THE UNIVERSITY OF CHICAGO

DECODING *CIS*-REGULATION OF THE *DROSOPHILA EVEN-SKIPPED* LOCUS WITH PHYSICAL-CHEMICAL MODELS AND SYNTHETIC ENHANCERS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS AND SYSTEMS BIOLOGY

BY
KENNETH A. BARR

CHICAGO, ILLINOIS
AUGUST 2017

"The only way of discovering the limits of the possible is to venture a little way past them into the impossible."

-Arthur C. Clarke

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my adviser John Reinitz. Before graduate school I already had an interest in the problem of gene regulation, but at the time the problem seemed far too complex to be comprehensible. I was particularly drawn to John's lab for his creative approach. From John I have learned how to make sense of complex systems without over simplification and under John's tutelage I have developed confidence as a both a scientist and programmer. I have greatly enjoyed our conversations, about science as well as many diverse shared interests.

I want to thank my committee, Martin Kreitman, Ilaria Rebay, and Kevin White for the experience, advice, and support they have given me. I especially want to thank my chair, Martin Kreitman. Our joint 'Kreitnitz' lab space has been a warm environment that has broadened my scientific perspective.

The members of the Reinitz lab, both present and past, have had a lasting positive impact. I want to thank AhRam Kim, whose intellectual legacy greatly influenced my project. My time sitting next to Carlos Martinez were some of the most enjoyable in graduate school. He helped me develop as a programmer and our extended conversations helped me develop my own scientific perspective. I want to thank Zhihao Lou for his help and expertise in global optimization. I also thank Jennifer Moran and Unjin Lee for their assistance with cloning the many constructs used in this work. I would like to thank collaborators, Ovidiu Radulescu and Alexandre Ramos. I want to thank Pengyao Ziang, who rotated the same quarter and defended her own thesis only two weeks before me. It has been great going through this process together.

Finally, I want to thank all my friends and family who have supported me in this process. I want to thank my parents who have encouraged my intellectual pursuits from a very early age. I want to thank my graduate class for their friendship and support. They have made Hyde Park a home all these years. Most importantly, I want to thank Katie Igartua for her love and support. She gave me much needed confidence when experiments did not go

as expected, listened to my concerns and complaints, and provided me with much needed encouragement that got me through the hardest times.

Too many people have helped me through these years to name, so to all the colleagues, friends and family not named, thank you for your assistance and the support you have given me throughout.

ABSTRACT

Within the non-coding portions of the genome lie sets of instructions that specify when, where, and how much genes should be expressed. This regulatory function of DNA is of paramount importance in the understanding of development, evolution and disease. These instructions take the form of docking sites for proteins known as transcription factors that bind to control regions known as enhancers. At every level of this process there are outstanding questions, from the logic of multiple transcription factors acting on a single enhancer, to the action of multiple enhancers interacting in a complex genetic locus. This dissertation presents theoretical and experimental work that connects the underlying structure of transcription factor binding sites to the complex regulation of the intact *Drosophila melanogaster* *even-skipped* locus, which is expressed in seven transverse stripes across developing embryos. Studies of this locus can take full advantage of previous work that has established a quantitative, single nucleus resolution spatiotemporal atlas of transcription factors levels and gene expression. Models based on physical-chemical interactions between transcription factors and DNA have been developed to explain how enhancers interpret these transcription factor levels in order to specify expression in the appropriate places and times during development. In this dissertation, I present new efficient algorithms and code that allow this type of modeling to be applied at scales that were previously impossible. I show that the enhancer structure of *even-skipped* arises out of competition between DNA fragments for interaction with the basal transcription machinery. This enhancer competition allows the function of a complex regulatory locus to be connected directly to DNA sequence. In order to test the extent to which the molecular interactions that are implemented in this model fully explain the activity of enhancers, I generate a panel of synthetic enhancers, designed *ab initio* to drive a single stripe across developing embryos. Analysis of these sequences reveals that the molecular interactions that determine the activity of enhancers have complex dependencies on the position and orientation of sites. Finally, I show that there is a relationship between the length of enhancers and their robustness to perturbation in both transcription factor

concentration and genetic mutation.

Three files are included this dissertation, including the code that implements the transcription model, the measured expression of *even-skipped* that the model is trained on, and the parameters of all fits to data.

CHAPTER 1

INTRODUCTION

1.1 The genetic regulatory code.

Discovery of the genetic regulatory code. In 1958, Francis Crick proposed the central dogma of molecular biology, that the flow of genetic information proceeds from nucleic acid to protein, but not in reverse[33]. In other words, DNA sequence is transcribed into RNA message, which is then translated into a protein. From this perspective, DNA is a complete blueprint for life, and the only information necessary to understand all life's complexities is the sequence of DNA and the code through which it is read. The discovery of the codon—a simple three nucleotide sequence which specifies one of 20 amino acids—was the first step in cracking this genetic code[137]. But even in the same year that Nirenberg identified the first codon, Jacob and Manod described an *E. coli* gene (a molecular unit of heredity) that did not code for protein but instead controlled, on the same DNA strand, the transcription of 'structural' genes (sequences that directly specify the amino acid sequence of proteins)[80]. They called this gene an operator. Later, this regulatory unit was determined to be controlled by proteins[55], known as transcription factors (TFs), binding to specific DNA sequences[56] immediately upstream of the transcription start site (TSS). Broadly, regulatory DNA sequences have become known as *cis*-regulatory elements (CREs) or modules (CRMs). While the genetic code for amino acid sequence was completed soon after the discovery of the first codon [136, 88], decades later there is still no general framework for reading the regulatory code of CREs.

The regulatory genetic code in the post-genomic era. Modern technology has highlighted the importance of cracking the regulatory code. It is now possible to computationally identify structural genes from sequenced genomes, which has led to the discovery of 20,000 such genes in humans, covering approximately 2% of the genome [5]. These protein

coding sequences are vastly outnumbered by hundreds of thousands of putative CREs[32]. These regions account for as much as 90% of of all trait-associated genetic variation in humans[72, 125], and many are under selection[201], making their study of paramount importance in the study of disease and evolution. Despite the ability to characterize protein-DNA interactions rapidly and on a genome wide scale[85], there is currently no available method to accurately predict the effects of functional variation on gene expression.

1.2 Enhancers

The operator paradigm, as described by Jacob and Manod, was not found in eukaryotes. Instead, the first breakthrough in metazoan gene regulation came decades later with the discovery of a sequence upstream of the simian virus 40 (SV40) which greatly enhanced expression of the β-globin gene, from upstream or downstream, in forward or reverse orientation, and at distances greater than 3 kb[11]. For these reasons, such sequences became known as enhancers. Enhancers were also found to direct tissue-specific expression of the immunoglobulin heavy chain[10, 57]. These properties—tissue specific expression and independence of position and orientation—have become the defining features of enhancers, which are now known to be a general feature of metazoan gene regulation[104, 108].

1.3 The *Drosophila* developmental system for the study of enhancers.

1.3.1 Drosophila *embryogenesis*

Drosophila embryogenesis is a rapid process, taking only 22 to 24 hours at 25 to generate a larva from an egg. The first three hours of development are characterized by rapid nuclear divisions in the absence of cytokinesis. These divisions define cleavage cycles, representing the time between the end of last mitotic division and the end of the next, such that the nuclear

cycle n is the time between mitosis $n-1$ and n. The first nine nuclear cycles take about eight minutes each and occur near the center of the egg. After the ninth nuclear division, the majority of nuclei migrate to the periphery, forming a hollow ellipsoid shell of nuclei called the syncytial blastoderm. These early nuclear cycles are characterized by minimal transcription occurring from zygotic nuclei. The next four nuclear cycles take progressively longer[54].

During the 14^{th} nuclear cycle (NC14), cellularization begins. Nuclei elongate along their basal-apical axis and cell membranes invaginate between nuclei. It takes approximately 45 minutes to complete cellularization, after which the embryo undergoes three dimensional restructuring, called gastrulation, to form the presumptive mesoderm, endoderm and ectoderm. This phase of NC14, between the 13^{th} nuclear division and the onset of gastrulation, is called cycle 14A (NC14A). During this cycle, maternally deposited mRNA and protein are degraded and zygotic transcription increases[44]. This process is referred to as the mid-blastula transition (MBT) or alternatively the maternal to zygotic transition (MZT).

Drosophila embryos have emerged as a powerful system for the study of enhancers. Within the syncytial blastoderm are TFs that collectively specify future body segments[4, 76]. Because each nucleus responds to different collective inputs, it is analogous to a single tissue. This property is important in the study of tissue-specific enhancers, where this spatial separation in *Drosophila* embryos allows for the function of enhancers to be observed across multiple tissues within a single organism. In the sections that follow, I describe the properties of gene regulation in this model system, and the regulatory mechanisms that have been discovered.

1.3.2 Maternal, gap, and pair-rule genes

The studies presented in this work will focus primarily on specification along the anterior-posterior (AP) axis of developing embryos. The molecular players in this process were largely defined by their mutant effects, where 'gap' gene mutants cause large gaps between body seg-

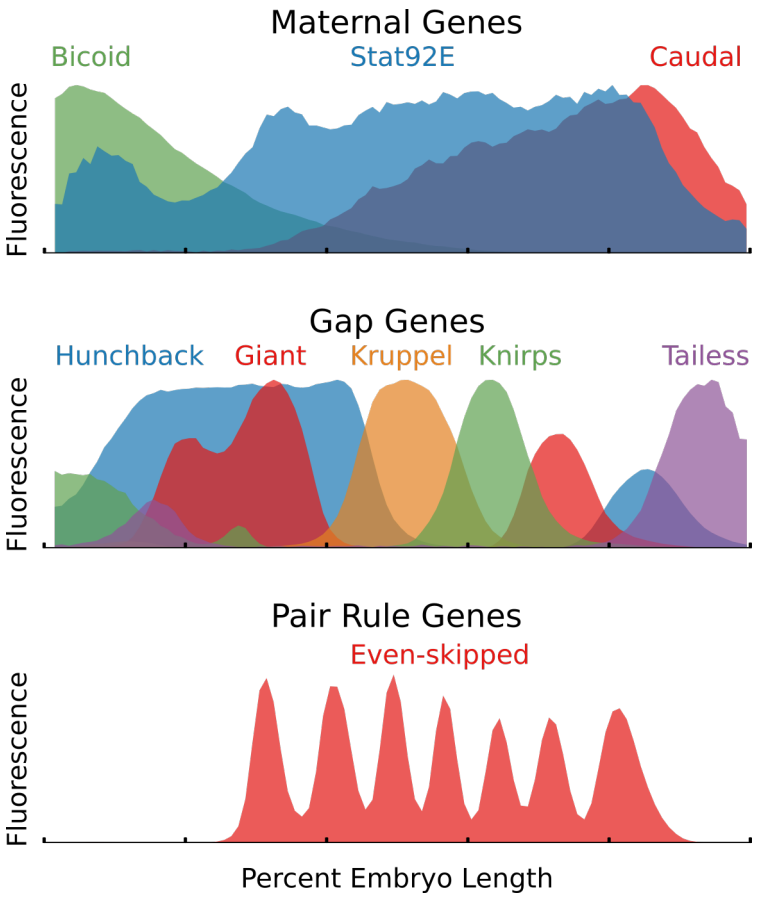


Figure 1.1. Maternal, gap and pair-rule expression. The expression levels of maternal (top), gap (middle) and the pair-rule gene *even-skipped* (bottom) along the anterior posterior axis. Expression is given by average fluorescence along a a 10% wide dorsal-ventral stripe from embryos stained by *in situ* hybridization for the indicated factor.

ments, and 'pair-rule' mutants cause every other segment to be missed[140]. Cell type specification is initiated by the protein Bicoid (Bcd), whose mRNA is deposited maternally[16] in what it defines to be the anterior of the embryo. This results in an exponentially declining gradient of Bcd protein concentration[38, 62]. Multiple spatial gradients form from ubiquitous maternally deposited factors[29], such as Caudal (Cad)[128, 115, 39] and Hunchback (Hb)[190, 189, 75, 78] (Fig 1.1, top). Together, these maternal factors set up the spatial domains of gap genes including Giant (Gt)[52, 151, 130], Kruppel (Kr)[202], and Knirps (Kni)[102] (Fig 1.1, middle). Finally, maternal and gap genes together specify the expression pattern of pair rule genes (Fig 1.1, bottom). Among these, the genes, *even-skipped* (*eve*) and *runt* (*run*) respond only to the spatial information present in maternal and gap genes, while other pair rule genes require cross regulation[76, 160].

1.3.3 Gap gene regulation of eve

By the end of NC14A, *eve* is expressed in seven transverse stripes across developing embryos[47, 48]. The fact that that only gap and maternal genes are responsible for the regulation of *eve* imposes logic on *eve* regulation. Gap genes express in broad domains, approximately 10-15 nuclei wide, yet determine the expression of narrow *eve* stripes that are approximately 3 nuclei wide. This fact eliminates the possibility that *eve* stripes are formed by silencing from gap genes, because this mechanism would silence *eve* in domains as wide as the gap gene domains. Similarly, it was unclear how activation could lead to narrow stripes. There was some suggestion that *eve* forms boundaries of gap gene domains, yet there were insufficient borders to explain the existence of all seven *eve* stripes[48].

1.3.4 Enhancers of eve

A potential solution was afforded by the discovery that *eve* is regulated by enhancers that respond independently to TFs[59, 1]. These elements were discovered through reporter assays, in which a DNA segment is placed upstream of a detectable and physiologically

inert marker. When the proximal 7.3 kb of *eve* was placed upstream of the reporter *lac*Z, which was subsequently detected with *in-situ* hybridization, the sequence drove expression overlapping *eve* stripes 2, 3 and 7 early in NC14A, and all seven stripes immediately preceding gastrulation[59]. Various deletions in this fragment identified three enhancers at 1 kb, 2.9 kb, and 4.65 kb upstream of the TSS that were necessary for stripe 2, stripe 3, and late seven stripe pattern, respectively. No deficiency led to the loss of stripe 7[59]. Expression driven by the seven stripe element was disrupted in embryos that were homozygous for null mutations in *run*, *h* or *eve*, but expression driven by the proximal 4.6 kb (containing the sequences responsible for early expression of stripes 2, 3 and 7) was unperturbed[59]. This indicated that separate *cis*-regulatory elements of *eve* could respond independently to TFs.

1.3.5 Regulation of a single eve *stripe*

Studies of *eve* expression in embryos homozygous for null mutations in gap genes suggested that *eve* stripe 2 is regulated by Hb, Kr and Gt. Stripe 2 was missing in Hb- embryos, indicating that Hb positively regultates stripe 2[48]. Similarly, stripe 2 was fused with stripe 1 or stripe 3 in Gt- or Kr- embryos respectively, indicating that these borders were formed by repression[48]. These results were ambigous due to cross-regulation of gap genes, which leads to shifts in the positions the remaining gap genes in gap gene mutant embryos[79]. Several experiments established a direct, causal link between these factors and the expression of stripe 2. First, DNAse footprinting identified direct binding of Gt, Kr, Hb and Bcd to the stripe 2 enhancer[177]. Second, cotransfection of stripe 2 sequences driving chloramphenicol acetytransferase (CAT), together with gap gene expression plasmids, in Schneider cells showed that this sequence responds positively and synergistically to Bcd and Hb, and negatively to Gt and Kr[177]. Finally, mutations to Gt and Kr binding sites within the stripe 2 enhancer caused anterior and posterior shifts in the border of stripe 2 respectively, indicating that the effect is due to direct interactions with DNA[180].

1.3.6 *The* eve *stripe 2 enhancer*

Collectively, these experiments defined the location and function of the stripe 2 enhancer. This sequence is is broadly activated by Bcd and Hb, and locally repressed by Gt and Kr. The 480 bp DNA fragment, between 1 kb and 1.5 kb upstream of the TSS, was both necessary and sufficient to drive the stripe 2 pattern of *eve* and was the smallest fragment that was able to do so[175]. For that reason this 480 bp fragment has become known as the minimal stripe 2 element (MSE2), though some work has focused on a longer sequence that contains several additional binding sites for key regulators and is known as the stripe 2 enhancer (S2E)[177, 113]. Initially, the function of MSE2 was defined by 12 footprinted binding sites for Bcd, Hb, Gt and Kr, though other TF binding sites have been since been identified in MSE2. For instance, Sloppy Paired 1 (Slp1) binds to MSE2 and is required to prevent ectopic expression of *eve* in the anterior[7].

1.3.7 *Short-ranged quenching permits enhancer autonomy*

The fact that the stripe 3 enhancer drives expression of *eve* at the peak of Kr binding to MSE2 shows that the action of Kr must be spatially restricted to MSE2. Repression through steric competition for binding to DNA could explain the short range action of repression, however other Drosophila TFs were found to repress at distances up to about 150 bp from the nearest transcriptional activator[61]. This distance is far enough to exclude the possibility of competition for binding. This short distance repression has been termed 'quenching', and its limited range explains how repressors may independently switch individual enhancers on and off.

1.3.8 *Regulation of* eve *stripes 3 and 7 by a single enhancer*

Reporter analysis narrowed down the DNA segment responsible for expression of *eve* stripe 3 to a 500 bp sequence, located between 3.8 kb and 3.3 kb upstream of the TSS[176]. Because

smaller fragments did not drive expression this has become known as the minimal stripe 3 element (MSE3). MSE3 also drives expression of *eve* stripe 7, and this function could not be separated, indicating that a single enhancer sequence drives both stripes. Components of the JAK-Stat signaling pathway were found to be required for optimal stripe 3 expression[20] and the broadly expressed Stat92E (Dst) was found to bind directly to MSE3[204]. This indicated that like MSE2, MSE3 is broadly activated. MSE3 drove expression of a fused stripe in Kni- embryos and the anterior border of stripe 3 expanded to the anterior in Hb- mutants[176]. This demonstrated that stripes 3 and 7 are formed through local repression by Hb and Kni.

1.3.9 Dual regulation by Hunchback

While MSE2 is positively regulated by Hb, stripe 3 is negatively regulated. Even more puzzling is the fact that stripe 7 is expressed within the posterior expression domain of Hb. This indicates that a single enhancer is able to respond differently to Hb within the anterior and posterior domains of the embryo. The phenomenon by which Hb switches between activating and repressing states is known as 'coactivation.' And the effect may be mediated by proximity to Bcd or Cad[93].

1.3.10 Bicoid binds DNA cooperatively

Reporter assays, in both *Drosophila* and yeast, with constructs that contain different arrangements of Bcd binding sites showed that spacing of sites was critical for activation driven by Bcd[66]. *In vitro* binding assays with *Drosophila* enhancers showed cooperativity of Bcd binding, offering a potential explaination for why enhancers might be sensitive to spacing between binding sites. Mutations in Bcd that disrupted the cooperativity of Bcd led to defects in defects in the expression of gap genes and *eve*[25, 101].

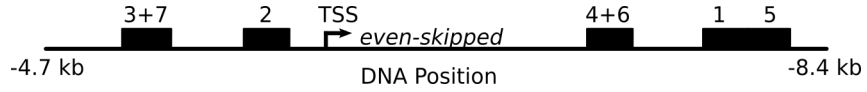


Figure 1.2. *eve* locus diagram. Five enhancers control the seven striped pattern of *eve*. The location of each enhancer is indicated by black rectangles and each is labeled according to the *eve* stripes driven by that sequence. The transcription start site (TSS) is indicated with an arrow.

1.3.11 Enhancers of eve *stripes 1, 4, 5 and 6*

The remaining enhancers, which drive stripes 1, 4, 5 and 6 were identified downstream of the coding region[50]. Stripes 4 and 6 are regulated by a single element, while stripes 1 and 5 are regulated by separable but adjacent elements. Collectively, *eve* expression is regulated by a 16 kb chromatin domain, with the early seven stripe pattern controlled by 5 enhancers that lie within a fragment spanning from 4.7 kb upstream to 8.4 kb downstream of the *eve* TSS[163](Fig 1.2).

1.3.12 Chromatin state.

Proper expression driven by *eve* enhancers also requires a chromatin state that is permissible to binding. Factors that establish permissible chromatin states in *Drosophila* embryogenesis were first indicated by the discovering of the TAGteam motif, which was found to control the timing of expression for developmental enhancers[21]. Zelda (Zld, previously called Vfl[181]) binds to the TAGteam motif and is essential for early activation of the zygotic genome[107]. Genomic analysis revealed that presence of Zld binding sites was more predictive of binding for other TFs than their motifs themselves[166], which indicated that Zld acts to establish a chromatin state that permits the binding of other TFs. This chromatin modifying role of Zld has been confirmed[169]. Zld binding has been identified on *eve* enhancers, where it is required for activation[183].

1.3.13 Multiple enhancers in a complex locus.

The observation that *eve* stripes 2, 3, and 7 were separable into stripes 2 and 3 + 7 led to the assumption that multiple enhancers have independent and additive effects[61]. The discovery of functionally redundant enhancers, also known as shadow enhancers, puts this assumption into question[149, 150, 46]. Shadow enhancers are a common feature of development[27] and are non-additive in their function [40, 24]. New regulatory mechanisms will be necessary to properly model the behavior of multiple enhancers within a complex locus.

1.4 Principles of enhancer organization

1.4.1 Enhancer logic

These experiments collectively defined a certain logic to the way arrays of transcription factor binding sites give rise to expression domains. *Drosophila* developmental enhancers contain binding sites for ubiquitously or broadly expressed activators, and achieve restricted expression through the localized repression. This repression requirs specific spatial arrangements of binding sites in order to achieve repression through binding site competition or quenching. Some TFs switch between activating and repressing roles, depending on the context, and others bind to DNA cooperatively. These mechanisms which set constraints on the number, order, spacing, and orientation of TF binding sites that leads to functional enhancers and forms part of what is known as the *cis*-regulatory logic of enhancers.

1.4.2 Types of cis*-regulatory logic*

Two prominent types of enhancer logic have been proposed: the enhanceosome and the billboard. The former refers to elements with high degrees of cooperative binding and high levels of constraint with respect to the arrangement of binding sites. The prototypical enhanceosome is the interferon β enhancer. This element is only 57 bp in length and is so highly structured that it has been crystallized[145]. Even small sequence insertions in this enhancer

led to near complete loss of enhancer activity[192]. In contrast, Kulkarni and Arnosti have proposed that *Drosophila* developmental enhancers function as an information display or 'billboard'[98]. Such enhancers are composed of multiple sub-elements that are independently read by the transcription machinery and could have varying degrees of constraint on the arrangement of binding sites.

1.4.3 Flexibility of eve *enhancers.*

Experiments with *eve* enhancers from different species suggests considerable flexibility in *cis*-regulatory logic. S2Es have been identified in multiple species[110, 113, 109, 68, 67]. These sequences drive conserved expression, yet lack sequence conservation. Combined with the fact that there must be accessible evolutionary paths between these sequences, this suggests that there are many permissible arrangements of binding sites that can drive stripe 2 expression. However, when 5' and 3' halves of enhancers from two separate species were combined to form chimeric enhancers, expression was disrupted[112]. This indicates that despite the flexibility in site arrangements over evolution, the evolved elements are structured and contain local molecular interactions that are disrupted in chimeric enhancers.

1.5 Quantitative analysis of enhancers.

Considering the vast number of TFs, binding sites, and distance-dependent mechanisms acting on an enhancer, the challenge is to determine which configurations of TF binding sites drive expression, and which do not. With multiple activating and repressing mechanisms simultaneously acting on the same sequence, it is hard to know which effect will prevail. Moreover, TFs act on enhancers in a dose dependent fashion. In such circumstances, quantitative data and methods are necessary. To this end, several groups have collected spatially and temporally resolved quantitative data on levels of TFs and mRNA during early *Drosophila* development[185, 184, 114, 45] and have also developed quantitative methods to integrate

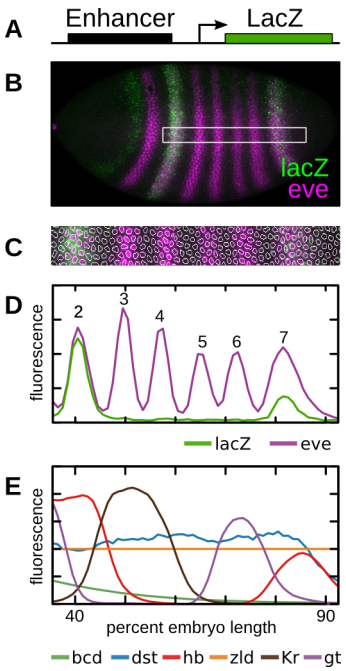


Figure 1.3. Enhancer quantification. (A) Regulatory sequences are cloned upstream of a *lac*Z reporter into the AttP2 site[63]. (B) Embryos are stained using fluorescent *in situ* hybridization (FISH) for *lac*Z and antibody staining for *eve* and imaged using confocal microscopy. (C) Nuclei are identified and levels of *lac*Z and *eve* are taken in a 10% stripe from 35.5% to 92.5% embryo length along the anterior-posterior axis at nuclear cycle 14, time class 6[82]. (D) Multiple embryos are quantified to give an average expression level for any given enhancer along the AP axis. (E) Previously quantified levels of transcriptional regulators are shown[185, 184, 93].

this data[158, 81, 171, 164, 92, 70, 93, 122, 167].

1.5.1 Generation of quantitative TF data

A quantitative understanding of gene regulation requires data on both the levels of TFs and the levels of mRNA driven by enhancers. The levels of TFs acting to control *eve* have previously been quantified along the AP axis in the FlyEx database[152]. This data was collected by detecting Eve in combination with other gap genes, maternal genes, or expression driven by enhancer elements using fluorescently labeled secondary antibodies (Fig. 1.3A). The embryos were then imaged on the lateral surface using confocal microscopy (Fig. 1.3B). Next, the images were segmented to determine the average fluorescence in individual nuclei.

This signal correlates linearly with the true concentration of protein[62]. The Eve staining was used to register the data. Finally, nuclei along a 10% wide DV stripe were averaged to yield an expression profile for each TF along the AP axis[82] (Fig. 1.3C,D). mRNA driven by enhancers can be detected with fluorescent *in situ* hybridization (FISH) and subsequent protein detection[81, 93, 122]. If reporter expression is measured simultaneously with Eve, this data can be registered against the atlas of TF concentration data. This data amounts to a set of quantitative single cell assays of transcription input and output in a native tissue context, providing an extraordinarily precise system for testing theoretical models. Several groups have used the Flyex Database to develop sequence to expression models of gene regulation.

1.5.2 Estimation of binding affinity with PWMs

The first step in connecting TF levels to gene regulation is identifying TF binding sites and their relative affinities. In the absense of DNAse footprints, binding sites can be identified using what are called position weight matrices (PWMs). Under the assumption that each DNA contact contributes independently to binding affinity, PWMs represent the relative energetic contributions of every possible nucleotide at each position in a binding site[182]. Thus, PWMs can be used to estimate the binding affinity of any TF for any sequence[15]. Typically, a PWM is used to score every nucleotide in a sequence, and only sites above a threshold are considered to be viable TF binding sites, though there is no consensus on how thresholds should be set, which will be discussed below and in Chapter 2.

1.5.3 Occupancy of TFs

With the concentration of TFs and binding affinity for sequence, it is possible to estimate the relative occupancy of any TF on any sequence. That is, we solve the simple chemical formula

$$[\mathrm{TF}] + [\mathrm{DNA}_{\mathrm{free}}] \overset{K}{\rightleftharpoons} [\mathrm{DNA}_{\mathrm{bound}}]. \tag{1.1}$$

By the definition of K,

$$K = \frac{[\mathrm{DNA_{bound}}]}{[\mathrm{TF}][\mathrm{DNA_{free}}]}. \tag{1.2}$$

The fraction of time that DNA is bound is given by,

$$f = \frac{[\mathrm{DNA_{bound}}]}{[\mathrm{DNA_{bound}}] + [\mathrm{DNA_{free}}]}. \tag{1.3}$$

Substituting Eq. 1.2, we find

$$f = \frac{K[\mathrm{TF}][\mathrm{DNA_{free}}]}{[\mathrm{DNA_{free}}] + K[\mathrm{TF}][\mathrm{DNA_{free}}]}. \tag{1.4}$$

Canceling out, we find

$$f = \frac{K[\mathrm{TF}]}{1 + K[\mathrm{TF}]}. \tag{1.5}$$

Thus, the fractional occupancy f is a quantity that is independent of the concentration of DNA. In a system in which the rates of binding or unbinding are significantly faster than downstream effects, the fraction of time any site is bound is a more relevant quantity than any individual binding event. This appears to be the case for eukaryotic TF binding events which have been estimated to last tenths to tens of seconds[91, 22], and is significantly shorter than the time scale of transcriptional changes, which lasts minutes to tens of minutes in *Drosophila* embryogenesis.

1.5.4 Thermodynamics for the calculation of occupancy.

Real enhancer sequences contain large numbers of transcription factors that can compete for binding or bind to DNA cooperatively. To calculate occupancy in such cases it is useful to adopt the notation of thermodynamics. This approach assigns statistical weights to particular configurations of bound or free binding sites and was first used by Ackers and Shea to describe regulation by the phage λ repressor[2]. This approach is equivalent to the example (Eq 1.5) using stoichiometry[161]. In this calculation, each binding site is assigned

a Boltzmann weight, equal to the product of the TF concentration and binding affinity. The probability of finding the system in any particular binding state is proportional to the product of Boltzmann weights of each bound site in that state. Fractional occupancies can then be derived by dividing the weight of all states that contain the bound site by the weight of all states. The time it takes to perform this calculation scales with the number possible binding states, which is equal to 2^n for n binding sites. We denote this relationship between computational time as a function of binding sites by the expression $O(2^n)$. This exponential scaling sets a limit on the size of problems that can be reasonably calculated.

1.5.5 Multiple models of Drosophila cis-*regulation*

While occupancy of transcription factors on any DNA sequence can be calculated from well described chemical laws, the real challenge of understanding gene regulation is in determining which particular configurations of bound transcription factors drive transcription and which do not. Any quantitative description of the *eve* locus should at minimum include all the known mechanisms acting on the *eve* locus, which includes competitive and cooperative binding, quenching, and coactivation. Several groups have published models of gene regulation for enhancers that drive expression along the *Drosophila* AP axis. While each model may at first glance appear to be roughly equivalent, and each uses the same FlyEx Database of TF levels, small differences in the underlying assumptions can lead to major differences in the implementation and conclusions.

1.5.6 A binding site model of eve *stripe 2*

In 2003, Reinitz and Sharp proposed the first model of transcriptional control that contained an explicit thermodynamic representation of occupancies[158], and this model was first applied to the expression driven by the proximal 1700 bp of *eve* by Janssens *et al*[81]. In this model, binding sites and affinities were calculated using PWMs. Occupancies were calculated using the previously described approach, however to improve computational efficiency

competition for binding was only calculated for the nearest neighboring binding site. After occupancies were calculated, quenching was treated with a multiplicative reduction in the effective occupancy of activator binding sites. The magnitude of quenching was mediated by the occupancy of nearby repressors. Finally, unquenched activators caused an exponential rise in transcription rate. Because this model only treated regulation of *eve* stripes 2 and 7, Hb was fixed as an activator. This model was able to predict the effects of mutations in both *cis* and *trans*.

1.5.7 A model of 44 Drosophila *developmental enhancers*

Two years later, Segal *et al.* published a model of 44 enhancers active along the AP axis. The approach differed from Janssens *et al.* in several ways. First, there was no quantitative data available for the enhancer expression patterns. Instead, the model was fit to digitized patterns generated by hand from previously published figures. Second, the PWMs used to identify binding sites were treated as parameters, that must be fit to data. Third, occupancies were never calculated. Instead, any given binding state contributed to expression according to a weighted sum of the activators and repressors bound in that state, where activators contribute positively and repressors negatively. This gives a single value for each binding state, which was then run through a sigmoid function to find the contribution of the state. The total activation driven by a sequence was treated as a weighted sum of the activation driven by each binding state, multiplied by the probability of finding the system in that state. This approach does not capture quenching, which limits the scope to small, independently acting, sequence elements. Similarly, the model did not include any coactivation mechanism, so it is not a surprise that it failed to describe the expression pattern driven by MSE2.

1.5.8 Using models to investigate mechanisms acting on enhancers

A fundamental goal of transcription models is to identify whether known mechanisms are sufficient to explain the activity of enhancers. With this goal in mind, He *et al.* built a set of

models that incorporated different sets of mechanisms[70]. They allowed repression to occur through direct interaction with the basal machinery, similar to the negative weights of Segal *et al*, or to act through short range quenching. Rather than calculate occupancies, they modeled quenching by introducing an additional thermodynamic state for each quencher. In one state, the repressor is bound, but inactive. In the other, the repressor is bound, and prevents the activity of nearby activators. They model activation as additive or synergistic, in which activators cooperatively recruitment. Finally, they incorporated cooperativity of binding. All of this was calculated using an algorithm that considers all possible thermodynamic states. Despite the apparent necessity of short range repression mechanisms in explaining the behavior of intact loci[61], the utility of short ranged mechanisms in explaining the behavior of individual enhancers was unclear. In contrast, cooperativity and synergistic activation both greatly improved the accuracy of models. The accuracy here was assessed using correlation coefficients, which is a poor measure for accuracy of fits to expression patterns[92].

1.5.9 Subsequent developments

The work of Janssens *et al.* and Reinitz and Sharp *et al.* has been expanded. In 2013, Kim *et al.* used the model to explain the behavior of fusions of the MSE2 and MSE3 enhancers[93]. When these enhancers are separated by a spacer, their activity is roughly independent and additive, however if the enhancers are fused there are dramatic changes in the expression pattern they drive[174]. These fusions proved to be a rigorous test of distance-dependent mechanisms in models of gene regulation, because differences in expression are due to differences in the distance between binding sites. Successful modeling of this behavior required the addition of coactivation of Hb, as well as cooperativity of Bcd binding. This model has also been used to reveal accessible evolutionary paths for diverged *eve* stripe 2 enhancers[122].

1.5.10 Optimizing models of transcription

All of these models include some number of unknown parameter values, called free parameters, that must be fit to the data. For instance, in Kim *et al*, scaling factors for relative to absolute transcription factor concentration, the efficiency of quenching, or activation or coactivation, the strength of Bicoid cooperativity, and the threshold for binding were all free parameters. Global optimization is used to finds the value parameters that minimizes the differences between the model output and data. The work here uses an optimization method called Lam-Delosme Simulated Annealing[159, 99, 100]. This optimization process requires testing millions of values parameters. Models of transcription must be implemented in a computationally efficient framework such that testing their output millions of times is not prohibitive.

1.5.11 Current limitations of enhancer models

While these previously published models of enhancer function have had broad successes in describing and predicting the behavior of regulatory DNA, there are currently limitations to their broad application. First, models must incorporate all the TF interactions necessary in the studied system. For instance, neither He *et al.* nor Segal *et al.* considered coactivation of Hb. For this reason, it was not possible to model the behavior of both MSE3 and MSE2 in the same model. Similarly, Segal *et al.* did not include a quenching mechanism making it impossible to model large pieces of DNA that contain multiple enhancers. At this point, no framework is able to simultaneously model the activity of the entire *eve* locus, and only the work of Kim *et al.* was able to model the activity DNA sequences containing more than one enhancer. Second, computational efficiency limits the scope of problems that can be solved. Calculations operating on thermodynamic states grow exponentially, so either new techniques or approximations are necessary to increase the scale of problems that can be addressed with modeling.

1.6 Chapter overview

The results in this thesis are presented in four additional chapters. In Chapter 2, I describe a new algorithm for the calculation of transcription factor occupancy. This calculation scales linearly with the number of binding sites. The resulting improvement in computational speed allows thermodynamic calculations to be extended to problems on a genomic scale. In Chapter 3, I describe the first sequence level model for the regulation of the entire *eve* locus. I show that the enhancer structure of *eve* arises out of competition between DNA segments for interaction with the basal transcription machinery. In Chapter 4, I use the model to generate and test a panel of synthetic enhancers designed to drive expression of *eve* stripe 2. The results of this test indicate that there are new, unmodeled TFs and distance-dependent interactions that were not included in the model. In Chapter 5, I test how molecular interactions within enhancers make them robust to changes in TF concentration or enhancer sequences. In Chapter 6, I summarize these results and their implications, and introduce new perspectives and future directions.

CHAPTER 2

EFFICIENT CALCULATION OF TF-DNA BINDING EQUILIBRIUM

2.1 Introduction

The first step in sequence-to-expression models is a calculation of fractional occupancies of TFs bound to DNA. It is especially important that this step is done correctly, as any errors will be propagated by all future steps in a feed forward model. Unfortunately, this step is also computationally expensive, and computation time scales exponentially with the number of binding sites. For that reason, groups have adopted approximations[81], used sampling[171], or have only considered high affinity sites[70]. Ideally, we would like to efficiently consider all possible binding states, including those with low-affinity sites. In this chapter, I fully describe a sequence level model of gene regulation that includes a new, efficient algorithm for the calculation of transcription factor occupancy. This time required to execute this algorithm grows linearly with the number of sites. I provide new code implementing the algorithm and demonstrate its efficiency compared to previous implementations of this model. I have made several modifications to the underlying model. First, this algorithm is able to consider all possible configurations of pairwise cooperatively bound transcription factors. In addition, the model as described here considers both sequence specific and non-specific interactions between TFs and DNA.

2.2 Gene Regulation Model

We calculate the rate of transcription under the control of a DNA sequence in six steps: (1) calculation of binding affinity across DNA sequence and designation of binding sites, (2) calculation of fractional occupancy of factors at each binding site, (3) coactivation of Hunchback by locally bound Bicoid or Caudal, (4) quenching of activators by locally bound repressors,

(5) weighted summation of unquenched activators over the length of DNA sequence. Finally, the rate of transcription induced by a DNA fragment is related to the number of bound, unquenched activators by a diffusion limited Arrhenius rate law. In the sections that follow, I go through the equations that implement each of these steps, and I highlight any changes I have introduce that were not used in Kim *et al.*[93].

2.2.1 PWM Scores and Binding Affinity

Before any calculation of transcription factor occupancy, binding sites must be identified and assigned binding affinities. To do this we use position weight matrixes (PWMs, introduced in 1.5.2), which describe the relative sequence preferences of DNA-binding proteins. Given a PWM, the log-likelihood that an observed sequence i—spanning bases m though n—is a binding site for transcription factor a (which we indicate with index $i[m,n;a]$) is given by

$$S_{i[m,n;a]} = \sum_{k=m}^{n} \ln\left(\frac{P_a(k-m, j_k)}{P_{\mathrm{bg}}(j_k)}\right) \tag{2.1}$$

where j_k is the nucleotide observed at position k, $P_a(k-m,j)$ is the probability of observing this nucleotide at position $k-m$ in a binding site for factor a, and $P_{\mathrm{bg}}(j_k)$ is the probability of observing this nucleotide in the null distribution. We call S the PWM score. For the background distribution we use the frequencies of nucleotides in the *Drosophila* genome ($P_{\mathrm{bg}}(A) = P_{\mathrm{bg}}(T) = 0.297$, $P_{\mathrm{bg}}(C) = P_{\mathrm{bg}}(G) = 0.203$).

The PWM score can be used to calculate the sequence-specific binding affinity $K_{i[m,n;a]}$, which is given by

$$K_{i[m,n;a]} = K_a^{\max} \exp\left(\frac{S_i - S_a^{\max}}{\lambda_a}\right) \tag{2.2}$$

where $S_a^{\max}$ is the maximum score possible for the PWM of factor a and λ_a is a proportionality constant[15].

2.2.2 Thresholds

The calculation above yields a score and affinity for every TF at every position on DNA. It is computationally intractable to consider binding to every possible nucleotide. Moreover, the vast majority of predicted affinities are very low and in enhancers like MSE2 there are only a few footprinted binding sites for each TF. Somehow, only a select number of binding sites must be selected for future calculations. Typically, this is done by setting some threshold, where only PWM scores above a threshold are considered binding sites. To ensure that all functional binding sites are included, the threshold should be low enough that sites at this threshold are of low affinity and do not contribute to model output, but there is no consensus on what a proper threshold is. He *et al.* used what they call a p-value[70], where the p-value represents the probability of drawing a sequence of equal or greater PWM score from the background distribution of nucleotides. This value has no basis in biology, because only the binding affinity affects occupancy, independent from whether a site has greater or lesser affinity than other possible sites.

Some works have treated the PWM thresholds as free parameters (values that are fit to date, introduced in 1.5.10) [93, 123, 122]. While this allows the thresholds to be naturally set from the data, it also introduces a free parameter for each TF. In order to minimize the number of free parameters, and to reduce the risk of overfitting, we would like to eliminate binding site thresholds. Optimizing thresholds can introduce instability. To illustrate, consider a model of the MSE2 enhancer including Dicheate (Model 06, reported in Kim *et al.*[93]). The model identifies a single binding site for Dicheate, and one that falls just below the PWM threshold (Fig 2.1A). The model does a good job of fitting the expression pattern of MSE2, but if we lower the threshold by only 0.1, a second Dicheate site is introduced which leads to a 4-fold loss of predicted expression (Fig 2.1B). If a threshold represents the point at which binding becomes irrelevant, a proper treatment of PWM thresholds should yield model parameters that are relatively insensitive to small changes in threshold. In other words, the threshold should be set sufficiently low, such that binding sites just above or

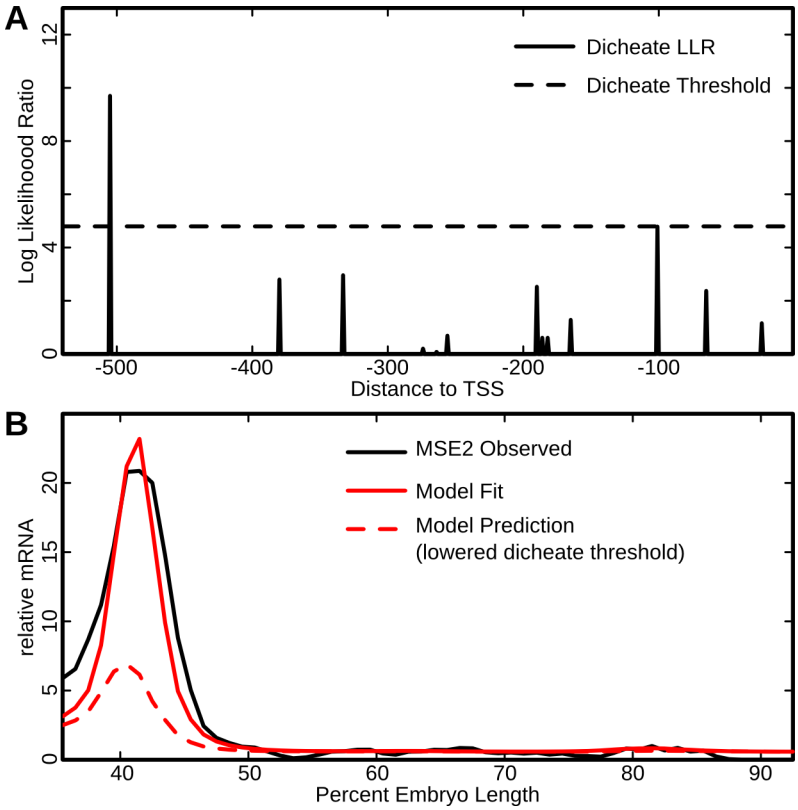


Figure 2.1. Instability introduced by PWM thresholds (A) The PWM score of binding for Dicheate over the length of MSE2. The PWM threshold of the optimized model is plotted with a dotted line. (B) The observed mRNA levels (black line) and output (red line) for an optimized model with threshold illustrated in A. When the PWM threshold is lowered by 0.1, enough to include the site 100 bp upstream of the TSS, the predicted output changes dramatically (dashed red line). The model is previously reported in Kim *et al.* as model 06[93].

below the threshold are of low occupancy and do not contribute to the output.

2.2.3 Specific and non-specific binding

Transcription factors have both sequence-specific and non-specific (sequence independent) interactions with DNA and the affinity of non-specific interaction provides a natural threshold for calling binding sites. For this reason, I added a new mechanism that accounts for non-specific interaction of TFs and DNA. Consider a single TF binding to a single binding site. The TF has specific affinity K and non-specific affinity K^{ns}. DNA can now exist in specifically bound, non-specifically bound, and free states, which we call $[\text{DNA}_{\text{sp}}]$, $[\text{DNA}_{\text{ns}}]$ and $[\text{DNA}_{\text{free}}]$ respectively. K and K^{ns} are defined by

$$K = \frac{[\text{DNA}_{\text{sp}}]}{[\text{DNA}_{\text{free}}][\text{TF}]}, \tag{2.3}$$

and

$$K^{\text{ns}} = \frac{[\text{DNA}_{\text{ns}}]}{[\text{DNA}_{\text{free}}][\text{TF}]}. \tag{2.4}$$

We also define the effective affinity, denoted K^{ef}, accounting for both non-specific and specific interactions.

$$K^{\text{ef}} = \frac{[\text{DNA}_{\text{sp}}]}{[\text{TF}]([\text{DNA}_{\text{free}}] + [\text{DNA}_{\text{ns}}])}. \tag{2.5}$$

Expanding the denominator and substituting Eqs. 2.3 and 2.4 we find

$$K^{\text{ef}} = \frac{K[\text{TF}][\text{DNA}_{\text{free}}]}{[\text{TF}][\text{DNA}_{\text{free}}] + K^{\text{ns}}[\text{TF}]^2[\text{DNA}_{\text{free}}]}. \tag{2.6}$$

Canceling out $[\text{TF}][\text{DNA}_{\text{free}}]$, we find

$$K^{\text{ef}} = \frac{K}{1 + K^{\text{ns}}[\text{TF}]}. \tag{2.7}$$

Non-specific binding affinity has been shown to be approximately three orders of magni-

tude smaller than the maximum specific binding energy for eukaryotic transcription factors[117], so for this work we set it one thousandth of the maximum binding affinity for a particular transcription factor.

For binding sites at or near the maximum affinity (i.e. a consensus binding site or good match to PWM), this adjustment makes only a small difference in the predicted affinity, and an even smaller difference in the predicted occupancy of a site. However, for binding sites with affinities below non-specific levels, the sites are more likely to be occupied non-specifically than specifically. For the work presented in Chapter 2, the PWM thresholds are set to 0. This means sites are considered that have affinity at or below the non-specific energy of binding. This adjustment for non-specific binding ensures that sites near the PWM threshold have low occupancy. In general we consider this adjustment to be conservative. In reality, the specifically bound site must compete with all other TFs present in the cell that can interact non-specifically with DNA.

2.2.4 Adjusting for true affinity and true concentration

While we do not know absolute concentrations of transcription factors, fluorescence scales linearly with concentration [62]. We need a parameter that scales fluorescence for factor a to true concentration v_a. We call this parameter $v_a^{\max}$. Wherever this parameter occurs it is always multiplied by a K, so we cannot separate the parameter $v_a^{\max}$ from $K_a^{\max}$. In light of this, we introduce a compound parameter, A_a, that scales the product of relative affinities and fluorescence measurements to true affinities and concentrations.

$$A_a = v_a^{\max} K_a^{\max}. \tag{2.8}$$

Combining Eqs. 2.2, 2.7 and 2.8, the effective affinity at any site is given by

$$K_{i[m,n;a]}^{\mathrm{ef}} = \frac{K_{i[m,n;a]}}{K_a^{\mathrm{ns}} v_a^{\mathrm{fl}} A_a + 1}. \tag{2.9}$$

2.2.5 Fractional Occupancy

Given a set of binding sites and their affinities, the next step is to determine the occupancy of each binding site as a function of concentration. The standard approach to calculating TF occupancy proceeds as follows. First, each binding site is assigned a Boltzmann weight, equal to the product of the of the binding affinity and the concentration of that site. Then, each possible binding state is considered, where there are 2^n possible states, representing all possible combinations of each binding site being occupied or empty. These states are assigned a Boltzmann weight equal to the product of the weight of each bound site in that configuration. States with overlapping binding sites are assigned a weight of zero. To find the occupancy of a given site, the weight of all states containing a site bound is divided by the weight of all states, known as the partition function Z. I illustrate this calculation in full for a simple problem with only 3 non-competing binding sites of equal affinity in Figure 2.2. While the binding of these three sites are independent, and it would be possible to solve the occupancy of each one without considering the others, the example is useful for illustration.

To demonstrate the computational cost of this algorithm as a function of increasing the number of binding sites, as well as to demonstrate the computational cost of lowering thresholds, I tested the effects of lower thresholds in the seven construct model reported by Kim *et al.*[93]. I tested the number of binding sites and time per cost function evaluation at different thresholds of binding. I specified thresholds according to the percent of the maximum PWM score. I found that as the threshold is lowered, there is an approximately exponential increase in the number of binding sites identified (Fig 2.3A). The time for a single cost function evaluation grows approximately exponentially with the number of sites (Fig 2.3B), and greater than exponentially with lowered thresholds (Fig 2.3C). The time per iteration represents the average over 100,000 iterations of the cost function, as assessed by the Unix 'time' command. All calculations were performed in serial on a Dell Precision Tower 7910 with 2 Intel(R) Xeon(R) E5-2670 v3 CPUs at 2.30GHz and 64 Gb PC4-17000 ECC RAM. Thresholds of zero could not be considered due to integer overflow in the number

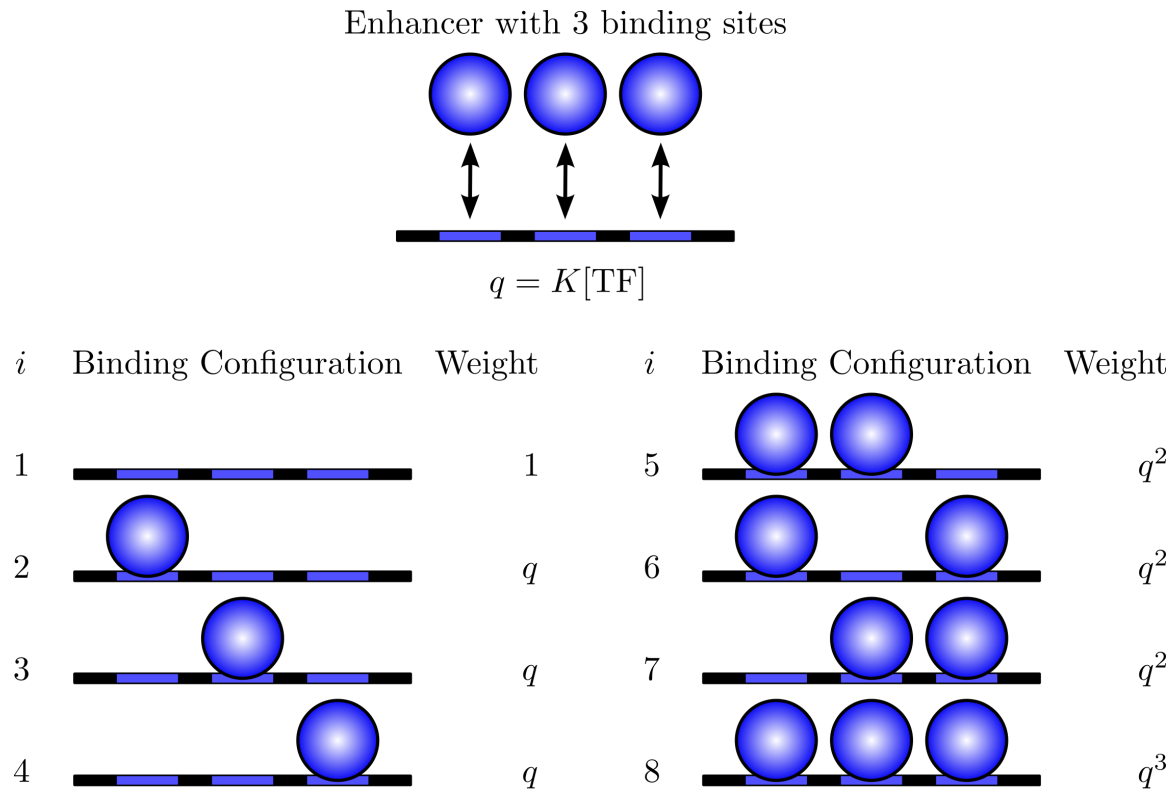


Figure 2.2. Standard computation of fractional occupancy: an example The calculation of fractional occupancy is illustrated for a simple sequence that contains three binding sites of equal affinity for the same transcription factor. The 2^n thermodynamic states are enumerated. For each state a Boltzmann weight is calculated. The fractional occupancy of each factor is given by the weight of all states that contain that site, divided by the weight of all states, the partition function Z.

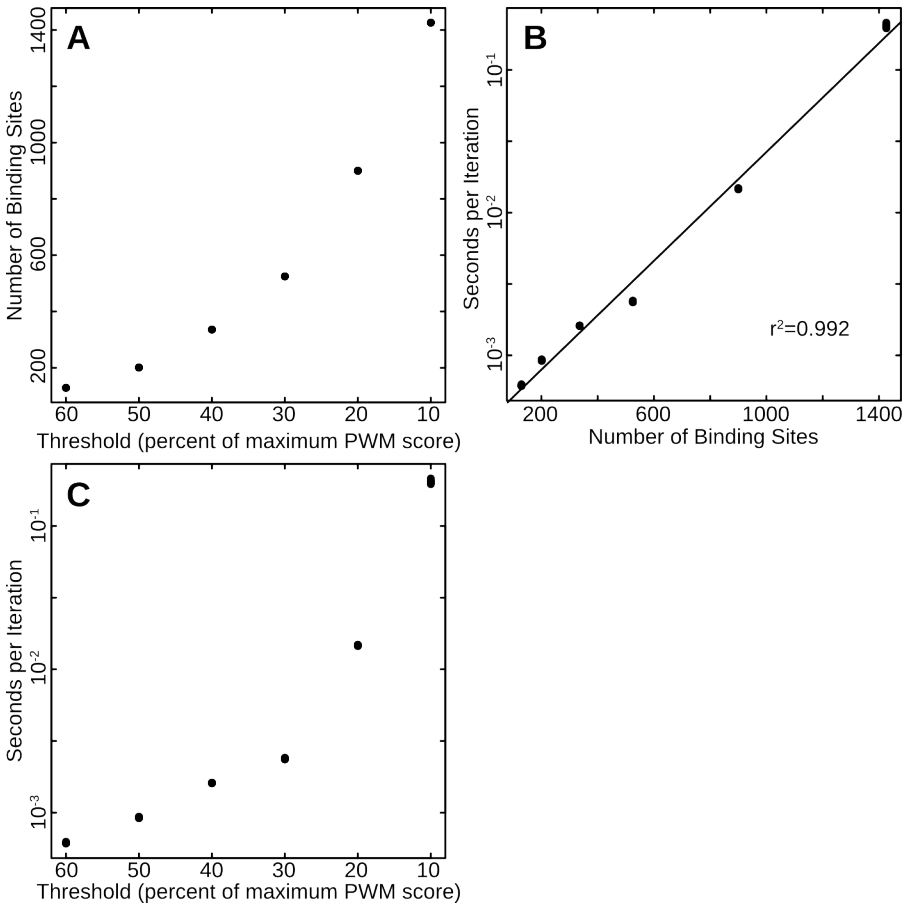


Figure 2.3. Computation time as a function of PWM threshold (A) The number of binding sites grows approximately exponentially with lowered thresholds. The number of binding sites in the seven construct model (Model 06, Kim *et al.*[93]) is plotted as a function of lowering PWM thresholds. Thresholds for each factor are set to a percent of the maximum PWM score for each factor. Thresholds of zero could not be tested due to integer overflow in number of binding states. (B) The time for a single cost function evaluation grows exponentially with the number of binding sites identified. The natural logarithm of the time per cost function evaluation is plotted as a function of the number of binding sites. Each point is the average time for 100,000 cost function evaluations. There are 10 repetitions for each test. The r^2 of a linear regression to this data is given, and the regression line is shown in black. (C) The time for a single cost function evaluation grows greater than exponentially with the PWM threshold. The the time per cost function evaluation is plotted as a function of the the PWM threshold. Each point is the average time for 100,000 cost function evaluations. There are 10 repetitions for each test.

of possible binding states.

2.2.6 Efficient calculation of fractional occupancy in a simple example

In order to calculate fractional occupancy of a given site, we only need to know the weight of all states that contain a site bound, and the weight of all states. This does not mean that the Boltzmann weight of every possible binding configuration must be calculated independently. Instead, we can do calculations on the weights of groups of states, similar to the way the number Z holds the weight of all states. In this section I illustrate this principle, demonstrate it with a simple problem, and build up to successively harder problems.

First, consider the simple example with all binding states enumerated in Figure 2.2. We name these sites sites 1, 2 and 3 in order of position 5' to 3' (see Fig 2.4). Consider an enhancer that only contains site 1. The partition function Z containing all the sum of Boltzmann weights of all binding configurations is simply $1 + q$. Now if we want to find the partition function for an enhancer containing sites 1 and 2, the partition function can be calculated by the sum of the partition function from the enhancer with only 1 binding site, plus all combinations of site 2 with site 1. The second part of this sum is just given by the product of the partition function of the single site enhancer and the weight of a single binding site. In other words, if we define the partition function of the single site enhancer as Z_1, then the partition function of the two site enhancer is given by

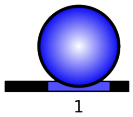
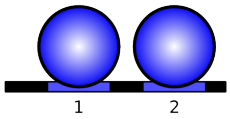
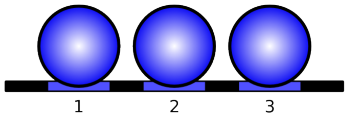
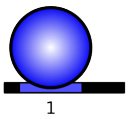
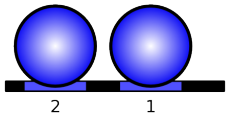
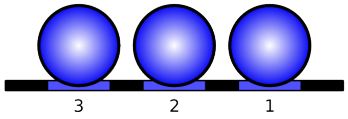
$$Z_2 = Z_1 + qZ1 \tag{2.10}$$

If we define $Z_0 = 1$, representing the weight of a configuration containing no binding, it is simple to extend this example to any number of binding sites.

$$Z_i = Z_i + qZ_{i-1} \tag{2.11}$$

This example demonstrates the following simple rules.

Initialize paritition function $Z_0 = 1$

i	Sites considered	Partial Partition	Total Partition
1		$Z_1^+ = qZ_0$ $= q$	$Z_1 = Z_0 + Z_1^+$ $= 1 + q$
2		$Z_2^+ = qZ_1$ $= q + q^2$	$Z_2 = Z_1 + Z_2^+$ $= 1 + 2q + q^2$
3		$Z_3^+ = qZ_2$ $= q + 2q^2 + q^3$	$Z = Z_2 + Z_3^+$ $= 1 + 3q + 3q^2 + q^3$
1		$Z_1^- = qZ_0$ $= q$	$Z_1 = Z_0 + Z_1^-$ $= 1 + q$
2		$Z_2^- = qZ_1$ $= q + q^2$	$Z_2 = Z_1 + Z_2^-$ $= 1 + 2q + q^2$
3		$Z_3^- = qZ_2$ $= q + 2q^2 + q^3$	$Z = Z_2 + Z_3^-$ $= 1 + 3q + 3q^2 + q^3$

$$f_1 = \frac{Z_1^+ Z_3^-}{qZ} = \frac{q+2q^2+q^3}{Z}; f_2 = \frac{Z_2^+ Z_2^-}{qZ} = \frac{q+2q^2+q^3}{Z}; f_3 = \frac{Z_3^+ Z_1^-}{qZ} = \frac{q+2q^2+q^3}{Z}$$

Figure 2.4. Dynamic computation of fractional occupancy: an example The dynamic calculation of fractional occupancy is illustrated for a simple sequence that contains three binding sites of equal affinity for the same transcription factor. Rather than enumerate all configurations, we take two linear passes along DNA, one in the forward and one in the reverse direction. We keep track the partial partition function in the forward (Z_i^+) or reverse (Z_i^-) direction, which contain the Boltzmann weights of all states through site i in which site i is bound. We also keep track of the total partition function Z which keeps track of the weight of all states through site i. The weight of all states in which site i is bound can be recovered with the function ($\frac{Z_i^+ Z_i^-}{q}$).

1 Given an ordered array of binding sites, the sum of Boltzmann weights of all binding states up through and including site i is given by the sum of weights through site $i-1$ plus the weight of all states that contain a binding site.

2 The sum of Boltzmann weights of all states up through and including site i, in which i is bound, is given by the the sum of all previous weights multiplied by the weight of the site alone.

Because these sums represent sums of Boltzmann weights, similar to the partition function Z, we call these partial partition functions. We define the following partial partition functions.

Z The partition function, containing the sum of Boltzmann weights of all possible binding configurations of an enhancer.

Z_i A partial partition function containing the sum of all Boltzmann weights of configurations through site i.

Z_i^+ A partial partition function containing the sum of all Boltzmann weights of configurations through site i, ordered from 5' to 3', in which site i is bound.

Z_i^- A partial partition function containing the sum of all Boltzmann weights of configurations through site i, ordered from 3' to 5', in which site i is bound.

The partition function Z can be calculated in either the 5' to 3' direction, or the 3' to 5' direction, by the recursive relation

$$\begin{aligned} Z_i &= Z_{i-1} + Z_i^+ \\ &= Z_{i-1} + qZ_{i-1} \end{aligned} \tag{2.12}$$

This calculation is illustrated in both the forward or reverse direction in Figure 2.4.

This calculation provides an efficient method to calculate the partition function, however this is only part of the calculation of occupancy. In addition, we need the sum of Boltzmann

weights of all binding configurations in which any site i is bound. This sum can be recovered from the partial partition functions Z_i^+ and Z_i^-. These numbers contain the weight of all configurations in which site i is bound in combination with sites in the 5' or 3' direction respectively. To recover the weight of all states in which i is bound in combination with sites in the 5' and 3' directions, we simply multiply these two quantities, and divide by the weight of site i bound, q.

$$f_i = \frac{Z_i^+ Z_i^-}{q} \tag{2.13}$$

This is true for the same reason we can calculate $Z_i^+ = qZ_{i-1}$, however because both Z_i^+ and Z_i^- are already multiplied by the weight of this site with all the combinations, we have double counted q and must divide though. This calculation is illustrated in Figure 2.4 and can be checked against the standard calculation in Figure 2.2.

2.2.7 Efficient calculation of fractional occupancy incorporating cooperativity and competition

The previous section gave an example of an efficient algorithm for calculating occupancy in a simple problem containing three non-competing binding sites. In this case, the binding of each site is independent. The process of separating these sites is called subgrouping, where a subgroup is defined as a minimum set of binding sites such that the calculation of occupancy is not independent. For the algorithm to be useful, we need to be able to consider cooperativity of binding and competition for binding. To be able to do so we must introduce new indexing and two new partial partition functions.

First, the binding sites, indexed by $i[m_i, n_i; a_i]$, are sorted such that i increases monotonically with n in the 5' to 3' direction. We define a new indexing function $h(i)$ that returns the index of the last site that does not compete for binding with site i

$$h\big(i[m,n;a]\big) = \max_{n_k \leq m_i} \big(k[m,n;a]\big). \tag{2.14}$$

We note that if binding sites have different widths, this indexing function is not identical in the 3' to 5' direction compared to the 5' to 3' direction.

Additionally, we define the variable q_i as

$$q_i = K^{\mathrm{ef}}_{i[m,n;a]} v^{\mathrm{fl}}_a A_a \tag{2.15}$$

We also define the partial partition functions:

$Z_i^{\mathrm{nc}+}$ A partial partition function containing the sum of all Boltzmann weights of configurations through site i, ordered from 5' to 3', in which site i is bound non-cooperatively.

$Z_i^{\mathrm{nc}-}$ A partial partition function containing the sum of all Boltzmann weights of configurations through site i, ordered from 3' to 5', in which site i is bound non-cooperatively.

$Z_i^{\mathrm{c}+}$ A partial partition function containing the sum of all Boltzmann weights of configurations through site i, ordered from 5' to 3', in which site i is bound cooperatively.

$Z_i^{\mathrm{c}-}$ A partial partition function containing the sum of all Boltzmann weights of configurations through site i, ordered from 3' to 5', in which site i is bound cooperatively.

Now, the partition function through site i is given by the sum of weights of all states through sites $i-1$, plus the weight of all states in which site i bound non-cooperatively, plus the weight of all states in which site i bound cooperatively.

$$Z_i = Z_{i-1} + Z_i^{\mathrm{nc}+} + Z_i^{\mathrm{c}+}, \tag{2.16}$$

The weight of all states through site i in which site i is bound non-cooperatively, $Z_i^{\mathrm{nc}+}$, is given by the weight of this binding site, multiplied by the weight of all prior non-competing binding states.

$$Z_i^{\mathrm{nc}+} = q_i Z_{h(i)} \tag{2.17}$$

We allow sites to bind cooperatively in pairs. This means that in any given binding state,

a cooperativly bound TF interacts with exactly one partner. However, If there are multiple potential partners, interactions with each parter define a separate viable state. This is an advance over work by Kim *et al.*[93], where only one potential cooperative parter was considered. To calculate the weight of of all states through site i, in which site i is bound cooperatively, $Z_i^{\mathrm{c}+}$, we must calculate the sum of weights of all potential interacting partners. The weight of the cooperative pair is given by $q_i q_k w(i,k)$, where $w(i,k)$ is the cooperative interaction strength between sites i and k. We define $w(i,k)$ to be zero if the distance between these sites is greater than 60 bp. This pair can bind in combinations with all states that do not compete with the first binding site in the pair, given by $Z_{h(k)}$. Altogether, we find

$$Z_i^{\mathrm{c}+} = \sum_{k=1}^{i-1} q_i q_k w(i,k) Z_{h(k)}. \tag{2.18}$$

We calculate the array of weight sums Z_i in both the forward (5' to 3') and reverse (3' to 5') direction across DNA and denote this Z_i^-, and Z_i^+. Additionally, we store partial partition functions in each direction $Z_i^{\mathrm{nc}+}$, $Z_i^{\mathrm{c}+}$, $Z_i^{\mathrm{nc}-}$, and $Z_i^{\mathrm{c}-}$. Now the fractional occupancy of site i can be recovered by calculating

$$f_{i[m,n;a]} = \frac{Z_i^{\mathrm{nc}+} Z_i^{\mathrm{c}-} + Z_i^{\mathrm{c}+} Z_i^{\mathrm{nc}-} + Z_i^{\mathrm{nc}+} Z_i^{\mathrm{nc}-}}{Z q_i}. \tag{2.19}$$

The quantity in the denominator is similar to in Eq. 2.13, however this now considers all states in which the site is bound cooperatively in the 3' direction, 5' direction, or neither. The multiple $Z_i^{\mathrm{c}+} Z_i^{\mathrm{c}-}$ would consider states in which site i has two cooperative partners and is deliberately excluded.

While calculations that explicitly consider the weight of every binding state scale must loop over all 2^n states, where n is the number of binding sites, the dynamic algorithm presented here only needs to perform n calculations. We note that this has a lot of similarities to an algorithm that was independently derived by Teif *et al.* [191]. The major difference is that Teif *et al.* allows cooperative interactions to polymerize along DNA.

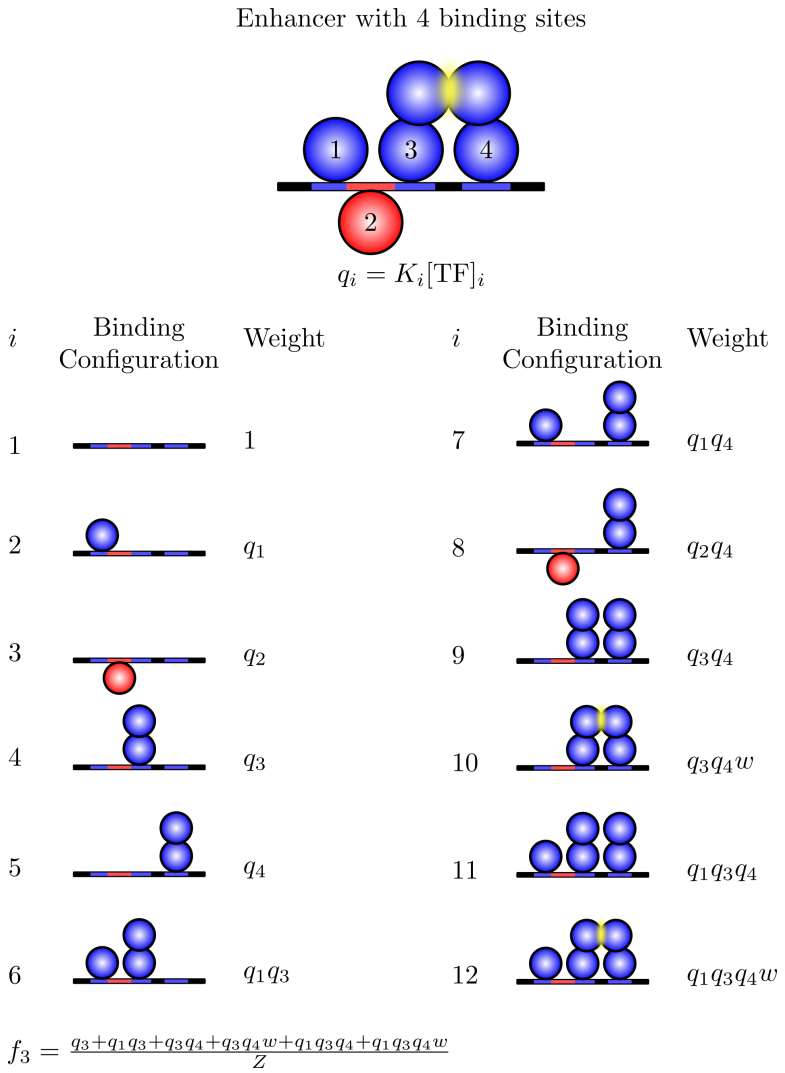


Figure 2.5. Standard computation of fractional occupancy in a complex example The standard calculation of fractional occupancy is illustrated for a complex example that contains four binding sites of different affinity or concentration. Site 2 competes for binding with sites 1 and 3, while sites 3 and 4 can bind cooperatively. All permissible binding states of this enhancer are enumerated, and the Boltzmann weight of each state is given. States that have site 2 bound with sites 1 or 3 are not permissible, have a weight of zero, and are not shown. The cooperativity of binding between sites 3 and 4 leads to an additional contribution of w to the weight cooperatively bound configurations. The calculated occupancy of site 3 is given.

I have included a demonstration of this algorithm for an enhancer that demonstrates both cooperative and competitive interactions using both the standard computation over all configurations (Figure 2.5) and using the new algorithm (Figure 2.6). For a single site, I showed the calculated occupancy and verified that this quantity is the same between the methods. Additionally, I have verified that the code gives identical results across a wide range of problems.

2.2.8 Action at a Distance

Once the occupancy of transcription factors is calculated, interactions between transcription factors are considered. TFs interact according to the distance between sites, where nearby sites have strong interactions and distant sites do not interact. In order to define rules governing these interactions, we define a function that determines the efficiency of interaction between sites i and k. We report the the distance between sites as

$$d(i,k) = \min(\left|m_i - n_k\right|, \left|n_i - m_k\right|), \tag{2.20}$$

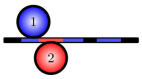
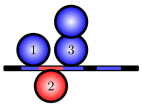
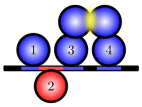
and the efficiency of interaction is given by

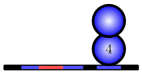
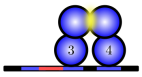
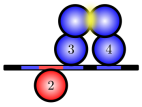
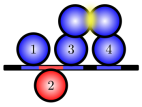
$$g(i,k,A,B) = \begin{cases} 1 & d(i,k) \leq A \\ 1 - \frac{d(i,k)-A}{B} & A < d(i,k) < A+B \\ 0 & A+B \leq d(i,k) \end{cases}, \tag{2.21}$$

where A and B govern the shape of this interaction.

2.2.9 Coactivation

We allow bound factors to take one of two possible states: an activating state f^A and a quenching state f^Q. For obligate repressors $f^Q_{i[m,n;a]} = f_{i[m,n;a]}$, and similarly for obligate

i	Sites considered	$h(i)$	$Z_i^{\mathrm{nc}+}$	$Z_i^{\mathrm{c}+}$	Z_i
1		0	q_1	0	$1+q_1$
2		0	q_2	0	$1+q_1+q_2$
3		1	$q_3(1+q_1)$	0	$1+q_1+q_2+q_3(1+q_1)$
4		3	$q_4(1+q_1+q_2+q_3(1+q_1))$	$(1+q_1)q_3q_4w$	Z

i	Sites considered	$h(i)$	$Z_i^{\mathrm{nc}-}$	$Z_i^{\mathrm{c}-}$	Z_i
4		0	q_4	0	$1+q_4$
3		4	$q_3(1+q_4)$	q_3q_4w	$1+q_3+q_4+q_3(1+q_4)$ $+q_3q_4w$
2		4	$q_2(1+q_4)$	0	$1+q_1+q_2+q_3(1+q_4)$ $+q_3q_4w+q_2(1+q_4)$
1		3	$q_1(1+q_3+q_4$ $+q_3(1+q_4)+q_3q_4w)$	0	Z

$$
\begin{aligned}
f_3 &= \frac{Z_3^{\mathrm{nc}+}Z_3^{\mathrm{c}-} + Z_3^{\mathrm{c}+}Z_3^{\mathrm{nc}-} + Z_3^{\mathrm{nc}+}Z_3^{\mathrm{nc}-}}{q_3 Z} \\
&= \frac{q_3(1+q_1)q_3q_4w + q_3(1+q_1)q_3(1+q_4)}{q_3 Z} \\
&= \frac{q_3q_4w + q_1q_3q_4w + q_3 + q_3q_4 + q_1q_3 + q_1q_3q_4}{Z}
\end{aligned}
$$

Figure 2.6. Dynamic computation of fractional occupancy in a complex example The dynamic calculation of fractional occupancy is illustrated for the enhancer in Figure 2.5. The values of the indexing function $h(i)$ and partial partition functions Z_i, Z_i^{nc} and Z_i^{nc}. are given at every iteration along the array of binding sites in both the 5' to 3' direction (top) as well as the 3' to 5' direction (bottom). Additionally, the calculated occupancy of site 3 is given, and gives the same answer as in Figure 2.5.

activators $f^A_{i[m,n;a]} = f_{i[m,n;a]}$. For proteins that can take on both possible states we allow state switching from a default state to an induced state based on the occupancy of nearby inducing factor. This is the case for the transcription factor Hunchback, a quencher that can be induced to activate by the presence of bound Bicoid or Caudal[93]. For sites i and k, we define a function that returns the efficiency of this interaction based on the distance between these two sites

$$f^Q_{i[m_i,n_i;a_i]} = f_{i[m_i,n_i;a_i]} \prod_k \big(1 - g(i,k,D_c,50) E^C_{a_k} f_{k[m_k,n_k;a_k]}\big), \tag{2.22}$$

$$f^A_i = f_i - f^Q_i \tag{2.23}$$

where D_c is a free parameter giving the maximum distance at which coactivation is 100% efficient and $E^C_{a_k}$ is a free parameter giving the maximum efficiency with which factor a_k induces activation of factor a_i. This product occurs over all k binding sites within the locus.

2.2.10 Quenching

Bound repressors quench the activity of bound activators. This results in an an effective occupancy of each each activator, F, which is given by

$$F_i = f^A_i \prod_k \big(1 - g(i,k,100,50) E^Q_{a_k} f^Q_k\big), \tag{2.24}$$

where $E^Q_{a_k}$ is a free parameter giving the efficiency with which factor a_k quenches. This product occurs over all k binding sites within the locus.

2.2.11 Summation of Recruited Transcriptional Adaptors

The remaining bound, unquenched activators are free to recruit transcriptional adaptors that enhance interaction with the promoter. The number of adaptors recruited to a sequence

bounded by bases p and q is given by

$$N_{[p,q]} = \sum_k F_k E^A_{a_k} I(k,p,q), \tag{2.25}$$

where the $E^A_{a_k}$ is the activating efficiency of factor a_k and $I(k,p,q)$ is a function that specifies whether site k falls between p and q, given by

$$I(k,p,q) = \begin{cases} 1 & m_k \geq p, n_k < q \\ 0 & \text{Otherwise} \end{cases} \tag{2.26}$$

In Kim *et al.*[93] and Martinez *et al.*[123, 122], the recruited adapters were allowed to simultaneously interact with the basal transcription machinery. In Chapter 3, in order to model large DNA segments, I will only allow specific sequences, bounded by p and q to interact at a single time, leading to competition between segments.

2.2.12 Three State Promoter Model

We allow a promoter to have three states: (1) an open state with no polymerase bound, (2) a paused, bound state, and (3) an actively elongating state. Only the binding and unbinding of polymerase is reversible. This model has been presented by

$$\text{Open} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \text{Bound} \overset{k_2}{\rightarrow} \text{Elongating} \overset{k_3}{\rightarrow} \text{Open}. \tag{2.27}$$

The probability of the system being in any state is given by the system of differential

equations

$$\begin{aligned} \frac{dP_1}{dt} &= -k_1 P_1 + k_{-1} P_2 + k_3 P_3 \\ \frac{dP_2}{dt} &= k_1 P_1 - (k_{-1} + k_2) P_2 \\ \frac{dP_3}{dt} &= k_2 P_2 - k_3 P_3. \end{aligned} \tag{2.28}$$

At steady state, all the derivatives vanish, leaving three equations in three unknowns. Solving for the P_is, we obtain

$$\begin{aligned} \bar{P_1} &= \frac{k_3(k_{-1} + k_2)}{q} \\ \bar{P_2} &= \frac{k_1 k_3}{q} \\ \bar{P_3} &= \frac{k_1 k_2}{q}, \end{aligned} \tag{2.29}$$

where $q = k_3(k_{-1} + k_2) + k_1(k_2 + k_3)$. The rate of transcription is the rate at which the system moves through state 3

$$\bar{P_2} k_2 = \bar{P_3} k_3 = \frac{k_1 k_2 k_3}{k_3 k_{-1} + k_3 k_2 + k_1 k_2 + k_1 k_3}. \tag{2.30}$$

Given that we are modeling a developmental promoter displaying paused polymerase, we assume that the rate limiting step is initiation. Note that as $k_2 \to \infty$, the rate of transcription goes to some value $R_{\max}$. From eqn. 2.30, as $k_2 \to \infty$,

$$k_1 k_3 = R_{\max}(k_1 + k_3). \tag{2.31}$$

Substituting this relation into eqn. 2.30 we find the transcription rate is given by

$$R = \frac{k_2}{1 + K_d + \frac{k_2}{R_{\max}}}, \tag{2.32}$$

where $K_d = \frac{k_{-1}}{k_1}$. We treat the reaction rate k_2 as a rate catalyzed by adaptor factors recruited to the promoter by enhancers, where each recruited adaptor gives a reduction in the energy barrier to transcription initiation. This rate is given by

$$k_2 = c \exp\Big(\frac{-\Delta A}{RT}\Big), \tag{2.33}$$

and

$$\Delta A = \phi - M \tag{2.34}$$

where ϕ is the barrier to initiation and M is the reduction $\Delta\Delta A$ in the activation energy. Multiplying Eq. 2.32 and substituting Eq. 2.33 for k_2, we obtain

$$R = R_{\max} \frac{c \exp(\frac{M-\phi}{RT})}{R_{\max}(1+K_d) + c \exp(\frac{N-\phi}{RT})}. \tag{2.35}$$

Dividing through by the numerator, we obtain

$$R = \frac{R_{\max}}{1 + \exp\big(\frac{\phi - M}{RT} + \ln(\frac{R_{\max}(1+K_d)}{c})\big)}. \tag{2.36}$$

We define the quantity $N = \frac{M}{RT}$ and $\theta = \frac{\phi}{RT} - \ln(\frac{R_{\max}(1+K_d)}{c})$. Substituting into Eq 2.36, we derive the transcription rate equation

$$R = \frac{R_{\max}}{1 + \exp(\theta - N)}. \tag{2.37}$$

2.3 Code implementing this model

To implement the new algorithm, I wrote a new program that implements the above equations. The code implementing this model is included in supplemental files. Several design principles were used to implement the new code, each of which led to substantial improvements in computational efficiency. These design principles are discussed below.

2.3.1 The C++ Programming language

While previous versions of the this transcription model were written in C, I selected C++ to implement the new code. C++ was selected because it uses an object-oriented programming paradigm and does not sacrifice speed compared to C. The model equations described above fit naturally into an object-orientated paradigm, where individual PWMS, TFs, DNA fragments, protein-protein interactions, and nuclei are objects.

2.3.2 The basic loop

The code must calculate the transcription rate for every construct and at every nucleus. Previous code placed nuclei in the outer loop and constructs in the inner loop. I have reversed this direction to minimize the cost of loading data into memory. To calculate the occupancy of a single site at various nuclei, the only variable that changes is the concentration of that TF. I can have everything else loaded into memory and available and simply pass an array of TF concentrations. In contrast, if nuclei are at the outer loop, the entire binding array must be loaded and unloaded several times per iteration. This is computationally expensive. Having constructs at the outer loop makes parallelizing or extending this model to large numbers of sequences trivial, though it makes extending the model to large numbers of cells or tissues more expensive.

2.3.3 Move functions

During optimization, the equations must be calculated millions of times, making the time to recalculate of paramount importance. The code I have generated in this work is designed with the principle that when a parameter is changed, only the data structures that will change should be updated. For instance, if we want to test a new value of Bicoid cooperativity, it is unnecessary to update the PWM scores binding sites. Similarly, if we update the binding affinity of Gt, we do not need to update the affinity of Kr sites. In practice, implementing

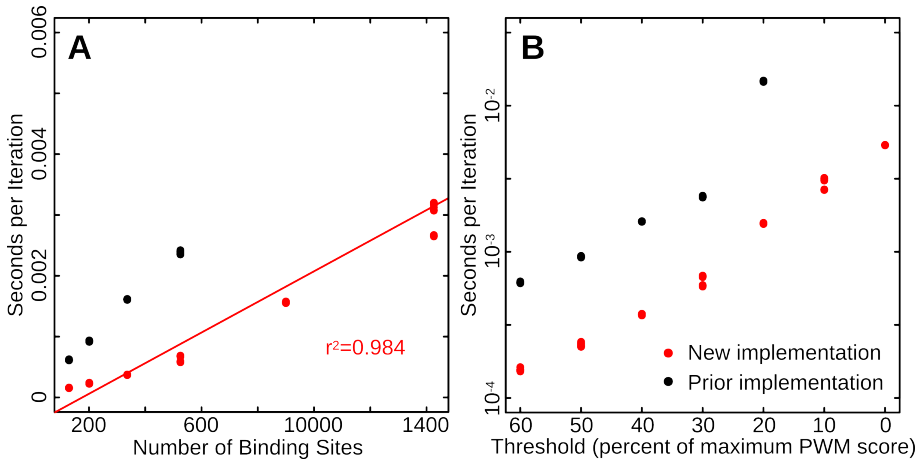


Figure 2.7. Benchmarks for dynamic algorithm Efficiency of new model implementation for the seven construct model presented in Kim *et al.*[93]. (A) The time for a single cost function evaluation grows approximately exponentially with number of binding sites with the previous algorithm (black dots) but linearly with the new algorithm (red dots). Each point is the average time for 100,000 cost function evaluations. There are 10 repetitions for each test. The time for the new algorithm was fit to a linear regression. The regression line is shown (red line) and the value of r^2 is given. (B) The time of a cost function evaluation is plotted as a function of PWM threshold. The cost function evaluation time for the new algorithm (red dots) grows exponentially with lowered thresholds, but is several orders of magnitude faster at low thresholds than the previous algorithm (black dots). Each point is the average time for 100,000 cost function evaluations. There are 10 repetitions for each test.

this requires a large number of functions that update values, and it places a burden on the programmer to know what values must be updated and what functions must be called. While this makes the code more difficult to maintain, there is also a substantial improvement in efficiency.

2.3.4 Efficiency of updated code

The new code, implemented with the above principles is substantially more efficient than previous code. The time per cost function evaluation no longer increases exponentially with the number of binding sites (Figure 2.7A). While computation time does increase exponentially with lowered threshold (Figure 2.7B), this is due to the exponential rise in identified sites as thresholds are lowered. The new code can be extended to thresholds of zero without

causing overflow. At a high threshold of 60% of the maximum PWM score there was an improvement of 3.9 fold, while at a low threshold of 10% of the maximum PWM score the improvement was 61 fold.

2.4 Discussion

Computation efficiency limits the scope or accuracy of problems that can be solved by models of gene expression. For instance, PWM thresholds introduce instability in models, yet were necessary approximations because models with low thresholds were intractable. In this work I have introduced a new body of code that makes substantial improvements in efficiency. These improvements are due to changes in both the algorithm and the implementation. Algorithmic improvements are due to a new dynamic algorithm for the calculation of fractional occupancy that scales linearly, rather than exponentially with the number of binding sites. Improved efficiency in the implementation comes from structure of loops that minimizes loading data into memory and as well as care taken to only recalculate the necessary data structures during parameter optimization. In addition, I have modified to model to include non-specific interactions, and all possible configurations of cooperatively bound pairs of factors. The overall improvement in computational efficiency means that problems that are orders of magnitude larger in scope can be solved, up to a genomic scale.

CHAPTER 3
A SEQUENCE LEVEL MODEL OF AN INTACT LOCUS PREDICTS THE LOCATION AND FUNCTION OF NONADDITIVE ENHANCERS

3.1 Abstract

Metazoan gene expression is controlled through the action of long stretches of noncoding DNA that contain enhancers—shorter sequences responsible for controlling a single aspect of a gene's expression pattern. Models built on thermodynamics have shown how enhancers interpret protein concentration in order to determine specific levels of gene expression, but the emergent regulatory logic of a complete regulatory locus shows qualitative and quantitative differences from isolated enhancers. Such differences may arise from steric competition limiting the quantity of DNA that can simultaneously influence the transcription machinery. We incorporated this competition into a mechanistic model of gene regulation and applied it to the regulation of *Drosophila even-skipped* (*eve*). This model finds the location of enhancers and identifies which factors control the boundaries of *eve* expression. This model predicts a new enhancer that, when assayed *in vivo*, drives expression in a non-*eve* pattern. Incorporation of chromatin accessibility eliminates this inconsistency.

The work presented in this chapter has been submitted to *PLoS Genetics* as a manuscript by myself and John Reinitz.

3.2 Summary

Only a small percentage of metazoan genomes directly codes for protein. It is now clear that much of the remaining non-coding genome is responsible for specifying levels of transcription of nearby genes. The regulatory code controlling this process is as yet poorly understood. Some understanding of this control exists at the level of individual enhancers, short segments

of DNA that direct expression in a particular spatial domain or tissue type. Considerably less understanding exists at the level of an intact genetic locus, the fundamental unit of genetic function. In this paper we show for the first time how the regulation of an entire locus and its structure of enhancers arise from the DNA sequence of the locus and the action of a small set of regulatory mechanisms. We show that this happens as a consequence of competition between enhancers activating transcription, a phenomenon not seen in the study of individual enhancers. Using a mathematical model embodying this and other regulatory mechanisms, we are able to correctly predict the locations and function of enhancers in the *even-skipped* locus of *Drosophila melanogaster*, a locus that has long served as a model system for studies of gene regulation.

3.3 Introduction

Understanding how genetic function arises from the structural properties of genes is a fundamental problem of molecular genetics. With respect to the non-coding portions of genes, in prokaryotes there is a clear relationship between chemical properties and genetic function. In the *lac* operon, for example, there is a one-to-one mapping between the functional genetic unit of the operator and the structural/chemical unit of the binding site for *lac* repressor [80]. This level of understanding is absent in metazoan genes. The expression of many such genes is under the control of *cis*-acting DNA sequence which can span tens[163] to hundreds of thousands[103] of nucleotides. The central feature of such genes is the presence of enhancers, also known as *cis*-regulatory modules (CRMs). These sequences, which typically span 500 to 1000 base pairs (bp), recruit sequence-specific transcription factors to drive a subset of a gene's full expression pattern[59, 50, 175, 176, 197]. Although enhancers are ubiquitous, how they arise from the underlying structure of genes remains obscure.

In this paper we address this problem by showing that under very general assumptions about underlying chemical mechanisms, the physical limitation that only a subset of distally bound transcription factors (TFs) can interact with the basal promoter complex at the same

time induces a modular structure on a genetic locus. We consider a well characterized locus in *Drosophila melanogaster* known as *even-skipped* (*eve*). The enhancer structure of this gene has been exceptionally well characterized experimentally [59, 50, 175, 176], and quantitative chemical models of the function of these enhancers are now well known [81, 171, 92, 70, 165, 93, 122].

The use of theoretical models in these studies is required because of the complexity of the chemical mechanisms underlying gene regulation. Experimental assays permit the dissection of a gene into its constituent parts and allow the properties of these parts to characterized in isolation. Models allow us to assay whether or not well defined interactions of these components give rise to the observed behavior of the intact system, and thus provide a minimal set of mechanisms required for understanding the biological phenomenon at hand. Theoretical models of whole loci have been constructed by assuming an underlying modular structure of enhancers and reconstructing the whole locus expression pattern from a weighted sum of of outputs of individual enhancers [164], but this does not address how this modular structure arises from underlying chemical interactions. Moreover, the fact that most developmental genes contain shadow enhancers [27, 149, 150, 46] that behave nonadditively [40, 24] suggests that important regulatory mechanisms exist at the level of an intact locus that are not seen in isolated enhancers.

In this work, we construct a quantitative theory to explore the consequences of steric limitations on the amount of transcription factors that can simultaneously interact with proximal transcription complex. This form of enhancer competition is added to a previously existing model of gene regulation[93, 123, 122]. When applied to the *Drosophila eve* locus, which drives seven transverse stripes across the syncytial blastoderm, the model is able to fit the expression pattern and discover the factors that form each stripe. Furthermore, the underlying enhancer structure of the locus emerges from the internal structure of the model when it is fit to data, without such structure being imposed in the assumptions of the model. The model also shows the importance of chromatin accessibility in driving gene

expression. Without consideration of such accessibility, the model predicts a new enhancer in the *eve* locus that, when assayed *in vivo*, drives expression in a non-*eve* pattern. Chromatin accessibility assays suggest that this fragment is inaccessible *in vivo* [106, 126]. A model that incorporates this accessibility data does not predict expression driven by this fragment within the intact locus.

3.4 Results

3.4.1 Sequence level model without enhancer competition and eve *expression*

In previous work, we generated a model of gene regulation that computes transcription rate from DNA sequence, transcription factor concentrations, and DNA-binding preferences in the form of position weight matrices (PWMs)[158, 81, 93, 123, 122]. In this model, we first calculate equilibrium transcription factor occupancy using thermodynamics. This calculation incorporates cooperative binding and repression through steric competition for binding. Second, we calculate context dependent switching between repressing and activating states, known as coactivation, wherein proteins activate only when bound in proximity to a bound coactivator. Third, we calculate the repressive effects of short-range quenching. Fourth, we calculate the number of transcriptional adaptors, proteins which interact with both DNA-bound transcription factors (TFs) and the transcription machinery [197], recruited to an enhancer by a weighted sum that represents the efficiency of adaptor recruitment for each activator. Finally, we treat these adaptors as catalysts that reduce the energy barrier to transcription and describe this in the form an a diffusion-limited Arrhenius rate law.

To test the ability of this model to describe the regulation of an entire locus, we applied this model to the *even-skipped* (*eve*) gene of *Drosophila melanogaster*. Confocal microscopy in *melanogaster* embryos has allowed quantification of both transcription factor levels and mRNA levels at single nucleus resolution along the anterior-posterior axis [185, 184]. These

data amount to a set of quantitative single cell assays of transcription input and output in a native tissue context, providing an extraordinarily precise testbed for theoretical models.

We attempted to model the whole locus behavior of *eve* by two methods. First, we trained the above model on levels of *eve* mRNA from 35.5% to 92.5% embryo length, encompassing stripes 2 through 7, driven by 13,150 bp of *eve* DNA extending from 4730bp upstream to 8420 bp downstream of the transcription start site (TSS). This DNA is sufficient to drive the early seven stripe pattern [59, 163]. The model was able to drive the desired pattern from the regulatory sequence used (Fig 3.1A), but when the parameters learned from this fit were confronted with smaller segments of sequence corresponding to the enhancers for each of the six stripes, none of the enhancers were predicted to drive expression (Fig 3.1B-E). Similarly, when we trained the model simultaneously on each individual enhancer driving its respective stripe pattern, we are able to achieve good fits (Fig 3.1F-I), but the parameters obtained predict that the intact locus will drive saturating expression across the entire embryo (Fig 3.1J). No fits were able to simultaneously describe the the action of both the intact locus and individual enhancers, leading us to conclude that at least one additional regulatory mechanism emerges at the level the intact locus and is necessary to model its behavior.

3.4.2 An enhancer competition model

One potential issue is the implicit assumption that factors bound to the entirety of modeled DNA simultaneously influence a promoter. Instead, only a finite length of α bp of DNA can simultaneously influence a gene's promoter within a short timespan. We expect activators bound to DNA on scales smaller than this length to synergistically activate transcription through cooperative action on the basal transcription machinery, while activators separated by larger scales will compete for promoter occupancy. There has been much focus on "minimal" enhancers—the smallest segments that are able to recapitulate a pattern *in vivo*—but much larger sequences may be able to influence a promoter. Indeed, we find that when the 480 bp minimal stripe 2 element of *eve* (MSE2) is extended by 219 bp there is a five fold

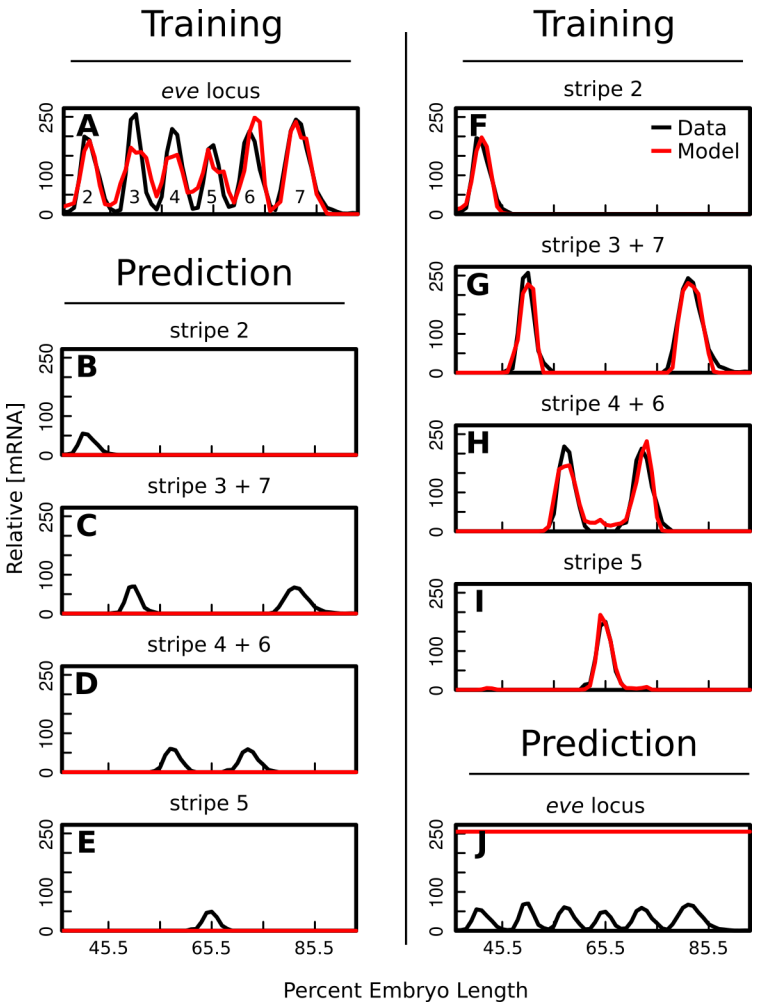


Figure 3.1. Model fits without enhancer competition. (A) The transcription model, given by Eqs 2.1-2.25 and Eq 3.3, was trained to the expression pattern of *even-skipped*. Percent embryo length(x-axis) is measured from the anterior pole. The identity of each *eve* stripe is indicated. The model (red line) is able to achieve good fits to data (black line). (B-E) Using the model shown in A, we predicted the [mRNA] driven by four enhancers that have previously been shown to drive each of the stripes (red lines). The identity of each sequence is labeled. Sequence coordinates for each enhancer are reported in Materials and Methods. The locus data that corresponds to each stripe is shown with black lines. (F-I) We trained the model to the four *eve* enhancers driving their respective portion of the locus pattern. This model output (red lines) achieves good fits to data (black lines). (J) We used the model shown in F-I to predict expression driven by the entire *eve* locus. Predicted output (red line); Data (black line).

increase in transcription rate (Fig 7.1). This suggests that sequences of up to $\alpha = 1$ kb are able to simultaneously influence a promoter, and we use this value for the rest of this work. However, the final results were completely insensitive to setting $\alpha = 500$ (Fig 7.2 and File 7.3).

While we expect activators bound within a region smaller than 1kb to synergistically activate transcription, how disparate elements compete for access to a promoter is currently unknown. Recently, it has been shown that transcription driven by *Drosophila* developmental enhancers occurs in bursts [23] and that forced enhancer-promoter looping in murine cell lines indicates that the frequency of bursts is determined by the frequency of interaction with a promoter[14]. Collectively, this demonstrates that transcription rates can be controlled at the level of burst size or burst frequency and these quantities correspond to the rate of transcription induced by enhancer-promoter interactions and the frequency of such interactions respectively. For tested *Drosophila* enhancers, these quantities are highly correlated [51]. Thus, we propose that the frequency of enhancer-promoter interaction and the rate induced by such interaction is proportional to the number of transcriptional activators bound to a DNA segment.

Specifically, we imagine that, for any DNA segment bounded by base pairs $[m, m+\alpha]$, where α, introduced above, represents the length of DNA that simultaneously influence the promoter (the "window size"), that $N_{[m,m+\alpha]}$ transcription adaptors are recruited (For calculation of N, see [93] and Chapter 2. The rate of mRNA synthesis, $R_{[m,m+\alpha]}$, driven when the segment interacts with the promoter is given by a diffusion-limited Arrhenius rate law

$$R_{[m,m+\alpha]} = \frac{R_{\max}}{1 + \exp(\theta - N_{[m,m+\alpha]})}, \tag{3.1}$$

where we assume without loss of generality that a single bound coactivator lowers the Arrhenius energy barrier, ΔA, to transcription initiation by one unit . The free parameter θ is the total energy barrier which sets the rate of transcription in the absence of activation. The scale of both N and θ are effectively set by the fit to data.

For a locus of length l, the fraction of time that any DNA segment $[m, m+\alpha]$ influences the promoter is given by

$$T_{[m,m+\alpha]} = \frac{\beta N_{[m,m+\alpha]}}{1 + \sum_{n=1-\alpha}^{l} \beta N_{[n,n+\alpha]}}, \tag{3.2}$$

where the free parameter β determines how much individual bound adaptors increase the frequency of interaction with the promoter. Note that the summation in the denominator is taken over every base position in the locus. The total rate of transcription driven by the locus is then given by the frequency-weighted sum of transcription due to each DNA segment $[m, m+\alpha]$, so that

$$R_{\text{total}} = \sum_{m=1-\alpha}^{l} R_{[m,m+\alpha]} T_{[m,m+\alpha]}. \tag{3.3}$$

Again, the summation occurs over all possible α subsequences of the *eve* locus iterated in single nucleotide increments. The half life of *lac*Z and *eve* mRNA is short compared to the timescale of changes in gene expression, so that

$$\frac{d[\text{mRNA}]}{dt} \propto [\text{mRNA}], \tag{3.4}$$

an observable quantity.

The calculation of transcription factor occupancy with full thermodynamics, which is used to calculate N, requires enumeration of all possible binding states. In previous work this was done using an explicit calculation on each configuration [93]. Such a calculation scales with 2^n where n is the number of binding sites on a sequence. When performing calculations on the entire locus, we identified 2920 binding sites with a log-odds score greater than 0—the threshold used for calling binding sites in this work. Explicit calculation of 2^{2920} states is computationally infeasible. In Chapter 2, I developed a new algorithm that uses dynamic programming. This new algorithm makes calculation of occupancy in the entire locus a tractable problem.

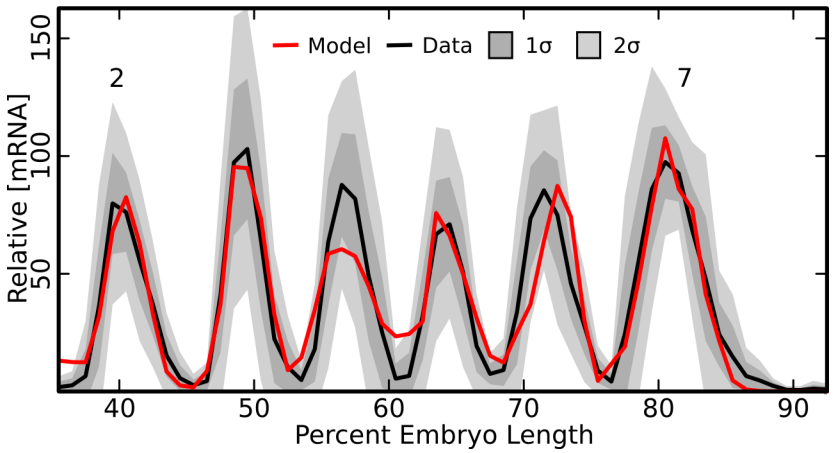


Figure 3.2. Model with enhancer competition trained on the *eve* locus. Observed mRNA levels (black line) are shown together with model output (red line). One (dark grey shading) and two (light grey shading) standard deviations about the mean of the data are shown. Data comes from 7 embryos for a total of 19 to 30 nuclei per embryo position. The axes are labeled; percent egg length is measured from the anterior pole.

3.4.3 Enhancer competition and eve *expression*

We trained the free parameters in the model given by Eqs. 2.1-2.25, Eqs. 3.1-3.2, and 3.3 to the expression of the *eve* locus from 35.5% to 92.5% embryo length, using the 13kb sequence described previously. We omitted stripe 1 from this study because its anterior border is controlled by transcription factors for which we do not have data. Additionally, anterior of stripe 1 the clean functional distinction between AP and dorsal-ventral (DV) patterning breaks down, and data along a single axis is inadequate.

The model was able to achieve a good fit to the expression pattern of *eve* stripes 2-7 (Fig 3.2). Specifically, the model was within two standard deviations of the data everywhere except at the 1-2 and 4-5 interstripes, and within one standard deviation of the data except at the two locations mentioned as well as at the peak of stripe 4, which is smaller than the data, and the margins of stripe 6, where the model produces a stripe displaced one nucleus to the posterior. Interestingly, the lag in the position of stripe 6 is consistent with the lag observed from the stripe 4+6 enhancer [179] indicating there may be reasons for the discrepancy between the enhancer and locus that are outside the scope of this model.

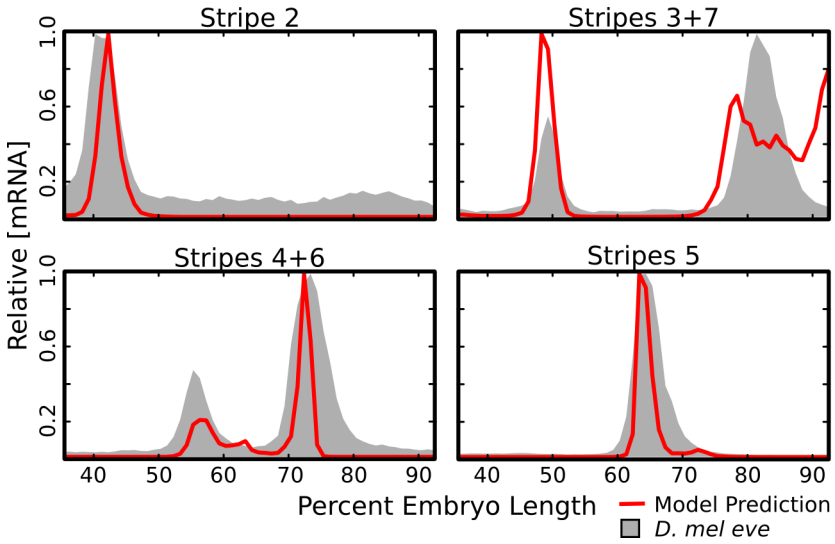


Figure 3.3. Predicted output of known *eve* enhancers. We used the trained model to predict the the transcription rate driven by four previously reported enhancers of *eve*. For each enhancer the predicted output was standardized such that 1 represents the maximum rate driven by that enhancer (red lines). The relative mRNA driven by each enhancer (gray shading) was obtained from Staller *et al.*[179]. This data, also standardized, is included for visual orientation within the embryo and levels are not commensurate with predicted enhancer output.

3.4.4 Identification of eve *enhancers*

The *de novo* identification of enhancer locations and activity is a major goal of gene regulatory models. We tested the ability of the model to identify known enhancers in two ways. First, we used the trained model to simulate the activity of known enhancers of *eve* *in silico* (Fig 3.3) and compared this to quantitative data on the expression driven by each enhancer[179]. Each is correctly predicted to drive expression of its corresponding stripes. Quantitatively, there are some discrepancies. For the enhancer of stripe 3+7 we predict reduced output from stripe 7, which is consistent with the initial reports on this enhancer[176], but not with quantitative data. We predict poor anterior repression of stripe 7 when driven by the 3+7 enhancer. Additionally, we observe weak expression from stripe 4 when driven by the 4+6 enhancer. Generally, the predicted expression patterns are narrower than observed patterns.

Similarly, we looked at expression contributions across the entire *eve* locus by looking

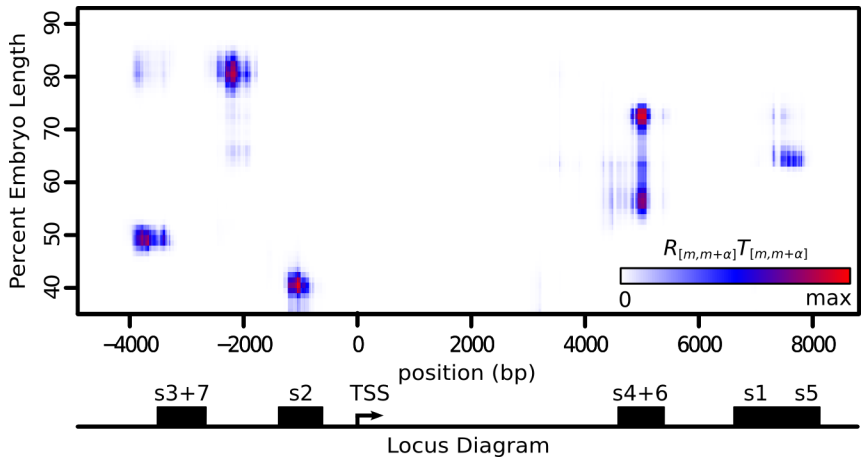


Figure 3.4. Expression contribution over space sequence and embryo length. We report a heat map of the quantity $R_{[m,m+\alpha]}T_{[m,m+\alpha]}$ (Eq. 3.3), which represents the amount each 1kb sequence, centered on base $m + \alpha/2$ (x-axis), contributes towards total expression at each position in the embryo (y-axis). The color scale is standardized to the range of the data. The x-axis is labeled with a map of the *eve* locus, displaying the transcription start site and locations of previously identified enhancers (black rectangles).

at the rate driven by every individual 1kb subsequence (Fig 3.4). We find that for stripes 2 through 6, the majority of activation is result of tightly clustered groups of sequences that have high overlap with the locations of previously reported enhancers. While stripes 2 through 6 have single clusters that drive their expression, we find that stripe 7 is driven not only by the stripe 3+7 enhancer, but also by DNA that lies 5' of the stripe 2 enhancer. Expression driven by parts of this region have previously been reported for constructs that contain varying lengths of DNA 5' of the stripe 2 enhancer[81, 179] and explains why deletions of the stripe 3+7 enhancer lead to loss of stripe 3, but not stripe 7[59].

3.4.5 Control of eve *stripe domains*

Three lines of evidence have been used to establish which factors control the boundaries of *eve* expression domains—mutations in *trans*, mutations in *cis*, and regulatory models—carried out in either the intact locus or enhancer-reporter constructs. In the best cases there is agreement between these techniques, for example in Giant (Gt) null embryos the anterior border of stripe 2 expands when driven by both native *eve* [48] and by a reporter for the

proximal 2.9kb of the locus[177]. Similarly, there is a stripe 2 expansion when Gt binding sites were removed from reporters for either the proximal 5.2kb of *eve*[180] or MSE2[175]. Collectively, these experiments provide strong evidence that Gt is responsible for forming the anterior boundary of *eve* stripe 2. Additionally, models of gene regulation have identified Gt as a key regulator of stripe 2 in both the locus[159, 164] and enhancers[81, 122].

For other *eve* borders there is conflicting evidence. For instance, in Kruppel (Kr) null embryos[48] the posterior of native *eve* stripe 2 expands, but the domain driven by MSE2 does not[175] indicating that other factors may contribute to this stripe border. Similarly, in Knirps (Kni) null embryos or after deletion of Kni sites, the minimal stripe 3 enhancer (MSE3) does not form a posterior border[176, 183], however stripe 3 forms normally in the intact locus[59, 48, 176]. Finally, the anterior border of stripe 7 appears to be regulated by Kni [176, 31, 183] when stripe 7 expression is driven by MSE3, or by Gt when expression is driven by the whole locus [48] or by an *eve* 2+7 enhancer [179, 81].

In each of the above cases there are conflicting results from experiments where expression is driven by separate enhancers compared to those in which it is driven by the intact locus. In order to resolve these conflicts we identified the factors responsible for stripe boundaries in both the locus and individual enhancers in a single, unified, model. Given a trained set of model parameters, we are able to quantitatively decompose the change in [mRNA] in adjacent nuclei into the effects due to the changes in concentration of each transcription factor in both the locus (Fig 3.5B) and individual enhancers (Fig 3.6). Within the locus, in wild type *D. melanogaster*, we find that single transcriptional repressors are responsible for forming the boundaries of each stripe (Fig 3.5B, summarized in Fig 3.5C). For the factors forming the borders of stripes 4 through 6, the model identifies the same factors (Fig 3.5B and Fig 3.6C-D) that have been previously identified through experiment[50, 31]. In agreement with previous literature[48, 175, 177, 180, 159], we find that Gt sets the anterior border of stripe 2 and that Kr defines the posterior boundary of that stripe in the intact locus. In contrast to the locus, we find that there is a significantly larger contribution from declining

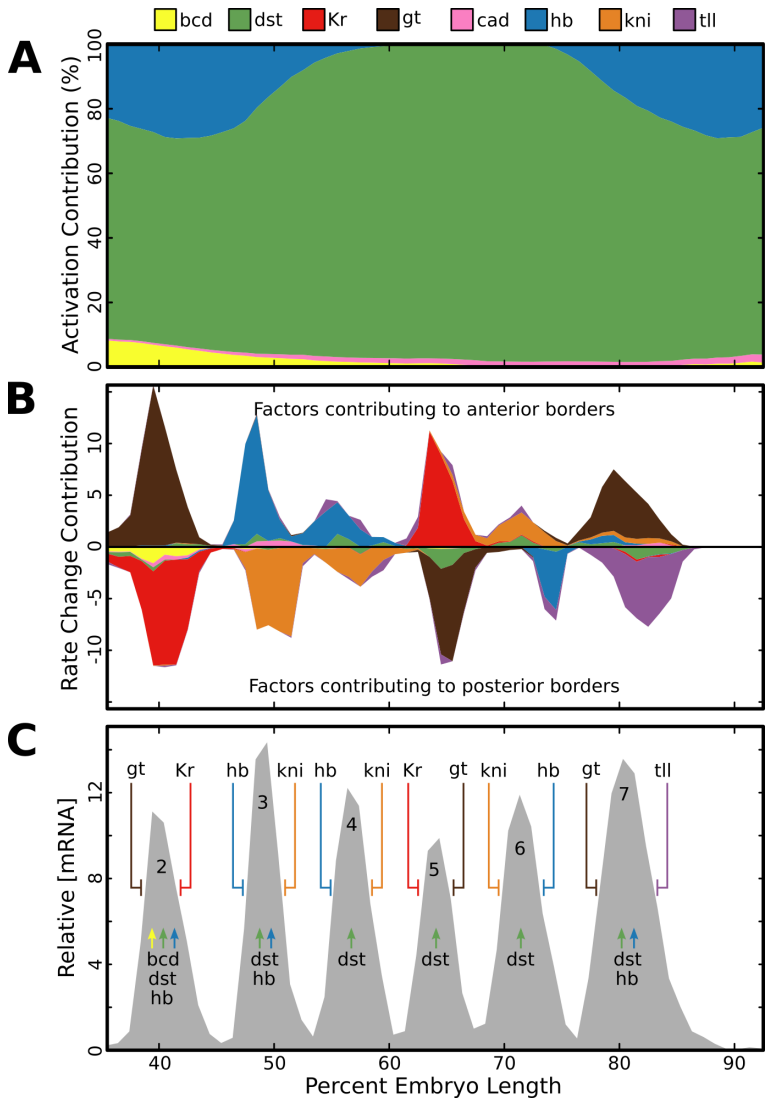


Figure 3.5. Mechanisms of activation and repression in the locus. (A) Cumulative line graph showing the amount of *eve* mRNA attributable to each TF (y-axis) at each embryo position (x-axis). We calculated the percent of transcriptional adaptors N that are recruited by each TF to the transcription machinery at each embryo position (x-axis) and scaled total output by this value. For calculation, see Materials and Methods. (B) Cumulative line graph showing the change in [mRNA] caused by a change in concentration of each TF (y-axis) at each embryo position (x-axis). The total sum gives the the change in [mRNA] at each embryo position. Thus, factors which contribute to anterior borders give positive values and those that contribute to posterior borders give negative values. For calculation, see Materials and Methods. (C) A summary of the factors responsible for each expression feature of *eve* as determined by A and B. Activators are indicated by arrows and repressors by T-bars.

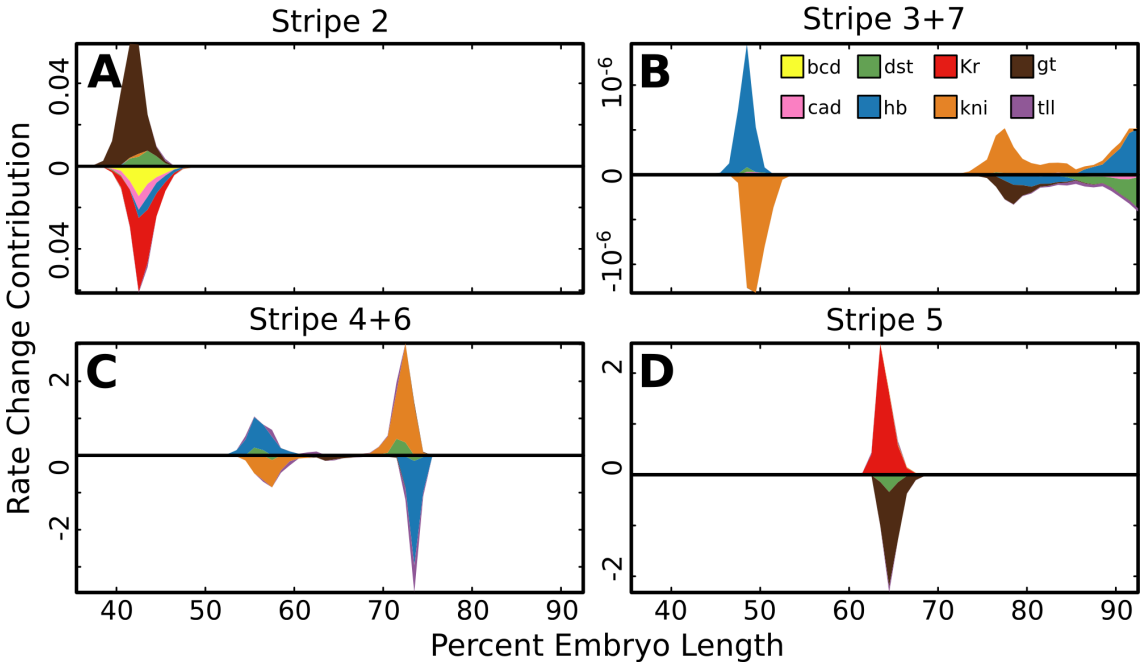


Figure 3.6. Mechanisms of repression in enhancers. For four previously reported *eve* enhancers the predicted contribution of each TF to a change in [mRNA], along the AP axis, was calculated as in Fig 3.5B and described in Materials and Methods.

Bicoid (Bcd) and Hunchback (Hb) levels on MSE2 (Fig 3.6A), which potentially explains why expression driven by MSE2 does not shift to the posterior in Kr null embryos[175].

Next we examined the regulation of stripes 3 and 7. We find that in the intact locus, stripe 3 has anterior and posterior borders set respectively by Hb and Kni in both the intact locus (Fig 3.5B) and in the stripe 3+7 enhancer (Fig 3.6B). This result is consistent with previous reports[176, 183, 31], but falls short of explaining how stripe 3 forms in Kni mutants[48, 176]. We do not detect a contribution from Kr as suggested by a previous model[159]. For stripe 7 we find that Gt sets the anterior border in the intact locus, but we also find that that Kni sets this border when expression is driven by the stripe 3+7 enhancer. Similarly, we find that the posterior border of stripe 7 is primarily set by Tailless (Tll) repression when that stripe is driven by the whole locus. but that the posterior border of stripe 7, when driven by the 3+7 enhancer, is set by Hb.

These numerical results are reminiscent of a recent experimental result showing that the locus and 3+7 enhancer respond differently to the ectopic expression of Hb driven by the

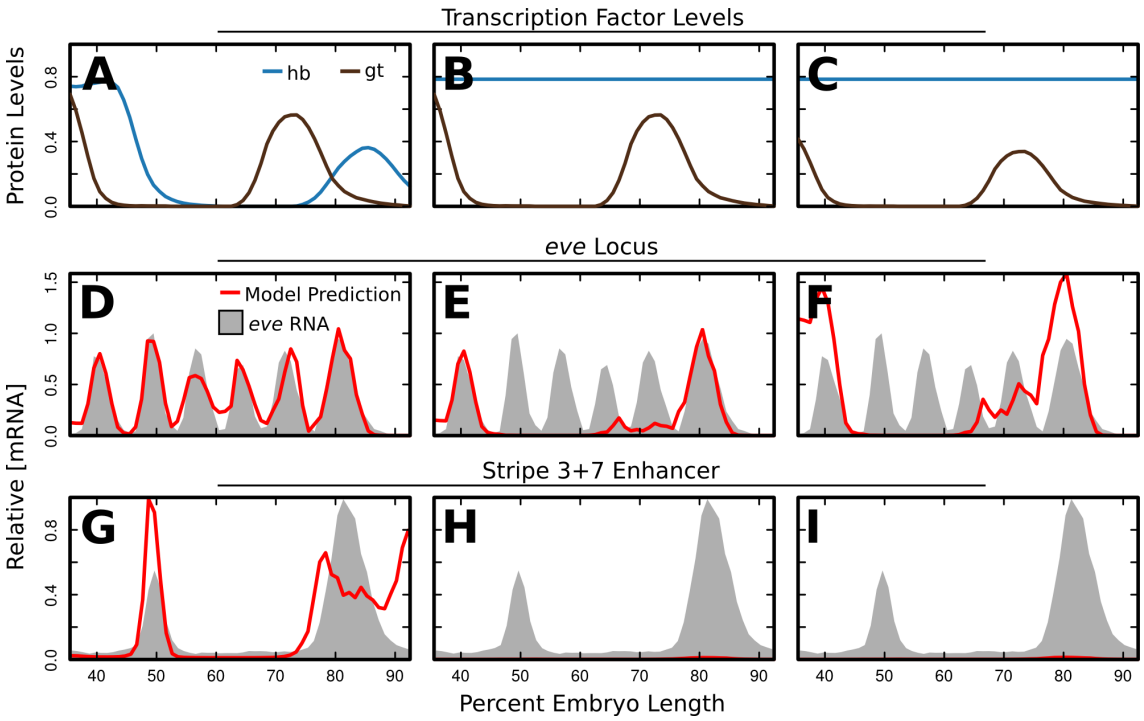


Figure 3.7. Predicted effects of ectopic Hb. (A) The measured relative levels of Hb and Gt (y-axis) from 35.5% to 92.5% embryo length (x-axis)[153, 184, 185, 152]. (B) Simulated relative levels of Hb and Gt. Hb is set to a spatially uniform value and Gt is unchanged from A. (C) Simulated relative levels of Hb and Gt. Hb is set to a spatially uniform value and Gt is reduced by 40%. (D-F) Predicted relative [mRNA] levels (red lines) driven by the *eve* locus under the TF levels indicated in A-C. Model output is standardized to the maximum rate driven by the locus in the wildtype *trans* environment. Data for relative [mRNA] of *eve* (gray shading) is included for visual orientation within the embryo and levels are not commensurate with predicted locus output. (G-H) Predicted relative [mRNA] levels (red lines) driven by the *eve* Stripe 3+7 enhancer under the TF levels indicated in A-C. Model output is standardized to the maximum rate driven by the enhancer in the wildtype *trans* environment. Data for relative [mRNA] driven by the stripe 3+7 enhancer (gray shading) is included for visual orientation within the embryo and levels are not commensurate with predicted enhancer output.

snail promoter[179]. Under ectopic expression of Hb, stripe 7 is lost when driven by the stripe 3+7 enhancer, however when driven by the intact locus, stripe 7 is not lost and expression expands towards the anterior. Ectopic Hb leads to complex changes in the *trans* environment [205] and specific levels of transcription factors are unknown, however we are able to simulate changes in *trans* to test whether our model is consistent with these results. To this end, we set Hb levels to a spatially uniform value (Figure 3.7B). We find that expression driven by the 3+7 enhancer is lost (Figure 3.7H), but stripe 7 is not lost when driven by the entire *eve* locus (Figure 3.7E). We do not observe the anterior expansion of stripe 7 when only Hb expression is changed, but ectopic expression of Hb has pleiotropic effects which act to reduce levels of both Gt and Kni in the posterior of the embryo [205]. A reduction in the level of Gt (Figure 3.7C) in addition to ectopic Hb is sufficient to drive the anterior expansion of this stripe (Figure 3.7F).

3.4.6 Activation by Hunchback and Stat92E

It has long been recognized that *eve* is activated by broadly distributed factors [159, 175, 176]. Our model included three transcriptional activators: Bcd, Caudal (Cad), and Stat92E (Dst). Additionally, the repressor Hb is able to activate when bound near Bcd or Cad [65, 174, 93], a phenomenon called coactivation. In order to determine which factors are responsible for activation we found the percent of adaptors recruited to the TSS by each transcriptional activator (Fig 3.5A). We find that the majority of activation is driven by Stat92E, with a significant contribution from Hb in the anterior and posterior portions. While we do not observe large direct contribution from Bcd and Cad, these factors are responsible for the activating activity of Hb through coactivation.

3.4.7 Behavior of a predicted cis*-regulatory element*

Our results indicate that most of the activation of stripe 7 is driven by a sequence upstream of the stripe 2 enhancer, between that and the stripe 3+7 enhancer elements (Fig 3.4). We

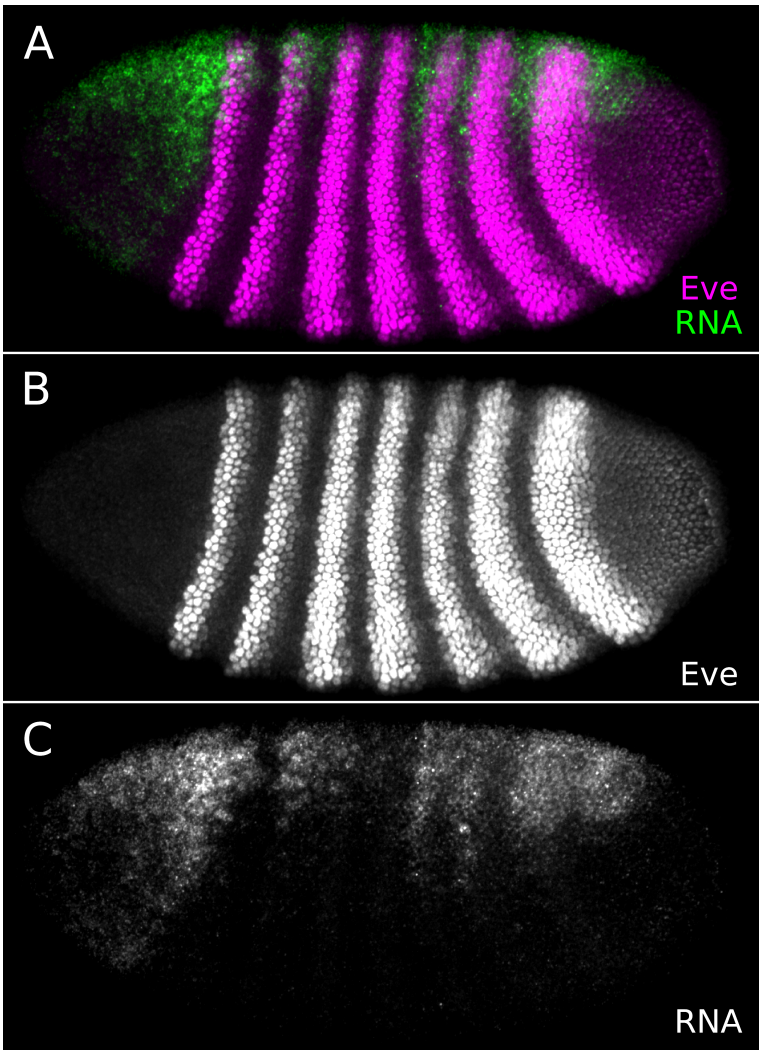


Figure 3.8. Expression driven by a predicted enhancer. A 900bp sequence, located between 3130 and 2230 bp upstream of the *eve* TSS, was placed upstream of a lacZ reporter. An embryo containing this construct at the AttP2 site[63] was stained by FISH and immunostaining with antisense lacZ probe and α–Eve antibody[95] respectively. The embryo was imaged in late nuclear cycle 14 with a 20x objective on a Zeiss 710 confocal microscope. Brightness and contrast have been increased uniformly across images. (A) Eve (magenta) and lacZ (green) (B) Eve in grayscale (C) lacZ in grayscale

took a 900bp fragment, located between 3130 and 2230 bp upstream of the TSS and centered on this region, and tested its ability to drive expression of *lac*Z *in vivo*. This sequence, which we call the 3130 element, drives expression dorsally overlapping stripe 2 and stripes 5 through 7 (Fig 3.8A,C). This fragment drives stronger and more ventral expression within the posterior interstripes than in the stripes themselves. Remarkably, this pattern is not observed in reporter assays for larger sequences that contain the 3130 element[59].

3.4.8 Incorporation of chromatin state

It is possible that the assay for the 3130 element is not faithful to *in vivo* expression because this fragment has been removed from its native chromatin state. Indeed, the 3130 element falls into inaccessible chromatin when assayed using either DNAse-seq [106] or FAIRE-seq [126]; moreover models of binding trained with DNAse-seq and ChIP-seq data do not predict binding in this region[90]. In order to incorporate this information, we defined accessible nucleotides to be those that are within accessible regions found using either DNAse-seq or FAIRE-seq(Fig 3.9B). Then we retrained the model, this time only scanning for transcription factor binding sites that were within accessible chromatin. After training, the best parameter set generated an equally good fit to data as those that did not incorporate chromatin status (Fig 3.9A). We no longer find expression driven by the 3130 element within the context of the intact locus (Fig 3.9B), but this fragment is still predicted to drive expression when removed from its native chromatin context (Fig 7.3). The DNA regions that contribute to activation overlap with their corresponding enhancers (Fig 3.9B), and when we simulated the activity of known enhancers *in silico* each enhancer is still correctly predicted to drive expression of its corresponding stripes (Fig 3.9C). The identified mechanisms of stripe border control (Fig 7.4) and predicted effects of ectopic Hb (Fig 7.5) did not change after inclusion of chromatin accessibility data.

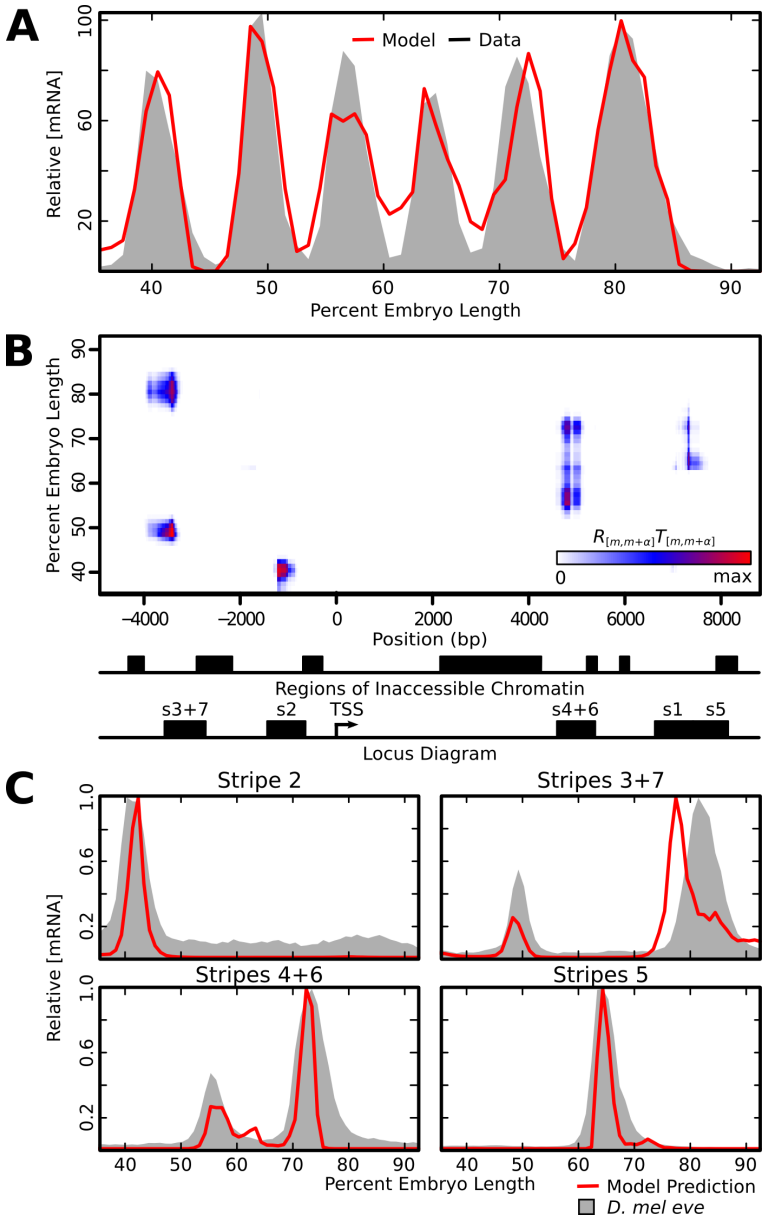


Figure 3.9. Model output after masking inaccessible chromatin. We excluded transcription factor binding within regions identified as inaccessible and retrained model parameters. (A) Observed mRNA levels (gray shading) are shown together with model output (red line) after inclusion of a chromatin mask. (B) Heatmap of the quantity quantity $R_{[m,m+\alpha]}T_{[m,m+\alpha]}$ at each nucleotide and embryo position, representing the amount each 1kb sequence, centered at that nucleotide, contributes towards total expression. The identified regions of inaccessible chromatin and locations of known enhancers are indicated on the x-axis. (C) We tested the relative output of the known *eve* enhancers *in silico* using the retrained model (red lines). The relative mRNA driven by each enhancer (gray shading), is included for visual orientation within the embryo and levels are not commensurate with predicted enhancer output.

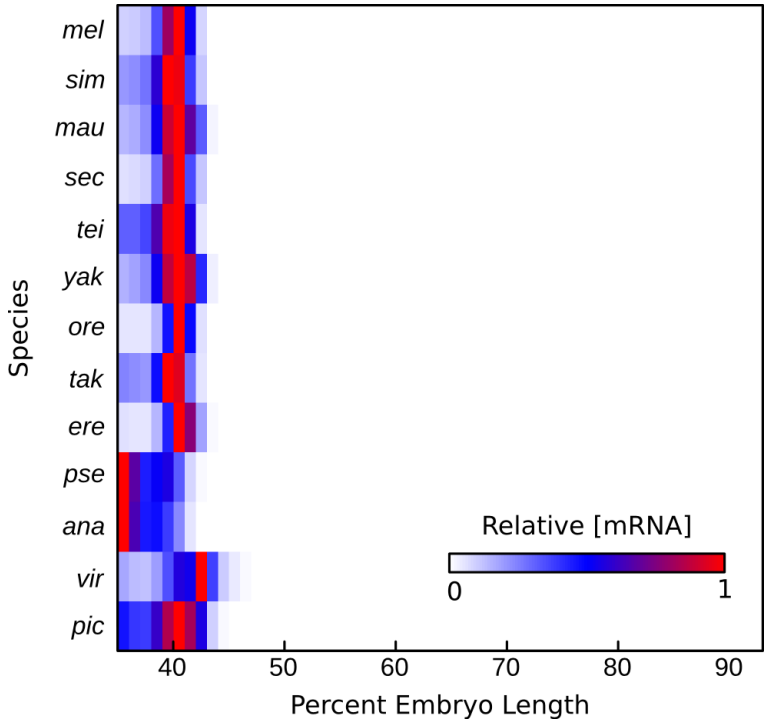


Figure 3.10. Model prediction of S2Es. The predicted relative activity of 13 *Drosophila* stripe 2 enhancers, previously reported in Kim *et al.*[93], is shown on a heatmap with each row representing an S2E from a different species over the AP axis. The predicted output of each enhancer is standardized on a 0 to 1 scale. The color represents the relative mRNA driven by that species at that embryo position.

3.4.9 Model predicts changes in cis

Enhancers of *Drosophila eve* are notable in that there is functional conservation in the absence of sequence conservation[113, 112, 68, 67, 122]. To test the ability of the trained model to predict the effects of chagnes to sequence, we tested the ability of the model to predict the function of S2Es from various drosophilids. Kim *et al.*[93] defined 13 drosophilid stripe 2 enhancers by homology. Using the sequences reported in that work, we tested the ability of the model with chromatin included to predict the function of these sequences. 11 of 13 sequences are predicted to drive proper expression of a stripe at 40% embryo length (Fig 3.10).

3.5 Discussion

The central result of this paper is the demonstration that the enhancer structure of *eve* arises because of competition between different regions of the proximal promoter for interaction with the basal complex (Eq. 3.2). The competition described may reflect kinetic statistics of interactions between distally bound adaptors and the basal complex. This competition differs from steric competition for a binding site in that $N_{[m,m+\alpha]}$, unlike q_i (Eq. 2.15), depends not only on thermodynamically described interactions of TFs with the DNA but also on the protein-protein interactions which convert repressors into activators by coactivation and quench activators (Eqs. 2.22 and 2.24).

Previously, the independent action of enhancers has been explained by quenching. This short range repression mechanism allows expression to be driven by one enhancer while transcriptional repressors bind to quenched enhancers only a few hundred nucleotides away[61]. This mechanism is indeed necessary to explain the action of *eve* enhancers, but if repression occurs over short distances then low levels of bound activators over sufficiently long pieces of DNA will eventually overcome repression. Some additional mechanism must exist to prevent this domination of activation over repression. Short range repression, together with

the competition of activators for interaction with the basal complex is sufficient to explain the independent action of *eve* enhancers in the context of the whole locus. Furthermore, such mechanisms may explain the nonadditive effects observed for shadow enhancers [40, 24] which are now known to be a common feature of developmentally important genes [27].

Advancements over previous work. One previous work has modeled the regulation of *eve* by its entire regulatory sequence [164]. These authors devised a two tiered model, in which the lower tier is a previously reported enhancer model [70] which uses a thermodynamic picture of protein binding, and is capable of modeling short range repression but not coactivation. The starting parameters of this lower tier component are determined by fitting to a set of expression data from approximately 40 enhancers from 27 genes, excluding the one to be modeled. In the second tier of the model, a collection of up to 5 DNA segments ("windows") for each expression domain is constructed as follows. Every possible DNA segment in the locus with starting points at 100 bp intervals and lengths between 500 and 2500 bp is considered. For each expression domain or stripe the 5 segments that give the best pattern for that domain are chosen. A model of the whole locus is then constructed from a weighted sum of the expression patterns driven by the segments chosen, and then the first tier parameters are retrained while keeping the DNA segments and their weights fixed. This cycle of training window weights and first tier parameters is continued until the score ceases to improve.

The chief difference between the work reported here and that reported by Samee and Sinha [164] is that those investigators started with the assumption that genes have an enhancer structure. This assumption had two consequences. First, the lower tier model of individual enhancers [70], alluded to above, is the starting point and an integral component of the model of the whole locus. The lower tier model was trained on expression data from isolated enhancers. Our previously reported models of isolated enhancers were not used to construct the whole locus model reported here, nor was expression data driven by isolated

enhancers used for training.

Second, the second tier of the previously reported model assumes a one-to-one mapping between contiguous segments of DNA and expression domains, a point that is integral to the fitting procedure described above and the weighting of expression driven by DNA segments. The weights were constant over the whole embryo and only assigned to the five segments of DNA which best matched expression domains. In this work, the weighting is done not in terms of expression domains in the embryo but rather in terms of activation on the distal promoter. This is done in such a way that strongly activating distal promoter regions have stronger interactions with the basal promoter. As a consequence, the relative contributions of individual segments varies from cell to cell as the concentrations of TFs vary. This leads to competition that extends to the interstripes, and may be a reason why expression in the interstripes is higher in the previously reported work, extending to about three quarters of peak stripe expression in the interstripe between stripes 2 and 3 [164, Fig. 4]. Moreover, in this work the weighting by activation is always performed at single nucleotide resolution over the whole locus (Eqs. 3.2 and 3.3) rather than being limited to five segments of DNA. Another difference between the models is with respect to coactivation, which we comment on below. Although the different treatment of coactivation affects biological conclusions about *eve*, this difference arises from prior work by both groups at the enhancer level [70, 93].

Transcriptional regulation of *eve*. A locus level understanding of gene regulation is complicated by the context dependent action of transcription factors. It has previously been shown that ectopic expression of Hb leads to the loss of *eve* stripe 7 when driven by the stripe 3+7 enhancer, but not when driven by the locus[179]. Our model includes a coactivation mechanism, where locally bound Bcd and Cad cause Hb to switch from repressor to activator [65, 174, 93]. This coactivation is required for the activation of *eve* stripe 7 within the posterior Hb domain. In the model reported here, higher spatially uniform levels of Hb expression, which presumably mimic the reported results[179], repress the stripe 3+7

enhancer by providing additional quenching from Hb sites distant from bound Cad (Fig. 3.7H). However, the locus is still able to drive stripe 7 through the action of DNA upstream of the stripe 2 enhancer. These results indicate that coactivation by Bcd and Cad is sufficient to explain dual regulation by Hb. This mechanism was not treated by Samee et al. [164].

We find that Dst has a major contribution towards the activation of *eve*. Evidence in favor of this finding is afforded by the observations that stripe 3+7 expression is reduced by Dst binding site mutations[183] and that *eve* RNA levels drop by a factor of greater than 6 in embryos that lack maternal Dst[195]. Some ambiguity remains, however. Embryos lacking maternal and zygotic Dst still express seven *eve* stripes when driven from the intact *eve* locus [74], presumably at reduced levels. However, these embryos fail to drive stripe 3 from the proximal 5.2kb of the *eve* promoter. In addition to highlighting another difference between fragments driving reporter expression and the intact locus, these results indicate the likely presence of other widespread activators. Possible candidates include Zelda[107], Trithorax-like[144] and Dicheate[131, 162]. Of these, Zelda has reported to act through modification of chromatin state[166, 169], which we have treated directly. In this work we did not include these additional wide spread activators because their functional roles cannot be distinguished without further experimental information. Such experimental information might take the form of quantitative assays of *eve* expression in embryos lacking maternal and zygotic contributions of each of these factors in various combinations. Alternatively, a defined synthetic promoter could be built up by systematically adding binding sites for one such factor at a time.

We find that enhancers do not necessarily follow the same regulatory logic as the intact locus. When a sub-sequence of the intact locus is placed into an enhancer-reporter assay it is removed from the context of the locus. The enhancer may not contain all the sites or regulatory interactions present in extended sequence and thus will not follow the same logic as the locus. In this work we identify a specific case with regards to the regulation of stripe 2. When driven by MSE2, the posterior border of stripe 2 is more strongly regulated

by Bcd and Hb than when stripe 2 is driven by the intact locus , which explains the lack of posterior expansion when Kr binding is disrupted [175]. This is consistent with previous reports on other enhancers. For instance, in Kni– embryos the posterior of stripe 3 and the anterior border of stripe 6 are abolished when expression is driven by MSE3, while these borders remain present when driven by the intact locus[48, 176]. Additionally, for features whose regulation is distributed, as in stripe 7, each CRM uses a separate set of factors to generate function. As such, changes to the environment in *trans* will have different effects than observed on either element alone.

Consequences of enhancer competition. Competition for the basal complex has direct consequences for *eve* expression. The expression patterns driven by individual enhancers are broader than than the same stripes driven by the intact locus[179]. These broad expression domains driven by individual enhancers overlap at the interstripes. If expression is additive at these positions, there will be poor repression in *eve* interstripes. However, competing enhancers will drive expression at levels less than the sum of the rates driven by either enhancer alone. Thus, enhancer competition is sufficient to explain how sharp stripes are driven by broadly expressed enhancers.

Studies of how transcription rate varies with respect to the positioning and separation between bound activators will be required to distinguish between different modes of enhancer competition, additivity, or cooperativity. Specifically, the model proposed in this work suggests that if two activators are bound adjacently, these activators will synergistically activate transcription through cooperative action on the basal complex. If these activators are then separated by increasing lengths of neutral DNA we expect that transcription rate will decline linearly up to a distance of α (Eqs. 3.1-3.3), at which point there will be steric hindrance preventing simultaneous interaction with the basal complex. If the rate does not decline, it implies that sequences do not compete, and that instead intervening DNA can be looped out. Such experiments could be performed by varying the relative positions and orientations

of shadow enhancers acting on a common promoter.

Latent enhancers. We identify a case in which a DNA fragment had regulatory activity in a reporter assay, but not in the intact locus. Our model predicted the existence of a new regulatory element in the *eve* locus, however when we tested the activity of this fragment *in vivo* we found that it drove expression in an unexpected pattern that is not a subset of the expression pattern driven by the *eve* locus. For that reason we conclude that this fragment is not active in the intact *eve* locus and that placing it upstream of a reporter has revealed latent function. We hypothesized that this fragment may lie within inaccessible chromatin. Indeed, when we only model binding sites within accessible chromatin, we no longer predict expression driven by this fragment in the intact locus. This result highlights the importance of studying intact loci in addition to isolated enhancers and indicates that incorporation of chromatin accessibility increases the accuracy and utility of regulatory models.

Generalizability of this approach. Our analysis of *eve* depended on having the DNA sequence and chromatin accessibility of the locus, expression data from the locus over a range of cell types comparable to what is seen *in vivo*, a complete set of regulators (although as we discussed above, there is some ambiguity as to the full set of activators), a set of PWMs for these regulators, an understanding of the extent of the locus, and knowledge of the functional roles of the TFs used in the models. Overall, most or all of this information is already available in numerous systems or can be obtained, at least in principle, by high-throughput techniques. Entire genome sequences are now available for a large number of organisms together with functional data including chromatin accessibility [32, 129] across numerous tissues and cell lines. Expression data could be achieved through RNA-seq on a carefully curated set of cell lines, or alternatively from single-cell techniques on more homogeneous tissues. The cells from which expression data is obtained must also be subjected to transcriptome or proteome analysis to reveal the TFs present. The extent of the locus can be obtained my mapping insulator elements [132] or using chromosome conformation

assays[36]. Curated sets of PWMs for TFs are readily available [124], and TF roles and interactions can be inferred from data [178] or learned by comparing the model results of all possible perturbations of functional roles[18]. While this list appears imposing, all of the assays mentioned are regularly performed and the challenge is to integrate them together in an effective high-throughput approach. The study reported here provides a proof of concept for such future investigations.

3.6 Materials and Methods

3.6.1 Model data inputs

Quantitative levels of *eve* mRNA along the AP axis have previously been reported for three lines in Drosophila Genetic Reference Panel (DGRP)[116][83]. Data from L1 (RAL-437) at time class T6 was used for this study and is reported in File 7.2. This data corresponded well to RNA data collected in 3D from CantonS[45]. Quanititative enhancer-reporter data was obtained from Staller et. al [179]. Relative transcription levels along the AP axis were obtained from the FlyEx database[153, 184, 185, 152]. PWMs were derived from SELEX for factors Bcd, Hb, Kr, and Gt[143], bacterial-one hybrid for Kni and Cad[139], and footprinted sites for Tll[156] and Dst (http://line.bioinfolab.net/webgate/help/dxp.htm#D-stat-223). These PWMs have been used in prior work[93, 122].

3.6.2 Sequence selection

The *eve* locus was taken from the *D. mel* assembly dm3 using coordinates 2R:5862089-5875238. A fragment spanning these bases was reported to drive the early nuclear cycle 14 seven stripe pattern [59, 163]. Multiple enhancers have been reported for *eve* stripes 2 and 3. For generation of figure 3.1 we selected the enhancer with the greatest length for testing, as the longer enhancer is more likely to contain all DNA which drives a particular stripe. For this figure, the stripe 2 enhancer, sometimes called S2E, is the 800bp sequence

spanning conserved blocks A and B reported in [113] and has dm3 coordinates 2R:5865217-5866014. The stripe 3+7 enhancer is the restriction fragment identified in [59] and has dm3 coordinates 2R:5863006-5863888. Both the stripe 4+6 enhancer and the stripe 5 enhancer were identified in [50]. These have respective dm3 coordinates of 2R:5871404-5872203 and 2R:5874230-5875033.

For the remainder of the work we used the sequences reported in Staller *et al.*[179] as these sequences can be directly compared quantitative to enhancer-reporter data from the same work.

3.6.3 Data Registration

All data in this work was registered against the transcription factor levels available in the FlyEx database. The *eve* RNA was imaged together with Eve protein so that the protein channel could be used for data registration. In order to use data from Staller et. al.[179] nuclei in a 10% DV strip along AP axis were registered to the FlyEx database using the *eve* RNA channel. All data was registered using the BREReA[96, 97] software. We used the mean levels at each percent embryo length for comparison to predicted reporter expression.

3.6.4 Parameter estimation

The model equations Eqs. 2.1-2.25, Eqs. 3.1-3.2, and 3.3 were implemented in C++ code. Optimization of model parameters was performed by minimizing the sum of squared differences (SSE) between the model and data using Lam-Delosme Simulated Annealing in serial [159, 99, 100]. Annealing parameters are given in File 7.3. Below we describe the search space and controls for accuracy and significance.

Search Space. The search space for each parameter was explicitly set (File 7.3) in terms of a range for each parameter. These ranges were set to ensure that biologically relevant parameter values could be achieved. A TF, a, will have 3 to 5 associated parameters depending

on its biological function. The first of these, A_a (2.8), converts activities observed as fluorescence and binding free energies obtained as PWM scores into chemical units. Intracellular activities of proteins range from about 1 to 1000 nM, and allowing for similar uncertainty in the affinities K, which always occur as products with v in the model equations (see 2.15), we allow A_a a range of somewhat more than 6 orders of magnitude, from 10^{-6} to 4×10^{0}. The next parameter, λ, scales differences in binding score to differences in relative affinity. Originally, values of 0.5 to 2 were proposed as a reasonable range for this parameter[15], but some PWMs used in this work were generated using multiple rounds of SELEX, which may under-represent low affinity binding sites. We extend the range of this parameter to be from 0.5 to 5 to allow for the possibility of over-specified PWMs. The range of the bicoid cooperativity ω was set to 1 to 1000. This corresponds to a ΔG of up to –7 kcal/mole, which is fully compatible with observed ranges for λ repressor and *Drosophila* TFs[84, 155].

The efficiency of transcription factors E^Q and E^C in repression or coactivation respectively always multiply the fractional occupancy f, and hence were fixed to their natural scale of 0 to 1. In contrast, the activation efficiency E^A also sets the scale of N and thus the steepness of promoter response to activation. We allowed E^A to vary from 0 to 25. At the high end of this range promoter response is sufficiently close to a step function that biologically undetectable changes in TF concentrations can switch the promoter between on and off states. θ ranged from 5 to 25 because it is subtracted from N, and values smaller than 5 allow for substantial transcription in absence of activation.

Optimization. Optimization was performed 20 times with κ [30] set to 1.6×10^{-4} and 80 times with κ set to 1.6×10^{-5}, where smaller values of κ give more accurate results at the cost of additional computational time. Each optimization run was started from a random set of initial parameter values.

In order to verify that our optimization procedure is able to find the global minimum we require a scenario in which this global minimum is known. We construct such a test problem

by replacing the data with the output of the parameter set reported in the main text of this work. In this case, there is a known global minimum at zero, where the learned parameters are the parameters of the fit used to construct the test problem. When we repeated this procedure 80 times, the best resulting parameter sets had scores several orders of magnitude lower than than those fit to data. The learned parameter sets were also well correlated with the parameter set used to generate the test problem. Spearmen ρ was 0.963, 0.952, and 0.934 for the best three fits respectively (File 7.3, sheet "Fit Known Optimum"). We obtained similar results when this control was repeated for models incorporating chromatin state (File 7.3, sheet "Fit Known Optimum Chromatin").

The lowest scoring run (Model 12) in the initial set of 20 runs with $\kappa = 1.6 \times 10^{-4}$ was selected for the analysis in this work. We verified optimiztion accuracy by performing 80 additional runs with $\kappa = 1.6 \times 10^{-5}$ and subjecting the best three of these to further analysis. These runs had a 4% improvement in summed square error and gave parameter values that were well correlated with Model 12, having Spearman ρ of 0.969, 0.969, and 0.924 respectively (File 7.3, sheet "Repeat"). These parameter sets did not differ significantly in their output or enhancer prediction (Figure 7.7). We also repeated this procedure for the model incorporating chromatin state. Again, the best three fits had similar properties, owing to a very high correlation in the achieved parameter sets, which had Spearman ρ of 0.958, 0.958 and 0.969 respectively (File 7.3, sheet "Repeat Chromatin"). These did not differ significantly in their prediction of enhancer location or output (Figure 7.8).

Overfitting is generally a concern when the number of parameters exceeds the number of data points. Here we are fitting 32 free parameters to 58 data points. Additionally, to confirm that this model was not overfit we tested whether permuted data could be used to drive the expression pattern. We permuted the non-coding sequence data and the best fits to this data had scores that were three times worse than the best fits to the *eve* locus (File 7.3). None of these fits drove all of the six *eve* stripes that were in the modeled region.

3.6.5 Calculation of contribution to stripe borders

At every AP position, the marginal contribution to transcription rate with respect to a change in each transcription factor concentration was calculated numerically by adding and subtracting from the concentration of each factor and calculating the predicted change in transcription rate while keeping all other parameters constant. Specifically, we estimate the quantity $\frac{\partial R_i}{\partial [A]_i}$ using numerical differentiation, where R_i is the predicted transcription rate at AP position i and $[A]_i$ is the concentration of factor A at the same position. Where $[A]$ is greater than 0 we use a symmetric difference quotient $\frac{f(x+h)-f(x-h)}{2h}$, otherwise we use Newton's difference quotient $\frac{f(x+h)-f(x)}{h}$. We used changes in concentration of orders of magnitude 10^1 to 10^{-11} to verify convergence of this estimate (Fig 7.6). To calculate the contribution of each transcription factor to a change in transcription rate between adjacent AP positions, we multiply $\frac{\partial R_i}{\partial [A]_i}$ by the amount that the transcription factor is changing at that position $\triangle[A]_i$, given by $\frac{[A]_{i-1}+[A]_{i+1}}{2}$.

3.6.6 Calculation of contribution to activation

To calculate the contribution of each factor towards the total transcription rate we first calculate the number of transcriptional adaptors recruited to each sequence by each factor $N_{i[m,m+\alpha;a]} = \sum_{k:a_i=a_k} F_k E^A_{a_k} I(k, m, m+\alpha)$. Next, we find the number of transcriptional adaptors recruited to the TSS by each factor by taking time weighted sum $N_a = \sum_i N_{i[m,m+\alpha;a]} T_i$. We report the percent of adaptors recruited to the TSS by each factor $100(N_a/\max(N_a))$.

3.6.7 Generation of reporter constructs

Reporter constructs where generated using a pCaSpeR backbone (GeneBank X81644.1) containing the promoter and first 22 amino acids of *eve* fused to *Lac*Z, generated by Small et al.[174]. An attB sequence was inserted into the multiple cloning site using the restric-

tion enzyme Xba1 for insertion in the AttP2 landing site on chromosome 3[63]. The enhancer sequence was extended by PCR primers containing overlap with this vector. The vector was then digested by enzymes EcoR1 and Xho1 and the enhancer was inserted using Gibson assembly[53]. The resulting vector was injected into flies of the genotype P{nosphiC31\int.NLS}X, P{CaryP}attP2 by Rainbow Transgenics. Quantitative data was collected from these lines as previously described [82].

3.6.8 Identification of accessible chromatin

Accessible chromatin regions defined by FAIRE-seq data by McKay 2013 [126] were obtained from GEO accession number GSE38727. Accessible chromatin regions defined by DNAse-seq data were obtained from Li 2011 [106] and were translated to dm3 coordinates using the UCSC genome browser LiftOver tool. Open chromatin regions were defined as the intersect of the two datasets.

CHAPTER 4

SYNTHETIC ENHANCER DESIGN BY COMPENSATORY EVOLUTION REVEALS FLEXIBILITY AND HIDDEN CONSTRAINTS ON *CIS*-REGULATION

4.1 Abstract

Models that incorporate specific chemical mechanisms have been successful in describing the activity of *Drosophila* developmental enhancers as a function of underlying transcription factor binding motifs. Despite this, the minimum set of mechanisms required to reconstruct an enhancer from its constituent parts is not known. Synthetic biology offers the potential to test the sufficiency of known mechanisms to describe the activity of enhancers, as well as to uncover constraints on the number, order, and spacing of motifs. In this work, we test the degree to which known mechanisms explain the activity of the *Drosophila even-skipped* stripe 2 element. Using *in silico* compensatory evolution, we generated 40 putative synthetic stripe 2 elements that have varying degrees of homology to the natural element. We show that functional models are able to balance the levels of binding for activators and repressors in order to maintain spatially regulated expression, even after more than a third of binding sites have been lost. We also show that a sequence about twice as small as the minimal stripe 2 enhancer can drive ten times greater expression levels. Sequences that contained the same arrangements of known motifs drove different levels of expression, indicating that undescribed factors likely act to modulate expression. Activation driven by Bicoid and Hunchback is highly sensitive to spatial organization of binding motifs, while activation driven by Stat92E or Zelda is flexible. Finally, we show that expression can be highly variable within embryos, and indicate potential mechanisms driving this variability.

The work presented in this chapter has been submitted to *Genetics* as a manuscript by myself, Carlos Martinez, Jennifer Moran, AhRam Kim, Alexandre Ramos, and John Reinitz.

Carlos Martinez and AhRam Kim generated the quantitative models that I used to design the putative enhancers. Jennifer Moran assisted in the design and cloning the reporter constructs. Alexandre Ramos generated the curves describing theoretical variation mRNA presented in Figure 4.8.

4.2 Introduction

Enhancers, also known as *cis*-regulatory modules (CRMs), are DNA segments that recruit sets of sequence-specific transcription factors (TFs) in order to control the spatiotemporal expression of genes. These elements are critical in controlling cell fate in development[104] and are under selection[109, 6, 201]. More recently, genetic variation within enhancers has been widely implicated in common human disease[72, 125]. Predicting the effects of this *cis*-regulatory variation on local gene expression remains a challenging task. Even in the best studied enhancers, there is evidence of unknown function[187, 7], and it is not yet possible to reconstruct these elements from their constituent parts [199, 86].

The enhancer which drives the second of seven transverse stripes of *even-skipped* (*eve*) in the developing *Drosophila* blastoderm is among the most studied enhancers in all of biology. A deletion of a 480 bp fragment located 1.1kb upstream of the transcription start site leads to loss of this stripe[59], and is the smallest known fragment sufficient to drive reporter expression in a stripe 2 pattern[175]. Footprinting, TF knockouts[177] and site-directed mutagenisis[180] of this minimal stripe 2 element (MSE2) have identified 4 TFs that act through 12 sites in order to direct the stripe 2 pattern. MSE2 is broadly activated in the blastoderm through the activators Bicoid (Bcd) and Hunchback (Hb) and forms a stripe through repression by the factors Giant (Gt) on the anterior and Kruppel (Kr) on the posterior [177, 175, 180, 48]. Despite being subject to such detailed molecular dissection, there are unexplained features of this enhancer. For instance, deletions of sequences outside the 12 footprinted sites all led to changes in function and additional TFs are required to prevent aberrant expression driven by this enhancer[7].

Enhancers integrate the simultaneous, opposing effects of both activators and repressors in order to determine specific expression levels. Thus, predicting the output of enhancers given any level of input requires quantitative methods. To address this, confocal microscopy has been used to generate spatial and temporal atlases of protein[185, 184] and mRNA[114, 45] levels at single nucleus resolution during the first 4 hours of *Drosophila* development. Using transgenesis of enhancers driving reporter expression, the precise input-output function of enhancers can be measured. Sequence-level models (SLMs) of gene regulation have been used to describe this function as an emergent property of underlying TF binding sites[81, 171, 164, 92, 70, 93, 122, 167]. Such models predict binding using thermodynamics and incorporate known, context-dependent rules of TF function, such as repression through short-range quenching[61, 60, 42]. SLMs have identified additional binding sites, regulators and interactions that are important in the control of MSE2[81, 93].

Experiments with enhancers across *Drosophila* species suggest that there is considerable flexibility in the architecture of stripe 2 enhancers. Sequences that have diverged over tens of millions of years still drive stripe 2, despite a lack of sequence conservation[113, 112, 68, 67, 122]. This functional conservation in the absence of sequence conservation indicates that there are many ways to construct a stripe 2 enhancer. SLMs trained on enhancer-reporter data from *D. melanogaster* have successfully predicted the activity of stripe 2 enhancers (S2Es) from distant Drosophilids[93] and identified accessible evolutionary paths that conserve expression through compensatory evolution[122]. This suggests that the context-dependent rules incorporated into SLMs are sufficient to describe the flexibility of MSE2 *cis*-regulatory logic.

While SLMs have been able to successfully describe the activity, evolution and flexible architecture of evolved enhancers, such enhancers represent only a small proportion of the sequences that are predicted to drive stripe 2[123]. Instead, selection may obfuscate many constraints on the order and arrangement of TF motifs that give rise to functional S2Es by removing nonfunctional motif configurations from natural populations. This is sug-

gested by the fact that sequences that lie outside of known binding motifs are necessary for expression[7, 187], as well as by past failures to generate synthetic *Drosophila* enhancers using reconstituted binding sites[199, 86].

Synthetic enhancers offer a potential means to address the extent to which known regulatory mechanisms represent a complete description of the *cis*-regulatory function, and to uncover hidden constraints on *cis*-regulatory architecture. While SLMs can be used to generate thousands of sequences that are predicted to drive virtually any pattern along the *Drosophila* anterior-posterior (AP) axis[123], the apparent success or failure of such sequences to drive the expected expression pattern is uninterpretable in the presence of a large number of changes from naturally selected sequence. Because the large number of sequence changes will tend to conflate those changes which are neutral with those which are critical, what is needed is a method in which functional changes can be attributed to a single or small number of sequence changes. Furthermore, because the minimum requirement for constructing a functional regulatory element is unknown, such changes must be made from the starting point of functional naturally selected sequence.

In this work we introduce a novel approach to the design of synthetic enhancers. Using synthetic compensatory evolution, we generated two series of S2Es with decreasing homology to MSE2. These series address the extent to which SLMs describe the eve stripe 2 regulatory function and provides informative data when results differ from predictions. In total, we tested the activity of 40 synthetic putative regulatory sequences using site-specific integration[63] in developing *Drosophila* embryos. We collected quantitative expression data from 8 synthetic enhancers. We found that an SLM was able to successfully balance the effects of activators and repressors in order to maintain a stripe even after over a third of binding motifs were lost. We showed that a synthetic sequence half the size of the previously minimal stripe 2 element is able to drive this stripe at more than 10 times the levels driven by MSE2. While homotypic arrays of the activators Zelda (Zld) and Stat92E (Dst) were able to drive expression, we found that activation driven by Bcd and Hb is sensitive to the

spacing, affinity, and orientation of sites. Additionally, we found that motif content not only controls mean expression levels, but also variability in expression within single embryos.

4.3 Results

4.3.1 Design of synthetic enhancers with decreasing homology to MSE2

The rational design of synthetic enhancers requires a quantitative model that is able to balance the action of numerous activators and repressors acting on the same DNA sequence. The developing *Drosophila* embryo permits the input-output function of regulatory DNA to be assayed with single nucleus precision. In this work we place test sequences upstream of a *lac*Z reporter using site specific integration in *Drosophila* embryos so that integration site effects are fixed and output can be quantitatively compared across lines (Fig. 1.3A). The embryos are subsequently imaged in nuclear cycle 14 (C14), timeclass 6 (T6)[185, 184] for nuclei, *lac*Z and Eve protein (Fig. 1.3B). At this point in development seven stripes of *eve* expression are clearly defined, but cross-regulation from other pair-rule genes is absent [59, 163]. Next, nuclei are segmented (Fig. 1.3C) and data from multiple embryos are averaged to yield an expression profile for each enhancer in a 10% wide stripe along the AP axis (Fig. 1.3D) [82]. We consider nuclei from 35.5% to 92.5% embryo length, where there is a clean functional distinction between the AP and dorsal-ventral axis. This profile is then registered to an atlas of protein levels [152]. The resulting dataset is a cell by cell assay of transcription under the control of quantified TFs, which can be used to obtain the input-output function of any DNA sequence (Fig. 1.3E).

In previous work, we have generated quantitative models that explain the *cis*-regulatory function of eve stripe 2 from multiple species, in which the kinetic parameters of an SLM are trained to the input output function of multiple enhancers (Fig. 4.1A-B). In one instance, fusions of the *Drosophila even-skipped* stripe 2 and stripe 3 enhancers give rise to novel expression patterns[174] that proved a rigorous training set for SLMs. An SLM trained on

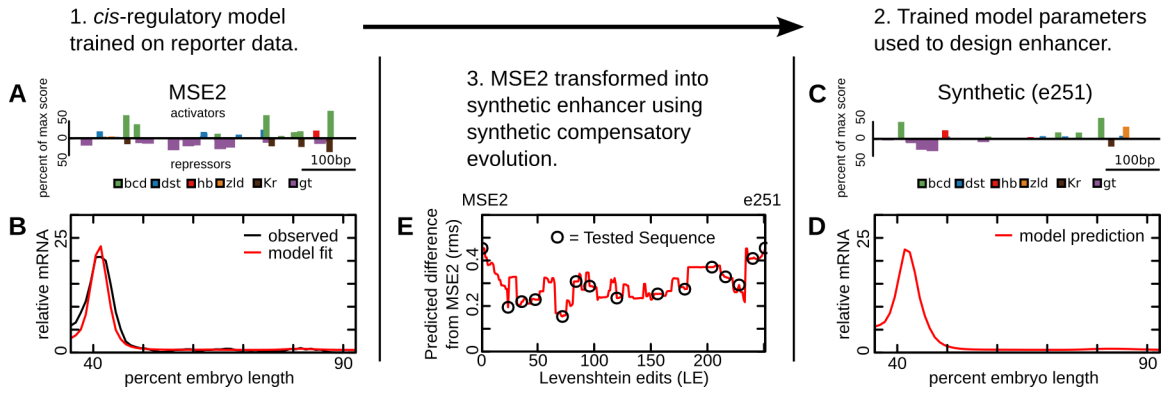


Figure 4.1. Design of a synthetic compensatory path. (A) The binding structure of the minimal stripe 2 element is shown. Height of bars represents the percent of the maximum PWM score of a motif at each position. Putative activators are plotted on the positive axis and putative repression on the negative axis. (B) Expression levels of enhancers are used to train a sequence level model (SLM) of gene regulation. The observed expression levels (black) and model fit (red) are shown along the AP axis. (C) The binding structure of a putative synthetic stripe 2 element designed *in silico*. (D) The predicted expression of the synthetic enhancer in C along the AP axis is shown. (E) A synthetic compensatory path is found that will transform MSE2 into the synthetic element while preserving predicted expression of a stripe. The root mean squared (rms) differences between predicted expression and MSE2 expression is shown as a function of number of sequence mutations. Each step along the x-axis represents a single nucleotide change to the previous sequence. A set of these elements are selected for validation *in vivo*.

this data was able to predict expression pattern driven by S2Es from distant Drosphilid and Sepsid flies[93]. In another example, a model trained on S2Es from multiple Drosophilids identified putative ancestral S2Es and accessible evolutionary paths between them[122].

The parameters of SLMs generated in these studies provides a starting point for the design of synthetic regulatory sequences. Keeping the kinetic parameters of the SLMs fixed, we optimized DNA sequence using simulated annealing such that it minimized the sum of squared differences between the expression of MSE2 and the output of one to seven models, each with its own set of kinetic parameters (see Fig. 4.1C-D and Materials and Methods). This yields a sequence that is designed *ab initio* to drive expression in the pattern of *eve* stripe 2. Two such *ab initio* enhancers were generated and tested *in vivo* in the present study.

In order to generate a set of sequences with decreasing homology to MSE2, we also generated sets of compensatory mutations that will mutate MSE2 into each of the *ab initio* synthetic enhancers while conserving stripe 2 expression. This was done by finding the single nucleotide changes required to mutate one sequence into the other, known as the Levenshtein edits (LE), and permuting their order such that predicted stripe 2 expression is conserved as much as possible at each edit [123, 122]. We call this set of edits a "neutral path." Each neutral path uses the same set of kinetic parameters as were used to design the *ab initio* synthetic construct at the end of the path. We selected paths of 15 and 11 putative enhancers respectively to test *in vivo*(Fig. 4.1E).

The two *ab initio* synthetic enhancers generated in this work are separated from MSE2 by 251 and 272 LE respectively, and we call them e251 and s272. Sequence e251 was designed to drive expression from well separated binding sites using the *in silico* strategy described above. In contrast, sequence s272 was designed to be a "sub minimal" stripe 2 element that was as short as possible subject to the constraint that all TFs known to regulate stripe 2 could bind and exert their regulatory effects by mechanisms known to operate in S2E. This was done by arranging consensus bindings motifs for known regulators of stripe 2 by hand

and then adjusting their affinity such that differences between stripe 2 expression and the model output of this enhancer were minimized. We discuss each of these sequences and the neutral paths in turn.

4.3.2 Expression along the e251 synthetic compensatory path

The first four tested sequences along the neutral path to e251—at 24 LE (e34), 36 LE (e36), 48 LE (e48), and 60 LE (e60) from MSE2—all successfully balanced activation and repression in order to maintain expression of a stripe at 40% embryo length (Fig. 4.2). As predicted, each of these four sequences expressed at levels greater than MSE2(Fig. 4.2A-E). The remaining 11 sequences at greater than 72 LE did not drive expression in the modeled region. In addition to this region, many embroys also drove expression within the anterior portion of the embryo (Fig. 7.9A-E). The vector used in this work has previously been reported to drive an ectopic stripe anterior to *eve* stripe 1[174]. To confirm that expression is driven by the vector, we generated a control construct that contained no enhancer. Embryos with this construct drove expression of an ectopic stripe (Fig. 7.10), albeit at levels considerably lower levels than in some tested synthetic enhancers.

To confirm that the sequence changes resulted in binding site turnover, we examined the gain and loss of binding motifs for the key regulators Bcd, Hb, Kr, and Gt. We identified a total of 32 binding motifs for these regulators at a PWM score[71] greater than zero. By 60 LE, 19 (59%) sites are conserved (Fig. 4.2F), 10 sites are gained (Fig. 4.2G), and 13 (41%) are lost (Fig. 4.2G). This level of binding site turnover is comparable to that seen between *D. mel* and *D. erecta* (Fig. 7.11), which are several million years diverged[188] and show quantitative differences in expression driven by S2E[122].

4.3.3 Known motifs cannot explain loss of expression after 60 LE

While sequences at 60 LE or less drove stripe 2 expression, sequences at 72 LE or more failed to drive expression. We sought to identify which of the 12 sequence mutations was

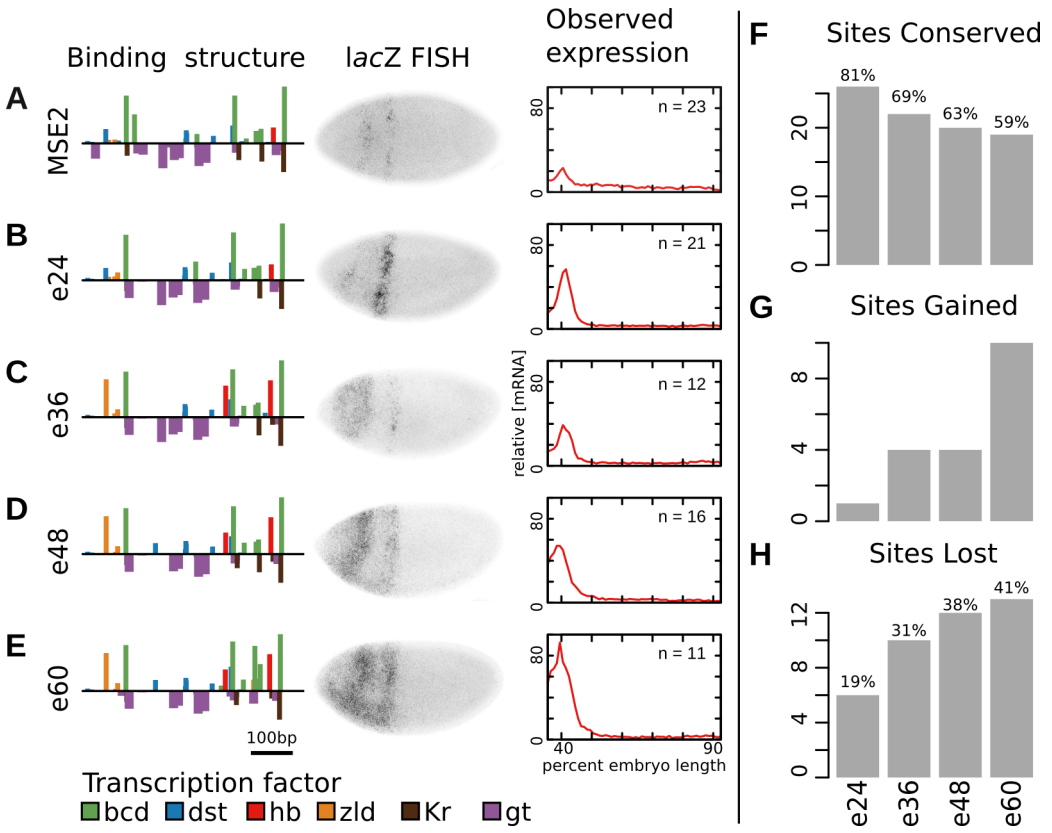


Figure 4.2. Expression along a synthetic compensatory path. (A-E) Synthetic enhancers 24 (e34), 36 (e36), 48 (e48) and 60 (e60) edits removed from MSE2. (Column 1) Binding structure of tested enhancers. Activators are plotted on the positive y-axis and repressors on the negative. Bar height is proportional to the PWM score of binding. (Column 2) The *lac*Z mRNA expression of a representative embryo. (Column 3) The quantitative level of mRNA expression for each line along a 10% DV stripe from 35.5% to 92.5% embryo length. Data represents an average fluorescence. The number of averaged images, n, is indicated. (F) The number of binding motifs at PWM score greater than 0 for the factors Bcd, Hb, Kr, and Gt that are conserved with MSE2 are shown for each synthetic enhancer. The percent conservation is given above each bar. (G) The number of binding motifs (PWM score >0) for the same 4 factors that are gained are shown for each synthetic enhancer. (H) The number of motifs (PWM score >0) for the 4 factors that are lost are shown for each factor. The percent of sites lost is given above each bar.

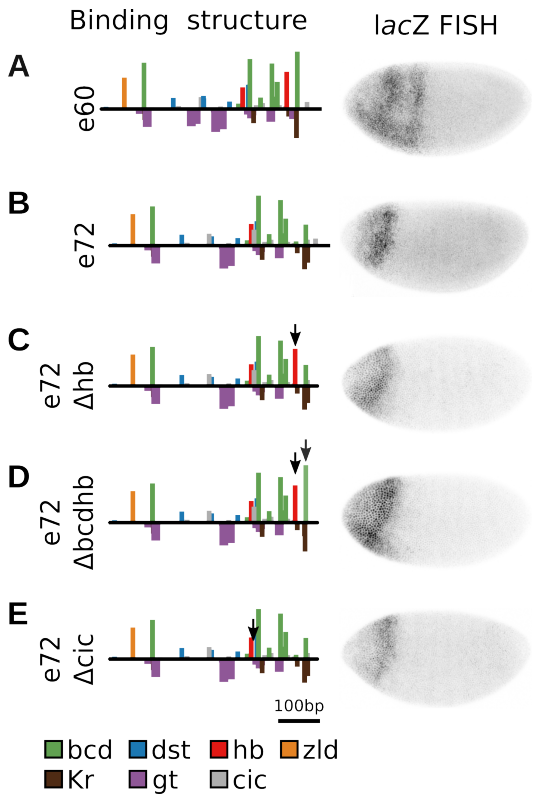


Figure 4.3. Known motifs cannot explain loss of expression after 60 LE. (A) The binding structure of the e60 construct and image of a representative embryo. Height of bars is proportional to PWM score of the binding motif. e60 drives expression of a stripe. (B) The binding structure end representative embryo from e72. This sequence is only 12 bp different than e60. (C) The binding structure of e72 with the affinity for a hb site (arrow) restored to levels in e60. The construct does not drive expression. (D) The binding structure of e72 with the affinity for a hb site and Bcd motif (arrows) restored to levels in e60. The construct does not drive expression. (E) The binding structure of e72 with the a cic motif (arrow) removed, as in e60. The construct does not drive expression.

responsible for this change. Most model parameters predicted a reduction in expression at 72 LE (Fig. 7.9) as a result of the loss of a Hb motif and reduced affinity for Bcd. Recent work has highlighted the importance of the lost Hb site [93], making it a prime candidate for the change which caused loss of expression. The loss of the Hb motif is a result of a single nucleotide A to T change (ATAAAAA to ATATAAA). We reversed this change in e72 to restore the Hb motif. This did not rescue expression in e72 (Fig. 4.3C).

In addition to loss of a Hb, loss of Bcd affinity can explain loss of expression at e72. A single nucleotide T to G change from e60 to e72 (GGATTA to GGATGA) disrupted a consensus Bcd motif. Hb, which is typically a repressor, is able to activate when bound near Bcd[65, 174, 93], so it is possible that the loss of Bcd affinity not only disrupted activation through Bcd, but also though this coactivation of Hb. To test this, we restored both the Bcd and Hb sites in e72. The resulting sequence did not rescue expression driven by e72 (Fig. 4.3D).

The remaining 10 nucleotide differences between e60 and e72 do not lead to appreciable differences in the predicted affinity for modeled TFs. We checked for predicted changes in binding preferences for factors within the Fly Factor Survey[208] that are maternally or ubiquitously expressed in the Berkeley Drosophila Genome Project[194]. A single candidate emerged. A single nucleotide T to A change (TGATTG to TGAATG) led to the creation of a site for the repressor Capicua (Cic) that is expressed throughout the entire modeled region. This repressor is reported to set borders of Bcd target genes[29], and its binding motif is present in all tested sequences at 72 or greater LE from MSE2. Removal of this site did not restore expression in e72 (Fig. 4.3E).

4.3.4 Expression along the s272 synthetic compensatory path

The sequence s272 was designed to contain the motifs and interactions currently known to be essential for stripe 2 expression (Fig. 4.4A). The sequence contains a single motif for Dst, which is known to be essential for expression of eve stripe 3 [183] and is a major activator of

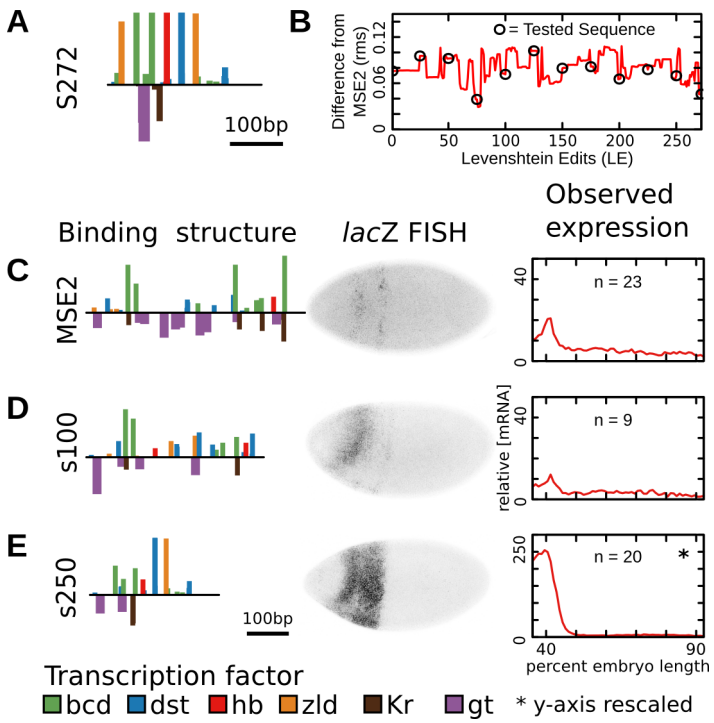


Figure 4.4. A 319bp synthetic enhancer drives stripe 2 at levels more than 10 times greater than MSE2. (A) The binding structure of a sequence designed, with model feedback, to drive expression of a stripe 2 pattern is shown. Height of bars is proportion to PWM score. Most sites represent consensus motifs for each factor. (B) The predicted root mean squared differences from MSE2 expression along a series of 272 edits that would transform MSE2 into the sequence in A. 10 sequences that were tested *in vivo* are shown. (C) The binding structure, a representative embryo, and quantitative levels driven by MSE2. The number of averaged images, n, is indicated. (D) The binding structure, a representative embryo, and quantitative levels driven by a sequence 100 edits from MSE2. This sequence drives expression of levels slightly less than MSE2. The number of averaged images, n, is indicated. (E) The binding structure, a representative embryo, and quantitative levels driven by a 319 bp sequence 250 edits from MSE2. This sequence drives expression at levels more than 10 times greater than MSE2. The number of averaged images, n, is indicated.

zygotic expression[195]. It contains two adjacent motifs for the activator Bcd, which binds to DNA cooperatively[66]. It contains a single Hb site, which activates expression of eve stripe 2 when bound near Bcd[81, 93]. These four activator motifs are flanked by two Zld motifs, which has been reported to open chromatin[166, 169]. Finally, the sequence contains motifs for the repressors Gt and Kr, which set the boundaries of stripe 2 expression[48, 180, 177, 175, 81]. In addition to this sequence, we designed and tested the activity of ten sequences in a set of 272 LE that mutate MSE2 into this synthetic enhancer while conserving stripe expression (Fig. 4.4B,7.12).

While the designed enhancer did not drive reporter expression (Fig. 7.12), two tested sequences drove expression of a stripe at 40% embryo length. A sequence at 100 LE (s100) drove weak expression of stripe 2, despite having lost 22 of 32 (69%) of binding motifs for the factors Bcd, Hb, Gt, and Kr (Fig. 4.4D). Another sequence, 250 LE from MSE2 (s250), drove very strong expression of a stripe at 40% embryo length despite having only 2 (6%) motifs conserved with respect to MSE2. At 319 bp in length, with the majority of motifs falling in a ¡200bp cluster, this sequence is significantly smaller than the 480bp minimal stripe 2 element, yet drives expression at levels more than 10 times greater than MSE2 (Fig. 4.4E).

4.3.5 Homotypic clusters of Zelda and Stat92E drive embryonic expression

The sequence s272 did not drive expression despite being separated by only 22 LE from a strong enhancer. The loss of expression could be due to an imbalance between activation and repression. To test this hypothesis, we generated a variant of the designed enhancer that eliminated motifs for Kr and Gt (s272ΔgtΔhb). The resulting sequence failed to drive expression (Fig. 4.5A).

In order to determine which TFs are capable of driving expression alone, we generated sequences with homotypic clusters of 6 motifs for Zld, Bcd, and Dst. Starting from the previous construct, we replaced each of the 6 motifs with a motif for the specified factor, keeping

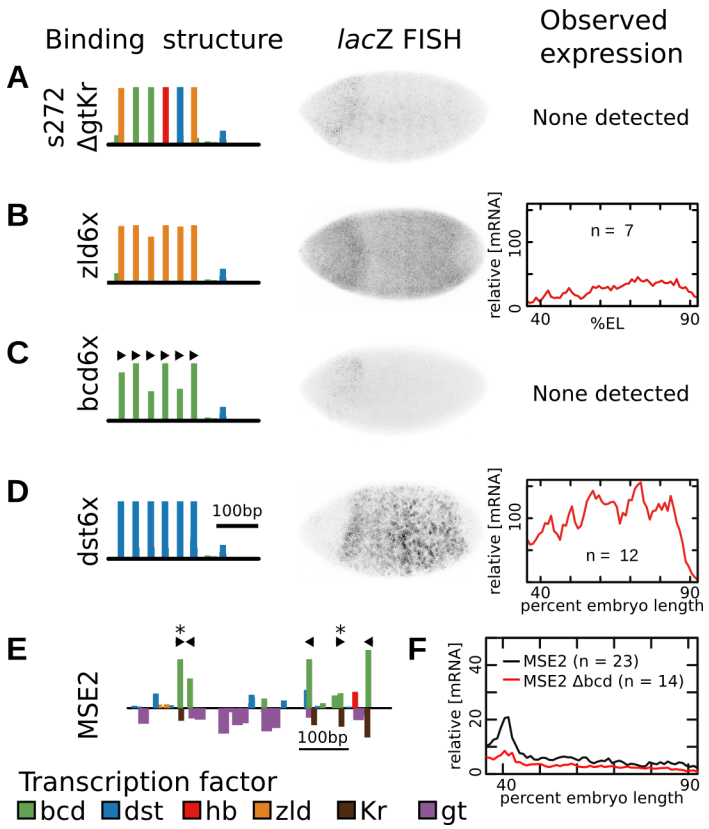


Figure 4.5. Homotypic clusters of Zld and Dst, but not Bcd drive embryonic expression. (A) The binding structure and representative embryo of sequence s272 with sites for repressors Gt and Kr removed are shown. (B) Each motif in A was replaced with a motif for Zld, preserving inter-motif sequences. The resulting enhancer drives expression across the entire length of the embryo (middle and right). The number of averaged images, n, is indicated. (C). Each motif in A was replaced with a motif for Bcd, preserving inter-motif sequences. Arrow represent the orientation of binding motifs. The resulting sequence did not drive expression (middle). (D) Each motif in A was replaced with a motif for Dst, preserving inter-motif sequences. The resulting enhancer drives expression across the middle of the embryo (middle and right). The number of averaged images, n, is indicated. (E) The binding structure of MSE2 is shown with arrows indicating the orientation of Bcd motifs in the sequence. (F) The expression driven by MSE2 and MSE2 with the motif orientations indicated in E reversed. The resulting sequence has the same predicted affinity for Bcd, but drives less expression than MSE2. The number of averaged images, n, is indicated.

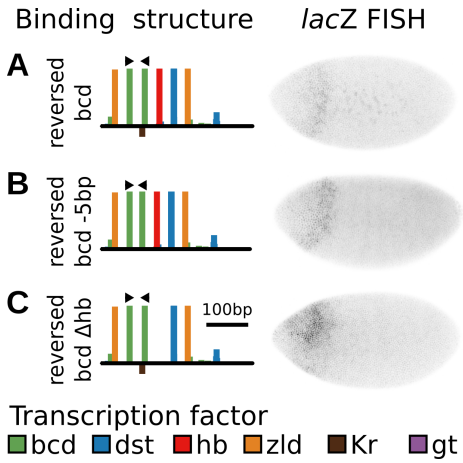


Figure 4.6. Bcd orientation and spacing does not rescure s272. (A) The motif structure of the s272 construct with the second Bcd motif orientation reversed. The resulting enhancer does does not drive expression (right). (B) The motif structure in A with the 5 bp of inter-motif spacer removed. The resulting enhancer does not drive expression (right). (C) The motif structre in A with the hb site deleted. The resulting enhancer does not drive expression.

any intermotif sequences constant. Some small differences from consensus motifs were necessary to prevent the creation of repressor motifs. Homotypic clusters of Zld drove moderate expression (Fig. 4.5B) and homotypic clusters of Dst drove strong levels of expression (Fig. 4.5D).

4.3.6 Bcd binding orientation is important for MSE2 function

While homotypic clusters Zld and Dst drove reporter expression in developing embryos, 6 Bcd motifs failed to drive expression(Fig. 4.5C). The fact that Bcd immunoprecipitation preferentially pulls down sequences with Bcd in the head to head orientation [49] suggests that Bcd orientation may be an important factor. This point is strengthened by the fact that orientation and spacing drove different levels of expression in fly embryos[66, 26]. In order to test the role of Bcd orientation in MSE2, we reversed the orientation of two Bcd sites, keeping affinities of all sites constant (Fig. 4.5E). The resulting sequence drove significantly lower levels of expression than MSE2 (Fig. 4.5F).

In order to test the role of Bcd orientation in s272ΔgtΔhb, we generated a sequence

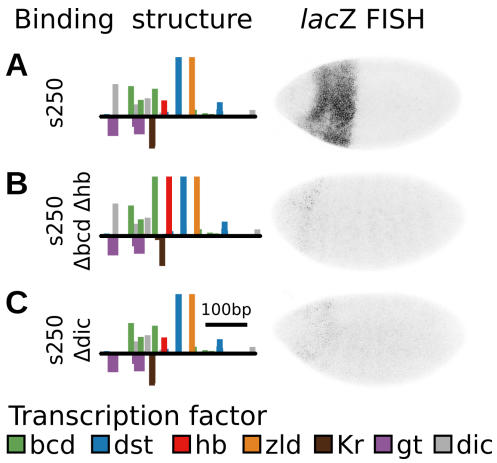


Figure 4.7. Bicoid, huncback, and dicheate are essential for expression driven by s250. (A) The binding structure of the s250 construct. The the factor dicheate has been included in the binding structure in grey. The enhancer drives strong expression (right) (B) The binding structure in A with a single Bcd and hb orientation reverted to the orientation present in s272. The resulting enhancer does not drive expression(right). (C) The binding structre in A with the a single dicheate site removed. The resulting enhancer does not drive expression (right)

that reversed the orientation of a Bcd motif (Fig. 4.6A). This sequence did not restore expression. In order to test whether helical orientation on DNA prevented these sites from binding cooperatively, we removed 5bp of inter-motif DNA. This sequence did not restore expression (Fig 4.6B). Finally, to test whether Hb was not being coactivated, and thus preventing expression through quenching, we tested a sequence that reversed a Bcd motif orientation and removed the Hb motif. This sequence did not restore expression (Fig. 4.6C).

4.3.7 Bicoid, Huncback, and Dicheate are essential for expression driven by s250

While various orientations of Bcd and Hb did not restore expression in s272, the sequence s250 drove strong expression despite being only 22 LE different in sequence. The fact that Bcd binding orientation is important in MSE2 suggested that the specific orientation of Bcd and Hb motifs in s250 is essential. Additionally, a strong binding motif for the TF Dicheate was lost between s250 and s272.

We tested whether each of these changes was responsible for loss of expression individually. Editing a single Bcd and single Hb site and their inter-motif sequence led to a complete loss of expression driven by the sequence (Fig. 4.7B). Surprisingly, a change of only 2bp that removes a single motif for Dicheate also led to complete loss of expression (Fig. 4.7C).

4.3.8 Motif content controls variability in expression within embryos

We noted a difference in the visual appearance of mRNA FISH for expression driven by homotypic clusters of Dst compared to Zld (Fig. 4.5B,D), which seemed to indicate a high level of variation in expression between adjacent nuclei in the construct driven by Dst. In order to investigate within embryo variation we first adjusted for differences in mean expression between embryos (Materials and Methods), and then considered the expression in individual nuclei across the AP axis when driven by homotypic clusters of Dst (Fig. 4.8A) and Zld (Fig. 4.8B). The levels of expression driven by Dst appears to have both a higher mean and greater variability than expression driven by Zld. To test whether the higher mean levels can explain variability, we tested the relationship between the mean and standard deviation in each line. We found that there was a linear relationship between the mean expression and standard deviation of expression in 1% bins along the AP axis, but the slope of this line for the Dst driven enhancer is nearly double that of the Zld driven enhancer (0.28 to 0.54) (Fig. 4.8C), indicating that greater expression variability cannot be explained by difference in the mean.

In order to observe the shape of the distribution independent of the mean, we divided the fluorescence values by the mean levels at each AP 1% bin from 60 to 80% embryo length. The resulting distribution in fluorescence about the mean is wider when driven by Dst than when driven by Zld (Fig. 4.8E-F).

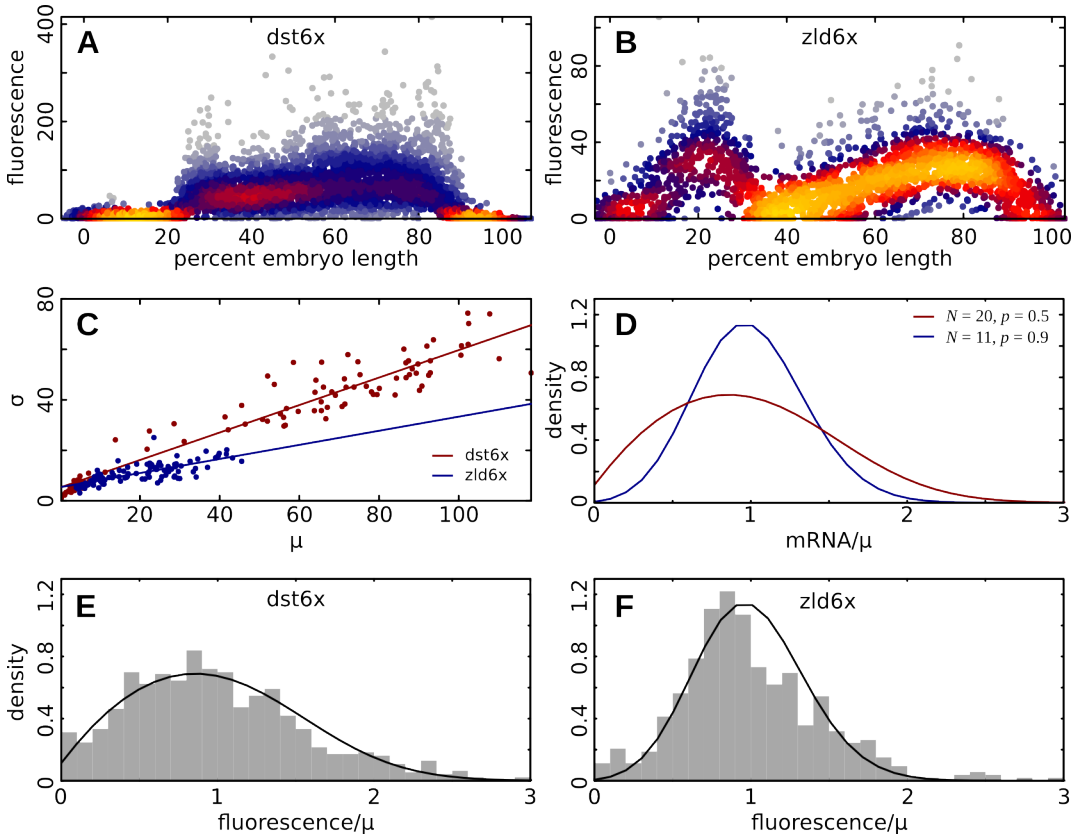


Figure 4.8. Motif structure controls variability in expression. (A) Individual nucleus fluorescence levels driven by the construct dst6x. Expression data on each embryo was scaled to minimize the the sum of squared pairwise differences in fluorescence intensity between embryos in 1% bins along the AP axis. Points are colored according to their local density. (B) Expression levels driven by the construct zld6x, plotted as in A. (C) To test whether differences in standard deviation can be explained by different mean expression levels, the standard deviation as a function of the mean expression is plotted for each 1% bin along the AP axis for both dst6x (red) and zld6x (blue). The relationship between standard deviation and mean is linear with different slopes for each construct. (D) Two distributions of mRNA count, divided by the mean, in a stochastic transcription model in which the number of transcripts and the ON-OFF state of the promoter are coupled random variables (see Materials and Methods). The parameter N represents the strength of transcription when the gene is in the ON state, and b is the probability of finding the gene in the ON state. (E) Fluorescence values of nuclei were divided by the mean levels in each 1% bin from 60% to 80% embryo length. The resulting distribution in fluorescence about the mean for dst6x is shown. The distribution with $N = 20$ and $p = 0.5$ from D is also shown (black line). (E) The distribution of fluorescence about the mean for zld6x is shown. There is a significant difference in the variance of the zld6x and dst6x distributions (Fligner-Killeen test, $p = 1.7 \times 10^{-8}$). Additionally, the distribution with $N = 11$ and $p = 0.9$ from D is shown (black line).

4.4 Discussion

Synthetic biology affords the opportunity to rigorously test constraints on the number, order, and types of TF binding sites required to drive specific spatial and temporal expression of genes. In this work we generated 40 synthetic sequences using a model of gene regulation that captures known chemical mechanisms and rules that govern the architecture of enhancers. These enhancers, which had varying degrees of homology to MSE2, were constructed in order to address the question of to what degree are these mechanisms and rules sufficient to describe the activity of enhancers? We found that while the model successfully predicted the activity of several enhancers, incongruities point to new molecular players and mechanisms that are required to predict regulatory function.

The mechanisms included in our model are DNA binding, steric competition for DNA binding, cooperative binding, short-range repression, direct repression, and coactivation of Hb by Bcd or Caudal. This set has been sufficient to explain the flexibility evident in enhancer sequence divergence over the course of evolution [93, 122]. The fact that S2E expression was maintained in the first tested synthetic compensatory path, even after 41% of binding sites for key regulators were lost, suggests that these mechanisms explain much, but not all, of the function of S2E.

4.4.1 Additional factors

Despite this initial success, only the first 4 of 15 tested synthetic sequences in the path to e251 successfully drove expression in nuclear cycle 14. e60 drove expression and e72 failed to drive expression. We analyzed the 12 nucleotide changes that led to loss of expression in e72. Of these 12 changes, 3 were in known binding motifs. Restoring these three changes did not restore expression. This result indicates that there are either unknown motifs which have been gained or lost, or that the edits resulted in other structural changes to DNA that disrupted function.

This evidence that there are new factors is supported by other work. For MSE2, an enhancer reconstituted with all footprinted sites failed to drive expression [199]. For other *Drosophila* enhancers, function was found to reside within most inter-motif sequences [86]. Additionally, DNA features such as GC content and dinucleotide content may affect reporter activity through structural effects[121].

4.4.2 Constraints on enhancer architecture

Every enhancer in the neutral paths considered in this work was constructed to contain a similar balance of bound activators and repressors. Despite this fact, these sequences drove vastly different levels of expression in developing embryos. This suggests that there are additional interactions between bound transcription factors that modulates their activity. We characterized one particular interaction in this work, Bcd cooperativity, that has a constraint not considered in the model. We found that changing the orientation of two Bcd sites within MSE2 disrupted activity of the enhancer, leading us to conclude that pairwise cooperative binding of Bcd requires a pair of sites with opposite orientation on the same strand.

We also discovered a molecularly uncharacterized component of interactions between Bcd and Hb. In sequences containing only six binding sites for known activators, we were unable to find combinations and orientations of Bcd and Hb that drove expression in developing embryos, even in the absence of known repressors, and despite the sequence being compatible with expression when the six motifs were substituted with either Zld or Dst. Despite this, changing 15 nucleotides between s250 and s250ΔbcdΔhb in such a way as to increase the affinity and spacing of a Bcd and Hb site was able to render the strongest enhancer assayed in this work non-functional (Fig. 4.7B). Collectively, this suggests that spacing and orientation of sites is critical in controlling levels of expression. While there are many permissible configurations, as evidenced by binding turnover in evolution, there may be many more configurations that are not permissible. Further work with synthetic sequences that exhaustively tests regulatory output as a function of distance and orientation will be required

to precisely define this interaction.

This conclusion that enhancers are highly sensitive to the arangement of binding sites is supported by other work with synthetic enhancers. In *Drosophila* embryos, synthetic sequences containing homotypic and heterotypic clusters of binding sites were sensitive to small changes in intermotif distances and motif orientations[41]. Furthermore, these constraints were tissue dependent, suggesting that changing concentrations of cofactors may effect constraints on *cis*-regulatory architecture. Similarly, synthetic sequences designed to probe the distance dependency of quenching found that this function is not monotonic [42] and depended on orientation[98]. In mouse, analysis of synthetic sequences containing various complexities of motif structure identified numerous pairwise synergistic interactions[178]. Moreover, synthetic sequences containing eight motifs drove highly variable expression depending on the arrangement[178].

4.4.3 Expression variability

We found that synthetic enhancers drove different levels of within-embryo expression variability that is independent of the mean (Fig, 4.8C,E-F). These distributions are reminiscent of distributions seen in a stochastic transcription model in which the number of transcripts and the ON-OFF state of the promoter are coupled random variables [77, 154]. We found that the width of distributions of mRNA levels, scaled to the mean, is altered by varying the probability that a gene is actively transcribing, p (Fig. 4.8D). While the mean expression level driven by six Dst motifs is far higher than that driven by six Zld motifs, the distribution driven by Zld is less variable than that driven by Dst. This indicates that the the probability that the promoter is in the ON state is lower when driven by Dst than by Zld. This difference could be explained by the role of Zld as a chromatin remodeler. While Dst can drive high levels of transcription, it must compete with nucleosomes for binding. Where Dst has displaced nucleosomes high levels of transcription are achieved, while adjacent nuclei are inactive. In contrast, if Zld can more easily displace nucleosomes, all nuclei will have

consistent levels of Zld binding. This leads to a high probability of the gene being in the ON state, even though Zld activates transcription more weakly than Dst.

4.5 Materials and Methods

4.5.1 Design of synthetic enhancer sequences

The method of optimizing sequence given a set of kinetic parameters is discussed in depth in Martinez et. al[123]. In brief, seven parameters sets, characterized in three previous works [93, 123, 122], were used to generate the sequences in this work. The parameters for these models are given in Table 7.1. Parameter sets 1, 2, and 3 are described Kim *et al.* [93], where they are called model 01, 06, and 07 respectively. Parameter set 4 was trained using all the data from Kim *et al.* as well as the expanded model in Martinez *et al.*[122]. This fit used the PWMs for Hb and Bcd, that are reported in Martinez *et al.* Parameter sets 5 and 6 were trained to the same data used by Kim et al. but used a different PWM for Bcd[69]. Parameter set 7 was obtained by training with the PWMs and data used in Kim et al, with the addition of Zld as a uniformly expressed activator at a constant level of 100. The Zld and Cic PWMs are from the Fly Factor Survey[208]. All PWMs used in this study are given in the Appendix.

The synthetic sequence e251 was designed using all seven parameter sets. In order to generate a sequence with well separated sites, we used the cost function

$$E = \Big(\sum_i (x_i - y_i)^2 \Big) + \beta o, \tag{4.1}$$

where x_i and y_i are the model output and data respectively, β is a configurable parameter, o is the number of overlapping motifs, and collectively βo is a penalty for overlapping binding motifs. We defined two motifs as overlapping if the end to end distance between their footprints was within 5 nucleotides, and we set β to be 1% of the maximum possible score,

corresponding to $x_i = 255$ for all x. To generate e251, we minimized the mean cost function

$$E_{\text{consensus}} = \frac{1}{7}\sum_{i=1}^{7} E_i, \tag{4.2}$$

where i denotes a parameter set from the set of seven described above. The initial sequence used was random.

The synthetic sequence s272 was designed by starting with a 258 bp DNA having the structure shown in Fig. 4.4A, but with each binding site having a maximum affinity consensus sequence. This sequence drove a predicted stripe 2 pattern a few nuclei anterior of the observed pattern when assessed using parameter set 2. The affinities for repressors Gt and Kr were then reduced such that the pattern was predicted to drive expression of a stripe at the position of stripe 2. This was accomplished with a single nucleotide change to the Gt consensus motif (TTACGCAAT to TTACGCAA<u>A</u>) and three changes to the Kr consensus (TAACCTTTC to <u>A</u>AACCC<u>A</u>TT<u>T</u>).

4.5.2 Design of enhancers by synthetic compensatory evolution

The method of generating synthetic compensatory paths is discussed in depth in two previous works [123, 122]. In brief, we select the synthetic sequence (e251 or s272), identify the number of single nucleotide edits required to mutate MSE2 into this sequence, then permute the order of edits such that at each step we minimize a cost function. We define the function

$$F = \sum_i \left(\frac{x_i}{\max_j x_j} - \frac{y_i}{\max_j y_j}\right)^2 \times \text{Penalty}, \tag{4.3}$$

where x_i is the predicted model output and y_i is the data. This function standardizes both data and output on a 0 to 1 scale. We penalize model predicted expression less than data

with the multiplicative penalty,

$$\text{Penalty} = \begin{cases} \frac{\max_i y_i}{\max_i x_i} & \max_i x_i < \max_i y_i \\ 1 & \max_i x_i \geq \max_i y_i \end{cases}. \tag{4.4}$$

For the path to s272, we minimize the the function

$$F_{\text{s272}} = \sum_{i=1}^{272} F_i, \tag{4.5}$$

where i denotes the sequence after i LE given the permutation being scored. Only model 2 was used in scoring.

For the path to e251, which uses consensus design, we minimized the function

$$F_{\text{consensus e251}} = \sum_{i=1}^{272} \sum_{j=1}^{7} F_{ij}, \tag{4.6}$$

where i denotes the sequence after i LE and j is a parameter set from the set of seven previously described.

4.5.3 Generation of reporter constructs

Reporter constructs where generated using a p*CaSpeR* backbone (GeneBank X81644.1) containing the promoter and first 22 amino acids of *eve* fused to *lac*Z, generated by Small et al.[174]. An AttB sequence was inserted into the multiple cloning site using the restriction enzyme *Xba*1 for insertion in the AttP2 landing site on chromosome 3[63]. The enhancer sequence was extended by PCR primers containing overlap with this vector (Appendix). The vector was then digested by enzymes *Eco*R1 and *Xho*1 and the enhancer was inserted using Gibson assembly[53]. The resulting vector was injected into flies of the genotype P{nos-phiC31\int.NLS}X, P{CaryP}attP2 by Rainbow Transgenics. Quantitative data was collected from these lines as previously described [82].

4.5.4 Sequences used in this work

The sequences of all 40 enhancers generated in this work are included in the Appendix. Additionally, expression data are provided in the Appendix.

4.5.5 Analysis of binding site conservation

In order to determine the number of binding sites gained, lost, or conserved between two sequences we first performed a pairwise alignment between two sequences using the R package Biostrings. The PWM score of binding was calculated at every position in each aligned sequence. Sequences were called binding sites if the PWM score was greater than zero. In order to accommodate gaps in sequence alignments, sites were considered conserved if they aligned within 3 bp. Sites were considered lost if there was no site with PWM score greater than 0 within 3 bp on the corresponding aligned sequence. For the background distribution we use the frequencies of nucleotides in the *Drosophila* genome ($P_{\mathrm{bg}}(A) = P_{\mathrm{bg}}(T) = 0.297$, $P_{\mathrm{bg}}(C) = P_{\mathrm{bg}}(G) = 0.203$).

4.5.6 Scaling of data for variation analysis

Variation in expression can be due to effects both within and between embryos. In order to remove the between embryo effects, we introduced a scaling factor for each embryo which multiplies the fluorescence measurements across the entire AP axis. We then optimized the scaling factors for each embryo in order to minimize the sum of squared differences in fluorescence measurements between embryos of the same genotype. This was subject to the constraint that the sum of scaling factors equals the number of embryos of that genotype. The scaled data from multiple embryos was then pooled for subsequent analysis.

4.5.7 Theoretical distribution of mRNA

The steady-state distribution of mRNA counts has been previously derived for a stochastic transcription model in which the number of transcripts and the ON-OFF state of the promoter are coupled random variables [157, Eq. 29]. This distribution is defined by three variables: p gives the probability of the promoter being in the ON state, N gives the transcription rate when the promoter is in the ON state, and b gives the rate of switching between ON and OFF states. Ramos *et al.*[157, Eq. 29] gives the distribution of mRNA when the promoter is in the OFF state, α_n, or ON state β_n. Here we report the total distribution $\phi_n = \alpha_n + \beta_n$, keeping the parameter b fixed at $b = 4$. The mean number of mRNA is given by $\mu = Np$.

CHAPTER 5

AN *IN SILICO* ANALYSIS OF GENE REGULATION LINKS ENHANCER LENGTH TO ROBUSTNESS

5.1 Abstract

Robustness assures that organisms can survive when faced with unpredictable environments or genetic mutations. Organisms must ensure that expression of genes is directed to the appropriate tissues at the correct times, but little is known about how gene regulatory systems are robust to perturbation. In this work we investigate the robustness of gene regulation using a sequence level model of the *Drosophila melanogaster* gene *even-skipped*. We find that gene regulation can be remarkably sensitive to changes in transcription factor concentrations at the boundaries of expression features, while it is robust to perturbation elsewhere. We also find that the length of sequence used to control an expression feature correlates with the number of nucleotides that are sensitive to mutation in both natural and *in silico* predicted enhancers. This indicates that sequence length and redundancy make gene regulatory systems more robust to genetic perturbation.

The work presented in this chapter is a manuscript to be submitted by myself, John Reinitz, and Ovidiu Radulescu. Ovidiu Radulescu developed the theory and perturbation schemes.

5.2 Introduction

Biological systems must be robust to perturbations, both environmental and genetic, in order to maximize the chance of survival in unpredictable circumstances[94]. Among such robust systems is organismal development, where canalization assures that all individuals arrive at the same phenotype despite individual variation[200, 173]. At the level of genetic networks, canalization has been observed within *Drosophila* development[120, 119]. The connections

in this network represent regulatory interactions which control levels of expression of genes through *cis*-regulatory elements called enhancers[108]. These sequences contain clusters[17, 141] of binding sites for transcription factors (TFs) that act in combination in order to direct gene expression at specific times or within specific tissues. While it is understood how the dynamics of developmental networks confers robustness to the system, it is less understood how or if the enhancers that control these networks contribute to robustness through their own internal structure.

The function of enhancers is the product of a complex set of interactions between bound TFs. Given this complexity, it is necessary to use quantitative methods and models in order to uncover the relationship between enhancer function and individual TF binding sites. Within *Drosophila* development, confocal microscopy has been used to generate spatial and temporal atlases of protein and mRNA levels at single nucleus resolution during the first 4 hours of development[114, 185, 184, 45]. This data provides the basis of sequence level models of gene regulation, which predict gene expression levels as a function of protein levels and TF binding sites [81, 171, 164, 92, 70, 93, 122, 167]. These models predict binding using thermodynamics and incorporate context-dependent rules of TF function such as competition for binding, short-range quenching[61, 60, 42], and coactivation[65, 174, 93]. Using such models, it is possible to address how the general principles of gene regulation, encompassed by these models, can confer robustness to enhancers.

In this work we use a previously reported model of gene regulation (Chapter 3) to model the robustness of the *Drosophila even-skipped* locus with respect to variation in both TF concentration and sequence mutation. We specifically assess two types of robustness: distributed robustness and r-robustness[58]. The former arises when many components contribute to the same output (i.e. the mean), while the latter refers to cases where only a subset of parameters are critical. We find that gene regulation can be extraordinarily sensitive to some changes in TF concentrations, a property that may help form sharp borders in expression domains. We find that the gene regulation is r-robust with respect to sequence mutation. Expression

is only sensitive to changes in a few nucleotides. We find that longer enhancers are sensitive to changes in fewer nucleotides in both natural and *in silico* generated enhancers, indicating that enhancer length confers robustness to genetic perturbation. Similarly, we find that longer enhancers also confer robustness to changes in TF concentration.

5.3 Results

5.3.1 Distinguishing types of robustness

Distributed and r-robustness arise in systems that contain many parameters. In the former case, the system is robust because because the it depends on not just one, but many parameters. In the later, weaker case, it is robust because it is insensitive to perturbation in some subset of parameters. Gorban and Radulescu[58] formalized these types of robustness that arise in systems with many parameters and investigated a the robustness of a well described signaling pathway. In this work we follow the definitions laid out in Gorban and Radulescu [58, Eqs. 1-2]. We consider the robustness of a function M, with parameters K indexed by i, $M = f(K_1, K_2, \cdots, K_n)$. The system is robust with respect to changes in these parameters if the variance in M is reduced compared to the variance in the parameters K. That is, if the variance in all parameters $\mathrm{Var}(\log K_i) = \mathrm{Var}(\log K)$, then we consider the system to be robust if

$$\mathrm{Var}(\log M) \ll \mathrm{Var}(\log K). \tag{5.1}$$

Similarly, if we consider an index of r parameters $I_r = i_1, i_2, \cdots, i_r$, which we multiply by positive independent random scales s_i ,we define M as r-robust if for any I_r

$$\mathrm{Var}(\log M) \ll \mathrm{Var}(\log s). \tag{5.2}$$

To distinguish between these types of robustness, it is useful to study the relationship between the variance of parameters and the variance in the output, or similarly to observe

the variance in the output given the number of parameters perturbed. For instance, consider a system that is r-robust. In this system, there are n parameters, of which n_0 parameters are sensitive to perturbation. If we select r of n parameters at random, the probability we did not select a sensitive parameter is $(1 - n_0/n)^r$. Then the probability that at least one sensitive parameter was selected is $1 - (1 - n_0/n)^r$. If sensitive mutation contributes V_0 to the variance, and the effect is not cumulative, then the variance in M with respect to r mutations is given by

$$V(r) = (1 - (1 - n_0/n)^r)V_0. \tag{5.3}$$

We call this system r-robust, and the variance saturates with respect to the number of perturbed parameters.

In contrast, consider the function $M = \text{mean}(K_1, K_2, \cdots, K_n)$. If all parameters have a variance of V_K, then variance will be cumulative, and the variance of M with respect to r will be

$$V(r) = \frac{V_K r}{n^2}. \tag{5.4}$$

This function grows linearly with respect to r. If all parameters are are perturbed, the variance of of M is simply

$$\text{Var}(M) = V_K/n. \tag{5.5}$$

and the variance grows linearly with the variance of individual parameters. This equation illustrates distributed robustness. While the mean gives simple linear functions on these planes, some robust functions give more complex results (for an example, see Gorban and Radulescu [58, Fig. 6]).

In the cases of r-robustness or distributed robustness we can assign a number that describes the robustness of the system. When there are critical parameters, n_0 describes the robustness of the system, and systems with lower n_0 are more robust. For distributed systems, robustness is described by the ratio of variance of the input to variance of the output, a parameter we call ρ. For much of this work we will be working with variables of unknown

scale. In light of this, we observe the variance in the fold-change of the input and the fold-change of the output. If the original values of M and K are M' and K' respectively, we call the fold change

$$\Delta M = M/M'; \Delta K = K/K'. \tag{5.6}$$

The robustness ratio ρ is then given by

$$\rho = \frac{\mathrm{Var}(\log(\Delta M))}{\mathrm{Var}(\log(\Delta M))}. \tag{5.7}$$

5.3.2 Robustness of mRNA levels with respect to transcription factor concentration

Enhancers interpret the local concentration of transcription factors in order to specify appropriate production of mRNA. In order to determine the degree to which known regulatory mechanisms acting on enhancers contribute robustness to fluctuations in TF concentration, we utilized a model that simulates the regulation of *Drosophila even-skipped* (*eve*)(Chapter 3), which is expressed in seven transverse stripes across developing embryos. This model incorporates several mechanisms: TF-DNA binding and competition for binding based on thermodynamics, short-ranged quenching of transcriptional activators[61, 60, 42], coactivation of repressors[65, 174, 93], and competition for interaction with the basal transcriptional machinery. The model is able to accurately fit the expression pattern of stripes between 35.5% and 92.5% embryo length (5.1A), identify the enhancers within the *eve* regulatory locus, and simulate the effects of mutations to the environment in *trans*.

To test how this system responds to fluctuations in TF levels, we simulated changes in TF levels by multiplying by $\exp(AX)$, where A is a parameter that sets the size of fluctuations, and X is a random uniform number between -1 and 1. Because TF levels are in arbitrary units, we observe the fold change in TF levels and the fold change in resulting mRNA (Eq 5.6) after simulating 10,000 fluctuations.

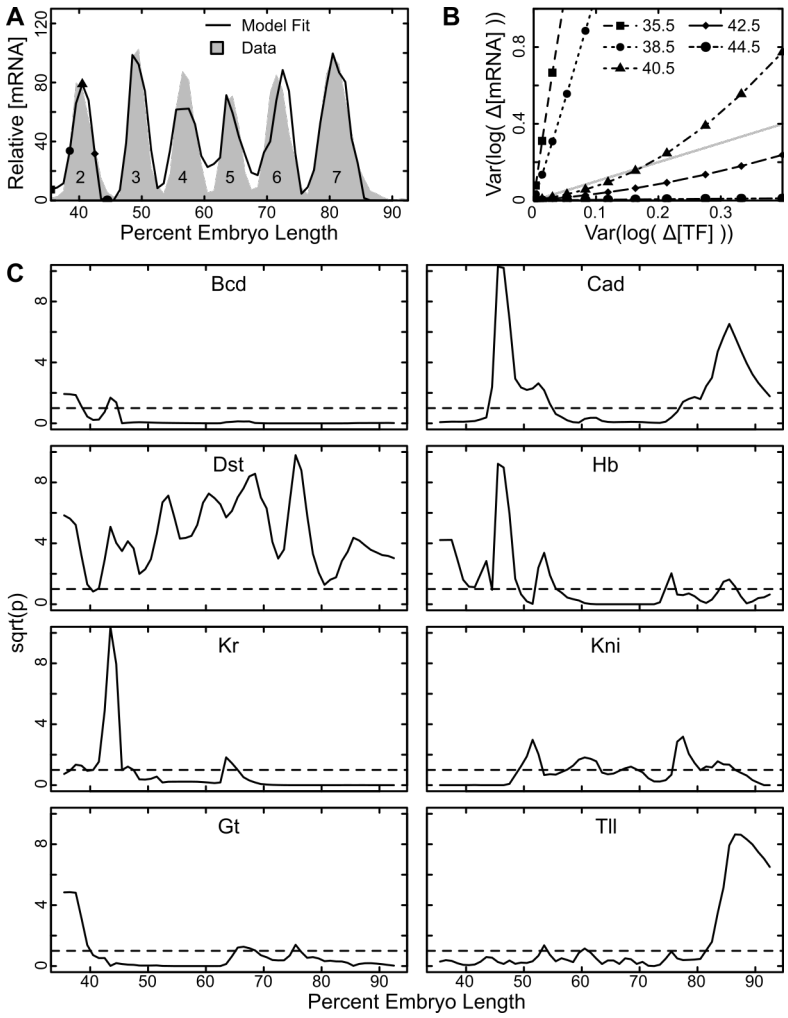


Figure 5.1. Robustness of *eve* expression to variation in TF concentration. (A) The relative expression of *eve* mRNA along a 10% wide stripe along the anterior posterior axis (gray shading) and the model fit to the same data (black line). *eve* stripe number is indicated. (B) The relationship between variation in fold-change TF concentration and fold-change mRNA levels (Eq 5.6). The points are plotted for the AP positions indicated. The line representing $\rho = 1$ (Eq 5.7) is indicated with a gray line. Points below this line are robust, while points above are sensitive. (C) The ratio of the variance in fold-change mRNA to the variance of fold change TF concentration ρ (Eq 5.7) for the indicated TF at each position in the embryo. ρ values have been square root transformed for better visual presentation. The dashed line indicates the value $\rho = 1$. Perturbation size A was set to 0.1.

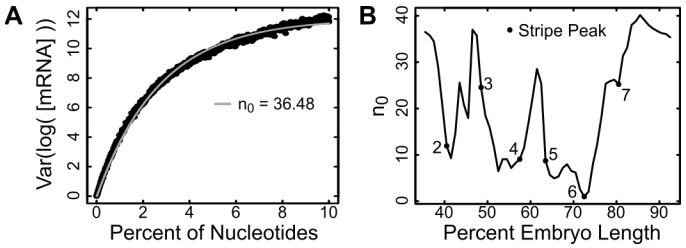


Figure 5.2. Robustness of *eve* expression to variation in DNA sequence. (A) The variation of *eve* expression at 35.5% embryo length at various values of r nucleotides that are mutated. The fit to data (Materials and Methods) is shown as a grey line, and the estimated number of sensitive nucleotides, n_0, is indicated (Eq 5.3). (B) The number of sensitive nucleotides, n_0 at every position along the AP axis.

We find that sensitivity to fluctuations in individual TFs varies with respect to position in the embryo. For instance, if we observe sensitivity to fluctuations in the TF Giant (Gt) at the interstripes, borders, and peak of the second *eve* stripe (positions indicated in Fig 5.1A), we find that at the anterior interstripe and border, expression is not robust to changes in Gt regardless of the magnitude of fluctuation (Fig 5.1B). In contrast, the posterior interstripe of stripe 2 is insensitive to fluctuations in Gt. Notably, at the peak of stripe 2, expression is robust against small fluctuations and sensitive to large ones (Fig 5.1B).

In general, we note that expression at interstripes is more sensitive to fluctuating TF levels than expression at stripe peaks (Fig 5.1C,7.13). For fluctuating repressors, the expression border that is set by each factor is sensitive to fluctuations in concentration.

5.3.3 The eve *locus is r-robust with respect to nucleotide changes*

Genetic systems may also be robust with respect to changes in DNA sequence. In order to investigate the robustness of the *eve* locus with respect to sequence perturbation, we simulated random mutations to r nucleotides 10,000 times, with r spanning 1 to 10% of all nucleotides. This results in a curve that saturates with increasing r (Fig 5.2A). This saturating curve is well described by a system with critical parameters. When we fit the Eq 5.3 at every position along the AP axis, we find that stripe peaks are more robust to perturbation than interstripes in that they have fewer critical parameters (Fig 5.2B).

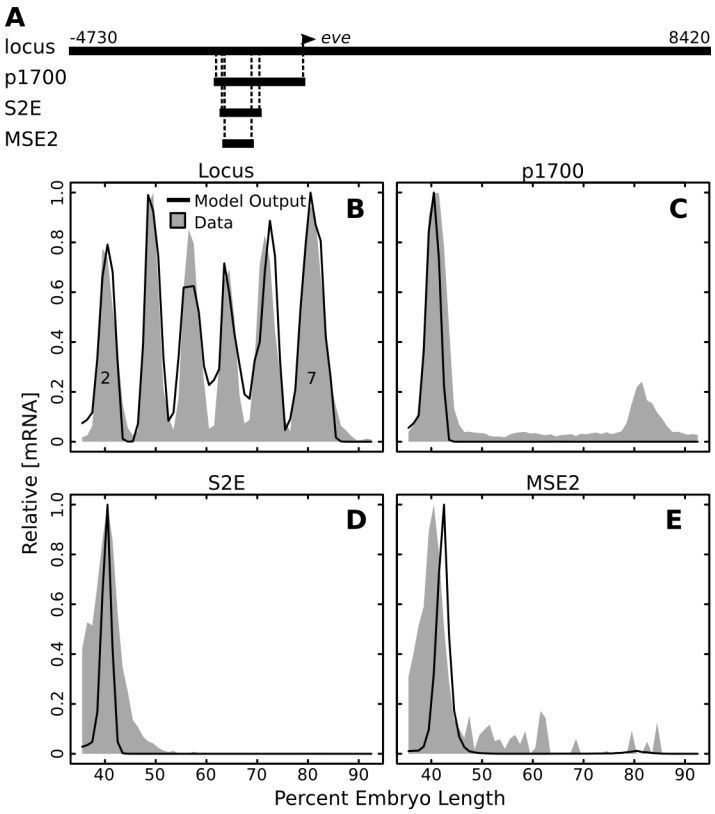


Figure 5.3. Predicted expression of successively smaller enhancers. (A) Cartoon diagram showing the entire *eve* locus and successively smaller sequences that all drive stripe 2. Where each sequence aligns within the locus is indicated with dashed lines. Position with respect to TSS is indicated. (B-E) The model predicted expression and actual expression for each of the sequences in (A) along the AP axis. Model output is in black lines and expression data is in gray shading. Because the true scale is unknown, relative expression on a 0 to 1 scale is reported.

5.3.4 Longer eve *enhancers are more robust to perturbation*

For the second stripe of *eve*, four different sequences have been reported to drive expression: the intact locus, the proximal 1700 bp, S2E, and MSE2. Each larger sequence contains the sequence of all smaller sequences (Fig 5.3A). In order to investigate whether additional sequence contributes additional robustness, we investigated the number of sensitive nucleotides n_0 in each of these four sequences at the peak of stripe 2 expression (40.5% embryo length). We find that that as sequence length grows, not only does the ratio of sensitive nucleotides decrease, the absolute number of sensitive nucleotides decreases from about 26 to 12 (Fig 5.4A-D).

We also investigated robustness to changes in TF concentration for each of the stripe 2 enhancers. For the majority of TFs, there was a relationship between the size of the enhancer and robustness to TF concentration (Fig 5.4E). In all cases, the intact locus was the most robust to changes in TF concentration. With respect to some TFs, MSE2 was more robust than S2E or the proximal 1700bp.

5.3.5 Robustness is a function enhancer length

In order to determine whether the relationship between sequence length and robustness to mutation is an evolved property or one inherent to the system, we generated 8010 putative stripe 2 elements. We generated 10 putative S2Es with each length from 200 bp to 1000 bp. All 8010 S2Es are predicted to to drive the correct expression pattern (Fig 5.5A). We estimated the number of sensitive nucleotides n_0 for each of these S2Es by simulating 10,000 sets of r sequence mutation with r spanning from 1 bp to 10% of the sequence length. We estimated $n0$ at the peak of stripe 2 expression (40.5% embryo length). We found that $n0$ decreases with enhancer length (Fig 5.5B).

We also investigated robustness to changes in TF concentration for each of these 8010 enhancers. We measured ρ from 10,000 simulations with $A = 0.1$. For most TFs, there is a relationship between enhancer length and robustness to TF concentration. Enhancer length

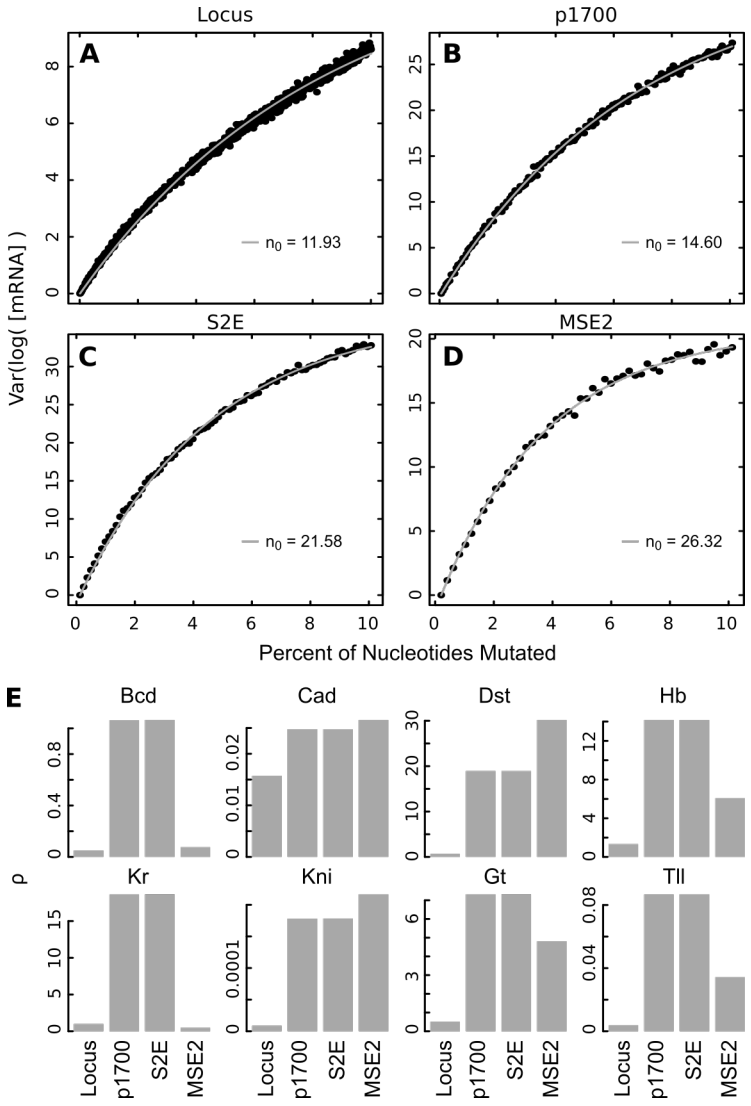


Figure 5.4. Robustness robustness of natural S2Es. (A-D) The variation in mRNA expression is reported for increasing subsets of r nucleotides mutated is reported for each natural element that drives *eve* stripe 2. The best fit curve and estimated number of sensitive nucleotides n_0 is indicated. (E) The ratio of variation in mRNA to variation in TF concentration ρ (Eq 5.7) for each TF and each S2E. A was set to $A = 0.1$ for simulations.

contributes to robustness for the factors Bcd, Hb, Kr, Kni, Gt, and Tll ($p < 2 \times 10^{-16}$), but was not significant for Cad or Dst (Fig 5.5C).

5.4 Discussion

In this work we assessed the robustness of enhancers with respect to changes in TF levels or sequence mutation in the context of a sequence level model of gene regulation. The approach here allows us to ask to what degree the mechanisms included in the model lend robustness to *cis*-regulation. The particular model used includes a calculation of TF occupancy on DNA, which includes competition for binding sites and cooperativity of binding. In addition, the model considers coactivation, quenching, synergistic activation through a diffusion-limited Arrhenius rate law, and competition between sequence elements for interaction with the basal transcription complex.

We found that the transcription rate has varying degrees of sensitivity with respect to the concentration of TFs, however the system is only robust where concentrations of TFs approach zero. Instead, transcription rate was incredibly sensitive to changes in the concentration of TFs, especially at stripe borders. For instance, the anterior border of *eve* stripe 2 is controlled by Gt. When Gt levels fluctuate at this embryo position, transcription levels fluctuate to a greater degree. Such sensitivity may be a necessary feature of the circuit, where high sensitivity to individual repressors allows the formation of extremely sharp borders. In contrast, we found that robustness to sequence mutation is well-described by a system with a small number of sensitive parameters.

Early experiments with enhancers sought to find the minimum fragments that could recapitulate an expression feature [174], however the true size of enhancers is largely unknown and sequence outside of minimal fragments may contribute to robustness. For well studied regulatory modules, TF binding sites outside of the minimal regions have been found to influence expression levels of reporters (Fig 7.1). It has been proposed that redundancy built into enhancers ensures robust expression in appropriate tissues without disrupting specificity[43].

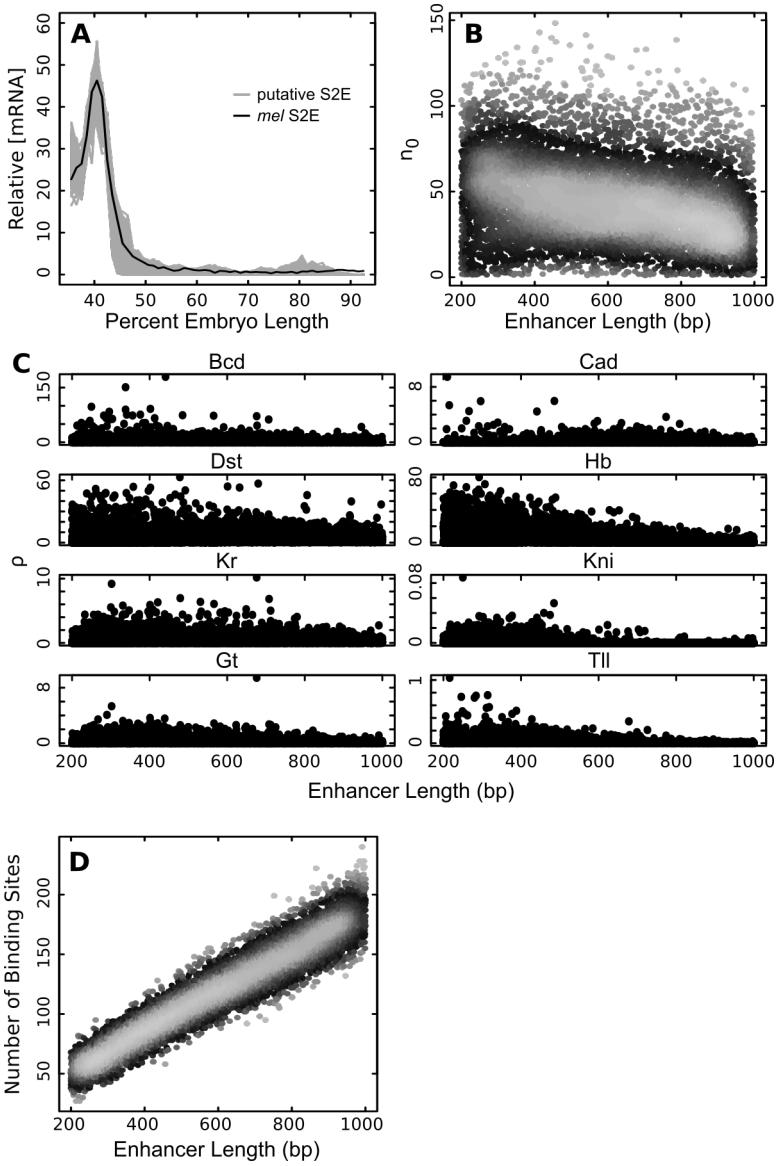


Figure 5.5. Sequence robustness of putative S2Es. (A) Predicted expression of 8010 putative S2Es with lengths from 200 bp to 1000 bp (gray lines). Each putative S2E is predicted to drive expression similar to the natural *D. mel* S2E. (B) The number of sensitive nucleotides n_0 *vs.* sequence length for each of the 8010 putative S2Es. The points are colored according to their local density. (C) The ratio of the variance in mRNA to the variance in TF levels, ρ (Eq 5.7) for each of the 8010 putative S2Es for each TF. ρ was estimated with $A = 0.1$. (D) The number of modeled binding sites in each putative S2E *vs.* sequence length for each of the 8010 putative S2Es. The points are colored according to their local density.

Indeed, while the minimal region of enhancers may be sufficient under normal conditions, environmental perturbation can disrupt the function of enhancers that lack sufficient levels of redundancy [111, 34]. Moreover, enhancer elements themselves are often redundant [73, 27, 40, 179], ensuring robust expression of genes [149, 46, 203].

When we investigated the relationship between sequence length and robustness we found that longer sequences were less sensitive to both changes in TF concentration and sequence mutation. This property was not unique to natural elements, as it was also true of a set of putative S2Es generated *in silico*. While not significantly different, Both MSE2 and S2E driven expression was sensitive to changes in fewer nucleotides (26.3 and 21.6) than *in silico* enhancers of the same length (average of 49.5 and 37.1), indicating that robustness to sequence perturbation may be an evolved property. The high correlation between length and number of sites (Fig 5.5D) indicates that redundancy in binding sites is probably the cause of increased robustness. Supporting this, we find that MSE2 and S2E have a higher density of sites (114 and 204) than the *in silico* enhancers of the same length (average of 100.6 and 141).

5.5 Materials and Methods

5.5.1 Model selection

The model used in this work is the same as reported in Chapter 3. The parameter set used was the best model including chromatin state information, called 'Repeat Chromatin #2' in that work.

5.5.2 Simulations of TF concentration perturbation

We perturbed TF concentration by first selecting a random uniform number X between -1 and 1. Then we multiplied the TF concentration along the entire AP axis by $A\exp(X)$, where A is a set parameter that scales the size of perturbation, and we observed the predicted change

in mRNA at all AP positions. If [TF] is the initial concentration and [TF]' is the perturbed concentration, we define Δ[TF] as the ratio [TF]'/[TF]. Similarly, we define Δ[mRNA] as the ratio [mRNA]'/[mRNA]. We repeated this calculation 10,000 times for each value of A between 0 and 3 in increments of 0.1, for a total of 310,000 simulations. This was repeated for each of the 8 TFs included in the model.

5.5.3 Simulations of sequence mutation

In the intact locus model, some nucleotides are not accessible for TF binding. We perturbed DNA sequence by selecting sets of r nucleotides only from the accessible set. The r nucleotides were then substituted by one of the remaining three possible nucleic acids with uniform probability. We then assessed model output at each AP position. This was repeated for every set r from 1 to 877, which represents 10% of the total accessible nucleotides. Similarly, we assessed the ratio of peak expression to each interstripe.

5.5.4 Estimation of sensitive nucleotides

To fit Eq. 5.3 to data, we used simulated annealing from the R package GenSA, using default parameters. n was set to the total length of accessible nucleotides, or 8765 for the locus, 1726 for p1700, 804 for S2E and 484 for MSE2.

5.5.5 Enhancer-reporter assays

The locus expression, as well as the expression of the S2E and MSE2 construct are reported in Barr and Reinitz(Chapter 3). The p1700 data is from Janssens *et al.*[81].

5.5.6 Generation of putative S2Es

To generate putative S2Es of different lengths, we fixed the kinetic parameters and optimized DNA sequence using previously described methods[123]. We used the expression of S2E as

the target. We started each optimization with a random sequence of the desired final length.

CHAPTER 6

DISCUSSION

6.1 Chapter overview

In Chapter 1, I introduced the concept of *cis*-regulatory DNA, the *cis*-regulatory code, and the enhancer, which directs tissue specific expression of structural genes independent of position and orientation. I discussed how interactions between TFs bound to enhancers gives rise to enhancer logic, or a specific set of rules for how any particular set of binding sites specifies a particular expression level. Finally, I introduced quantitative data and physical-chemical models of gene regulation and discussed how they have been applied to *Drosophila* enhancers.

In Chapter 2, I discussed the computational complexity of calculating the occupancy of transcription factors binding to DNA. Past approaches used a calculation with a computation time that scaled exponentially with the number of binding sites. This was especially problematic when attempting to model systems with large numbers of TFs or when using low thresholds for identification of binding sites. To resolve this, I introduced an algorithm that is able to compute TF occupancy in linear time with respect to the number of binding sites identified. The improved algorithm is several times faster for typical problems, can calculate all possible pairwise cooperative interactions, and is efficient enough to be applied at the genomic scale.

In Chapter 3, I extended a model of gene regulation to the intact *eve* locus. Typically, enhancers have been considered additive in their activity, yet sites within an enhancer act synergistically. I introduced a new mechanism that naturally separates these regimes. When steric competition limits the quantity of DNA that can interact with the basal transcription machinery, spatially separated TF binding sites cannot synergistically interact because they never simultaneously interact with the basal complex. I called this new mechanism 'enhancer competition'. When I incorporated this mechanism into the transcription model described

in Chapter 2, the enhancer structure of *eve* naturally arose from the structure of underlying binding sites.

In Chapter 4, I tested the sufficiency of interactions included in the Reinitz lab transcription model. I accomplished this through the synthesis of 40 synthetic putative enhancers of *eve* stripe 2, each with varying degrees of homology to natural elements. If the mechanisms were entirely sufficient, it would be possible to use them to design enhancers *ab initio*. Instead, a subset of enhancers, especially those with closer homology to MSE2, drove expression, while the majority of synthetic sequences failed to drive detectable expression. Analysis of these enhancers revealed that the context dependent rules that determine enhancer logic may be far more complex than initially treated by the model. I highlight Bcd cooperativity and Hunchback coactivation as two mechanisms that have more complex interaction.

In Chapter 5, I tested the degree to which the mechanisms in transcription model contribute robustness to enhances. While I found that the *eve* locus is quite sensitive to changes in TF concentration, especially at stripe borders, I found that the locus is robust with respect to sequence mutation. I showed that there is a relationship between the length of enhancer and their robustness to changes in concentration of transcription factors and sequence mutation.

6.2 Possible mechanisms acting on enhancers

The central goal the this work presented in this dissertation has been to interpret the cellular instructions provided by enhancers according to physical-chemical laws. While models of enhancers have been broadly successful in predicting the behavior of evolved sequences, as evidenced by the success of the intact locus model presented in Chapter 3, I show in Chapter 4 that using such models to design enhancers *ab initio* gives inconsistent results. The key question that remains is why these models have been successful in describing evolved sequence[93, 122], but fail to predict the activity of synthetic sequence. In the sections that follow I will discuss potential avenues of investigation. For each topic, I will discuss how

it is currently treated in the model, if at all, and how the current treatment may differ from reality. Where there is ambiguity, I will propose experiments that could resolve that ambiguity. Presumably, some of these mechanisms could be included in future modeling work, leading to more complete and predictive models of gene regulation.

6.2.1 Additional transcription factors

There are likely more transcription factors interacting with MSE2 than the ones considered in this work. This was indicated by the lack of known motifs that might explain the loss of expression from e60 to e72. Additional factors acting on MSE2 has also been highlighted by other work[7]. The challenge is to identify what these factors are. This is an especially difficult problem for ubiquitously expressed activators, where loss of binding causes a loss of expression everywhere. This problem is somewhat mitigated by more complete information on the binding preferences of all *Drosophila* TFs, combined with data to say whether any TF is expressed in an embryo or not.

6.2.2 Position weight matrices

The first step in any model of gene regulation is to identify binding sites for TFs. If sites are not assigned proper binding affinities, all downstream calculations will be affected by the propagated error. The fundamental assumption underlying the PWM is that the individual nucleotides contribute independently to the affinity. While some TFs appear to have multiple modes of recognition[9], the vast majority of binding events appear to be well described by simple PWMs[207]. Regardless, for PWM scores to yield true affinities, it is extremely important to obtain high quality binding data from sites with known energetic distributions. Methods like SELEX[142], which rely on multiple rounds of affinity purification and selection, lose information about the changes in frequencies of motifs from one round to the next, making it impossible to know the true relative affinities. Modern high-throughput analogs [206, 143] can rectify this problem, as can microfluidic platforms

that measure true affinity[117], as long as care is taken to ensure physiological salt levels are maintained. Such approaches have already been used on many *Drosophila* and human TFs[87, 138]. The PWMs currently used in the model probably do not perfectly represent true *in vivo* binding preferences, and future work on *eve* could greatly benefit from such high-quality binding data.

6.2.3 Non-specific affinity and thresholds

Interactions between TFs and DNA can be specific or non-specific, where the former is sequence-dependent and the later comes from simple electrostatic interactions with negatively charged DNA, independent of sequence. It is currently unclear what, if any, contribution non-specifically bound TFs have towards regulating expression of genes. There is some evidence that TFs can be recruited to DNA by specifically bound TFs[89]. In this work, I introduced a new method that considered each site to be competing between specific and non-specific states, where the non-specific interaction was one thousandth of the maximum specific energy, an approximation from binding data[117]. This approach was likely conservative, and overestimates specific binding at low affinity sites. In reality, the TF not only needs to compete with non-specific binding at that site, but also at each adjacent site that can sterically compete, as well as with every other TF that can also bind non-specifically. With estimates that 90% of TFs are bound to the genome non-specifically[19], competition for non-specifically bound TFs probably represents a significant barrier to binding at low affinity sites. Unfortunately, calculation of occupancy that fully considers non-specific binding requires considering every single nucleotide a potential binding site for every expressed TF, providing a substantial computational burden.

6.2.4 Thermodynamic equilibrium

Among the fundamental assumptions of sequence to expression models is that of thermodynamic equilibrium. This assumption is valid provided there is sufficient time for the system

to reach equilibrium and energy is not being expended. It is clear from our data, and the fact that transcription rates can double or half within 6.5 minute timeclasses, that transcription occurs in the order of minutes. There are relatively few eukaryotic transcription factors for which true residency times have been measured, but where it has been measured it is in the tenths to tens of seconds[91, 22]. This indicates that binding and unbinding to DNA happens at sufficient frequency to justify calculating average factor occupancy. This is in stark contrast to residency time for *LacI*, which binds to specific operators for 5 minutes[64]. More data must be acquired for residency times of metazoan TFs, and if times prove to be longer a simple occupancy model will be inadequate. Already, such models may run into problems if nucleosomes are to be included. While these rapidly switch between wound and unwound states with DNA at sufficiently fast rates[105], these molecules are modified in an energy dependent process. For non-equilibrium processes it is possible to consider a stochastic master-equation approach[3], though it is unclear whether such an approach could be efficiently computer for complex metazoan enhancers.

6.2.5 Nucleosomes

Transcription factors are not the only complexes that occupy DNA. Large histone octamers, known as nucleosomes, also occupy 147 bp stretches of DNA[35]. The fact that TFs must displace nucleosomes for DNA binding lends indirect cooperativity to all TF-DNA interactions [127]. It is unlikely that thermodynamic estimates of TF-DNA binding are accurate, especially for low affinity binding sites that must compete with nucleosomes, unless the calculation takes nucleosomes into account. Fortunately, sequence-dependent binding models of nucleosomes exist[196], as do efficient frameworks for incorporating them into thermodynamic models in both fully and partially wound states[191].

6.2.6 Quenching, Coactivation, and Protein-Protein Interactions

In this work I used a simple distance-dependent function to determine the efficiency of protein-protein interactions. In this function, efficiency is 100% up to a fixed distance, then declines linearly with increasing distance. The distance function for quenching is based off relatively few examples[61]. More recently, it has been shown that this relationship is non-monotonic and is affected by the organization of binding sites[42]. In other words, sites that fall between a quencher and its target modulate the efficiency of quenching. It may require significantly more observed quenching interactions to determine the true distance-dependence of quenching.

Some clues may come from understanding how quenching works. The majority of quenchers in this model work by recruiting the transcriptional corepressor C-terminal Binding Protein (CtBP)[135, 134, 133]. The specific molecular function of CtBP is still unclear. While CtBP is able to recruit histone deacetylases (HDACs)[172], the repressive function is still present in HDAC null[118]. CtBP could act through chromatin modulation or through steric competition with transcriptional adaptors. If the former, it is unclear whether the feed-forward modeling approach is appropriate. In the latter case, we might expect the quenching function to be periodic with respect to position along the DNA double helix and the persistence length of DNA. In either case, it would be valuable to further define the quenching function at high resolution. The ability to generate fusion proteins with corepressors that can be specifically targeted to DNA could provide a useful tool to dissect the action of corepressors independent from DNA-binding TFs[148].

There is even less data to suggest a molecular mechanism underlying coactivation of Hunchback (Hb) by Bicoid (Bcd) or Caudal (Cad). One hypothesis is that Hb and Bcd can cooperatively recruit a transcriptional adaptor. While alone, Hb might have higher affinity for a corepressor, but when bound in proper stoichiometry with respect to Bcd the corepressor is displaced by cooperative recruitment of an adaptor. Such a system could be highly sensitive with respect to orientation, which would explain my inability to generate

synthetics that displayed Hb coactivation in Chapter 4. Again, decoding this function will require exhaustive testing of all distances and orientations between Bcd and Hb.

I suspect that both quenching and coactivation interactions are mediated by cofactors. It is possible that the concentrations of these molecules vary from tissue to tissue. This is indicated by the fact that in a panel of synthetic enhancers, the sensitivity to position and orientation of binding sites varied from tissue to tissue[41]. If adaptors and coactivators do vary between nuclei, modeling could substantially benefit from their inclusion. At least two such cases exist where AP patterning TFs have modulated effects due to the presence of cofactors: The ubiquitin ligase Degringolade (Dgrn) modifing the repressive effects of Hairy[13], and the F-box protein Dampened (Dmpd) decreasing Bcd-driven activation in the head.

6.2.7 Synergistic activation

Two general approaches have been used to model synergistic activation. This work and its intellectual predecessors rely on an Arrhenius rate law, where a linear reduction in an activation energy barrier leads to an exponential increase in transcription rate. The sigmoid function used by Segal and colleagues[171] is functionally equivalent. In contrast, the multiplicative synergy used by He and colleagues[70] is based on the cooperative recruitment of polymerase. Both approaches have limitations. For instance, it unclear how two proteins of the same type could catalyze the same kinetic rate step in such a way as to cause an additive reduction in the activation barrier. Similarly, if we imagine that TFs cooperatively recruit polymerase, at its limit TFs would bind polymerase so strongly as to prevent its clearance from promoters.

How synergy might arise from multiple TF binding has been the subject of theoretical analysis[168]. Depending on the number of kinetic rate steps catalyzed, multiple TF binding can take on varying degrees of synergy. If different TFs act on different kinetic rate steps, multimerization of single TF binding sites might cause saturation of transcription levels,

but if a site for a single TF that acts on a different rate step is added, transcription levels could be higher than the previous saturation point. This type of synergy can explain why in panels of synthetic enhancers in mouse cells, the highest transcription rates were achieved from enhancers containing the highest diversity in types of TF binding sites[178]. Sequences that test the effects of different numbers and combinations of activator binding sites could uncover the number of kinetic rate steps that are controlled.

6.2.8 Enhancer competition

In this work I treated the interaction frequency of enhancers and promoters as a variable that is coupled with the strength of enhancers. While this perspective was effective for the *eve*, necessary to limit the number of free parameters, and is broadly consistent with literature[51], this interaction is likely far more complicated than treated here. Different TFs may have different abilities to promote interaction with the promoter, indicated by the fact that some sequences are required for action of enhancers at a distance, but are dispensable when the enhancer is adjacent to the promoter[187]. Moreover, the competition described here does not explain how multiple enhancers could drive synergistic expression, which has been observed for some shadow enhancers[24]. Additionally, interactions may exist between enhancers themselves and not simply between enhancers and promoters. Three approaches could be used to tackle this problem. First, 5C[37], or other variations on chromatin capture could be used to collect high-resolution data on enhancer-promoter interactions. Second, imaging techniques can be used to directly observe enhancer- promoter interactions[28]. Finally, synthetic enhancers could be test both adjacent to the TSS and at a distance to find sequence elements responsible for action at a distance.

6.2.9 Chromatin State

In Chapter 3, I showed that chromatin state information is necessary to model the behavior of the intact *eve* locus, and that repressive chromatin prevents ectopic expression due to

latent enhancers. Such an approach was possible in this work because DNAse-seq[106, 193] and FAIRE-seq[126] datasets were available. The role of ubiquitously expressed TFs in establishing this environment suggests that this chromatin state does not vary along the AP axis. Other groups have also incorporated chromatin information into models of gene regulation[147] or occupancy[90] however, these approaches rely on a sigmoid function that effectively classifies chromatin into open and closed domains and are functionally similar to the binary approach presented here.

Ideally, chromatin state would be inferred from the underlying structure of binding sites. This information would be necessary if accessibility changed from cell to cell or tissue to tissue, as well as in tissues where data has not been collected. Genomic data suggests that Zelda and Trithorax-like both act to establish open chromatin at enhancers and promoters, respectively[166]. More data will be required to precisely establish how Zelda establishes open chromatin and constructs that vary the position and number of Zelda sites with respect to an enhancer could help establish the reach and dynamics of chromatin remodeling.

It is possible that chromatin remodeling propagates. Such a phenomenon has been described for repression by Hairy, where the distance at which repression propagates varies with respect to the number of Hairy binding sites[12]. Dynamic models of front propagation have been proposed to explain these phenomena[170]. Using such models, it may be possible to solve for the equilibrium position of chromatin domains as a function of binding sites for chromatin openers or silences.

6.2.10 The cis*-regulatory code.*

The use of the word 'code' in the *cis*-regulatory code suggests that that there is a system of rules that can be used to understand the instructions present on DNA. Given the complexity of the molecular mechanisms acting on enhancers, it is possible that there is no general framework applicable to all enhancers. Instead, each enhancer could implement an entirely unique molecular solution to the problem that evolution has presented. If this fatalistic

perspective were correct, we would expect enhancers to be both highly conserved and highly fragile. If the molecular solution requires very precise binding site positioning such that it is unique to a single element, any change in that positioning should disrupt function. In other words, we might expect all enhancers to be enhanceosomes. The existence of a *cis*-regulatory code is supported by the fact that many enhancers are poorly conserved and rapidly evolve [110, 113, 109, 112, 68, 67, 186, 198], and by the success of modeling approaches towards understanding naturally selected enhancers[81, 171, 70, 93, 122].

6.3 Future directions

There are many potential avenues for expanding the modeling approach presented in this work, including non-equilibrium dynamics, non-specific binding and recruitment, direct treatment of cofactors, multiple types of activation, and chromatin dynamics. In the absence of hard data, it is difficult to know which avenue is the best step forward. This is especially true when trying to understand the activity of naturally selected elements that contain dozens of TF binding sites, and may contain sites for as of yet undiscovered factors.

Piecing apart these interactions is most easily accomplished with small sequences containing only known binding sites, where regulatory interactions can be isolated. Situations where the binding sites are fixed, but differences in distances change, are particularly powerful, and was one of the major reasons that a study of enhancer fusions proved to be a good data set with which to train the transcription model[93]. Unfortunately, simple sequences with only known binding sites may not drive any expression. What is needed is a system that is sensitive to additional activation or repression. Such a system could be accomplished by adding sites to an existing enhancer. With some activation provided by the enhancer, any small contribution from an added pair of sites would be easily detectable.

An extraordinary number of genetic constructs will be required to piece apart even simple pairwise interactions. In the ideal experiment, every single distance up to about 200 bp should be tested. Single nucleotide resolution ensures that effects due to helical orientation

can be sufficiently resolved. To detect interactions that can span longer stretches of DNA, distances longer than the persistence length of DNA should also be tested. If these experiments are repeated for different orientations of TFs, in head to head, head to tail, and tail to tail orientations, the number of required constructs triples. Finally, intervening DNA that separates binding sites might introduce unintended binding for other factors. In the absence of known binding preferences for all expressed TFs, such ectopic binding is unavoidable. The best solution is to repeat experiments with different intersite sequences, as random sequences are unlikely to create the same site or effect twice. In total, it might take 1200 sequences to rigorously test a single interaction. While significantly smaller subsets could provide informative data, it is clear that testing this number of constructs in *Drosophila* embryos would be impractical with existing technology.

6.4 Emerging technologies

It is now possible to perform reporter assays in parallel in cell lines. Two technologies have emerged, the Massively Parallel Reporter Assay (MPRA)[146] and STARR-seq[8]. The former technology takes advantage of the independence of enhancer position and places the enhancer downstream of a TSS such that it drives transcription of itself. The later technology involves synthesized sequences that contain a cloning site and a barcode. A promoter is then cloned in between the enhancer and barcode such that the enhancer drives transcription of the barcode. In both cases, a complex library of reporter constructs cloned into a cell line and RNA-seq can be used to count the relative number of transcripts driven by any given enhancer. Tens of thousands of enhancers can be tested in a single experiment.

MPRAs and STARR-seq offer the potential to exhaustively test pairwise interactions between transcription factors. While current technology does not allow these experiments to be performed in *Drosophila* embryos, experiments in cell lines could provide valuable information about the types and complexity of interactions that occur on enhancers, which could later be verified in live organisms. Sea squirts are now emerging as a powerful model

organism for the study of enhancers. It is possible to introduce MPRAs into sea squirts through electroporation, allowing unprecedented parallelization while preserving the ability to study enhancer activity across diverse cell types in a live organism[43].

The the ability to generate orders of magnitude more data in the form of reporter constructs provides an exciting new avenue for future investigations. The efficient code presented in Chapter 2 is capable of analyzing these larger datasets. Moreover, this code is most efficient when parallelized over large numbers of constructs, rather than large numbers of tissues. Carefully constructed assays, combined with the modeling framework presented in this work, will allow the logic of *cis*-regulation to be deconstructed at unprecedented resolution.

REFERENCES

[1] Autoregulatory and gap gene response elements of the *even-skipped* promoter of *Drosophila*. *The EMBO Journal*, 8:1205–1212, 1989.

[2] 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 USA*, 79:1129–1133, 1982.

[3] T. Ahsendorf, F. Wong, R. Eils, and J. Gunawardena. A framework for modelling gene regulation which accommodates non-equilibrium mechanisms. *BMC Biology*, 12:102, 2014.

[4] Michael Akam. The molecular basis for metameric pattern in the *Drosophila* embryo. *Development*, 101:1–22, 1987.

[5] R. P. Alexander, G. Fang, J. Rozowsky, M. Snyder, and M. B. Gerstein. Annotating non-coding regions of the genome. *Nature Reviews Genetics*, 11:559–571, 2010.

[6] P. Andolfatto. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature*, 437:1149–1153, 2005.

[7] L. P. M. Andrioli, V. Vasisht, E. Theodosopoulou, A. Oberstein, and S. Small. Anterior repression of a *Drosophila* stripe enhancer requires three position-specific mechanisms. *Development*, 129:4931–4940, 2002.

[8] C. D. Arnold, D. Gerlach, C. Stelzer, L. M. Boryn, M. Rath, and A. Stark. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, pages 1074–1077, 2013.

[9] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C. F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. Diversity and complexity in dna recognition by transcription factors. *Science*, 324:1720–1723, 2009.

[10] J. Banerji, L. Olsen, and W. Shaffner. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, 33:729–740, 1983.

[11] J. Banerji, S. Rusconi, and W. Schaffner. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27:299–308, 1981.

[12] S. Barolo and M. Levine. Hairy mediates dominant repression in the *Drosophila* embryo. *The EMBO Journal*, 16:2883–2891, 1997.

[13] K. C. Barry, M. Abed, D. Kenyagin, T. R. Werwie, O. Boico, A. Orian, and S. M. Parkhurst. The *Drosophila* STUbL protein Degringolade limits HES functions during embryogenesis. *Development*, 138:1759–1769, 2011.

[14] C. R. Bartman, S. C. Hsu, C. S. S. Hsiung, A. Raj, and G. A. Blobel. Enhancer regulation of transcriptional bursting parameters revealed by forced chromatin looping. *Molecular Cell*, 62:237–247, 2016.

[15] G. Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology*, 193:723–50, 1987.

[16] T. Berleth, M. Burri, G. Thoma, D. Bopp, S. Richstein, G. Frigerio, M. Noll, and C. Nüsslein-Volhard. The role of localization of *bicoid* RNA in organizing the anterior pattern of the *Drosophila* embryo. *The EMBO Journal*, 7:1749–1756, 1988.

[17] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proceedings of the National Academy of Sciences USA*, 99:757–762, 2002.

[18] E. Bertolino, J. Reinitz, and Manu. The analysis of novel distal Cebpa enhancers and silencers using a transcriptional model reveals the complex regulatory logic of hematopoietic lineage specification. *Developmental Biology*, 413:128–144, 2016.

[19] M. D. Biggin. Animal transcription networks as highly connected, quantitative continua. *Developmental Cell*, 21:611–626, 2011.

[20] R. Binary and N. Perrimon. Stripe-specific regulation of pair-rule genes by hopscotch, a putative Jak family tyrosine kinase in *Drosophila*. *Genes and Development*, 8:300–312, 1994.

[21] J. R. Bosch, J. A. Benavides, and T. W. Cline. The TAGteam DNA motif controls the timing of *Drosophila* pre-blastoderm transcription. *Development*, 133:1967–1977, 2006.

[22] D. Bosisio, I. Marazzi, A. Agresti, N. Shimizu, M. E. Bianchi, and G. Natoli. A hyper-dynamic equilibrium between promoter-bound and nucleoplasmic dimers controls NF-kappaB-dependent gene activity. *EMBO*, 25:798–810, 2006.

[23] J. P. Bothma, H. Garcia, E. Esposito, G. Schlissel, T. Gregor, and M. Levine. Dynamic regulation of *eve* stripe 2 expression reveals transcriptional bursts in living *Drosophila* embryos. *Proceedings of the National Academy of Sciences USA*, 111:10598–10603, 2014.

[24] J. P. Bothma, H. Garcia, S. Ng, M. W. Perry, T. Gregor, and M. Levine. Enhancer additivity and non-additivity are determined by enhancer strength in the *Drosophila* embryo. *eLife*, 4:e07956, 2015.

[25] D. S. Burz and S. D. Hanes. Isolation of mutations that disrupt cooperative DNA binding of the *Drosophila* Bicoid protein. *Journal of Molecular Biology*, 305:21–230, 2001.

[26] David S. Burz, Rolando Rivera-Pomar, Herbert Jaeckle, and Steven D. Hanes. Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent gene activation in the *Drosophila* embryo. *The EMBO journal*, 17:5998–6009, 1998.

[27] E. Cannavò, P. Khoueiry, D. A. Garfield, P. Geeleher, T. Zichner, H. E. Gustafson, L. Ciglar, J. O. Korbel, and E. E. M. Furlong. Shadow enhancers are pervasive features of developmental regulatory networks. *Current Biology*, 26:38–51, 2016.

[28] H. Chen and M. Fujioka T. Gregor. Direct visualization of transcriptional activation by physical enhancer-promoter proximity. *bioRxiv*, 099523, 2017.

[29] H. Chen, X. Zhe, C. Mei, D. Yu, and S. Small. A system of repressor gradients spatially organizes the boundaries of bicoid-dependent target genes. *Cell*, 2:618–629, 2012.

[30] K. W. Chu, Y. Deng, and J. Reinitz. Parallel simulated annealing by mixing of states. *The Journal of Computational Physics*, 148:646–662, 1999.

[31] D. E. Clyde, M. S. Corado, X. Wu, A. Pare, D. Papatsenko, and S. Small. A self-organizing system of repressor gradients establishes segmental complexity in *Drosophila*. *Nature*, 426:849–853, 2003.

[32] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

[33] F. Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

[34] J. Crocker, N. Abe, L. Rinaldi, A. P. McGregor, N. Frankel, S. Wang, A. Alsawadi, P. Valenti, S. Plaza, F. Payre, R. S. Mann, and D. L. Stern. Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell*, 160:191–203, 2015.

[35] C. A. Davey, D. F. Sargent, K. Luger, A. W. Maeder, and T. J. Richmond. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 A resolution. *Journal of Molecular Biology*, 319:1097–1113, 2002.

[36] J. Dekker, M. A. Marti-Renom, and L. A. Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14:390–403, 2013.

[37] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16:1299–1309, 2006.

[38] W. Driever and C. Nüsslein-Volhard. A gradient of Bicoid protein in *Drosophila* embryos. *Cell*, 54:83–93, 1988.

[39] J. Dubnau and G. Struhl. RNA recognition and translational regulation by a homeodomain protein. *Nature*, 379:694–649, 1996.

[40] L. Dunipace, A. Ozdemir, and A. Stathopoulos. Complex interactionsbetween cis-regulatory modules in native conformation are critical for *Drosophila* snail expression. *Development*, 138:4075–4084, 2011.

[41] J Erceg, T.E. Sauders, C. Girardot, D.P. Devos, L. Hufnagel, and E.E.M. Furlong. Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. *PLoS Genetics*, 10:e1004060, 2014.

[42] Walid D. Fakhouri, Ahmet Ay, Rupindar Sayal, Jacqueline Dresch, Evan Dayringer, and David N. Arnosti. Deciphering a transcriptional regulatory code: modeling short-range repression in the *Drosophila embryo*. *Molecular Systems Biology*, 6:341, 2010.

[43] E. K. Farley, K. M. Olson, W. Zhang, A. J. Brandt, D. S. Rokhsar, and M. S. Levine. Suboptimization of developmental enhancers. *Science*, 350:325–328, 2015.

[44] V. E. Foe and B. M. Alberts. Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in *Drosophila* embryogenesis. *The Journal of Cell Science*, 61:31–70, 1983.

[45] C. C. Fowlkes, C. L. Luengo Hendricks, S. V. E. Keränen, G. H. Weber Oliver Rübel, M. Huang, S. Chatoor, A. H. DePace, L. Smirenko, C. Henriquez, A. Beaton, R. Weiszmann, S. Celnicker, B. Hamann, D. W. Knowles, M. D. Biggin, M. B. Eisen, and J. Malik. A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm. *Cell*, 133:364–374, 2008.

[46] N. Frankel, G. K. Davis, D. Vargas, S. Wang, F. Payre, and D. L. Stern. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, 466:490–493, 2010.

[47] M. Frasch, T. Hoey, C. Rushlow, H. J. Doyle, and M. Levine. Characterization and localization of the even-skipped protein of *Drosophila*. *The EMBO Journal*, 6:749–759, 1987.

[48] M. Frasch and M. Levine. Complementary patterns of *even-skipped* and *fushi-tarazu* expression involve their differential regulation by a common set of segmentation genes in *Drosophila*. *Genes and Development*, 1:981–995, 1987.

[49] D. Fu, C. Zhao, and J. Ma. Enhancer sequences influence the role of the amino-terminal domain of Bicoid in transcription. *Molecular and Cellular Biology*, 23:4439–4448, 2003.

[50] 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*, 126:2527–2538, 1999.

[51] T. Fukaya, B. Lim, and M. Levine. Enhancer control of transcriptional bursing. *Cell*, 166:358–368, 2016.

[52] J. P. Gergen and E. Wieschaus. Localized requirements for gene activity in segmentation of *Drosophila* embryos: analysis of *armadillo*, *fused*, *giant* and *unpaired* mutations in mosaic embryos. *Roux's Archives of Developmental Biology*, 195:49–62, 1986.

[53] D. Gibson. One-step enzymatic assembly of DNA molecules up to several hundred kilobases in size. *Protocol Exchange*, doi:10.1038/nprot.2009.77, 2009.

[54] S. F. Gilbert. *Developmental Biology*. Sinauer Associates, Sunderland, MA, seventh edition, 2003.

[55] W. Gilbert and B. Müller-Hill. Isolation of the Lac repressor. *PNAS*, 56:1891–1898, 1966.

[56] W. Gilbert and B. Müller-Hill. The Lac operator is DNA. *PNAS*, 58:2415–2421, 1967.

[57] S. D. Gillies, S. L. Morrison, V. T. Oi, and S. Tonegawa. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell*, 33:717–728, 1983.

[58] A. N. Gorban and O. Radulescu. Dynamical robustness of biological networks with hierarchical distribution of time scales. *IET Ststems Biology*, 1:238–246, 2007.

[59] T. Goto, P. MacDonald, and T. Maniatis. Early and late periodic patterns of *even-skipped* expression are controlled by distinct regulatory elements that respond to different spatial cues. *Cell*, 57:413–422, 1989.

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

[61] S. Gray, P. Szymanski, and M. Levine. Short-range repression permits multiple enhancers to function autonomously within a complex promoter. *Genes and Development*, 8:1829–1838, 1994.

[62] T. Gregor, E. F. Wieschaus, A. P. McGregor, W. Bialek, and D. W. Tank. Stability and nuclear dynamics of the Bicoid morphogen gradient. *Cell*, 130:141–152, 2007.

[63] A.C. Groth, M. Fish, R. Nusse, and M. P. Calos. Construction of transgenic *Drosophila* by using the site-specific integrase from phage phic31. *Genetics*, 166:1775–1782, 2004.

[64] P. Hammar, M. Walldén, D. Fange, F. Persson, Ö. Baltekin, G. Ullman, P. Leroy, and J. Elf. Direct measurement of transcription factor dissociation excludes a simple operator occupancy model for gene regulation. *Nature Genetics*, 46:405–408, 2014.

[65] K. Han, M. Levine, and J. L. Manley. Synergistic activation and repression of transcription by *Drosophila* homeobox proteins. *Cell*, 56:573–583, 1989.

[66] S. D. Hanes, G. Riddihough, D. Ish-Horowicz, and R. Brent. Specific DNA recognition and intersite spacing are critical for action of the bicoid morphogen. *Molecular and Cellular Biology*, 14:3364–3375, 1994.

[67] E. E. Hare, B. K. Peterson, and M. B. Eisen. A careful look at binding site reorganization in the *even-skipped* enhancers of *Drosophila* and sepsids. *PLoS Genetics*, 4(11):e1000268, 2008.

[68] E. E. Hare, B. K. Peterson, V. N. Iyer, R. Meier, and M. B. Eisen. Sepsid *even-skipped* enhancers are functionally conserved in *Drosopila* despite lack of sequence conservation. *PLoS Genetics*, 4:e1000106, 2008.

[69] B. Z. He, A. K. Holloway, S. J. Maerkl, and M. Kreitman. Does positive selection drive transcription factor binding site turnover? a test with *Drosophila cis*-regulatory modules. *PLoS Genetics*, 7:e1002053, 2011. PMCID: PMC3084208.

[70] X. He, M. A. H. Samee, C. Blatti, and S. Sinha. Thermodynamics-based models of transcriptional regulation by enhancers: The roles of synergistic activation, cooperative binding and short-range repression. *PLoS Computational Biology*, 6:e1000935, 2010.

[71] G. Z. Hertz and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15:563–577, 1999.

[72] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences USA*, 106:9362–9367, 2009.

[73] J. W. Hong, D. Hendrix, and M. Levine. Shadow enhancers as a source of evolutionary novelty. *Science*, 321:1314, 2008.

[74] X. S. Hou, M. B. Melnick, and N. Perrimon. *marelle* Acts Downstream of the *Drosophila* HOP/JAK Kinase and Encodes a Protein Similar to the Mammalian STATs. *Cell*, 84:411–419, 1996.

[75] M. Hülskamp, C. Schröder, C. Pfeifle, H. Jäckle, and D. Tautz. Posterior segmentation of the *Drosophila* embryo in the absence of a maternal posterior organizer gene. *Nature*, 338:629–632, 1989.

[76] P. W. Ingham. The molecular genetics of embryonic pattern formation in *Drosophila*. *Nature*, 335:25–34, 1988.

[77] G. C. F. Innocentini and J. E. M. Hornos. Modeling stochastic gene expression under repression. *Journal of Mathematical Biology*, 55:413–431, 2007.

[78] V. Irish, R. Lehmann, and M. Akam. The *Drosophila* posterior-group gene *nanos* functions by repressing *hunchback* activity. *Nature*, 338:646–648, 1989.

[79] H. Jäckle, D. Tautz, R. Schuh, E. Seifert, and R. Lehmann. Cross-regulatory interactions among the gap genes of *Drosophila*. *Nature*, 324:668–670, 1986.

[80] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *The Journal of Molecular Biology*, 3:318–356, 1961.

[81] H. Janssens, S. Hou, J. Jaeger, A. R. Kim, E. Myasnikova, D. Sharp, and J. Reinitz. Quantitative and predictive model of transcriptional control of the *Drosophila melanogaster even skipped* gene. *Nature Genetics*, 38:1159–1165, 2006.

[82] H. Janssens, D. Kosman, C. E. Vanario-Alonso, J. Jaeger, M. Samsonova, and J. Reinitz. A high-throughput method for quantifying gene expression data from early *Drosophila* embryos. *Development, Genes and Evolution*, 215:374–381, 2005.

[83] P. Jiang, M. Z. Ludwig, M. Kreitman, and J. Reinitz. Natural variation of the expression pattern of the segmentation gene *even-skipped* in *Drosophila melanogaster*. *Developmental Biology*, 123:106–113, 2015. doi:10.1016/j.ydbio.2015.06.019; PMCID:PMC4529771.

[84] A. D. Johnson, B. J. Meyer, and M. Ptashne. Interactions between DNA-bound repressors govern regulation by the λ phage repressor. *Proceedings of the National Academy of Sciences USA*, 76:5061–5065, 1979.

[85] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, 316:1497–1502, 2007.

[86] L. A. Johnson, Y. Zhao, K. Golden, and S. Barolo. Reverse-engineering a transcriptional enhancer: a case study in *Drosophila*. *Tissue Engineering Part A*, 14:1549–1559, 2008.

[87] A. Jolma1, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge1, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152:327–339, 2013.

[88] O. W. Jones and M. W. Nirenberg. Degeneracy in the amino acid code. *Biochimica et Biophysica Acta*, 119:400–406, 1966.

[89] G. Junion, M. Spivakov, C. Girardot, M. Braun, E. H. Gustafson, E. Birney E, and E. E. Furlong. A transcription factor collective defines cardiac cell fate and reflects lineage history. *Roux's Archives of Developmental Biology*, 148:473–486, 2012.

[90] T. Kaplan, X. Y. Li, P. J. Sabo, S. Thomas, J. A. Stamatoyannopoulos, M. D. Biggin, and M. B. Eisen. Quantitative models of the mechanisms that control genome-wide patterns of franscription factor binding during early *Drosophila* development. *PLoS Genetics*, 7:e1001290, 2011.

[91] T. S. Karpova, M. J. Kim, C. Spriet, K. Nalley, T. J. Stasevich, Z. Kherrouche, L. Helliot, and J. G. MnNally. Concurrent fast and slow cycling of a transcriptional activator at an endogenous promoter. *Science*, 319:466–469, 2008.

[92] M. Kazemian, C. Blatti, A. Richards, M. McCutchan, N. Wakabayashi-Ito, A. S. Hammonds, S. E. Celniker, S. Kumar, S. A. Wolfe, M. H. Brodsky, and S. Sinha. Quantitative analysis of the *Drosophila* segmentation regulatory network using pattern generating potentials. *PLoS Biology*, 8:e1000456, 2010.

[93] A. R. Kim, C. Martinez, J. Ionides, A. F. Ramos, M. Z. Ludwig, N. Ogawa, D. H. Sharp, and J. Reinitz. Rearrangements of 2.5 kilobases of noncoding DNA from the *Drosophila even-skipped* locus define predictive rules of genomic *cis*-regulatory logic. *PLoS Genetics*, 9:e1003243, 2013. PMCID: PMC3585115.

[94] M. Kitano. Biological robustness. *Nature Reviews Genetics*, 5:826–837, 2004.

[95] D. Kosman, S. Small, and J. Reinitz. Rapid preparation of a panel of polyclonal antibodies to *Drosophila* segmentation proteins. *Development, Genes and Evolution*, 208:290–294, 1998.

[96] K. Kozlov, E. Myasnikova, A. Pisarev, M. Samsonova, and J. Reinitz. A method for two-dimensional registration and construction of the two-dimensional atlas of gene expression patterns in situ. *In Silico Biology*, 2:125–141, 2002.

[97] K. N. Kozlov, E. Myasnikova, A. A. Samsonova, S. Surkova, J. Reinitz, and M. Samsonova. GCPReg package for registration of the segmentation gene expression data in *Drosophila*. *Fly*, 3:151–156, 2009. PMCID: PMC3171190.

[98] Meghana M. Kulkarni and David N. Arnosti. *cis*-Regulatory logic of short-range transcriptional repression in *Drosophila melanogaster*. *Molecular and Cellular Biology*, 25:3411–3420, 2005.

[99] J. Lam and J.-M. Delosme. An efficient simulated annealing schedule: Derivation. Technical Report 8816, Yale Electrical Engineering Department, New Haven, CT, September 1988.

[100] J. Lam and J.-M. Delosme. An efficient simulated annealing schedule: Implementation and evaluation. Technical Report 8817, Yale Electrical Engineering Department, New Haven, CT, September 1988.

[101] D. Lebrecht, M. Foehr, E. Smith, F. J. P. Lopes, C. E. Vanario-Alonso, John Reinitz, D. S. Burz, and S. D. Hanes. Bicoid cooperative DNA binding is critical for embryonic patterning in *Drosophila*. *Proceedings of the National Academy of Sciences USA*, 102:13176–13181, 2005.

[102] R. Lehmann. Phenotypic comparison between maternal and zygotic genes controlling the segmental pattern of the *Drosophila* embryo. *Development (Supplement)*, 104:17–27, 1988.

[103] L. A. Lettice, S. Heaney, J. H. Simon, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill, and E. de Graaff. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, 12:1725–1735, 2003.

[104] M. Levine. Transcriptional enhancers in animal development. *Current Biology*, 20:R754–R763, 2010.

[105] G. Li, ¡. Levitus, C. Bustamante, and J. Widom. Rapid spontaneous accessibility of nucleosomal DNA. *Nature Structural and Molecular Biology*, 12:46–53, 2004.

[106] X. Y. Li, S. Thomas, P. J. Sabo, M. B. Eisen, J. A. Stamatoyannopoulos, and M. D. Biggin. The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding. *Genome Biology*, 12:1–17, 2011.

[107] H. L. Liang, C. Y. Nien, H. Y. Liu, M. M. Metzstein, N. Kirov, and C. Rushlow. The zinc-finger protein Zelda is a key activator of the early zygotic genome in *Drosophila*. *Nature*, 456:400–404, 2008.

[108] H. K. Long, S. L. Prescott, and J. Wysocka. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell*, 167:1170–1187, 2016.

[109] M. Z. Ludwig, C. M. Bergman, N. H. Patel, and M. Kreitman. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature*, 403:564–567, 2000.

[110] M. Z. Ludwig and M. Kreitman. Evolutionary dynamics of the enhancer region of *even-skipped* in *Drosophila*. *Molecular Biology and Evolution*, 12:1002–1011, 1995.

[111] M. Z. Ludwig, Manu, R. Kittler, K. P. White, and M. Kreitman. Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. *PLoS Genetics*, 7:e1002364, 2011. doi:10.1371/journal.pgen.1002364 PMCID: PMC3213169.

[112] M. Z. Ludwig, A. Palsson, E. Alekseeva, C. M. Bergman, J. Nathan, and M. Kreitman. Functional evolution of a *cis*-regulatory module. *PLoS Biology*, 3(4):e93, 2005.

[113] M. Z. Ludwig, N. H. Patel, and M. Kreitman. Functional analysis of *eve* stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Development*, 125:949–958, 1998.

[114] Cris L. Luengo-Hendriks, Soile V.E. Keranen, Charless C. Fowlkes, Lisa Simirenko, Gunther H. Weber, Clara Henriquez, David Kaszuba, Bernd Hamann, Michael Eisen, Jitendra Malik, Damir Sudar, Mark D. Biggin, and David W. Knowles. 3D morphology and gene expression in the *Drosophila* blastoderm at cellular resolution I: Data acquisition pipeline. *Genome Biology*, 7:R123, 2006.

[115] P. M. Macdonald and G. Struhl. A molecular gradient in early *Drosophila* embryos and its role in specifying the body pattern. *Nature*, 324:537–545, 1986.

[116] T. F. C. Mackay, S. Richards, and E. A. Stone et al. The *Drosophila melanogaster* genetic reference panel. *Nature*, 482:173–178, 2012.

[117] S. J. Maerkl and S. R. Quake. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315:233–237, 2007.

[118] M. Mannervik and M. Levine. The Rpd3 histone deacetylase is required for segmentation of the *Drosophila* embryo. *Proceedings of the National Academy of Sciences USA*, 96:6797–6801, 1999.

[119] Manu, S. Surkova, A. V. Spirov, V. Gursky, H. Janssens, A. Kim, O. Radulescu, C. E. Vanario-Alonso, D. H. Sharp, M. Samsonova, and J. Reinitz. Canalization of gene expression and domain shifts in the *Drosophila* blastoderm by dynamical attractors. *PLoS Computational Biology*, 5:e1000303, 2009. doi:10.1371/journal.pcbi.1000303 PMCID: PMC2646127.

[120] Manu, S. Surkova, A. V. Spirov, V. Gursky, H. Janssens, A. Kim, O. Radulescu, C. E. Vanario-Alonso, D. H. Sharp, M. Samsonova, and J. Reinitz. Canalization of gene expression in the *Drosophila* blastoderm by gap gene cross regulation. *PLoS Biology*, 7:e1000049, 2009. doi:10.371/journal.pbio.1000049 PMCID: PMC2653557.

[121] B. B. Maricque, J. D. Dougherty, and B. A. Cohen. A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of *cis*- regulatory activity in neural cells. *Nucleic Acids Research*, 45:e16, 2017.

[122] Carlos Martinez, Ah-Ram Kim, Joshua S. Rest, Michael Ludwig, Martin Kreitman, Kevin White, and John Reinitz. Ancestral resurrection of the *Drosophila* S2E enhancer reveals accessible evolutionary paths through compensatory change. *Molecular Biology and Evolution*, 31:903–916, 2014. PMCID: PMC3969564.

[123] Carlos A. Martinez, Kenneth A. Barr, Ah-Ram Kim, and John Reinitz. A synthetic biology approach to the development of transcriptional regulatory models and custom enhancer design. *Methods*, 62:91–98, 2013. PMCID: PMC3924567.

[124] A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C. Chen, A. Chou, H. Ienasescuand J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 42:D142–D147, 2014.

[125] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, and R. Kaul and. J A. Stamatoyannopoulos. Systematic Localization of Common Disease Associated Variation in Regulatory DNA. *Science*, 337:1190–1195, 2012.

[126] Daniel J. McKay and Jason D. Lieb. A common set of DNA regulatory elements shapes *Drosophila* appendages. *Developmental Cell*, 27:306–318, 2013.

[127] L. A. Mirny. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences USA*, 121:22534–22539, 2010.

[128] M. Mlodzik, A. Fjose, and W. J. Gehring. Isolation of *caudal*, a *Drosophila* homeo box- containing gene with maternal expression, whose transcripts form a concentration gradient at pre-blastoderm stage. *The EMBO Journal*, 4:2961–2969, 1985.

[129] The modENCODE Project Consortium. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, 330:1787–97, 2010.

[130] J. Mohler, E. D. Eldon, and V. Pirrotta. A novel spatial transcription pattern associated with the segmentation gene, *giant*, of *Drosophila*. *The EMBO Journal*, 8:1539–1548, 1989.

[131] P. A. Nambu and J. R. Nambu. The *Drosophila fish-hook* gene encodes a HMG protein essential for segmentation and CNS development. *Development*, 122:3467–3475, 1996.

[132] N. Negre, C. D. Brown, P. K. Shah, P. Kheradpour, C. A. Morrison, J. G. Henikoff, X. Feng, K. Ahmad, S. Russel, R. A. H. White, L. Stein, S. Henikoff, M. Kellis, and K. P. White. A comprehensive map of insulator elements for the *Drosophila* genome. *PLoS Genetics*, 6:e1000814, 2012.

[133] Y. Nibu and M. Levine. CtBP-dependent activities of the short-range Giant repressor in the *Drosophila* embryo. *Proceedings of the National Academy of Sciences USA*, 98:6204–6208, 2001.

[134] Y. Nibu, H. Zhang, E. Bajor, S. Barolo, S. Small, and M. Levine. dCtBP mediates transcriptional repression by Knirps, Krüppel and Snail in the *Drosophila* embryo. *The EMBO Journal*, 17:7009–7020, 1998.

[135] Y. Nibu, H. Zhang, and M. Levine. Interaction of short-range repressors with *Drosophila* CtBP in the embryo. *Science*, 280:101–104, 1998.

[136] M. W. Nirenberg, O. W. Jones, P. Leder, B. F. C. Clark, W. S. Sly, and S. Pestka. On the coding of genetic information. *Cold Spring Harbor Symposia on Quantitative Biology*, 28:104–109, 1963.

[137] M. W. Nirenberg and J. H. Matthaei. The dependence of cell-free protein synthesis in *e. coli* upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences USA*, 47:1588–1602, 1961.

[138] K. R. Nitta, A. Jolma, Y. Yin, E. Morgunova, T. Kivioja, J. Akhtar, K. Hens, T. Toivonen, B. Deplancke, E. E. Furlong, and J. Taipale. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife*, 4:e04837, 2015.

[139] M. B. Noyes, X. Meng, A. Wakabayashi, S. Sinha, M. H. Brodsky, and S. A. Wolfe. A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic Acids Research*, pages 1–14, 2008.

[140] C. Nüsslein-Volhard and E. Wieschaus. Mutations affecting segment number and polarity in *Drosophila*. *Nature*, 287:795–801, 1980.

[141] A. Ochoa-Espinosa, G. Yucel, L. Kaplan, A. Pare, N. Pura, A. Oberstein, D. Papatsenko, and S. Small. The role of binding site cluster strength in Bicoid-dependent patterning in *Drosophila*. *Proceedings of the National Academy of Sciences USA*, 102:4960–4965, 2005.

[142] A. R. Oliphant, C. J. Brandl, and K. Struhl. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random- sequence oligonucleotides: analysis of yeast GCN4 proteins. *Molecular Cell Biology*, 9:2944–2949, 1989.

[143] N. Orgawa and M. D. Biggin. High-Throughput SELEX Determination of DNA Sequences Bound by Transcription Factors In Vitro. *Methods in Molecular Biology*, 786:51–63, 2012.

[144] S. Pagans, M. Ortiz-Lombardia, M. L. Espinas, J. Bernues, and F. Azorin. The *Drosophila* transcription factor *tramtrack* (ttk) interacts with *trithorax-like* (gaga) and represses gaga-mediated activation. *Nucleic Acids Res*, 30:4406–4413, 2002.

[145] D. Panne, T. Maniatis, and S. C. Harrison. An atomic model of the interferon-beta enhanceosome. *Cell*, 129:1111–1123, 2007.

[146] R. P. Patwardhan, J. B. Hiatt, D. M. Witten, M. J. Kim, R. P. Smith, D. May, C. Lee, J. M. Andrie, S. I. Lee, G. M. Cooper, N. Ahituv, L. A. Pennacchio, and J. Shendure. Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nature Biotechnology*, 30:265–270, 2012.

[147] P. C. Peng, M. A. H. Samee, and S. Sinha. Incorporating chromatin accessibility data into sequence- to-expression modeling. *Biophysics Journal*, 105:1257–1267, 2015.

[148] P. Perez-Pinera, D. D. Kocak, C. M. Vockley, A. F. Adler, A. M. Kabadi, L. R. Polstein, P. E. Thakore, K. A. Glass, D. G. Ousterout, K. W. Leong, F. Guilak, G. E. Crawford, T. E. Reddy, and C. A. Gersbach. RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nature Methods*, 10:973–976, 2013.

[149] M. Perry, A. N. Boettiger, J. P. Bothma, and M. Levine. Shadow enhancers foster robustness of *Drosophila* gastrulation. *Current Biology*, 20:1562–1567, 2010.

[150] M. Perry, A. N. Boettiger, and M. Levine. Multiple enhancers ensure precision of gap gene-expression patterns in the *Drosophila* embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 108:13570–13575, 2011.

[151] J. P. Petschek, N. Perrimon, and A. P. Mahowald. Region-specific defects in *l(1)giant* embryos of *Drosophila melanogaster*. *Developmental Biology*, 119:175–189, 1987.

[152] A. Pisarev, E. Poustelnikova, M. Samsonova, and J. Reinitz. FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic Acids Research*, 37:D560–D566, 2008. PMCID: PMC2686593.

[153] E. Poustelnikova, A. Pisarev, M. Blagov, M. Samsonova, and J. Reinitz. Flyex database. `http://urchin.spbcas.ru/flyex`, 2005.

[154] G. N. Prata, J. E. Hornos, and A. F. Ramos. Stochastic model for gene transcription on *Drosophila melanogaster* embryos. *Physical Reviews E*, 93:022403, 2016.

[155] F. Qiao, H. Song, C. A. Kim, M. R. Sawaya, J. B. Hunter, M. Gingery, I. Rebay, A. J. Courey, and J. U. Bowie. Derepression by depolymerization ; structural insights into the regulation of yan by mae. *Cell*, 118:163–173, 2004.

[156] N. Rajewsky, M. Vergassola, U. Gaul, and E. D. Siggia. Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics*, 3:30, 2002.

[157] A. F. Ramos, G. C. P. Innocentini, F. M. Forger, and J. E. M. Hornos. Symmetry in biology: from genetic code to stochastic gene regulation. *IET SYSTEMS BIOLOGY*, 4:311–329, 2010.

[158] J. Reinitz, S. Hou, and D. H. Sharp. Transcriptional control in *Drosophila*. *ComPlexUs*, 1:54–64, 2003.

[159] J. Reinitz and D. H. Sharp. Mechanism of *eve* stripe formation. *Mechanisms of Development*, 49:133–158, 1995.

[160] J. Reinitz and D. H. Sharp. Gene Circuits and Their Uses. In J. Collado, B. Magasanik, and T. Smith, editors, *Integrative Approaches to Molecular Biology*, chapter 13, pages 253–272. MIT Press, Cambridge, Massachusetts, USA, 1996.

[161] J. Reinitz and J. R. Vaisnys. Theoretical and experimental analysis of the phage lambda genetic switch implies missing levels of cooperativity. *The Journal of Theoretical Biology*, 145:295–318, 1990.

[162] S. R. Russell, N. Sanchez-Soriano, C. R. Wright, and M. Ashburner. The dichaete gene of *Drosophila melanogaster* encodes a sox-domain protein required for embryonic segmentation. *Development*, 122:3669–3676, 1996.

[163] C. Sackerson, M. Fujioka, and T. Goto. The *even-skipped* locus is contained in a 16-kb chromatin domain. *Developmental Biology*, 211:39–52, 1999.

[164] M. A. H. Samee and S. Sinha. Quantitative modeling of a gene's expression from its intergenic sequence. *PLoS Computational Biology*, 10:1–21, 2014.

[165] Md Abul Hassan Samee and Saurabh Sinha. Evaluating thermodynamic models of enhancer activity on cellular resolution gene expression data. *Methods*, 62:79–90, 2013.

[166] R. Satija and R. K. Bradley. The TAGteam motif facilitates binding of 21 sequence-specific transcription factors in the *Drosophila* embryo. *Genome Research*, 22:656–665, 2012.

[167] R. Sayal, J. M. Dresch, I. Pushel, B. R. Taylor, and D. Arnosti. Quantitative perturbation-based analysis of gene expression predicts enhancer activity in early *Drosophila* embryo. *eLife*, 5:e08445, 2016.

[168] C. Scholes, A. H. DePace, and A. Sánchez. Combinatorial gene regulation through kinetic control of the transcription cycle. *Cell Systems*, 4:97–108, 2017.

[169] K. N. Schulz, E. R. Bondra, A. Moshe, J. E. Villalta J. D. Lieb, T. Kaplan, D. J. McKay, and M. M. Harrison. Zelda is differentially required for chromatin accessibility, transcription factor binding, and gene expression in the early *Drosophila* embryo. *Genome Research*, 25:1715–1726, 2015.

[170] M. Sedighi and A. M. Sengupta. Epigenetic chromatin silencing: bistability and front propagation. *Physical Biology*, 4:246–255, 2007.

[171] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*, 451:535–540, 2008.

[172] Y. Shi, J. Sawada, G. Sui, B. el Affar, J. R. Whetstine, F. Lan, H. Ogawa, M. P. Luke, Y. Nakatani, and Y. Shi. Coordinated histone modifications mediated by a CtBP co-repressor complex. *Nature*, 422:735–738, 2003.

[173] M. L. Siegal and A. Bergman. Waddington's canalization revisited: Developmental stability and evolution. *Proceedings of the National Academy of Sciences USA*, 99:10528–10532, 2002.

[174] S. Small, D. N. Arnosti, and M. Levine. Spacing ensures autonomous expression of different stripe enhancers in the *even-skipped* promoter. *Development*, 119:767–772, 1993.

[175] S. Small, A. Blair, and M. Levine. Regulation of *even-skipped* stripe 2 in the *Drosophila* embryo. *The EMBO Journal*, 11:4047–4057, 1992.

[176] S. Small, A. Blair, and M. Levine. Regulation of two pair-rule stripes by a single enhancer in the *Drosophila* embryo. *Developmental Biology*, 175:314–324, 1996.

[177] S. Small, R. Kraut, T. Hoey, R. Warrior, and M. Levine. Transcriptional regulation of a pair-rule stripe in *Drosophila*. *Genes and Development*, 5:827–839, 1991.

[178] R. P. Smith, L. Taher, R. P. Patwardhan, M. J. Kim, F. Inoue, J. Shendure, I. Ovcharenko, and N. Ahituv. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics*, 45:1021–1028, 2013.

[179] M. V. Staller, B. J. Vincent, D. J. Meghan, T. Lydiard-Martin, Z. Wunderlich, J. Estrada, and A. H. DePace. Shadow enhancers enable hunchback bifunctionality in the *Drosophila* embryo. *Proceedings of the National Academy of Sciences USA*, 112:785–790, 2015.

[180] D. Stanojevic, S. Small, and M. Levine. Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo. *Science*, 254:1385–1387, 1991.

[181] N. Staudt, S. Fellert, H. R. Chung, H. Jäckle, and G. Vorbrüggen. Mutations of the *Drosophila* zinc finger- encoding gene *vielfältig* impair mitotic cell divisions and cause improper chromosome segregation. *Molecular Biology of the Cell*, 17:2356–2365, 2006.

[182] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *e. coli*. *Nucleic Acids Research*, 10:2997–3011, 1982.

[183] P. Struffi, M. Corado, L. Kaplan, D. Yu, C. Rushlow, and S. Small. Combinatorial activation and concentration-dependent repression of the *Drosophila even skipped* stripe 3+7 enhancer. *Development*, 138:4291–4299, 2011.

[184] S. Surkova, D. Kosman, K. Kozlov, Manu, E. Myasnikova, A. Samsonova, A. Spirov, C. E. Vanario-Alonso, M. Samsonova, and J. Reinitz. Characterization of the *Drosophila* segment determination morphome. *Developmental Biology*, 313(2):844–862, 2008. PMCID: PMC2254320.

[185] S. Surkova, E. Myasnikova, H. Janssens, K. N. Kozlov, A. Samsonova, J. Reinitz, and M. Samsonova. Pipeline for acquisition of quantitative data on segmentation gene expression from confocal images. *Fly*, 2:58–66, 2008. PMCID: PMC2803333.

[186] C. I. Swanson, D. B. Schwimmer, and S. Barolo. Rapid evolutionary rewiring of a structurally constrained eye enhancer. *Current Biology*, 21:1186–1196, 2011.

[187] C. L. Swanson, N. C. Evans, and S. Barolo. Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Developmental Cell*, 18:359–370, 2010.

[188] K. Tamura, S. Subramanian, and S. Kumar. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Molecular and Biological Evolution*, 21:36–44, 2004.

[189] D. Tautz. Regulation of the *Drosophila* segmentation gene *hunchback* by two maternal morphogenetic centres. *Nature*, 332:281–284, 1988.

[190] D. Tautz, R. Lehmann, H. Schnürch, R. Schuh, E. Seifert, A. Kienlin, K. Jones, and H. Jäckle. Finger protein of novel structure encoded by *hunchback*, a second member of the gap class of *Drosophila* segmentation genes. *Nature*, 327:383–389, 1987.

[191] V. B. Teif, F. Erdel, D. A. Beshnova, Y. Vainshtein, J. Mallm, and K. Rippe. Taking into account nucleosomes for predicting gene expression. *Methods*, 62:26–38, 2013.

[192] D. Thanos and T. Maniatis. Virus induction of human IFN beta gene expression requires assembly of an enhanceosome. *Cell*, 83:1091–1100, 1995.

[193] S. Thomas, X. Y. Li, P. J. Sabo, R. Sandstrom, R. F. Thurman, T. K. Canfield, E. Giste, W. Fisher, S. E. Celniker, M. D. Biggin, and J. A. Stamatoyannopoulos. Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. *Genome Biology*, 12:R43, 2011.

[194] P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S. E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. E. Celniker, and G. M. Rubin. Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biology*, 3(12):RESEARCH0088, 2002.

[195] A Tsurumi, F. Xia, J. Li, K. Larson, R. LaFrance, and W. X. Li. STAT is an essential activator of the zygotic genome in the early *Drosophila* embryo. *PLoS Genetics*, 72:e1002086, 2011.

[196] T. van der Heijdena, J. J. F. A. van Vugta, C. Logieb, and J. van Noorta. Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. *Proceedings of the National Academy of Sciences USA*, 109:E2514–E2522, 2012.

[197] D. Vernimmen and W. A. Bickmore. The hierarchy of transcriptional activation: from enhancer to promoter. *Trends in Genetics*, 31:696–708, 2015.

[198] D. Villar, C. Berthelot, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli, T. J. Park, R. Deaville, J. T. Erichsen, A. J. Jasinska, J. M. A. Turner, M. F. Bertelsen, E. P. Murchison, P. Flicek, and D. T. Odom. Enhancer evolution across 20 mammalian species. *Cell*, 160:554–566, 2015.

[199] B. J. Vincent, J. Estrada, and A. H. Depace. The appeasement of Doug a synthetic approach to enhancer biology. *Integrative Biology*, 8:475–484, 2016.

[200] C. H. Waddington. *The Strategy of Genes*. George Allen & Unwin, London, 1957.

[201] L. D. Ward and M. Kellis. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*, 337:1675–1678, 2012.

[202] E. Wieschaus, C. Nüsslein-Volhard, and H. Kluding. *Krüppel*, a gene whose activity is required early in the zygotic genome for normal embryonic segmentation. *Developmental Biology*, 104:172–186, 1984.

[203] Z. Wunderlich, M. D. J. Bragdon, B. J. Vincent, J. A. White, J. Estrada, and A. H. DePace. *Krüppel* expression levels are maintained through compensatory evolution of shadow enhancers. *Cell Reports*, 12:1740–1747, 2015.

[204] R. Yan, S. Small, C. Desplan, C. R. Dearolf, and J. E. Darnell Jr. Identification of a *stat* gene that functions in *Drosophila* development. *Cell*, 84:421–430, 1996.

[205] D. Yu and S. Small. Precise registration of gene expression boundaries by a repressive morphogen in *Drosophila*. *Current Biology*, 18:868–876, 2008.

[206] Y. Zhao, D. Granas, and G. D. Stormo. Inferring binding energies from selected binding sites. *PLoS Computational Biology*, 5:e1000590, 2009.

[207] Yue Zhao and Gary D. Stormo. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature Biotechnology*, 29:480–483, 2011.

[208] L. J. Zhu, R. G. Christensen, M. Kazemian, C. J. Hull, M. S. Enuameh, M. D. Basciotta, J. A. Brasefield, C. Zhu, Y. Asriyan, D. S. Lapointe, S. Sinha, S. A. Wolfe, and M. H. Brodsky. FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Research*, 39:D111–D117, 2011.

CHAPTER 7

APPENDIX

7.1 Appendix Figures

7.1.1 Appendix Figures for Chapter 3

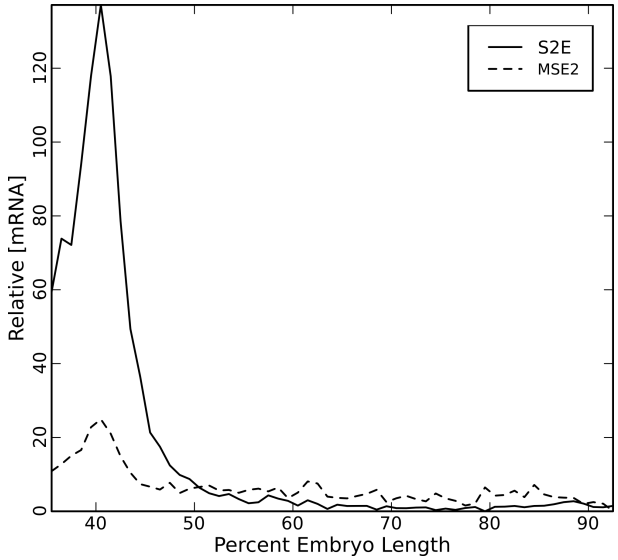


Figure 7.1. Rate driven by stripe 2 enhancers MSE2 and S2E. The 480 bp MSE2 fragment and the 698bp S2E (-dm3 coordinates 2R:5865217-5865913) were placed upstream of lacZ and cloned into the AttP2 site in *Drosophila*. Mean fluorescent in-situ hybridization (FISH) intensity at nuclear cycle 14 timepoint 6 is reported with S2E in solid lines and MSE2 in dashed lines. 15 embryos containing S2E were imaged, giving between 47 and 63 nuclei per AP position. 8 embryos containing MSE2 were imaged, giving between 26 and 37 nuclei per AP position. Peak expression of S2E is 5.5 times greater than that of MSE2, despite only containing 218 additional bases.

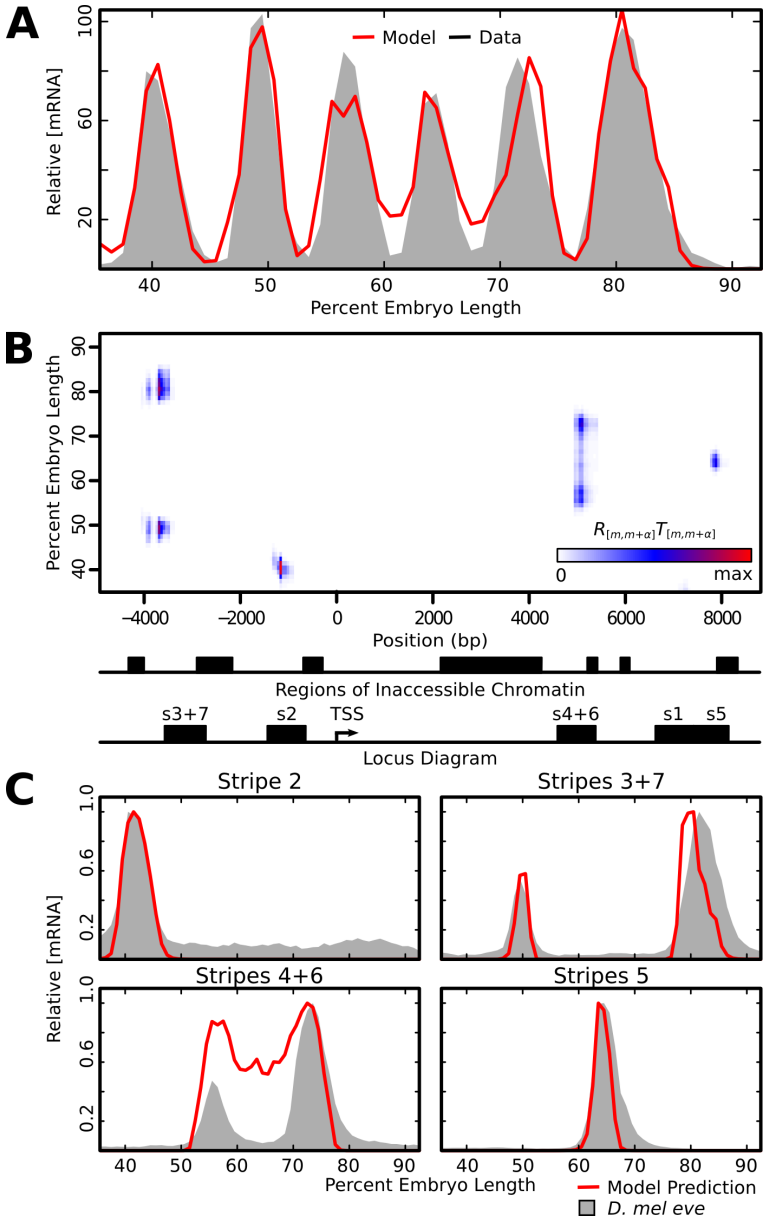


Figure 7.2. Best model fit using 500bp window. (A) the model output (red line) and data (gray shading) for the best fit to data. (B) Heatmap of the quantity quantity $R_{[m,m+\alpha]}T_{[m,m+\alpha]}$ at each nucleotide and embryo position, representing the amount each 1kb sequence, centered at that nucleotide, contributes towards total expression. The locations of known enhancers are indicated on the x-axis. (C) We tested the relative output of the known *eve* enhancers *in silico* using the retrained model (red lines). The relative mRNA driven by individual enhancers (gray shading), is included for visual orientation within the embryo and levels are not commensurate with predicted enhancer output.

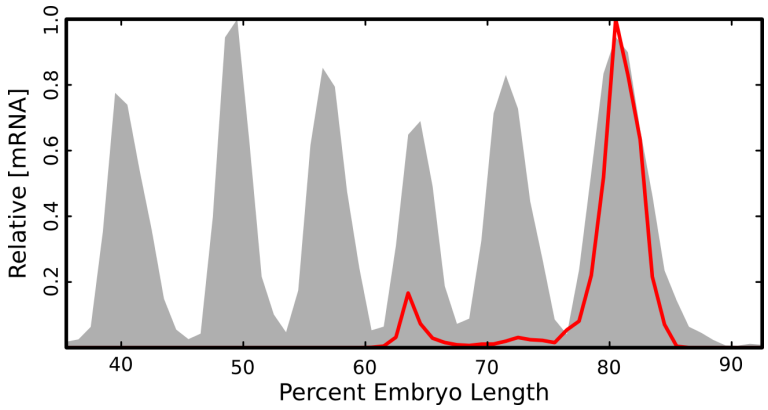


Figure 7.3. Prediction of e3130 element in model with accessibility. The model was trained as described in Fig 3.2, except binding sites were only called within regions of accessible chromatin. We predicted the activity of the 3130 element *in silico* to test its activity outside of its native chromatin context. The relative model output (red line) is plotted with *eve* mRNA. The relative mRNA driven by the locus (gray shading) is included for visual orientation within the embryo and levels are not commensurate with predicted enhancer output.

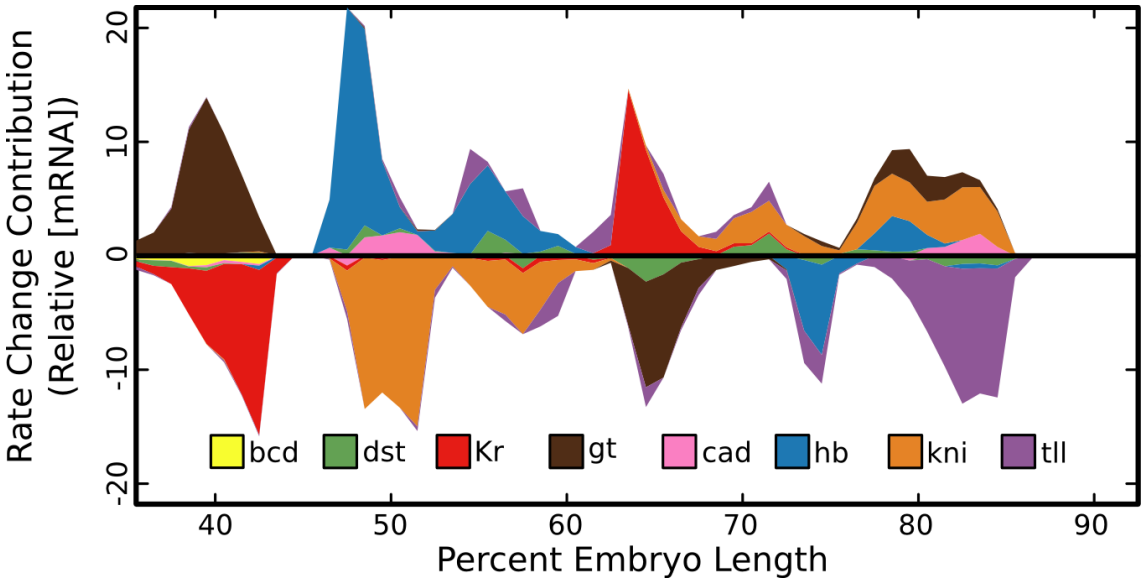


Figure 7.4. Mechanisms of repression in locus model with accessibility. The model was trained as described in Fig 3.2, except binding sites were only called within regions of accessible chromatin. Cumulative line graph showing the change in [mRNA] caused by a change in concentration of each TF (y-axis) at each embryo position (x-axis).

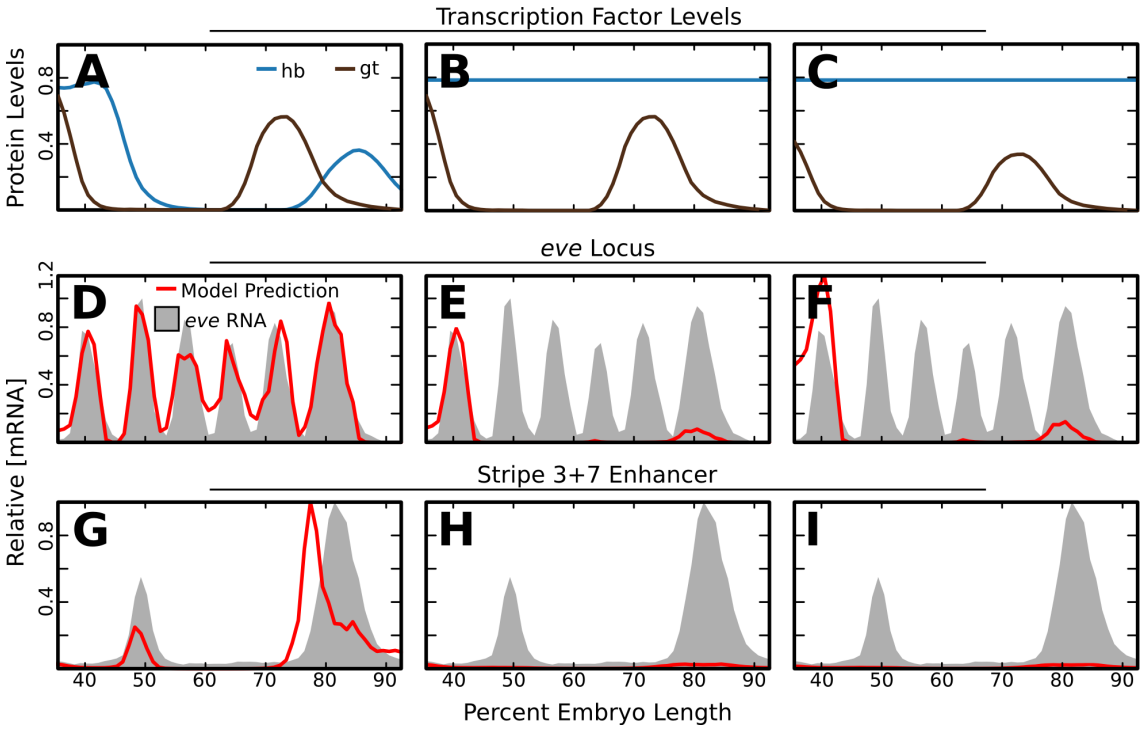


Figure 7.5. Predicted effects of ectopic Hb in model with accessibility. The model was trained as described in Fig 3.2, except binding sites were only called within regions of accessible chromatin. (A) The measured relative levels of Hb and Gt (y-axis) from 35.5% to 92.5% embryo length (x-axis). (B) Simulated relative levels of Hb and Gt. Hb is set to a spatially uniform value and Gt is unchanged from A. (C) Simulated relative levels of Hb and Gt. Hb is set to a spatially uniform value and Gt is reduced by 40%. (D-F) Predicted relative [mRNA] levels (red lines) driven by the *eve* locus under the TF levels indicated in A-C. Model output is standardized to the maximum rate driven by the locus in the wildtype *trans* environment. Data for relative [mRNA] of *eve* (gray shading) is included for visual orientation within the embryo and levels are not commensurate with predicted locus output. (G-H) Predicted relative [mRNA] levels (red lines) driven by the *eve* Stripe 3+7 enhancer under the TF levels indicated in A-C. Model output is standardized to the maximum rate driven by the enhancer in the wildtype *trans* environment. Data for relative [mRNA] driven by the stripe 3+7 enhancer (gray shading) is included for visual orientation within the embryo and levels are not commensurate with predicted enhancer output.

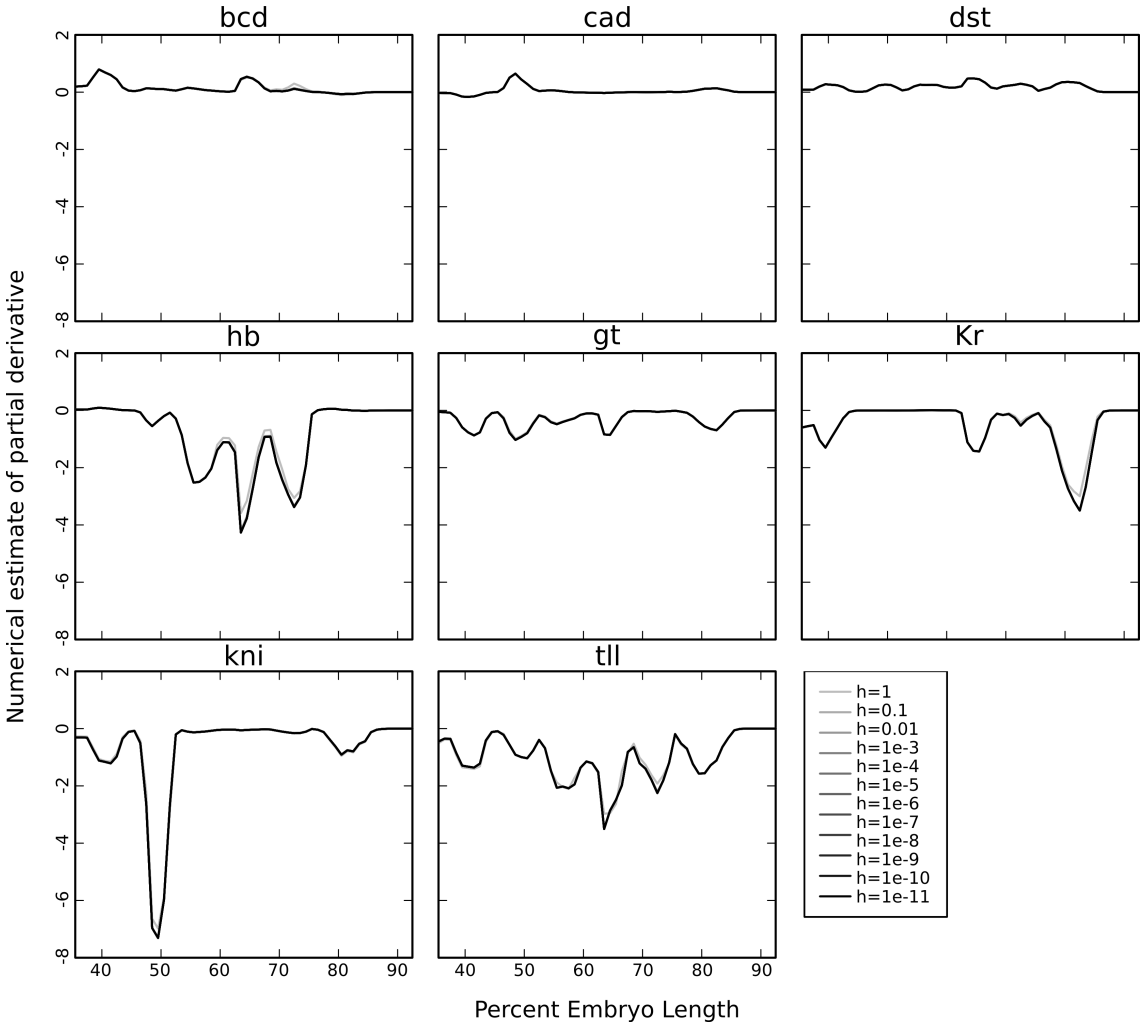


Figure 7.6. Numerical partial derivative estimates. The partial derivative $\frac{\partial R}{[\mathrm{TF}]}$ was estimated for each modeled TF using the symmetric difference quotient $\frac{f(x+h)-f(x-h)}{2h}$, at each position in the embryo, where h is the change in fluorescence of the TF in question over adjacent nuclei. Estimates are robust over values of h from 10^{-1} through 10^{-11}.

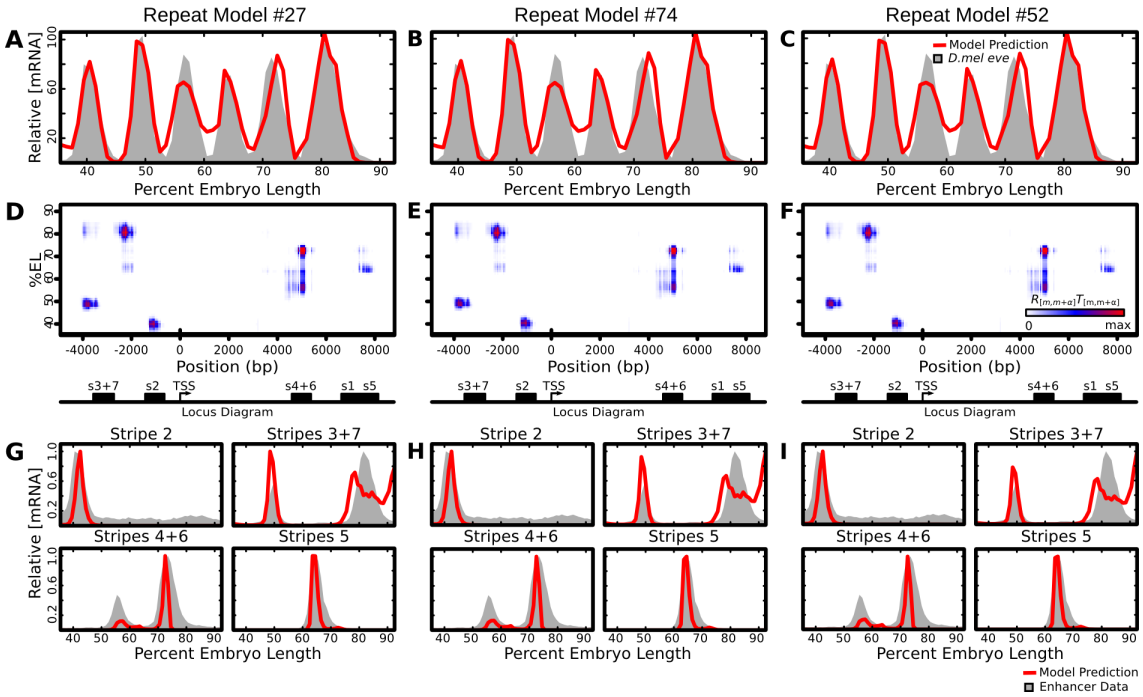


Figure 7.7. Best three model fits after repeating the optimization procedure. We repeated the optimization procedure an additional 80 times. The best three model fits have similar predictions to the model used to generate figures in the main text. We report predictions for the three parameter sets with the lowest score. (A-C) the model output (red line) and data (gray shading) for the top three parameter sets respectively. (D-F) Heatmap of the quantity quantity $R_{[m,m+\alpha]}T_{[m,m+\alpha]}$ at each nucleotide and embryo position, representing the amount each 1kb sequence, centered at that nucleotide, contributes towards total expression. The locations of known enhancers are indicated on the x-axis. (G-I) We tested the relative output of the known *eve* enhancers *in silico* using the retrained model (red lines). The relative mRNA driven by individual enhancers (gray shading), is included for visual orientation within the embryo and levels are not commensurate with predicted enhancer output.

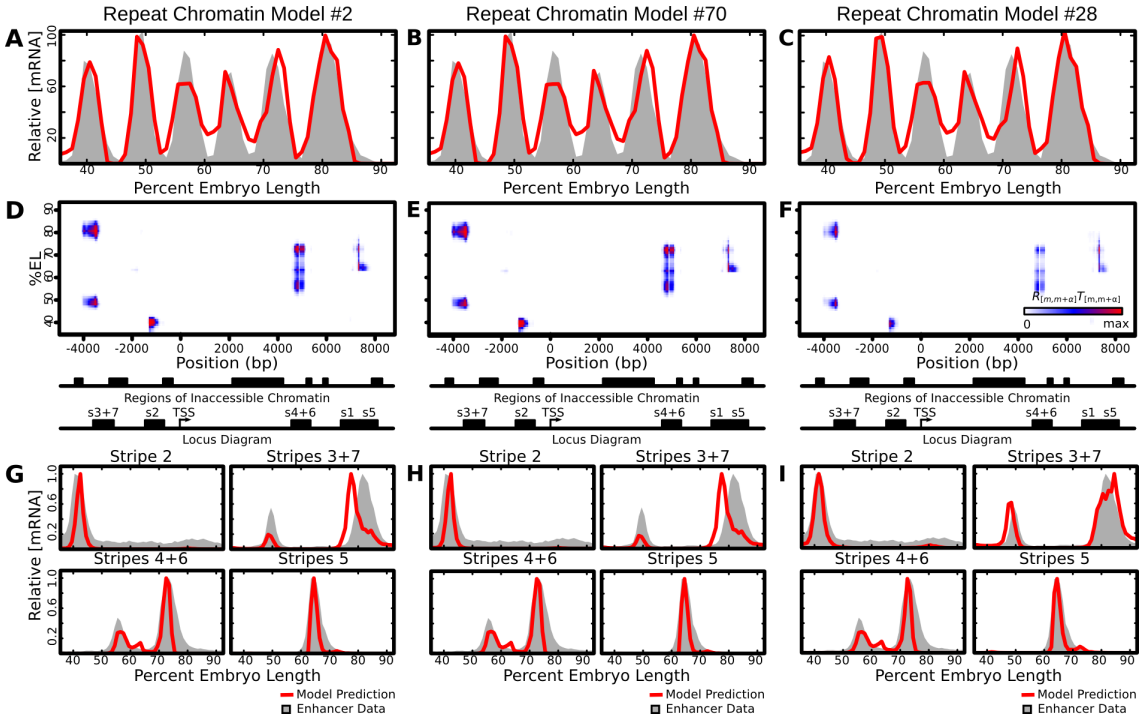


Figure 7.8. Best three model fits incorporating chromatin after repeating the optimization procedure. We repeated the optimization procedure an additional 80 times for fits incorporating chromatin data. The best three model fits have similar predictions to the model used to generate figures in the main text. We report predictions for the three parameter sets with the lowest score. (A-C) the model output (red line) and data (gray shading) for the top three parameter sets respectively. (D-F) Heatmap of the quantity quantity $R_{[m,m+\alpha]}T_{[m,m+\alpha]}$ at each nucleotide and embryo position, representing the amount each 1kb sequence, centered at that nucleotide, contributes towards total expression. The identified regions of inaccessible chromatin and locations of known enhancers are indicated on the x-axis. (G-I) We tested the relative output of the known *eve* enhancers *in silico* using the retrained model (red lines). The relative mRNA driven by individual enhancers (gray shading), is included for visual orientation within the embryo and levels are not commensurate with predicted enhancer output.

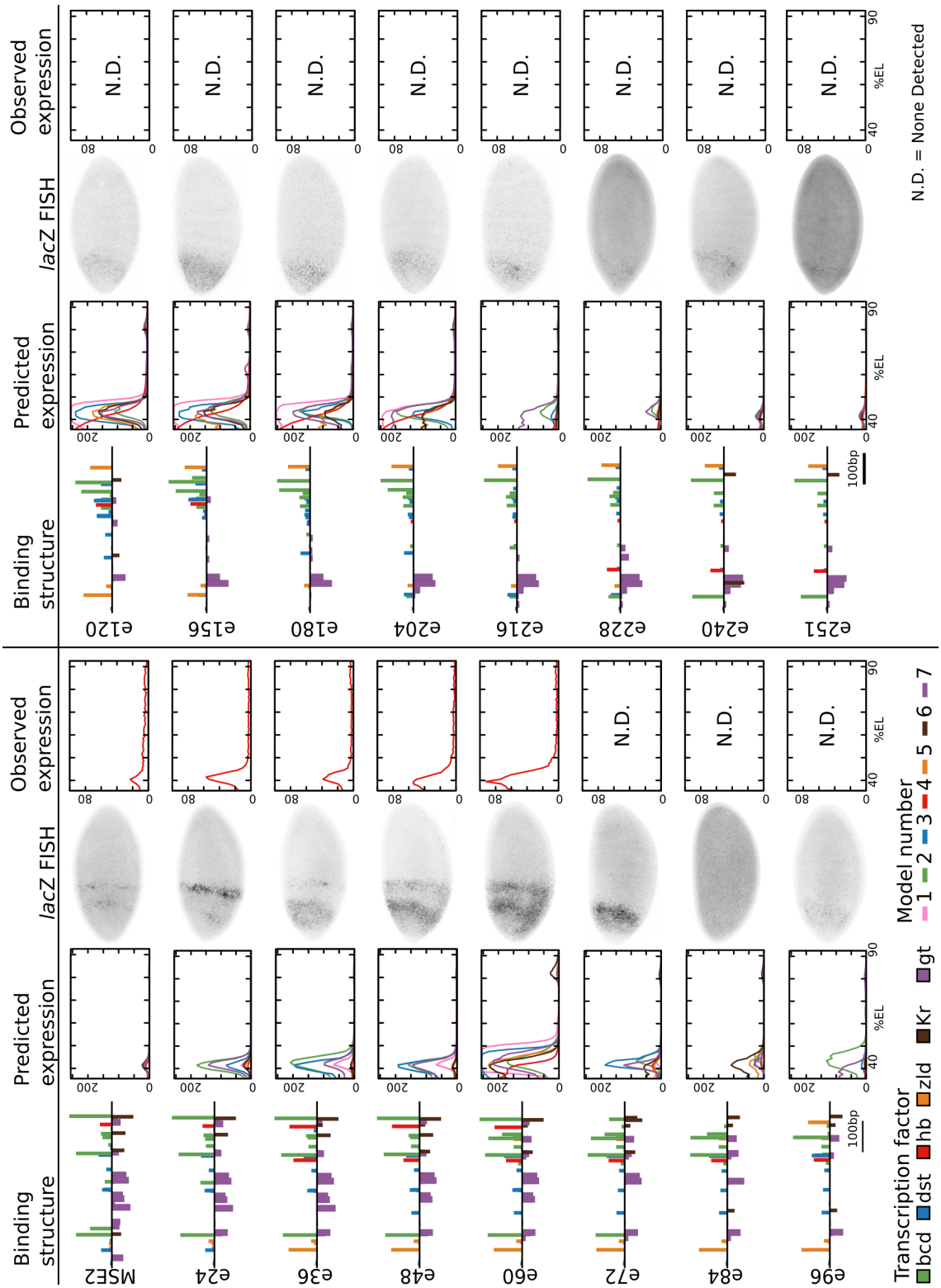


Figure 7.9. Continued on next page...

Figure 7.9. Expression and prediction on path to e251. Continued from previous page. For all tested enhancers the following are shown: (Column 1) Binding structure of tested enhancers. Activators are plotted on the positive y-axis and repressors on the negative. Bar height is proportional to the PWM score of binding. (Column 2) The predicted output in each of 7 models used in the design of the sequences. (Column 3) The *lac*Z mRNA expression of a representative embryo. (Column 4) The quantitative level of mRNA expression for each line along a 10% DV stripe from 35.5% to 92.5% embryo length.

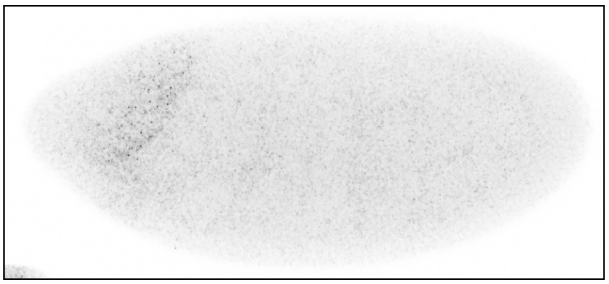


Figure 7.10. mRNA FISH of a reporter containing no enhancer. We generated a vector, as previously described, that contains no inserted putative enhancer and inserted the sequence into the Attp2 landing site. The resulting construct drives weak expression of an ectopic stripe anterior to *eve* stripe 1, confirming that this expression pattern is driven, at least in part, by sequences present on the vector.

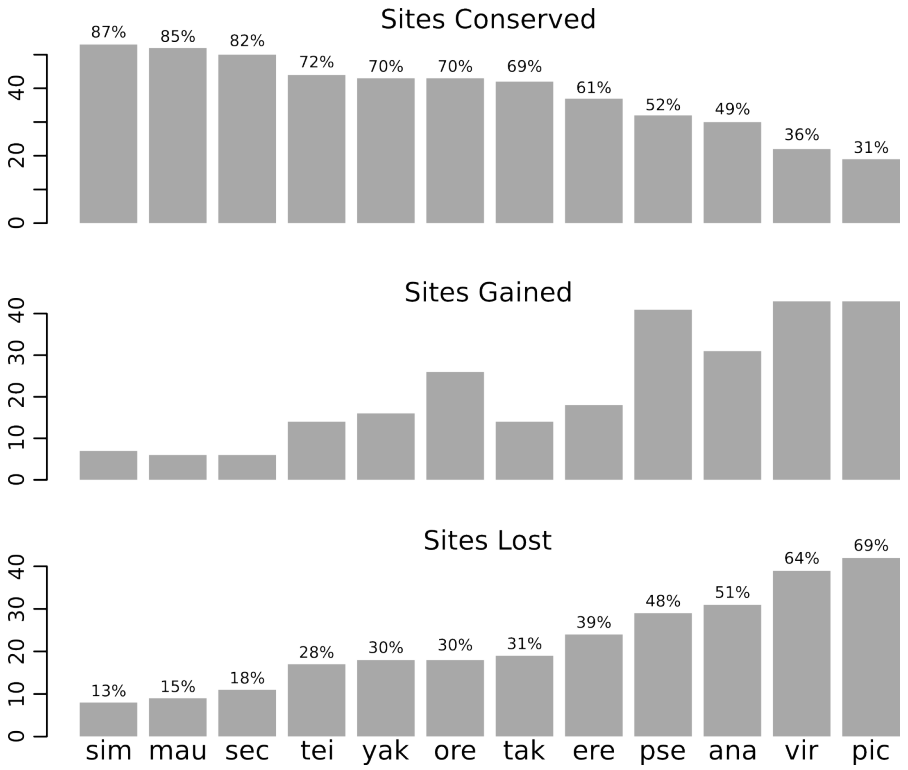


Figure 7.11. Conservation of motifs in S2Es. (Top) The number of binding motifs at PWM score greater than 0 for the factors Bcd, hb, Kr, and gt that are conserved with S2E are shown for each putative S2E. The percent conservation is given above each bar. (Middle) The number of binding motifs (score >0) for the same 4 factors that are gained are shown for each putative S2E. (Bottom) The number of motifs (score >0) for the 4 factors that are lost are shown for each factor. The percent of sites lost is given above each bar. The sequences of each are reported in Kim *et al.*[93].

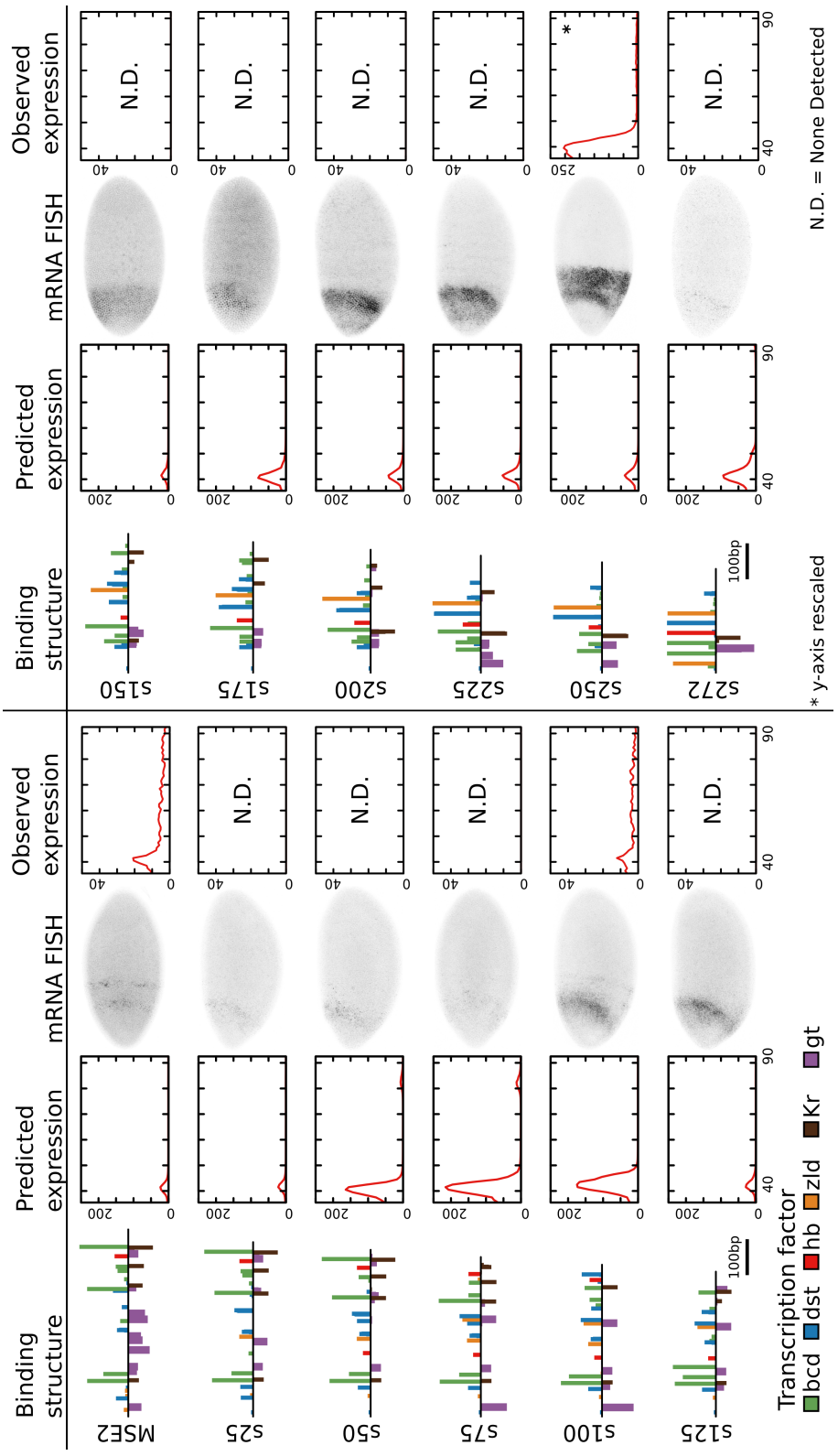


Figure 7.12. Continued on next page...

Figure 7.12. Expression and prediction along a path to s272. Continued from previous page. For all tested enhancers the following are shown: (Column 1) Binding structure of tested enhancers. Activators are plotted on the positive y-axis and repressors on the negative. Bar height is proportional to the PWM score of binding. (Column 2) The predicted output in the model used in the design of the sequences. (Column 3) The *lac*Z mRNA expression of a representative embryo. (Column 4) The quantitative level of mRNA expression for each line along a 10% DV stripe from 35.5% to 92.5% embryo length.

		Model						
Parameter	TF	1	2	3	4	5	6	7
A	Bicoid	0.0240	0.0378	0.0761	0.0021	0.0898	0.0630	0.4899
λ	Bicoid	1.9345	4.9920	2.1623	0.5918	0.5900	0.5900	3.4410
E^A	Bicoid	2.6774	0.0629	0.0002	16.695	0.0001	0.0001	0.0001
E^C	Bicoid	0.9999	0.2177	0.3503	0.6833	0.2358	0.1949	0.1398
D^C	Bicoid	150.66	158.19	150.63	182.60	186.89	183.88	184.81
ω	Bicoid	499.46	189.21	127.11	34.655	43.990	499.67	9.9926
Threshold	Bicoid	1.7100	1.7100	1.7100	-1.200	-1.000	-1.000	1.7100
PWM	Bicoid	selex	selex	selex	mitomi	mitomi	mitomi	selex
A	Caudal	0.0715	0.0524	0.0264	0.0184	0.0034	0.0084	0.0143
λ	Caudal	4.9982	4.9769	3.1853	9.9910	4.9999	4.9996	4.9969
E^A	Caudal	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002
E^C	Caudal	0.9999	0.3372	0.8877	0.9999	0.9999	0.9999	0.5127
D^C	Caudal	16.932	70.875	22.310	11.549	69.293	36.338	69.615
Threshold	Caudal	2.5013	3.0610	2.0651	2.9920	2.1084	2.1637	2.7288
A	Stat92E	3.9906	1.0995	3.9991	3.9990	3.9997	3.9990	3.9995
λ	Stat92E	0.8949	2.5858	0.6967	0.6936	0.7439	0.7352	1.3093
E^A	Stat92E	0.0002	0.0002	19.995	19.997	19.999	19.999	0.0004
Threshold	Stat92E	4.0387	2.8364	3.6314	2.9772	3.5897	4.0988	3.8904
A	Dicheate	3.9539	3.8938	0.1800	0.0512	3.9914	0.2712	1.1636
λ	Dicheate	1.3104	1.9834	4.5480	2.4845	0.9298	4.9927	2.5160
E^A	Dicheate	0.0002	0.0001	0.4538	0.0008	0.2821	0.0002	0.0001
Threshold	Dicheate	4.0033	4.7946	2.9628	3.8131	4.9091	4.3845	2.4163
A	Hunchback	0.0898	2.7242	0.0415	0.0465	0.0597	0.0544	1.1728

Table 7.1. Continued on next page...

λ	Hunchback	4.7353	1.8368	4.9999	0.5918	4.5542	4.9995	1.8391
E^Q	Hunchback	0.9693	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
E^A	Hunchback	19.547	20.417	29.999	18.684	24.630	27.349	29.979
E^D	Hunchback	0.4418	0.3239	0.5818	0.9769	0.0955	0.0002	0.4905
Threshold	Hunchback	0.6300	0.6300	0.6300	-3.250	0.6300	0.6300	0.6300
PWM	Hunchback	selex	selex	selex	mitomi	selex	selex	selex
A	Zelda	NA	NA	NA	0.0662	NA	NA	0.0022
λ	Zelda	NA	NA	NA	9.9999	NA	NA	3.7559
E^A	Zelda	NA	NA	NA	0.5582	NA	NA	19.983
Threshold	Zelda	NA	NA	NA	1.4878	NA	NA	-0.754
A	Kruppel	3.9978	0.0387	2.9529	0.8651	0.0252	0.0590	0.0554
λ	Kruppel	0.8886	4.0403	0.9808	1.2661	4.9971	2.8633	4.9993
E^Q	Kruppel	0.9999	0.9028	0.9999	0.9999	0.9999	0.9999	0.9999
E^D	Kruppel	0.9997	0.9997	0.9997	0.2613	0.4148	0.7248	0.7588
Threshold	Kruppel	0.3353	2.1164	0.0727	1.3488	0.4456	0.5366	0.4097
A	Knirps	0.2089	2.2331	0.2723	0.0608	0.2672	0.3967	0.0948
λ	Knirps	4.9988	2.4875	1.5688	3.8794	1.9085	1.4414	3.1533
E^Q	Knirps	0.1688	0.0637	0.9992	0.9999	0.9994	0.9998	0.9997
E^D	Knirps	0.0005	0.1236	0.9995	0.0898	0.3789	0.0340	0.9588
Threshold	Knirps	4.3219	2.2344	4.8570	5.6116	4.1055	3.7391	2.2623
A	Giant	0.1254	2.5308	0.0399	0.1105	3.9966	0.0549	3.9954
λ	Giant	4.9983	1.7136	4.9992	2.3216	1.2730	4.9993	1.3821
E^Q	Giant	0.3861	0.7236	0.7396	0.7956	0.4925	0.4266	0.8581
E^D	Giant	0.9746	0.1714	0.9997	0.0002	0.9995	0.9999	0.1710
Threshold	Giant	0.5855	0.5912	0.5044	2.6119	0.7235	0.6465	0.6532
A	Tailless	0.4225	0.0236	1.9470	0.0135	1.0147	0.2827	0.0637
λ	Tailless	1.3041	4.9804	0.9629	9.7618	1.1325	1.4912	2.7105
E^Q	Tailless	0.9995	0.9995	0.9999	0.0001	0.9998	0.9988	0.9999
E^D	Tailless	0.0006	0.0003	0.0002	0.0002	0.0002	0.0002	0.0002
Threshold	Tailless	1.9685	1.9773	1.9722	2.5069	1.8130	1.9693	1.9747
θ	NA	6.1477	6.0856	5.9558	7.4780	5.7273	5.7319	6.2430

Table 7.1. Continued on next page...

Table 7.1. Model Parameters. Continued from previous page. The model parameters and PWMs that were used generate the sequences in this work are given. All parameters are as described in Kim *et al.*[93] and models 1,2, and 3 are discussed in that work where they are called Model 01, Model 06, and Model 07 respectively. Models 5, 6, and 7 were trained as described in Kim *et al.* [93]. Model 4 was trained as described in Martinez *et al.* [122]. PWMs are specified for factors where the PWM chosen was different among fits.

7.1.3 Appendix Figures for Chapter 5

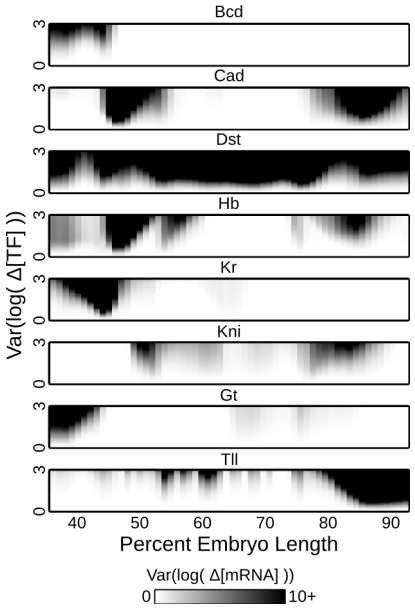


Figure 7.13. Robustness of *eve* expression to variation in TF concentration. A heatmap comparing the variance in fold-change input to fold-change output (Eqs 5.6-5.7) Var(log(Δ[mRNA])) at different positions within the embryo as well as different sizes of perturbation to TF concentration, indicated by Var(log(Δ[TF])). Darker shading represents increasing variation in mRNA levels.

7.2 Position Weight Matrices Used

>	Bcd_selex													
A	83	74	108	48	5	381	379	4	0	5	72	61	65	68
C	114	159	127	149	0	0	0	0	383	340	136	174	166	158
G	106	72	114	11	0	1	4	4	0	2	132	60	52	49
T	80	78	34	175	376	0	0	374	0	34	43	88	100	108

> cad_1hyb

A	9	12	2	3	12	38	0	3	21	1
C	9	6	3	0	0	0	0	0	0	8
G	4	3	3	0	1	0	0	7	14	10
T	11	16	29	34	24	0	38	27	0	1

> dst_selex

A	0	1	2	1	0	5	3	0	24	28	27	5
C	0	0	1	27	20	16	3	1	3	0	1	7
G	0	1	0	1	6	8	22	27	1	1	0	6
T	28	28	27	1	3	1	2	0	2	1	2	11

> dic_1hyb

A	1	0	0	20	0	0	2	0	1	3	6
C	7	25	17	0	0	0	0	0	2	10	1
G	7	0	0	0	0	3	27	0	4	6	1
T	13	3	1	9	28	26	0	28	22	9	21

> hb_selex

A	53	1	0	1	0	0	0	280	31	20
C	6	5	1	0	1	2	1	0	43	100
G	224	3	0	0	0	0	0	2	78	109
T	7	279	288	288	288	287	288	5	138	61

> Kr_selex

A	17	187	158	0	1	0	8	0	2	44
C	73	5	39	194	194	197	22	1	34	109
G	6	0	0	0	0	0	6	0	1	15
T	101	5	0	2	2	0	161	195	159	29

> kni_1hyb

A	19	25	16	5	0	21	0	17	1	0	25	5
C	0	0	0	9	4	0	0	0	3	26	0	12
G	1	0	0	6	0	5	26	7	18	0	0	7
T	3	0	10	5	21	0	0	1	4	0	0	2

>	gt_selex								
A	85	11	776	8	82	0	1020	1105	15
C	61	107	24	762	19	555	88	0	378
G	18	359	275	64	996	0	0	0	85
T	941	630	32	274	11	553	0	2	631

>	tll_matrix								
A	11	1	1	5	1	11	1	0	0
C	7	2	2	1	2	1	17	2	2
G	0	2	1	0	15	5	0	1	2
T	0	15	16	14	0	3	2	17	15

>	zld_1hyb							
A	54	7	756	0	1	4	756	76
C	0	749	1	0	0	0	5	75
G	478	0	1	731	750	3	0	566
T	229	5	3	30	10	754	0	44

>	Bcd_mitomi					
A	0.4446859	1.024804	0.000000	1.308060	1.212380	0.0000000
C	0.8241404	1.212753	0.901402	1.386241	0.819161	1.1815924
G	0.0000000	0.000000	1.211622	1.455025	1.407177	1.1908577
T	0.8472205	1.141426	1.261173	0.000000	0.000000	0.8257633

>	Hb_mitomi							
A	0.2557999	1.8157733	2.896855	2.954505	3.928484	3.637427	1.759458	0.0000000
C	0.6691087	0.8469678	2.268623	1.556663	4.230409	2.748917	3.135718	1.5068606
G	0.0000000	2.5883169	2.128065	1.824146	3.205319	3.789288	2.927774	0.8407924
T	0.2229569	0.0000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.4966308

>	cic_1hyb							
A	1	0	0	18	0	0	0	18
C	10	11	18	0	0	0	10	0
G	0	0	0	0	0	0	8	0
T	6	7	0	0	18	18	0	0

7.3 Sequences

The sequences generated in this work are given below.

>MSE2

CCGGTACTGCATAACAATGGAACCCGAACCGTAACTGGGACAGATCGAAAAGCTGGCCTGGTTTCTCGCTGTGTGTGCCGTGTT
AATCCGTTTGCCATCAGCGAGATTATTAGTCAATTGCAGTTGCAGCGTTTCGCTTTCGTCCTCGTTTCACTTTCGAGTTAGACTTTAT
TGCAGCATCTTGAACAATCGTCGCAGTTTGGTAACACGCTGTGCCATACTTTCATTTAGACGGAATCGAGGGACCCTGGACTATAAT
CGCACAACGAGACCGGGTTGCGAAGTCAGGGCATTCCGCCGATCTAGCCATCGCCATCTTCTGCGGGCGTTTGTTTGTTTGTTTGCT
GGGATTAGCCAAGGGCTTGACTTGGAATCCAATCCCGATCCCTAGCCCGATCCCAATCCCAATCCCAATCCCTTGTCCTTTTCATTAG
AAAGTCATAAAAACACATAATAATGATGTCGAAGGGATTAGGGG

>e24

CCGGTACTGAATACCAGTGGATTCCGAACCGTAACAGGGACAGATCGAAAAGCTGGCCTGGTATCTGCTGTGTGTGCCGTGTT
AATCCGTTGCCACAGCGAGATTGTTAGTCACTTGCAGTTACAGTATTTCGCTGTCGTCCTCGTTTCACTTTCGAGTTAGACTTTATT
GCAGCATCTAGAACAATCGTCGCAGTTCGGTAACACGCTGTGCCATACTATCATTTAGACGGAATCGAGGGACCCTGGACCATAATC
GCACAACGAGACCGGGTTGCGAAGTCAGGGCATTCCGCCGATCTAGCCATCGCCATCTTCTCCGGGTGTTTGTTTGTTTGTTTGCTG
GGATTAGCCAAGGGCTGACTTGGAATCCAATCCCGTCCCTAGCCCGATCCCAATCCCAATCCCAATCCCTTGTCCTTTTCATTAGAAA
GTCATAAAAACACATAATAATGATGTCGAAGGGATTAGGGG

>e36

CCGGTACTGAATACCAGTGGATTCCGAACCGTAACAGGTACAGATCGAAAAGCTGGCCTGGTATCTGCTGTGTGAGCCGTGTT
AATCCGTTGCCACAGCGAGATTGTTAGTCACTTGCAGTTACAGTATTTCGCTGTCGTCCTCGTTTCACTTTCGAGTTAGACTTTATT
GCAGCATCTATAACAATCGTCGCAGTTCGGTAACACGCTGTGCCATACTATCATTTAGACGGAATCGAGGGACCCTGGACCATAGTC
GCACAACGAGACCAGTTGCGAAGTCAGGGCATTCCGCCGATCTAGCCATCGCCATCTTCTCCGCGTGTTTTTTGTTTGTTTGCTGGG
ATTAGCCAAGGGCTGACTTGGAATCCAATCCCGTCCCTAGCCCGATCCCGATCCCAATCCCAATCCCTTGTCCTCTTCATTAGAAACT
CATAAAAACCCATAATAATGATGTCGAAGGGATTAGGGG

>e48

CCGGTACTGAATACCAGTGGATTCCGAACCGTAACAGGTACAGATCGAAAAGCTGGTCTGGTATCTGCTGTGTGAGCCCTGTTA
ATCCGTTGCCACAGCGAGATTGTAAGTCACTTGCAGTTACAGTATTTCGCTGTCGTCCTCGTTTCACTTTCGAGTCTAGACTTTAAT
GCAGCATCTATAACAATCGTCGCAGTTCGGTAACACGCTGTGCCATATCTATCATTTAGACGGAATCGAGGGACCCTGGACCATAGT
CGCACAACGAGACCAATTGCGAAGTCAGGGCATTCCGCCGATCTAGCCATCGCCATCTTCTCCGCGTGTTTTATGATTGTTTGCTGG
GATTAGCCAAGGGTGACTTGGAATCCAATCCCGTCCCTAGCCCGATCCCGATCCCAATCCCAATCCCTTGTCCTCTTCATTTGAAACT
CATAAAAACCCATAATAATGATGCGAAGGGATTAGGGG

>e60

CCGGTACTGAATACCAGTGGATTCCGGACCGTAACAGGTACAGATCGAAAAGCTGGTCTGGTATCTGCTGTGTGACCCCTGTTA
ATCCGTTGCCACAGCGAGAATGTAAGTCACTTGCAGTTACAGTATTTCGCTGTCGTCCTCGTTTCACTTTCGAGTCTAGACTTTAAT
GCAGCATCTATACCAATCGTCGCAGTTCGGTAACACGCTGTGCCATATCTATCATTTAGACGGAATCGAGGGCGCTGGACCATAGTC
GCACAACGAGACCAATTGCGAAGTCAGGGCATTCCGCCGATCTAGCCATCGCCATCTTATCCGCGTGTTTTATGATTGTTTGCGGGA
TTAGCCAAGGGTGACTTGGAATCCAATCCCGTCCATAGCCCGATCCCGACCTAATCCCAATCCCTAGTCCTCTTCATTTGAAACTCAT
AAAAACCCATAATAATGATGCGAAGGGATTAGGGG

>e72

CCGGTACTGAATACCAGTGGATTCCGGACCGTAACAGGTACAGATCGAAAAGCTGGTCTGGTATCTGCTATGTGACCCCTGATA
ATCCGTTGCCACAGCGAGAATGTAAGACACTTGCAGTTACAGTATTTCGCTGTCGTCCTCGTTTCACTTTCGAGTCTAGACTTTAAT
GCAGCATCTATACCAATCGTCGCAGTTCGTTAACACGCTGTCCATATCTATCATTTAGATGGAATCGAGGGCGCTGGACCATAGTCG
CACAACGAGACCAATTGCGAAGTCAGGGCATTCCGCCGATCTAGCCATCGCTCATCTTATCCGCGTGTTTTATGAATGTTTGCGGGA
TTAGCCAAGGGTGACTTGGAATCCAATCCCGTCCATAGCCCGATCCCGACCTAATCCCAATCCCTAGTCCTCTTCATTTGAAACTCAT
ATAAACCCATAATACTGATGGGAAGGGATGAGGGG

>e84

CCGGTACTGAATACCAGTGGATTCCGGACCGTAACAGGTACAGATCGAAAAGCTGGTCTGGTATCTGCTATGTGACCCCTGAG
GATGCGTTGCCACAGCGAGAATGTAAGACACTTGCAGTTACAGTATTTCGCTGTCGTCCTCGTTTCACTTTCGAGTCTAGCCTTTAA
TGCAGCATCTATACCAATCGTCGCAGTTCGTTAACACGCTGTCCATATCTATCATTTAGATGGAATCGAGGGCGCTGGACCATAGTC
GCACAACGAGACCAATTCGAAGTCAGGGCATTCCGCCGATCTAGCCATCAGCTCATCTTATCCGCGTGTTTTATGAAAGTTTGCGGG
ATTAGCCAATGGTGACTTGGAATCCAATCCCGTCCATAGTTCCCGATCCCGACCTAATCCCAACCCTAGTCCTCTTCATTTGAAACTC
ATATAAACCCATAATACTGATGGGAAGTGATGAGGGG

>e96

CCGGTACTGAATACCAGTGCATTCCGGACCGTAACAGGTACAGATCGAAAAGCTAGGTCTGGTATCTGCTATGCGACCCCTGAG
GATGCGTTGCCACAGCGAGAATGTAAGACACTTGCAGTTACAGTATTTCGCTGTCGTCCTCGTTTCACTTTCGAGTCTAGCCTTTAA

TGCAGCATCTATACCAATCTGTCGCAGTTCGTTAACACGCTGTCCATATCTATCATTTAGATGGAATCGAGGGCGCTGGACCATAGT
CGCACGACGAGACCAATTCGAAGTCAGGGCATTCCGCCGATCTATCCATCAGCTCATCTTATCCGCGTGTTTTATGAAAGTTTGCGG
GATTCGCCAATGGGACTTGGAATCCAATCCCGTCCATAGTTCCCGATCCCGACCTAATCCCAACCCTAGTCATCTTCATTTGAAACTC
ATATAAACCCATAATACCTGATGGGACACTGATGAGGGG
>e120
CGGTACTGAATTCCAGTGCATTCCGGACCGTAACAGGTACAGATCGAAAAGCTAGGTCTGGTATCTGCTATGCGACCCCTGAGG
ATGCGTTGCCACAGCGAGAATGTAAGACACTTGCAGTTACAGTATTTCGACTGTCGTCCTCGTTTCACTTTCGAGTCTAGCCTTTAA
TGCAGCATCTATACCAATCTGTCGCAGTTCGTTAACACGCTGTCCATATCTATCATTTAGATGGAATCGAGCGCGCTGGACCAGAGT
CGCACGACGAGACAATTCGAAGTTAGGGGATTCCGCCGAACATTTCCATCAGCTCACCTTATCCGCGTGTTTTATGAAAGTTTGCGG
GATTCGCCGATGGGACTTGGAATCTAATCCCGTCCATCAGTTCCCGTTCCCGACTAATCCCAACCCAGTCATCTTCATTTGATACTCA
TATATACCCATGATACCTGATGGGACACTCATGAGAGCG
>e156
CAGTACTGAATTCCGAGTGCATTCCGGACCGTTACAGGTACAGCTCGATAAGCTAAGGTCTGGTATCTGCTATGCAACCCCTGA
GGATGCGTTGCCACAGCGAGAATGTAAGACGACTTGCAGTTACAGTATTTCGACTGTCGTCCTCGTTGCACTTTCAAGTCTAGCCTT
GAATGCAGCATCTATACCAATCTGTCGACCCGTACCTTCACACGTTGTCCATATCAATATTCCGATGTGAATCGACCGCGCTGTACCA
GAGTCGCACGACAGAAATTCGAAGTTAAGGGAATCCGCCGAACATTTCCATCAGCTCACTTTATCCGCGTGTTTTATGAAAGTTTGC
GGGATTCGCCGATGGGATTGGAATCTAATCCCGTCCATCAGTTCCCGTTCCCGACTAATCCAACCCAGTAATCTTCATTTGATACTG
ATATATACCCGTGATACCTGATGGGACACTCATGAGACGCG
>e180
CAGTACTGAATTCCGAGTGCATTCTGGACCGTTACAAGTACAGCTCGATAAGCTAAGGTCTGGTATCTGCTATGCAACCCCTGA
GGATGCGTTGCCACAGCGAGAATGTAAGACGACTTTCATGTCTACAGTATTTCGAACGGTCGTCCTCGTTGCACTTTCAAGTCTAGC
CTCGAATGCATGATCTATACCAATCTGTCGACCCGTACCTTCAACGTTGTCCATATCAATATTCCGCTGTGGATGACCGCGCTGTACC
AGAGTCCCACGACAGAAAATCTAAGTTAAGGGAATCCGCCGACACATTTCCATCAGTCTCACTTTATCCGCGTGTTTCATGAAAGTA
TGCGGGATTCGCCGATGGGATTGGAATCTAATCCCGTCCATCAGTTCCCGTTCTCGACTAATCCAACACAGAATCTTCATTTGATAC
TGATATATACCCGTGATACCTGATGGGACACTCATGAGACGCG
>e204
CAGTGCTGAATTCCGAGTGCATTCTGGACCGATTACAAGTAAAGCTGATGATCTAAGGTATGGTATCTGCTATGCAACCCCTGA
GGATGCGTTGCACAGCGAGAATGTAAGACGAGCTTTCATGATCTACAGTATTTCGAACGGGCGTCCTCGTTGCACTTTCAAGTCTAG
CCTCGAATGCATGATCTATCCCAATCTGTCGACCCGTACCTTCAACGTAGTTCATATCAATATGCCGCTGTGGATGACCGCGCTGTA
CCAGAGGCCCACGACAGAAAATCTAAGTTAAGGGAATCCGCCGACACACTTCCATCAGTCTCACTTTATCCGCGTGTTTCATGAAAG
TATGCGGGATTCGCCGATGGGATTGGAATCTAATCCCGTACATCAGTTCCCATTCTCGACTAATCCAACACAGAATCTTAATGTGAT
ACCGATATATACCCGTGATACCTGATGGGACCCTCATGAGACGACG
>e216
CAGTGGTAATTCCGAGTGCATTCTGGACCGATTACAAGTAAAGCTGATGATCTAAGGTATGGTATCTGCTATGCAACCCCTGAG
GATGCGTTGCACAGCGAGAATGTAAGACGAGCTTTCATGATCTACAGTATTTCGAACGGGCGTCCTCAGTTGCACTTTCAAGTCCAG
CCACGAATGTCATGATCAATCCCAATCTGTCGACCCGTACCTTCAACGTAGTTCATATCAATATGCCGCTGTGGATGACCGCGCTGT
ACCAGAGGCCCACGACAGAAAATCTAAGTTAAGGGAATCCTCCGACACACTTCCATCAGTCTCACTTTATCCGCGTGTTTCATGAAA
GTATGCGGGATTCGCGATGGGATTGGAATCTAACTCCCGTACATCAGTTCCCATTCTCGACTAATCCAACACAGAATCTTAAGGTGA
TACCGATATATACCCGTGATACCTGTGGGACCCTCATGAGACGACG
>e228
CAGTGGTAATTCCGAGTGCATTCTGGGCCGATTACAAGTAAAGCTGATGATCTAAGGTATGGTATCTGCTATGCAACCCCTGAG
GATGCGTTGCACAGCGAGAATGTAAGACGAGCTTTCTATGATCTACAGTATTTCGAACGGGCGTTCCTCAGTTGCGCTTTCAAGTCC
AACCACGAATGTCATGAGCAATCCCAATCTGTCGACCCGTACCTTCAACGTAGTTCATATCAATATGCCGCTGTGGATAACCGCGCC
GTGCCAGAGGCCCACGACAGAAAATCTAAGTTAAAGGAATCCTCCGACACACTTCCATCAGTCTCACTTTATCCGCGTGTTTCATGA
AAGTATGCGGGATTCGCGATGGGATTGGAATCTAACTCCCGTACATCAGTTCACCATTCTCGACTAATCCAACACAGAATCTTAAGG
TGATACCGATATATGCCCGTGATACCTGTGGGACCCTCATGAGACGACG
>e240
CAGTGGTAATTCCGAGTGCATTCTGGGCAGATTAAAGTAAAGCTGATGATCTAAGGTATGGTATCTGTATGCAACCCCTGAGG
ATGCGTTGCACAGGAGAATGTAAGACGAGCTTTCTATGATCTACAGTAATTCGAACGGGAGTTCCTCAGTTAGCGCTGTCAAGTCCA
ACCACGAATGTCATGAGCAATCCCAATCTGTCGACCCGTACCTTCAACGTAGTTCATATCAATATGCCGCTGTGGATAACCGCGCCG
TGCCAGAGGCCCACGACAGAAAATCTAAGTTAAAGGAATCCTCCGACACACTTCCATCAGTCTCACTTTATCCGCGTGTTTCATGAA
AGTATGCGGCATTCGCGATGGGATTGGAACTAACACCCGTACATCAGTTCACCATTCTCGACTAATCCAACACAGAATCTTAAGGGT
GATACCGATATATGCCCGTGATACCTGTGGGACCCTCATGAGACGACG
>e251
CAGTGGTAATTCCGAGTGCATTCTTGGCAGATTAAAGTAAAGCTGATGATCTAAGGTATGGATCTGTATGCAACACCTAGGATG
CGTTGCACAGGAGAATGCAAAACGAGCTTTCTATGATCTACAGTAATTCGAACGGGAGATTCCTCAGTTAGCGGTGTCAAGTCCAAC

CACGAATGTCATGAGCAATCCCAATCTGTCGACCCGTACCTTCAACGTAGTTCATATCAATATGCCGCTGTGGATAACCGCGCCGTG
CCAGAGGCCCACGACAGAAAATCTAAGTTAAAGGAATCCTCCGACACACTTCCAACAGTACTCACTTTATCCGCGTGTTTCATGAAA
GTATGGGCATTCGCGATGGGATTGGAACTAACACCCGTACATCAGTTCACCATTCTCGACTAATCCAACACAGAATCTTAAGGGTGA
TACCGATATATGCCCGTGATACCTGTGGGACCCTCATGAGACGACG
>s25
CCGGTACGCATACAATGGAACCCGAACCGTAACTGGGACAGATCGAAAAGCTGGCCCGGTTTCCCGCTGTGTGTGCCGTGTTAA
TCCGTTTGCCATCAGCGAGATTATTATCAATGCAGTTGCAGCGTTTCGCCTTCGTCCTCGTTTCACTTCGAGTTAGATTTATTCAGCA
TCTTGAAAATCGTCGCAGTTTGGTAACACGCTGTGCCTACTTTCATTTAGACGGAATCGAGGGACCCTGGACTATAATGCACACGAG
ACCGGGTGGAAGTCAGGGCATTCCGCCGAATAGCCATCGCCATCTTCTGCGGCGTTTGTTGTTTGTTTGCAAGGGATTAGCCAAGGC
TTGACTTGGAATCCAATCCCGATCCCAGCCCGATCCCAATCCCAATCCCAATCCCTTGTCCTTTTCATTAGAAGTCATAAAAACACAT
AATAATGATGTCGAAGGGATTAGGGG
>s50
CCGTACGCATACAATGGAACCCGAACCGTAACTGGACAGATCGAAAAGCTGGCCCGGTTTCCCGCTGTGTGTGCCGTGTTAATC
CGTTTGCCATCAGCGAGATTATTATCAATGCATTGCAGCGTTTCGCCTTCGTCCTCGTTTCACTTCGAGTTAGATTTTTCACATCTTG
AAAATCGTCGCAGTTGGTAACACGCTGTGCCTACTTTCGTTTAGACGGAATCAGGGACCCTGGACTATAATGCACACGAGCCGGGTA
AGTCAGGGCATTCCGCCGAATAGCCATCGCCACTTCGCGCGTTTGTTGTTTGATTGCAAGGGATTAGCCAAGGCTTGATTGGAATCC
ATCCCGATCCCAGCCCGATCCGAATCCCAATCAATCCCTTGCCTTTCATTAGAAGTCATAAAAACACATTAATGATGTCGAAGGGATT
AGGG
>s75
CCGTACGCAACAATGGAACGCGAACCGTAACTGGACAGATCGAAAAGCTGGCCCGGTTTCCCGCTGTGTGTGCGTGTTAATCCG
TTTGCCATCAGCGAGATTATTATCAATGCATTGCAGCGTTTCCCTTCGTCCTCATTTACTCGAGTTAGATTTTTCACATCTTGAAAAT
CGTCGCAGTTGGTAACACGCTGTGCCTACTTTCGTTTAGACGGAATCAGACCCTGGACTATAATGCACACGAGCCGGGTAAGTCAGG
CCATTCCGCCGAATGCCTAGCCACTTCGCGCTTTGTTGTTTGATTGCAAGGGATTAGCCGGCTTGATGATCCATCCCGATCCCAGCC
CGATCGAATCCCAATCAATCCCTTGCCTTTCATTAGAAGCTAAAAACACATTAGATGTCGAAGGGATAGGG
>s100
CCTACGCAACATGGAACGCGAACCGTAACTGGACAGATCGAAAGCTGGCCCGGTTTCCCGCTGTTGTGCGTGTTAATCCGTTTG
CCATCAGCGGATTATTATCTGCATTGCAGCGTTCCCTTCGTCCTCATTTGCTCGAGTTAGATTTTTCACATCTGAAAACGTCGCAGTT
GGTAACACGCTGTGCCTACTTTCGTTTAGACGGAATCAGACCCTGGACTATAATGCACACGAGCCGGGTAGTCAGGCCATTCCGCCG
AATGCCTAGCCACTCGCGCTTGTTGTTTGACTGCAAGGAATTAGCCGGCTTGATGATCCATCCCGATCCCAGCCCGATCGAATCCAT
CAATCCCTTGCCTTCATTAGAGCTAAAAACACATAGTGTCGAGGATAGG
>s125
CCTACGCACATGGAACGCGACCGTAACTGGACAGATCGAAAGCTGGCCCGGTTTCCCGCTGTTGTGCGTGTTAATCCGTTTGCC
ATCAGCGGATTATTATCTGCATTGCAGCGTTCCCTTAATCCTCATTTGCTCGAGTTAGATTTTTCACATCTGAAACGTCGCAGTGGT
ACACTGTGCTACTTTGTTTAGACGGAATAGACCCTGGACTAGAATGCACACGAGCCGGGTAGTCAGGCCATTCCGCCGAATGCCTAG
CCACTCGCGCTTGTTGTTTGACCGCAAGGAATTAGCCGGCTTGATGATCATCCCATCCCAGCCGATCGAATCCATCAATCCCTTGCTT
CAGTAGAGCTAAACACATGTCGAGATAGG
>s150
CCTACGCACATGGAAGTGACCGTAACTGGACAGTCGAAGCTGGCCCTATTTCCCGCTGTTGTGCGTGTAATCCGTTTGACAGCG
GCTTATTATCTCATTGCAGCTTCCCTTAATCCTCATTTGCTCGAGTTAGATTTTTCACATCTGAAACGTCGAGTGGTACACTGTGCAC
TTTGTTTAGACGGAAAGACCCTGGACTAGACTGCACACGAGCCAGGTAGTCTGGCCATTCCGCCGAATGCCTAGCCACTCCGCTTGT
TGTTTGACCGCAAGGAATTAGCCGGCTTGATGATCATCCCATCCCAGCCGATCGGATCCATCAATCCCTTGCTCAGTAGAGCTAAAC
ACATCGAATAGG
>s175
CCTACGCACATGGAAGTGACCGTAACTGGACGTCGAAGCTGGCCCTATTTCCCGCTGTTGTGCGTGTAATCCTTGACAGCGGCT
TATTATCCATTGCAGCTCCCTTAATCCTCATTTGCTCGGTTGTTTTTCACATCTGAAACGTCAGTGGACACTGCACTTGTTTGACGGA
AAGACCTGGACTAGACTGCACATGAGCCAGGTAGTCTGGCCATTCCGCCGAATGCCTAGCACCCCTGTTGTTTGACCGCAAGGAATT
AGCCGGCTTGATGATCATCCATCCCAGCCGATCGGATCATCAATCCCTTGCAGTAGAGCTAAACACTCGAATAGG
>s200
CCTACGCACATGGTAGTGACCGTAACTGGACGTCGAAGCTGGCCCTATTTCCCGCTGTTGTGCGTGTAATCCTGCAGCGGCTTA
TTATCCTTGCACCCCTTAATCCTCATTTGCTGGTGTTTTTCACATCTGAAACTCAGTGGACACGCACTTGTTTGACGGAAAGACCTG
GACTAGACTGCACATGAGCAGGTAGTCTGGCCATTCCGCCAATGCTAGCACCCCTGTTGTTAGACCGCAAGGAATTAGCCGGCTTGA
TATATCCATCCAGCCGATGGGATACAATCCCTGCAGTAGAGCAAACACTGAATG
>s225
CCTACGCCATGGTAGTGACCGTGACTGACGTCGAAGCTGGCCCTAATTCCCCCTGTTGTGCGTGTAATCCGCAGCGGCTTATTA
TCCTCACCCCTTAATCCTCATTTGCTGGTGTTTTTAATCTAAACTAGGGACACGCACTTGTTTGCCGGAAAGACTGGACTGACTGCA

CTGAGCAGGTAGTCTGGCCATTCCGCCAATGCTAGCACCCCTGTAGTTAGACCCCAAGAAATTAGCCGGTGATATATCCATCCAGCC
GATGGGATACGATCCCTGGTAGAGCAAACACTGAATG

>s250

CCACGACCATGGTAGTGACCGTGACTGACTCGACTGGCCCTAATTCCCCCTGTTGTGCGGTAATCGCAGCGCTTATTATCCTCA
CCCCTTAAACCTCATTTGCTGGTGTTTTTATCAAACTAGGGACACGCACTGTTTCCCGGAAAGACTGACTGACTGACTGAGCAGGTA
GTCTGGCCAGTCCTCCAATGCTAGACCCCTGAGTTAGACCCCAAGAAATTTGCGGTGATATATCATCCAGCCGATGGGATAGATCCC
GGTAGAGAAGCACTGAATG

>s272

ACGCAGGTAGTGACTGACTGACTGACTGGCCCTAATCCCCCCTGTTTGCGTAATCGCAGCCCTAATCCCCCCCTGAAACCCATT
TGCTGGTTTTTTATCAAACTAGGGCACGCACTGTTTCCCGGAAAGACTGACTGACTGACTGAGCAGGTAGTCTGGCCAGTCCTCCAA
TGCTAGACCCCTGAGTTAGACCCCAAGAAATTTGCGGTGATATATCATCGCCGATGGGATAGATCCCGGTAGAGAAGCACTGAATG

>s272 Δgt ΔKr

ACGCAGGTAGTGACTGACTGACTGACTGGCCCTAATCCCCCCTGTTCGCGTACTCGCAGCCCTAATGCCCCCCTGAATACCAAT
TGATGGTTTTTAGTCTAACCGTATAACGCACTGTTTCCCGGAAAGACTGACTGACTGACTGAGCAGGTAGTCTGGCCAGTCCTCCAA
TGCTAGACCCCTGAGTTAGACCCCAAGAAATTTGCGGTGATATATCATCGCCGATGGGATAGATCCCGGTAGAGAAGCACTGAATG

>s272 reverse bcd

ACGCAGGTAGTGACTGACTGACTGACTGGCCCTAATCCCCCCAGTTTGCTTACTCGCAGGGGGGATTAGGGCCTGAAACCGAG
TTGCTGGTTTTTTATCAAACTACGAAGCGCACTGTTTCCCGGAAAGACTGACTGACTGACTGAGCAGGTAGTCTGGCCAGTCCTCCA
ATGCTAGACCCCTGAGTTAGACCCCAAGAAATTTGCGGTGATATATCATCGCCGATGGGATAGATCCCGGTAGAGAAGCACTGAATG

>s272 reverse bcd -5bp

ACGCAGGTAGTGACTGACTGACTGACTGGCCCTAATCCCCCCAGTTGTTCTGCGGGGGGATTAGGGCCTGAAACCGAGTTGCT
GGTTTTTTATCAAACTACGAAGCGCACTGTTTCCCGGAAAGACTGACTGACTGACTGAGCAGGTAGTCTGGCCAGTCCTCCAATGCT
AGACCCCTGAGTTAGACCCCAAGAAATTTGCGGTGATATATCATCGCCGATGGGATAGATCCCGGTAGAGAAGCACTGAATG

>s272 reverse bcd Δhb

ACGCAGGTAGTGACTGACTGACTGACTGGCCCTAATCCCCCCAGTTTGCTTACTCGCAGGGGGGATTAGGGCCTGAAACCGAG
TTGCTGGTTCTGTATCAAACTACGAAGCGCACTGTTTCCCGGAAAGACTGACTGACTGACTGAGCAGGTAGTCTGGCCAGTCCTCCA
ATGCTAGACCCCTGAGTTAGACCCCAAGAAATTTGCGGTGATATATCATCGCCGATGGGATAGATCCCGGTAGAGAAGCACTGAATG

>zld6x

ACGCAGGTAGTGACTGACTGACTGACTGGCCGCAGGTAGCCCTGTTCGCGTACTCGCTGCGGCATGTAGCCCCTGAATACCAAT
TGATCGGCAGGTAGCTAACCGTATAACGCACTCTTGCAGGTAGAGACTGACTGACTGACTGAGCAGGTAGTCTGGCCAGTCCTCCAA
TGCTAGACCCCTGAGTTAGACCCCAAGAAATTTGCGGTGATATATCATCGCCGATGGGATAGATCCCGGTAGAGAAGCACTGAATG

>bcd6x

CCGCTAAGCCCCCCGGACTGACTGACTGGCCCTAATCCCCCCTGTTCGCGTACTCGCAGCCCTAATGCCCCCCTGAATACCAAT
TCACCATAATCCGCCCCACCGTATAGGGAACCCCCGAATCCGCCCCCGGACTGACTGACCCACTAATCCGCCCGGCCAGTCCTCCAAT
GCTAGACCCCTGAGTTAGACCCCAAGAAATTTGCGGTGATATATCATCGCCGATGGGATAGATCCCGGTAGAGAAGCACTGAATG

>dst6x

TTTCCCGGAAAGACTGACTGACTGACTGGTTTCCCGGAAACCTGTTCGCGTACTCGCAGTTTCCCGGAAACCCTGAATACCGAT
TGATTTCCCGGAAACTAACCGTATAACGCACTGTTTCCCGGAAAGACTGACTGACTGACTTTTCCCGGAAACTGGCCAGTCCTCCAA
TGCTAGACCCCTGAGTTAGACCCCAAGAAATTTGCGGTGATATATCATCGCCGATGGGATAGATCCCGGTAGAGAAGCACTGAATG

>e72 Δhb

CCGGTACTGAATACCAGTGGATTCCGGACCGTAACAGGTACAGATCGAAAAGCTGGTCTGGTATCTGCTATGTGACCCCTGATA
ATCCGTTGCCACAGCGAGAATGTAAGACACTTGCAGTTACAGTATTTCGCTGTCGTCCTCGTTTCACTTTCGAGTCTAGACTTTAAT
GCAGCATCTATACCAATCGTCGCAGTTCGTTAACACGCTGTCCATATCTATCATTTAGATGGAATCGAGGGCGCTGGACCATAGTCG
CACAACGAGACCAATTGCGAAGTCAGGGCATTCCGCCGATCTAGCCATCGCTCATCTTATCCGCGTGTTTTATGAATGTTTGCGGGA
TTAGCCAAGGGTGACTTGGAATCCAATCCCGTCCATAGCCCGATCCCGACCTAATCCCAATCCCTAGTCCTCTTCATTTGAAACTCAT
AAAAACCCATAATACTGATGGGAAGGGATGAGGGG

>e72 Δhb Δbcd

CCGGTACTGAATACCAGTGGATTCCGGACCGTAACAGGTACAGATCGAAAAGCTGGTCTGGTATCTGCTATGTGACCCCTGATA
ATCCGTTGCCACAGCGAGAATGTAAGACACTTGCAGTTACAGTATTTCGCTGTCGTCCTCGTTTCACTTTCGAGTCTAGACTTTAAT
GCAGCATCTATACCAATCGTCGCAGTTCGTTAACACGCTGTCCATATCTATCATTTAGATGGAATCGAGGGCGCTGGACCATAGTCG
CACAACGAGACCAATTGCGAAGTCAGGGCATTCCGCCGATCTAGCCATCGCTCATCTTATCCGCGTGTTTTATGAATGTTTGCGGGA
TTAGCCAAGGGTGACTTGGAATCCAATCCCGTCCATAGCCCGATCCCGACCTAATCCCAATCCCTAGTCCTCTTCATTTGAAACTCAT
AAAAACCCATAATACTGATGGGAAGGGATTAGGGG

>e72 Δcic

CCGGTACTGAATACCAGTGGATTCCGGACCGTAACAGGTACAGATCGAAAAGCTGGTCTGGTATCTGCTATGTGACCCCTGATA
ATCCGTTGCCACAGCGAGAATGTAAGACACTTGCAGTTACAGTATTTCGCTGTCGTCCTCGTTTCACTTTCGAGTCTAGACTTTAAT

GCAGCATCTATACCAATCGTCGCAGTTCGTTAACACGCTGTCCATATCTATCATTTAGATGGAATCGAGGGCGCTGGACCATAGTCG
CACAACGAGACCAATTGCGAAGTCAGGGCATTCCGCCGATCTAGCCATCGCTCATCTTATCCGCGTGTTTTATGATTGTTTGCGGGA
TTAGCCAAGGGTGACTTGGAATCCAATCCCGTCCATAGCCCGATCCCGACCTAATCCCAATCCCTAGTCCTCTTCATTTGAAACTCAT
ATAAACCCATAATACTGATGGGAAGGGATGAGGGG

>MSE2 reversed bcd

CCGGTACTGCATAACAATGGAACCCGAACCGTAACTGGGACAGATCGAAAAGCTGGCCTGGTTTCTCGCTGTGTGTGCCGTGTg
gattaGTTTGCCATCAGCGAGATTATTAGTCAATTGCAGTTGCAGCGTTTCGCTTTCGTCCTCGTTTCACTTTCGAGTTAGACTTTATT
GCAGCATCTTGAACAATCGTCGCAGTTTGGTAACACGCTGTGCCATACTTTCATTTAGACGGAATCGAGGGACCCTGGACTATAATC
GCACAACGAGACCGGGTTGCGAAGTCAGGGCATTCCGCCGATCTAGCCATCGCCATCTTCTGCGGGCGTTTGTTTGTTTGTTTGCTG
GGATTAGCCAAGGGCTTGACTTGGAATCCATCCCGATCCCTAGCCCGATCCCAATCCCAcaagggattgggatTCCTTTTCATTAGAAAGTC
ATAAAAACACATAATAATGATGTCGAAGGGATTAGGGG

>s250 Δdic

CCACGCCACGGAAGTGACCGTGACTGACTCGACTGGCCCTAATTCCCCCTGTTGTGCGGTAATCGCAGCGCTTATTATCCTCAC
CCCTTAAACCTCATTTGCTGGTGTTTTTATCAAACTAGGGACACGCACTGTTTCCCGGAAAGACTGACTGACTGACTGAGCAGGTAG
TCTGGCCAGTCCTCCAATGCTAGACCCCTGAGTTAGACCCCAAGAAATTTGCGGTGATATATCATCCAGCCGATGGGATAGATCCCG
GTAGAGAAGCACTGAATG

>s250 Δbcd Δhb

CCACGCCATGGTAGTGACCGTGACTGACTCGACTGGCCCTAATTCCCCCTGTTGTGCGGTAATCGCAGCGCTTATTATCCTCAC
GCCCTAATCCCCCCCTGAAACCCATTTGCTGGTTTTTTATCAAACTAGGGACACGCACTGTTTCCCGGAAAGACTGACTGACTGACT
GAGCAGGTAGTCTGGCCAGTCCTCCAATGCTAGACCCCTGAGTTAGACCCCAAGAAATTTGCGGTGATATATCATCCAGCCGATGGG
ATAGATCCCGGTAGAGAAGCACTGAATG

>5' Extention Primer

TGGGTTTTATTAACTTACATACATACTAGAATTCGAGCTCGCCCGGGGATC

>3' Extention Primer

GTTGTTGACTGTGCGGCGGTCACAGCTCGAGTGTGCTGCTCTCAGCCACCCCGCGCCCTTTTATACCGCTGCGCTC

7.4 Appendix Files

File 7.1. A zip file containing the C++ implementation of the transcription model.

File 7.2. Florescence levels and standard deviations for *even-skipped* mRNA at T6.

File 7.3. Excel file contain all parameter sets and model scores, as well as parameter search space for all fits in Chapter 3.