A schematic of the decomposition. The probabilities a_k, b_k, c_k, and d_k depend only on the distributions of both adjacent chains X_k and X_{k+1}, while the behavior of X between pinch points p_{k-1} and p_k only depends on the distribution of X_k.

Once X^k leaves s_k, there are several possible behaviors, and we need to introduce chains that behave as X^k conditioned on each behavior. For each k, let $\nu_s^k(\cdot)$ denote the distribution of X^k after the first jump from s_k, namely, if T is the time of the first jump, then

$$\nu_s^k(i) = \mathbb{P}\{X_T = i \,|\, X_0 = s\}.$$

Similarly, let $\nu_f^k(\cdot)$ denote the probability distribution of X^k after the first jump from f_k. Write $X^k_{\rightarrow}$ for a Markov chain on $\mathcal{X}^k$ that has the distribution of X^k begun with distribution ν_s^k and conditioned to hit f_k before returning to s_k; also write $X^k_{\circlearrowright}$ for the chain that has the distribution of X^k begun with distribution ν_s^k and conditioned to return to s_k before hitting f_k. Define $X^k_{\leftarrow}$ and $X^k_{\circlearrowleft}$ similarly but with the roles of s_k and f_k reversed. Define the following four random *traversal times*

$$\begin{aligned} \tau^k_{\rightarrow} &= \tau_{f_k}(X^k_{\rightarrow}) \\ \tau^k_{\leftarrow} &= \tau_{s_k}(X^k_{\leftarrow}) \\ \tau^k_{\circlearrowright} &= \tau_{s_k}(X^k_{\circlearrowright}) \\ \tau^k_{\circlearrowleft} &= \tau_{f_k}(X^k_{\circlearrowleft}). \end{aligned} \tag{3.5}$$

Denote the pinch points $p_0, \ldots, p_n$, where $p_0 = s_1$, $p_n = f_n$, and $p_k = \{f_k \text{ identified with } s_{k+1}\}$ for $1 \leq k \leq n-1$. If the chain X is at the k^{th} pinch point p_k, for $1 \leq k \leq n-1$, then it has four options with the following corresponding probabilities

$$\begin{aligned} a_k &= \mathbb{P}\{ \text{ hit } p_{k+1} \text{ without returning to } p_k \ \}, \\ b_k &= \mathbb{P}\{ \text{ hit } p_{k-1} \text{ without returning to } p_k\}, \\ c_k &= \mathbb{P}\{ \text{ move into } \mathcal{X}^{k+1} \text{ but return to } p_k \text{ before hitting } p_{k+1} \ \}, \\ d_k &= \mathbb{P}\{ \text{ move into } \mathcal{X}^{k} \text{ but return to } p_k \text{ before hitting } p_{k-1} \ \}. \end{aligned} \tag{3.6}$$

If X is at either of the pinch points p_0 or p_n it has only two options. Once we choose one of these options, X then moves like a conditioned X^k chain until it hits a pinch point again. For instance, if the event with probability a_k happens, then p_{k+1} will be the next pinch point hit, and until p_{k+1} is hit the chain X moves like the chain $X^k_{\rightarrow}$. We compute the probabilities a_k, b_k, c_k, d_k in Subsection 3.5.4.

When the chain leaves a pinch point and returns, it could have done so in either direction, so we combine $\tau_{\circlearrowleft}$ and $\tau_{\circlearrowright}$ to form an additional traversal time. For each $1 \leq k \leq n$ let $\tau^k_{\circ}$ be a mixture of $\tau^{k+1}_{\circlearrowright}$ and $\tau^k_{\circlearrowleft}$, defined by

$$\tau^k_{\circ} = \begin{cases} \tau^{k+1}_{\circlearrowright} & \text{with probability } \frac{c_k}{c_k+d_k}, \\ \tau^k_{\circlearrowleft} & \text{with probability } \frac{d_k}{c_k+d_k}. \end{cases} \tag{3.7}$$

If $c_k = 0$ and $d_k > 0$ then $\tau^k_{\circ} = \tau^{k+1}_{\circlearrowright}$, if $c_k > 0$ and $d_k = 0$ then $\tau^k_{\circ} = \tau^k_{\circlearrowleft}$, and if $c_k = d_k = 0$ then we define $\tau^k_{\circ} = 0$ (although this will not enter into the computations).

The glue that joins the above modules together is the "pinch chain" Z, defined to be the discrete-time Markov chain that records the order in which X visits the pinch points. Formally, Z is a Markov chain on the state space

$\{0, 1, \ldots, n\}$ that at each step either moves up by one, down by one, or stays put, and the transition probabilities are, for $0 \le k \le n$,

$$\mathbb{P}\{Z_{k+1} = j \mid Z_k = i\} = \begin{cases} a_i, & \text{if } j = i + 1 \le n, \\ b_i, & \text{if } j = i - 1 \ge 0, \\ c_i + d_i, & \text{if } j = i, \\ 0, & \text{otherwise.} \end{cases} \tag{3.8}$$

We define P to be the transition matrix for the chain Z stopped upon hitting n, so that

$$\begin{aligned} &P_{ij} = \mathbb{P}\{Z_1 = j \mid Z_0 = i\}, \quad \text{for } 0 \le i \le n-1,\ 0 \le j \le n, \\ &P_{nj} = 0, \quad \text{for } 0 \le j \le n-1, \\ &P_{nn} = 1. \end{aligned} \tag{3.9}$$

We discuss computation of P in Subsection 3.5.4.

Finally, for each pinch point $0 \le k \le n-1$, define an independent random variable S^k with the exponential distribution

$$\mathbb{P}\{S^k > t\} = \exp\left\{-t(r^k(f) + r^{k+1}(s))\right\}, \tag{3.10}$$

where $r^k(f)$ is the jump rate out of f_k for X^k, and $r^{k+1}(s)$ is the jump rate out of s_{k+1} for X^{k+1}. The random variable S^k has the distribution of the amount of time X spends at p_k before moving.

Computing system noise properties

Now we are ready to state our theoretical results. First we give the form of the Laplace transform and the moments of the assembly time in terms of the transition probabilities between modules and the distributions of the traversal times. In Subsection 3.5.4 we discuss how to compute the transition probabilities, and in Subsection 3.5.4 we discuss how to compute the relevant quantities for the traversal times.

Theorem Recall the matrix P from (3.9). For $0 \le j \le n-1$ and $0 \le k \le n$ set

$$\tau_{jk} = \begin{cases} \tau^j_{\leftarrow} + S^j, & \text{if } k = j - 1 \text{ and } P_{jk} > 0, \\ \tau^j_{\circ} + S^j, & \text{if } k = j \text{ and } P_{jk} > 0, \\ \tau^{j+1}_{\rightarrow} + S^j, & \text{if } k = j + 1 \text{ and } P_{jk} > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where the various random variables are as defined as in (3.5), (3.7) and (3.10). Furthermore, put

$$\phi_{jk}(\lambda) = \mathbb{E}[\exp(-\lambda \tau_{jk})],$$

and consider the $n \times n$ matrix W with entries $W_{jk} = \phi_{jk}(\lambda) P_{jk}$ for $0 \leq j, k \leq n-1$. Let $v(\lambda)$ be a vector with $v_n(\lambda) = 1$ and

$$v_i(\lambda) = \sum_{j=0}^{n-1} (I - W)^{-1}_{ij} \phi_{jn}(\lambda) P_{jn}$$

for $0 \leq i \leq n-1$.

Then, the total time to assembly, $\tau = \inf\{t > 0 : X_t = f_n\}$, has Laplace transform

$$\mathbb{E}[\exp(-\lambda\tau) \mid X_0 = p_i] = v_i(\lambda)$$

for $0 \leq i \leq n$.

Corollary Define matrices M, Σ, and R by

$$M_{ij} = P_{ij}\mathbb{E}[\tau_{ij}], \quad 0 \leq i \leq n-1,\ 0 \leq j \leq n, \tag{3.11}$$

$$\Sigma_{ij} = P_{ij}\mathbb{E}[\tau_{ij}^2], \quad 0 \leq i \leq n-1,\ 0 \leq j \leq n, \tag{3.12}$$

$$R_{ij} = \begin{cases} (I - P_{-n})^{-1}_{ij}, & 0 \leq i \leq n-1,\ 0 \leq j \leq n-1, \\ 0, & i = n,\ 0 \leq j \leq n-1. \end{cases} \tag{3.13}$$

Then, the first and second moments of the random time τ are

$$\begin{aligned} \mathbb{E}[\tau \mid X_0 = p_i] &= (RM\mathbf{1})_i \\ \mathbb{E}[\tau^2 \mid X_0 = p_i] &= \left(R\Sigma\mathbf{1} + 2(RM)^2\mathbf{1}\right)_i \end{aligned} \tag{3.14}$$

for $0 \leq i \leq n-1$.

Remark[Random starting state] We have treated the starting state as fixed, but this need not be the case. If, for instance, after transcription is completed, the PIC returns to an intermediate state, then the delay between subsequent transcription events could be modeled as the time to transcription begun at a random state added to the (in general random) time required for actual transcription. So, if after transcription the chain waits time W and independently jumps to state I, then the time delay between transcriptions is $D = W + \tau^{(I)}$, where $\tau^{(I)}$ has the distribution of τ if $X_0 = p_I$, and $I = i$

with probability w_i, for some probabilities w_i with $\sum_{i=0}^{n-1} w_i = 1$. Then we have that $\mathbb{E}[D] = \mathbb{E}[W] + \mathbb{E}[\tau^{(I)}]$, and $\text{Var}[D] = \text{Var}[W] + \text{Var}[\tau^{(I)}]$, and

$$\begin{aligned} \mathbb{E}[\tau^{(I)}] &= \sum_i w_i \left(RM\mathbf{1}\right)_i \\ \mathbb{E}\left[\left(\tau^{(I)}\right)^2\right] &= \sum_i w_i \left(R\Sigma\mathbf{1} + 2(RM)^2\mathbf{1}\right)_i \end{aligned} \tag{3.15}$$

for $0 \le i \le n-1$.

Theorem 3.5.4 is the solution we needed to compute the Laplace transform of the total transition time, τ, from the transition times between the modules of the larger chain. The corollary will be useful in computing moments without having to recompute the derivatives of the Laplace transform of τ for each model one examines. We will next prove both the theorem and the corollary. The reader primarily interested in the method and not the proof may jump to Section 3.5.4, where we show how the Laplace transforms of the transitions between modules can be computed from the rate matrices for those modules.

Proof[Proof of Theorem 3.5.4] To prove Theorem 3.5.4, we decompose the path of X by first looking at the order in which X traverses the pinch points — the sample path of the pinch chain Z — and then according to the path that Z takes, add in the appropriate random amounts of time for each step. The Laplace transform of the assembly time will be put together from two pieces: the joint probability generating function of the transition counts of the pinch chain and the Laplace transforms of the relevant traversal times.

Set $Z_0 = 0$, and write T for the first time that Z hits n, after which Z stays fixed. Define for each pair of states (j, k) the transition count

$$N_{jk} = \#\{1 \le i \le T : Z_{i-1} = j \text{ and } Z_i = k\}.$$

That is, N_{jk} is the number of times the chain Z moves from j to k. If k is not one of $j-1$, j, or $j+1$, then N_{jk} will be zero.

The following lemma giving the joint probability generating function of the transition counts is proved in Subsection 3.5.4.

Lemma Let $\{z_{ij}\}$ be a set of dummy variables with $z_{ij} \in [0, 1]$ for all $0 \le i, j \le n$. Define the matrix $P(z)_{jk} = z_{jk}P_{jk}$ and define the vector $v(z)$ by $v(z)_n = 1$ and

$$v(z)_i = \sum_{j=0}^{n-1} \left((I - P(z)_{-n})^{-1}\right)_{ij} P(z)_{jn} \quad \text{for } 0 \le i \le n-1, \tag{3.16}$$

where $P(z)_{-n}$ is the matrix $P(z)$ with the last row and column removed. Then,

$$\mathbb{E}\left[\prod_{jk} z_{jk}^{N_{jk}} \,\middle|\, Z_0 = i\right] = v_i(z).$$

Suppose $Z_i = k$, indicating that X is in state p_k. The amount of time before X leaves p_k has the distribution of S^k, so we need to add an independent copy of S^k. If $Z_{i+1} = k+1$, then X will hit p_{k+1} before returning to p_k. By construction, the amount of time this takes has the same distribution as $\tau_{\rightarrow}^{k+1}$, so we need to add on a copy of $\tau_{\rightarrow}^{k+1}$, whose value is independent of everything else. Similarly, if $Z_i = k$ and $Z_{i+1} = k-1$, we need to add on a copy of $\tau_{\leftarrow}^{k}$. If $Z_i = Z_{i+1} = k$, then this corresponds to a single excursion of X from the k^{th} pinch point that could have been in either direction. In this case, we need to add a random time $\tau_{\circ}^{k}$ that is a mixture of the distribution of $\tau_{\circlearrowleft}^{k+1}$ with probability $c_k/(c_k+d_k)$ and the distribution of $\tau_{\circlearrowright}^{k}$ with probability $d_k/(c_k+d_k)$, as defined in (3.7).

Let the total time to assembly be denoted τ, and for each k let $\tau_{\rightarrow,1}^{k}, \tau_{\rightarrow,2}^{k}, \ldots$ be an infinite sequence of independent copies of $\tau_{\rightarrow}^{k}$; define $\tau_{\leftarrow,m}^{k}$ and $\tau_{\circ,m}^{k}$ for $m \geq 1$ similarly. Also let $S_{1,m}^{k}$, $S_{2,m}^{k}$, and $S_{3,m}^{k}$ be three infinite sequences of independent copies of S^k. Our decomposition in terms of the path of Z tells us that τ is distributed as

$$\sum_{k=0}^{n}\left(\sum_{m=1}^{N_{k,k-1}}\left(\tau_{\leftarrow,m}^{k}+S_{1,m}^{k}\right)+\sum_{m=1}^{N_{k,k}}\left(\tau_{\circ,m}^{k}+S_{2,m}^{k}\right)+\sum_{m=1}^{N_{k,k+1}}\left(\tau_{\rightarrow,m}^{k+1}+S_{3,m}^{k}\right)\right).$$

Therefore, by conditioning on Z, we get

$$\begin{aligned}\mathbb{E}\left[e^{-\lambda\tau}\right] &= \mathbb{E}\left[\prod_{k=0}^{n}\left(\prod_{m=1}^{N_{k,k-1}} e^{-\lambda(\tau_{\leftarrow,m}^{k}+S_{1,m}^{k})}\prod_{m=1}^{N_{k,k}} e^{-\lambda(\tau_{\circ,m}^{k}+S_{2,m}^{k})}\prod_{m=1}^{N_{k,k+1}} e^{-\lambda(\tau_{\rightarrow,m}^{k+1}+S_{3,m}^{k})}\right)\right]\\ &= \mathbb{E}\left[\prod_{k=0}^{n}\left(\mathbb{E}[e^{-\lambda(\tau_{\leftarrow}^{k}+S^{k})}]^{N_{k,k-1}}\mathbb{E}[e^{-\lambda(\tau_{\circ}^{k}+S_{k})}]^{N_{k,k}}\mathbb{E}[e^{-\lambda(\tau_{\rightarrow}^{k+1}+S_{k})}]^{N_{k,k+1}}\right)\right].\end{aligned}$$

This proves Theorem 3.5.4. □

Note that since, for instance, S^k and $\tau_{\circlearrowright}^{k}$ are independent, we may compute their Laplace transforms and moments separately. Furthermore,

$$\mathbb{E}\left[e^{-\lambda\tau_{\circ}^{k}}\right] = \frac{1}{c_k+d_k}\left(c_k\mathbb{E}\left[e^{-\lambda\tau_{\circlearrowleft}^{k+1}}\right]+d_k\mathbb{E}\left[e^{-\lambda\tau_{\circlearrowright}^{k}}\right]\right). \qquad (3.17)$$

Proof[Proof of Corollary 3.5.4.] Without loss of generality, take $X_0 = p_0$. We leave this implicit and write, for instance, $\mathbb{E}[\tau] = \mathbb{E}[\tau \mid X_0 = p_0]$.

Note that by differentiating the result of Lemma 3.5.4 we get

$$\begin{aligned} \mathbb{E}[\tau] &= \sum_{j=0}^{n-1} \sum_{k=0}^{n} \mathbb{E}[\tau_{jk}] \partial_{z_{jk}} v_0(1), \\ \mathbb{E}[\tau^2] &= \sum_{j=0}^{n-1} \sum_{k=0}^{n} \mathbb{E}[\tau_{jk}^2] \partial_{z_{jk}} v_0(1) \qquad (3.18) \\ &\quad + \sum_{j=0}^{n-1} \sum_{k=0}^{n} \sum_{\ell=0}^{n-1} \sum_{m=0}^{n} \mathbb{E}[\tau_{jk}] \mathbb{E}[\tau_{\ell m}] \partial_{z_{jk}} \partial_{z_{\ell m}} v_0(1). \end{aligned}$$

We compute the derivatives of v at $z = 1$. For ease of notation, write $\partial_{z_{jk}}$ as ∂_{jk}. Because $P(z)v(z) = v(z)$,

$$\partial_{jk} v(z) = (\partial_{jk} P(z)) v(z) + P(z) \partial_{jk} v(z).$$

Now, since $v(1) = (1, 1, 1, \ldots, 1)^T$ and

$$(\partial_{jk} P(z))_{qr} = \begin{cases} P_{jk}, & \text{if } q = j, \text{ and } r = k, \\ 0, & \text{otherwise,} \end{cases}$$

$\partial_{jk} v(1)$ solves the set of linear equations $(I - P)\partial_{jk} v(1) = P_{jk} e_j$, where e_j is the j^{th} standard basis vector. More explicitly,

$$\partial_{jk} v(1)_i - \sum_r P_{ir} \partial_{jk} v(1)_r = \begin{cases} P_{jk}, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases}$$

Since we require that $v(z)_n = 1$, we have $\partial_{jk} v(z)_n = 0$.

For the moment, write I_m for the identity matrix of order m. Because $I_{n+1} - P$ is the transition matrix for an irreducible continuous-time Markov chain stopped upon reaching n, it follows that the matrix $I_n - P_{-n}$ is invertible. Define R to be $(I_n - P_{-n})^{-1}$ with an extra row of zeros at the bottom, as in the statement of the corollary, and let w be the j^{th} column of R. It is easy to check that $(I - P)w = e_j$, and that this solution is unique over vectors whose last entry is zero. Thus,

$$\partial_{jk} v(1)_i = R_{ij} P_{jk}$$

for $0 \le i \le n-1$ and $0 \le j,k \le n$.

Differentiating the identity a second time, we get

$$\begin{aligned}\partial_{jk}\partial_{lm}v(z) &= (\partial_{jk}^2 P(z))v(z) + (\partial_{jk}P(z))\partial_{lm}v(z) \\ &\quad + (\partial_{lm}P(z))\partial_{jk}v(z) + P(z)\partial_{jk}^2 v(z).\end{aligned}$$

Now let $u = \partial_{jk}\partial_{lm}v(1)$. Since $\partial_{jk}\partial_{lm}P(z) = 0$, the vector u satisfies $(I_{n+1} - P)u = P_{jk}\partial_{lm}v(1)_k e_j + P_{lm}\partial_{jk}v(1)_m e_l$, or, using our solution for the first derivative,

$$u_i - \sum_r P_{ir}u_r = \begin{cases} P_{jk}P_{lm}R_{kl}, & \text{if } i = j, \\ P_{lm}P_{jk}R_{mj}, & \text{if } i = l, \\ 0, & \text{otherwise.} \end{cases}$$

By linearity, we can use our solution from above to solve this system. Thus,

$$\partial_{jk}\partial_{lm}v(z)_i = P_{jk}P_{lm}\left(R_{il}R_{mj} + R_{ij}R_{kl}\right).$$

Evaluating the sums in (3.18) gives (3.14). □

Transition counts for the pinch-chain

Proof[Proof of Lemma 3.5.4.] Define a matrix $P(z)$ by $P(z)_{jk} = z_{jk}P_{jk}$. It is easy to see that the joint probability generating function for the transition counts N_{jk} is given by

$$\mathbb{E}\left[\prod_{jk} z_{jk}^{N_{jk}}\right] = \lim_{m\to\infty}(P^m(z))_{0,n}.$$

For this to be nonzero, we must take $z_{nn} = 1$. Also, since the matrix $P(z)$ is substochastic and the last row of $P(z)$ is zero except for a 1 on the diagonal, $P(z)$ has a single eigenvalue of value 1, with left eigenvector $\pi = (0, 0, \ldots, 1)$. Any tridiagonal real matrix (a_{ij}) satisfying $a_{i,i+1}a_{i+1,i} > 0$ for all i is similar to a symmetric matrix and, in particular, such a matrix has a full complement of real eigenvalues with corresponding left and right eigenvectors [65]. Therefore, $P(z)$ has a unique right eigenvector with eigenvalue 1, that we call $v(z)$ and normalize so that $v_n(z) = 1$. The other eigenvalues are strictly less than one, so

$$\lim_{m\to\infty}(P^m(z))_{jk} = \pi_k v_j(z),$$

whence

$$\mathbb{E}\left[\prod_{jk} z_{jk}^{N_{jk}}\right] = v_0(z).$$

By Lemma 3.5.4, the unique solution of the eigenvector equation $P(z)v(z) = v(z)$ with the normalization $v_n(z) = 1$ is

$$v(z)_i = \sum_{j=0}^{n-1} (I - P_{-n})_{ij}^{-1} P_{jn}.$$

Note that, by conditioning on the first step of the chain, $v_i(z) = \mathbb{E}\left[\prod_{jk} z_{jk}^{N_{jk}} \,\middle|\, Z_0 = i\right]$ is a solution to $P(z)v(z) = v(z)$. □

Transition probabilities between modules

Here we compute the transition probabilities (a_k, b_k, c_k, d_k), defined in (3.6). Consider the tridiagonal matrix P with entries

$$P = \begin{bmatrix} 1-a_0 & a_0 & & & \\ b_1 & 1-(b_1+a_1) & a_1 & & \\ & b_2 & 1-(b_2+a_2) & a_2 & \\ & & \ddots & & \\ & & b_{n-1} & 1-(b_{n-1}+a_{n-1}) & a_{n-1} \\ & & & 0 & 1 \end{bmatrix}.$$

Suppose that X^k jumps out of f_k at rate r_f^k, X^{k+1} jumps out of s_{k+1} at rate r_s^{k+1}, and the probability that X^{k+1} begun at s_{k+1} reaches f_{k+1} before returning to s_{k+1} is q. Then,

$$a_k = q r_s^{k+1}/(r_f^k + r_s^{k+1}). \tag{3.19}$$

Similarly, if the probability that X^k begun at f_k reaches s_k before returning to f_k is q', then

$$b_k = q' r_f^k/(r_f^k + r_s^{k+1}).$$

Also $c_k = (1-q)r_s^{k+1}/(r_f^k + r_s^{k+1})$ and $d_k = (1-q')r_f^k/(r_f^k + r_s^{k+1})$. We need only compute q and q' for each component chain X^k separately.

For the remainder of this section, let Y be an irreducible continuous-time Markov chain with distinguished states s and f, and matrix of transition rates G, where $G_{ii} = -\sum_{j\neq i} G_{ij}$ as usual.

Define G^{**} to be the matrix of transition rates for Y stopped at both s and f. That is,

$$G^{**}_{ij} = \begin{cases} G_{ij}, & \text{if } i \neq s, f, \\ 0, & \text{otherwise.} \end{cases}$$

Note that if we define

$$x_i = \mathbb{P}\{Y \text{ hits } f \text{ before } s \mid Y_0 = i\}, \tag{3.20}$$

then it is well-known that [37]

$$\sum_j G^{**}_{ij} x_j = 0 \quad \text{for all } i.$$

In other words, x is the unique right eigenvector of G^{**} corresponding to the zero eigenvalue that satisfies the boundary conditions $x_s = 0$ and $x_f = 1$.

By Lemma 3.5.4, if we denote by G_{-sf} the submatrix obtained from G by removing rows and columns corresponding to both s and f, then, for $i \notin \{s, f\}$,

$$x_i = \sum_{j\notin\{s,f\}} (-G_{-sf})^{-1}_{ij} G_{jf}. \tag{3.21}$$

We require $\mathbb{P}\{Y \text{ hits } f \text{ before } s \mid Y_0 = s\}$, i.e. the probability q in (3.19). We find this probability by conditioning on the state the chain goes to at the time it first leaves the state s. Let S be the time that Y first leaves s. This random variable has an exponential distribution with rate $-G_{ss} = \sum_{j\neq s} G_{sj}$. Thus,

$$\begin{aligned} \mathbb{P}\{Y \text{ hits } f \text{ before } s \mid Y_0 = s\} &= \sum_i \mathbb{P}\{Y_S = i\} \\ &\qquad \times \mathbb{P}\{Y \text{ hits } f \text{ before } s \mid Y_0 = i\} \\ &= \sum_i \frac{G_{si}}{-G_{ss}} x_i. \end{aligned}$$

We summarize the above computations. Let (G^k, x^k) be the objects discussed above that are associated with the k^{th} chain and let (a_k, b_k, c_k, d_k)

be defined as in (3.6). By definition, $b_0 = d_0 = a_n = c_n = 0$. If we take $G^0_{ff} = G^{n+1}_{ss} = 0$, then

$$\begin{aligned} a_k &= \frac{G^{k+1}_{ss}}{G^{k+1}_{ss} + G^k_{ff}} \sum_{i \neq s} \frac{G^{k+1}_{si}}{-G^{k+1}_{ss}} x^{k+1}(i) \\ &= -\sum_{i \neq s} \frac{G^{k+1}_{si}}{G^{k+1}_{ss} + G^k_{ff}} x^{k+1}(i), \qquad \text{for } 0 \leq k \leq n-1, \\ &= \frac{1}{-G^{k+1}_{ss} - G^k_{ff}} \left(G^{k+1}_{sf} + \sum_{i \notin \{s,f\}} \sum_{j \notin \{s,f\}} G^{k+1}_{si} (G^{k+1}_{-sf})^{-1}_{ij} G^{k+1}_{jf} \right) \end{aligned}$$

and similarly

$$\begin{aligned} b_k &= -\sum_{i \neq f} \frac{G^k_{fi}}{G^{k+1}_{ss} + G^k_{ff}} (1 - x^k(i)), \qquad \text{for } 1 \leq k \leq n, \\ c_k &= -\sum_{i \neq s} \frac{G^{k+1}_{si}}{G^{k+1}_{ss} + G^k_{ff}} (1 - x^{k+1}(i),) \qquad \text{for } 0 \leq k \leq n-1, \\ d_k &= -\sum_{i \neq f} \frac{G^k_{fi}}{G^{k+1}_{ss} + G^k_{ff}} x^k(i), \qquad \text{for } 1 \leq k \leq n. \end{aligned}$$

Traversal times within modules

Here we show how to compute quantities related to the traversal times. Again let Y be an irreducible Markov chain with transition matrix G, and let G^{**} be the transition matrix for Y stopped upon hitting either s or f, so that $G^{**}_{ij} = G_{ij}$ for $i \notin \{s, f\}$ and $G_{sj} = G_{fj} = 0$.

Lemma Let $\tau_{\rightarrow}$, $\tau_{\leftarrow}$, $\tau_{\circlearrowright}$, and $\tau_{\circlearrowleft}$ be defined as in (3.5) for a chain with matrix of transition probabilities G. Let $x_s = 0$, $x_f = 1$, and

$$x_i = \sum_{j \notin \{s,f\}} (-G_{-sf})^{-1}_{ij} G_{jf}.$$

The Laplace transforms are then given by

$$\mathbb{E}[e^{-\lambda\tau_{\rightarrow}}] \quad = \quad \sum_{i\notin\{s,f\}:x_i>0} \frac{G_{si}}{-G_{ss}} \frac{\lambda\left((\lambda - G^{**})^{-1}\right)_{if}}{x_i}, \tag{3.22}$$

$$\mathbb{E}[e^{-\lambda\tau_{\leftarrow}}] \quad = \quad \sum_{i\notin\{s,f\}:x_i<1} \frac{G_{fi}}{-G_{ff}} \frac{\lambda\left((\lambda - G^{**})^{-1}\right)_{is}}{(1-x_i)}, \tag{3.23}$$

$$\mathbb{E}[e^{-\lambda\tau_{\circlearrowright}}] \quad = \quad \sum_{i\notin\{s,f\}:x_i<1} \frac{G_{si}}{-G_{ss}} \frac{\lambda\left((\lambda - G^{**})^{-1}\right)_{is}}{(1-x_i)}, \tag{3.24}$$

$$\mathbb{E}[e^{-\lambda\tau_{\circlearrowleft}}] \quad = \quad \sum_{i\notin\{s,f\}:x_i>0} \frac{G_{fi}}{-G_{ff}} \frac{\lambda\left((\lambda - G^{**})^{-1}\right)_{if}}{x_i}. \tag{3.25}$$

Corollary The moments of the traversal times $\tau_{\rightarrow}$ and $\tau_{\circlearrowright}$ are

$$\begin{aligned} \mathbb{E}[\tau_{\rightarrow}^m] &= m! \sum_{i,j\notin\{s,f\}:x_i>0} \frac{G_{si}(-G_{-sf})_{ij}^{-(m+1)}G_{jf}}{(-G_{ss})x_i} \\ \mathbb{E}[\tau_{\circlearrowright}^m] &= m! \sum_{i,j\notin\{s,f\}:x_i<1} \frac{G_{si}(-G_{-sf})_{ij}^{-(m+1)}G_{js}}{(-G_{ss})(1-x_i)} \end{aligned} \tag{3.26}$$

The moments of $\tau_{\leftarrow}$ and $\tau_{\circlearrowleft}$ are found by exchanging the roles of s and f, which also interchanges x_i and $(1-x_i)$.

Note that $\mathbb{E}[e^{-\lambda\tau_{\circ}^{k}}]$ is obtained by substituting the results of the corollary into (3.17).

Remark To use these in Theorem 3.5.4 we need to translate the τ_{jk} defined there into combinations of the above traversal times. For convenience, we record here which entries of the matrices ϕ, M or Σ depend on the probability distributions of which traversal times. The following $(n+1)\times(n+1)$ matrix is tridiagonal, and the $(j,k)^{\text{th}}$ entry contains the random variables on whose distributions the $(j,k)^{\text{th}}$ entry of ϕ, M, or Σ depend.

$$\phi, M, \Sigma \quad \text{depend on} \quad \begin{bmatrix} \tau_{\circlearrowright}^1 & \tau_{\rightarrow}^1 & & & \\ \tau_{\leftarrow}^1 & (\tau_{\circlearrowleft}^1, \tau_{\circlearrowright}^2) & \tau_{\rightarrow}^2 & & \\ & \tau_{\leftarrow}^2 & (\tau_{\circlearrowleft}^2, \tau_{\circlearrowright}^3) & \tau_{\rightarrow}^3 & \\ & & \ddots & & \\ & & \tau_{\leftarrow}^{n-1} & (\tau_{\circlearrowleft}^{n-1}, \tau_{\circlearrowright}^n) & \tau_{\rightarrow}^n \\ & & & 0 & 0 \end{bmatrix}.$$

The empty entries are identically zero.

Also, recall that S^k is exponentially distributed with rate $-G_{ss}^{k+1} - G_{ff}^k$. Thus,

$$\mathbb{E}[e^{-\lambda S^k}] = \frac{-G_{ss}^{k+1} - G_{ff}^k}{\lambda - G_{ss}^{k+1} - G_{ff}^k}$$

and

$$\mathbb{E}[(S^k)^n] = \frac{n!}{(-G_{ss}^{k+1} - G_{ff}^k)^n}.$$

Remark At first sight, (3.22) does not appear to give the right answer at $\lambda = 0$. However, recall that G^{**} is not invertible. As we show in Lemma 3.5.4, $[\lim_{\lambda\to 0} \lambda(\lambda - G^{**})^{-1}]_{if} = x_i$, so

$$\begin{aligned}\lim_{\lambda\to 0} \mathbb{E}[e^{-\lambda\tau_\to}] &= \sum_{i\neq s} \frac{G_{si}}{-G_{ss}} \frac{x_i}{x_i} \\ &= 1.\end{aligned}$$

Proof[Proof of Lemma 3.5.4 and Corollary 3.5.4.] Let $Y^\to$ denote the chain Y conditioned to hit the state f before hitting s, and let Y^{**} be the chain Y stopped upon hitting either s or f. Denote by $A_\to$ the event that Y hits f before hitting s. For $i \notin \{s, f\}$,

$$\begin{aligned}\mathbb{P}\{Y_t^\to = j \,|\, Y_0^\to = i\} &= \frac{\mathbb{P}\{Y_t = j, A_\to \,|\, Y_0 = i\}}{\mathbb{P}\{A_\to \,|\, Y_0 = i\}} \\ &= \frac{\mathbb{P}\{Y_t^{**} = j \,|\, Y_0^{**} = i\}\mathbb{P}\{A_\to \,|\, Y_0 = j\}}{\mathbb{P}\{A_\to \,|\, Y_0 = i\}} \\ &= \left(e^{tG^{**}}\right)_{ij} \frac{x_j}{x_i}.\end{aligned}$$

Therefore, if $\tau_\to$ is the first time that $Y^\to$ hits f and S is the first time that Y leaves s, then, by conditioning on S and Y_S,

$$\begin{aligned}\mathbb{E}[e^{-\lambda\tau_\to}] &= \sum_{i\neq s} \frac{G_{si}}{-G_{ss}} \lambda \int_0^\infty \mathbb{P}\{Y_t^\to = f \,|\, Y_0^\to = i\} e^{-\lambda t} dt \\ &= \sum_{i\neq s} \frac{G_{si}}{-G_{ss}} \frac{\lambda\left((\lambda - G^{**})^{-1}\right)_{if}}{x_i} \\ &= \sum_{i\neq s} \frac{G_{si}}{-G_{ss}} \frac{\lambda\left((\lambda - G^{**})^{-1}\right)_{if}}{x_i}.\end{aligned}$$

Note by a quick computation with Lemma 3.5.4 that if we define x by

$$x_i = \lim_{\lambda \to 0} \lambda(\lambda - G^{**})^{-1}_{if}$$

then x is the unique solution to $G^{**}x = 0$ with $x_f = 1$ and $x_s = 0$, and so coincides with our definition of x in (3.20).

Differentiating and using Lemma 3.5.4 gives (3.26). □

Inverses and singular matrices

Lemma Let A be a block upper triangular matrix of the form

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline 0 & 0 \end{array}\right],$$

where the dimensions of A_{11}, A_{12} and A are respectively $m \times m$, $m \times k$ and $(m+k) \times (m+k)$, and suppose that $(\lambda - A_{11})$ is invertible for all $\lambda \in [0, \epsilon)$ for some $\epsilon > 0$. Then,

$$\lim_{\lambda \to 0} \lambda(\lambda - A)^{-1} = \left[\begin{array}{c|c} 0 & -A_{11}^{-1}A_{12} \\ \hline 0 & I \end{array}\right],$$

and

$$(-1)^n \partial_\lambda^n \lambda(\lambda - A)^{-1}\Big|_{\lambda=0} = \left[\begin{array}{c|c} -n!(-A_{11})^{-n} & n!(-A_{11})^{-(n+1)}A_{12} \\ \hline 0 & 0 \end{array}\right].$$

Furthermore, if c is a vector of length k, then the unique solution to

$$\begin{aligned} Ax &= 0 \\ (x_{m+1}, \ldots, x_{m+k}) &= c \end{aligned}$$

is

$$x = \left[\begin{array}{c} -A_{11}^{-1}A_{12}c \\ \hline c \end{array}\right].$$

Proof[Proof of Lemma 3.5.4.] By the block inversion formula for a 2×2 block matrix,

$$(\lambda - A)^{-1} = \left[\begin{array}{c|c} (\lambda - A_{11})^{-1} & \frac{1}{\lambda}(\lambda - A_{11})^{-1}A_{12} \\ \hline 0 & \frac{1}{\lambda}I \end{array}\right].$$

Using the following identity for differentiating the inverse of a matrix

$$\partial_t B(t)^{-1} = -B(t)^{-1}(\partial_t B(t))B(t)^{-1},$$

and differentiating each entry, we see that

$$(-1)^n \partial_\lambda^n \lambda(\lambda - A)^{-1} = \left[\begin{array}{c|c} n!(\lambda - A_{11})^{-(n+1)} A_{11} & n!(\lambda - A_{11})^{-(n+1)} A_{12} \\ \hline 0 & 0 \end{array}\right].$$

Since A_{11} is invertible, we can take the limit as $\lambda \to 0$ from above.

That x solves $Ax = 0$ is obvious; we need only justify that it is the unique solution. This follows since A_{11} is invertible, and so A has rank m.

□

3.5.5 Decomposition into parallel modules

In the elongation regulated model, there are two processes that must come to completion for transcription to occur: promoter assembly and enhancer recruitment. These two processes evolve independently of one another (in parallel) and transcription may begin only when both are in the correct state simultaneously. The sequential decomposition method does not apply to the elongation model, so in this section we introduce tools for simplifying the analysis of such parallel compositions of chains. A *parallel* composition of chains is a collection of noninteracting Markov chains, each with a distinguished final state, each of which must be in its final state for transcription to occur. In general, it does not seem possible to express the first two moments of the traversal time for the composite chain with only the first two moments of each component chain or similar quantities, as the example in Subsection 3.5.5 will show. It is still possible to compute quantities for the composite chain in terms of the component chains through a spectral decomposition, which we discuss in this section. In Subsection 3.5.5, we discuss a simple approximation for the case of a two–state enhancer chain.

Formally, we again have a sequence of continuous-time Markov chains X^k on a sequence of state spaces $\mathcal{X}^k$, for $k \in \{1, 2, \ldots, n\}$, each with two distinguished (and distinct) states s_k and f_k. We assume that each has at most one absorbing state, and so has a generator that can be diagonalized by an invertible matrix. The composite Markov chain X is simply the product chain on the Cartesian product $\prod_{i=1}^n \mathcal{X}^k$ given by $X_t = (X_t^1, \ldots, X_t^n)$, where $X^1, \ldots, X^n$ evolve independently. A state x for the product chain X is of the form $x = (x_1, x_2, \ldots, x_n)$, where $x_k \in \mathcal{X}^k$ for each $1 \leq k \leq n$. We denote by G the transition rate matrix of the full chain X, and G^k for the transition rate matrix of X^k.

First we review a few facts about these composite matrices. The generator, G, is defined as follows. If states x and y only differ in a single entry: $x_i \neq y_i$, but $x_j = y_j$ for all $j \neq i$, then $G_{xy} = G^i_{x_i y_i}$. If x and y differ in more than one entry, then $G_{xy} = 0$; and $G_{xx} = -\sum_{y \neq x} G_{xy}$. Since the chains are independent, the transition probability matrix $P(t)$ for the composite chain is the Kronecker product of the transition probability matrices $P^i(t)$ for each subchain: $P(t)_{xy} = \prod_i P^i(t)_{x_i y_i}$.

Below, we will want to compute $(\lambda I - G)^{-1}$, which we can do using information about only the component chains. Suppose each G^k has eigenvalues λ^k_i with corresponding left and right eigenvectors ℓ^k_i and r^k_i, for $1 \leq i \leq m_k$. If an eigenspace has dimension greater than one (as will be the case if f_k is absorbing) then any choice of of eigenvectors that spans the eigenspace may be made as long as ℓ^k_i is orthogonal to r^k_j for $i \neq j$. If for each G^k we pick some right eigenvector and form a vector in the product space in the natural way, then the resulting product vector will be a right eigenvector of G with eigenvalue equal to the product of the respective eigenvalues. Under our assumptions, all right eigenvectors of G are formed in this way and they span the product space. Formally, we know that for each $i_1, \ldots, i_n$ with $1 \leq i_k \leq m_k$, the product $\lambda_{i_1,\ldots,i_n} = \prod \lambda_{i_j}$ is an eigenvalue for G, with corresponding left and right eigenvectors $\ell_{i_1,\ldots,i_n}(x) = \prod_k \ell^k_{i_k}(x_k)$ and $r_{i_1,\ldots,i_n}(x) = \prod_k r^k_{i_k}(x_k)$, where $x = (x_1, \ldots, x_n)$. To be clear about notation, $\ell = \ell_{i_1,\ldots,i_n}$ is a vector in the product space $\prod \mathcal{X}^k$, and so is indexed by elements $x \in \prod \mathcal{X}^k$ of the form $x = (x_1, \ldots, x_n)$. We form the product vector ℓ be saying that $\ell_{i_1,\ldots,i_n}(x) = \prod_k \ell^k_{i_k}(x_k)$. Furthermore, this provides a spectral decomposition of G, giving the result that

$$(\lambda I - G)^{-1}_{xy} = \sum_{i_1,\ldots,i_n} (\lambda - \prod_k \lambda^k_{i_k})^{-1} \prod_k r^k_{i_k}(x_k) \ell^k_{i_k}(y_k),$$

where the sum is over distinct n-tuples of indices with $1 \leq i_k \leq m_k$.

We suppose that transcription (or the jump to the next stage) occurs at rate ρ while X is in state $f = (f_1, \ldots, f_n)$. Thus, we are interested in the time until death of the chain X if it is killed at rate ρ while in state f. The Feynman-Kac formula gives a way to compute the Laplace transforms and moments of the killing times — for an excellent discussion, see [38]. If Π is the projection matrix with $\Pi_{ff} = 1$ and $\Pi_{ij} = 0$ otherwise, and if τ is the killing time, then

$$\mathbb{E}^x\left[\exp(-\lambda\tau)\right] = \rho\left(\lambda I - G + \rho\Pi\right)^{-1}_{xf}$$

This is equation (48) in [38] (but beware the differences in notation).

Since G is of product form, and we can find its spectral decomposition in terms of the spectral decompositions of the component chains, it would be nice to compute $(\lambda I - G + \rho\Pi)^{-1}$ in terms of $(\lambda I - G)^{-1}$. This turns out to be possible, thanks to the following lemma, which is a special case of the Matrix Inversion Lemma, also known as the Sherman-Morrison-Woodbury formula [57]. This allows us to compute an explicit expression for $\mathbb{E}[e^{-\lambda\tau}]$, if we have a spectral decomposition of each product chain.

Lemma Let B be an invertible $m \times m$ matrix, and let u and v be m-dimensional vectors such that $v^t Bu \neq -1/\rho$. Then,

$$(B + \rho uv^t)^{-1} = B^{-1} - \frac{\rho}{1 + \rho v^t B^{-1} u} B^{-1} uv^t B^{-1}.$$

Remark The lemma allows us to compute the inverse of a rank-one correction to B easily using only B^{-1} — if u and v are the i^{th} and j^{th} basis vectors respectively, then $(B^{-1}uv^tB^{-1})_{xy} = B^{-1}_{xi}B^{-1}_{jy}$, while $v^tB^{-1}u = B^{-1}_{ij}$.

Using this lemma, if we let $q = (\lambda I - G)^{-1}_{ff}$, we may write

$$(\lambda I - G + \rho\Pi)^{-1}_{xy} = (\lambda I - G)^{-1}_{xy} - \frac{\rho}{1 + \rho q}(\lambda I - G)^{-1}_{xf}(\lambda I - G)^{-1}_{fy},$$

and hence

$$\begin{aligned}\mathbb{E}^x\left[\exp(-\lambda\tau)\right] &= \rho(\lambda I - G)^{-1}_{xf}\left\{1 - \frac{\rho}{1 + \rho q}(\lambda I - G)^{-1}_{ff}\right\} \\ &= \frac{\rho}{1 + \rho(\lambda I - G)^{-1}_{ff}}(\lambda I - G)^{-1}_{xf}.\end{aligned}$$

Remark If we take $\rho \to \infty$, we get the expression for Laplace transform of the first hitting time of f (denoted here by τ_f) as the ratio of two terms in the resolvent

$$\mathbb{E}^x[\exp(-\lambda\tau_f)] = \frac{(\lambda I - G)^{-1}_{xf}}{(\lambda I - G)^{-1}_{ff}}.$$

This relation may, of course, also be obtained by recognizing that $\lambda(\lambda I - G)^{-1} = \int_0^\infty \lambda e^{-\lambda t} P_t dt$ and using the strong Markov property.

A simple approximation for a parallel two-state enhancer

We now consider a special case. Let X be a Markov chain with distinguished states s and f, and further assume that once X is in the final state f, it does not leave. Let Y be an independent two-state chain Y that takes values in $\{0, 1\}$, that moves from 0 to 1 at rate α and moves from 1 to 0 with rate β. The transition probabilities for Y are

$$\mathbb{P}\{Y_t = 1 \,|\, Y_0 = 0\} = \frac{\alpha}{\alpha+\beta}(1 - e^{-(\alpha+\beta)t}).$$

We construct a chain on the product space by saying that transcription occurs once X is in state f and Y is in state 1. Let τ be the first time this occurs,

$$\tau = \inf\{t \geq 0 : X_t = f \text{ and } Y_t = 1\}.$$

Let τ_X be the first time that X hits the state f, and let W be an independent exponential random variable with rate α. Then, since X does not leave state f,

$$\tau \overset{d}{=} \tau_X + (1 - Y_{\tau_X})W \tag{3.27}$$

whence

$$\begin{aligned}\mathbb{E}[\exp(-\lambda\tau)] = &\left(\frac{\alpha}{\alpha+\beta} + \frac{\beta}{(\alpha+\beta)(\alpha+\lambda)}\right)\mathbb{E}[\exp(-\lambda\tau_X)] \\ &- \frac{\alpha}{\alpha+\beta}\left(1 + \frac{1}{\alpha+\lambda}\right)\mathbb{E}[\exp(-(\lambda+\alpha+\beta)\tau_X)]\end{aligned}$$

and (working directly from (3.27))

$$\begin{aligned}\mathbb{E}[\tau] &= \mathbb{E}[\tau_X] + \frac{1}{\alpha+\beta}\mathbb{E}\left[1 - e^{-(\alpha+\beta)\tau_X}\right] \\ \mathbb{E}[\tau^2] &= \mathbb{E}[\tau_X^2] + 2\frac{1}{\alpha+\beta}\mathbb{E}\left[\tau_X\left(1 - e^{-(\alpha+\beta)\tau_X}\right)\right] + \frac{2}{\alpha(\alpha+\beta)}\mathbb{E}\left[1 - e^{-(\alpha+\beta)\tau_X}\right].\end{aligned}$$

Therefore, it seems that computing the moments of τ requires the full Laplace transform of τ_X. However, if τ_X is reasonably large relative to $\alpha + \beta$, then a good approximation is

$$\begin{aligned}\mathbb{E}[\tau] &\approx \mathbb{E}[\tau_X] + \frac{1}{\alpha+\beta} \\ \mathbb{E}[\tau^2] &\approx \mathbb{E}[\tau_X^2] + 2\frac{1}{\alpha+\beta}\mathbb{E}[\tau_X] + \frac{2}{\alpha(\alpha+\beta)}.\end{aligned}$$

3.5.6 Experimental Approaches to testing hypothesis about total mRNA numbers

A central prediction of our work is the potential effect of elongation regulation on cell-to-cell variation in the total number of transcripts. Recent advances in molecular imaging could be adapted to test this hypothesis directly in the appropriate system, by direct counting of individual, fluorescently labeled cytoplasmic mRNAs in each cell in a fixed embryo, and comparing the count between all sister cells in a common tissue.

Current developments in labeling and imaging technology allow for sensitive detection of individual molecules [126, 125, 124]. If the concentration of mRNA is sufficiently large that individual molecules are within half a wavelength of the detection light, they can be resolved using Stochastic Optical Reconstruction Microscopy (STORM), a sub-diffraction limited method of imaging wherein a small fraction of the labeled samples are photo-switched into the detectable emission spectra at a time, imaged until bleaching, and then a new subset is photo-switched by a short pulse into the detectable spectra [67]. Individual Gaussian or Airy functions are then fit to the large collection of (overlapping) diffraction limited spots to find their centers, thereby allowing the spots be separated and individually counted at a 10–100nm resolution, depending on the set-up. For review of this technique, we direct the reader to Bates 2008 [5]. For a combined perspective on single molecule imaging and its application to transcription, we direct the reader to “Single-molecule approaches to stochastic gene expression”, Raj and van Oudenaarden, *Annual review of biophysics* [126].

Chapter 4

Shadow Enhancers and Robust Control

4.1 Background:

4.1.1 Too many enhancers

Whole genome transcription factor binding data for Twist, Dorsal and Snail, important transcription factors which control dorsal-ventral patterning in Drosophila [171] facilitated the discovery of a range of new enhancers [171, 63].

As a more complete assembly of early enhancers emerged for these dorsal-vental patterning genes, it became apparent that many of the newly discovered enhancers had already been discovered. That is to say, entirely different genomic sequences in the vicinity of the same gene had already been shown to drive expression in a similar set of cells during that same time in development. For example, the peak of Twist and Dorsal binding in the intron of the Tim172b locus drives an essentially identical mesodermal expression pattern as the neighboring gene, *snail.* This expression pattern is also produced by a 2.8 kb fragment of the five-prime sequence of the *snail* gene. Pol-II CHiP-chip binding data show that the Tim17b2 gene is off. It might have been that these newly identified enhancers only actually work as enhancers when moved artificially near a heterologous promoter, as with a standard enhancer testing assay. However subsequent experiments by Mike Perry and I using bacterial artificial chromosomes (BAC) to integrate a 20kb section around the *snail* locus spanning both enhancers show that this ‘shadow’ en-

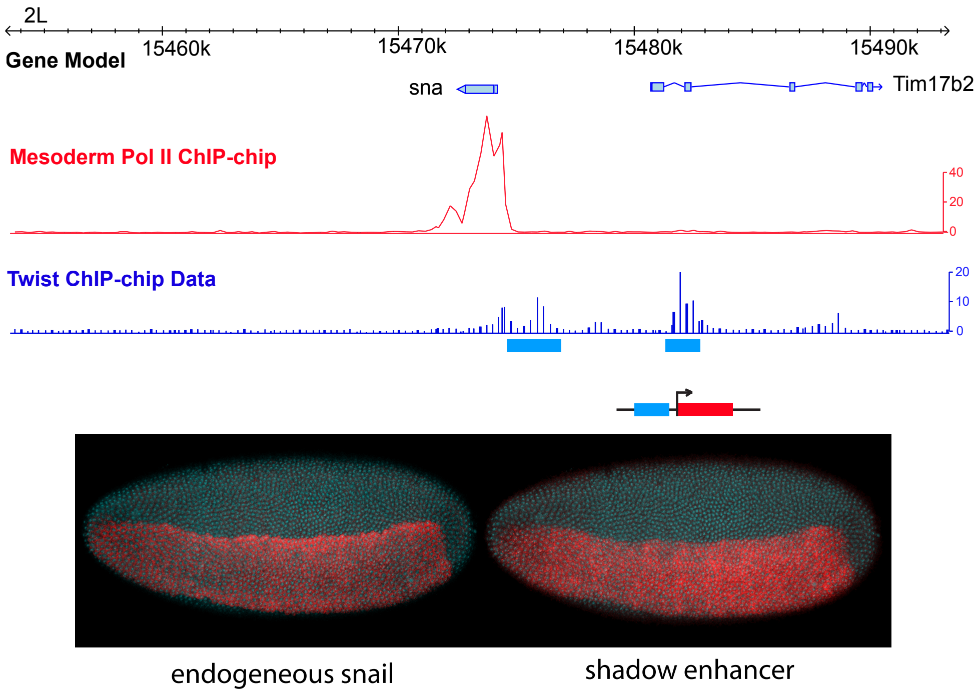

Figure 4.1: **CHiP-chip data reveals extra *snail* enhancers:** Twist CHiP-chip data picks out the bipartate *snail* proximal enhancer and the distal ‘shadow’ enhancer inside the intron of the neighboring gene, Tim17b2. Though *snail* is highly transcribed in this mesodermal tissue, Tim17b2 is off, as clearly seen from the Pol II CHiP-chip. Below, embryos comparing the endogenous *snail* expression pattern and the reporter pattern driven by the reporter (stains are false colored for comparison, but were imaged in the same embryo in orthologous channels (Alexa555 and Alexa488 respectively).

hancer does indeed drive expression from the *snail* promoter even when kept in its endogenous position on the far side of the Tim17b2 promoter, as I will discuss in more detail below.

Recent work has established that many other systems are rife with these ‘shadow enhancers’ (for example, Mike Perry’s analysis of the gap gene system, which I joined to test possible function or shadow enhancers for neurogenic expression of the transcription factor COE in Ciona intestanalis (Blair Gaineous, unpublished data). In the following chapter I present our investigations into possible functional mechanisms which may account for this apparent redundancy, and discuss some of the newly discovered and confirmed additional ‘shadows’ along the way.

4.1.2 Note on terminology

Any new enhancer which drove gene expression in a pattern for which an enhancer had previously been characterized was termed a 'shadow enhancer'. At various points we hypothesized that 'shadow enhancer' were evolutionarily younger [64], more distant and consequently weaker, and several other somewhat arbitrary additional distinctions (like being in an intron). During this time we referred to the previously discovered partner as the 'primary enhancer', meaning first discovered rather than necessarily more important enhancer. Deeper investigations into conservation by Michael Perry (looking at the Drosophilids) and Jessica Cande (looking out as far as beetles and mosquitos) [21] showed in fact no clear correlation between evolutionary conservation and whether or not an enhancer had been termed a shadow.

I performed quantitative in situ hybridizations on reporters for these various enhancers and found no consistent pattern with the apparent 'strength' of the enhancer (as determined by the fraction of cells simultaneously transcribing the reporter in the induced region, and whether it was the 'shadow enhancer' or not. For example, for *Kruppel* and *sog* (at low temperature), the shadow drives a more complete pattern of activation. For *hunchback* and *knirps* the proximal enhancer drives a more complete expression pattern (discussed more in depth below). Due to the historical approach of enhancer finding by blind bashing of five-prime sequences, all of the newly discovered shadows are further away from the promoter than the originally discovered enhancers. As a result is has also been convenient to simply refer to the two enhancers as 'distal' and 'proximal', which does not imply any subordination of one to the other.

Due to this lack of subordinate function which many feel is implied by the term 'shadow', I now favor the simple terminology distal and proximal. However it is confusing and insufficient to refer to a gene which has shadow enhancers as a gene that has 'distal enhancers'. Consequently I retain for this purpose the use of the word 'shadow', as it is more convenient to refer to genes as simply 'having a shadow enhancer' rather than the 'having multiple enhancers which drive similar patterns of expression at overlapping times during development'. I also believe this terminology is less confusing than 'having apparently redundant enhancers'. It also plays appropriate homage to the historical precedent that started these investigations.

4.2 Dual enhancers support robust *snail* expression and gastrulation

4.2.1 Introduction

Recent whole-genome analyses in Drosophila suggest that critical developmental control genes sometimes contain shadow enhancers [64]. These can be located in remote positions, including the introns of neighboring genes. They nonetheless produce patterns of gene expression that are the same or similar as those produced by primary enhancers. It was suggested that shadow enhancers help foster robustness in gene expression in response to environmental or genetic perturbations [17, 120]. Here, we critically test this hypothesis by employing a combination of BAC recombineering and quantitative confocal imaging methods [17, 162]. Evidence is presented that the *snail* gene is regulated by a distal shadow enhancer located within the neighboring Tim17b2 locus. *snail* encodes a zinc finger transcription factor that has been implicated in epithelial/mesenchyme transitions (EMT) in a broad spectrum of developmental processes and cancers [89, 83, 4]. Removal of the proximal primary enhancer does not significantly perturb *snail* function, including the repression of neurogenic genes and formation of the ventral furrow during gastrulation at normal temperatures of development. However, at elevated temperatures there is sporadic loss of *snail* expression and coincident disruptions in gastrulation. Similar defects are observed at normal temperatures upon reductions in the levels of Dorsal, a key activator of *snail* expression (reviewed in [63]). Altogether, these results suggest that shadow enhancers represent a novel mechanism of canalization, whereby complex developmental processes bring about one definite end-result regardless of minor variations in conditions [164].

Despite both intrinsic and environmental sources of noise, which introduce variability in complex developmental processes, the patterning of the Drosophila embryo unfolds with high fidelity (e.g., [51]). It has been postulated that gene interactions in developmental regulatory networks can channel these variable inputs into faithful outcomes, as a ball bouncing inside of a funnel is channeled to the center, a process termed canalization [164]. Here we present evidence that shadow enhancers [64] are important mediators of canalization, ensuring reliable and robust expression of critical patterning genes.

4.2.2 Results: redundant enhancer provide robust *snail* expression

snail is a key determinant of dorsal-ventral patterning [89, 83, 20, 69]. It encodes a zinc finger repressor that establishes a sharp boundary between the presumptive mesoderm and neurogenic ectoderm, and is essential for the formation of the ventral furrow and invagination of the mesoderm. Whole genome ChIP-chip assays identified a cluster of Dorsal and Twist (key activators of *snail* expression) binding sites in the immediate 5 flanking region of the *snail* transcription unit that coincide with the known enhancer [69, 171]. Unexpectedly, these studies also identified a second cluster of binding sites within the neighboring Tim17b2 locus, located ≈7 kb upstream of snail. A small genomic DNA fragment (≈1 kb) encompassing this second cluster of binding sites was attached to a lacZ reporter gene and expressed in transgenic embryos (Fig. 4.2). The fusion gene exhibits localized expression in the presumptive mesoderm, similar to that seen for the endogenous gene (e.g., Fig. 4.2) or obtained with the proximal enhancer (the first 2.8 kb of the 5 flanking region; see ref. [69]). We arbitrarily refer to the newly identified distal enhancer as the shadow enhancer and the original, proximal enhancer as the primary enhancer [64].

A *snail* fusion gene containing only the primary enhancer rescues the gastrulation of at least some *snail* mutants in a population of mutant embryos [60]. Since *snail* is essential for the coordinated invagination of the mesoderm during early gastrulation, variability in expression could lead to occasional disruptions in morphogenesis. Perhaps the additional enhancer provides a mechanism for suppressing such variability, thereby ensuring robust expression in large populations of embryos. This hypothesis was motivated in part by previous preliminary evidence that neurogenic genes with shadow enhancers show less sensitivity to changes in activator concentration than similar genes lacking shadows [17].

An alternative view is that the proximal and shadow enhancers are primarily responsible for controlling distinct dynamic aspects of the *snail* expression pattern, rather than functioning in an overlapping manner during mesoderm invagination. An expectation of the former robustness hypothesis is that transgenes containing either enhancer alone should be sufficient to induce gastrulation in *snail* mutant embryos. We tested this possibility by creating a series of recombineered BACs [162, 161] containing a 25 kb genomic interval encompassing the *snail* and Tim17b2 loci (Fig. 4.2).

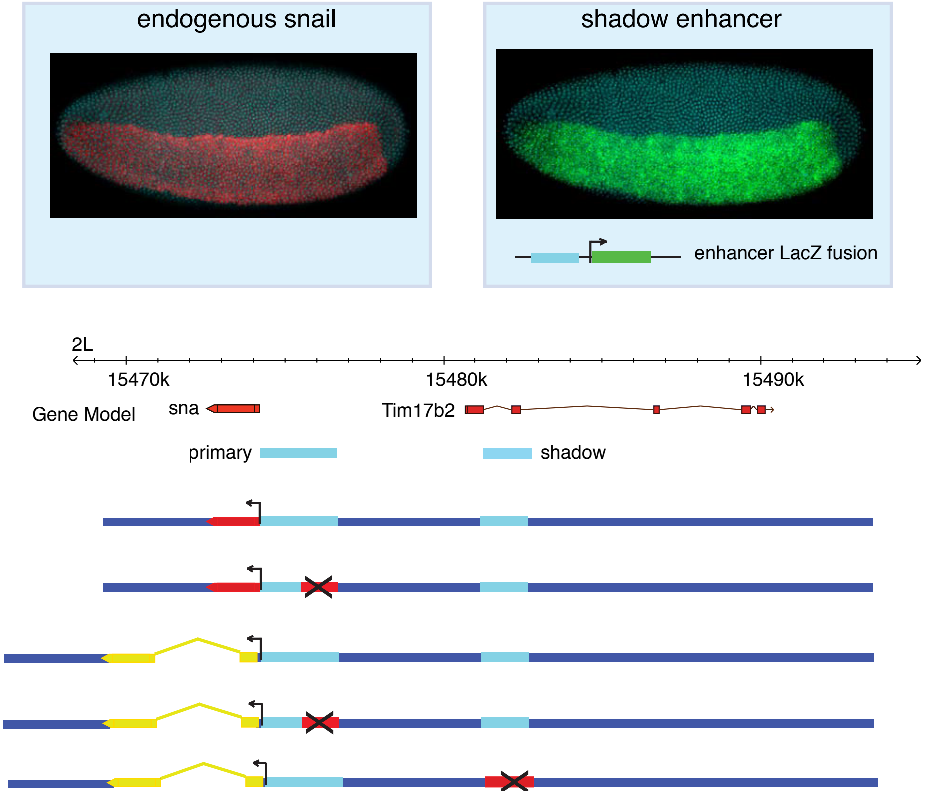

Figure 4.2: **Identification of a *snail* shadow enhancer.** The *snail* gene is expressed in the presumptive mesoderm, (top left in red). An intronic region in neighboring Tim17b2 was shown to be bound by transcription factors that regulate *snail* [171] and here is shown to drive expression of a lacZ fusion gene in the mesoderm in a pattern qualitatively similar to the endogenous gene (upper right panel). Below, schematic representations of the BAC constructs used in subsequent experiments are aligned to the gene model. In all figures, anterior is to left, dorsal is at top, unless indicated.

Comparable BACs were prepared that either contain or lack the proximal enhancer. This enhancer was not simply deleted, but a 1 kb segment containing critical Dorsal activator elements was replaced with a random DNA sequence (see Experimental Procedures) in order to retain normal spacing of the regulatory region.

To measure the effect that different enhancers have on transcriptional activity we developed a reporter system for detecting nascent transcripts. The endogenous yellow gene is not transcribed until late in development and contains a large intron (e.g., [72, 114]), making it an ideal reporter for the detection of de novo transcripts by in situ hybridization. In contrast, the *snail* transcription unit lacks introns and is therefore not amenable to quantitative in situ hybridization methods that rely on intronic probes. Consequently, a series of BACs were created that contain yellow in place of snail. These BACs contain both enhancers or have either the primary or shadow enhancer replaced with random DNA (Fig 4.2). All of the aforementioned BACs were inserted in the same chromosomal location on 2L using phiC31 targeted integration [162, 52, 16].

BACs containing the *snail* gene were crossed into a mutant background with a deletion spanning the entire *snail* transcription unit (Df (2L)osp29) along with a marked balancer to identify homozygous *snail* null mutants . As noted earlier, the reciprocal situation, proximal enhancer without shadow, can sometimes rescue gastrulation [60]. Mutant embryos homozygous for the *snail* deficiency chromosome (osp29) are easily recognized by the absence of *snail* expression and ectopic single-minded (sim) expression, a key regulator of midline formation within the central nervous system that is normally excluded from the mesoderm by the Snail repressor [78, 109] (Fig. 4.3E,F).

There is neither a ventral furrow nor subsequent ingression of the mesoderm in these mutants (e.g., [89, 83]). BAC transgenes containing both enhancers (Fig. 4.3A) or just the shadow enhancer alone (Fig. 4.3B) rescue gastrulation of mutant embryos (Fig. 4.3C,D; compare with E,F). In both cases, a complete ventral furrow is formed, followed by invagination of the mesoderm indistinguishable from that seen in wild-type embryos. Both BACs restore *snail* expression in the presumptive mesoderm, and sim transcripts are restricted to lateral regions that form the ventral midline of the CNS after gastrulation. These observations, along with previous studies (e.g., [60]), indicate that neither the primary nor shadow enhancer is necessary for the gastrulation of embryos raised at optimal, permissive conditions.

Although the shadow enhancer is sufficient for generating a qualitatively

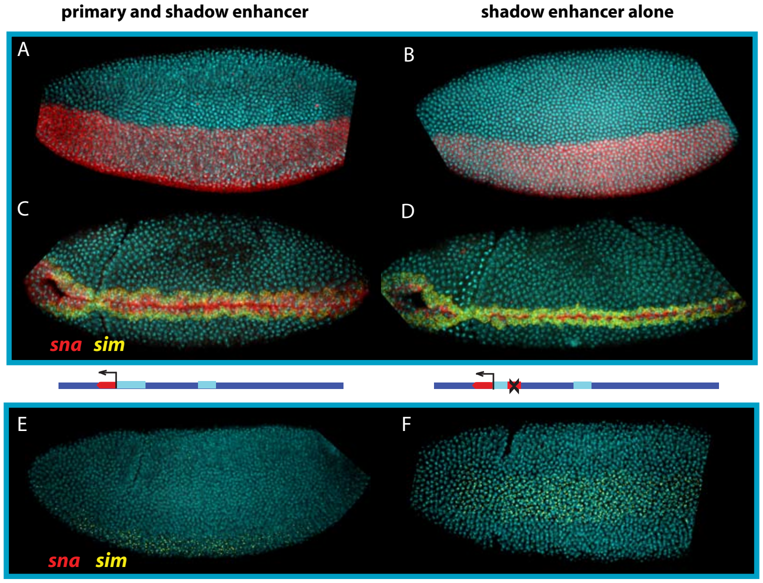

Figure 4.3: **The *snail* shadow enhancer rescues gastrulation.** A. The rescue BAC construct in a *sna* mutant background drives *sna* expression (red) uniformly throughout the mesoderm in cycle 14 embryos. B. The pattern driven only by the BAC with the primary enhancer deleted is qualitatively similar. C. During gastrulation, all *sna* expressing cells migrate into the interior of the embryo. A single row of cells flanking the *sna* domain express sim, shown in yellow. D. *sna* driven without the primary enhancer is sufficient to induce normal gastrulation and normal sim expression when these embryos are raised at 22C. E. In embryos lacking the *snail* BAC rescue construct, no *sna* is expressed. Instead, sim is expressed throughout the ventral region. Without *sna* there is no mesodermal invagination. Lateral view. F. Embryo as in (E), mesodermal view.

normal pattern of *snail* expression, additional assays were done to determine whether there might be subtle changes in expression. Quantitative confocal imaging methods were used to investigate this possibility (see [17]). As mentioned earlier, BAC transgenes were prepared that contain the yellow reporter gene in place of the *snail* transcription unit. In situ hybridization assays with intronic probes permit direct detection of yellow de novo transcripts, and hence, precise measurements of *snail* transcription with single cell (nucleus) resolution (see Fig 4.4). At normal culturing temperatures, 22C, there is no discernible difference in the initial de novo transcription patterns of BAC transgenes containing both enhancers (Fig. 4.5A) or just a single enhancer, either the primary enhancer or shadow enhancer (Fig. 4.5B). In the majority of cases more than 90% of the nuclei in the presumptive mesoderm express yellow transcripts.

Less reliable expression is observed for BAC transgenes containing a single enhancer at elevated temperatures, 30C (Fig. 4.5C,D). More than 20% of the nuclei in the presumptive mesoderm lack yellow transcripts in over half of the embryos expressing the BAC transgene without the shadow enhancer. This effect is even more pronounced upon removal of the primary enhancer. The same cut-off value, absence of yellow transcripts in at least 20% of all mesodermal nuclei, occurs in over three-fourths of these embryos (Fig 4.5). In contrast, the BAC transgene containing both the primary and shadow enhancers continues to display nearly complete patterns of de novo transcription at the elevated temperature.

Similar results were obtained in response to genetic perturbations (Fig. 4.7A,B). For example, the yellow transgene BAC containing both enhancers exhibits a normal pattern of expression in embryos derived from *dl*/+ mothers containing half the normal dose of the Dorsal gradient (Fig. 4.7A). The distribution of nuclei failing to maintain active expression is similar to that seen for wild-type embryos (Fig 4.7C). However, the comparable BAC transgene containing only the shadow enhancer exhibits erratic patterns of activation in these embryos, particularly in lateral regions (Fig. 4.7B, quantification in C). These results, along with the preceding analysis of embryos grown at elevated temperatures, suggest that the *snail* shadow enhancer helps ensure accurate and reproducible patterns of gene expression in large populations of embryos subject to genetic and environmental perturbations.

The preceding results document quantitative changes in the variability and reliability of *snail* expression upon removal of the primary or shadow enhancer. We next asked whether such variation causes changes in cellular

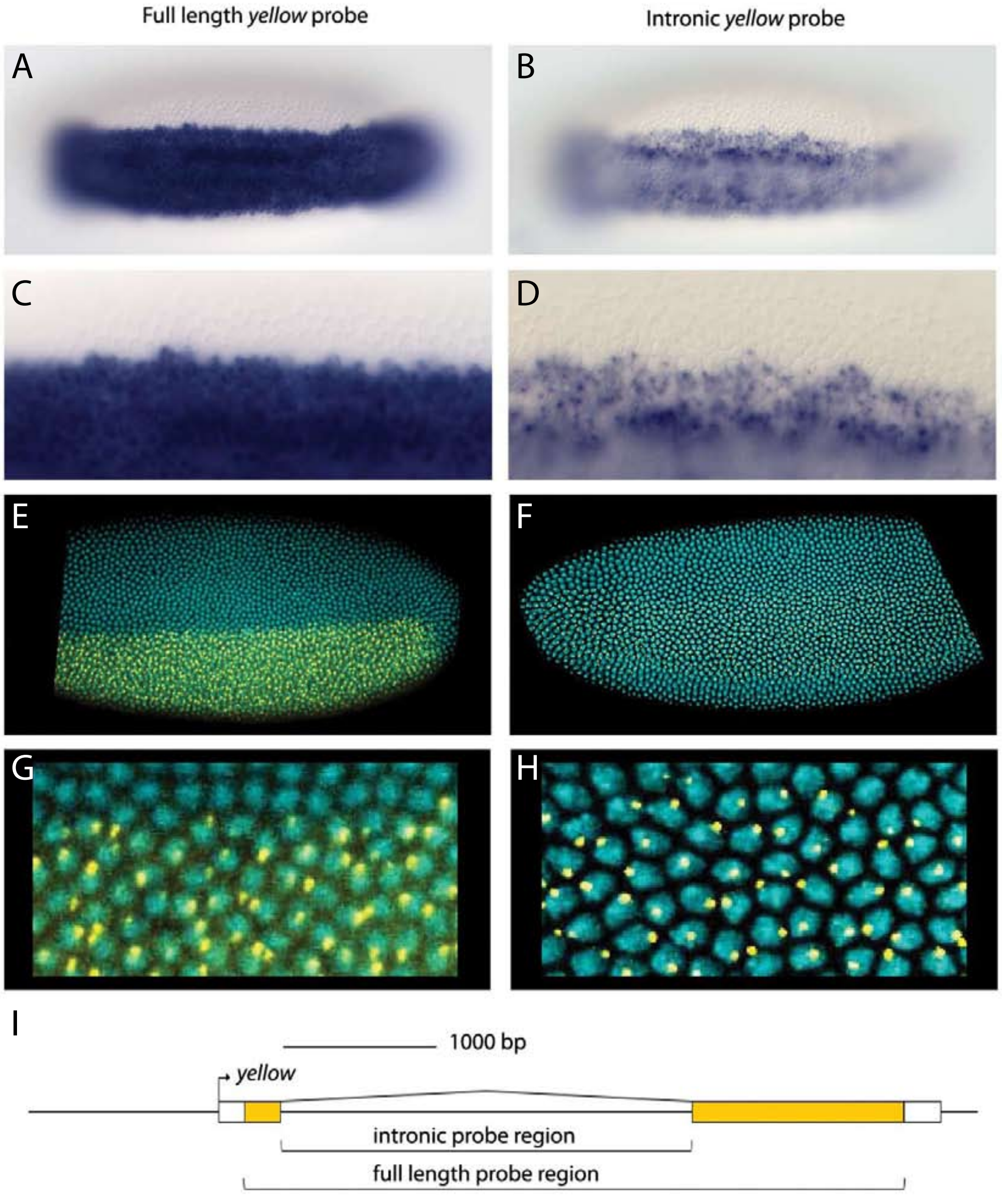

Figure 4.4: Comparison of full-length yellow probe (A,C,E,G) to intronic yellow probe (B,D,F,H) stained both colormetrically (A-D) and fluorescently (E-H)

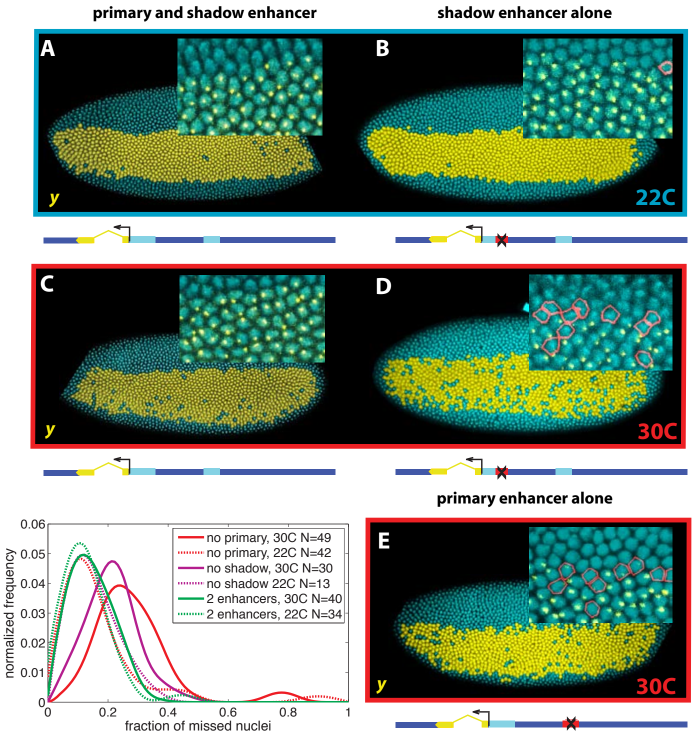

Figure 4.5: **Multiple enhancers ensure robust gene expression under different thermal conditions.** A. Visualization of expression of the yellow reporter gene from the BAC containing the *sna* locus, stained for the yellow intron. Cells actively transcribing the reporter are shown in yellow. Intronic probes show a single bright point of transcription inside actively transcribing nuclei (inset); all embryos are heterozygous and have one copy of the reporter. Nuclei that express the endogenous gene but not the reporter are outlined in red. A schematic representation of the BAC is shown below the embryo. B. At 22C a similar degree of uniform expression is exhibited by embryos carrying a yellow BAC lacking the primary enhancer. C. At 30C embryos with both enhancers still show straight boundaries and a small percent of inactive nuclei. D. Embryos lacking the primary enhancer at 30C show substantially more ragged boundaries of expression and a greater percent of inactive cells in the mesoderm. E. Embryos lacking the shadow enhancer are similar to those lacking the primary at both temperatures. Lower left: Frequency distributions of the fraction of cells in the *sna* expressing region that lack yellow expression are plotted for each of the 6 different embryo populations. N indicates the number of embryos in each population sample.

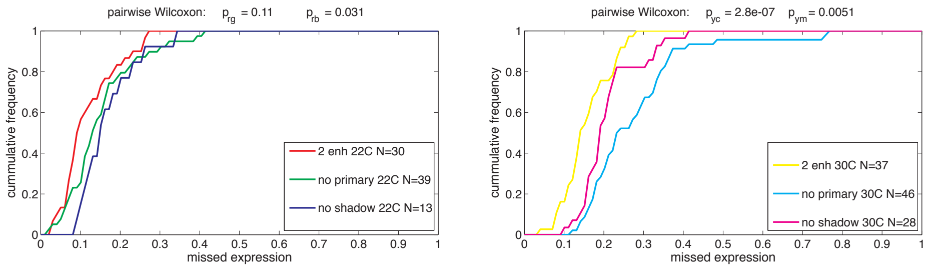

Figure 4.6: **Thermal Stress: Cumulative frequency plots for snail-yellow expression** at 22C and 30C. p-values from pairwise wilcoxon comparison for the indicated tracks are labeled on the figures (p_{rg} gives the p-value comparing the red and green curves).

morphogenesis, particularly the formation of the ventral furrow and subsequent invagination of the mesoderm (Fig. 4.7D,E). *snail* mutant embryos carrying BACs with both enhancers (Fig. 4.7D) or just the shadow enhancer (Fig. 4.7E) were grown at elevated temperatures, 30C. Embryos carrying the transgene with both enhancers exhibit normal patterns of gastrulation (Fig. 4.7D). In contrast, comparable embryos lacking the primary enhancer display erratic patterns of gastrulation, including the formation of incomplete ventral furrows that do not extend along the entire germband (Fig. 4.7E) and disruptions in the symmetry of the involuted mesodermal tube (see Fig. 4.9). As shown earlier, such defects are not observed at normal temperatures, 22C (Fig. 4.3 and 4.9).

4.2.3 Discussion: shadow enhancers may ensure robustness

We have presented evidence that the *snail* shadow enhancer located within the Tim17b2 locus helps ensure reliable and reproducible patterns of *snail* expression in the presumptive mesoderm during gastrulation. BAC transgenes lacking either the primary enhancer or the shadow enhancer display erratic patterns of de novo transcription at elevated temperatures. We propose that shadow enhancers come to be fixed in populations by ensuring robustness in the activities of key patterning genes such as snail. Increases in temperature should cause less stable occupancy of critical binding sites, but an additional enhancer could suppress this noise by increasing the probability of gene activation. This increased time of active transcription per cell might increase the

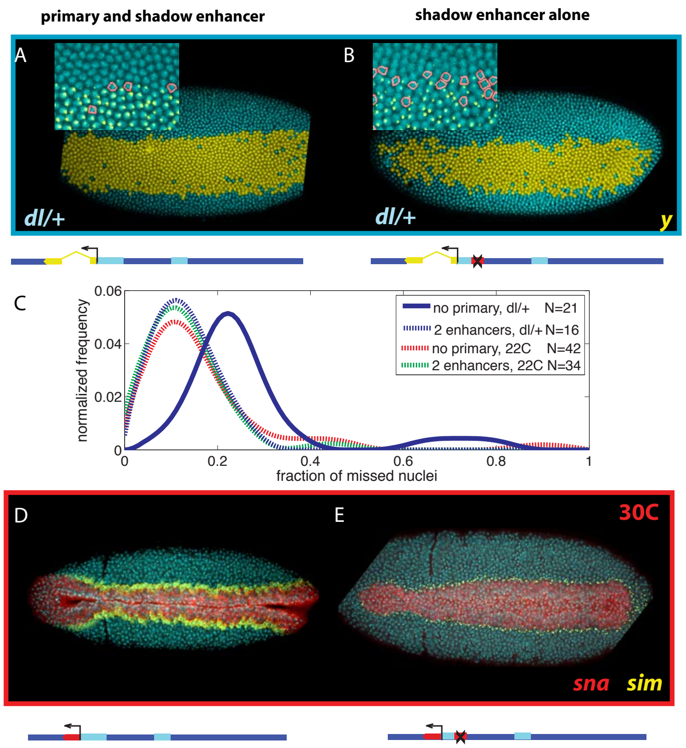

Figure 4.7: **The effect of intrinsic and extrinsic variability.** A. Embryos from dorsal heterozygote mothers raised at 25C show uniform yellow expression when driven with both enhancers. Only a few cells are lacking active expression (inset). B. Embryos with a single enhancer in this background show substantially greater loss of expression and ragged boundaries. C. The distribution nuclei which fail to maintain active transcription shifts to the right in the dorsal heterozygote background only for embryos lacking one of the enhancers. D. All observed embryos raised at 30C (N=28) from a population heterozygous for the BAC-constructs containing both enhancers gastrulate normally, forming a straight ventral furrow; note stage of development by presence of cephalic furrow. E. Some embryos from a similar population, but with only the single enhancer and raised at 30C show various defects in gastrulation (N=10 of 14). Note embryo stage by presence of cephalic furrow, yet lack of significant mesodermal invagination. The number of *snail* expressing cells anterior to the cephalic furrow is also reduced. Figure 4.9 shows a range of defects observed in these embryos; a narrower pattern of anterior expression may result in delays in involution of anterior regions, and some exhibit a more erratic midline.

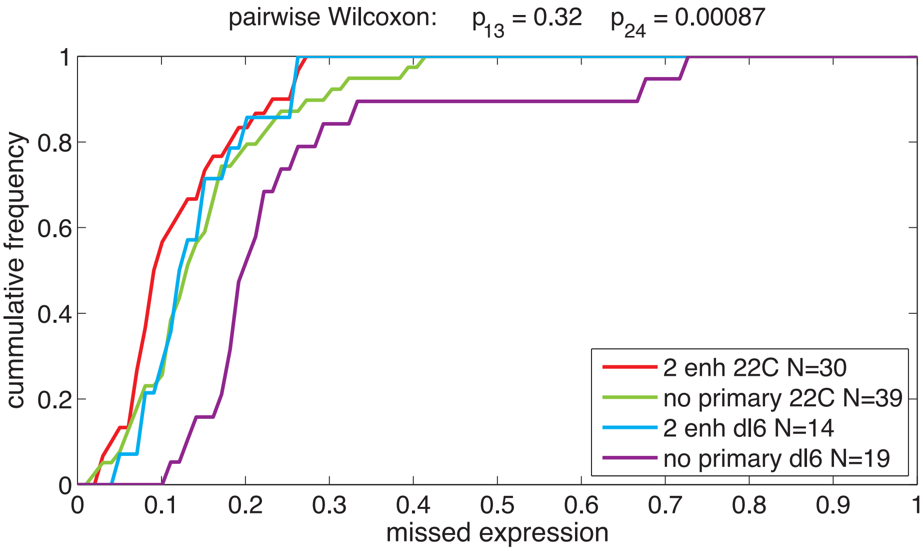

Figure 4.8: **Genetic Stress: Cumulative frequency plots for snail-yellow expression in dorsal heterozygotes vs. wildtype embryos**. p-values from pairwise wilcoxon comparison for the indicated tracks are labeled on the figures (p_{13} gives the p-value comparing the first and third tracks, as listed in the legend).

overall levels of expression, which could be an important function of shadow enhancers.

Other critical dorsal-ventral determinants also contain shadow enhancers, including *brinker*, *vnd*, and *sog* [64]. The recent analysis of *shavenbaby* suggests that shadow enhancers are essential for the reliable morphogenesis of embryonic bristles in older embryos [41]. There is also evidence that shadow enhancers might be a common feature of vertebrate systems, such as zebrafish [80].

Shadow enhancers appear to represent a novel mechanism of canalization [164], whereby complex developmental processes lead to a fixed outcome despite genetic and environmental perturbations. Other mechanisms of canalization have been suggested, including recursive wiring of gene regulatory networks and capacitors such as *hsp90* that suppress both altered folding of mutant proteins and transpositioning of mobile elements [96, 102, 138, 147].

It is conceivable that primary and shadow enhancers mediate overlap-

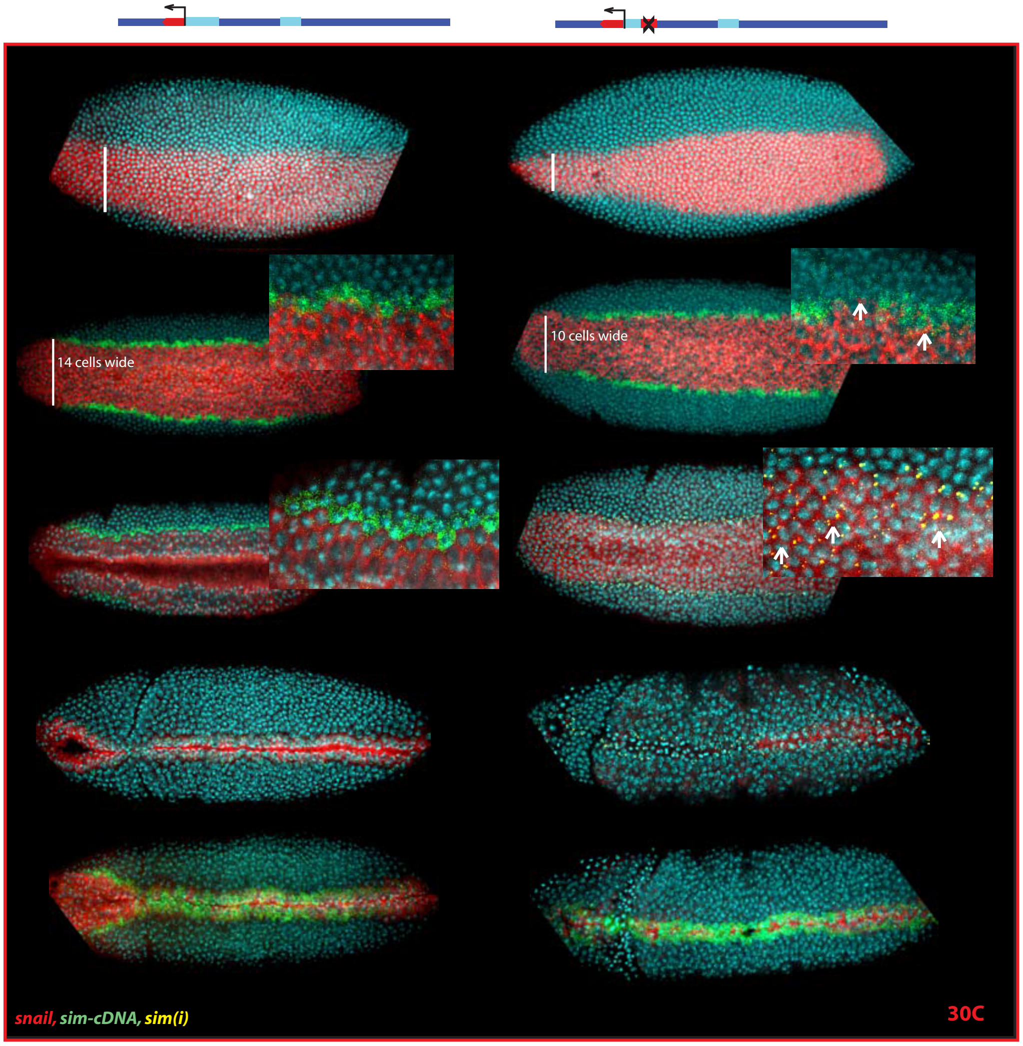

Figure 4.9: Selection embryos from early *snail* expression through gastrulation (top panels to bottom panels), raised at 30C. Control rescues show wildtype expression and invagination patterns, (left panels). Some embryos at all stages of this expression show subtle defects in expression and the coordination of gastrulation (right panels).

ping patterns of activity only during early embryogenesis. They might come to possess distinctive regulatory activities at later stages of development. Nonetheless, during the time when their activities coincide during gastrulation, they maintain reliable patterns of *snail* expression in response to environmental and genetic variability. Although either enhancer might be sufficient, both enhancers are required for accurate and reliable patterns of expression in response to variability. This precise patterning enables rapid development, without delays arising from corrective feedback mechanisms. It is easy to imagine that delays in embryogenesis would result in selective disadvantages to the resulting larvae, which must compete for limiting sources of food. Regardless of the specific mechanisms that select for shadow enhancers, the occurrence of such enhancers provides an opportunity for the evolution of novel patterns of gene expression. As long as the two enhancers maintain overlapping activities during developmental hotspots such as gastrulation, they can drift or be selected to produce divergent patterns of gene expression.

4.2.4 Experimental Procedures

Fly genetics

Positive BAC line males (labeled with w+) were crossed to yw; wg[Sp]/CyO; Pr,Dr/TM3,Sb,Ser virgins. Homozygous BAC lines were created by selfing the red-eyed, Sb,Ser flies from the F1 generation. Males from the BAC lines carrying the yellow reporter constructs were crossed to yw, dl[6]/CyOvirgins.

To test for rescue of the BAC constructs, we generated a white eyed, double balancer strain carrying a CyO linked hunchback-LacZ reporter by crossing and back crossing wnt4/CyO, hb-lacZ (BSC 6650) to yw; wg[Sp]/CyO; Pr,Dr/TM3,Sb,Ser virgins. Positive BAC males were crossed into this line to create w; +/CyO, hb-lacZ; BAC[snail,w+]/TM3,Sb,Ser virgins. Simultaneously, w; Df (2L)osp29/CyO, (BSC 3078) flies carrying a deletion spanning the *snail* gene were crossed to yw; wg[Sp]/CyO; Pr,Dr/TM3,Sb,Ser virgins. The Df (2L)osp29/wg[Sp], +/TM3,Ser males were crossed to the virgins containing the labeled balancer and the BAC. The progeny were selfed to create homozygous stable lines for the BAC carrying the *snail* deletion over the hb-lacZ marked CyO balancer. Populations still containing the Ser balancer or a wildtype chr III were analyzed also analyzed to test the effect of single copy rescue. The labeled balancer allowed for the reliable identification of embryos lacking a functional copy of endogenous sna.

Recombineering and transgenesis

Recombineering was performed as described previously [162, 161, 87, 95, 167] with modifications described in Supplemental Experimental Procedures in the published version of this manuscript. These supplemental sections also describe construction of the yellow intronic reporter, the use of plasmids with a conditional origin of replication to reduce recombineering colony background [31], and preparation of modified BAC constructs for microinjection. Supplemental Table 1 lists primers used to make BAC modifications; the *sna* primary enhancer sequence was replaced using an ampicillin resistance cassette as a non-regulatory spacer. BAC CH321-18I14 [161] was used as the basis for all other modifications.

Fluorescent in situ hybridization and quantitative imaging methods

Fluorescent in situ hybridization was performed as described in [84]. Embryos were imaged on a Leica Scanning Confocal SL microscope as a 14-20 section z-stack through the nuclear layer at 1/2 micron intervals, with scanning resolution of approximately 250 nm/pixel. Images were maximum projected and computationally segmented to localize and count nuclei, mRNA expression domains, and nascent transcripts.

4.3 Gap Gene Shadows

4.3.1 Introduction

Recent studies identified shadow enhancers for genes engaged in the dorsal-ventral patterning of the early Drosophila embryo [64]. These enhancers are sometimes located within neighboring genes, and along with conventional, proximal enhancers, they produce robust patterns of gene expression in early embryos under stress [41, 119]. For example, the *snail* gene exhibits erratic patterns of activation in embryos raised at 30C when either the proximal or shadow enhancer is removed [119]. It was proposed that shadow enhancers represent a mechanism of canalization [164], whereby populations of embryos develop normally even when subject to variations in temperature or genetic background.

In the present study we provide evidence that many of the genes controlling anterior-posterior patterning contain multiple enhancers with overlapping activities, including head patterning genes and gap genes, which initiate the segmentation gene network [24, 71]. For example, the gap genes *hunchback* (hb), *Kruppel* (Kr), and *knirps* (kni) are each regulated by two distinct enhancers that control the initial bands of gene expression within the presumptive head, thorax, and abdomen. Evidence is presented that the two enhancers work together (enhancer synergy) to ensure uniform expression within correct spatial limits.

In some cases, individual enhancers fail to recapitulate authentic features of the endogenous expression pattern. For example, the classical *hb* enhancer located at the proximal promoter mediates ectopic expression at the anterior pole of precellular embryos [154]. The newly identified distal enhancer is silent in anterior regions, and when present in a common BAC transgene, it helps impose an authentic *hb* expression pattern, including attenuated expression at the anterior pole. Similarly, one of the *kni* enhancers (intronic) produces an abnormally broad pattern of expression [142], but when combined with the distal 5 enhancer, it helps produce uniform expression and correct spatial limits.

Previous studies have documented examples of enhancer autonomy and enhancer interference. Multiple enhancers often produce additive patterns of gene expression, as seen for the 7-stripe even skipped (eve) expression pattern arising from 5 separate enhances (2 located 5 of the eve transcription unit and 3 located downstream of the gene) [145, 146, 43]. Sometimes, multiple enhancers interfere with one another when placed within a common regulatory region. For example, ventral repressors that delineate the intermediate neuroblasts defective (ind) expression pattern block the activities of a neighboring eve stripe 3 enhancer, and conversely, repressors that establish the posterior limit of the stripe 3 pattern interfere with ind [150].

In the present study, evidence is presented that combining multiple enhancers in a common regulatory region can produce sharper and more homogenous patterns of gene expression. We discuss potential mechanisms for such enhancer synergy and suggest that minimal enhancers producing aberrant patterns of gene expression might nonetheless contribute to authentic expression profiles in the context of their native loci.

4.3.2 Results and Discussion

Every Gap Gene Contains Multiple Enhancers for a Single Gap Pattern

Candidate gap enhancers were identified using ChIP-chip data [93]. Specifically, clustered binding sites for maternal and gap proteins were identified within 100 kb of every gap gene (see Experimental Procedures). This survey identified each of the known enhancers, as well as putative shadow enhancers [35, 62, 136, 10, 142, 111]. For example, a potential distal shadow enhancer was identified for hb, located 4.5 kb upstream of the proximal transcription start site (designated P2 in [104]) and upstream of the later-acting distal promoter (designated P1) (Fig. 4.10C).

A 400 bp genomic DNA fragment from this newly identified region was attached to a lacZ reporter gene and expressed in transgenic embryos (Fig. 4.10B). The resulting hb/lacZ fusion gene exhibits localized expression in anterior regions of the embryo similar to that seen for the endogenous gene and classical enhancer identified over 20 years ago [35, 153] (Fig. 4.10B; compare with A). The classical proximal and distal shadow enhancers exhibit similar responses to increasing Bicoid copy number (Fig. 4.11A).

ChIP-chip data also identified potential pairs of enhancers for *Kr* (Fig. 4.10D-F) and *kni* (Fig. 4.10G-I). There are two distinct clusters of transcription factor binding sites upstream of Kr. The previously identified *Kr* CD2 enhancer contains the proximal enhancer but also part of the distal binding cluster [62]. Subsequent lacZ fusion assays identified each ChIP-chip peak and underlying binding sites as separable proximal and distal enhancers (Fig.

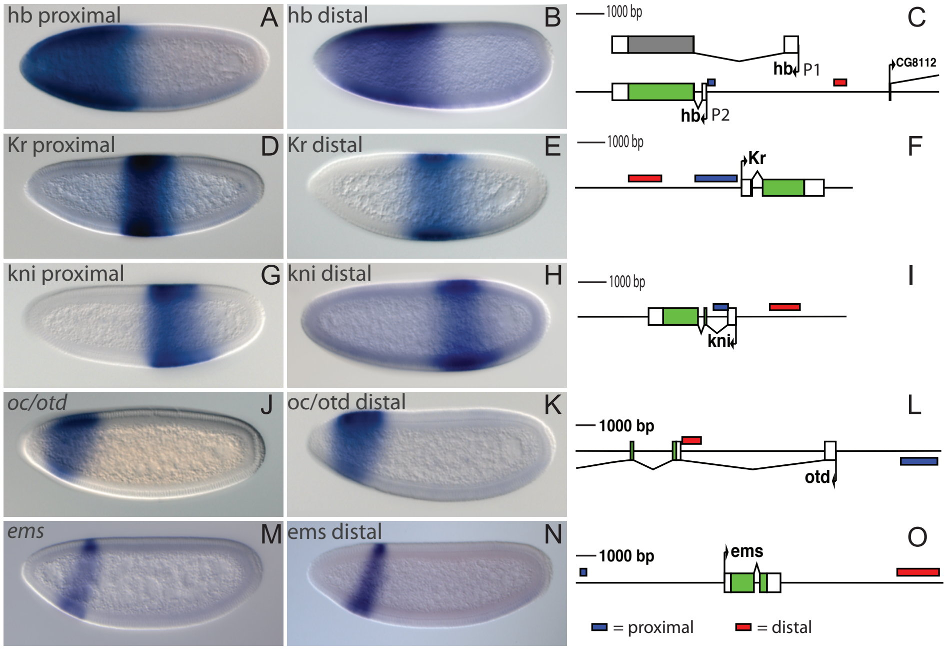

Figure 4.10: **Activities of gap enhancers identified by in situ hybridization** (A-B) hb/lacZ transgenes containing the (A) proximal (classical) or (B) newly identified distal enhancer (B). The locations of these enhancers are shown in (C). (D-E) Kr/lacZ transgenes containing the (D) proximal or (E) distal enhancer. The locations of these enhancers are shown in (F). (G-H) kni/lacZ transgenes containing either the (G) proximal intronic enhancer or (H) the distal 5 enhancer (H). The locations of these enhancers are shown in (I). (J-K) Expression of endogenous oc/otd (J); oc/lacZ transgene containing an intronic enhancer (K). The locations of the oc/otd enhancers are shown in (L). Expression of endogenous ems (M); ems/lacZ transgene containing a distal enhancer (N). Locations of the enhancers shown in (O).

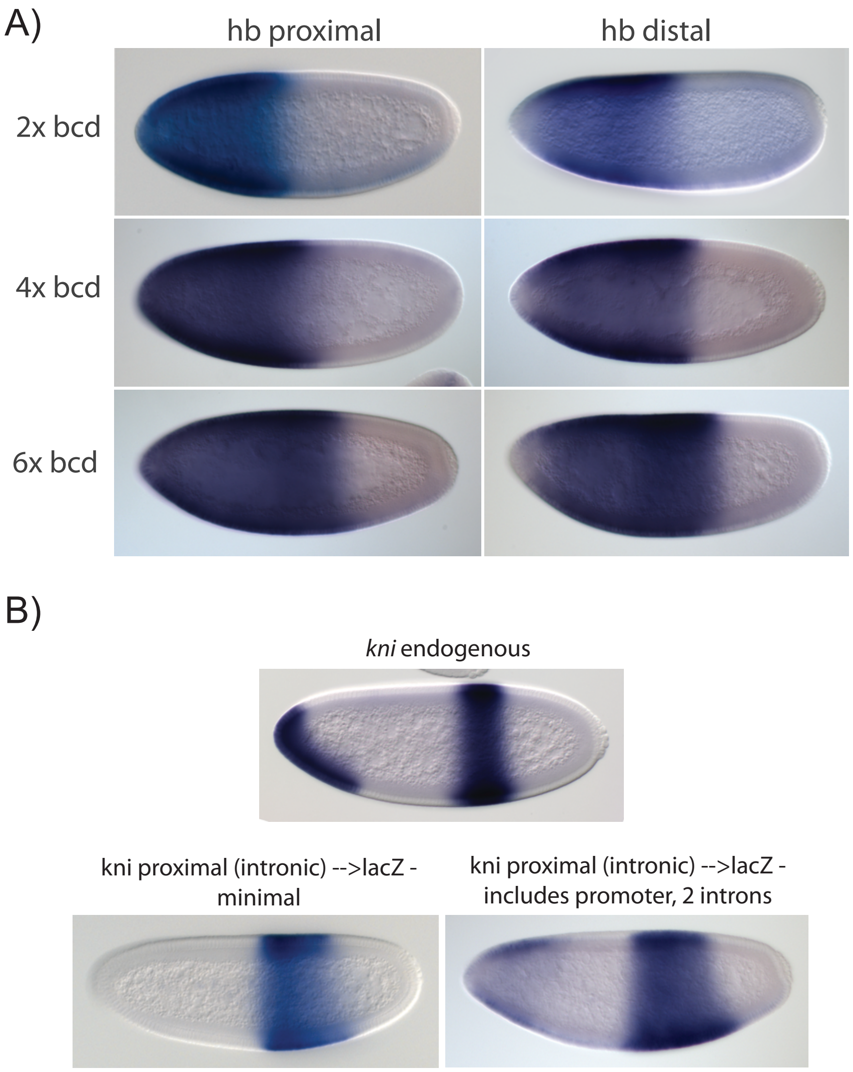
A)
hb proximal
hb distal
2x bcd
4x bcd
6x bcd
B)
kni endogenous
kni proximal (intronic) -->lacZ - minimal
kni proximal (intronic) -->lacZ - includes promoter, 2 introns

Figure 4.11: **Enhancer validation: (A)** both the proximal and distal *hunchback* enhancer respond to changes in the bicoid concentration by extending posteriorly, as has been reported for the endogenous gene. **(B)** The *knirps* intronic enhancer is always too broad as long as the distal enhancer is missing, independent of promoter choice, as seen by the comparison of the endogenous pattern to the minimal intronic or the promoter and both introns.

4.10D-F). Similarly, more refined limits were determined for the *kni* intronic enhancer (Fig. 4.10G,I; Fig. 4.10B; [142]), in addition to the previously identified 5 distal enhancer ([112]). Both the distal *Kr* enhancer and the intronic *kni* enhancer produce somewhat broader patterns of expression than the endogenous gene (Fig 4.10E,G; and 4.11B). Additional gap enhancers were also identified for giant, including an additional distal enhancer located 35 kb downstream within a neighboring gene (Fig. 4.12A and [142]).

The survey of gap and maternal binding clusters was extended to include the so- called head and terminal gap genes, critical for the differentiation of head structures and the non-segmented termini of early embryos (Fig. 4.10J-O; Fig. 4.12B-J). Additional enhancers were identified for empty-spiracles (ems) (original enhancer identified in Hartmann et al., 2001) (Fig. 4.10M-O), huckebein (hkb) (original enhancer in Hader et al 2000 [55] (Fig. 4.12E-G), and forkhead (fkh) (original enhancer in Schroeder et al. 2004 [142]) (Fig. 4.12B-D). More refined limits were also determined for the previously identified ocelliless/orthodenticle (oc/otd) intronic enhancer [142] (Fig 4.10J-L). For simplicity, we will hereafter refer to the two enhancers regulating a given gap gene as proximal and distal, based on their relative locations to the transcription start site.

Multiple *hb* Enhancers Produce Authentic Expression

BAC recombineering [162, 161], phiC31 targeted genome integration [52, 16], and quantitative in situ hybridization assays (Boettiger and Levine, 2009; Perry et al., 2010) were used to determine the contributions of the proximal and distal enhancers to the *hb* expression pattern (Fig. 4.14). BACs containing 20 kb of genomic DNA encompassing the *hb* gene and flanking sequences were integrated into the same position in the Drosophila genome. The *hb* transcription unit was replaced with the yellow gene, which permits quantitative detection of nascent transcripts using an intronic hybridization probe (see Perry et al., 2010[119]; Experimental Procedures and Fig. 4.13). The modified BAC retains the complete *hb* 5' and 3' UTRs. Additional BACs were created by inactivating the proximal or distal enhancers by substituting critical regulatory elements with random DNA sequences (see diagrams above panels in Fig. 4.14A-C and Experimental Procedures).

BAC transgenes lacking either the distal (Fig. 4.14A) or proximal (Fig. 4.14B) enhancer continue to produce localized patterns of transcription in anterior regions of transgenic embryos in response to the Bicoid gradient.

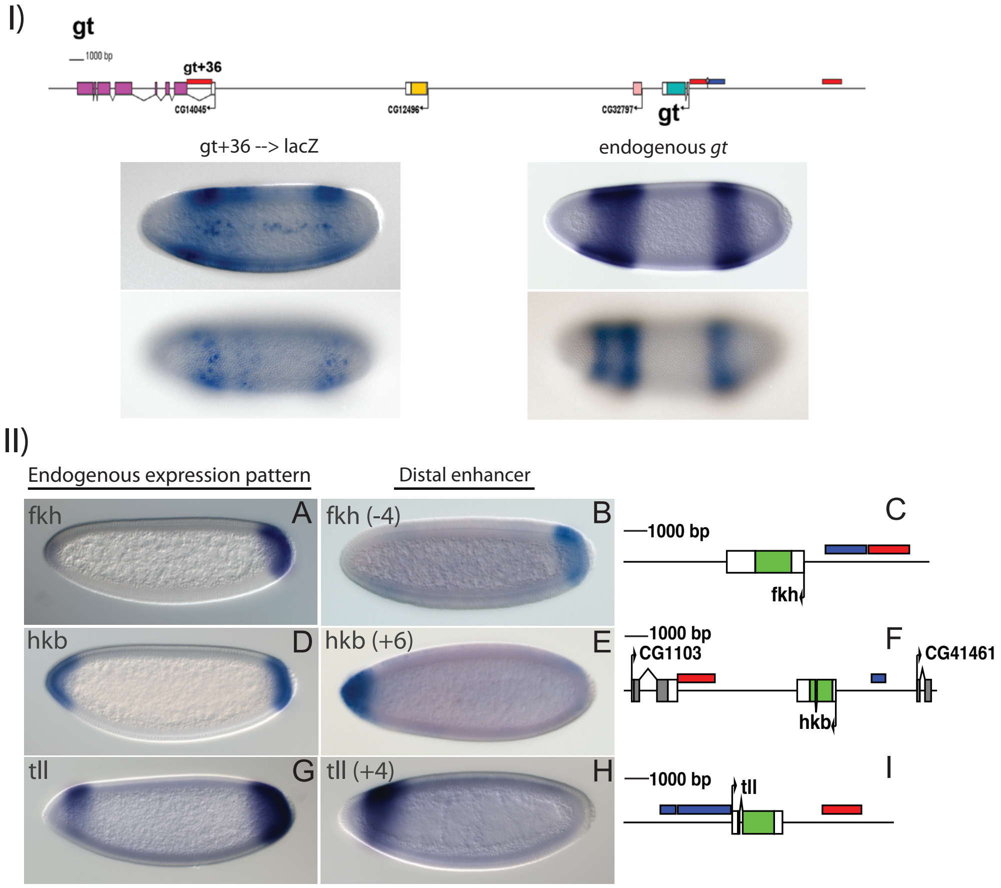

Figure 4.12: **Additional gap shadow enhancers (I)** 35 kb downstream of *gt* is another gap enhancer for both the anterior and posterior pattern of *gt*. **(II)** The head gap genes *fkh* A-C, *hkb* D-F, and *tll* G-I also have shadow enhancers. The originally discovered proximal enhancers are marked in blue (C,F,I), but images are not reproduced here. The position of the more recently discovered shadow enhancers are denoted in red.

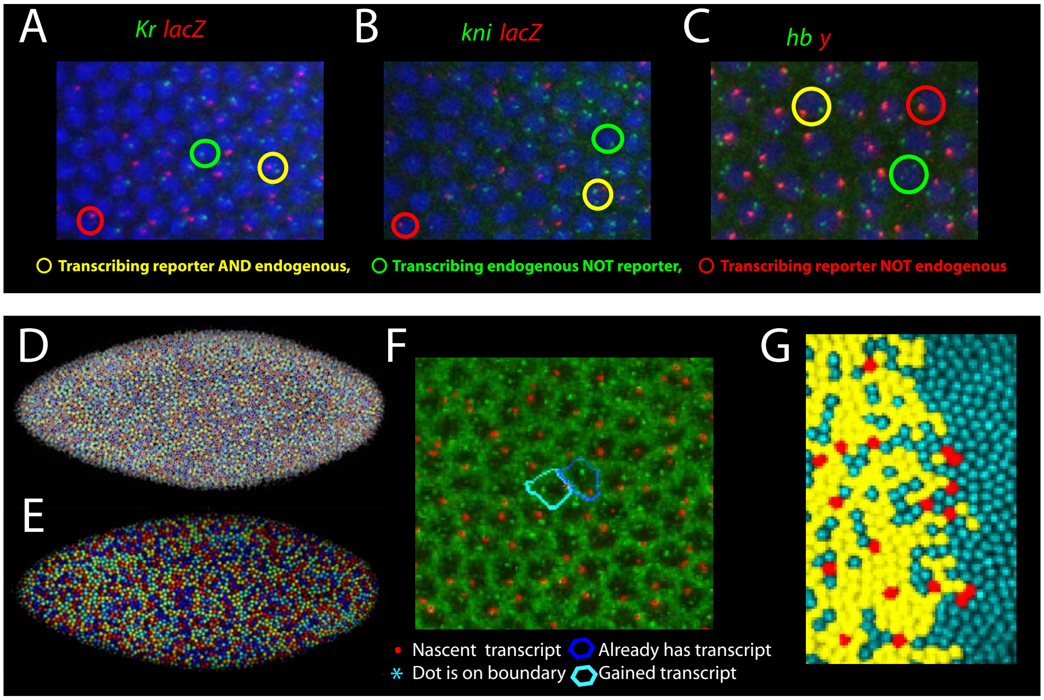
A
Kr lacZ
B
kni lacZ
C
hb y
○ Transcribing reporter AND endogenous,
○ Transcribing endogenous NOT reporter,
○ Transcribing reporter NOT endogenous
D
E
F
G
● Nascent transcript
Already has transcript
* Dot is on boundary
Gained transcript

Figure 4.13: **Image Analysis Method** A Nascent transcription foci detected for endogenous *Kr* (green channel) and for lacZtransgene (red channel) detected by a fluorescent in situ using Kr-full length mRNA probe and LacZ full length mRNA probe. Nuclei are counterstained in blue. Double staining allows detection of nuclei which transcribe both loci (yellow circles), just the endogenous gene (green circles), or just the reporter gene (red circles). B as in (A), with endogenous kni-full length probe. C as in (B), with intronic yell0w-reporter in red channel and endogenous full length hb-probe in green. Since the *hb* embryos are imaged in cell cycle 13 instead of 14, the nuclei are also less dense. D Computational image processing algorithm segments all the different nuclei (representing unique nuclei by separate, randomly chosen colors). E These masks in (D) are then expanded in a space filling way so every pixel of the embryo is assigned to a nucleus. F The individual transcripts are segmented by a filter (red dots), and assigned to the containing nucleus (blue polygons). An extra filtering step resolves dots which lie on the border between two cells by comparing the presence of dots in neighboring cells. G The results are plotted by assigning the cell mask computed in (E) to a color code which represents its composite transcriptional state (i.e. transcribing reporter AND endogenous vs transcribing reporter NOT endogenous).

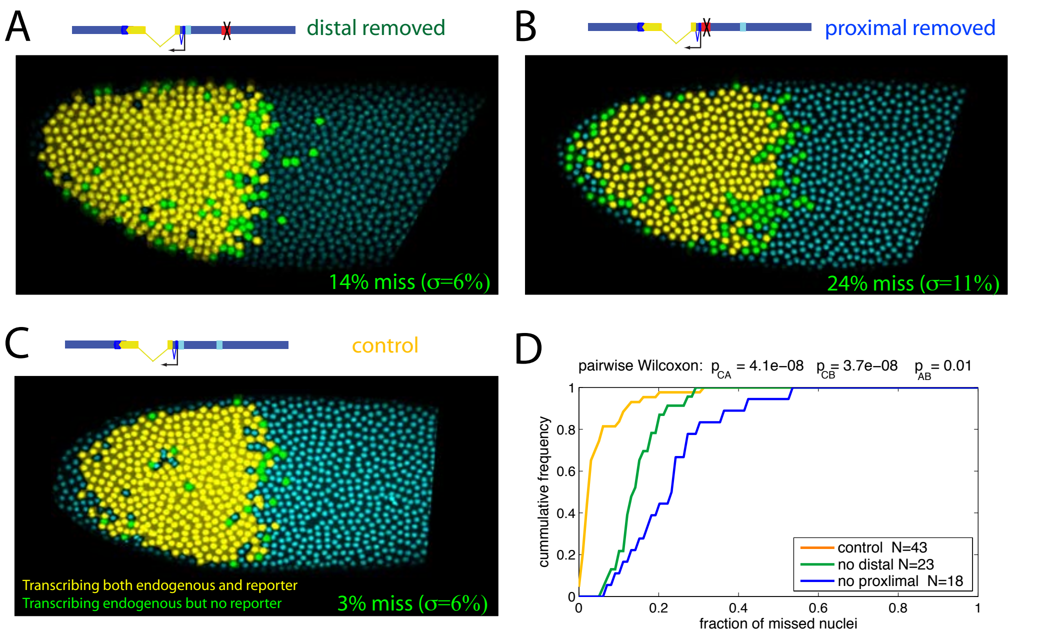

Figure 4.14: **Function of *hb* enhancers via BAC transgenesis** An 20 kb BAC containing genomic DNA encompassing the *hb* locus was modified to remove specific proximal or distal enhancers. The *hb* transcription unit was replaced with the yellow reporter gene in order to identify de novo transcripts via in situ hybridization using a probe directed against the yellow intron (see diagram above images in panels A-C). (A) hb-BAC with distal enhancer inactivated (see X in diagram), (B) hb-BAC with proximal enhancer inactivated, and (C) hb-BAC with both enhancers intact. The median ratio of discordance is indicated beneath each image. This is the fraction of nuclei that express endogenous *hb* nascent transcripts, but not yellow transcripts. (D) Cumulative frequency distributions for the fraction of missing nuclei in the three populations of embryos. The ordinate axis gives the probability of observing an embryo from this population with fewer than the abscissa fraction of nuclei transcribing the endogenous gene but not the reporter. Statistical comparisons between the distributions are presented above the panel, with subscripts matching the panel labels (i.e. pCA is the p value from the pair wise comparison of the distribution of embryos with the *hb* control BAC, (C), to those with the distal enhancer removed (A)).

However, the patterns are not as faithful as compared with the BAC transgene containing both enhancers (Fig. 4.14C). Embryos were double-labeled to detect both yellow and *hb* nascent transcripts. During nuclear cleavage cycle (cc) 13, a substantial fraction of nuclei (14%) expressing *hb* nascent transcripts lack yellow transcription upon removal of the shadow enhancer (Fig. 4.14A). An even higher fraction of nuclei (24%) lack yellow transcription when the proximal enhancer is removed (Fig. 4.14B). Control transgenic embryos containing both enhancers exhibit more uniform patterns of transcription, whereby only an average of 3% of nuclei fail to match the endogenous pattern of transcription (Fig. 4.14C). The distribution of missing nuclei across the population of cc13 embryos is plotted in Fig 4.14D.

The pairwise Wilcoxon rank sum test (also called the Mann-Whitney U-test) was used to determine the significance of the apparent variation in gene expression resulting from the removal of either the proximal or distal enhancer (Fig. 4.14D). Control embryos containing the *hb* BAC transgene with both enhancers exhibit some variation in the number of nuclei that lack yellow nascent transcripts. Despite this variation, the statistical analyses indicate that the loss of either the proximal or distal enhancer results in a significant increase in the variability of yellow transcription patterns as compared with the control BAC transgene (p=4E-8).

The Distal *hb* Enhancer Mediates Dominant Repression

The preceding analyses suggest that multiple enhancers produce more uniform patterns of de novo transcription than individual proximal or distal enhancers. Additional studies were done to determine whether multiple enhancers also help produce authentic spatial limits of transcription (Fig. 4.15).

hb expression normally diminishes at the anterior pole of cc13-14 embryos. This loss in expression has been attributed to attenuation of Bcd activity by Torso RTK signaling (e.g., [81]. However, the proximal enhancer fails to recapitulate this loss (Fig. 4.10A). In contrast, the distal enhancer is inactive at the anterior pole (Fig. 4.10B), and the two enhancers together produce a pattern that is similar to endogenous expression, including reduced expression at the pole (Fig. 4.14C).

To examine the relative contributions of the proximal and distal enhancers in this repression, yellow nascent transcripts were measured in transgenic embryos expressing BAC reporter genes containing one or both *hb* enhancers (see Fig. 4.14). Particular efforts focused on the early phases of cc14, when

Figure 4

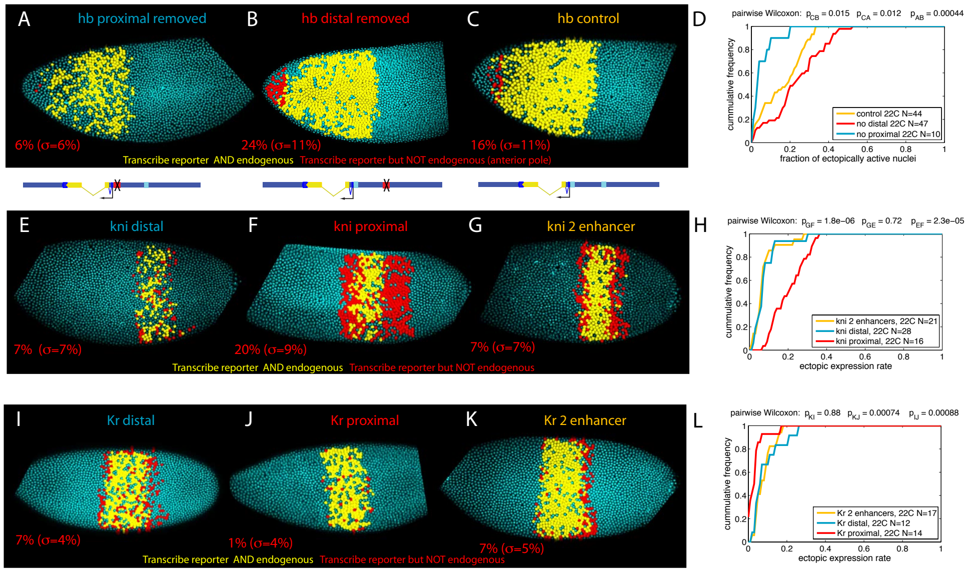

Figure 4.15: **Enhancer synergy produces correct spatial limits.** Discordance of yellow (A-C) or lacZ (E-G, I-K) transgenes and endogenous gap gene nascent transcripts. Nuclei exhibiting ectopic transgene expression are indicated in red. Sites of concordant expression are indicated in yellow. (A-C) BAC transgenes lacking the proximal (A) or distal (B) enhancer, or containing both enhancers intact (C). Nuclei in the anterior third of the hb-expression region which transcribe the reporter but not endogenous *hb* are shown in red. (D) Cumulative frequency of nuclei in the anterior third of the hb-expression domain containing yellow, but not endogenous hb, nascent transcripts. Median and standard deviations are shown on the corresponding panels. (E-F) kni/lacZ reporter genes driven by (E) distal enhancer, (F) proximal enhancer or (G) both enhancers. Median fractions of nuclei transcribing lacZ but not the endogenous gene are indicated below each image. (H) Cumulative frequency distributions for the fraction of ectopically active nuclei. (I-L) Similar analysis of Kr/lacZ transgenes containing the distal (I), proximal (J), or both (K) enhancers.

repression of endogenous *hb* transcripts is clearly evident (Fig. 4.15). For the transgene lacking the proximal, classical enhancer, but containing the newly identified distal enhancer, a median of 6% (std 6%) of nuclei contains yellow nascent transcripts but lack expression of the endogenous gene (Fig 4.15A). In contrast, a median of 24% (std 11%) of nuclei exhibits a similar discordance upon removal of the distal enhancer. In control embryos, 16% (std 11%) of nuclei exhibit yellow, but lack hb, nascent transcripts (Fig 4.15B-C). It should be noted that the BAC transgene lacking the proximal enhancer exhibits super-repression due to reduced activation at the anterior pole ($p = 0.012$) (Fig 4.15D).

These observations suggest that the distal enhancer contains repression elements that function in a dominant manner to attenuate the activities of the proximal enhancer at the anterior pole. There is a loss of repression when the distal enhancer driving lacZ is crossed into a torso mutant background (Fig 4.16). This observation implicates one or more repressors functioning downstream of Torso signaling, including Tailless and Huckebein, which have been shown to function as long-range dominant repressors [29, 47, 48]. The persistence of *hb* expression in anterior regions has been shown to be detrimental, causing defects in mouth parts and malformation of the gut (Janody et al., 2000).

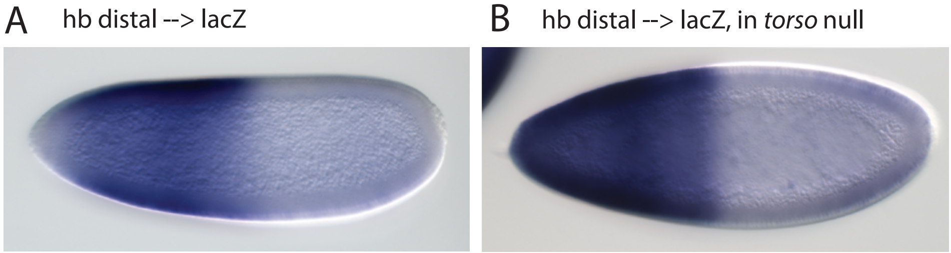

Figure 4.16: **Torso signaling is required for anterior *hb* repression.** **A** In a wildtype background the *hb* distal enhancer, like the endogenous gene, is repressed in the anterior tip. **B** In torso null mutant, reporter expression from the distal enhancer extends all the way to the anterior tip (as does the endogenous gene expression, data not shown.

Combining Two Enhancers Corrects the *kni* Expression Boundaries

Kr/lacZ and kni/lacZ fusion genes containing either one or two enhancers were inserted into the same position in the Drosophila genome (Fig. 4.15). Transgenic embryos were double-labeled to detect the expression of the transgene (lacZ) as well as the endogenous gap gene.

The *kni* proximal (intronic) enhancer alone produces an abnormally broad pattern of expression (Fig. 4.15F; see Fig. 4.10G and [142]). In contrast, the *kni* distal (5) enhancer produces erratic lacZ activation within nearly normal spatial limits (Fig. 4.15E). An essentially normal pattern of lacZ transcription is observed with both enhancers (intronic enhancer 5 and distal enhancer 3 of lacZ; Fig. 4.15G, Fig. 4.17G). It appears that lacZ transcription is slightly broader than the endogenous pattern, but considerably narrower than the pattern observed for the proximal enhancer alone (Fig. 4.15J) ($p = 1.8E\text{-}6$), and not statistically different from the expression limits of the distal enhancer alone ($p = 0.72$) (Fig. 4.15L). There is no significant narrowing of the Kr/lacZ expression pattern when both the distal and proximal enhancers are combined within the same transgene (Fig. 4.15K,L) ($p = 1.0$). Perhaps additional *Kr* regulatory elements are required for the type of narrowing observed for the *kni* intronic enhancer.

As discussed earlier, long-range repressors bound to the distal *hb* enhancer might inhibit the activities of the proximal enhancer at the anterior pole of precellular embryos. The distal *kni* enhancer might function in a similar manner to sharpen the expression limits of the intronic enhancer. Gap gene expression limits depend on cross-repressive interactions (e.g. [85, 101, 102]). The *kni* intronic enhancer might lack critical gap repression elements since it produces an abnormally broad expression pattern.

The mechanism of dominant repression is unknown. Perhaps the distal *kni* and *hb* enhancers modify the target promoter in regions where the respective genes are inactive (e.g., posterior or anterior pole, respectively) (Fig 4.17C compare to D). Indeed, recent studies suggest that long-range repressors might lead to the positioning of nucleosomes at the core promoter, thereby limiting Pol II activity [92].

We propose that multiple enhancers for a common gap gene expression pattern function synergistically to produce accurate and robust patterns of transcription (summarized in Fig. 4.17). The exact mechanism is uncertain. For example, the looping of one enhancer might render the promoter regions of target genes more accessible to interactions with the second enhancer. A

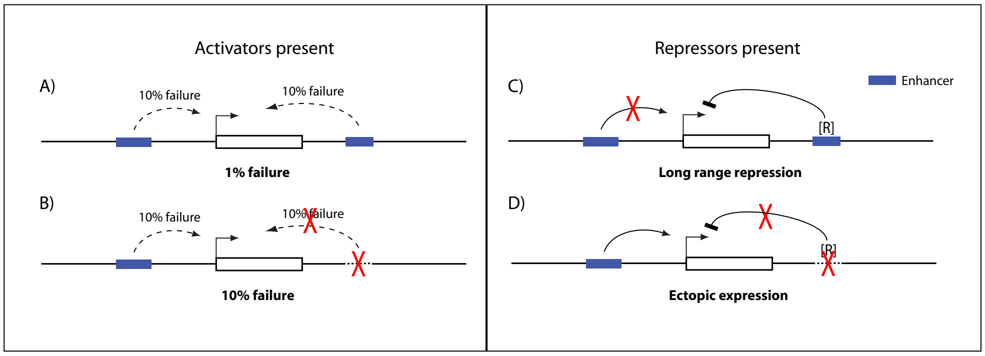

Figure 4.17: **Models for enhancer synergy.** (A-B) Activation of one promoter by two enhancers. If two independent enhancers each have a 10% failure rate in activating expression and TF binding or enhancer looping is rate limiting, then the two enhancers should have a combined failure rate of 10% x 10% = 1% (A). Removing one enhancer increases the failure rate to 10% (B). (C-D) The binding of a long range, dominant repressor to one enhancer is sufficient to inactivate the other. (C). Removal of this enhancer results in ectopic expression (D).

nonexclusive alternative is that two enhancers increase the probability that at least one engages the promoter within a given time window (Fig 4.17A compare to B). Regardless of mechanism, the combined action of multiple enhancers explains why an individual enhancer sometimes fails to recapitulate the endogenous expression pattern when taken from its native context. Such enhancers nonetheless help ensure homogenous and robust patterns of gene expression.

4.3.3 Methods

Enhancer Identification and Testing

Prospective enhancers were identified near genes of interest using a combination of ChIP-chip data (provided for various maternal, gap, and pair rule genes by the Berkeley Drosophila Transcription Network Project, [93] and sequence-based binding site cluster analysis. The cluster analysis was performed using the software ClusterDraw2 [113]. The program and binding motif models used are available online at http://flydev.berkeley.edu/cgi-bin/cld/submit.cgi.

Candidate regions were tested in vivo using traditional lacZ reporter assays combined with targeted phiC31 transgenesis as adapted for use in

Drosophila [52, 16]. An nE2G backbone with insulators [105] modified for targeted integration was used to test potential enhancers by placing them upstream of an eve-lacZ fusion gene. The same construct was used for the one vs. two enhancer experiments for *Kr* and kni; the second enhancer for the two enhancer constructs was added into a BstBI restriction site downstream of lacZ 5kb away from the first enhancer. The landing site 51D [16], Bloomington Stock Center number 24483, was used for lacZ assays.

The two *hb* enhancer-lacZ constructs were crossed into a 4x or 6x maternal Bicoid copy number background using the BB9+16 fly line [153].

Recombineering and Transgenesis

Recombineering was performed as described previously in Perry et al., 2010 (see also [162, 161, 87, 95]. The yellow reporter (used to detect sites of nascent transcript by using an intronic in situ probe) was integrated as a yellow-kanamycin fusion that left the native *hb* UTRs intact. The bcd binding site clusters and surrounding regions of the primary or shadow enhancers were removed via replacement with an ampicillin resistance cassette taken from pBlueScript. Primers used for construct building and recombineering will be listed in the Supplemental Data of the published manuscript. BAC CH322-55J23 [161] was the basis for all subsequent modifications. All BACs were integrated into landing site VK37 on chromosome 2 [162], Bloomington Stock Center number 24872.

Whole-Mount in situ Hybridization

Embryos were fixed using standard methods. Fluorescent or colormetric in situ hybridization was performed as described in [84, 119]. Probes were generated with the primers listed in Supplemental Data and in vitro transcription. Reporter genes were labeled with digoxigenin-tagged antisense probes, sheep anti-dig primary antibodies (Roche), and donkey anti-sheep Alexa 555 secondary antibodies (Invitrogen). Endogenous genes hb, *Kr* and *kni* were labeled with biotin-tagged probes, mouse anti-bio primary antibodies (Roche), and donkey anti-mouse Alexa 488 secondaries (Invitrogen). Nuclei were counterstained with DRAQ5 (Biostatus Ltd.).

Confocal Image Acquisition and Computational Image Processing

1024x1024 3-color image stacks were acquired using a Leica SL Laser Scanning Confocal microscope as described in Perry et al., 2010. Image segmentation and analysis was performed as described in Perry et al. 2010, with minor modification. Nuclei were segmented using a Difference of Gaussians filter optimized with size selection (Fig S3D). A space-filling, segment dilation algorithm was used to assign all pixels in the embryo to one of the segmented nuclei, created a final nuclear mask (Fig S3E). All nuclear masks were manually checked to confirm accurate segmentation. Nascent transcripts were localized for both the reporter and the endogenous genes, also using Difference of Gaussians filters, this time optimized to detect the bright transcripts corresponding to sites of transcription, see Fig. S3. Intensity thresholds and dot size thresholds reduced spurious counts. Segmentation results were curated by the user. This segmentation enabled the nucleus by nucleus analysis of transcriptional state of reporter and endogenous gene as described in the text and shown in figures 2-3. This analysis scripts were wrapped in a Graphical User Interface implemented through Matlab's software GUI Design Environment (GUIDE). The source code for this analysis is available in the supplemental material of Perry et al. 2010 and on http://alistairboettiger.info/home/Software.html. Updated versions of our image segmentation routines can be found on our Github page for image processing tools: https://github.com/AlistairBoettiger/Image-Analysis. All of the source code used to compute and plot the results from this publication are available on the Github source page for this project on shadow enhancers: https://github.com/AlistairBoettiger/Shadow-Enhancers.

4.4 Beyond Site Occupancy: Effects of Enhancer Multiplicity on Transcription

4.4.1 Introduction

Recent investigations have revealed that many developmental patterning genes are controlled by independent regulatory sequences (enhancers) which drive expression in the same subset of cells during overlapping periods in development. It has been shown in a few cases that this apparently redundant activity is necessary to ensure reliable gene expression and patterning under mild stress conditions [41, 119]. Further investigations have proposed an alternative explanation – that extra enhancers provide a refining role, whereby unintended induction by one enhancer is kept in check by the repressive action of its counterpart, as has been observed for *knirps* and *hunchback* [121].

In this work, we ask how commonly does the overlapping activity of two enhancers effect the completeness of the expression pattern in the target tissue. Previous work has tested the effect of multiple enhancers acting on the same promoter by removing one of the elements from its native locus and replacing it with non-regulatory sequence [119, 121]. It is not known however if the synergistic effect of multiple enhancers is dependent upon the chromatin context, whether the choice of promoter is important for the effect, and whether the change observed upon replacing one enhancer with alternative sequence does not result in some sort of abnormal regulatory interference (as opposed to simple loss of function). We investigate each of these questions by using a synthetic arrangement of the positions of the enhancers and heterologous promoters, for 4 genes previously shown to have so called "shadow" enhancers, *hb*, *Kr*, *kni* and *sog*. We show that for each of these genes, the overlapping activity of two enhancers even in a synthetic arrangement of positions results in a more complete, homogeneous activation of transcription, more closely resembling the patterns of expression seen with the endogenous gene than exhibited by reporters driven by only one of the relevant enhancers.

The effect of enhancer number on homogeneity of transcriptional activation for several very different classes of early developmental patterns has several implications for our understanding of the control of gene expression in the early embryo. The consistency of this effect among a broad range of expression patterns suggests the effect is not specific to a specialized mode of activation – since the enhancers examined span a considerable range of regulatory architectures. Moreover, the increase in transcriptional activation upon a simple increase in enhancer number suggests that in that enhancer promoter interactions and/or waiting times from transcription factor site recognition are rate limiting regulatory events. This observation challenges the traditional view that transcriptional state is directly determined by site-occupancy of the enhancer, as has been proposed by numerous groups, [1, 163, 19, 14, 15, 174, 173]. Were such the case, there should be no effect of adding an additional enhancer at a position where it does not interfere with the original enhancer, except for a possible case where the second cluster of binding sites might *reduce* the transcription rate by spreading the transcription factors too thin between the two enhancers (a case expected for very cooperative enhancer models to occur near the switch point). That we observe the opposite effect argues for a critical role in the rate of TF binding and/or long range chromatin interactions in determining whether a gene becomes transcriptionally active.

4.4.2 Results

Native chromatin context is not required for enhancer synergy

We first asked if the two identified enhancers for the anterior expression pattern of the gap gene *hunchback* were alone sufficient to create as uniform a pattern of transcriptional activation as created by the endogenous gene. Previous work showed that replacing either of these enhancers with a bacterial antibiotic resistance gene slightly alters the frequency of detecting nascent transcription events among the induced nuclei [121]. We tested if these two elements were alone sufficient to give that expression pattern by placing the proximal element upstream and the distal element downstream of a LacZ reporter gene. The expression from this construct was compared to expression from transgenic reporters containing only the proximal or distal element. All elements were inserted into the same landing site on chromosome II (J28). A heterologous eve promoter was used in place of the endogenous promoter,

Comparative analysis of nascent transcripts of the reporter genes to the endogenous *hb* transcription among embryos in cell cycle (cc) 13 produced quantitatively similar results for adding back enhancers as was previously observed for removing them [121]. Embryos with reporter expression driven by the proximal enhancer alone showed a median of 10% of nuclei expressing the endogenous gene but not the reporter, (standard deviation 8%), see figure 4.18A. Though driven by a minimal regulatory sequence of some 500 bp (?) this expression is not statistically distinguishable from embryos in which only the distal enhancer (1kb) of the 20 kb locus was removed. In embryos with only the distal enhancer, some 25% of nuclei in cc 13 show expression of the endogenous gene and not the reporter (standard deviation 8%), (see figure 4.18B), again statistically indistinguishable from embryos where only 1 kb around the primary enhancer is removed [121]. In embryos where the transgene is flanked by the proximal and distal enhancer, a median of 5% of nuclei fail to match the endogenous pattern, only slightly less than observed for reporter expression driven by an otherwise intact 20 kb locus (see figure 4.18C).

Uniform patterns of transcriptional activation is a general property of shadow enhancers

We used a similar approach to analyze the function of overlapping enhancer activity reported for the genes *knirps* (*kni*) and *Kruppel* (*Kr*). Two independent elements near the *kni* locus drive early *kni* like patterns, one which is located in the intron of the gene itself and the second which is more distal five-prime of the promoter. The intronic enhancer alone drives a pattern of transcriptional activity broader than that observed for the endogenous gene – which is corrected back towards a more native-like pattern when both enhancers are available to act on the promoter [121]. We asked if in addition to a role in correcting the boundaries, if the additional enhancer affected the fraction of actively transcribing cells in the induced region, as had been observed for *snail* and *hb*. We also asked if a similar function could be observed for the overlapping enhancers at the *Kr* locus.

Transgenes containing a single enhancer exhibit erratic, incomplete patterns of transcriptional activation (Fig. 4.19A,B;E,F), whereby 20-34% of the nuclei exhibiting endogenous expression of *Kr* or *kni* lack lacZ nascent transcripts. More uniform patterns of lacZ transcription are observed for transgenes containing both *Kr* or *kni* enhancers, with 15% of nuclei exhibiting discordance of the lacZ and endogenous patterns. In these experiments, solo enhancers were placed immediately upstream of lacZ (Fig. 4.19A,B;E,F). The distal enhancer was placed downstream of lacZ in transgenes containing both enhancers (Fig. 4.19C,G), and the two enhancers are separated by 5 kb. Pairwise Wilcoxon rank sum tests suggest that the fidelity of transgene

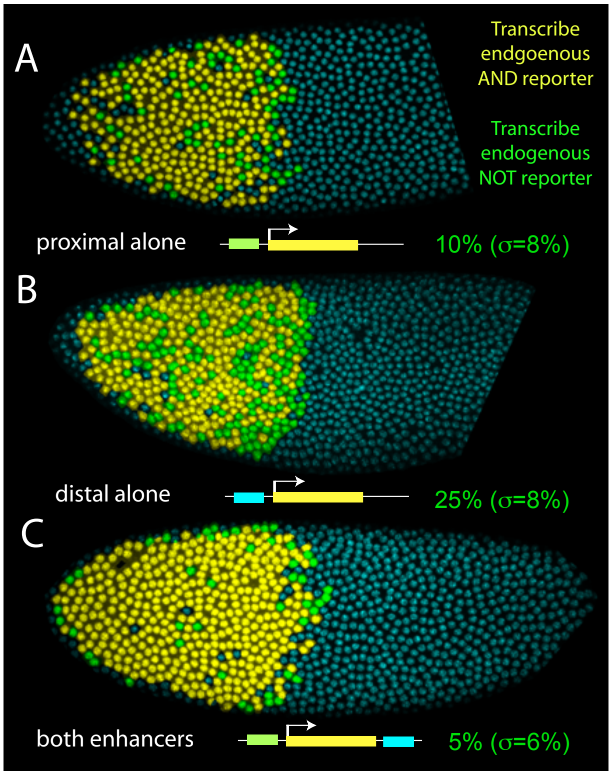

Figure 4.18: **Native chromatin context is not required for enhancer synergy**. A comparison of reporter expression and endogenous gene expression for embryos containing only the proximal enhancer (A), distal enhancer (B), or both enhancers, with the proximal upstream and the distal downstream of the LacZ reporter gene (C). Yellow cells are actively transcribing both the endogenous gene and the reporter, green cells are transcribing the endogenous gene but failing to transcribe the reporter. Inactive nuclei are uncolored and stained in cyan. Median percentages of cells in the expression region failing to transcribe the reporter are given below each embryo, with standard deviation in parentheses.

expression is significantly improved with two enhancers as compared with either the proximal or distal enhancer alone (Fig 4.19D,H).

In addition to cell-by-cell comparisons of the pattern of active transcription, it is instructive to measure the fraction of the expression region which is transcribing the gene at any given time. Taken independently, each enhancer activates transcription in substantially fewer cells of the expression domain, (60% and 40% for the proximal and distal respectively), but together the endogenous median frequency of 70% is achieved (Fig 4.19E-H). Cell-by-cell comparisons of transcription state (as done with hb), provide a more precise measure missed activation. Embryos with two enhancers show no more discordance in expression than expected from the independent promoters (e.g. 2 separate promoters active in 80% of cells should show 16% mean discordance), whereas embryos with only one of the enhancer pair show significantly more discordance: p¡0.05, (See figure 4.19D,H).

Embryos containing a transgene driven by the *Kr* proximal enhancer exhibit simultaneous transcription in only a median of 50% of cell within the *Kr* domain, whereas the endogenous gene exhibits transcription in 80% of cells in the *Kr* domain. The distal enhancer provides a slightly more complete pattern of expression (70% of the domain). When added downstream of a reporter containing upstream the proximal enhancer the median transcribing fraction increases to 80% (Fig S4A-D). All three reporter constructs use the same heterologous promoter and targeted landing site and are thus readily comparable, though care should be taken in comparing the reporters directly to the endogenous, since promoter, enhancer position and genomic context are different.

Shadow enhancers have been reported for genes responding to a very broad collection of activation signals which drive expression in most regions of the early embryo (e.g. anterior half, central band, posterior band, mesoderm, neuorgenic ectoderm). We have shown for four of these five that one consequence of this overlapping activity is to generate more uniform patterns of transcription. So we next analyzed the neurogenic ectoderm gene *sog* and its shadow enhancers [64] to see if this synergy depended strongly on the nature of the signals to which the enhancers were responding (and hence the subset of cells in which expression is induced).

Using a similar arrangement with the native proximal enhancer (normally intronic) located upstream and the distal shadow (normally 20 kb upstream on the opposite side of the neighboring gene, see figure 4.20), we tested the effect of dual regulation. In embryos raised in unstressed laboratory

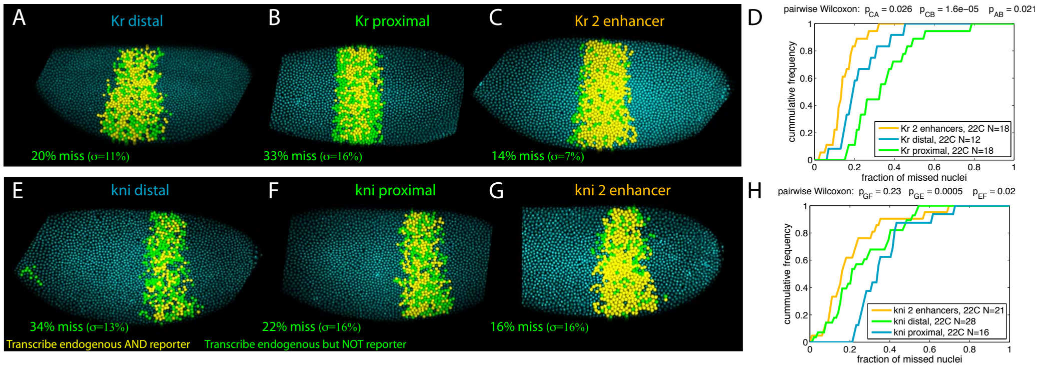

Figure 4.19: **Fidelity of reporter activation, Kr,kni**. **A-C** Comparison of transcriptional state of the endogenous *Kr* locus to a LacZ reporter gene driven by either the *Kr* distal enhancer (A), *Kr* proximal enhancer (B), or both enhancers, proximal upstream, distal downstream (C). Yellow cells are transcribing both the endogenous and reporter gene. Green cells are transcribing the endogenous gene but not the reporter. **D** Cumulative frequency distributions for the three populations of embryos with statistical comparison. **E-G** Comparison of transcriptional state of the endogenous *kni* locus to a LacZ reporter gene driven by either the *kni* distal enhancer (E), *kni* proximal enhancer (F), or both enhancer, proximal upstream, distal downstream (G). **H** Cumulative frequency distributions for the three populations of embryos with statistical comparison.

condition there is very little difference in transcriptional activation profile between the embryos (see figure 4.21A-C,G,H). In each of the transgenic backgrounds, around 20-30% of cells transcribing the endogenous gene fail to simultaneously transcribe the reporter, with slightly higher rates observed when lacking the proximal (31%) or distal (26%) enhancer than when both are present. When challenged with a mild heat-stress condition at 30C, there is a a further reduction in the concordance of reporter and endogenous gene expression in both backgrounds with only a single enhancer. In embryos with only the proximal some 36% (std 9%) of nuclei transcribe the endogenous gene but not the reporter, a weakly significant change (Fig 4.21H). In embryos with only the distal enhancer a median of 50% of nuclei (std 15%) transcribe the endogenous but not the reporter. The substantial difference in thermal sensitivity between these otherwise remarkably similar enhancers is a point to which we will return later. In notable contrast to the thermal response of the individual enhancers, embryos with both enhancers raised at 30C have expression patterns that are statistically indistinguishable from those raised at 22C (see Fig 4.21D-H).

Thus, with some quantitative differences in the degree of change, it appears to be commonly true that adding the presence of a second regulatory element with overlapping potential to induce transcription does ensure a more uniform pattern of transcriptional activity. There are two possible consequences of this difference in frequency of active transcription. If the switch between actively transcribing and not transcribing is fast, then embryos which have a smaller fraction of cells transcribing at any given time will not accumulate as much total transcript. As the particular subset of cells which are transcribing or not changes frequently in this case, after a substantial time window each cell will have gone through several rounds of activity and silence, with total time spent in the transcribing phase equal to the fraction of cells observed transcribing at any given instant. Thus expression levels will be uniform, and reduced by the same fraction as the frequency of active transcription.

The alternative situation is that switching between the actively transcribing and transcriptionally silent modes is slow, and happens only a few times while the gene is activated. In this case the fraction of cells which are transcribing at the instant of fixation are expected to have been transcribing for some time and likely to have continued in the induced mode after the fixation time, and similarly the silent cells have likely been silent for some previous time as well (since switching between the modes is slow). As such, some

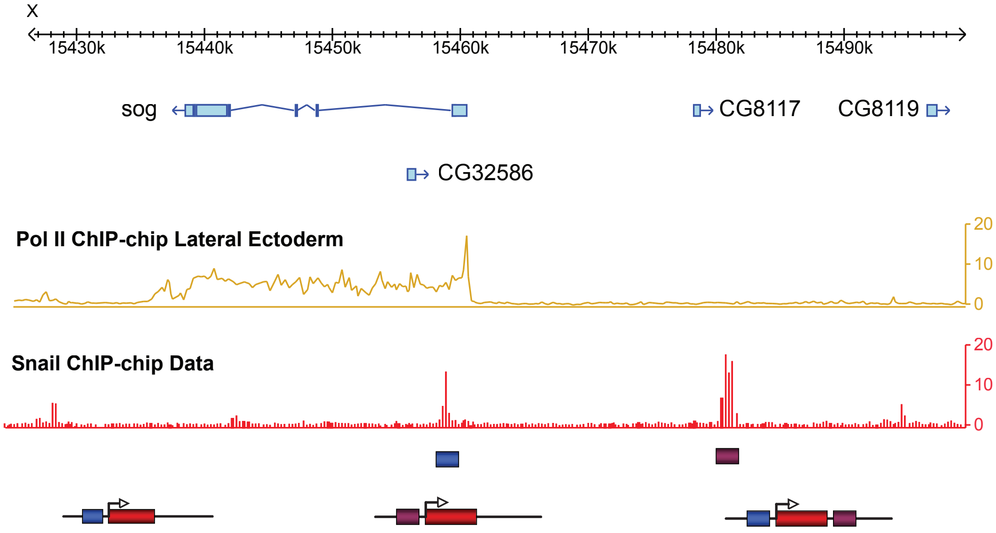

Figure 4.20: Location of multiple enhancers of the gene *sog.* Shown in yellow is pol II CHiP-chip data from toll[rm9]/toll[rm10] embryos, which produce primarily neurogenic ectoderm tissue at the expense of dorsal ectoderm and mesoderm. Shown in red is whole embryo *snail* transcription factor binding CHiP-chip, a strong repressor of *sog* expression in the mesoderm. The proximal (blue box) and distal (purple box) enhancers are marked at their respective locations, and schematics of the transgenic reporters are shown below.

cells may produce substantial levels of transcript (similar to that observed in the endogenous case), while those that fail to switch into the active case will produce substantially less.

We undertook to distinguish these cases by comparing the relative intensity of the cytoplasmic mRNA around each nucleus for the case of the gene *sog* (Fig 4.22). To quantify this difference we compute the median intensity of staining in each cell, and plot the coefficient of variation (standard deviation over mean) for all cells in the expression region, in each of the transgenic backgrounds.

In embryos with only the distal enhancer raised at 22C moderate variation in intensity even among immediate neighbor cells is detectable (see Fig 4.22A), with a coefficient of variation around 0.45. Embryos with only the proximal *sog* enhancer show similar degrees of variation (Fig 4.22B), which are statistically indistinguishable from those with the distal enhancer alone (Fig 4.22H). Substantially less cell-cell variation is detectable in embryos with both enhancers ($p < 0.002$), even though the difference in the frequency of detected nascent transcript is small. Dramatically greater variation is observed between cells in embryos raised at 30C with just the distal enhancer (Fig 4.22D). Surprisingly, little change is observed in the heterogeneity of transcript accumulation for embryos with just the proximal enhancer when raised at this elevated temperature (Fig 4.22E), though the variation is still significantly greater than when both enhancers are present (Fig 4.22G,H).

Effects of thermal stress

The differential behavior of the two *sog* enhancers with respect to temperature stress raises several questions. The sensitivity only under stressed conditions was one of the early features attributed to 'shadow enhancers' [17, 120, 119, 41]. Under genetic stress genes which had shadow enhancers showed less response than those which lacked them. Similarly in the case of *shavenbaby* and *snail* appreciable effects on gene expression and resulting phenotypes were only observable under mild stress conditions, in particular, temperature stress. Under ideal, controlled laboratory conditions, the population of embryos with single enhancers were indistinguishable from those with two.

Deeper analysis has now shown that some genes which have shadow enhancers require the activity of both even in laboratory conditions to quantitatively match the transcriptional activity profiles of the endogenous gene (e.g.

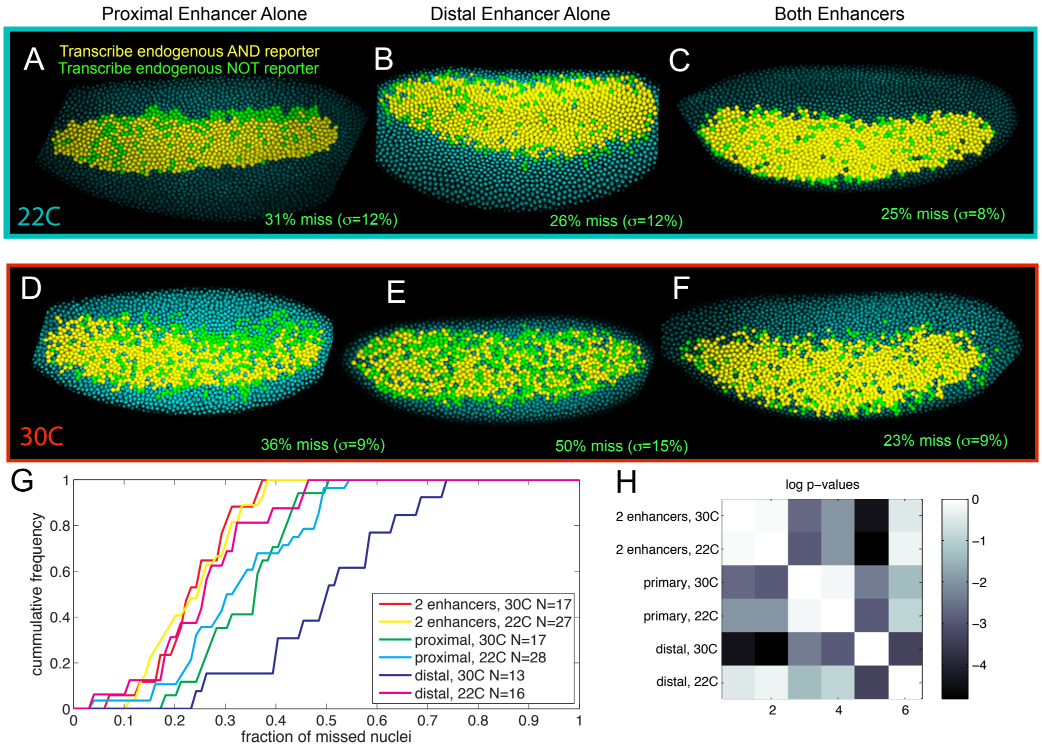

Figure 4.21: **Fidelity of reporter activation,*sog*]**. **A-C** Comparison of transcriptional state of the endogenous *sog* locus to a LacZ reporter gene in cycle 14 embryos at 22C, driven by either the *sog* proximal enhancer (A), *sog* distal enhancer (B), or both enhancers, proximal upstream, distal downstream (C). Yellow cells are transcribing both the endogenous and reporter gene. Green cells are transcribing the endogenous gene but not the reporter. **D-F**, as above, but embryos raised at 30C. **G** Cumulative frequency distributions for the three populations of embryos with statistical comparison. **H** log p-values from pairwise Wilcoxon rank-sum test of statistical significance for the difference in the distributions.

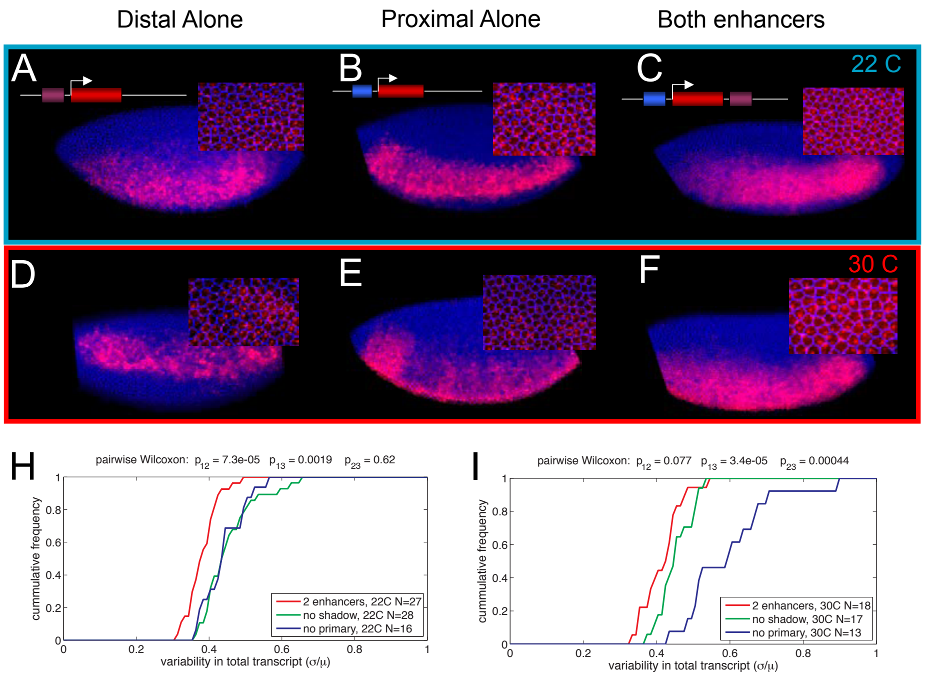

Figure 4.22: **Effect of stochastic transcription patterns on total mRNA synthesis** **A-C** variation in intensity of total mRNA among cycle 14 embryos at 22C, with reporter expression driven by either the *sog* proximal enhancer (A), *sog* distal enhancer (B), or both enhancers (C). Color patches represent median intensity in each cell. Insets show raw staining with overlay of cell boundaries. **D-F** As in A-C but embryos raised at 30C. **G** Cumulative frequency plot of the coefficient of variation for all embryos raised at 22C. **H** As in (G) but embryos raised at 30C.

hb. Interestingly, the same mild temperature stress for *hunchback* does not substantially alter the expression pattern of any of the transgenic reporters lines for *hb.* This suggests that the temperature stress response is likely a specific consequence of certain enhancer architectures, and a separable feature the more common phenomenon of multiple overlapping enhancers effecting the uniformity of expression. The experiments with *sog* demonstrate moreover that differential sensitivity to temperature change may even be observed in the same tissue with different enhancers for the same gene.

Evolutionary conservation

One expectation from the association of extra enhancers with robustness to thermal perturbations is that species which live in more thermally variable ranges may have a greater frequency of such enhancers (though as we have seen, thermal stress is not necessary or sufficient for the enhancer number to have a quantifiable effect on transcription). Investigations into the conservation of the *sog* enhancers provide a suggestive example. Consistent with the hypothesis that these elements play a functional role in development despite the apparent redundancy, both enhancers are conserved across 40 million years of evolution at least, from *Drosophila melanogaster* out through *Drosophila virilis* (see Fig 4.23). However, the virilis proximal enhancer provides only a partial expression pattern, compared with that produced by the distal enhancer or in the other species like *willistoni* (see Fig 4.23).

Drosophila virilis is reported to have a narrower thermal range compared to the more broadly ranging *melanogaster*, generally prefer a more restricted range of cooler temperatures. The proximal enhancer which produces a less developed pattern in *virilis* is also slightly weaker than the distal enhancer at the cooler temperatures of 22C (see Fig 4.21A-B). However under mild temperature stress this enhancer is more robust. *virilis*' restricted thermal range might put less pressure to evolve (or keep) a fully complimentary proximal enhancer than required by the more traveled *melanogaster*.

4.4.3 Discussion

As methods for identifying enhancers have improved with the rapid expansion of post-genome tools and techniques [150, 63, 64, 121, 92] an increasing number of enhancers with apparently redundant functions are being discovered. Here we have shown that may of these enhancers in the early *Drosophila*

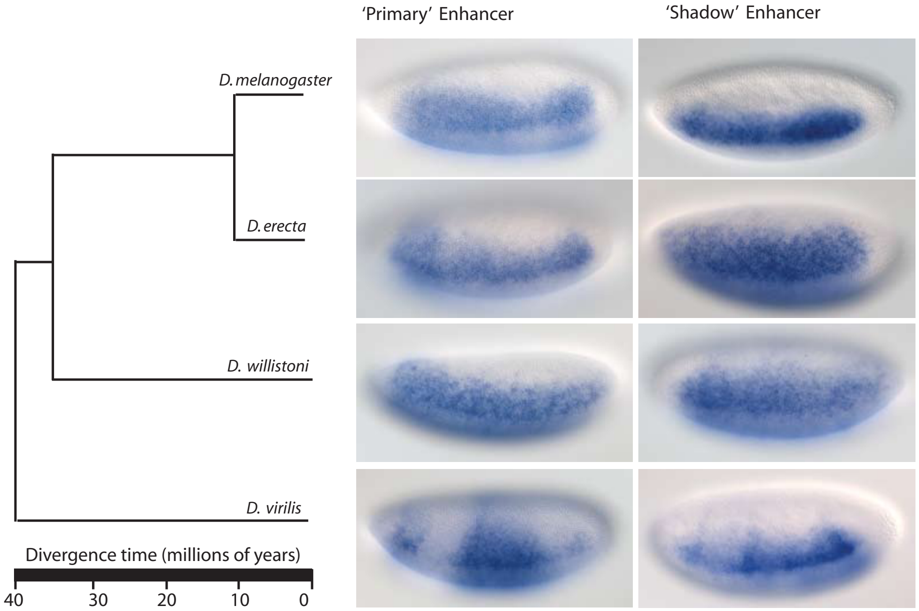

Figure 4.23: ***sog* lateral ectoderm enhancers across the Drosophilids**. Left, phylogenetic tree showing species divergence. Right, *sog* lateral ectoderm enhancers isolated from each of the species and tested in transgenic reporter assays in *Drosophila melanogaster*. Enhancers from the 4 species shown here we identified, tested, and imaged by Michael Perry.

lead to the creation of more uniform patterns of transcriptional activation by increasing the frequency of transcription for each cell. This observation that enhancer number itself actually effects transcription frequency itself prompts some adjustment and revision to the existing paradigm for gene regulation. It has been proposed that transcription activity is a direct reflection of enhancer site-occupancy [1, 163, 19, 14, 15, 174, 173]. These approach assumes that TF binding is fast, such that site occupancy reflects nuclear TF concentration. Then the probability that the gene is transcriptionally active is proportional to fraction of activation. While a useful framework for explaining observed dose-response relationships, threshold responses, and some gene network interactions, this approach provides no conceptual framework for understanding observed effects of multiple enhancers regulating the same promoter in the same cells – a consequence of some of its simplifying assumptions, as we shall see.

If the number of total transcription factors is small, the occupancy of an individual enhancer at any point in time may not reflect the occupancy expected from the average nuclear concentration of the transcription factor (due to Brownian motion and the associated Poisson fluctuations in the number of molecules in a small volume at any time) [11, 51, 8]. In this case, the additional enhancer can increase the effective sensor size to detect the low concentration of activator, increasing the probability of a sufficient number of binding events to activate transcription.

Alternatively, even if transcription factor binding to DNA is relatively fast and thus enhancer occupancy readily reflects the nuclear concentration of TF, as long as enhancer promoter interactions are sufficiently slow, a given nucleus may have fully occupied enhancers without initiating transcription. In this case as well, adding a second enhancer would increase the frequency of transcription by increasing the probability that one of the enhancers interacts with the promoter. The ability to modify transcription frequency by changing enhancer number therefore indicates a critical role in either time for binding or the rate of chromatin interactions between distal enhancer and their cognate promoters. Understanding the molecular components that effect these to processes will be essential to developing a predicative understanding of gene regulation.

Modification of the frequency of transcription through the number of enhancers acting on the gene may have two different effects: modulating the levels uniformly across the patterned field, or modifying the heterogeneity between cells in expression. Which of these two conditions results will depend on the relative kinetics of activation. From our experiments with the *sog* locus and in light of the rapid development of the *Drosophlia* embryo, we propose that the dominant effect is to reduce heterogeneity in expression.

Under thermal stress the embryo develops more rapidly and the relative stability of weak chemical interactions is reduced by the elevated thermal noise. Under these conditions one might expect more variable patterning. It is possible that the frequency of overlapping enhancer function ('shadow enhancers') in the *melanogaster* genome may buffer this variation and facilitate its survival in the wide range of thermal environments in which the species thrives today.

Chapter 5

Cross-regulation creates precise boundaries

Thus far we have considered only compact cis regulatory mechanisms that effect the expression of the gene in a feed-forward manner (promoter choice, number of enhancers). Much of the precision and robustness of gene expression I suspect also comes from the organization of the interactions between the regulatory genes. In this chapter I discuss my research into the regulatory interactions that establish one of the most robust and impressive cell-fate decisions in the embryo – the delineation of mesoderm from neurogenic ectoderm, the first true tissue boundary in the embryo which correlates with striking differences not only in gene expression but mechanical cell behavior on either side of the boundary.

This has been an outstanding problem and one of interest in the Levine lab for many years. While this work will not lay to rest all of the questions, I hope to provide compelling evidence that the process is one driven by cross-repressive feedback, and not a threshold induced by cooperative behavior as is commonly believed – a mechanisms I will argue that is both less robust and less precise for generating boundaries.

In the search for early embryonic transcription factors that could repress *snail* I screened over sixty different mutant and mutant combinations over two years. This works was greatly facilitated by Madhurima Promod, a dedicated undergraduate who worked with me from January 2009 through May 2010, and performed many embryo collection/fixations and in situs. At the time of writing these findings on cross-regulation and its role in mesoderm specification are all unpublished.

5.1 Introduction

The first true tissue boundary in the Drosophila embryo delineates the presumptive mesoderm from the neurogenic ectoderm. This boundary is marked and determined by the expression pattern of the gene snail, which is transcribed at uniform high levels throughout the mesoderm cells and not at all in their neurogenic neighbors by mid cell-cycle 14. This mesoderm tissue undergoes a collective epithelial to mesycnhemal transition (EMT) and invaginates to form a hollow tube inside the embryo (which will later collapse and spread out to form the various muscle tissues of the embryo).

It is an open question by what molecular mechanisms this sharp boundary is formed. It has been proposed that this switch like behavior arises from cooperative binding of the activators Dorsal and Twist to the *snail* regulatory sequences with such a high degree of cooperativity as to produce an apparent switch like response. In this section I will present definitive evidence that Dorsal and Twist binding, though important for the process, do not of themselves instruct this sharp border. Instead, this boundary arises from feedback interactions which require the function of the *snail* gene itself.

5.2 Results

5.2.1 Snail expression refines from shallow to sharp

By cell cycle 11, transcription of *snail* is first detectable in the ventral nuclei as bright puncta of nascent transcription (see Fig 5.1A). This expression is more easily detected in embryos where the *snail* coding sequence has been replaced by the yellow reporter gene [119], where the addition of around 2 kb of intronic sequence allows for sensitive detection of nascent transcript [17]. Cytoplasmic mRNA is only detectable at low levels in these early cycles. (Fig 5.1A-C). The Dorsal boundary of the expression region is very poorly defined and rather non-uniform. Some *snail* target genes, like *sog*, are also expressed during these cell cycles. *sog* expression spans the mesoderm and is readily detectable in the same cells which are transcribing snail, suggesting that Snail protein has not accumulated to substantial levels.

By mid cell cycle 13, *snail* mRNA has accumulated in a bell-shaped profile, centered on the ventral midline, and spanning some 16 cells across (Fig 5.1C). This expression profile expands to around 18 cells wide. Repression

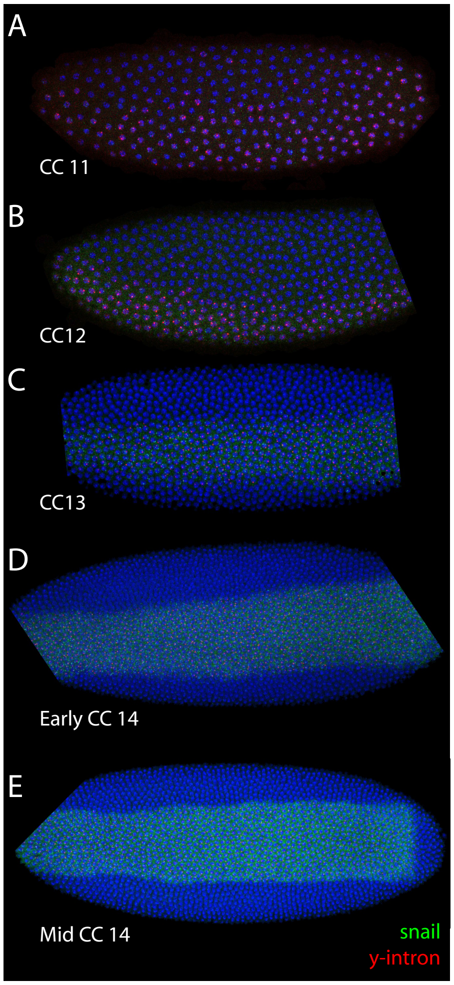

Figure 5.1: **Dynamics of early *snail* expression**

of *sog* is first detectable in ventral cells in late cycle 13, indicating that the protein concentration has accumulated to sufficient concentration to begin to effect target genes. At the beginning of cycle 14 enough Snail protein and mRNA have decayed during the brief mitotic interval that *sog* expression is generally detected throughout the mesoderm once more. Transcription of *snail* is observed in a band some 18 to 20 nuclei wide, and somewhat narrower just anterior to where the cephalic furrow will form (Fig 5.1D, 5.2G-O). The

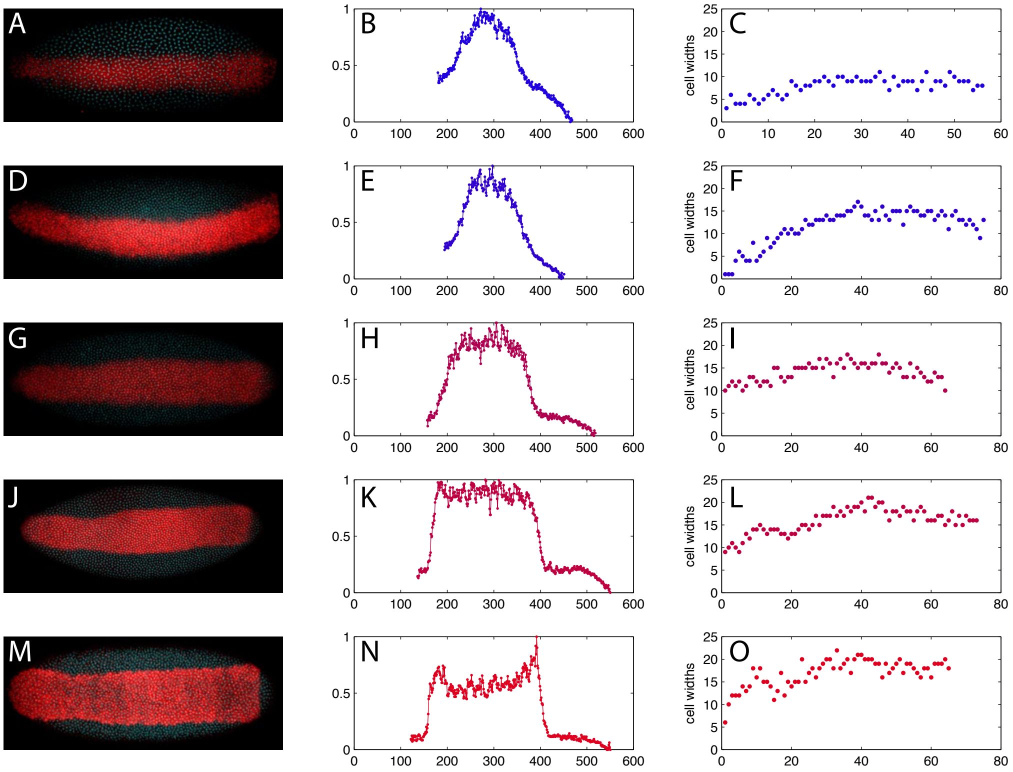

Figure 5.2: **Dynamics of cell cycle 14 *snail* expression**

total mRNA still forms a graded expression profile. By mid cycle 14 this bell-shaped profile of *snail* has refined to a box shape, with uniform expression throughout the now delineated mesoderm some 18 to 20 cells wide (Fig 5.1E). After this sharp border is formed, *single-minded* (*sim*) expression is first detected in adjacent midline cells in an initially incomplete band that gradually fills out to include all midline cells just prior to the onset of gastrulation (see Fig 5.3A,C).

It appears that several of the midline cells that at the point of gastrulation have undetectable levels of *sna* mRNA previously transcribed the gene at a lower rater while the profile was more Gaussian shaped, though this transition is difficult to evidence clearly in a wildtype background without live imaging. Unfortunately available live mRNA imaging techniques dramatically increase mRNA stability, making them unsuited for this analysis.

5.2.2 Snail activity is required for normal *snail* expression

In embryos containing a recessive lethal complete loss of function allele for *snail* (*sna*[18]), early gene expression patterns are indistinguishable from wildtype expression. Though *snail* in these embryos is initially expressed throughout the ventral regions of the embryo, by partway through cell cycle 14 transcription slows dramatically, as evidenced by the loss of nascent transcripts. The border of mRNA expression remains graded and highly variable in its dorsal extend (compare Fig 5.3A and B). Expression of sim appears in patches in mesoderm (Fig 5.3D), but is not activated in the presumptive mesectoderm cells where it is normally expressed (Fig 5.3C). By the time the cephalic furrow has formed ventral furrow formation and gastrulation, which normally occur at this point in wildtype embryos (Fig 5.3E) do not show any sign of commencing (as previously observed for *snail* mutants [89]). Also as expected, lateral ectoderm genes such as *vnd* and *sog* are robustly expressed through the mesoderm (see Fig 5.3F). Surprisingly, *snail* expression itself is further effected, levels of mRNA are substantially lower, nascent transcription is rarely detectable, and the accumulated mRNA from the early expression starts to disappear (compare Fig 5.3E and F).

In addition to these effects on the expression pattern of *snail* mRNA, loss of Snail function results in decrease or complete loss of expression of several other mesodermal genes. Using double labeling with the *snail* mRNA antisense probes and intronic antisense probes for other mesodermal genes, we analyzed *sna*[18] mutant embryos for changes in transcription activity. In these embryos by mid cycle 14 *htl* transcription was almost entirely halted (see Fig 5.4A, compare B). Similarly transcription of Mes2 ceased in essentially all nuclei (Fig 5.4C-D), and transcription of *hbr* appears to be reduced, though less dramatically than Mes2 or *htl* (Fig 5.4E-F). By the time the cephalic furrow is fully formed and mesoderm invagination would have completed in wildtype embryos, staining of twist mRNA is still detectable throughout ventral regions in a broad band appears to be somewhat weaker and with a less well defined border in these mutant embryos. *snail* expression by comparison is more substantially reduced, detectable only in weak patches more ventrally restricted than *twist* (see Fig 5.4G). Normally by the onset of gastrulation both genes are highly expressed in congruent cell populations, as readily seen in the green red double labeling experiment in fig 5.4G.

It should be noted that the snail-null point mutant, (*sna*[18]), created through EMS mutagenesis, probably exhibits a small degree of nonsense read-through translation (perhaps 10%), resulting in low levels of functional Snail. This causes this null mutation to fail in some respects to fully phenocopy the deletion alleles. The gene *sim*, normally expressed in the midline cells adjacent to the mesoderm is expressed in the ventral-most nuclei. However in *sna*[18] mutant embryos (also called *sna*[IIG] in the earlier literature) expression is a rather patchy fashion, unlike the broad, uniform band of sim expression throughout the presumptive mesoderm observed with the deletion allele[60]. It is possible even very weak levels of *snail* are sufficient to partially repress the expression of *sim*, (Fig 5.3D), but are not sufficient to repress other ectodermal genes (Fig 5.3F).

Since a normal *sna* boundary fails to form in the absence of Snail activity, this boundary must require additional input and interaction than simply a cooperative readout of Dorsal and Twist (contrary to what has previously been proposed) [83, 89]. Such a mechanism might also help explain why the dorsal border of *snail* expression does not move in *dorsal* heterozygotes [119], which have substantially reduced levels of Dorsal [23] and do dramatically effect the expression of Dorsal responsive transgenes [69]. One possible role is that a lateral ectoderm expressed gene represses mesodermal targets a case Leptin *et al* noted is not excluded by their analysis of *twist* and *snail* mutants [89], of which *snail* itself may be a prime candidate (a situation not previously considered). Such a gene (or genes) being also repressed by *snail*, could easily establish through mutual repression a bistable condition which could explain the sharp boundaries observed in this expression. Moreover, if the repressor of *snail* was activated by Dorsal and Twist, it might allow for maintenance of the boundary position (see below).

5.2.3 Ectopic activation of Dorsal and Twist repress regions of *snail* expression

To test the hypothesis that the Dorsal and Twist gradients may activate repression of *snail*, we examined embryos in which an ectopic gradient of Dorsal and Twist is induced from the anterior pole and asked how this affects *snail* expression. This ectopic activity was induced by maternal expression

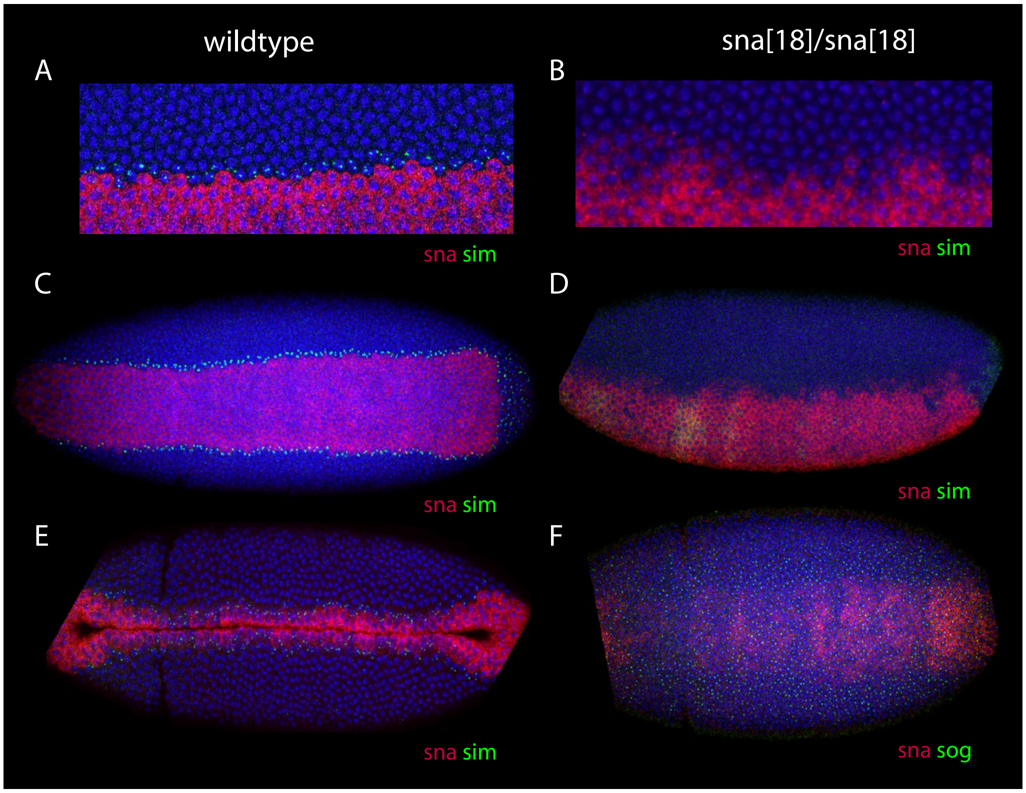

Figure 5.3: **Snail activity is required for normal *snail* expression.** A. endogenous *snail* forms a sharp border traced by a mutually exclusive pattern of *sim* expression. B. In *snail* mutants, the boundary of *snail* mRNA expression is considerably less straight and somewhat less sharp. C. *sna* (red) and *sim* (green, intronic probe revealing only active transcription foci) expression in wildtype embryos. D exonic *sim* and *snail*, in *snail* mutant embryos. E. Wildtype embryos form a uniform ventral furrow and gastrulate almost completely by the time the cephalic furrow is mature. *snail* expression remains strong with sharp boundaries F. *sog* is actively transcribed throughout the mesoderm and lateral ectoderm in *sna* mutants.

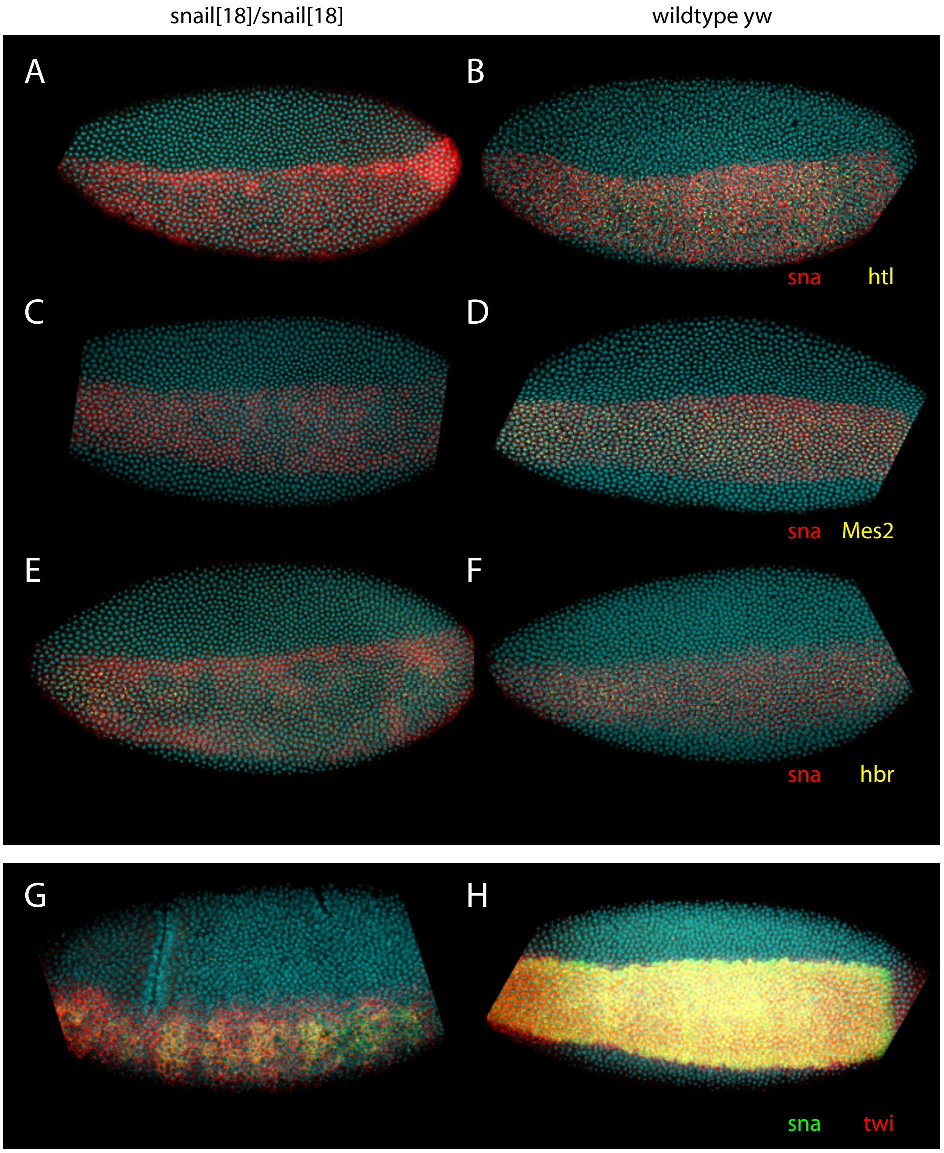

Figure 5.4: **Snail dependent gene expression** A-F Expression of several other mesodermal genes is reduced or lost in a *snail* mutant background *snail* mRNA is shown in red and intronic probes for transcription of the chosen mesodermal gene are in yellow, *htl* (A,B), Mes2 (C,D), and *hbr* (E,F). G-H. Snail expression (now in green) is also dramatically reduced in a *snail* mutant background, and *twist* expression (here stained in red) appears a little weaker and more patchy.

of mRNA (using the hsp83 enhancer sequence) encoding a constuitively active form of the toll receptor (toll[10b]), localized to the anterior pole of the embryo using the bcd 3'-UTR [66]. The subsequent activation of Dorsal is sufficient to activate mesodermal genes in the anterior regions, neuroectoderm genes in the middle regions, and dorsal ectoderm genes in posterior regions of the embryo [66]. This gives rise to a 'T' shaped expression pattern of mesodermal marker *snail* (or 'L' shaped in a lateral view, see Fig 5.5A). However, in approximately half of the embryos from this population, the endogenous ventral expression of *snail* is repressed (see Fig 5.5B-D), producing a gap between the anterior ectopic activation of *snail* and the usual mesodermal activation of *snail*.

In all cases the *snail* expression pattern by mid cell cycle 14 has sharp boundaries of expression. The position of the anterior-posterior boundary of the induced *snail* expression is highly variable (see Fig 5.5A,B). Additionally the width of the gap when a gap is induced varies considerably (Fig 5.5C,D). This variation is suggestive of a bistable system in which small variations in the initial concentrations of the two mutual repressive factors through feedback develop into highly divergent stable states.

We then generated a gradient of just Twist from the anterior pole using the same maternal driver and bcd-3' UTR localization of the *twi* mRNA to test if Twist alone produced any interference with the endogenous *snail* patterning. Twist is not sufficient to activate *snail* in the absence of Dorsal and therefore no ectopic *snail* expression is observed in these embryos. However, the endogenous ventral expression of *snail* narrows dramatically along the DV-axis (see Fig 5.5F compare to E). This surprising for two reasons. First, Twist is a known activator of *snail*, an is required for proper *snail* expression [83, 89, 69]. Secondly the ectopic expression is induced along the AP axis, but the primary effect of expression is re-oriented along the DV axis. This is reminiscent of the ability of Sog to reorient and ectopically induced AP pattern of *dpp* (expressed under control of the *eve* stripe 2 enhancer) to nonetheless rescue mostly normal DV patterning in *dpp* mutants [166] (*dpp* is normally expressed in a DV pattern restricted to the dorsal ectoderm). The expression domain of *twist* narrows concurrently with the change in the domain of *sna* (Fig 5.6A-D). Note also the ectopic expression results in higher levels of nuclear localized twist than the endogenous twist in the ventral mesoderm (Fig 5.6B).

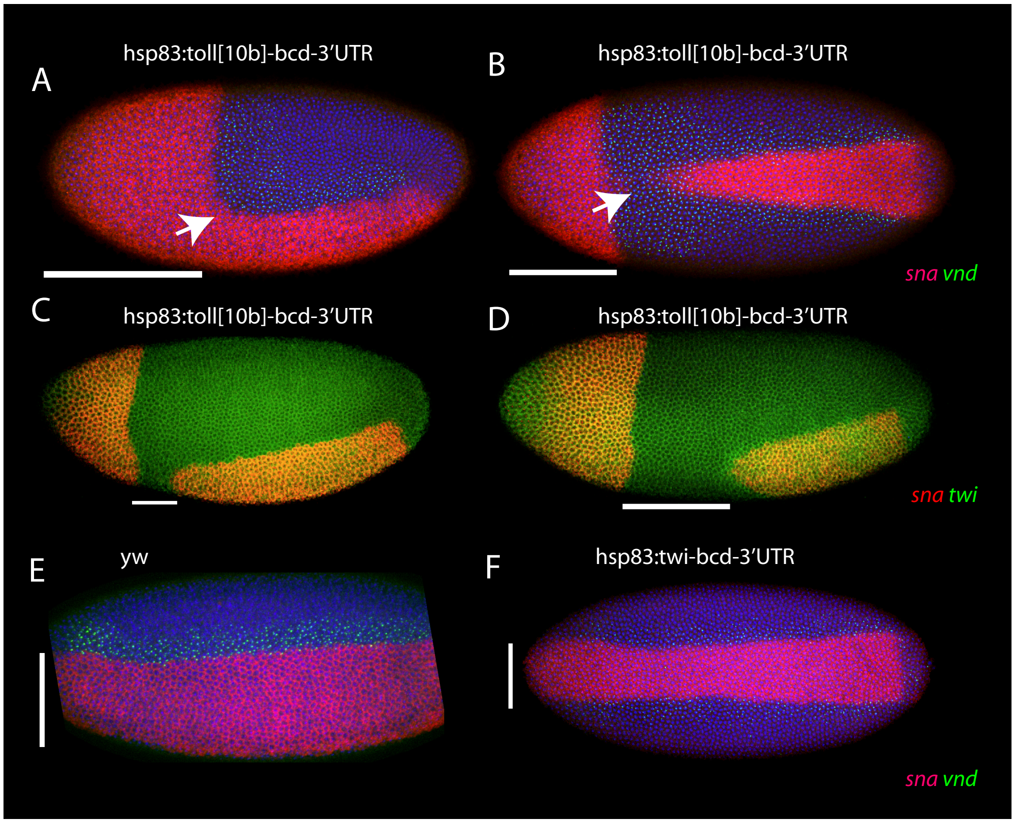
A
hsp83:toll[10b]-bcd-3′UTR
B
hsp83:toll[10b]-bcd-3′UTR
sna vnd
C
hsp83:toll[10b]-bcd-3′UTR
D
hsp83:toll[10b]-bcd-3′UTR
sna twi
E
yw
F
hsp83:twi-bcd-3′UTR
sna vnd

Figure 5.5: **Evidence for *snail* repression** The pattern of *snail* is now around 10 cells wide at mid-embryo (white bar). A-B *snail* and *vnd* expression in embryos expressing constitutively active toll allele at the anterior pole. Note the variation in the posterior boundary of the *snail* domain and the presence of a gap in expression. C-D overlapping *snail* and *twist* expression in embryos from the same background. White bars denote the variation in the width of the gap in *snail* expression. E wildtype expression of and and *snail* in cell cycle 14. The mature *snail* expression pattern is around 20 cells wide at mid-embryo (white bar). F expression of *snail* is narrower when *twist* is over-expressed from the anterior pole.

5.2.4 Ectopic *snail* gradients evolve into all-or-none expression profiles

We then asked if this repression acts in a concentration dependent fashion, turning off *snail* expression in moderate or weakly induced cells while maintaining peak production in strongly induced cells. Such a mechanism could

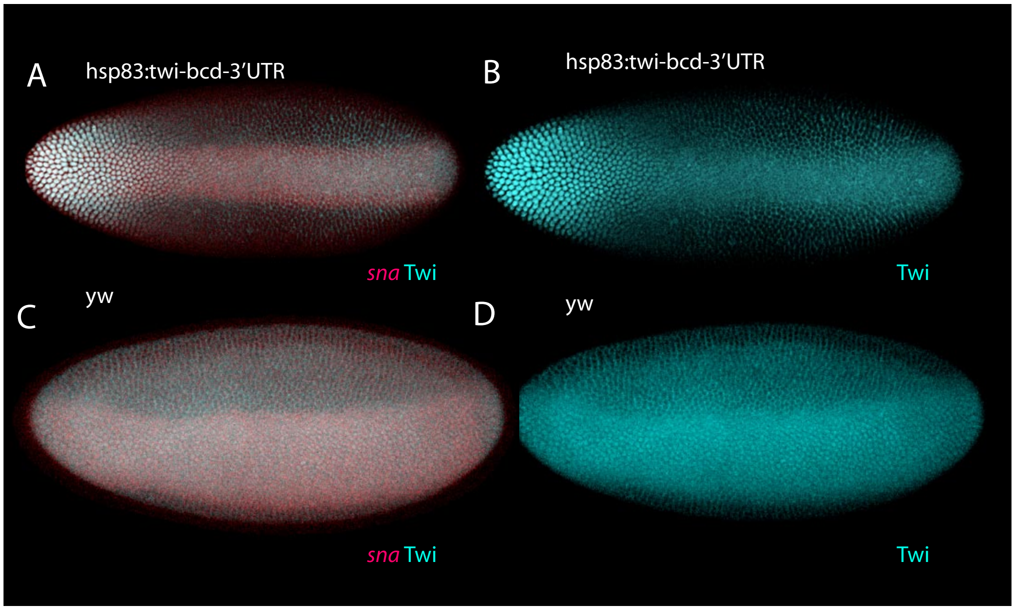

Figure 5.6: **Ectopic Twist represses endogenous *snail* and *twist*** A-B embryo from hsp83:twi-bcd-3'UTR mother expressing ectopic twist from the anterior pole, stained for twist protein (cyan) and *snail* mRNA. C-D comparison of broader expression in wildtype embryos.

help explain the change from a graded bell shaped profile seen at the beginning of cell cycle 14 (Fig 5.2B) into the steplike profile which emerges in mid to late cell cycle 14 (Fig 5.2K).

To test this hypothesis, we induced graded expression of *snail* in cycle 13 embryos and asked how this profile changes in response to any endogenous feedback and cross regulation which might refine the pattern into a bimodal expression distribution. To test refinement in response to ectopic expression is was important not to induce *snail* expression under direct control of heterologous enhancer elements, since such an element would lack the any feedback regulatory sequences which the endogenous locus may contain. Therefore we generated an ectopic gradient through misexpression of both Dorsal and Twist. We created embryos which have uniform, moderate levels of Dorsal and an aterior-posterior gradient of Twist by driving maternal expression of *twist* with a 3-prime Bcd UTR under the hsp83 driver in females with the toll alleles toll[rm9]/toll[rm10] (see methods).

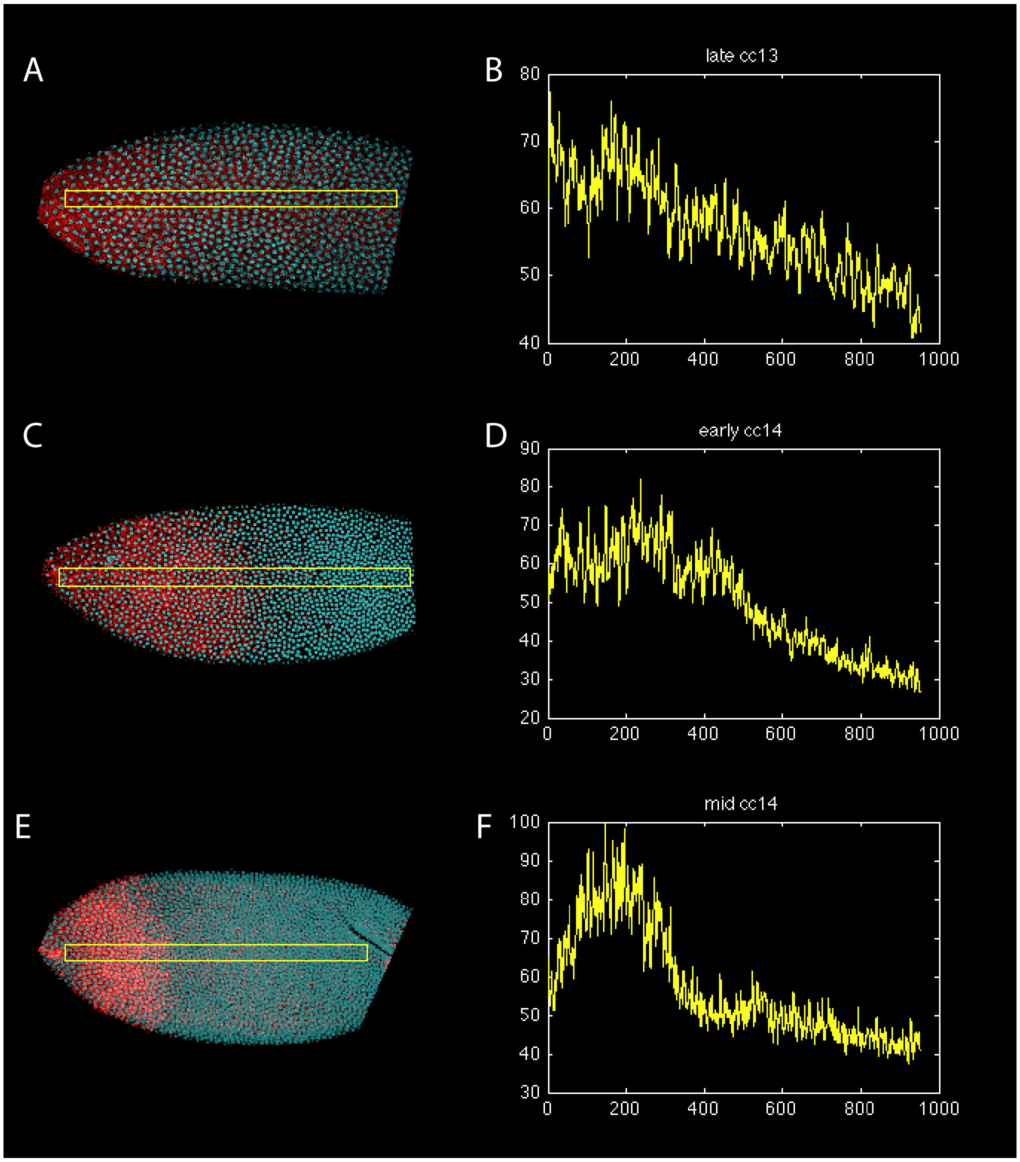

Figure 5.7: **Ectopic *snail* gradients refine A-B** young cc13 embryo, **C-D** early cc14 embyro, **E-F** mid cc14 embryo. B A,C,E, *snail* is stained in red, nuclei are in cyan. Yellow boxes indicate regions used to measure gradient, plotted in B,D,F

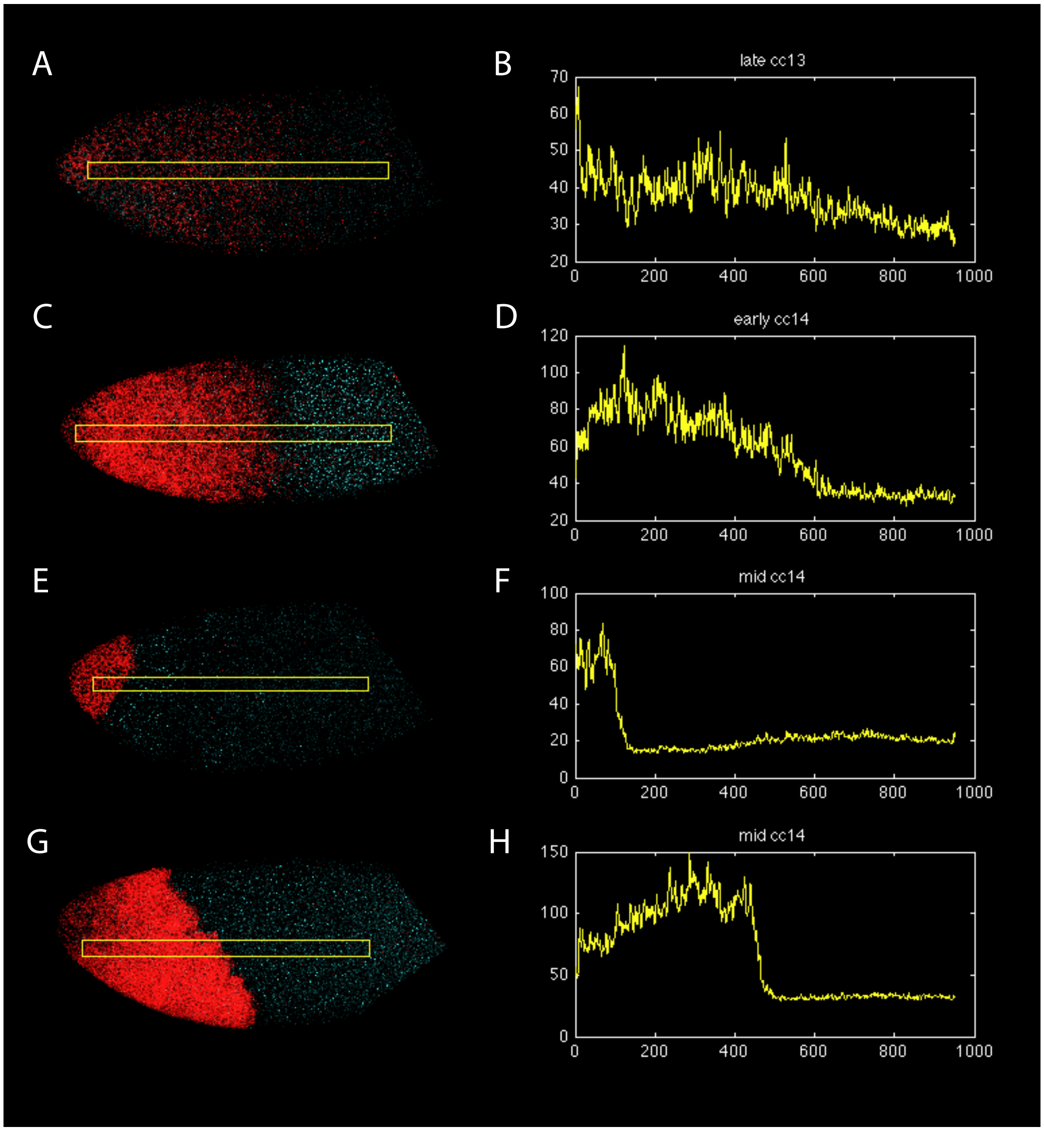

Figure 5.8: **Ectopic *snail* gradients refine** A,C,E,G, *snail* is stained in red, *sog* is stained in cyan. Yellow boxes indicate regions used to measure gradient, plotted in B,D,F,H.

Embryos which contain an anterior gradient of Twist alone in an otherwise wildtype background do not show any ectopic anterior expression of *snail* (see Fig 5.7). However when increased levels of nuclear Dorsal are also induced through the *toll[rm9]/toll[rm10]* background, *snail* is expressed in anterior regions. As expected, the initial expression pattern is broad, and transcription can be detected along almost the entire length of the embryo (see Fig 5.7A). Expression levels are not uniform however, with higher rates of transcription in anterior regions leading to a smooth gradient of expression maximal at the anterior pole and minimal at the posterior pole see Fig 5.7B). The gradient is substantially shallower than that observed for Bcd along this same axis. By the beginning of cell cycle 14, the smooth gradient has started to refine, with an inflection near the middle of the embryo (Fig 5.7C-D). by mid cycle 14, *snail* expression is substantially sharper still, dropping sharply over three of four cell instead of the more gradual change observed in younger embryos. This sharpens further to a discrete pattern of on and off expression on a border whose posterior extent varies over many cell diameters in with a local correlation (Fig 5.7C-D). The steady state boundary position is highly variable among clonal embryos (e.g. Fig 5.8E-H). The variation within an embryo along the AP position can be equally substantial, resulting in oblique boundaries (compare Fig 5.8E and G). Snail is also repressed by *hkb* at the anterior pole (this repression is normally seen only in the posterior pole, where early *snail* expression overlaps the endogenous *hkb* expression region). This repression causes *sim* to be induced throughout the anterior region, inverting the usual expression order of these genes, as previously reported, [152].

5.3 Discussion

The formation of the mesoderm boundary is the first critical cell-fate decision in the developing embryo and one that is essential for establishing normal gastrulation behavior. Here we have presented evidence that this boundary is established through cross-repressive interactions of *snail* with lateral ectoderm repressors, in place of the previously proposed system of threshold readout through site occupancy of cooperatively binding activators.

There is a well established belief that in Drosophila development tissue boundaries are established by threshold readout of morphogen gradients through cooperative binding of activating factors [83, 73, 70, 75, 136, 137, 97, 51, 9, 59, 3, 32]. While an attractive model and one routinely reasserted, accumulating evidence suggests that the fate decisions arise from cross-regulatory interactions where the relative levels of transcription factors are more important than absolute concentrations suggested by morphogen gradient models. These two possible mechanisms, threshold readouts and cross-regulatory interactions, differ in several fundamental ways. The most important difference to the study of precision and robustness in gene expression, is that in the threshold readout scheme, the difference between activated and inactivated cells for a particular gene comes down to whether or not the regulatory sequence ever 'sees' enough of the inducing signal to turn on. This means that boundary cells are forced to make a decision at the point where there is maximal uncertainty in the signal – when it as weak as it can be and still be detectable. In contrast, no such condition is required in a cross-regulatory scheme, and all the molecular noise involved with minimal probability binding can be avoided.

An accumulating body of evidence argues against the threshold readout model as the mechanism for determining cell fate decisions. For example in embryos with globally reduced concentrations of bcd (1x *bcd*, *vas* null, *exu* null), the *hunchback* expression can still be induced in cells experiencing what for wildtype embryos would be sub-threshold levels of bcd. Similarly in embryos with ectopically enriched gradients of bcd, cells experiencing substantially greater than the normal level of Bcd required for activation still do not express head genes like *otd*, *ems* or *btd* [110]. Extensive analysis of Bcd responsive enhancers has also demonstrated a lack of correlation between strength of Bcd binding site clusters an the AP boundary position [111]. Additional work by the Small and Reinitz labs [24, 101, 102] has shown that a system of cross repressive interactions between the gap genes is responsible for converting the position information from the Bcd and Hb morphogen gradients into sharply deliniated boundaries of different cell fates required in body plan segmentation.

We argue that ultimately cross-repressive regulatory interactions establish the cell fate decisions along the dorsal-ventral axis as well, and that such a mechanism is a general approach to cell fate decision in development. As with the gap patterning system, cross repression in the DV patterning is initated by differential response of repressive transcription factors to different levels of the morphogen signal. Where Bcd, Hb, and Cad provide the morphogenic signals along the AP system [35, 169, 136], Dorsal, Twist, and Dpp provide morphogen signals along the DV axis. With only lower affinity Dorsal binding sites [69, 75] *snail* responds to the high concentration of Dorsal in ventral regions. Lower levels of Dorsal and Twist activate repressors in the lateral ectoderm. These are repressed by Snail ventral regions, but themselves repress expression of *snail* in more lateral regions, preventing *snail* expression from expanding. Repressors from the Enhancer of Split complex may play a role in mediating this repression.

5.4 Methods

5.4.1 Fly Genetics

The hsp83:toll[10b]-Bcd-3'UTR [66] and hsp83:twi-Bcd-3'UTR [149] lines are marked with mini-white cassette and are female sterile. They were maintained through the male line by crossing to *y,w* virgins each generation. To make females of genotype *y,w*; hsp83:twi-Bcd-3'UTR; *toll[rm9]/toll[rm10]*, we introduced the dominant third-chromosome markers *Pr*, *Dr* and balancer TM3, *Sb*, *Ser*. Similarly we introduced a marker *wg[Sp]* over the CyO balancer into the *toll[rm9]* and *toll[rm10]* lines (balanced over TM3 *Ser*), and *y,w* X chromosomes to allow for eye marker selection. Males of genotype *y,w*; hsp83:twi-Bcd-3'UTR; Pr,Dr/TM3 *Sb*,*Ser* were crossed to virgin females *y,w*; *Sp*/CyO; toll[rm9]/TM3 *Ser* and male progeny of genotype *y,w*; hsp83:twi-Bcd-3'UTR/CyO; toll[rm9]/TM3 *Sb*,*Ser* were collected and crossed to virgin females *y,w*; *Sp*/CyO; toll[rm10]/TM3 *Ser*. Females *y,w*; hsp83:twi-Bcd-3'UTR; *toll[rm9]/toll[rm10]*, were collected and crossed to *y,w* males for embryo collection.

5.4.2 Quantitative Analysis of Expression Patterns

Custom image processing scripts were written in Matlab and to segment the nuclei using a difference of Gaussians filtering approach and to identify the *snail* expressing cells based on threshold detection, as previously described [119]. Width of *snail* expression was measured both in pixels and in terms of nuclei spanning the pattern. To count the nuclei across the width of the expression pattern we first created an adjacency neighbor matrix of all nuclei from our nuclear segmentation. We then identified all boundary cells from pre-oiriented images of ventral views. A graph shortest path algorithm was used to determine the number of cells between each boundary cell on the top half of the expression pattern and its cognate partner on the bottom half. These analysis steps were carried out through a custom Graphical User Interface to streamline the segmentation process. The code for this analysis can be found online at https://github.com/AlistairBoettiger/Image_Analysis.

Acknowledgements

I have been fortunate to have the support of many wonderful and insightful people throughout my research.

I would like to thank Kate Senger, Joung-Woo Hong, and Jessica Cande for technical assistance with the experimental setup and design for my investigations on Paused Polyermase, and Dave Hendrix for help with genome browser for viewing the CHiP-chip data. Jessica Cande provided the hsp83:bcd-3-UTR fly line and the dl[6] null allele fly line used in these experiments. I would also like to thank Alberto Ponce, a REU student who worked with me during summer 2008 and successfully designed and tested several probes to look at gene activation.

Peter Ralph and Steven Evans really made the mechanistic modeling investigations possible. I had been struggling with simulations and deterministic models when my course work in Stochastic Processes convinced me there were better tools to deal with these problems, if I could become a bit more proficient at them. Peter and Steve made this possible, not only helping to analyze the complex processes I was interested in, but also teaching me with great patience the tools and logic behind the process.

I thank Joung-Woo, Rob Zinzen, and Jessica Cande for fly lines for my early investigations into the expression and fidelity of shadow enhancers. Most of the constructs I ultimately analyzed were designed and built by Michael Perry, whose dedication, prodigious construction of difficult transgenic lines, and lively discussion have been a critical part of all our studies of shadow enhancers. Jacques Bothma provided many useful suggestions and discussions for code development, test piloted code for publication, and developed some creative and powerful tools for image segmentation – including the strategy of recoloring nuclei to represent presence or absence of nascent transcripts rather than trying to trace and count dots by eye.

In this project I would also like to thank Chiahao Tsui, Madhurima Pramod, and Emilia Esposito for technical support, Nipam Patel, Jessica Cande, Benjamin Haley, Vivek Chopra, and other members of the Levine lab for helpful suggestions. Russell Vance for recombineering advice and reagents, Tony Ip for providing the sna2.8 enhancer lines and the osp[29] line, Steve Small for providing the tll (+4) enhancer line and Gary Struhl for providing the BB9+16 bcd extra copy line.

My qualifying exam committee (Adam Arkin, Dan Rokhsar, George Oster, and Steven Evans), provided both critical feedback and encouraging

support at the design stage of my research projects, and continued in the capacity of my thesis committee to provide feedback and suggestions. George has been a mentor and advisor throughout, and I have greatly valued the ability to have an ongoing dialogue about research, science, and graduate school with George outside of the structured meetings of the qual and thesis committees. He is unfailing generous with his time and good advice, and has also helped connect me with several exciting faculty at other universities with whom I may pursue my post-doctoral research.

I had the pleasure to have Adam not only as the chair of my qual committee (which was great for everything except scheduling) but also as the instructor for the Principles of Synthetic Biology course for which I was head GSI. His energy, creativity and encouragement were a source of inspiration and his insightful comments a source of guidance. The semester we spent together designing a new course for an emerging discipline helped new ways of thinking about both my research, from experimental design through research aims, and my teaching and emerging roles for technology in improving the experience.

I would like to thank all members of the biophysics program for being such a wonderful, supportive and creative community. In especial Kate Chase, for her constant support, information, and help with any paperwork and administrative requirements which the process of my degree required sorting out. Also the Program Chair Udi Isacoff, for helping build and run such a strong, flexible and stimulating program, and our new chair Eva Nogales for her dedication to improving the program and involvement of students in the decision making involved. Susan Marqusee, our Graduate Head of Studies was always responsive and flexible, and I am grateful for the latitude, trust and encouragement she always provided.

I similarly like to the thank the members of Computational Biology Designated Emphasis and seminar community. This was an invaluable resource for me to get more exposure to the breadth and depth of genomic research going on, and to meet more computational and statistically minded scientists to provide feedback on my research aims. In particular I'd like to thank Brian McClendon for his excellent administration of this program – allowing maximal latitude and opportunity to the students while handling all the complicated elements from scheduling to funding to advertising of the seminars, retreats, lunches, and other program activities.

My most consistent collaborator has been my twin brother, Carl. We discuss science most weeks, which help with everything from solving explicit mathematical derivations to big ideas about the direction of academic science and the new tools which will revolutionize its methodologies. Most of the tools which have had the biggest impact on the way in which I conduct my research I have learned from my brother, including electronic and web based notebooks, library and database management, RSS publication tracking, code version management, and data organization and distribution.

Throughout the process my parents, David and Barbara, have been a constant source of support. I am grateful for the passions they instilled in me from a young age, a love of scientific discussion and a love of original inquiry. Through their own lives they have been for me role models and a source of inspiration and in my life a source constant support and encouragement. Thankyou mum and dad.

Most of all I thank my thesis advisor and mentor Mike Levine. Whether it was some crazy new experimental idea or some forgotten detail from the early days of developmental genetics, Mike was happy to discuss it and provide either useful suggestions or rattle off a collection of useful references which would have the details. He was unfailing source of motivation (needed or not), and maintained a research environment at once both comic and serious – a fusion only members of the lab will really understand. His clarity of expression when discussing scientific ideas and his appreciation of the truly insightful observations from the 'archival advances' have been for me a model and an aspiration, and definitely something I will value in my future work as much as the technical skills I acquired while working in his lab. In sum, it has been an unforgettable experience to live and work with this lab and I will treasure both the lessons learned and the stories spun from this unique environment.

Bibliography

[1] G K Ackers, A D Johnson, and M A Shea. Quantitative model for gene regulation by lambda phage repressor. *Proceedings of the National Academy of Sciences of the United States of America*, 79(4):1129–33, February 1982.

[2] Karen Adelman, M T Marr, J Werner, A Saunders, Z Ni, E D Andrulis, and John T Lis. Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. *Molecular Cell*, 17:103–112, 2005.

[3] Naama Barkai and Ben-Zion Shilo. Robust generation and decoding of morphogen gradients. *Cold Spring Harbor perspectives in biology*, 1(5):a001990, November 2009.

[4] Alejandro Barrallo-Gimeno and M Angela Nieto. The Snail genes as inducers of cell movement and survival: implications in development and cancer. *Development*, 132(14):3151–61, July 2005.

[5] Mark Bates, Bo Huang, and Xiaowei Zhuang. Super-resolution microscopy by nanoscale localization of photo-switchable fluorescent probes. *Current Opinion in Chemical Biology*, 12(5):505–14, 2008.

[6] Mary K Baylies and A M Michelson. Invertebrate myogenesis: looking back to the future of muscle development. *Current opinion in genetics & development*, 11(4):431–9, August 2001.

[7] A Becskei, Alexander van Oudenaarden, and B B Kaufmann. Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nat. Genet.*, 37:937–944, 2005.

[8] H C Berg and E M Purcell. Physics of chemoreception. *Biophysical journal*, 20:193–219, 1977.

[9] Sven Bergmann, O Sandler, H Sberro, S Shnider, E Schejter, B Z Shilo, and Naama Barkai. Pre-steady-state decoding of the Bicoid morphogen gradient. *PLoS Biol*, 5:e46, 2007.

[10] Benjamin P Berman, Yutaka Nibu, Barret D Pfeiffer, Pavel Tomancak, Susan E Celniker, Michael Levine, Gerald M Rubin, and Michael B Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. USA*, 99(2):757–62, January 2002.

[11] William Bialek and Sima Setayeshgar. Cooperativity, Sensitivity, and Noise in Biochemical Signaling. *Physical Review Letters*, 100(25):1–4, 2008.

[12] Frédéric Biemar, David a Nix, Jessica Piel, Brant Peterson, Matthew Ronshaugen, Victor Sementchenko, Ian Bell, J Robert Manak, and Michael Levine. Comprehensive identification of Drosophila dorsal-ventral patterning genes using a whole-genome tiling array. *Proc. Natl. Acad. Sci. USA*, 103(34):12763–8, August 2006.

[13] Ethan Bier, L Y Jan, and Y N Jan. rhomboid, a gene required for dorsoventral axis establishment and peripheral nervous system development in Drosophila melanogaster. *Genes & Development*, 4(2):190–203, 1990.

[14] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, Thomas Kuhlman, and Rob Phillips. Transcriptional regulation by the numbers: applications. *Current opinion in genetics & development*, 15(2):125–35, April 2005.

[15] Lacramioara Bintu, Nicolas E Buchler, Hernan G Garcia, Ulrich Gerland, Terence Hwa, Jané Kondev, and Rob Phillips. Transcriptional regulation by the numbers: models. *Current opinion in genetics & development*, 15(2):116–24, April 2005.

[16] Johannes Bischof, Robert K Maeda, Monika Hediger, François Karch, and Konrad Basler. An optimized transgenesis system for Drosophila

using germ-line-specific phiC31 integrases. *Proc. Natl. Acad. Sci. USA*, 104(9):3312–7, February 2007.

[17] Alistair N Boettiger and Michael Levine. Synchronous and stochastic patterns of gene activation in the Drosophila embryo. *Science*, 325(5939):471–3, July 2009.

[18] 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*, page 21, March 2011.

[19] S15. Bolouri, H, and E H Davidson. Transcriptional regulatory cascades in development: Initial rates, not steady state, determine network kinetics. *Proc. Natl. Acad. Sci. USA*, 100:9371–9376, 2003.

[20] J L Boulay, C Dennefeld, and A Alberga. The Drosophila developmental gene snail encodes a protein with nucleic acid binding fingers. *Nature*, 330(6146):395–8, 1987.

[21] Jessica Cande, Yury Goltsev, and Michael Levine. Conservation of enhancer location in divergent insects. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34):14414–9, August 2009.

[22] Kunitoshi Chiba, Junichi Yamamoto, Yuki Yamaguchi, and Hiroshi Handa. Promoter-proximal pausing and its release: molecular mechanisms and physiological functions. *Experimental cell research*, 316(17):2723–30, October 2010.

[23] Kwanghun Chung, Yoosik Kim, Jitendra S Kanodia, Emily Gong, Stanislav Y Shvartsman, and Hang Lu. A microfluidic array for large-scale ordering and orientation of embryos. *Nature methods*, 8(2):171–6, February 2011.

[24] Dorothy E Clyde, Maria S G Corado, Xuelin Wu, Adam Paré, Dmitri Papatsenko, and Stephen J Small. A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. *Nature*, 426(6968):849–53, December 2003.

[25] Mathieu Coppey, Alexander M Berezhkovskii, Yoosik Kim, Alistair N Boettiger, and Stanislav Y Shvartsman. Modeling the bicoid gradient: diffusion and reversible nuclear trapping of a stable protein. *Developmental biology*, 312(2):623–30, 2007.

[26] Mathieu Coppey, Alistair N Boettiger, Alexander M Berezhkovskii, and Stanislav Y Shvartsman. Nuclear trapping shapes the terminal gradient in the Drosophila embryo. *Current biology*, 18(12):915–9, 2008.

[27] Leighton J Core and John T Lis. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science*, 319(5871):1791–2, 2008.

[28] Antoine Coulon, Olivier Gandrillon, and Guillaume Beslon. On the spontaneous stochastic dynamics of a single gene: complexity of the molecular interplay at the promoter. *BMC systems biology*, 4:2, January 2010.

[29] A J Courey and S Jia. Transcriptional repression: the long and the short of it. *Genes & development*, 15(21):2786–96, November 2001.

[30] Xavier Darzacq, Yaron Shav-Tal, Valeria de Turris, Yehuda Brody, Shailesh M Shenoy, Robert D Phair, and Robert H Singer. In vivo dynamics of RNA polymerase II transcription. *Nature structural & molecular biology*, 14(9):796–806, September 2007.

[31] K A Datsenko and B L Wanner. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc. Natl. Acad. Sci. USA*, 97(12):6640–5, June 2000.

[32] Aitana Morton de Lachapelle and Sven Bergmann. Precision and scaling in morphogen gradient read-out. *Molecular systems biology*, 6(351):351, January 2010.

[33] Stefano De Renzis, Robert P Zinzen, J Yu, and Eric Wieschaus. Dorsal-ventral pattern of Delta trafficking is established by a Snail-Tom-Neuralized pathway. *Developmental cell*, 10(2):257–64, 2006.

[34] Tatjana Degenhardt, Katja N Rybakova, Aleksandra Tomaszewska, Martijn J Moné, Hans V Westerhoff, Frank J Bruggeman, and Carsten

Carlberg. Population-level transcription cycles derive from stochastic timing of single-cell transcription. *Cell*, 138(3):489–501, 2009.

[35] W Driever, G Thoma, and C Nüsslein-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, August 1989.

[36] William Feller. *An introduction to probability theory and its applications, Vol. I.* John Wiley & Sons Inc., New York, second edi edition, 1971.

[37] William Feller. *An introduction to probability theory and its applications, Vol. II.* John Wiley & Sons Inc., New York, second edi edition, 1971.

[38] P Fitzsimmons and Jim Pitman. Kac's moment formula and the FeynmanKac formula for additive functionals of a Markov process. *Stochastic Processes and their Applications*, 79(1):117–134, January 1999.

[39] V Francois, M Solloway, J W O'Neill, J 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(21):2602–2616, November 1994.

[40] L H Frank and C Rushlow. A group of genes required for maintenance of the amnioserosa tissue in Drosophila. *Development*, 122:1343–1352, 1996.

[41] Nicolás Frankel, Gregory K. Davis, Diego Vargas, Shu Wang, François Payre, and David L. Stern. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, pages 1–5, May 2010.

[42] Nicholas J Fuda, M Behfar Ardehali, and John T Lis. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461(7261):186–92, 2009.

[43] M Fujioka, Y Emi-Sarker, G L Yusibova, T Goto, and J B Jaynes. Analysis of an even-skipped rescue transgene reveals both composite and

discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. *Development*, 126(11):2527–38, June 1999.

[44] Eileen E M Furlong, E C Andersen, B Null, K P White, and M P Scott. Patterns of gene expression during Drosophila mesoderm development. *Science*, 293(5535):1629–33, August 2001.

[45] Daniel A Gilchrist, Gilberto Dos Santos, David C. Fargo, Bin Xie, Yuan Gao, Leping Li, and Karen Adelman. Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation. *Cell*, 143(4):540–551, November 2010.

[46] David S Gilmour. Promoter proximal pausing on genes in metazoans. *Chromosoma*, 118(1):1–10, February 2009.

[47] R E Goldstein, G Jiménez, O Cook, D Gur, and Z Paroush. Huckebein repressor activity in Drosophila terminal patterning is mediated by Groucho. *Development*, 126(17):3747–55, September 1999.

[48] Susan Gray and Michael Levine. Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in Drosophila. *Genes & Development*, 10:700–710, 1996.

[49] Thomas Gregor, David W Tank, Eric F Wieschaus, and William Bialek. Probing the limits to positional information. *Cell*, 130(1):153–64, July 2007.

[50] Thomas Gregor, Eric Wieschaus, William Bialek, De Ruyter Van Steveninck, R R, and D W Tank. Diffusion and scaling during early embryonic pattern formation. *Proc. Natl. Acad. Sci. USA*, 102:18403–18407, 2005.

[51] Thomas Gregor, Eric F Wieschaus, Alistair P McGregor, William Bialek, and David W Tank. Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell*, 130(1):141–52, July 2007.

[52] Amy C Groth, Matthew Fish, Roel Nusse, and Michele P Calos. Construction of transgenic Drosophila by using the site-specific integrase from phage phiC31. *Genetics*, 166(4):1775–82, April 2004.

[53] Tanja Gryzik and H-Arno J Müller. FGF8-like1 and FGF8-like2 encode putative ligands of the FGF receptor Htl and are required for mesoderm migration in the Drosophila gastrula. *Current Biology*, 14(8):659–67, April 2004.

[54] M G Guenther, S S Levine, L A Boyer, R Jaenisch, and R A Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130:77–88, 2007.

[55] T Häder, D Wainwright, T Shandala, R Saint, H Taubert, G Brönner, and H Jäckle. Receptor tyrosine kinase signaling regulates different modes of Groucho-dependent control of Dorsal. *Current biology*, 10(1):51–4, January 2000.

[56] Gordon L Hager, James G McNally, and Tom Misteli. Transcription dynamics. *Molecular cell*, 35(6):741–53, 2009.

[57] William W. Hager. Updating the Inverse of a Matrix. *SIAM Review*, 31(2):221, 1989.

[58] Diana C Hargreaves, Tiffany Horng, and Ruslan Medzhitov. Control of inducible gene expression by signal-dependent transcriptional elongation. *Cell*, 138(1):129–45, 2009.

[59] Feng He, Ying Wen, Jingyuan Deng, Xiaodong Lin, Long Jason Lu, Renjie Jiao, and Jun Ma. Probing intrinsic properties of a robust morphogen gradient in Drosophila. *Developmental cell*, 15(4):558–67, 2008.

[60] Kirugaval Hemavathy, Xiaodi Hu, Shovon I Ashraf, Stephen J Small, and Y Tony Ip. The repressor function of snail is required for Drosophila gastrulation and is not replaceable by Escargot or Worniu. *Developmental biology*, 269(2):411–20, May 2004.

[61] David a Hendrix, Joung-Woo Hong, Julia Zeitlinger, Daniel S Rokhsar, and Michael Levine. Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *Proc. Natl. Acad. Sci. USA*, 105(22):7762–7, June 2008.

[62] Michael Hoch, C Schröder, E Seifert, and H Jäckle. cis-acting control elements for Krüppel expression in the Drosophila embryo. *The EMBO journal*, 9(8):2587–95, August 1990.

[63] Joung-Woo Hong, David a Hendrix, and Michael Levine. Shadow enhancers as a source of evolutionary novelty. *Science*, 321(5894):1314, September 2008.

[64] Joung-Woo Hong, David a Hendrix, Dmitri Papatsenko, and Michael Levine. How the Dorsal gradient works: insights from postgenome technologies. *Proc. Natl. Acad. Sci. USA*, 105(51):20072–6, December 2008.

[65] R. A. Horn and C. R Johnson. *Matrix analysis.* Cambridge University Press, Cambridge, UK, corrected edition, 1990.

[66] A M Huang, Jannette Rusch, and Michael Levine. An anteroposterior Dorsal gradient in the Drosophilaembryo. *Genes & Development*, 11(15):1963–1973, 1997.

[67] Bo Huang, Sara A Jones, Boerries Brandenburg, and Xiaowei Zhuang. Whole-cell 3D STORM reveals interactions between cellular structures with nanometer-scale resolution. *Nature Methods*, 5(12):1047–52, December 2008.

[68] F Imam, D Sutherland, W Huang, and M A Krasnow. a Drosophila gene required for fibroblast growth factor (FGF)-directed migrations of tracheal and mesodermal cells. *Genetics*, 152:307–318, 1999.

[69] Y Tony Ip, Michael Levine, R E Park, David Kosman, and Ethan Bier. The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive neuroectoderm of the Drosophila embryo. *Genes & Development*, 6:1728–1739, 1992.

[70] Y Tony Ip, R E Park, David Kosman, K Yazdanbakhsh, and Michael Levine. dorsal-twist interactions establish snail expression in the presumptive mesoderm of the Drosophila embryo. *Genes & Development*, 6(8):1518–1530, August 1992.

[71] Johannes Jaeger, Svetlana Surkova, Maxim Blagov, Hilde Janssens, David Kosman, Konstantin N Kozlov, Manu, Ekaterina Myasnikova,

Carlos E Vanario-Alonso, Maria Samsonova, David H Sharp, and John Reinitz. Dynamic control of positional information in the early Drosophila embryo. *Nature*, 430(6997):368–71, July 2004.

[72] Sangyun Jeong, Mark Rebeiz, Peter Andolfatto, Thomas Werner, John True, and Sean B Carroll. The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. *Cell*, 132(5):783–93, March 2008.

[73] Jin Jiang, T Hoey, and Michael Levine. Autoregulation of a segmentation gene in Drosophila: combinatorial interaction of the even-skipped homeo box protein with a distal enhancer element. *Genes & Development*, 5(2):265–277, 1991.

[74] Jin Jiang, David Kosman, Michael Levine, and Y Tony Ip. The dorsal morphogen gradient regulates the mesoderm determinant twist in early Drosophila embryos. *Genes & Development*, 5(10):1881–1891, October 1991.

[75] Jin Jiang and Michael Levine. Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell*, 72(5):741–52, March 1993.

[76] Jin Jiang, Christine Rushlow, Stephen J Small, Qin Zhou, and Michael Levine. Individual dorsal morphogen binding sites mediate activation and repression in the Drosophila embryo. *EMBO*, 11(8):3147–3154, 1992.

[77] Tamar Juven-Gershon, Jer-Yuan Hsu, Joshua Wm Theisen, and James T Kadonaga. The RNA polymerase II core promoter - the gateway to transcription. *Current opinion in cell biology*, 20(3):253–9, 2008.

[78] Y Kasai, J R Nambu, P M Lieberman, and S T Crews. Dorsal-ventral patterning in Drosophila: DNA binding of snail protein to the single-minded gene. *Proc. Natl. Acad. Sci. USA*, 89(8):3414–8, April 1992.

[79] M Keaveney and K Struhl. Activator-mediated recruitment of the RNA polymerase II machinery is the predominant mechanism for transcriptional activation in yeast. *Molecular cell*, 1(6):917–24, May 1998.

[80] Hiroshi Kikuta, Mary Laplante, Pavla Navratilova, Anna Z Komisarczuk, Pär G Engström, David Fredman, Altuna Akalin, Mario Caccamo, Ian Sealy, Kerstin Howe, Julien Ghislain, Guillaume Pezeron, Philippe Mourrain, Staale Ellingsen, Andrew C Oates, Christine Thisse, Bernard Thisse, Isabelle Foucher, Birgit Adolf, Andrea Geling, Boris Lenhard, and Thomas S Becker. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome research*, 17(5):545–55, May 2007.

[81] Yoosik Kim, Mathieu Coppey, Rona Grossman, Leiore Ajuria, Gerardo Jiménez, Ze'ev Paroush, and Stanislav Y Shvartsman. MAPK substrate competition integrates patterning signals in the Drosophila embryo. *Current biology*, 20(5):446–51, March 2010.

[82] Roger D Kornberg. The molecular basis of eukaryotic transcription. *Proc. Natl. Acad. Sci. USA*, 104(32):12955–12961, 2007.

[83] David Kosman, Y Tony Ip, Michael Levine, and Kavita Arora. Establishment of the Mesoderm-Neuroectoderm Boundary in the Drosophila Embryo. *Science*, 254(5028):118–122, 1991.

[84] David Kosman, Claudia M Mizutani, Derek Lemons, W Gregory Cox, William McGinnis, and Ethan Bier. Multiplex detection of RNA expression in Drosophila embryos. *Science*, 305(5685):846, August 2004.

[85] R Kraut and Michael Levine. Mutually repressive interactions between the gap genes giant and Krüppel define middle body regions of the Drosophila embryo. *Development*, 111(2):611–21, February 1991.

[86] Chanhyo Lee, Xiaoyong Li, Aaron Hechmer, Michael Eisen, Mark D Biggin, Bryan J Venters, Cizhong Jiang, Jian Li, B Franklin Pugh, and David S Gilmour. NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila. *Molecular and cellular biology*, 28(10):3290–300, May 2008.

[87] E C Lee, D Yu, J Martinez de Velasco, L Tessarollo, D a Swing, D L Court, N a Jenkins, and N G Copeland. A highly efficient Escherichia coli-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. *Genomics*, 73(1):56–65, April 2001.

[88] H Lee, K W Kraus, M F Wolfner, and John T Lis. DNA sequence requirements for generating paused polymerase at the start of hsp70. *Genes & Development*, 6(2):284–295, February 1992.

[89] Maria Leptin. twist and snail as positive and negative regulators during Drosophila mesoderm development. *Genes & Development*, 5(9):1568–1576, September 1991.

[90] Maria Leptin and Markus Affolter. Drosophila gastrulation: identification of a missing link. *Current Biology*, 14(12):R480–2, 2004.

[91] Michael Levine and E H Davidson. Gene regulatory networks for development. *Proc. Natl. Acad. Sci. USA*, 102:4936–4942, 2005.

[92] Li M Li and David N Arnosti. Long- and Short-Range Transcriptional Repressors Induce Distinct Chromatin States on Repressed Genes. *Current biology*, 21:1–7, February 2011.

[93] Xiao-yong Li, Stewart MacArthur, Richard Bourgon, David Nix, Daniel a Pollard, Venky N Iyer, Aaron Hechmer, Lisa Simirenko, Mark Stapleton, Cris L Luengo Hendriks, Hou Cheng Chu, Nobuo Ogawa, William Inwood, Victor Sementchenko, Amy Beaton, Richard Weiszmann, Susan E Celniker, David W Knowles, Tom Gingeras, Terence P Speed, Michael B Eisen, and Mark D Biggin. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS biology*, 6(2):e27, February 2008.

[94] John T Lis. Imaging Drosophila gene activation and polymerase pausing in vivo. *Nature*, 450(7167):198–202, November 2007.

[95] Pentao Liu, Nancy a Jenkins, and Neal G Copeland. A highly efficient recombineering-based method for generating conditional knockout mutations. *Genome research*, 13(3):476–84, March 2003.

[96] Susan E Lott, Martin Kreitman, Arnar Palsson, Elena Alekseeva, and Michael Z Ludwig. Canalization of segmentation and its evolution in Drosophila. *Proc. Natl. Acad. Sci. USA*, 104(26):10926–31, June 2007.

[97] X Ma, D Yuan, K Diepold, T Scarborough, and J Ma. The Drosophila morphogenetic protein Bicoid binds DNA cooperatively. *Development*, 122(4):1195–206, April 1996.

[98] Hédia Maamar, Arjun Raj, and David Dubnau. Noise in gene expression determines cell fate in Bacillus subtilis. *Science*, 317(5837):526–9, July 2007.

[99] Paul M Macdonald and G Struhl. cis-acting sequences responsible for anterior localization of bicoid mRNA in Drosophila embryos. *Nature*, 336(6199):595–8, December 1988.

[100] S N Maity and B de Crombrugghe. Role of the CCAAT-binding protein CBF/NF-Y in transcription. *Trends in biochemical sciences*, 23(5):174–8, May 1998.

[101] Manu, Svetlana Surkova, Alexander V Spirov, Vitaly V Gursky, Hilde Janssens, Ah-Ram Kim, Ovidiu Radulescu, Carlos E Vanario-Alonso, David H Sharp, Maria Samsonova, and John Reinitz. Canalization of gene expression and domain shifts in the Drosophila blastoderm by dynamical attractors. *PLoS computational biology*, 5(3):e1000303, March 2009.

[102] Manu, Svetlana Surkova, Alexander V Spirov, Vitaly V Gursky, Hilde Janssens, Ah-ram Kim, Ovidiu Radulescu, Carlos E Vanario-Alonso, David H Sharp, Maria Samsonova, and John Reinitz. Canalization of gene expression in the Drosophila blastoderm by gap gene cross regulation. *PLoS biology*, 7(3):e1000049, March 2009.

[103] Thanasis Margaritis and Frank C P Holstege. Poised RNA polymerase II gives pause for thought. *Cell*, 133(4):581–4, 2008.

[104] Jonathan S Margolis, M L Borowsky, E Steingrímsson, 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, September 1995.

[105] Michele Markstein, Robert Zinzen, Peter Markstein, Ka-Ping Yee, Albert Erives, Angela Stathopoulos, and Michael Levine. A regulatory code for neurogenic gene expression in the Drosophila embryo. *Development*, 131(10):2387–94, May 2004.

[106] a M Michelson, S Gisselbrecht, E Buff, and James B Skeath. Heartbroken is a specific downstream mediator of FGF receptor signalling in Drosophila. *Development*, 125(22):4379–89, November 1998.

[107] Alan M Moses, Daniel a Pollard, David a Nix, Venky N Iyer, Xiao-Yong Li, Mark D Biggin, and Michael B Eisen. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS computational biology*, 2(10):e130, 2006.

[108] Ginger W Muse, Daniel A Gilchrist, S Nechaev, R Shah, J S Parker, S F Grissom, Julia Zeitlinger, and Karen Adelman. RNA polymerase is poised for activation across the genome. *Nat. Genet.*, 39(12):1507–1511, 2007.

[109] Yutaka Nibu. Interaction of Short-Range Repressors with Drosophila CtBP in the Embryo. *Science*, 280(5360):101–104, 1998.

[110] Amanda Ochoa-Espinosa, Danyang Yu, Aristotelis Tsirigos, Paolo Struffi, and Stephen Small. Anterior-posterior positional information in the absence of a strong Bicoid gradient. *Proceedings of the National Academy of Sciences of the United States of America*, 106(10):3823–8, March 2009.

[111] Amanda Ochoa-Espinosa, Gozde Yucel, Leah Kaplan, Adam Pare, Noel Pura, Adam Oberstein, Dmitri Papatsenko, and Stephen J Small. The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. *Proc. Natl. Acad. Sci. USA*, 102(14):4960–5, April 2005.

[112] M J Pankratz, M Busch, M Hoch, E Seifert, and H Jäckle. Spatial control of the gap gene knirps in the Drosophila embryo by posterior morphogen system. *Science (New York, N.Y.)*, 255(5047):986–9, February 1992.

[113] Dmitri Papatsenko. ClusterDraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors. *Bioinformatics*, 23(8):1032–4, April 2007.

[114] S M Parkhurst and V G Corces. Interactions among the gypsy transposable element and the yellow and the suppressor of hairy-wing loci in Drosophila melanogaster. *Molecular and cellular biology*, 6(1):47–53, January 1986.

[115] Jeffrey D Parvin and Richard A Young. Regulatory targets in the RNA polymerase II holoenzyme. *Current opinion in genetics & development*, 8:565–570, 1998.

[116] J Peccoud and B Ycart. Markovian Modeling of Gene-Product Synthesis. *Theoretical Population Biology*, 48(2):222–234, October 1995.

[117] Juan M Pedraza and Johan Paulsson. Effects of molecular memory and bursting on fluctuations in gene expression. *Science*, 319(5861):339–43, 2008.

[118] Juan M Pedraza and Alexander van Oudenaarden. Noise propagation in gene networks. *Science*, 307(5717):1965–9, 2005.

[119] Michael W Perry, Alistair N Boettiger, Jacques P Bothma, and Michael Levine. Shadow enhancers foster robustness of Drosophila gastrulation. *Current Biology*, 20(17):1562–7, September 2010.

[120] Michael W Perry, J D Cande, a N Boettiger, and Michael Levine. Evolution of Insect Dorsoventral Patterning Mechanisms. *Cold Spring Harbor symposia on quantitative biology*, pages 275–279, October 2009.

[121] Michael W Perry, Michael Levine, and Alistair N Boettiger. Multiple Enhancers Ensure Precision of Gap Gene Expression Patterns in the Drosophila Embryo. *in review in Dev Cell*, 2011.

[122] M Ptashne and A Gann. Transcriptional activation by recruitment. *Nature*, 386, 1997.

[123] Arjun Raj, Charles S Peskin, Daniel Tranchina, Diana Y Vargas, and Sanjay Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):e309, October 2006.

[124] Arjun Raj, Scott a Rifkin, Erik Andersen, and Alexander van Oudenaarden. Variability in gene expression underlies incomplete penetrance. *Nature*, 463(7283):913–8, February 2010.

[125] Arjun Raj, Patrick van Den Bogaard, Scott Rifkin, Alexander van Oudenaarden, and Sanjay Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, 5(10):877–879, 2008.

[126] Arjun Raj and Alexander van Oudenaarden. Single-molecule approaches to stochastic gene expression. *Annual review of biophysics*, 38:255–70, 2009.

[127] Tiina Rajala, Antti Häkkinen, Shannon Healy, Olli Yli-Harja, and Andre S Ribeiro. Effects of transcriptional pausing on gene expression dynamics. *PLoS computational biology*, 6(3):e1000704, January 2010.

[128] P Ramain, P Heitzler, M Haenlin, and P Simpson. pannier, a negative regulator of achaete and scute in Drosophila, encodes a zinc finger protein with homology to the vertebrate transcription factor GATA-1. *Development*, 119(4):1277–91, December 1993.

[129] Jonathan M Raser and Erin K O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–4, 2004.

[130] Jonathan M Raser and Erin K O'Shea. Noise in gene expression: origins, consequences, and control. *Science*, 309(5743):2010–3, 2005.

[131] Eric B Rasmussen and John T Lis. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc. Natl. Acad. Sci. USA*, 90(September):7923–7927, 1993.

[132] Haluk Resat, Linda Petzold, and Michel F Pettigrew. Kinetic modeling of biological systems. *Methods in molecular biology (Clifton, N.J.)*, 541:311–35, January 2009.

[133] Andre S Ribeiro. Stochastic and delayed stochastic models of gene expression and regulation. *Mathematical biosciences*, 223(1):1–11, January 2010.

[134] Andre S Ribeiro, Olli-Pekka Smolander, Tiina Rajala, Antti Häkkinen, and Olli Yli-Harja. Delayed stochastic model of transcription at the single nucleotide level. *Journal of computational biology : a journal of computational molecular cell biology*, 16(4):539–53, April 2009.

[135] Sara Ricardo and Ruth Lehmann. An ABC transporter controls export of a Drosophila germ cell attractant. *Science*, 323(5916):943–6, 2009.

[136] R Rivera-Pomar, X Lu, N Perrimon, H Taubert, and H Jäckle. Activation of posterior gap gene expression in the Drosophila blastoderm. *Nature*, 376(6537):253–6, July 1995.

[137] Jannette Rusch and Michael Levine. Threshold responses to the dorsal regulatory gradient and the subdivision of primary tissue territories in the Drosophila embryo. *Current opinion in genetics & development*, 6:416–423, 1996.

[138] S L Rutherford and S Lindquist. Hsp90 as a capacitor for morphological evolution. *Nature*, 396(6709):336–42, November 1998.

[139] Michael S Samoilov and Adam P Arkin. Deviant effects in molecular reaction pathways. *Nature biotechnology*, 24(10):1235–40, 2006.

[140] Thomas Sandmann, Charles Girardot, Marc Brehme, Waraporn Tongprasit, Viktor Stolc, and Eileen E M Furlong. A core transcriptional network for early mesoderm development in Drosophila melanogaster. *Genes & development*, 21(4):436–49, February 2007.

[141] D S Schneider, K L Hudson, T Y Lin, and K V Anderson. Dominant and recessive mutations define functional domains of Toll, a transmembrane protein required for dorsal-ventral polarity in the Drosophila embryo. *Genes & Development*, 5(5):797–807, May 1991.

[142] Mark D Schroeder, Michael Pearce, John Fak, HongQing Fan, Ulrich Unnerstall, Eldon Emberly, Nikolaus Rajewsky, Eric D Siggia, and Ulrike Gaul. Transcriptional control in the segmentation gene network of Drosophila. *PLoS biology*, 2(9):E271, September 2004.

[143] Robert J Sims, Rimma Belotserkovskaya, and Danny Reinberg. Elongation by RNA polymerase II: the short and long of it. *Genes & development*, 18(20):2437–68, 2004.

[144] James B Skeath, G F Panganiban, and S B Carroll. The ventral nervous system defective gene controls proneural gene expression at two distinct steps during neuroblast formation. *in Drosophila. Development*, 120:1517–1524, 1994.

[145] Stephen J Small, A Blair, and Michael Levine. Regulation of even-skipped stripe 2 in the Drosophila embryo. *The EMBO journal*, 11(11):4047–57, November 1992.

[146] Stephen J Small, A Blair, and Michael Levine. Regulation of two pair-rule stripes by a single enhancer in the Drosophila embryo. *Developmental biology*, 175(2):314–24, May 1996.

[147] Valeria Specchia, Lucia Piacentini, Patrizia Tritto, Laura Fanti, Rosalba D'Alessandro, Gioacchino Palumbo, Sergio Pimpinelli, and Maria P Bozzetti. Hsp90 prevents phenotypic variation by suppressing the mutagenic activity of transposons. *Nature*, 463(7281):662–5, February 2010.

[148] Rebekka O Sprouse, Tatiana S Karpova, Florian Mueller, Arindam Dasgupta, James G McNally, and David T Auble. Regulation of TATA-binding protein dynamics in living yeast cells. *Proc. Natl. Acad. Sci. USA*, 105(36):13304–8, September 2008.

[149] Angelike Stathopoulos and Michael 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, July 2002.

[150] Angelike Stathopoulos and Michael Levine. Localized repressors delineate the neurogenic ectoderm in the early Drosophila embryo. *Developmental biology*, 280(2):482–93, April 2005.

[151] Angelike Stathopoulos, Bergin Tam, Matthew Ronshaugen, Manfred Frasch, and Michael Levine. pyramus and thisbe: FGF genes that pattern the mesoderm of Drosophila embryos. *Genes & development*, 18(6):687–99, 2004.

[152] Angelike Stathopoulos, Madeleine Van Drenth, Albert Erives, Michele Markstein, and Michael Levine. Whole-Genome Analysis of Dorsal-Ventral Patterning in the Drosophila Embryo. *Cell*, 111:687–701, 2002.

[153] G Struhl. Differing strategies for organizing anterior and posterior body pattern in Drosophila embryos. *Nature*, 338(6218):741–4, April 1989.

[154] G Struhl, P Johnston, and P a Lawrence. Control of Drosophila body pattern by the hunchback morphogen gradient. *Cell*, 69(2):237–249, April 1992.

[155] Mukund Thattai and Alexander van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci. USA*, 98(15):8614–9, July 2001.

[156] Mukund Thattai and Alexander van Oudenaarden. Attenuation of noise in ultrasensitive signaling cascades. *Biophysical journal*, 82(6):2943–50, June 2002.

[157] Mary C Thomas and Cheng-Ming Chiang. The general transcription machinery and general cofactors. *Critical reviews in biochemistry and molecular biology*, 41(3):105–78, 2006.

[158] Gasper Tkacik, Thomas Gregor, and William Bialek. The role of input noise in transcriptional regulation. *PloS one*, 3(7):e2774, 2008.

[159] Susanna Valanne, Jing-Huan Wang, and Mika Rämet. The Drosophila toll signaling pathway. *Journal of immunology (Baltimore, Md. : 1950)*, 186(2):649–56, January 2011.

[160] Jeroen S van Zon, Marco J Morelli, Sorin Tnase-Nicola, and Pieter Rein ten Wolde. Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophysical journal*, 91(12):4350–67, 2006.

[161] Koen J T Venken, Joseph W Carlson, Karen L Schulze, Hongling Pan, Yuchun He, Rebecca Spokony, Kenneth H Wan, Maxim Koriabine, Pieter J de Jong, Kevin P White, Hugo J Bellen, and Roger A Hoskins. Versatile P[acman] BAC libraries for transgenesis studies in Drosophila melanogaster. *Nature methods*, 6(6):431–4, June 2009.

[162] Koen J T Venken, Yuchun He, Roger a Hoskins, and Hugo J Bellen. P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in D. melanogaster. *Science*, 314(5806):1747–51, December 2006.

[163] P H von Hippel and O G Berg. Facilitated target location in biological systems. *The Journal of biological chemistry*, 264(2):675–8, January 1989.

[164] C H Waddington. Canalization of development and genetic assimilation of acquired characters. *Nature*, 183(4676):1654–5, June 1959.

[165] Xiaoling Wang, Chanhyo Lee, David S Gilmour, and J Peter Gergen. Transcription elongation controls cell fate specification in the Drosophila embryo. *Genes & development*, 21(9):1031–6, 2007.

[166] Yu-chiun Wang and Edwin L Ferguson. Spatial bistability of Dpp-receptor interactions during Drosophila dorsal-ventral patterning. *Nature*, 434(7030):229–34, March 2005.

[167] Søren Warming, Nina Costantino, Donald L Court, Nancy a Jenkins, and Neal G Copeland. Simple and highly efficient BAC recombineering using galK selection. *Nucleic acids research*, 33(4):e36, January 2005.

[168] Jie Yao, M Behfar Ardehali, Christopher J Fecko, Watt W Webb, and John T Lis. Intranuclear distribution and local dynamics of RNA polymerase II during transcription activation. *Molecular cell*, 28(6):978–90, 2007.

[169] Danyang Yu and Stephen Small. Precise registration of gene expression boundaries by a repressive morphogen in Drosophila. *Current biology*, 18(12):868–76, June 2008.

[170] Julia Zeitlinger, Alexander Stark, Manolis Kellis, Joung-Woo Hong, Sergei Nechaev, Karen Adelman, Michael Levine, and Richard a Young. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nature genetics*, 39(12):1512–6, 2007.

[171] Julia Zeitlinger, Robert P Zinzen, Alexander Stark, Manolis Kellis, Hailan Zhang, Richard A Young, and Michael Levine. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. *Genes & Development*, 21(4):385–90, 2007.

[172] Min Yan Zhu, Robert Wilson, and Maria Leptin. A screen for genes that influence fibroblast growth factor signal transduction in Drosophila. *Genetics*, 170(2):767–77, 2005.

[173] Robert P Zinzen and Dmitri Papatsenko. Enhancer responses to similarly distributed antagonistic gradients in development. *PLoS computational biology*, 3(5):e84, 2007.

[174] Robert P Zinzen, Kate Senger, Michael Levine, and Dmitri Papatsenko. Computational models for neurogenic gene expression in the Drosophila embryo. *Current Biology*, 16(13):1358–65, 2006.