Enhancer Interactions in Developmental Gene Regulation

Enhancer Interactions in Developmental Gene Regulation

A dissertation presented

by

Kelly Marie Biette

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

May 2019

Dissertation Advisor: Dr. Angela DePace Kelly Marie Biette

Enhancer Interactions in Developmental Gene Regulation

Abstract

When and where a gene is expressed during development is a critical determinant of cell identity and transcriptional mis-regulation is a common driver of diverse disease states. The spatial and temporal expression of animal genes is controlled by enhancers, sequences of DNA that are bound by transcription factors (TFs) and direct the pattern, timing, and level of gene expression. Many developmental genes are surrounded by multiple enhancers, each of which directs a subset of the overall gene expression pattern, allowing the gene to be turned on at different stages or in different tissues throughout the lifetime of the organism. However, it remains unclear how the promoter integrates information from across a gene locus such that select enhancers are active at the right times, in the right cells, and in the right combinations.

Enhancers have long been described as "modules" that function independent of distance from and orientation to the promoter and other regulatory sequences. However, there is increasing evidence in the field that interactions between enhancers and other regulatory sequences are more common than initially appreciated. The current challenge is to understand the prevalence and functional consequences of these interactions on gene expression, both at the level of a single locus and at longer genomic scales.

We challenged this canonical view of enhancer modularity with computational models and quantitative experiments using two enhancers of the *even-skipped* locus in *Drosophila melanogaster* blastoderm embryos. Using controlled molecular biology and high-resolution imaging, we moved enhancers relative to each other in reporter constructs and deleted them from the endogenous locus to demonstrate that interactions between enhancers have functional

consequences on gene expression. These results argue that mechanisms of gene regulation operate at the locus-level to control the precise expression of this key developmental gene. The evidence presented herein suggests that the classical model by which enhancers function independently of surrounding sequences is too simplistic and lays the groundwork for future studies that identify the mechanisms by which information from many enhancers is integrated by a single promoter to produce precise and robust gene expression patterns throughout development.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

ACKNOWLEDGEMENTS

If I've learned anything during the last six years, it is that science is a deeply human endeavor and these relationships are the most important thing that I leave graduate school with. I am extremely grateful to the following people for improving my experience and making this work possible.

First, I would like to thank the Department of Systems Biology (special shout out to Jennie Epp, Stephanie Badger, and Kathy Buhl) who make doing science and working at Harvard easier and more fun. I'd also like to thank my Dissertation Advisory Committee -- Drs. Mitzi Kuroda, Mike Springer, and Debbie Marks -- for your thoughtful feedback and encouragement during the entire process. Finally, thanks to Drs. Sean Megason, Karen Adelman, and Juan Fuxman Bass for serving on my defense committee, and an extra-special thanks to Mike Springer for serving on both. It's no small amount of time, and I really appreciate that you prioritize students when you certainly don't have to. Thanks also to Susan Dymecki, Tom Bernhardt, Anne O'Shea, Kate Hodgins, and Danny Gonzales. Thank you for working hard to improve the BBS program and the experience of all the students in it. To the smart and rigorous scientists in the Systems Biology department, most especially Charlotte Strandkvist and Evi Van Italie, thank you for your support, encouragement, and friendship. I feel very lucky to have trained with such inspiring and strong female scientists.

I would also like to thank our collaborators for making the experiments and ideas presented in this thesis better. To Hassan Samee and Saurabh Sinha, thank you for helpful discussions about the philosophy of merging modeling and experiments, and for your critical feedback on our work. To the members of the Zeitlinger Lab, especially Wanqing Shao, Cindi Staber, Elisa Ahern, Melanie Weilert, and Sara Jackson thank you for patiently teaching me new techniques, for sharing buffers and scripts, and for being extremely welcoming. Cindi and Elisa, I wouldn't have lasted six hours (let alone six weeks!) without your sunflowers, football parties, and line dancing lessons. Finally, thank you to Minhee Park and Mo Khalil for selflessly sharing

your reagents and ideas. Our work together taught me how fun it is to do science with friends. Minhee, I learned so much about science and hard work from you. Thank you for always being willing to try another experiment and to stare at data we don't understand; working with you has made graduate school so fun.

Second, I wouldn't be defending or in graduate school at all without a fantastic support system, for which I am exceedingly grateful. To Orrin Stone, thank you for encouraging me to think about graduate school all those years ago and for being a long time cheerleader and constant source of encouragement. To Patrick Vaughn, Lauren Ko, Ben Arevalo and Ana Frackman - thank you for being so welcoming all those years ago and letting me be on your volleyball team. "Safe Sets" kept me in Boston and in graduate school when I wanted to leave the most and playing with you over the years has been great fun.

To the GN+1 Journal Club - thank you for sharing the journey with me. Your generosity with feedback and ideas is admirable and I so value you as a community of rigorous thinkers. To Kelsey Tyssowski and Aurora Zhang, thank you for teaching me the finer points of modern style and the value of a two step plan, for always having good advice and not letting me take things too seriously. I'm lucky to have friends who are in constant pursuit of smashing the patriarchy and have taught me that no matter how things are going, the best part of the process is the friends you find to get through it with. Oh, and that pizza rolls taste better over the campfire. To Ian Hill, thank you for always being a friend that I could count on and for showing me that doing rigorous science is only part of what we came here to learn. You inspire me to look out for the people around me, and I admire your pursuit of mental wellness and better resources for our community. And thank you for marrying Adrienne Yates, in whom I have found a real kindred spirit. I'm so thankful for your friendship and am looking forward to the years ahead.

To Meghan Bragdon, Emily Dehmer, Ashley Nolan, Julia Bonadio, Mikaila Waters, Dori Gills, and Marissa Marcotte - thank you for being bright, beautiful stars in my life and for relentlessly pursuing your dreams. Thank you for tea times and phone calls, for beers in the sunshine, and afternoon bike rides. To Jennifer Wirth and Hayley Crossman, thank you for all of

the above and for the adventures of a lifetime. You inspire me to live out loud, dream big, and go after the things that are important to me, even when they seem impossible to achieve.

To my family, thank you for your support in pursuing my goals, and for all of your help moving over the years. Mom and Dad, thank you for encouraging us to think like scientists from an early age, for a childhood exploring the outdoors, and for not disowning me when I didn't go to Duke. Thank you for always being willing to provide advice when I have to make difficult choices, and for prioritizing fun times together, which have always been a great source of happiness. Katy and Sean, thank you for your optimism, happiness, and support over all the years, watching you achieve your goals has motivated me to pursue mine. I feel quite lucky to be your sister.

To Benjamin Judge, thank you for rubbing my face on the couch when I'm stressed, for your punny sense of humor, and for your unwavering cheerfulness. Thank you for support and love in a time that otherwise would have been stressful and difficult. I'd go to graduate school all over again for the chance to meet you on a springtime walk and I'm looking forward to more bear hugs, bike rides, and frozen burrito fueled adventures together.

Finally, I'd like to thank the members of the DePace lab who have made all the hard work worth it and without whom I would be a much less rigorous scientist. To Olivia Foster, thank you for reading drafts of everything, your heroic antibody staining efforts, and for always reminding me that whatever I'm feeling is ok. Thank you also for your graphics work; you are extremely talented. To Benjamin Vincent, thank you for the free therapy and life advice. Thank you for being an excellent example of mentoring trainees and of how to look out for the people around us. To MK Howard, thank you for making me cooler with your excellent playlists. To Qiu Wu, Lauren Bush, Ceejay Lee, Jacob Shenker, and Max Baas-Thomas, thank you for your patience and flexibility, and for sharing your exciting ideas. They made my projects better. To Timothy Harden, thanks for your stoke, both in science and outside. For your encouragement to keep plotting data and thoughtful feedback on my work. Long live secret projects and secret pow stashes. Finally, to Clarissa Scholes, thank you for years of coffee chats, moral support, dance

partying and collaboration. Thank you for being a partner in science and a wonderful friend. I'm looking forward to many more barefoot, book-reading days in the sunshine with you.

Finally, to Angela, it has been a great privilege to work with you. Thank you for running a lab where people are more important than papers and small victories are celebrated every day. Thank you for living by your values and encouraging us to do the same. You have built a lab that feels like a family; a place where people are celebrated for their strengths and supported in their weaknesses. Thank you for setting an example of outstanding integrity, building a culture based on feedback and growth, and for showing us that we don't have to compromise what is most important to us to be successful in science.

And to all of the DePace Lab -- including Angela, Ben Vincent, Francheska Lopez-Rivera, Meghan Bragdon, Zeba Wunderlich, Max Staller, Clarissa Scholes, Jeehae Park, Javi Estrada, Ed Pym, Lauren Bush, Olivia Foster, Anna Cha, Erik Clark, Lital Bentovim, Tim Harden, Rosa Corral, Karla Martinez, MK Howard -- thank you so much for making the lab a place that is fun to come to work in, for thinking critically about how to improve my experiments, and for your generous feedback on projects and presentations. Thank you for always being willing to have a coffee (or tea!) and hash things out. Working with you has been one of the great pleasures of my life and I am so grateful for your both selfless scientific collaboration and your genuine friendship.

This dissertation is dedicated to the synchronized swimmer, the tortured poet, the tattooed hipster, and the most interesting man in BBS. Thank you for your bottomless intellectual generosity and for making grad school tremendously fun.

Chapter 1 : Introduction

Overview

Enhancers are cis-regulatory sequences that direct the pattern, level, and timing of gene expression in animals (Long et al., 2016; Shlyueva et al., 2014; Spitz and Furlong, 2012). On the order of a few hundred base pairs in length, enhancer sequences are a scaffold to recruit transcription factors (TF) and chromatin remodeling proteins to regulate gene activation or repression. Enhancers play a critical role in development, where they direct cell-type specific gene expression to build complicated tissue types and body plans (Levine, 2010). Understanding how enhancers function is thus essential for understanding morphological evolution, as changes in regulatory networks underlie much of species diversity (Carroll, 2008). It is also essential for understanding disease susceptibility, as many disorders are caused by over- or under-expression of important genes; indeed sequence variation in enhancer sequences is common, as the vast majority of mutations occur in regulatory DNA (Maurano et al., 2012, 2015).

However, gene regulation is not governed by individual enhancers alone. Enhancers in the endogenous context are surrounded by a complex regulatory environment that includes other enhancers, promoters, insulating and silencing sequences, as well as the regulatory DNA for nearby genes; these pieces of regulatory DNA together create the gene locus (reviewed in Maston et al., 2006). If expression is controlled at the level of the gene locus, each of these other elements could influence the output of a given enhancer. Therefore, when considering the regulatory mutations, we must consider both how sequence variants within enhancers affect their function, which has received a great deal of attention (Frankel et al., 2011; Huang and Ovcharenko, 2015; Khoueiry et al., 2017; Wittkopp and Kalay, 2012), and how enhancers interact with the other sequences surrounding them, which has received much less. These interactions are particularly important to consider for developmental and regulatory genes that are used in multiple contexts; such genes contain multiple enhancer elements where each individual enhancer is thought to control some portion of a gene's spatial and temporal activity (Levine, 2010; Shlyueva et al., 2014; Spitz and Furlong, 2012).

In the last decade we have amassed a vast library of annotated enhancers using functional genomics. Using sequencing based methods that annotate regions of high transcription factor occupancy, DNA accessibility, diagnostic chromatin modifications, or a combination of these features, tens of thousands of enhancers are annotated in flies (Kvon et al., 2014) and hundreds of thousands of enhancers in humans (Jin et al., 2013; Pennacchio et al., 2013). Furthermore, chromatin-conformation studies have suggested that the average enhancer contacts multiple other enhancers and promoters in both flies (Ghavi-Helm et al., 2014) and human cells (Jin et al., 2013). But this information is not adequate for predicting gene expression at the locus level in two key ways. First, the identification of enhancers using empirical methods remains cell-type specific. Thus, there may be cryptic, unannotated information in any given locus. Second, the rules and consequences of enhancer interactions with other types of regulatory DNA remain to be defined.

Historically, the field of enhancer biology has considered enhancers to act independently, with a chromatin loop bringing a distal "active" enhancer into contact with a target promoter in a simple pairwise interaction (Benabdallah and Bickmore, 2015; Furlong and Levine, 2018). In this model, "inactive" enhancers are left out of the loop and rendered non-functional. As the regulatory DNA surrounding a gene can span thousands, or even millions, of base pairs in flies and humans (Levine, 2010), it is difficult to test all regions for activity in different developmental stages, resulting in this simple model of enhancer function. Furthermore, a focus on identifying and minimizing individual enhancers that drive only a subset of the gene's total expression pattern made it possible to break this complicated problem down, which has certainly provided insight into how TFs bind enhancers and turn on a gene (Arnosti et al., 1996a, 1996b; Li and Arnosti, 2011; Small et al., 1991, 1992). However, whether observed interactions between enhancers and other regulatory sequences are functional and important for endogenous expression has historically received less attention.

If enhancers do not act independently of the sequence context they are embedded in, studies of individual enhancers in isolation will not accurately predict how gene expression is

controlled in complex, endogenous loci. This predictive understanding is critical for identifying which changes in regulatory sequence will affect downstream organismal phenotypes, including disease susceptibility. Regulatory sequence variants occur at multiple scales, from SNPs that may impact TF binding to structural variants that affect locus-level properties, such as moving or deleting entire enhancers (Maurano et al., 2015; Weischenfeldt et al., 2013). In this thesis, I argue that we must think about how enhancers interact with other enhancers and other types of regulatory sequences if the ultimate goal is a predictive understanding of regulatory DNA function. In this chapter, I will discuss how historical studies of enhancers gave rise to the definition of enhancers as independent modules and recent work that is pushing the field towards a context-dependent view of enhancer function. Finally, I argue for the value of a single locus case study in this complicated field that often relies on genome-wide measurements and introduce our experiments that use two enhancers of a single gene to dissect the interactions between them.

Enhancers have long been defined as independent modules

The gold standard for identifying and validating enhancers has long been their ability to drive expression in a reporter construct, where the region of interest is placed upstream of a minimal promoter and the resulting expression is measured in embryos or cell lines (Catarino and Stark, 2018; Halfon, 2018). This assay has provided tremendous insight into how transcription is activated or repressed by the TFs bound to each enhancer. Given its utility, recent work has dramatically increased the throughput of reporter assays which enables the testing of many sequences for activity (Arnold et al., 2013, 2014; Inoue and Ahituv, 2015; Kheradpour et al., 2013; Melnikov et al., 2012). However, the regulatory information that controls each gene may span millions of base pairs and so it remains difficult, if not impossible, to assay all regions for activity by this classical assay in multiple cell types. But even more fundamentally, this assay cannot identify sequences that do not drive expression on their own (so-called “modifying” sequences), or capture any functional interactions with other sequences

present in the endogenous context, such as other enhancers, insulator elements, or architectural proteins that control locus topology (Cubeñas-Potts et al., 2017; Van Bortle and Corces, 2013; Van Bortle et al., 2014).

Early experiments in reporters led to the following definition of an enhancer, often cited throughout the literature: “the … enhancer can act over very long distances, and independent of its orientation” (Banerji et al., 1981). While enhancers do demonstrate remarkable flexibility in their ability to drive expression, this definition is often paraphrased to describe enhancers as independently functioning modules, which has both experimental and computational consequences. For example, enhancers are often minimized because, by their very nature, these reporter assays select for small compact regions to test for activity. While it may be possible to identify the minimum amount of sequence necessary to recapitulate a particular expression pattern, flanking sequences may be important for other functions of the enhancer, such as robustness to environmental perturbation (as is the case for the Drosophila even-skipped stripe 2 enhancer (Ludwig et al., 2011; Small et al., 1992)). In addition, genomic-scale measurements of features marking developmental enhancers, such as TF binding and chromatin accessibility, demonstrate that enhancer boundaries are likely not rigid, despite strict annotation cut-offs for minimal enhancer sequences (Li et al., 2008).

The experiments that initially identified enhancers claimed that : they function independent of distance and orientation with respect to the promoter. However, this claim is based on an experiment where a viral enhancer (from SV40) is placed upstream of a plasmid containing a basal promoter and measures the resulting expression qualitatively (Banerji et al., 1981). While the idea that a sequence could activate expression over large distances was exciting, these early studies had limited quantitative resolution. Despite this, some distance-dependent effects were observable when enhancers were placed at various distances from the promoter (Moreau et al., 1981; Wasylyk et al., 1984), suggesting that enhancers are capable of acting across distances and at both orientations, but that the precise expression levels they drive may be sensitive to these types of changes in context. The position and orientation dependence of

enhancer function has now been noted in a variety of places in the literature (Wilson et al., 1990); this dependence is another aspect of context dependence that is intrinsic to the organization of the locus. Supporting the functional importance of locus organization, the location of enhancers relative to their target gene and other enhancers is conserved across diverse insect species (Cande et al., 2009; Engström et al., 2007), suggesting that the spatial arrangement of enhancers is critical for the overall output of the gene. Moreover, there are examples of developmental phenotypes that arise from structural variants in regions surrounding individual genes (Klopocki et al., 2008; Kurth et al., 2009).

The ability of enhancers to activate expression at a distance combined with the functional validation of small fragments that drive a portion of a total gene expression pattern led to a simple working definition of enhancers as independent modules, which individually drive expression of a subset of the total pattern of the gene. The pattern was thought to result from a simple sum of the output of the individual enhancers. One canonical example of a modular developmental locus is the *even-skipped* (*eve*) gene in Drosophila development. During the blastoderm stage, *eve* is expressed in seven stripes which are controlled by five enhancers (Figure 1.1 A-B), with each enhancer directing expression of one or two stripes (Fujioka et al., 1999; Small et al., 1991, 1996).

This modular view of enhancer function is convenient because it enables complex gene regulatory control circuits and the reuse of genes at multiple times throughout development without pleiotropy (Carroll, 2000; Kirschner and Gerhart, 1998; Levine, 2010). For example, the same gene could be used in early developmental patterning and then re-expressed stably to specify a cell type in the adult, presumably through binding of distinct sets of TFs to different enhancer regions surrounding the gene. In *eve,* for example, additional enhancers aside from the five enhancers active in the blastoderm drive *eve* expression in other tissues, including in the anal plate (Frasch and Levine, 1987) and in the developing nervous system (Fujioka et al., 1999). Importantly, distinct enhancer regions surrounding a single gene allow for TF inputs to be

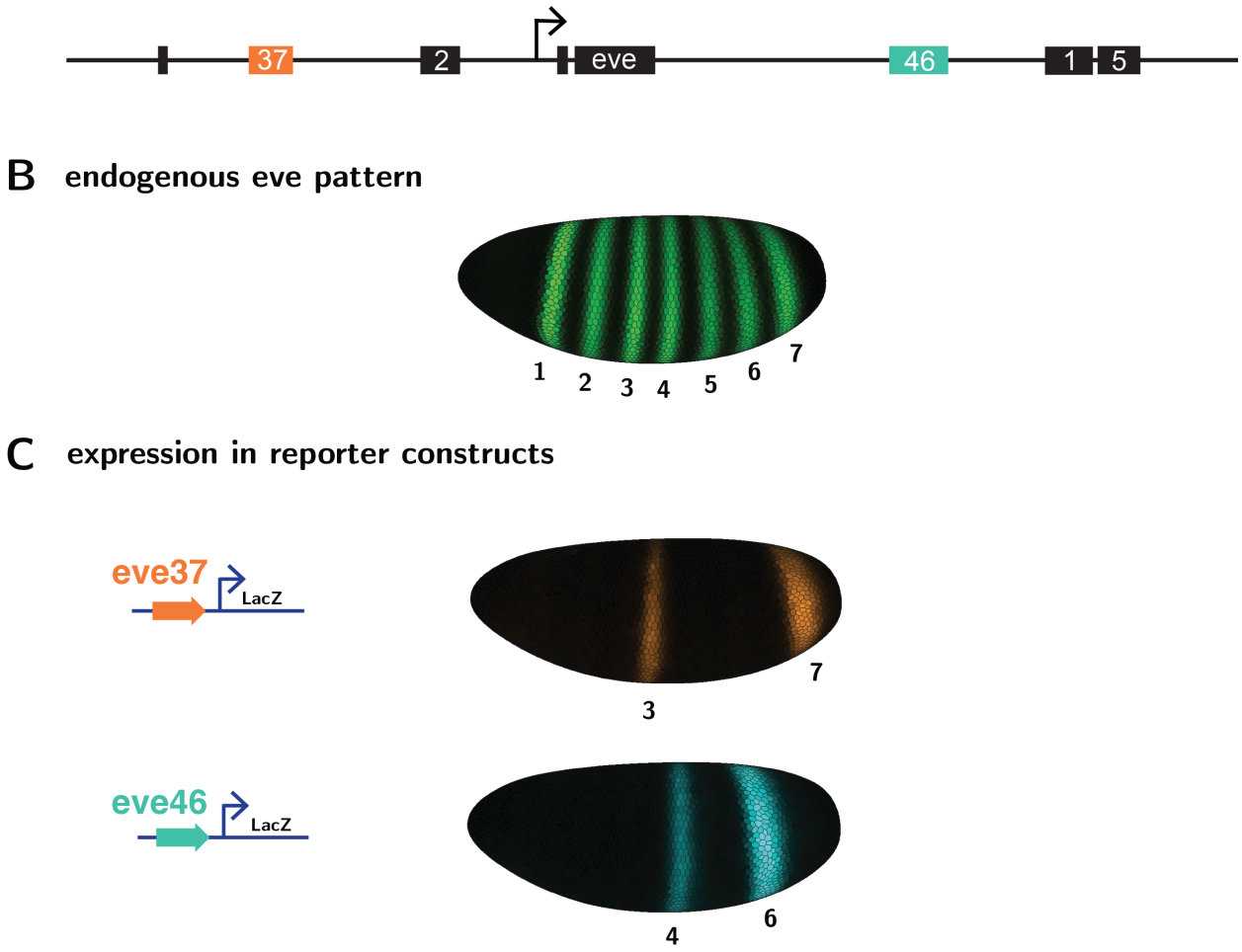


Figure 1.1 : The *even-skipped* (*eve*) locus contains five enhancers that drive expression of seven stripes. (A) The eve locus is contained in a 16kb chromatin domain. The relative positions of the five enhancers and eve coding sequence are shown, with the corresponding stripe numbers listed on each enhancer. Eve37 and eve46 are highlighted in orange and blue, respectively. (B) A computational rendering of the endogenous seven stripe *eve* mRNA expression pattern (green). Data is from Fowlkes et al., 2008. (C) A computational rendering of expression driven by an eve37-LacZ reporter construct is shown in orange and expression driven by an eve46-LacZ reporter construct is shown in blue. Throughout this thesis we used in situ hybridization to detect expression of transgenic reporters.

integrated under different contexts, rather than having all regulation channeled through the same sequences for different functions.

One complexity for this modular view is when two enhancers are active in the same cell. Termed “shadow” enhancers, they are found in both flies and higher mammals and their output is not simply additive (Dunipace et al., 2011; Frankel et al., 2010; Hong et al., 2008; Lam et al., 2015; Perry et al., 2010, 2011; Staller et al., 2015; Wunderlich et al., 2015). A recent survey of Drosophila mesoderm development found that 64% of gene loci contained shadow enhancers, suggesting that they are pervasive during fly development (Cannavò et al.). Shadow enhancers are thought to convey robustness to environmental or genetic perturbation due to their partial redundancies, and only just recently have studies began to suggest mechanisms by which they might interact (Bothma et al., 2015; Scholes et al., 2019).

Furthermore, we should also consider the potential complexity introduced by “inactive” enhancers affecting “active” enhancers. In the most common conceptual cartoon of enhancer function, an active enhancer is bound by TFs and looped to the promoter, while inactive enhancers are relegated to the sideline and have no effect. However, measurements of three-dimensional genome organization suggest that there is a higher order structure to most loci, whereby enhancers are organized into topologically associated domains (TADs) that bring specific regions into close proximity, such that they preferentially contact one another (de Laat and Duboule, 2013). This higher-order structure suggests that even inactive enhancers are in close physical proximity to active enhancers and could contribute to their activation or actively need to be held in an “inactive” state.

Measurements of TF binding and chromatin accessibility in Drosophila embryos suggest that it is difficult to distinguish between “active” and “inactive” enhancers during the same developmental time point using functional genomic features, but this is likely due to the low spatial resolution of these assays in whole embryos (Li et al., 2008; Thomas et al., 2011). A recent study using ATAC-seq on embryos sliced in half along the anterior-posterior axis found that genome-wide accessibility was similar between the two halves, even around promoters that are exclusively active in one half of the embryo (Haines and Eisen, 2018). As the technology and

resolution behind these assays continue to improve, it will be more feasible to use them to measure enhancer states and how dynamically they change over developmental time.

Multiple examples contradict the modular view

As the regulatory elements surrounding more loci are studied, there are more examples of enhancers not behaving as simple modules. In flies, there are genes controlled by more than one enhancer that cannot be minimized into separable elements (Klingler et al., 1996; Ludwig et al., 2005; Yuh and Davidson, 1996; Yuh et al., 1998). In the *sloppy-paired* (*slp1*) locus, two enhancers work together to drive the endogenous expression pattern, but they do so in a non-additive way: one enhancer produces some slp1 stripes, while the other expresses all stripes and expresses ectopically between stripes. Only when both enhancers are present is ectopic expression repressed and endogenous expression recapitulated (Prazak et al., 2010). In the *Ultrabithorax (Ubx)* locus, several enhancer elements are distributed over ~100kb, and interactions between them are critical for proper expression. A subset of these enhancers are responsible for the early temporal pattern, and are then repressed by a Polycomb-response element while a late set of enhancers, that are not responsive to this silencing, take over (Pirrotta et al., 1995; Qian et al., 1991, 1993). How the Polycomb response element silences certain enhancers but not others, and how "handoff" between these elements occurs is fully understood. In Ubx patterning in the legs, it has been impossible to identify an enhancer responsible for pupal expression in the second femur (Davis et al., 2007), suggesting that some loci are controlled by completely distributed elements that do not work outside their endogenous context. Similar examples of complex regulation by multiple elements exist in other species (Maekawa et al., 1989; Proudhon et al., 2016; Stine et al., 2011).

Genome wide functional data also reveals contradictions to the simple modular view by which a single enhancer drives expression independent of its surroundings. Enhancers physically interact with each other as well as the promoter (Chen et al., 2018; Ghavi-Helm et al., 2014; Proudhon et al., 2016), with active promoters contacting at least two enhancers on

average in human cells (Jin et al., 2013). Many developmental genes are surrounded by clusters of highly active enhancers, termed “super-enhancers,” that may function together to synergistically control gene expression (Lovén et al., 2013; Pott and Lieb, 2015; Siersbæk et al., 2014; Whyte et al., 2013). The effects of deleting component enhancers from these clusters are highly variable, and both synergistic and additive effects have been observed (Bahr et al., 2018; Huang et al., 2018; Moorthy et al., 2017; Suzuki et al., 2017). These results are often enhancer-specific, as studies that delete multiple component enhancers from the same super-enhancer report that some elements are critical for the overall gene expression pattern, while others are dispensable for function (Carleton et al., 2017; Hay et al., 2016; Le Noir et al., 2017).

These new results are pushing the field towards a paradigm of “transcriptional hubs,” where enhancers and other regulatory sequences are sequestered in preformed topologies that enable functional interactions between these elements. In this model, many relevant regulatory regions are all in close proximity, which allows for a high local concentration of TFs, PolII, and other protein complexes required for activation of the target gene (reviewed in (Furlong and Levine, 2018). While still speculative, this framework incorporates many of the recently puzzling observations of transcription factor and Pol II dynamics in flies (Mir et al., 2017; Tsai et al., 2017) and mammalian cells (Cho et al., 2018; Chong et al., 2018; Cisse et al., 2013; Sabari et al., 2018). For example, the enhancers driving the Drosophila *shavenbaby* (*svb*) locus contain low-affinity binding sites for the transcription factor Ubx. These enhancers colocalize, despite being physically separated in linear space, and produce a microenvironment with a high concentration of Ubx and its cofactor Homothorax (Hth), which is thought to allow Ubx to utilize low affinity binding sites and precisely activate expression of this key developmental gene (Tsai et al., 2017). These highly dynamic transcriptional environments appear to be more common than first appreciated, and allow for physical and functional interactions between enhancers that the previous modular view did not explain. The challenge that remains is to decipher which proteins or other features mediate these interactions, and the consequences on gene expression of disrupting them.

The value of a careful case study

Given the state of the field, it is important to ask whether these examples of enhancers not behaving modularly remain the exception or if we need to update our definitions of enhancer function to include context-dependent nuance. While the idea that enhancers may interact with each other and additional classes of regulatory sequences is becoming more accepted, the picture remains complicated and few labs are studying this problem using mechanistic experiments. Rather than working at the genome-scale to identify all putative interactions between enhancers, we argue for a careful case study to dissect the phenomenology of enhancer interactions in the canonical modular locus, *eve*. We chose two enhancers in this locus, and use experiments in reporter constructs and the endogenous locus to move them around relative to each other and the promoter. Much like experiments using single enhancers produced many of the rules that govern TF-TF interactions, we hope that working at the single locus scale will allow us to interrogate the sequences within and between enhancers that are responsible for modulating gene expression. Our goal was to use *eve* to test mechanistic hypotheses about how enhancers work together to control gene expression, with the hope that our results might inform what sequences people look for in studies of enhancer-interactions genome wide.

We chose to work in *eve* because of several advantages. First, we wanted a locus that is controlled by multiple enhancers, each of which has been studied in reporter constructs and is well-defined in the literature. These enhancers should behave modularly, each driving a subset of an overall gene expression pattern that can be summed together to produce to the total pattern, given that our null hypothesis in these experiments was modular enhancer function. Additionally, we wanted the input TFs of these enhancers to be well studied, such that the effect of perturbations is predictable and can be compared to existing datasets. The Drosophila blastoderm is an excellent system for studies of enhancer function, as it is patterned by an extremely well studied transcriptional network where the expression patterns and DNA binding preferences of all input TFs are known (Frasch and Levine, 1987; Jaeger, 2011; Small and

Levine, 1991; Small et al., 1991). The blastoderm is unmatched in our ability to precisely measure gene expression in intact, developing animals with high spatial and temporal resolution (Fowlkes et al., 2008, 2011; Keranen et al., 2006; Luengo Hendriks et al., 2006). Finally, both genetic and biochemical manipulations are highly tractable, including transgenic reporters and genome editing with CRISPR (Fish et al., 2007; Lamb et al., 2017), enabling us to precisely manipulate and quantitatively measure the effects of these changes.

My thesis tests the hypothesis that there are higher-order interactions between enhancers at the locus scale, using two enhancers of *eve* as a case study. *Eve* is expressed during blastoderm development in Drosophila embryos in a seven-stripe pattern that is driven by five enhancers (Frasch and Levine, 1987; Fujioka et al., 1999; Goto et al., 1989; Sackerson et al., 1999). As a pair-rule gene, it is expressed in seven stripes that eventually prefigures the body segments of the larvae. *Eve* inputs and outputs are well studied and have been defined by decades of classical genetics (Fujioka et al., 1999; Jaeger, 2011; Janssens et al., 2006). We chose two *eve* enhancers for our experiments, one that directs expression of stripes 3 and 7 (eve37) and one that directs expression of stripes 4 and 6 (eve46). These enhancers are separated by approximately 9kb, and eve37 is located 5' of the *eve* coding sequence while eve46 is the first 3' enhancer following the end of the 3' untranslated region (Figure 1.1 A, C).

Eve37 and eve46 have unique advantages for testing the limits of enhancer modularity. First, they are extremely well characterized, and the regulators they respond to are known (Clyde et al., 2003; Frasch and Levine, 1987; Fujioka et al., 1999; Struffi et al., 2011). Second, having a four stripe readout allows us to observe subtle phenotypic effects if the position or level of expression of any of these stripes changes during our experiments. Furthermore, these enhancers are both bound by the repressors Hunchback (Hb) and Knirps (Kni), but respond with different sensitivities, allowing us to test the hypothesis that enhancers interact to facilitate TF sharing across the locus when multiple enhancers are all bound by partially overlapping sets of input regulators. Finally, these enhancers are located on opposite sides of the endogenous

locus, which is ideal for interrogating the role of architectural proteins and looping factors in determining the DNA topology required for precise expression of these stripes.

Previous studies have perturbed the *eve* enhancers, but focused mainly on the 5' end of the locus. Removal of the 1.7kb spacer sequence that separates the two enhancers that drive *eve* stripe 2 (eve2) and *eve* stripes 3 and 7 (eve37) produced abnormal stripe levels and ectopic expression (Small et al., 1993). The authors speculate that these differences are due to TF interactions at the junctions of fused enhancers, however these effects do not disappear when short (160 and 300bp) spacer sequences are placed between the enhancers. In our experiments, we test more arrangements that cover all the junctions between enhancers and use computationally designed "neutral" spacers to remove binding sites for blastoderm TFs from these sequences.

More recently, studies have used live imaging of transcription to assess the consequences of deleting single *eve* enhancers from bacterial artificial chromosomes (BACs) and from a reporter integrated into the endogenous gene (Berrocal et al., 2018; Lim et al., 2018a, 2018b). Their work demonstrates that the *eve* pattern refines into stripes during the course of nuclear cycle 14, with stripes shifting anteriorly over time. Enhancer-enhancer interactions appear to be important for the precision of this patterning, as an endogenous deletion of the *eve* stripe 1 enhancer leads to earlier and more anterior expression of *eve* stripe 2 (Lim et al., 2018b). Another study using an *eve* BAC replaced the stripe 1 and stripe 2 enhancers with bacterial sequences, and they did not report significant effects on stripes driven by other enhancers, although quantitative effects are possible (Li and Eisen). These examples partially support the hypothesis that *eve* enhancers may not behave independently, and that interactions may be enhancer-specific.

Challenging the canonical definition of enhancer modularity

Given that there are many examples of enhancers not behaving modularly, we dissect the limits of this definition in two ways. First, we focus on how expression driven by one enhancer is

affected by adding a second enhancer to the same reporter construct, as modular enhancers should not be affected by the presence of an "inactive" enhancer sequence nearby. Second, we also investigate the "distance and orientation independence" of *eve* enhancers by moving them around relative to each other and the promoter in reporter constructs. To complement these experiments, we compare our data to computational models of gene expression and measure the effects of deleting enhancers from the endogenous locus.

Using rearrangements of these two enhancers in reporter constructs, we observe complex interactions between them that we do not yet understand. In Chapter 2, we fused the eve37 and eve46 enhancers together and show that all combinations of these two enhancers drive four stripes, but the level of each stripe changes depending on the position of the enhancer in the fusion. When we apply computational models of enhancer function to these data, the model that performs better splits the enhancer sequence into multiple "modules," despite having sequence characteristics typical of single enhancers. These results suggest that modules are defined and integrated at the locus-level, and argue for new types of experiments to decipher the mechanisms by which this process occurs.

To explore the interaction between these fused enhancers more extensively, we tested a series of variants of eve37 and eve46 in reporter constructs. We show that interactions between eve37 and eve46 cannot be explained simply by orientation or distance from the promoter, or by interactions between TFs at the enhancer junctions. Interestingly, eve37 and eve46 still affected each others output even when separated by large chunks of neutral sequence, or by an entire coding region, suggesting that interactions between these enhancers are configuration dependent and occur over long range. Finally, we delete eve37 and eve46 individually from the endogenous locus and measure the consequences on expression level and position of stripes driven by other enhancers. Deletion of each enhancer demolishes expression of the cognate *eve* stripes, and in the case of eve37, also affects stripe expression driven by other enhancers. These results, described in Chapter 3, argue that interactions between *eve* enhancers are important for the precise expression of this key developmental gene.

Interrogating interactions using synthetic biology

Finally, we believe that if we truly understand something we should be able to build it. In Chapter 4, I describe the synthetic system we built to investigate combinatorial control of transcription by multiple activators. The design of this study arose from thinking about how enhancers interact in a gene locus to control transcription, but, given their complexity, this is difficult to study in isolation. Since enhancers are made up of binding sites for many different transcription factors, and the identity and combination of TFs that they bind dictate the overall output of the enhancer, considering how combinations of TFs work together is an analogous problem at a smaller scale.

This effort began with theoretical work that challenged the canonical mode of combinatorial control, whereby TFs work together via direct physical interactions to recruit each other and RNA polymerase. In this mathematical framework, all transcript regulation is channeled through this single step (Bintu et al., 2005; Gertz et al., 2009; He et al., 2010). However, gene regulation in eukaryotes is known to be controlled on a cycle (Fuda et al., 2009; Nechaev and Adelman, 2011), and TFs may work on distinct steps of this process to control transcription kinetically (Beagrie and Pombo, 2016; Herschlag and Johnson, 1993). Another graduate student in the lab demonstrated that a kinetic model of combinatorial control can produce the same regulatory computations as a thermodynamic model, but without requiring TFs to physically interact on the DNA (Scholes et al., 2016). Because of our common interest in how enhancers and activators work together to control transcription, we collaborated on the design of an experimental platform to interrogate combinatorial control by multiple TFs.

We built a synthetic platform to recruit TFs alone and in combination to a reporter gene promoter in mammalian cells. Using synthetic zinc fingers, we can fuse activation domains of interest to a binding domain that specifically recognizes an engineered site that is orthogonal to the mammalian genome (Keung et al., 2014; Khalil et al., 2012; Park et al., 2018). With this system, we fuse activation domains to identical DNA binding domains such that only one TF can be bound to the promoter at a time, and we measure expression driven by each TF alone and

both together. In our pilot experiments, we see evidence for functional, rather than physical, interactions between TFs that depend on the affinity of the binding domain and biochemical role of the activators.

This platform is unique in that it enables us to interrogate the function of many TFs; its modular and flexible design can be used to ask a range of questions about TF, promoter, and enhancer function. The zinc-finger sites enable us to recruit many kinds of TFs in any combination, and the dropping cost of DNA synthesis makes it possible to test many site configurations in a highly controlled manner. In the long term, this modular platform could provide a way to interrogate locus level control in a simplified system using synthetic "enhancers" that consist of multiple kinds of sites or multiple arrays at varying distances from the promoter. Finally, experiments using a similar system are being done in yeast and could be done in flies, enabling us to compare regulatory principles across multiple species (yeast, flies, human). I discuss these experiments and other future directions in the discussion (Chapter 5).

References

Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074–1077.

Arnold, C.D., Gerlach, D., Spies, D., Matts, J.A., Sytnikova, Y.A., Pagani, M., Lau, N.C., and Stark, A. (2014). Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nat. Genet. *46*, 685–692.

Arnosti, D.N., Gray, S., Barolo, S., Zhou, J., and Levine, M. (1996a). The gap protein knirps mediates both quenching and direct repression in the Drosophila embryo. EMBO J. *15*, 3659–3666.

Arnosti, D.N., Barolo, S., Levine, M., and Small, S. (1996b). The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development *122*, 205–214.

Bahr, C., von Paleske, L., Uslu, V.V., Remeseiro, S., Takayama, N., Ng, S.W., Murison, A., Langenfeld, K., Petretich, M., Scognamiglio, R., et al. (2018). A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. Nature *553*, 515–520.

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

Beagrie, R.A., and Pombo, A. (2016). Gene activation by metazoan enhancers: Diverse mechanisms stimulate distinct steps of transcription. Bioessays.

Benabdallah, N.S., and Bickmore, W.A. (2015). Regulatory Domains and Their Mechanisms. Cold Spring Harb. Symp. Quant. Biol. *80*, 45–51.

Berrocal, A., Lammers, N.C., Garcia, H.G., and Eisen, M.B. (2018). Kinetic sculpting of the seven stripes of the Drosophila even-skipped gene.

Bintu, L., Buchler, N.E., Garcia, H.G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005). Transcriptional regulation by the numbers: models. Curr. Opin. Genet. Dev. *15*, 116–124.

Blau, J., Xiao, H., McCracken, S., O'Hare, P., Greenblatt, J., and Bentley, D. (1996). Three functional classes of transcriptional activation domain. Mol. Cell. Biol. *16*, 2044–2055.

Bothma, J.P., Garcia, H.G., Ng, S., Perry, M.W., Gregor, T., and Levine, M. (2015). Enhancer additivity and non-additivity are determined by enhancer strength in the Drosophila embryo. Elife *4*.

Cande, J., Goltsev, Y., and Levine, M.S. (2009). Conservation of enhancer location in divergent insects. Proc. Natl. Acad. Sci. U. S. A. *106*, 14414–14419.

Cannavò, E., Khoueiry, P., Garfield, D.A., Geeleher, P., Zichner, T., Hilary Gustafson, E., Ciglar, L., Korbel, J.O., and Furlong, E.E.M. Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. Curr. Biol.

Carey, M. (1998). The enhanceosome and transcriptional synergy. Cell *92*, 5–8.

Carleton, J.B., Berrett, K.C., and Gertz, J. (2017). Multiplex Enhancer Interference Reveals Collaborative Control of Gene Regulation by Estrogen Receptor α-Bound Enhancers. Cell Syst *5*, 333–344.e5.

Carroll, S.B. (2000). Endless forms: the evolution of gene regulation and morphological diversity. Cell *101*, 577–580.

Carroll, S.B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell *134*, 25–36.

Catarino, R.R., and Stark, A. (2018). Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. Genes Dev. *32*, 202–223.

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., and Gregor, T. (2018). Dynamic interplay between enhancer-promoter topology and gene activity. Nat. Genet.

Cho, W.-K., Spille, J.-H., Hecht, M., Lee, C., Li, C., Grube, V., and Cisse, I.I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science eaar4199.

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G.M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X., et al. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. Science *361*.

Cisse, I.I., Izeddin, I., Causse, S.Z., Boudarene, L., Senecal, A., Muresan, L., Dugast-Darzacq, C., Hajj, B., Dahan, M., and Darzacq, X. (2013). Real-time dynamics of RNA polymerase II clustering in live human cells. Science *341*, 664–667.

Clyde, D.E., Corado, M.S.G., Wu, X., Paré, A., Papatsenko, D., and Small, S. (2003). A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. Nature *426*, 849–853.

Cubeñas-Potts, C., Rowley, M.J., Lyu, X., Li, G., Lei, E.P., and Corces, V.G. (2017). Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. Nucleic Acids Res. *45*, 1714–1730.

Davis, G.K., Srinivasan, D.G., Wittkopp, P.J., and Stern, D.L. (2007). The function and regulation of Ultrabithorax in the legs of Drosophila melanogaster. Developmental Biology *308*, 621–631.

Duarte, F.M., Fuda, N.J., Mahat, D.B., Core, L.J., Guertin, M.J., and Lis, J.T. (2016). Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation. Genes Dev. *30*, 1731–1746.

Dunipace, L., Ozdemir, A., and Stathopoulos, A. (2011). Complex interactions between cis-regulatory modules in native conformation are critical for Drosophila snail expression. Development *138*, 4075–4084.

Engström, P.G., Ho Sui, S.J., Drivenes, O., Becker, T.S., and Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res. *17*, 1898–1908.

Fish, M.P., Groth, A.C., Calos, M.P., and Nusse, R. (2007). Creating transgenic Drosophila by microinjecting the site-specific phiC31 integrase mRNA and a transgene-containing donor plasmid. Nat. Protoc. *2*, 2325–2331.

Fowlkes, C.C., Hendriks, C.L.L., Keränen, S.V.E., Weber, G.H., Rübel, O., Huang, M.-Y., Chatoor, S., DePace, A.H., Simirenko, L., Henriquez, C., et al. (2008). A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. Cell *133*, 364–374.

Fowlkes, C.C., Eckenrode, K.B., Bragdon, M.D., Meyer, M., Wunderlich, Z., Simirenko, L., Luengo Hendriks, C.L., Keränen, S.V.E., Henriquez, C., Knowles, D.W., et al. (2011). A conserved developmental patterning network produces quantitatively different output in multiple species of Drosophila. PLoS Genet. *7*, e1002346.

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

Frankel, N., Erezyilmaz, D.F., McGregor, A.P., Wang, S., Payre, F., and Stern, D.L. (2011). Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. Nature *474*, 598–603.

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

Fuda, N.J., Ardehali, M.B., and Lis, J.T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature *461*, 186–192.

Fujioka, M., Emi-Sarker, Y., Yusibova, G.L., Goto, T., and Jaynes, J.B. (1999). 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.

Furlong, E.E.M., and Levine, M. (2018). Developmental enhancers and chromosome topology. Science *361*, 1341–1345.

Gertz, J., Siggia, E.D., and Cohen, B.A. (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature *457*, 215–218.

Ghavi-Helm, Y., Klein, F.A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E.E.M. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. Nature *512*, 96–100.

Goto, T., Macdonald, P., and Maniatis, T. (1989). 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.

Haines, J.E., and Eisen, M.B. (2018). Patterns of chromatin accessibility along the anterior-posterior axis in the early Drosophila embryo. PLoS Genet. *14*, e1007367.

Halfon, M.S. (2018). Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. Trends Genet.

Hay, D., Hughes, J.R., Babbs, C., Davies, J.O.J., Graham, B.J., Hanssen, L.L.P., Kassouf, M.T., Marieke Oudelaar, A., Sharpe, J.A., Suciu, M.C., et al. (2016). Genetic dissection of the [alpha]-globin super-enhancer in vivo. Nat. Genet. *48*, 895–903.

He, X., Samee, M.A.H., Blatti, C., and Sinha, S. (2010). Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput. Biol. *6*.

Herschlag, D., and Johnson, F.B. (1993). Synergism in transcriptional activation: a kinetic view. Genes Dev. *7*, 173–179.

Hong, J.-W., -W. Hong, J., Hendrix, D.A., and Levine, M.S. (2008). Shadow Enhancers as a Source of Evolutionary Novelty. Science *321*, 1314–1314.

Huang, D., and Ovcharenko, I. (2015). Identifying causal regulatory SNPs in ChIP-seq enhancers. Nucleic Acids Research *43*, 225–236.

Huang, J., Li, K., Cai, W., Liu, X., Zhang, Y., Orkin, S.H., Xu, J., and Yuan, G.-C. (2018). Dissecting super-enhancer hierarchy based on chromatin interactions. Nat. Commun. *9*, 943.

Inoue, F., and Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. Genomics *106*, 159–164.

Jaeger, J. (2011). The gap gene network. Cell. Mol. Life Sci. *68*, 243–274.

Janssens, H., Hou, S., Jaeger, J., Kim, A.-R., Myasnikova, E., Sharp, D., and Reinitz, J. (2006). Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. Nat. Genet. *38*, 1159–1165.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.

Keranen, S.V.E., Fowlkes, C.C., Luengo Hendriks, C.L., Sudar, D., Knowles, D.W., Malik, J., and Biggin, M.D. (2006). Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution II: dynamics. Genome Biol *7*, R124.

Keung, A.J., Bashor, C.J., Kiriakov, S., Collins, J.J., and Khalil, A.S. (2014). Using targeted chromatin regulators to engineer combinatorial and spatial transcriptional regulation. Cell *158*, 110–120.

Khalil, A.S., Lu, T.K., Bashor, C.J., Ramirez, C.L., Pyenson, N.C., Joung, J.K., and Collins, J.J. (2012). A synthetic biology framework for programming eukaryotic transcription functions. Cell *150*, 647–658.

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. *23*, 800–811.

Khoueiry, P., Girardot, C., Ciglar, L., Peng, P.-C., Hilary Gustafson, E., Sinha, S., and Furlong, E.E.M. (2017). Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. eLife Sciences *6*, e28440.

Kirschner, M., and Gerhart, J. (1998). Evolvability. Proc. Natl. Acad. Sci. U. S. A. *95*, 8420–8427.

Klingler, M., Soong, J., Butler, B., and Gergen, J.P. (1996). Disperse versus Compact Elements for the Regulation of runt stripes in Drosophila. Dev. Biol. *177*, 73–84.

Klopocki, E., Ott, C.-E., Benatar, N., Ullmann, R., Mundlos, S., and Lehmann, K. (2008). A microduplication of the long range SHH limb regulator (ZRS) is associated with triphalangeal thumb-polysyndactyly syndrome. J. Med. Genet. *45*, 370–375.

Kurth, I., Klopocki, E., Stricker, S., van Oosterwijk, J., Vanek, S., Altmann, J., Santos, H.G., van Harssel, J.J.T., de Ravel, T., Wilkie, A.O.M., et al. (2009). Duplications of noncoding elements 5′ of SOX9 are associated with brachydactyly-anonychia. Nat. Genet. *41*, 862.

Kvon, E.Z., Kazmar, T., Stampfel, G., Yáñez-Cuna, J.O., Pagani, M., Schernhuber, K., Dickson, B.J., and Stark, A. (2014). Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature *512*, 91–95.

de Laat, W., and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. Nature *502*, 499–506.

Lam, D.D., de Souza, F.S.J., Nasif, S., Yamashita, M., López-Leal, R., Otero-Corchon, V., Meece, K., Sampath, H., Mercer, A.J., Wardlaw, S.L., et al. (2015). Partially redundant enhancers cooperatively maintain Mammalian pomc expression above a critical functional threshold. PLoS Genet. *11*, e1004935.

Lamb, A.M., Walker, E.A., and Wittkopp, P.J. (2017). Tools and strategies for scarless allele replacement in Drosophila using CRISPR/Cas9. Fly *11*, 53–64.

Latchman, D.S. (2003). Eukaryotic Transcription Factors (Academic Press).

Le Noir, S., Boyer, F., Lecardeur, S., Brousse, M., Oruc, Z., Cook-Moreau, J., Denizot, Y., and Cogné, M. (2017). Functional anatomy of the immunoglobulin heavy chain 3 ' super-enhancer needs not only core enhancer elements but also their unique DNA context. Nucleic Acids Res. *45*, 5829–5837.

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. Curr. Biol. *20*, R754–R763.

Li, L.M., and Arnosti, D.N. (2011). Long- and short-range transcriptional repressors induce distinct chromatin states on repressed genes. Curr. Biol. *21*, 406–412.

Li, X.-., and Eisen, M.B. Mutation of sequences flanking and separating transcription factor binding sites in a Drosophila enhancer significantly alter its output.

Li, X.-Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C.L., et al. (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol. *6*, e27.

Lim, B., Heist, T., Levine, M., and Fukaya, T. (2018a). Visualization of Transvection in Living Drosophila Embryos. Mol. Cell *70*, 287–296.e6.

Lim, B., Fukaya, T., Heist, T., and Levine, M. (2018b). Temporal dynamics of pair-rule stripes in living Drosophila embryos. Proc. Natl. Acad. Sci. U. S. A.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell *167*, 1170–1187.

Lovén, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320–334.

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

Ludwig, M.Z., Manu, Kittler, R., White, K.P., and Kreitman, M. (2011). Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. PLoS Genet. *7*, e1002364.

Luengo Hendriks, C.L., Keränen, S.V.E., Fowlkes, C.C., Simirenko, L., Weber, G.H., DePace, A.H., Henriquez, C., Kaszuba, D.W., Hamann, B., Eisen, M.B., et al. (2006). Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution I: data acquisition pipeline. Genome Biol. *7*, R123.

Maekawa, T., Imamoto, F., Merlino, G.T., Pastan, I., and Ishii, S. (1989). Cooperative function of two separate enhancers of the human epidermal growth factor receptor proto-oncogene. J. Biol. Chem. *264*, 5488–5494.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet. *7*, 29–59.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.

Maurano, M.T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., and Stamatoyannopoulos, J.A. (2015). Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nat. Genet.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr, Kinney, J.B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. *30*, 271–277.

Mir, M., Reimer, A., Haines, J.E., Li, X.-Y., Stadler, M., Garcia, H., Eisen, M.B., and Darzacq, X. (2017). Dense Bicoid hubs accentuate binding along the morphogen gradient. Genes Dev. *31*, 1784–1794.

Moorthy, S.D., Davidson, S., Shchuka, V.M., Singh, G., Malek-Gilani, N., Langroudi, L., Martchenko, A., So, V., Macpherson, N.N., and Mitchell, J.A. (2017). Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. Genome Res. *27*, 246–258.

Moreau, P., Hen, R., Wasylyk, B., Everett, R., Gaub, M.P., and Chambon, P. (1981). The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. Nucleic Acids Res. *9*, 6047–6068.

Nechaev, S., and Adelman, K. (2011). Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. Biochim. Biophys. Acta *1809*, 34–45.

Park, M., Patel, N., Keung, A.J., and Khalil, A.S. (2018). Engineering Epigenetic Regulation Using Synthetic Read-Write Modules. Cell.

Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A., and Bejerano, G. (2013). Enhancers: five essential questions. Nat. Rev. Genet. *14*, 288–295.

Perry, M.W., Boettiger, A.N., Bothma, J.P., and Levine, M. (2010). Shadow enhancers foster robustness of Drosophila gastrulation. Curr. Biol. *20*, 1562–1567.

Perry, M.W., Boettiger, A.N., and Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. Proc. Natl. Acad. Sci. U. S. A. *108*, 13570–13575.

Pirrotta, V., Chan, C.S., McCabe, D., and Qian, S. (1995). Distinct parasegmental and imaginal enhancers and the establishment of the expression pattern of the Ubx gene. Genetics *141*, 1439–1450.

Pott, S., and Lieb, J.D. (2015). What are super-enhancers? Nat. Genet. *47*, 8–12.

Prazak, L., Fujioka, M., and Gergen, J.P. (2010). Non-additive interactions involving two distinct elements mediate sloppy-paired regulation by pair-rule transcription factors. Dev. Biol. *344*, 1048–1059.

Proudhon, C., Snetkova, V., Raviram, R., Lobry, C., Badri, S., Jiang, T., Hao, B., Trimarchi, T., Kluger, Y., Aifantis, I., et al. (2016). Active and Inactive Enhancers Cooperate to Exert Localized and Long-Range Control of Gene Regulation. Cell Rep. *15*, 2159–2169.

Qian, S., Capovilla, M., and Pirrotta, V. (1991). The bx region enhancer, a distant cis-control element of the Drosophila Ubx gene and its regulation by hunchback and other segmentation genes. EMBO J.

Qian, S., Capovilla, M., and Pirrotta, V. (1993). Molecular mechanisms of pattern formation by the BRE enhancer of the Ubx gene. EMBO J. *12*, 3865–3877.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science.

Sackerson, C., Fujioka, M., and Goto, T. (1999). The even-skipped locus is contained in a 16-kb chromatin domain. Dev. Biol. *211*, 39–52.

Scholes, C., DePace, A.H., and Sánchez, Á. (2016). Combinatorial Gene Regulation through Kinetic Control of the Transcription Cycle. Cell Syst.

Scholes, C., Biette, K.M., Harden, T.T., and DePace, A.H. (2019). Signal Integration by Shadow Enhancers and Enhancer Duplications Varies across the Drosophila Embryo. Cell Reports *26*, 2407–2418.e5.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.

Siersbæk, R., Rabiee, A., Nielsen, R., Sidoli, S., Traynor, S., Loft, A., Poulsen, L.L.C., Rogowska-Wrzesinska, A., Jensen, O.N., and Mandrup, S. (2014). Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. Cell Rep. *7*, 1443–1455.

Small, S., and Levine, M. (1991). The initiation of pair-rule stripes in the Drosophila blastoderm. Curr. Opin. Genet. Dev. *1*, 255–260.

Small, S., Kraut, R., Hoey, T., Warrior, R., and Levine, M. (1991). Transcriptional regulation of a pair-rule stripe in Drosophila. Genes Dev. *5*, 827–839.

Small, S., Blair, A., and Levine, M. (1992). Regulation of even-skipped stripe 2 in the Drosophila embryo. EMBO J. *11*, 4047–4057.

Small, S., Arnosti, D.N., and Levine, M. (1993). Spacing ensures autonomous expression of different stripe enhancers in the even-skipped promoter. Development *119*, 762–772.

Small, S., Blair, A., and Levine, M. (1996). Regulation of Two Pair-Rule Stripes by a Single Enhancer in theDrosophilaEmbryo. Dev. Biol. *175*, 314–324.

Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet. *45*, 1021–1028.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Staller, M.V., Vincent, B.J., Bragdon, M.D.J., Lydiard-Martin, T., Wunderlich, Z., Estrada, J., and DePace, A.H. (2015). Shadow enhancers enable Hunchback bifunctionality in the Drosophila embryo. Proc. Natl. Acad. Sci. U. S. A. *112*, 785–790.

Stine, Z.E., McGaughey, D.M., Bessling, S.L., Li, S., and McCallion, A.S. (2011). Steroid hormone modulation of RET through two estrogen responsive enhancers in breast cancer. Hum. Mol. Genet. *20*, 3746–3756.

Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature.

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

Struhl, K. (1991). Mechanisms for diversity in gene expression patterns. Neuron *7*, 177–181.

Suzuki, H.I., Young, R.A., and Sharp, P.A. (2017). Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis. Cell *168*, 1000–1014.e15.

Thomas, S., Li, X.-Y., Sabo, P.J., Sandstrom, R., Thurman, R.E., Canfield, T.K., Giste, E., Fisher, W., Hammonds, A., Celniker, S.E., et al. (2011). Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. Genome Biol. *12*, R43.

Tsai, A., Muthusamy, A.K., Alves, M.R., Lavis, L.D., Singer, R.H., Stern, D.L., and Crocker, J. (2017). Nuclear microenvironments modulate transcription from low-affinity enhancers. Elife *6*.

Van Bortle, K., and Corces, V.G. (2013). The role of chromatin insulators in nuclear architecture and genome function. Curr. Opin. Genet. Dev. *23*, 212–218.

Van Bortle, K., Nichols, M.H., Li, L., Ong, C.-T., Takenaka, N., Qin, Z.S., and Corces, V.G. (2014). Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biol. *15*, R82.

Wasylyk, B., Wasylyk, C., and Chambon, P. (1984). Short and long range activation by the SV40 enhancer. Nucleic Acids Res. *12*, 5589–5608.

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nat. Rev. Genet. *14*, 125–138.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Wilson, C., Bellen, H.J., and Gehring, W.J. (1990). Position effects on eukaryotic gene expression. Annu. Rev. Cell Biol. *6*, 679–714.

Wittkopp, P.J., and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. *13*, 59–69.

Wunderlich, Z., Bragdon, M.D.J., Vincent, B.J., White, J.A., Estrada, J., and DePace, A.H. (2015). Krüppel Expression Levels Are Maintained through Compensatory Evolution of Shadow Enhancers. Cell Rep. *12*, 1740–1747.

Yuh, C.H., and Davidson, E.H. (1996). Modular cis-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. Proceedings of the.

Yuh, C.-H., Bolouri, H., and Davidson, E.H. (1998). Genomic Cis-Regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene. Science *279*, 1896–1902.

Chapter 2: Quantitative Measurement and Thermodynamic Modeling of Fused Enhancers Support a Two-Tiered Mechanism for Interpreting Regulatory DNA

Md. Abul Hassan Samee*, Tara Lydiard-Martin*, Kelly M. Biette*, Ben J. Vincent, Meghan D. Bragdon, Kelly B. Eckenrode, Zeba Wunderlich, Javier Estrada, Saurabh Sinha, and Angela H. DePace

This chapter was published in Samee et al., 2017. Quantitative Measurement and Thermodynamic Modeling of Fused Enhancers Support a Two-Tiered Mechanism for Interpreting Regulatory DNA. Cell Reports, 21, pp.236-245. *denotes co-first authorship.

Author Contributions for Samee, Lydiard-Martin, Biette et al., 2017:
T.L.M. and A.H.D. designed research; T.L-M., K.M.B., M.A.H.S., B.J.V., M.D.B., K.B.E performed research; T.L-M., M.A.H.S., K.M.B., analyzed the data with input from A.H.D. and S.S.; Z.W. and J.E. processed imaging data; M.A.H.S., K.M.B, S.S., and A.H.D. wrote the paper.

Abstract

Computational models of enhancer function generally assume that transcription factors (TFs) exert their regulatory effects independently, modeling an enhancer as a "bag of sites." These models fail on endogenous loci that harbor multiple enhancers, and a "two-tier" model appears better suited: in each enhancer TFs work independently, and the total expression is a weighted sum of their expression readouts. Here, we test these two opposing views on how cis-regulatory information is integrated. We fused two Drosophila blastoderm enhancers, measured their readouts, and applied the above two models to these data. The two-tier mechanism better fits these readouts, suggesting that these fused enhancers comprise multiple independent modules, despite having sequence characteristics typical of single enhancers. We show that short-range TF-TF interactions are not sufficient to designate such modules, suggesting unknown underlying mechanisms. Our results underscore that mechanisms of how modules are defined and how their outputs are combined remain to be elucidated.

Introduction

An important goal in regulatory genomics is to understand how gene transcription is regulated by enhancers, which are classically defined as "discrete DNA elements that contain specific sequence motifs with which DNA-binding proteins interact and transmit molecular signals to genes" (Blackwood and Kadonaga, 1998). Experiments can identify the transcription factors (TFs) that bind to an enhancer and provide insight on how TFs influence the transcriptional readout of an enhancer. However, we need computational models to unify this experimental knowledge into a mechanistic framework to predict the transcriptional readouts of arbitrary sequences. Such models have been under development in multiple prokaryotic and eukaryotic systems for decades (Ay and Arnosti, 2011). The majority of models have focused on readouts of single enhancers, although the field has recently begun to address how multiple enhancers control the expression of a single gene (Bothma et al., 2015; Perry et al., 2010; Staller et al., 2015). Below, we describe the critical distinctions between modeling sequence readouts at the level of a single enhancer, or at the level of a locus that contains multiple enhancers.

Enhancer-level models assume that enhancers function independent of their genomic context (Banerji et al., 1981; Bulger and Groudine, 2011; Khoury and Gruss, 1983), giving rise to their common designation as cis-regulatory "modules." This assumption greatly simplifies our task in building a model, since we can focus on TF binding to the enhancer, while ignoring TF binding to the rest of the genome. But even after this simplification, we must decide how TFs bound to sites within an enhancer exert their regulatory effects together. This too is a question of modularity (Dynan, 1989). The simplest assumption is that each bound TF acts independently and the enhancer behaves as a "bag of sites." This "bag of sites" assumption and its extension to incorporate local TF-TF interactions underlie a range of enhancer-level models, e.g., log-linear models (Bussemaker et al., 2001; Wunderlich et al., 2012), logistic regression models (Ilsley et al., 2013; Kazemian et al., 2010), activator-site occupancy models (Fakhouri et al., 2010; Zinzen et al., 2006), and more complex thermodynamics-based models (He et al., 2010; Segal et al., 2008).

To build *locus-level models* that predict the readout of an arbitrary locus, which can harbor multiple enhancers, we must also define modularity over longer DNA-length scales. In the simplest case, we could extend the "bag of sites" assumption to the entire locus, treating the entire locus as one long enhancer, but we have shown that this is not effective (Samee and Sinha, 2014). In Kim et al. (2013), local interactions between TFs were sufficient to model even-skipped (eve) expression in Drosophila from known regulatory sequences in the *eve* locus. In Samee and Sinha (2014), we found it necessary to delineate the binding sites in a locus as being partitioned into separate enhancers and assume that each enhancer acts as a module, independently from the rest of the locus. We modeled the readouts of 30 Drosophila developmental genes from their loci utilizing this model. We address the differences in these modeling approaches in the Discussion.

How do we decide when to apply these various assumptions about how TFs control gene expression? The "bag of sites" model works well on individual enhancers, but not when modeling an entire locus. Assuming that TFs locally interact may improve models, but it is not clear whether it can predict locus level regulation in all cases. Resolving this issue is critical not only for modeling entire loci but also for modeling enhancers themselves. Enhancer lengths vary between 100 and 1,000 bps, making it formally possible that a long enhancer sequence contains multiple independent sub-modules. If these sub-modules exist within a long enhancer, does it require that the enhancer be modeled using the same principles found to be effective to model an entire locus?

Here, we challenged thermodynamic models that operate at the enhancer level and the locus level by constructing a series of reporter constructs driven by enhancer fusions, and fitting these models to the resulting quantitative data. We concatenated two enhancers, resulting in a contiguous piece of regulatory DNA that resembles a single enhancer in length (1 Kbp long) and binding site distribution; we then asked if their readout follows the enhancer-level model, validating the bag of sites assumption, or the two-tier model, supporting the assumption that an enhancer might function as a collection of sub-modules.

We chose to work with eve37 and eve46, two enhancers of the developmental gene even-skipped (eve), a textbook example of enhancer modularity (Clyde et al., 2003; Fujioka et al., 1999; Struffi et al., 2011) (Figure 2.1A). These two enhancers contain binding sites for the same nine TFs (Figure S2.1), but primarily differ in the affinity of the binding sites for two repressor TFs, Hunchback (Hb) and Knirps (Kni) (Figure 2.1B). Because of this special property, which distinguishes our study from a previous study of fused enhancers (Small et al., 1993), it is not clear what expression pattern an eve37/eve46 fusion should direct. If an enhancer-level model is operational, one would expect two broad stripes as the readout. This is because the enhancer-level model interprets the collection of Hb sites in the fusion construct as a single repressive influence (of strength dependent on the number and affinities of those sites) on gene expression, and likewise for the Kni sites present in the construct. If a two-tier model is operational in the fusion constructs, with the two constituent enhancers as the two independent sub-modules, then one would expect four stripes (Figure 2.1C).

We contextualize the output of our reporter constructs with three different thermodynamics-based models, each of which makes a different assumption about the existence of sub-modules within an enhancer and about the mechanisms to delineate those sub-modules. We use the GEMSTAT model (He et al., 2010), which operates under the "bag of sites" assumption; the GEMSTAT-SRR model, which allows for the delineation of sub-modules based on distance-dependent interactions between TFs; and the two-tier model GEMSTAT-GL (GEMSTAT for gene locus) (Samee and Sinha, 2014), which delineates transcriptionally active segments in a gene locus and linearly combines the readouts of the active segments to model gene expression.

We found that GEMSTAT fits the readouts of the individual enhancers of eve with high accuracy, but fails completely to model the fusion constructs even though they are similar in length to a typical enhancer. Given the expressiveness of GEMSTAT and the flexibility allowed in estimating the model parameters (see the Materials and Methods), we interpreted this as

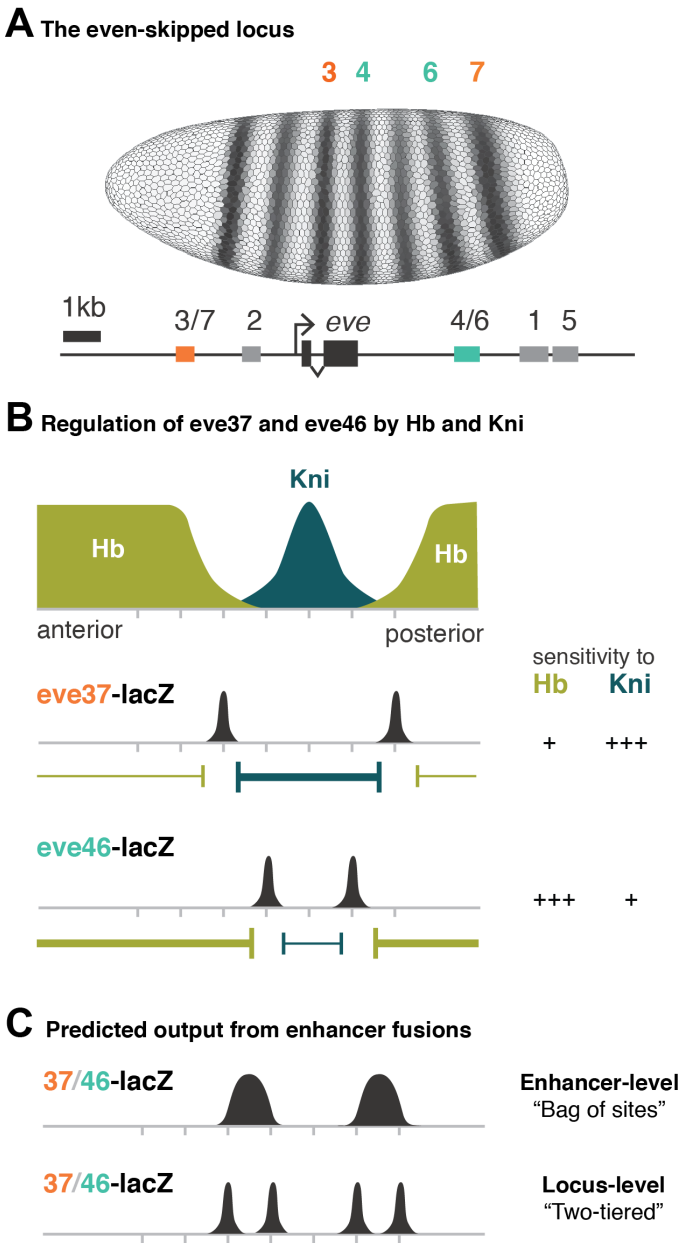


Figure 2.1: eve37 and eve46 Respond Differently to the Same Repressors. (A) Seven-striped expression of the *even-skipped* (*eve*) gene (top) and the genomic region containing the enhancers known to drive this pattern (bottom). Each enhancer is annotated with the stripe representing where it drives *eve* expression. (B) *eve* is regulated by the repressors Hb (green) and Kni (blue); these TFs have different spatial concentrations in the blastoderm embryo. The boundaries of eve37 and eve46 expression (black peaks) are set by differential sensitivities to Hb and Kni. (C) Information integration at multiple length scales predicts different outputs of enhancer fusions. If the enhancer-level model is operational and the fused enhancers are read as a "bag of sites," we expect two broad stripes (top). If a two-tier model is operational and the component enhancers remain autonomous, we expect four stripes (bottom). The mechanism under a two-tier model for maintaining inter-stripe gaps between stripes 3 and 4 and stripes 6 and 7 is unknown. See also Appendix A, Figure S2.1.

implying that current enhancer-level models are insufficient to explain the readouts of these enhancer-length sequences. Next, we attempted to fit the data using GEMSTAT-GL. GEMSTAT-GL uses all the thermodynamic parameters from GEMSTAT and uses additional parameters to model the weights of the independent regulatory segments mentioned above. To reduce the possibility that GEMSTAT-GL overfits the data by leveraging its additional parameters, we adopted a constrained strategy for parameter estimation (see the Materials and Methods). Despite such constraints, for each of the fusion constructs GEMSTAT-GL selected independent regulatory segments whose readouts when linearly aggregated fit the construct's readout with much higher accuracy, thus resolving the motivating question of this study. An enhancer-length regulatory sequence may indeed have within it sub-modules that are best thought of as functioning independently and ought to be modeled as such.

Our observation of multiple independent sub-modules within an enhancer length sequence suggests regulatory effects of DNA-bound TFs are somehow localized. One well-studied mechanism that results in such localization is that of "quenching" or "short-range repression" (SRR) (Kulkarni and Arnosti, 2005), whereby a repressor TF inhibits activator binding within a short range (100– 150 bps). We investigated this mechanism within our modeling framework and noted that it improves fits compared to the enhancer-level GEMSTAT model but fails to capture several salient features of data that GEMSTAT-GL was able to model. In summary, our results of modeling a unique dataset strongly argue that even enhancer-length sequences might work through independently functioning smaller segments present within the sequence.

Results

An Enhancer-Level Model Explains the eve Enhancer Readouts but Fails to Explain Readouts of Fusion Constructs

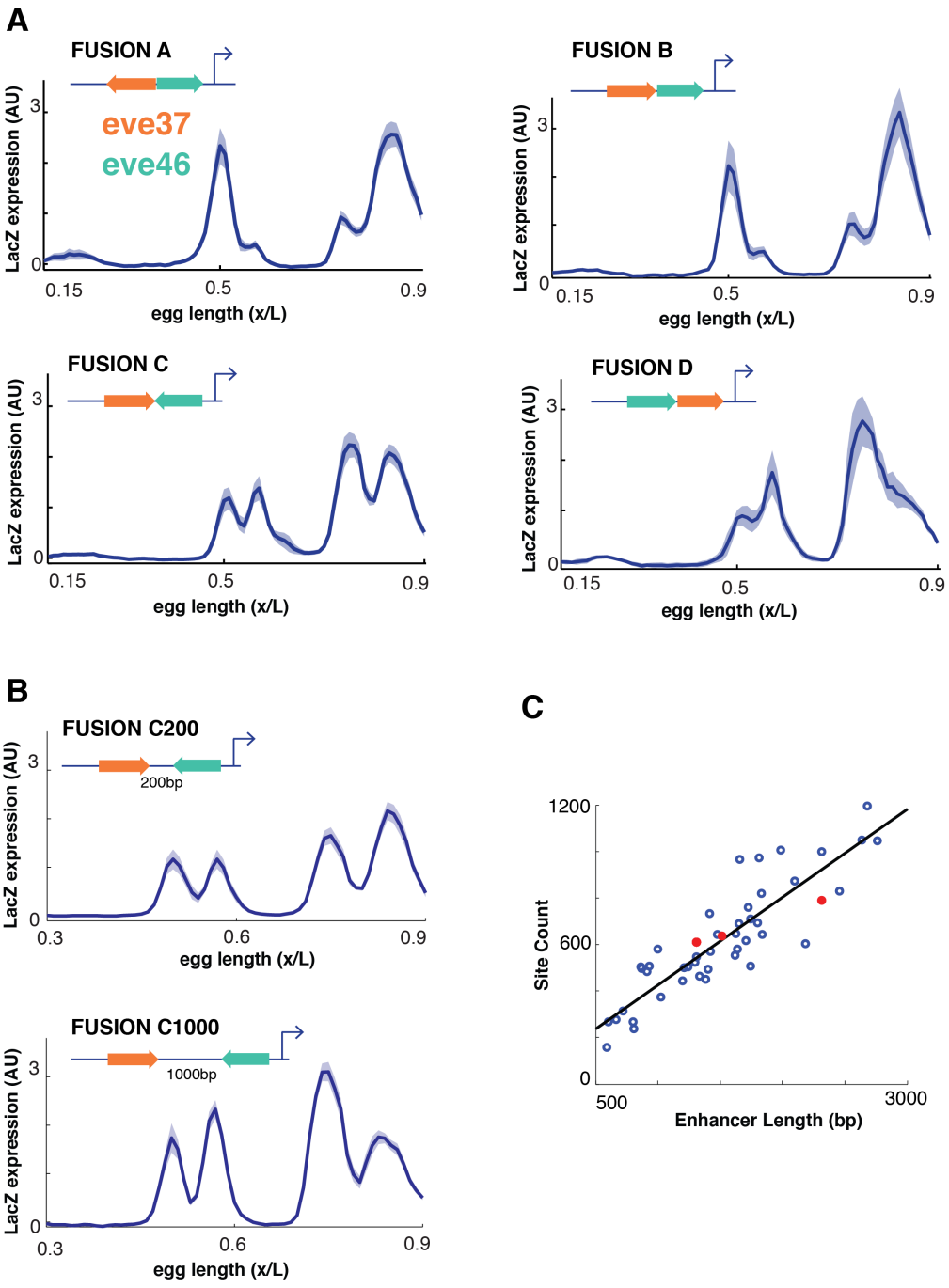


Figure 2.2: Expression Profiles for Fused Enhancers. (A) We measured LacZ expression normalized to a *hkb* co-stain driven by eve37/eve46 enhancer fusions in multiple configurations. The average expression for each transgenic line is displayed as a function of A/P position with the shadow representing SEM. (B) For one fusion configuration, "Fusion C," we placed 200-bp and 1,000-bp spacers between the enhancers and measured LacZ expression as above. (C) The length and binding site content of these synthetic constructs (orange) are comparable to other Drosophila developmental enhancers (blue). See also Appendix A, Figure S2.2.

We fused eve37 and eve46 together in multiple orientations driving expression of a LacZ reporter gene (Figure 2.2; see the Materials and Methods). All four possible junctions between the two enhancers are represented. The constructs drive patterned expression when placed next to a reporter gene (Figure 2.2A). Each enhancer fusion drove an expression pattern with more than two stripes; Fusion C drove a pattern with four distinct stripes. Because our goal was to test the limits of short-range interactions to define which transcription factor binding sites are included in a module, we also constructed two additional versions of Fusion C; one with a 200-bp spacer ("Fusion C200") and one with a 1,000-bp spacer ("Fusion C1000"), each of which should place the enhancers beyond the range of short-range repressors (typically 150 bp). These fusion constructs with spacers drive highly similar patterns to Fusion C (Figure 2.2B). All of these sequences are 1.5–2.5 Kbp in length, comparable to the average length of other developmental enhancers in Drosophila (Figure 2.2C).

We then tested whether the "bag of sites" assumption underlying enhancer-level models holds for six fusion constructs, by applying GEMSTAT to these data (Figure 2.3). The enhancer-level GEMSTAT model can accurately model the readouts of 40 A/P enhancers of the Drosophila embryo, using a common parameter setting (He et al., 2010); it can also accurately model all five A/P enhancers of eve (He et al., 2010; Samee and Sinha, 2014) (Figure 2.3A).

In our first strategy, we fit GEMSTAT on the eve3/7, eve5, and eve4/6 enhancers and utilized the trained parameters as the initial parameter estimates to optimize GEMSTAT individually for each of the four fusion constructs. Since all sequences being modeled here were eve enhancers and derivatives thereof, we followed a constrained optimization strategy in both steps of this exercise (see the Materials and Methods). Although the endogenous enhancer readouts were fit reasonably well, for all fusion constructs GEMSTAT predicted only two broad domains of expression spanning *eve* stripes 3–4 and 6–7, thus failing to capture any salient feature of those readouts (Figure 2.3B; Table S2.2). To confirm that this failure of enhancer-level assumptions was not an artifact of the parameter optimization process, we next fit GEMSTAT individually for each fusion construct following an unconstrained strategy (see the

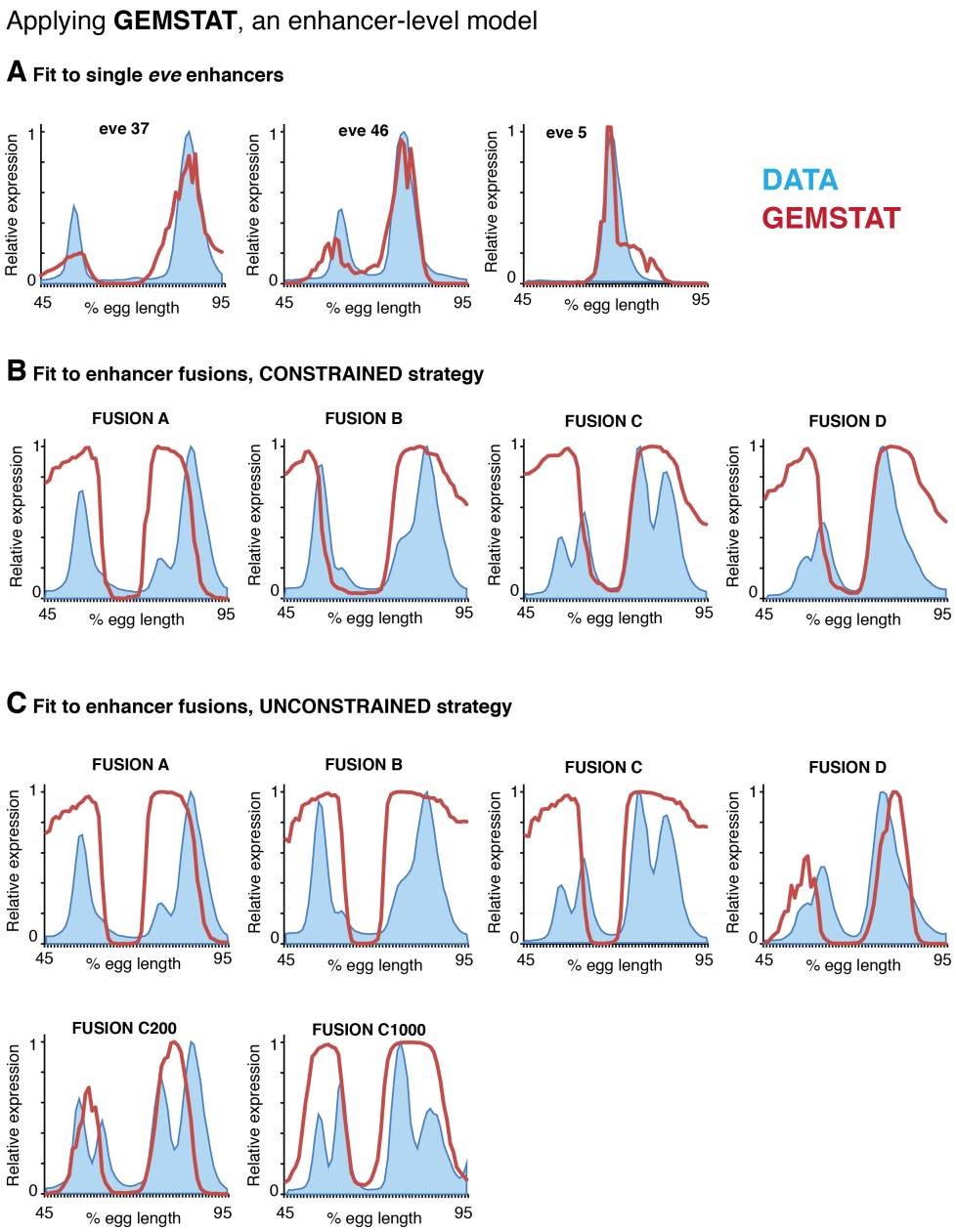


Figure 2.3 : The Enhancer-Level GEMSTAT Fails to Explain Readouts of Fused Enhancers. (A) We applied GEMSTAT to three single *eve* enhancers. In all panels, experimentally measured expression profiles are shown in blue and GEMSTAT output in red. For this approach, we simultaneously fitted the model on 30 developmental enhancers, excluding all enhancers of *eve*. We then fit the model on three enhancers of *eve*, starting from (P) initial parameters and letting (Q) denote the optimized parameters following a constrained strategy (see the Materials and Methods). (B) We used GEMSTAT to fit a separate model for each of the fused enhancers, starting from (Q) as the initial estimate and using a constrained fitting strategy. (C) We also used GEMSTAT to fit a separate model for six fused enhancers using an unconstrained fitting strategy. See also Appendix A, Tables S2.1 and S2.2.

Materials and Methods). Although this setup provided considerable flexibility in parameter fitting (in comparison to a typical enhancer modeling setup it is flexible to an extent that might lead to data overfitting), GEMSTAT failed again to model any construct's readout with satisfactory accuracy (Figure 2.3C; Table S2.2). This failure to model the readout of any fusion construct, despite an unconstrained fitting procedure and rigorous search over the parameter space, strongly suggests that assumptions of an enhancer-level model do not hold for these constructs.

A Two-Tier Model Can Explain Readouts of the Fused Enhancers by Identifying Independent Regulatory Segments within Each Construct

The failure of the enhancer-level GEMSTAT model, as noted above, is similar to our previous failure to model the expression pattern of a gene from its intergenic locus using GEMSTAT (Samee and Sinha, 2014). For genes expressed in multiple stripes (e.g., eve in Figure 2.1A), the GEMSTAT model in that study predicted a broad expression pattern spanning all stripes. This led us to develop the GEMSTAT-GL model, which assumes that a gene locus comprises several independent segments (i.e., putative enhancers) and models the readout of the locus as a weighted sum of the readouts of the segments. Readouts of individual segments are modeled using GEMSTAT, and GEMSTAT-GL has the ability to identify the segments in the course of modeling the data.

Noting the similar nature of these failures, we asked if GEMSTAT-GL could model the six fusion constructs. Note that applying the GEMSTAT-GL model to these data amounts to assuming that a fusion construct may function more like a gene locus, with independently functioning sub-modules within, even though the construct's length is typical of a single enhancer. By also testing constructs with spacers, we can address the length scale beyond which a sequence starts functioning as a collection of sub-modules. We optimized the GEMSTAT-GL parameters under a constrained setting to avoid overfitting (see the Materials and Methods),

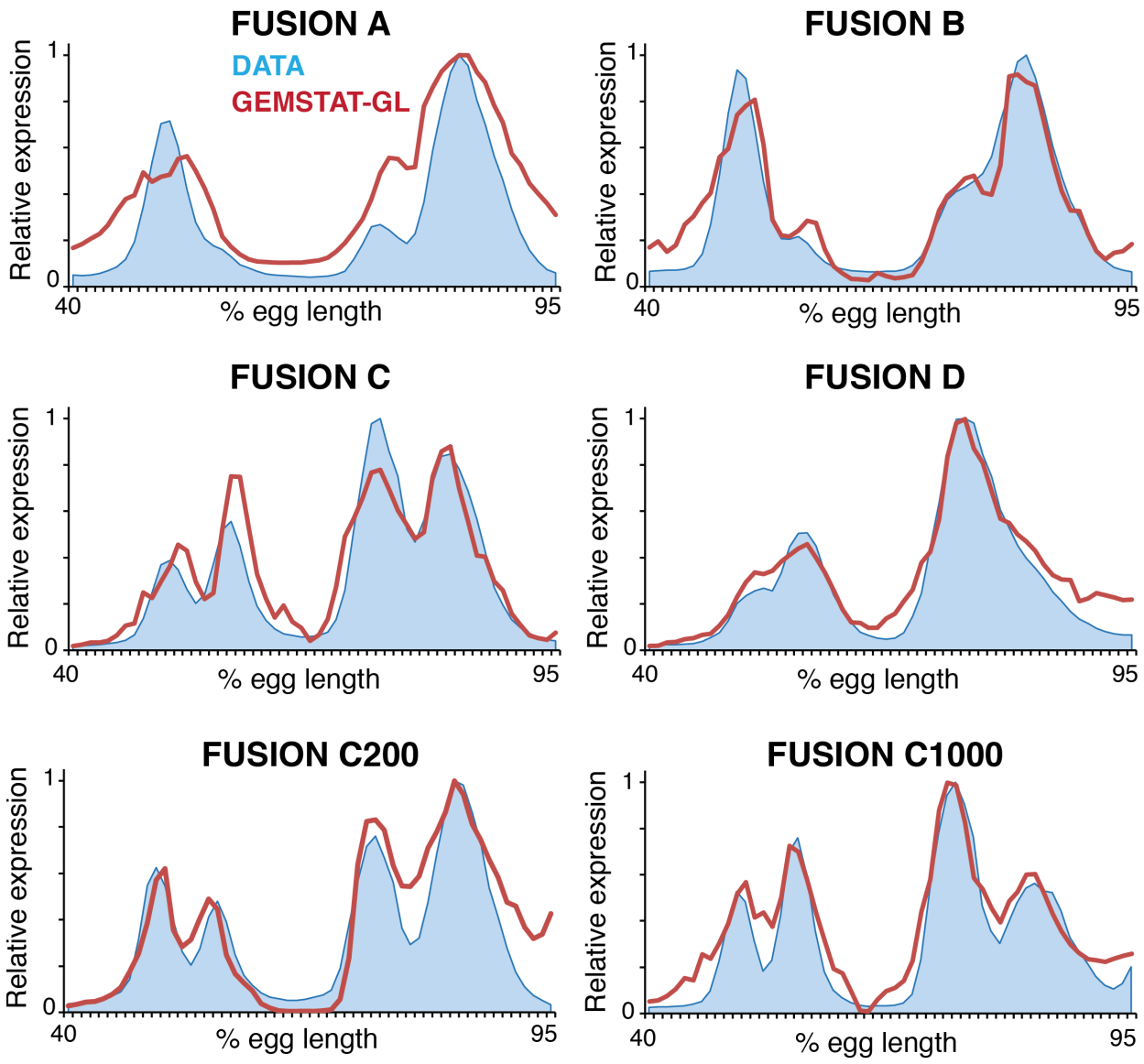


Figure 2.4 : The Two-Tier GEMSTAT-GL Model Captures the Readouts of Enhancer Fusions. We applied GEMSTAT-GL (for gene locus) to the six fusion constructs. Experimentally determined expression profiles (blue) and model predictions (red) are shown as a function of egg length along the A/P axis. See also Appendix A, Tables S2.1, S2.2, and S2.3.

and obtained improved fits to the readouts of all constructs (Figure 2.4). GEMSTAT-GL was successful in capturing both inter-stripe gaps observed in constructs "Fusion C," "Fusion C200," and "Fusion C1000" and the broad domains of overlapping stripe patterns observed in constructs "Fusion A," "Fusion B," and "Fusion D." Each fusion construct thus appears to comprise two independent sub-modules: Hb and Kni binding sites mediate repression only within the corresponding sub-module. The efficacy of this model supports the idea that there are two levels of information integration in regulatory DNA—one where modules are defined and another where the information from modules is integrated—even when the regulatory DNA in question resembles a single enhancer in its length and binding site content.

Interestingly, GEMSTAT-GL always selected one contributing regulatory segment from each of the two constituent enhancers in a fusion construct (Table S2.3). That it did not select multiple segments from any of the two constituent enhancers is consistent with the literature: prior experimental attempts to identify smaller functional segments within these constituent enhancers showed loss of function (Fujioka et al., 1999). The GEMSTAT-GL model never selected a segment that straddles across the boundary of the constituent enhancers, even though its optimization procedure had that flexibility. This raises the intriguing possibility of yet unidentified mechanisms that confine the submodules within the individual enhancers although they were fused without any spacer.

A Model in which Repressors Act Only over Short Ranges Can Explain Readouts of Fusion Constructs with Spacer Sequences but Fails for Other Constructs

The above exercise utilizing GEMSTAT-GL points to the existence of at least two independent regulatory segments within each construct. However, it is not clear how these segments are determined. Previous studies have suggested that some repressor TFs, by inhibiting activator binding within short ranges around their binding sites (typically 100-150 bps), may give rise to sub-modules within enhancer-length sequences (Gray et al., 1994; Gray and Levine, 1996; Kim et al., 2013; Small et al., 1993). We therefore considered this short-range

repression (SRR) mechanism as a potential explanation for the existence of independent segments within our fusion constructs. In a synthetic construct, two enhancers will function independently under the SRR mechanism when they are sufficiently far apart so that repressors bound to one enhancer do not interact with activators in the other enhancer. In the fusions, however, the constituent enhancers may interfere with each other's regulatory effect, as has been shown in Kim et al. (2013) and Small et al. (1993). In our previous work with the two-tier model we found the SRR mechanism to be insufficient to model a gene's expression pattern as the readout of its locus. A different implementation of the SRR model was used in Kim et al. (2013) to explain the seven-striped pattern of *eve* from its locus, but the authors had chosen to model only those pieces of the *eve* locus that have known enhancer activity for *eve*. Hence, it was not clear to us *a priori* if the SRR mechanism would prove to be a satisfactory explanation for the independent action of the two enhancers in our fusions.

We therefore asked whether GEMSTAT implemented with SRR could explain the following two features of our dataset. First, the four fusion constructs without spacers produce more than two peaks, indicative of underlying sub-modules, but the stripes are not always entirely distinct. This may be due to local interference between TFs bound at the junction of the enhancers. Second, fusion construct C does produce four stripes, and this is maintained with spacers, i.e., Fusion C200 and Fusion C1000. To fit these data, we allowed GEMSTAT-SRR the same flexibilities in unconstrained training as were allowed to the baseline GEMSTAT model (see the Materials and Methods). We found that for the constructs with spacers, especially the construct with 1,000-bp spacer sequence, the model's accuracy was comparable to that of the two-tier GEMSTAT-GL model (Figure 2.5). However, in cases where the two constituent enhancers were fused adjacent to each other, GEMSTAT-SRR's fits were not as accurate as GEMSTAT-GL. Fusion C is a particular example where GEMSTAT-SRR significantly fell short of modeling the data. For this construct, GEMSTAT-SRR failed to model both the expression at the domain of stripe 3 and the gap between stripes 3 and 4. Overall, GEMSTAT-SRR improved over the baseline GEMSTAT model for these constructs, yet the fits were not so accurate that we

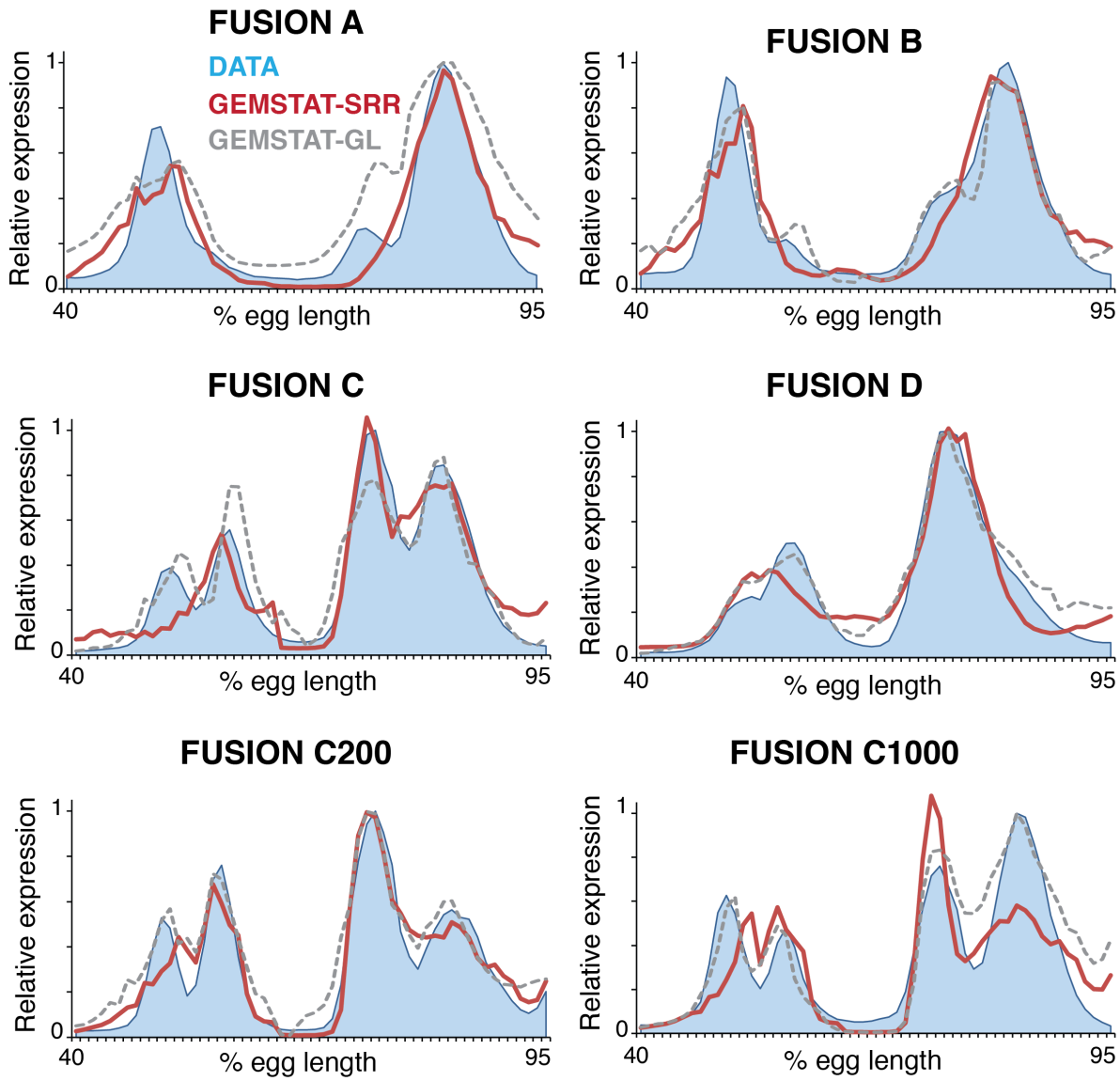


Figure 2.5 : GEMSTAT with Short-Range Repression (GEMSTAT-SRR) Less Accurately Predicts Salient Features of Fused Enhancer Expression. We applied GEMSTAT-SRR to the six fusion constructs and compared the predictions to those made by GEMSTAT-GL. Experimentally determined expression profiles are in blue, GEMSTATSRR predictions are in red, and GEMSTAT-GL predictions are represented by gray dashes. See also Appendix A, Tables S2.1 and S2.2.

could attribute the disrupted stripe formation to SRR-driven interference of TF activities at the junctions.

These observations imply that the SRR model, at least as implemented in GEMSTAT, is not mechanistically rich enough to explain the expression patterns driven by these fusion constructs. We therefore favor the GEMSTAT-GL model's findings as the explanation to the readouts of our fusion constructs: linear aggregation of individual enhancer output gives rise to the observed expression patterns.

Discussion

We have identified a gap in our current assumptions about enhancer readout by applying computational analysis to quantitative measurements of output from a set of fused enhancers. In current models of enhancer readout (enhancer-level model), TFs bound to an enhancer are assumed to act as a "bag of sites," exerting their effects independently. The effects of spacing between sites and their relative arrangement are also sometimes included, such that the TFs act somewhat coordinately within an enhancer. Although two decades of quantitative studies have championed this model, we found it failed for a set of fused enhancers. To explain the readouts of these fusions, we instead needed to assume that an enhancer might comprise multiple segments where information from bound TFs is interpreted independently within each segment. This is analogous to the two-tier model we proposed in previous work (Samee and Sinha, 2014), where the output of entire loci is comprised of a weighted sum of output from its component enhancers. Here, we show that this same two-tier mechanism captures the output of enhancer-length sequences as well (Figure 2.6).

It is important to consolidate this idea with previous reports on computational modeling of enhancer sequences where short-range effects were assumed for repressors (Kim et al., 2013; Small et al., 1993). We must first clarify where the short-range repression model stands in relation to the enhancer-level and the two-tier models we use here. It has been proposed that short-range repression could effectively partition bound TFs into multiple enhancers (Gray and

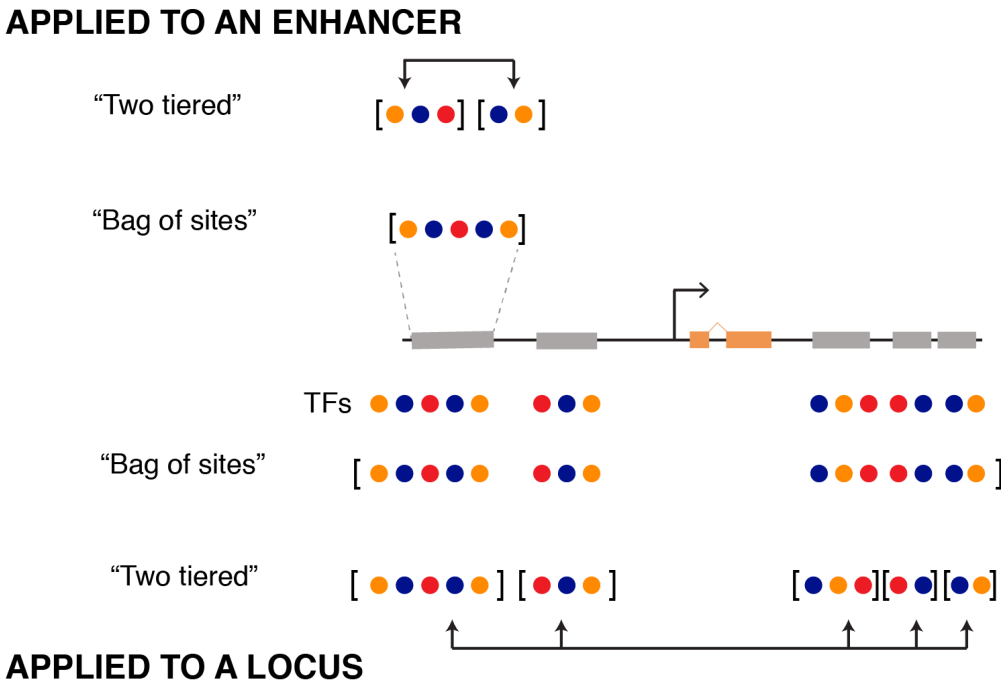


Figure 2.6 : A "Two-Tiered" Mechanism May Define and Integrate Submodules of Regulatory Sequence at the Level of Single Enhancers and Entire Loci. (Top) Enhancer sequences contain binding sites for different TFs that function by activating or repressing their target gene. Enhancer-level models capture each TF input independently, representing the enhancer as one "bag of sites." Two-tier models, such as GEMSTAT-GL, also can be applied to enhancer length sequences by first separating TF inputs into multiple regulatory segments and then integrating their weighted output to predict expression. (Bottom) Two-tier and enhancer-level models can be applied to an entire locus. Enhancer-level models consider TF binding across the locus as a large "bag of sites," without considering individual enhancers as separate regulatory entities. We can also apply the two-tiered model to a gene locus. This approach first subdivides the regulatory sequence around a gene into smaller modules and then integrates the regulatory information from each module to predict expression.

Levine, 1996; Gray et al., 1994; Kim et al., 2013; Small et al., 1993); thus it could be considered in the middle of a hierarchy of models between our enhancer-level and two-tier models. We showed that our enhancer-level model fails to explain the fusions, but that our two-tier model is successful. Therefore it is a relevant question whether short-range repression can account for the effective partitioning of the fusions into multiple segments, each of which adheres to enhancer-level rules. A related study found that short-range repression was sufficient to explain

the output of part of a locus, implying that it may serve as a mechanism to partition enhancers from longer sequences (Kim et al., 2013). However, we found here that short-range repression was inadequate to explain the partitioning of some of the fusion constructs in our dataset.

For practical considerations, it is also important to discuss when one class of modeling assumptions, e.g., the "bag of sites" in GEMSTAT, the "two-tier" approach of GEMSTAT-GL, or the short-range repression assumption of GEMSTAT-SRR, should be considered more suitable than the other two classes. As with any other type of statistical modeling, the choice depends on the type of questions being pursued; and the answers should be analyzed with a clear understanding of the limitations of the chosen model. We note that, GEMSTAT-GL is more flexible than GEMSTAT and GEMSTAT-SRR in terms of selecting/redefining active modules, and was successful in this study in modeling locus-level gene expression where both GEMSTAT and SRR had failed. However, GEMSTAT-GL cannot identify any mechanistic event that could have demarcated the modules it selected. The SRR model is more mechanistically grounded than both GEMSTAT and GEMSTAT-GL, but implements only one possible mechanism of distance-dependent repressor function. We speculate that it may be due to this limited view of repressor function that SRR underperformed compared to GEMSTAT in our previous study on 40 developmental enhancers (He et al., 2010). In our experience, GEMSTAT is often good as a first approximation. The same has been shown in many publications using different realizations of the bag of sites model (Bussemaker et al., 2001; Ilsley et al., 2013; Kazemian et al., 2010; Segal et al., 2008; Wunderlich et al., 2012). However, as we obtain more data on enhancer activity and improve our understanding of TF-TF and TF-DNA interactions, we believe mechanistically grounded models like SRR will eventually improve and may become more appropriate than the two other models for most datasets.

Our study suggests mechanisms other than short-range repression delineate active modules within enhancer-length sequences. What mechanisms might delineate such segments? We do not know of any direct interaction between the relevant TFs that can explain such delineations, despite nearly two decades of study of these proteins. DNA shape itself may

impose some modularity on enhancer sequences at short length scales (Peng and Sinha, 2016). Another possibility is that insulator and architectural proteins play a role. Weak binding sites for such proteins are prevalent throughout the *eve* enhancers; our fusion constructs do not create any new sites for Trl/GAF, su(Hw), CTCF, Cp190, or BEAF-32 at the junctions (Figure S2.2). Finally, indirect interaction between TFs could occur through binding to large co-factor complexes, such as mediator or CBP (Merika et al., 1998) or through chromatin state (Voss et al., 2011).

Our results are relevant for interpreting the increasing volume of high-throughput functional genomic data annotating active regulatory DNA in the genome; in particular it will be important to assess whether contiguous sequences act as a single enhancer, or as multiple functional sub-modules, like our fusions. In general, we emphasize that future computational modeling of enhancer readouts should not overlook the examples where the models perform poorly, and rather should systematically consider the existence of unknown mechanisms of enhancer function, such as the two-tiered effects we reveal here.

Materials and Methods

Transgenic Fly Lines

We used RedFly to identify coordinates of the D. melanogaster even-skipped stripe 3/7 and stripe 4/6 enhancers (Gallo et al., 2011). The eve_stripe_3+7 element is 510 bp (release 5 coordinates 2R: 5863006-5863516) (Small et al., 1996), whereas the eve_stripe4_6 element is 800 bp (release 5 coordinates 2R:5871404-5872203) (Fujioka et al., 1999). Note that the stripe 4/6 enhancer coordinates from REDfly contain an extra 208 bp on the 30 end compared to the construct tested in Fujioka et al. (1999). Enhancers were PCR amplified from genomic DNA from w118 Drosophila melanogaster flies, sequence verified, and inserted into the multiple cloning site of the pBOY vector (Hare et al., 2008) using isothermal assembly (Gibson et al., 2009), which leaves scarless junctions. A list of all enhancers and spacer sequences is given in Supplemental Materials and Methods. pBOY contains an *eve* core promoter 20 bp downstream of the multiple cloning site that drives an eve/lacZ fusion transcript. The vector also contains an attB site for phiC31 site-specific integration (Fish et al., 2007) and the mini-white gene for selection of transformants. Each plasmid was injected into attP2 flies (Markstein et al., 2008) by BestGene and homozygosed using the mini-white eye color marker.

Embryo Collection and In Situ Hybridization

Embryo collection and whole-mount in situ hybridization was performed as previously described (Luengo Hendriks et al., 2006). Briefly, 0–4 hr embryos (25C) were collected, dechorionated in 50% bleach, fixed in a 1:4 mixture of 10% formaldehyde to heptane, and devitellinized in heptane and methanol by shaking. Embryos were post-fixed in formaldehyde and a formaldehydebased hybridization buffer. Hybridizations were performed at 56C with three full-length cDNA probes: a DIG-labeled probe for fushi tarazu (ftz), a DNPlabeled lacZ, or *eve* probe and a DNP-labeled probe against huckebein (hkb). The probes were detected by successive antibody staining using anti-DIG-HRP (anti-DIG-POD; Roche, Basel, Switzerland) and anti-DNPHRP (Perkin-Elmer TSA Kit, Waltham, MA, USA) and labeled by reactions with

coumarin- and Cy3-tyramide (Perkin-Elmer). Embryos were treated with RNase and incubated with Sytox Green (Invitrogen, Carlsbad, CA, USA) to stain nuclei. Finally, embryos were dehydrated in ethanol and mounted in DePex (Electron Microscopy Sciences, Hatfield, PA, USA), using #1 coverslips to form a bridge to preserve 3D embryo morphology.

Imaging and Image Processing

Embryos were imaged and computationally segmented for further analysis (Fowlkes et al., 2008). A three-dimensional image stack of each embryo was acquired on a Zeiss LSM Z10 with a plan-apochromat 203 0.8 NA objective using two-photon microscopy. Embryos were binned into six time points of approximately 10 min windows using the extent of membrane invagination under phase-microscopy as a morphological marker. Time points correspond to 0%–3%, 4%–8%, 9%–25%, 26%–50%, 51%–75%, and 76%–100% membrane invagination along the side of the embryo that has progressed most. We imaged embryos from all age ranges and display data from the early blastoderm (4%–31% membrane invagination) when *hkb* normalization is used. Image files were processed into PointCloud representations containing the coordinates and fluorescence levels for each nucleus.

hkb Normalization

Normalization to a *hkb* co-stain was performed to test the variation in absolute levels of expression across reporter lines (Wunderlich et al., 2014). Embryos were stained with a mixture of lacZ- or eve-DNP and hkb-DNP probe. For each embryo, background was calculated as the mode of the fluorescence distribution. After subtracting background, mean *hkb* fluorescence was calculated as the geometric mean of the anterior and posterior expression domains. We noted that *eve* stripe 7 slightly overlaps with the posterior expression domain of *hkb*, and so chose to use the geometric mean of anterior and posterior rather than solely the posterior domain like in Wunderlich et al. (2014) to limit the impact of overlapping expression. The fluorescence in each nucleus was then divided by the mean *hkb* fluorescence to yield a normalized expression level.

Input Data for Models

Our sequence dataset, expression readouts, TF motifs, and TF concentrations are described in Supplemental Materials and Methods. Of note, since *eve* stripes 3 to 7 span 40%–95% of the anterior-posterior (A/P) axis of the Drosophila embryo, here we have considered sequence readouts and TF concentrations only within this spatial range. Below, we outline the previously published GEMSTAT and GEMSTAT-GL models (He et al., 2010; Samee and Sinha, 2014) and detail our parameter estimation strategy. We optimized all models utilizing weighted pattern generating potential (w-PGP) as the objective function. w-PGP was introduced in Samee and Sinha (2014), and we give an overview of w-PGP in Supplemental Materials and Methods.

GEMSTAT and GEMSTAT-GL

In GEMSTAT, three major components interact to regulate gene expression during transcription: (1) DNA sequence, (2) TF molecules, and (3) the basal transcriptional machinery (BTM). A TF molecule may bind the sequence at any binding site, with site-specific affinity computed from the TF's PWM (Supplemental Materials and Methods). The BTM may initiate transcription when bound at the core promoter of the gene. Interactions between bound TF molecules and the BTM determine the occupancy, i.e., probability of binding, of the BTM at the promoter. The level of gene expression is assumed proportional to BTM occupancy. The model fits two free parameters for each TF. The first parameter is a product of two unknowns: the TF's strength of DNA binding and a proportionality constant that scales the relative concentration of the TF to an absolute value. Note that our TF concentration data are derived from fluorescence intensity and thus are a relative measurement. The second parameter represents a bound TF's potency for activating or repressing BTM binding. The model can also include cooperative TF binding, which requires one free parameter for each pair of cooperatively binding TFs. In the default configuration of GEMSTAT, repressors exert their effects by acting over long-range to directly reduce BTM occupancy at the promoter. In an alternate configuration, repressors act

only over short ranges by interfering with activator occupancy at neighboring sites. We call the latter model GEMSTAT-SRR (short-range repression).

GEMSTAT-GL models a gene's expression as a weighted summation of expression driven by several enhancers within its locus, where each enhancer's output is predicted by GEMSTAT. From the intergenic locus of a gene, GEMSTAT-GL automatically selects a handful of segments that together generate the gene's expression. The number, lengths, and locations of contributing segments, as well as the weight of each segment's contribution are free parameters in the model. A constrained parameter estimation strategy, as described below, is adopted to avoid overfitting.

Parameter Estimation

In estimating our model parameters, we followed either a "constrained" or an "unconstrained" strategy. In the constrained strategy (adopted while modeling the *eve* enhancers and the fusion constructs collectively, and for fitting the GEMSTAT-GL model; see below) we first construct 1,000 models for a dataset comprising 30 A/P enhancers whose readouts fall within 40%–95% of the A/P axis. Of note, we had excluded all *eve* enhancers from this dataset. For this dataset, we first randomly sampled 1 million points (each denoting a different parameterization of the model) from the parameter space (Samee et al., 2015). Then we considered each of the top 1,000 models from the sampled collection, one at a time, as the initial parameterization of GEMSTAT and re-estimated parameters to optimize the model for the 30 enhancers. The optimized parameters of these models then become the initial parameters for optimizing GEMSTAT or GEMSTATGL on *eve* enhancers and/or the fusion constructs. For GEMSTAT we then construct 1,000 models starting from each of the 1000 models optimized on the 30 enhancers; since optimization of GEMSTAT-GL is timeconsuming, we optimize 10 GEMSTAT-GL models starting from the top 10 among the 1,000 models. In course of optimization, we constrain the value of each parameter to vary by at most 2-fold from its initial value. In GEMSTAT-GL, we additionally constrain each window's weight to vary by, at most, 2-fold from the weights of the other windows. As we have discussed in Samee and Sinha (2014),

this strategy can ensure that the final model trained on the enhancers of a single gene is largely consistent with a model that reflects other regulatory parts of the genome.

In the unconstrained strategy, we do not use the 30 enhancers mentioned above: we randomly sample 1 million points directly for the enhancer being modeled. Then, starting from the top 1,000 models from the sampled collection, one at a time, we optimize 1,000 models for the enhancer. We do not impose any constraints during this optimization.

References

Ay, A., and Arnosti, D.N. (2011). Mathematical modeling of gene expression: a guide for the perplexed biologist. Crit. Rev. Biochem. Mol. Biol. 46, 137–151.

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

Blackwood, E.M., and Kadonaga, J.T. (1998). Going the distance: a current view of enhancer action. Science 281, 60–63.

Bothma, J.P., Garcia, H.G., Ng, S., Perry, M.W., Gregor, T., and Levine, M. (2015). Enhancer additivity and non-additivity are determined by enhancer strength in the Drosophila embryo. eLife 4, 4.

Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. Cell 144, 327–339.

Bussemaker, H.J., Li, H., and Siggia, E.D. (2001). Regulatory element detection using correlation with expression. Nat. Genet. 27, 167–171.

Clyde, D.E., Corado, M.S.G., Wu, X., Pare´ , A., Papatsenko, D., and Small, S. (2003). A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. Nature 426, 849–853.

Dynan, W.S. (1989). Modularity in promoters and enhancers. Cell 58, 1–4.

Fakhouri, W.D., Ay, A., Sayal, R., Dresch, J., Dayringer, E., and Arnosti, D.N. (2010). Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo. Mol. Syst. Biol. 6, 341.

Fish, M.P., Groth, A.C., Calos, M.P., and Nusse, R. (2007). Creating transgenic Drosophila by microinjecting the site-specific phiC31 integrase mRNA and a transgene-containing donor plasmid. Nat. Protoc. 2, 2325–2331.

Fowlkes, C.C., Hendriks, C.L.L., Kera¨ nen, S.V.E., Weber, G.H., Rubel, O., € Huang, M.-Y., Chatoor, S., DePace, A.H., Simirenko, L., Henriquez, C., et al. (2008). A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. Cell 133, 364–374.

Fujioka, M., Emi-Sarker, Y., Yusibova, G.L., Goto, T., and Jaynes, J.B. (1999). 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.

Gallo, S.M., Gerrard, D.T., Miner, D., Simich, M., Des Soye, B., Bergman, C.M., and Halfon, M.S. (2011). REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. Nucleic Acids Res. 39, D118–D123.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods 6, 343–345.

Gray, S., and Levine, M. (1996). Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in Drosophila. Genes Dev. 10, 700–710.

Gray, S., Szymanski, P., and Levine, M. (1994). Short-range repression permits multiple enhancers to function autonomously within a complex promoter. Genes Dev. 8, 1829–1838.

Hare, E.E., Peterson, B.K., Iyer, V.N., Meier, R., and Eisen, M.B. (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet. 4, e1000106.

He, X., Samee, M.A., Blatti, C., and Sinha, S. (2010). Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput. Biol. 6, e1000935.

Ilsley, G.R., Fisher, J., Apweiler, R., De Pace, A.H., and Luscombe, N.M. (2013). Cellular resolution models for even skipped regulation in the entire Drosophila embryo. eLife 2, e00522.

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

Khoury, G., and Gruss, P. (1983). Enhancer elements. Cell 33, 313–314.

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

Kulkarni, M.M., and Arnosti, D.N. (2005). cis-regulatory logic of short-range transcriptional repression in Drosophila melanogaster. Mol. Cell. Biol. 25, 3411–3420.

Luengo Hendriks, C.L., Kera¨ nen, S.V.E., Fowlkes, C.C., Simirenko, L., Weber, G.H., DePace, A.H., Henriquez, C., Kaszuba, D.W., Hamann, B., Eisen, M.B., et al. (2006). Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution I: data acquisition pipeline. Genome Biol. 7, R123.

Markstein, M., Pitsouli, C., Villalta, C., Celniker, S.E., and Perrimon, N. (2008). Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. Nat. Genet. 40, 476–483.

Merika, M., Williams, A.J., Chen, G., Collins, T., and Thanos, D. (1998). Recruitment of CBP/ p300 by the IFN b enhanceosome is required for synergistic activation of transcription. Mol. Cell 1, 277–287.

Peng, P.-C., and Sinha, S. (2016). Quantitative modeling of gene expression using DNA shape features of binding sites. Nucleic Acids Res. 44, e120.

Perry, M.W., Boettiger, A.N., Bothma, J.P., and Levine, M. (2010). Shadow enhancers foster robustness of Drosophila gastrulation. Curr. Biol. 20, 1562– 1567.

Samee, M.A., and Sinha, S. (2014). Quantitative modeling of a gene's expression from its intergenic sequence. PLoS Comput. Biol. 10, e1003467.

Samee, M.A., Lim, B., Samper, N., Lu, H., Rushlow, C.A., Jime´ nez, G., Shvartsman, S.Y., and Sinha, S. (2015). A systematic ensemble approach to thermodynamic modeling of gene expression from sequence data. Cell Syst. 1, 396–407.

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

Small, S., Arnosti, D.N., and Levine, M. (1993). Spacing ensures autonomous expression of different stripe enhancers in the even-skipped promoter. Development 119, 762–772.

Small, S., Blair, A., and Levine, M. (1996). Regulation of two pair-rule stripes by a single enhancer in the Drosophila embryo. Dev. Biol. 175, 314–324.

Staller, M.V., Vincent, B.J., Bragdon, M.D., Lydiard-Martin, T., Wunderlich, Z., Estrada, J., and DePace, A.H. (2015). Shadow enhancers enable Hunchback bifunctionality in the Drosophila embryo. Proc. Natl. Acad. Sci. USA 112, 785–790.

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

Voss, T.C., Schiltz, R.L., Sung, M.-H., Yen, P.M., Stamatoyannopoulos, J.A., Biddie, S.C., Johnson, T.A., Miranda, T.B., John, S., and Hager, G.L. (2011). Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. Cell 146, 544–554.

Wunderlich, Z., Bragdon, M.D., Eckenrode, K.B., Lydiard-Martin, T., Pearl-Waserman, S., and DePace, A.H. (2012). Dissecting sources of quantitative gene expression pattern divergence between Drosophila species. Mol. Syst. Biol. 8, 604.

Wunderlich, Z., Bragdon, M.D., and DePace, A.H. (2014). Comparing mRNA levels using in situ hybridization of a target gene and co-stain. Methods 68, 233–241.

Zinzen, R.P., Senger, K., Levine, M., and Papatsenko, D. (2006). Computational models for neurogenic gene expression in the Drosophila embryo. Curr. Biol. 16, 1358–1365.

Chapter 3 : Gene expression is modulated by interactions between enhancers in the Drosophila *even-skipped* locus

Kelly M. Biette, Tara Lydiard-Martin, Anna Cha, Mary Kathryn Howard, Ben J. Vincent, Meghan D. Bragdon, Timothy T. Harden, Javier Estrada, Zeba Wunderlich, and Angela H. DePace

Author Contributions for Chapter 3:

K.M.B., T.L.M. and A.H.D designed the study; K.M.B., T.L.M., A.C., M.K.H., M.D.J.B, and K.E. collected the data; K.M.B., T. T. H., Z.W. and J.E. processed the data; K.M.B. and T.L.M. analyzed the data with input from B.J.V. and Z.W.; K.M.B. and A.H.D. wrote the manuscript.

Abstract

Animal transcription is controlled by enhancers – DNA sequences that are bound by transcription factors and direct the pattern, timing, and level of gene expression. Enhancers have long been described as "modules" that function independent of distance and orientation with respect to the promoter and other regulatory sequences. We demonstrate that this canonical view is too simplistic by investigating two enhancers of the *even-skipped* gene in *Drosophila melanogaster* blastoderm embryos. We show that these enhancers influence each other's activity when placed in the same reporter construct in a configuration-dependent manner. Moreover, deleting eve37 and eve46 from the endogenous locus perturbs not only the cognate *eve* stripes, but also the position and level of stripes driven by other enhancers. Our results indicate that mechanisms of gene regulation operate at the locus-level to control precise expression of developmental genes.

Introduction

Metazoan transcription is controlled by enhancers – sequences of regulatory DNA bound by transcription factors (TFs) that direct the pattern, level, and timing of gene expression. Enhancers are widely considered to be "modules" whose function is independent of context (Banerji et al., 1981; Bulger and Groudine, 2010; Spitz and Furlong, 2012). However, many classes of cis-regulatory elements, including enhancers, silencers, insulators, and targeting sequences can interact in complex ways (Cubeñas-Potts et al., 2016; Prazak et al., 2010; Spitz, 2016; Spitz and Furlong, 2012; Swanson et al., 2010). For example, shadow enhancers, which

target the same gene and drive overlapping patterns of gene expression, interact unexpectedly (Bothma et al., 2015; Dunipace et al., 2011; El-Sherif and Levine; Perry et al., 2011). Moreover, enhancers physically interact with each other as well as the promoter (Chen et al., 2018; Ghavi-Helm et al., 2014; Proudhon et al., 2016), and having multiple active enhancers in close proximity can lead to high levels of transcriptional activation (Crocker et al., 2017; Dowen et al., 2014; Hnisz et al., 2013). Together, these examples emphasize that enhancers may not function independently in the dense regulatory landscapes surrounding developmental genes.

The *even-skipped* (*eve*) gene of *Drosophila melanogaster* (*D.mel*) is an ideal case study for testing concepts of enhancer modularity. In textbooks, *eve* is often used as an example to illustrate modularity: the seven stripe pattern in the blastoderm embryo is driven by five enhancers, each of which directs expression of one or two stripes (Clyde et al., 2003; Fujioka et al., 1999; Struffi et al., 2011) (Figure 3.1). This experimental system, and *eve* in particular, has long been a flagship system for studies of enhancer function. Decades of developmental genetics have identified the *eve* enhancers and their regulators (Arnosti et al., 1996; Fujioka et al., 1999; Small et al., 1991, 1992, 1996; Struffi et al., 2011) and experimental advances including CRISPR enable precise manipulation of *eve* regulators and genomic regulatory sequences (Gratz et al., 2013, 2014; Ni et al., 2011; Staller et al., 2013). Finally, enhancer activity can be measured at cellular resolution in intact, developing embryos (Fowlkes et al., 2008; Luengo Hendriks et al., 2006; Wunderlich et al., 2014).

The dense regulatory landscape surrounding *eve* is representative of other developmental genes. Regulatory genes, such as transcription factors and signaling proteins, are often expressed in multiple contexts throughout development (Barolo and Posakony, 2002; Carroll, 2008; Kirschner and Gerhart, 2006). Classically, individual enhancers were thought to act independently to drive expression in a particular cell type. However, many developmental genes are surrounded by clusters of highly active enhancers, termed “super-enhancers,” that may function together to synergistically control gene expression (Lovén et al., 2013; Pott and Lieb, 2015; Siersbæk et al., 2014; Whyte et al., 2013). The effects of deleting component

enhancers from these clusters are highly variable, and both synergistic and additive effects have been observed (Bahr et al., 2018; Huang et al., 2018; Moorthy et al., 2017; Suzuki et al., 2017). These results are often enhancer-specific, as studies that delete multiple component enhancers from the same super-enhancer report that some elements are critical for the overall gene expression pattern, while others are dispensable for function (Carleton et al., 2017; Hay et al., 2016; Le Noir et al., 2017).

To measure the consequences of manipulating enhancers, it is useful to image patterns at high spatial resolution over time, as developmental patterns are highly dynamic. Recent studies have used the MS2 system to image the transcriptional dynamics driven by the entire *eve* locus in live embryos, from bacterial artificial chromosomes (BACs) and from a reporter integrated into the endogenous gene (Lim et al., 2018; Berrocal et al., 2018). Their work demonstrates that the *eve* pattern refines into stripes during the course of nuclear cycle 14, with stripes shifting anteriorly over time. Enhancer-enhancer interactions may be important for the precision of this patterning, as an endogenous deletion of the *eve* stripe 1 enhancer leads to earlier and more anterior expression of *eve* stripe 2 (Lim et al., 2018). Another study using an *eve* BAC replaced the stripe 1 and stripe 2 enhancers with bacterial sequences; they did not report significant effects on stripes driven by other enhancers, although quantitative effects are possible (Li and Eisen). These examples support the hypothesis that enhancers may not behave independently, and that interactions may be enhancer-specific.

We previously demonstrated that the enhancers that drive expression of *eve* stripes 3 and 7 (eve37) and stripes 4 and 6 (eve46) influence each other's output when "fused" together in the same reporter construct (Samee et al., 2017). This interaction is not easily explained by interaction of TFs at the boundaries of the fusions, as it persists even when enhancers are placed at a distance. Here, we systematically explore the interaction between the eve37 and eve46 enhancers by manipulating their arrangement and spacing in reporter constructs and deleting them from the endogenous *eve* locus. Our null hypothesis was that the *eve* enhancers each behave as independent modules, each driving a portion of the 7 stripe pattern in the blastoderm

embryo without influencing one another. This is supported by the ability of each eve enhancer to drive expression of a qualitatively accurate stripe in reporters; we were motivated to challenge it by the ample evidence of enhancer/enhancer interactions in the literature and our own previous study of *eve* enhancers, described above. If our null hypothesis is contradicted and we observe evidence for functional interactions between *eve* enhancers, *eve* may serve as a powerful case study to dissect the mechanistic underpinnings of enhancer function in the context of an intact locus.

Our results demonstrate that the function of these two enhancers is not entirely independent. Rather, eve37 and eve46 are functionally interacting with each other over long range, by a mechanism that is not yet understood. First, the level of expression driven by single and fused enhancers can vary significantly with orientation and distance to the promoter. Second, separating eve37 and eve46 by either a spacer sequence or an *eve-LacZ* coding sequence does not result in an expression pattern equal to the additive output from two component enhancers. This interaction is dependent on enhancer configuration. Finally, deletion of eve37 but not eve46 from the endogenous locus affects the position and level of stripes driven by other *eve* enhancers in a stripe-specific manner. Our results motivate future studies to identify mechanisms at the locus-scale for how *eve* enhancers interact with each other and additional classes of regulatory sequences in the *eve* locus.

Results

Constructing and imaging eve enhancer reporters

We constructed a series of enhancer reporter constructs containing the minimal eve37 enhancer (Small et al., 1996) and minimal eve46 enhancer (Fujioka et al., 1999) upstream of the *eve* coding sequence and a *LacZ* reporter (Figure 3.1). When a spacer sequence was required, we used sequences that were computationally designed to be devoid of binding sites for *eve* regulators (Estrada et al., 2016), except in initial experiments that use a portion of the *LacZ*

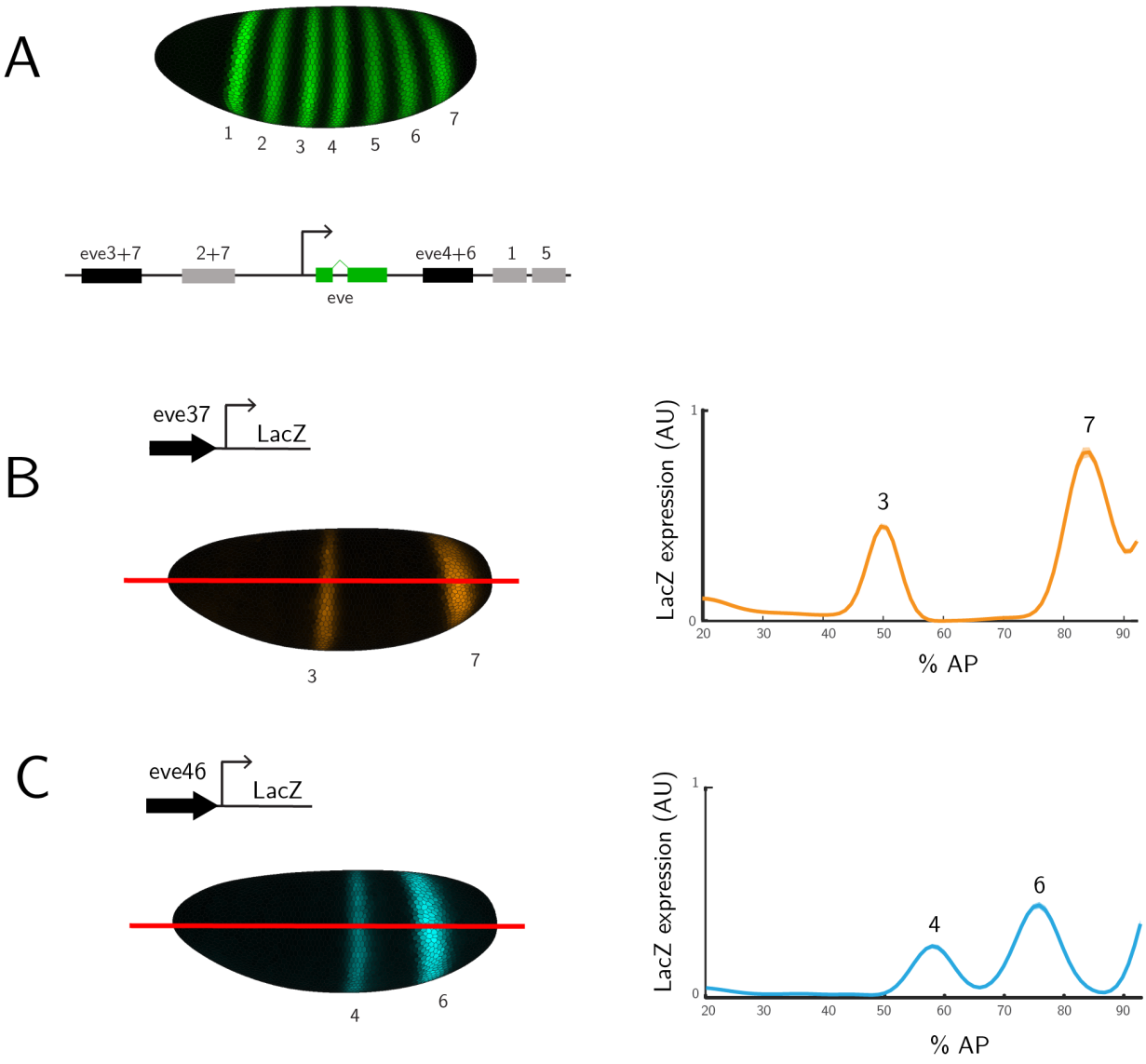


Figure 3.1 : The *even-skipped* (eve) locus contains five enhancers that direct expression of seven stripes. (A) A computational rendering of the endogenous *eve* mRNA expression pattern (green) with stripe numbers noted below (Fowlkes et al., 2008). The relative positions of the five enhancers and coding sequence are shown, with the corresponding stripe numbers listed above each enhancer. (B) A computational rendering of expression driven by an eve37 *LacZ* reporter construct is shown in orange. Throughout this study, we used *in situ* hybridization with a *hkb* co-stain to detect expression levels of reporters. Average signal from each line is plotted as a function of anterior-posterior (AP) position along the lateral side of the embryo. n ≥ 10 embryos with shadow representing standard error of the mean (SEM). (C) The expression pattern driven by an eve46 *LacZ* reporter construct is shown in blue. n ≥ 10 embryos. Data generated and plotted as described in (B).

coding sequence chosen to minimize predicted binding sites for the regulators of these two enhancers. We used the phiC31 system (Fish et al., 2007) to integrate the reporters into the same genomic location and collected blastoderm stage embryos (0-4 hrs after egg laying) from homozygous transgenic lines.

We measured expression from each reporter construct in blastoderm embryos using fluorescent *in situ* hybridization against the *LacZ* reporter and an endogenously expressed fiduciary marker, *fushi-tarazu* (*ftz*). To normalize levels of expression across reporter constructs, we co-stained reporter lines with the endogenously expressed *huckebein* (*hkb*) gene in the same channel as *LacZ* (Wunderlich et al., 2014). We imaged entire embryos at cellular resolution using two-photon confocal microscopy, as previously described (Luengo Hendriks et al., 2006). Three-dimensional image stacks were computationally segmented and processed into PointCloud representations containing the coordinates and fluorescence levels for each nucleus (Luengo Hendriks et al., 2006). For simplicity, we present the data as a lateral line trace, which is the average expression of a five-nuclei wide dorsal-ventral (D/V) strip along the anterior-posterior (A/P) axis (Figure 3.1 B-C).

Interactions between fused enhancers are not explained by orientation or proximity to the promoter

Our previous work on fusions of eve37 and eve46 indicated that these two enhancers can interfere with each other's regulatory function (Samee et al., 2017). To explore this interaction more extensively, we tested a series of variants that altered 1) the order of the enhancers in the fusion, 2) the orientation of the fusions to the promoter, and 3) the distance of the fusions to the promoter. All configurations tested drove patterned expression of more than two stripes, but none of these constructs produced the endogenous four stripe pattern that would be expected from enhancers behaving independently of other regulatory sequences (Figure 3.2). Varying the order of the enhancers in the fusion resulted in the reduced expression from the promoter-

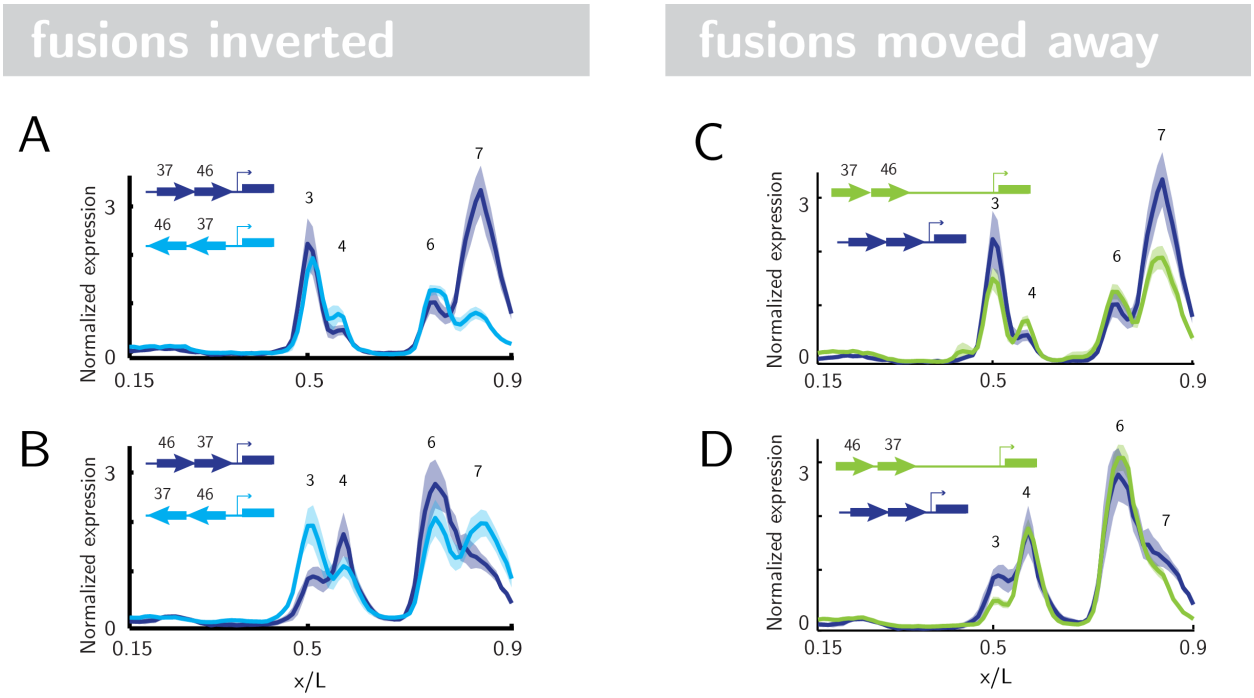


Figure 3.2 : Fused enhancers still interact when inverted or moved away from the promoter. (A-B) Orientation relative to the promoter influences level of expression driven by enhancer fusions. We compared expression of the fused enhancers (dark blue) to a complete inversion of the entire fusion sequence (light blue). Average expression for each transgenic line is displayed as a function of AP position with the shadow indicating standard error of the mean (SEM). n ≥ 5 embryos. (C-D) We tested the importance of promoter proximity in determining the expression level driven by enhancer fusions by introducing a 1000bp *LacZ* spacer between fused enhancers and the promoter. The expression pattern driven by the fusions at a distance (green) is similar to expression adjacent to the promoter (dark blue) for both enhancer configurations, with a decrease in level of stripes 3 and 7. n ≥ 5 embryos. Plotting as described in (A-B).

proximal enhancer (compare dark blue line in Figure 3.2A and 3.2B). To test whether the interactions we observed were orientation-dependent relative to the promoter, we inverted the entire sequence of two enhancer fusions. This experiment switches the enhancer closest to the promoter but keeps the enhancer sequences relative to one another the same. When inverted, the relative level of expression driven by the two enhancers is reversed (Figure 3.2 A-B). This effect cannot be accounted for simply by the change in distance of each enhancer relative to the promoter that occurs in the inversion (see Figure S3.3). When eve37 is moved from the distal to the proximal position, the level of stripe 7 decreases dramatically but the level of stripe 3

remains unchanged. This is surprising considering the regulatory logic for expression of stripes 3 and 7 is distributed within the same enhancer sequence (Small et al., 1996). In contrast, when eve46 is moved from the distal to the proximal position, the level of stripes 4 and 6 decrease slightly and the levels of stripe 3 and 7 increase. These results suggest that the different regulatory logic of individual enhancers, and within each enhancer, are differentially sensitive to position within the locus.

The repression of the promoter-proximal enhancer described above could be explained by direct interactions of enhancer bound TFs on either the promoter bound basal transcriptional machinery, or the promoter-proximal chromatin state. If this hypothesis is true, increasing the distance between the enhancer fusions and the promoter should diminish the promoter-proximal repression. To test this we designed constructs that separated each enhancer fusion from the promoter with a 1kb *LacZ* spacer sequence. These constructs show that increasing the distance between the enhancer fusions and the promoter does not diminish the promoter-proximal repression (Figure 2.2C-D). These results suggest that interactions between the enhancers themselves affect expression, rather than simply an interaction between one enhancer and the promoter.

Separating fused enhancers by a spacer sequence does not restore modular stripe patterns

It has been proposed that two TFs that regulate eve37 and eve46, Hunchback (Hb) and Knirps (Kni), act as short-range repressors over distances of less than 150bp (Clyde et al., 2003; Courey and Jia, 2001; Fakhouri et al., 2010; Gray and Levine, 1996; Li and Arnosti, 2011; Struffi et al., 2011). If short range repression by Hb or Kni is responsible for the loss of stripe levels in the fusions, this effect should be alleviated when the two enhancers are moved apart.

We first separated eve37 and eve46 by a 200bp “neutral” DNA spacer. This spacer separates the enhancers beyond the proposed range of short acting repressors, but we observed that eve37 and eve46 still affect each other’s function (Figure 3.3A and E, Figure S3.1). The

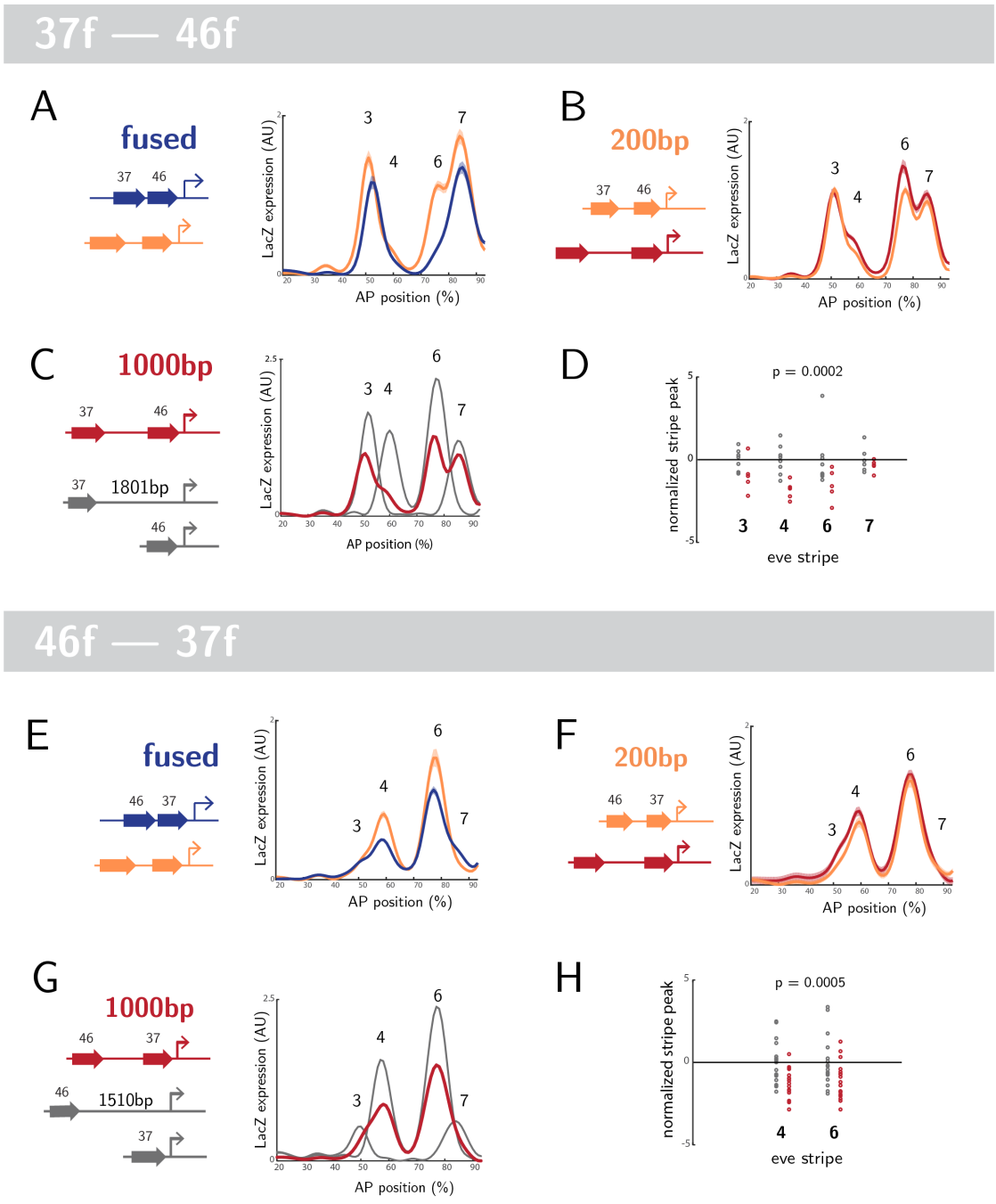


Figure 3.3 : Enhancers drive non-additive expression when placed in the same reporter construct. We measured expression driven by a series of reporter constructs containing eve37 and eve46 in two configurations with different size spacers. (Top) Constructs with eve37 upstream of eve46. n ≥ 5 for all lines. (A) We compared expression driven by fused enhancers (dark blue) to enhancers separated by a 200bp spacer (orange). Average expression for each transgenic line is displayed as a function of AP position with the shadow indicating standard error of the mean (SEM). Peak expression (95th percentile) of each stripe from individual embryos is plotted in Figure S3.1. (B) We compared expression driven by the two enhancers separated by a 200bp (orange) or 1kb (red) spacer sequence. Line traces are displayed as above and individual embryos are displayed in Figure S3.1. Differences

Figure 3.3 (continued).
between the orange trace in A and B are likely due to differences in age distribution of embryos (see Methods). (C) Enhancers separated by a 1kb spacer. We compared expression from reporters containing both enhancers (red) to expression driven by each single enhancer at the same distance from the promoter (gray). (D) We compared the peak expression (95th percentile) of each stripe in individual embryos. The mean peak values for each control are centered on zero to facilitate comparison (see Methods). Colors as in (C). When eve37 is distal to eve46, the pattern is different from a single sum of the individual controls (p-value = 0.0002 by ANOVA). (Bottom) We reversed the orientation of enhancers in the combined reporters so that eve46 is upstream of eve37 and conducted the same series of experiments (E-H). Values are plotted as above. In this reverse configuration, discrete peaks for stripe 3 and 7 are lost and the level of stripes 4 and 6 decreases significantly (p-value = 0.0005 by ANOVA). $n \geq 5$ for all lines. In both cases, the constructs containing two enhancers drive lower expression levels than the single enhancer controls.

degree of effect depends upon the enhancer configuration and spacer sequence (Figure S3.2).

We further separated these enhancers by a 1kb neutral spacer sequence which places them far outside the influence of short-range repression. Even at this distance, the enhancers still demonstrate promoter-proximal repression to a similar degree as constructs with the 200bp spacer (Figure 3.3B and 3.3F, Figure S3.2). To control for the increase in distance of the distal enhancer from the promoter these constructs were compared with distance-matched single enhancer controls (Figure 3.3C and G). While distance from the promoter can affect mRNA levels (Figure S3.3), in this case the effects are not dramatic.

Separating eve37 and eve46 by a coding sequence does not restore modular stripe expression

The standard assay for measuring enhancer function uses reporter constructs that place the putative enhancer upstream of, and close to, a generic promoter. Our results above with enhancer fusions were performed in a similar manner by placing the fusions 5' of the promoter. However within the endogenous *eve* locus, eve37 and eve46 are separated by approximately 9kb because eve37 is upstream of the coding sequence and eve46 is downstream. Considering that these 2 enhancers maintain an effect on each other over 1kb, we asked if placing them in a similar context to their endogenous location, on either side of a gene, would still result in interference with each other's regulatory function (Figure 3.4). We designed reporter constructs

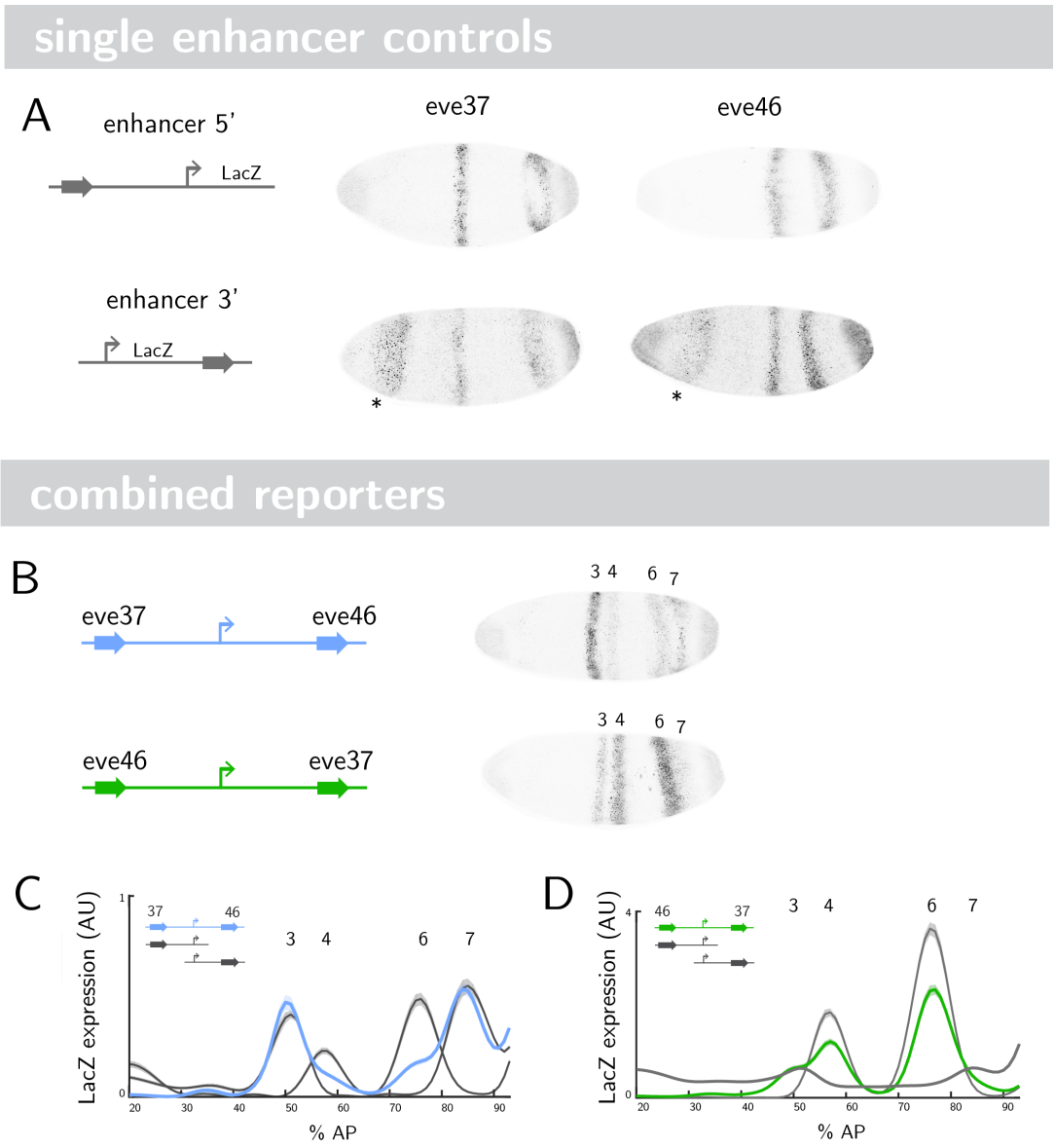


Figure 3.4 : Eve37 and eve46 still interact when separated by a coding sequence. We generated reporter constructs with eve37 and eve46 on opposite sides of the *eve* promoter driving expression of *LacZ*. (A) Expression driven by single enhancer controls varies with position relative to the coding sequence. Maximum projections of *LacZ* signal is shown for eve37 and eve46 upstream (top) and downstream (bottom) of the *LacZ* coding sequence. When either eve37 or eve46 is downstream of the coding sequence, ectopic expression is observed in the anterior of the embryo (denoted by * at 20-30% AP). (B) Reporter constructs containing both enhancers in their endogenous (light blue) and switched (green) positions do not drive expression of four stripes. (C-D) Quantification of comparison between constructs containing both enhancers (endogenous position in light blue and switched in green) compared to single enhancer controls (gray). Average expression for each transgenic line is displayed as a function of AP position with the shadow indicating standard error of the mean (SEM). In both cases, the constructs containing both enhancers do not match the sum of the patterns driven by single enhancer controls.

that place these two enhancers on opposite sides of the *eve* promoter driving a *LacZ* coding sequence and compared expression driven by these constructs to distance-matched controls. Each single enhancer control drove expression of two stripes (Figure 3.4A). Interestingly, both eve37 and eve46 drove stronger and more robust expression when placed 5' of lacZ even though eve46 is 3' of the coding sequence in its endogenous position.

Placing both enhancers within the same reporter construct in this configuration failed to result in the expected expression of all four stripes. In the endogenous orientation (eve37 upstream and eve46 downstream), stripes 3 and 7 are not affected but stripes 4 and 6 are decreased compared to the single enhancer controls. When the enhancers are switched relative to their endogenous positions, the levels of stripes 4 and 6 are decreased and stripe 7 is lost (Figure 3.4B and C). Stripe 7 is especially sensitive to position affects, as it drives robust expression regardless of the presence of the eve46 enhancer when it is upstream of the promoter, but cannot drive expression from the 3' position. These results highlight that even reporter constructs that maintain enhancer position relative to the promoter can fail to recapitulate endogenous expression patterns, and that there are likely additional mechanisms to explain position preference encoded within the sequence that we do not yet understand.

Deleting eve46 and eve37 from the endogenous locus has stripe-specific effects

The interactions we see between eve37 and eve46, even when placed 9kb apart and on opposite sides of a gene, led us to ask if they also functionally impact each other within the endogenous locus. To test this we used the CRISPR/Cas9 system to separately delete each enhancer from the endogenous locus. Importantly, the eve37 deletion is lethal and therefore we carry the allele in a balanced line. The deletion of eve46 results in decreased viability, but a small number of homozygotes survives. To compare these two deletions, we also carry the eve46 deletion in a balanced line.

In the resulting population, there are embryos homozygous for the deletion (expected ¼), heterozygous for the deletion (expected ½) and wildtype embryos (expected ¼). We

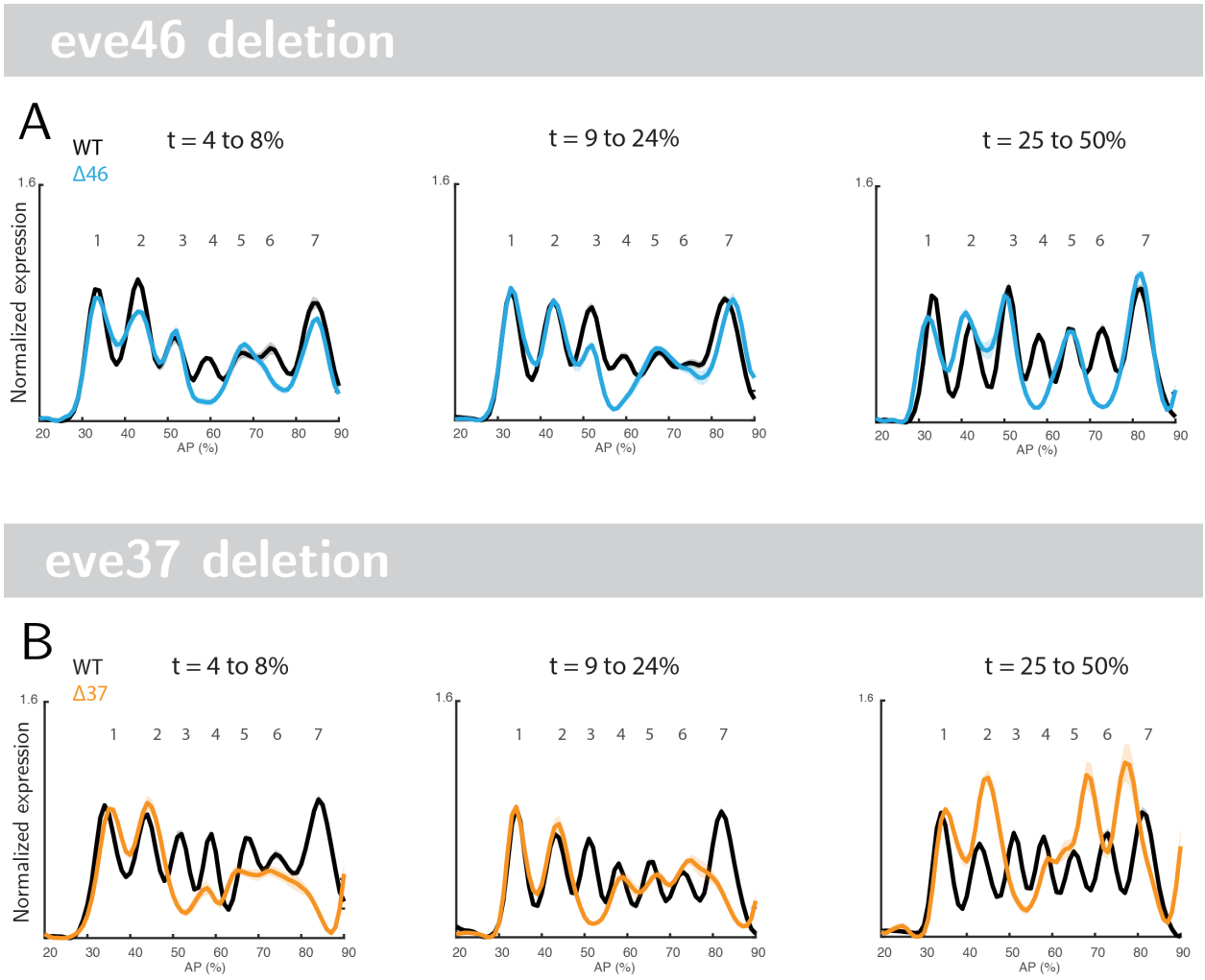


Figure 3.5 : Deleting eve46 and eve37 from the endogenous locus affects the position and level of stripes controlled by other *eve* enhancers. We deleted eve46 and eve37 from the endogenous locus and stained embryos from balanced lines for *eve* mRNA. Average peak expression is displayed as a line trace along the AP axis normalized to stripe 1 expression. Data is separated by time point through developmental stage 5, with WT lines is shown in black, Δeve46 in blue, and Δeve37 in orange. Shadow indicates SEM. (Top) eve46 deletion resulted in small effects on the position and level of stripes 2, 3, and 7. (Bottom) eve37 deletion alters the position of stripes 2 and 6 and the level of all other stripes. Traces from individual embryos are displayed in Figure S3.4. n ≥ 5 embryos for all time points.

collected fixed embryos from these balanced lines and stained them for endogenous *eve* mRNA and another pair-rule gene, *ftz*. We curated embryos into WT/heterozygous and homozygous

groups based on their expression pattern in processed images; embryos containing homozygous deletions express fewer than 7 clear stripes. 24 of 133 (18%) embryos and 19 of 108 (17.6%) processed embryos contained an eve46 or eve37 homozygous deletion, respectively. This number is lower than the expected 1 in 4, and may be due to pattern defects that cause drop-out during image processing. To compare levels of expression in homozygous deletions to wildtype *eve*, in each individual embryo, we normalized to the level of stripe 1 expression; any changes to stripe 1 expression level are therefore obscured, but changes in position are still visible. We display the average traces in Figure 3.5, and traces from individual embryos in the supplement. Importantly, the wild-type traces differ in each plot because they are generated from all embryos in the age range that are not missing stripes. This means that embryos heterozygous for the deletion and homozygous wild-type are all used to generate the wild-type trace because it is impossible to separate their genotypes using our current balancer chromosomes. Furthermore, each wild-type trace is generated from a minimum of five embryos per time point, but could contain a different distribution of embryo ages, resulting in slightly different stripe expression levels because the *eve* pattern in early stage 5 is highly dynamic (i.e. the 4-8% wild-type traces in Figure 3.5A and 3.5B differ slightly because the Δeve46 wild type trace includes a higher proportion of younger embryos).

In embryos lacking eve46, stripes 4 and 6 are absent as expected (Figure 3.5A, Fujioka 2002). There are subtle effects on the position and level of stripes 1, 2 and 3 at early time points, but by the midpoint of stage 5 the other stripes appear normal. In contrast embryos lacking eve37 (Figure 3.5B) are missing stripe 3 and only a small remnant of stripe 7 remains in some embryos. This may be due to the shadow enhancer for stripe 7, known as eve27 (Goto et al., 1989; Harding et al., 1989; Small et al., 1991, 1992, 1996; Staller et al., 2015; Vincent et al., 2018). However *eve* expression in stripes other than 3 or 7 is also altered. At early time points stripes 4, 5, and 6 are reduced in level, and have altered positioning. The level of these stripes partially recovers with time, but stripe position does not. By the midpoint of stage 5, the level of stripes 2, 5, and 6 is higher than in wild type embryos, and the stripe position is skewed

posteriorly. The effect is not identical in all embryos, and embryo-to-embryo variability in stripe position and level is much higher in embryos without eve37 (Figure S3.4). To investigate the effects of the eve46 and eve37 deletions on downstream patterning and organismal phenotype, we are performing cuticle preps and staining for genes downstream of *eve*, including *engrailed*.

Three features of the data differed from our expectations in this experiment. First, we observe different effects on the overall *eve* pattern due to deletions of eve37 and eve46, which suggests that enhancer-enhancer interactions are stripe-specific. Second, we do not observe any late stage 5 embryos lacking full stripes. This may be due to the influence of the *eve* late element, which turns on at the end of stage 5 and is sufficient to drive all 7 stripes. Finally, the eve37 enhancer deletion affects the position and level of stripes driven by other enhancers, but the phenotype is quite subtle. This suggests that there are likely other sequences in the endogenous *eve* locus outside the minimal eve37 enhancer that can buffer the effects of perturbing it, such as the sequence between eve37 and other enhancers, or other portions of the eve 5' locus, which is known to be sufficient to drive seven stripes (Fujioka et al., 1995). Despite this, our results suggest that eve37 interacts with other enhancers in the *eve* locus and these enhancer interactions are important for the precision of *eve* expression in the blastoderm.

Discussion

Here we present evidence that the *eve* enhancers do not function independently of one another. This is clear from the consequences of deleting eve37 from the endogenous *eve* locus: expression of other *eve* stripes is affected. The conclusion is further supported by our observation that *eve* enhancers interact in reporter constructs in a configuration dependent manner.

The discrepancy between expression driven by reporters and endogenous deletions is especially informative, as it reveals the influence of modifying sequences that likely buffer the effect of enhancer deletions in the endogenous context but are not present in the reporter constructs. Reporters offer the possibility of uncovering these sequences with "add-back" style

experiments that would allow us to quantify the influences of sequences that do not have activity on their own. Finally, the effect of removing the entire eve37 enhancer on the remainder of the *eve* pattern is quite subtle, and would not have been convincing on its own without the results from reporter constructs.

We observe different effects on the overall *eve* pattern due to deletions of eve37 and eve46. These results suggest that enhancer-enhancer interactions are stripe-specific. This is further supported by the conflicting results of deleting and replacing the *eve* stripe 1 (Lim et al., 2018) and 2 (Li and Eisen; Ludwig et al., 2011) enhancers from BAC reporters and the endogenous locus. The endogenous deletion of *eve* stripe 1 was shown to affect the both temporal and spatial precision of *eve* stripe 2 expression, and the authors hypothesize that the anterior-most stripe may function as the "dominant" stripe to coordinate the anterior-to-posterior shift of *eve* pattern observed during blastoderm embryogenesis (Lim et al., 2018). In contrast, another study replaced the *eve* stripe 1 and 2 enhancers with bacterial sequences in a BAC reporter and does not report significant alterations to the pattern or level of stripes driven by other enhancers (Li and Eisen). Finally, Ludwig et al. replaced the minimal *eve* stripe 2 enhancer in the endogenous locus with a fragment of the *mini-white* gene (Ludwig et al., 2011). When we stained for *eve* mRNA in this line, we observed effects on the position of stripes 3 and 7 and the level of stripe 6 (Figure S3.5). Together, these experiments indicate that the mechanisms underlying enhancer interactions likely depend on enhancer-specific features and *eve* will clearly be a powerful system for dissecting locus-level interactions.

Enhancer-enhancer interactions have been observed in other studies in mammalian systems that measured chromatin contacts and gene expression following enhancer deletions (Le Noir et al., 2017; Proudhon et al., 2016). In another system, constituent enhancers were deleted from the a-globin super-enhancer, the individual enhancers appeared to act modularly, suggesting that some enhancers may interact more than others (Hay et al., 2016). Genome-wide studies of physical interactions between regulatory elements using 3C-based chromatin conformation techniques in Drosophila have reported both enhancer-promoter and enhancer-

enhancer contacts (Beagrie et al., 2017; Ghavi-Helm et al., 2014). These studies demonstrate that enhancer interactions are likely sensitive to location in the locus or bound regulatory proteins, and are not simply a nonspecific byproduct of the crowded nuclear environment. As the resolution of these assays continues to improve, it will become possible to look at enhancer-enhancer and enhancer-promoter interactions within a single locus, ideally with high spatial resolution.

There are several sequences in the *eve* locus with functions that might be critical for integrating the inputs from multiple stripe enhancers, such that endogenous gene expression is relatively modular but two enhancers isolated in reporter constructs do not behave as such. This could include sequence between the enhancers or binding sites for looping or architectural proteins that are present in both the enhancers themselves and other regions, such as the promoter or 3' UTR. The locus is flanked by two insulator sequences, "Homie" and "Nhomie" that are thought to pair to separate *eve* into its own topologically associated domain, which may be critical for promoter communication with the 3' enhancers (Fujioka et al., 2009, 2013, 2016). There are a number of known insulator proteins with binding sites in both the *eve* promoter and throughout the locus, including GAGA factor (GAF), Pleiohomeotic (Pho), and Drosophila CTCF (dCTCF) that organize chromatin topology into structural domains that enable precise gene expression patterns (Chetverina et al., 2017; Kaye et al., 2017; Maksimenko and Georgiev, 2014; Postika et al., 2018). Because we used synthetic "neutral" spacers between the eve37 and eve46 enhancers, our reporter assays did not include these sequences, and now serve as a useful platform to add them back in controlled way to measure their effects.

Our experiments in reporter constructs demonstrate that the presence of one enhancer affects expression driven by another enhancer in an orientation-dependent manner. This places a strong constraint on possible mechanisms, as simple short-range repression or protein-protein interactions are not likely orientation-dependent at these length scales. Instead, the nature of this interaction points towards directional or steric mechanisms such as enhancer transcription (Henriques et al., 2018; Mikhaylichenko et al., 2018) or DNA topology (Beagrie et al., 2017;

Huang et al., 2018; Ing-Simmons et al., 2015; Proudhon et al., 2016). Finally, the five enhancers of *eve* are bound by partially overlapping sets of TFs and cofactors, which may contribute to the stripe-specific effects we observe if they are used by some, but not all, *eve* enhancers. This influence may be exerted directly, when TFs are bound, or indirectly, by increasing the local concentration of bound TFs. Recent studies have demonstrated that TFs may form phase-separated regions through their unstructured domains, which could lead to elevated local concentration and a shared pool of TFs (Cho et al., 2018; Chong et al., 2018; Sabari et al., 2018; Tsai et al., 2017).

Enhancer competition has been proposed as one model for how promoters integrate information from multiple enhancers in the same locus (Bothma et al., 2015; Scholes et al., 2018). While competition could explain the reduction in level we observe when enhancers are placed in the same reporter construct, it can not explain the differences in stripe boundaries observed when eve37 (this study) and the *eve* stripe 1 enhancer (Lim et al., 2018) are deleted. A change in boundary position when these enhancers are removed suggests that the promoter processes input TFs differently, as expression is now on in cells when it had previously been kept off. Our results argue that describing enhancer function in a given cell must include not only interactions between active enhancers and the promoter, but also inactive enhancers that are traditionally disregarded.

If we cannot consider active enhancers in isolation, how should we read regulatory information from the genome? How much context should we include and how does it influence gene expression? Our results call for new types of experiments to decipher locus-level modes of transcriptional control, including the role of enhancer-enhancer interactions and modifying sequences in the eve locus. Computational models have considered how regulatory sequence governs gene expression within enhancers (Bussemaker et al., 2001; Fakhouri et al., 2010; He et al., 2010; Ilsley et al., 2013; Janssens et al., 2006; Kazemian et al., 2010; Segal et al., 2008; Wunderlich et al., 2012; Zinzen et al., 2006) and in entire loci (Kim et al., 2013; Samee and Sinha, 2014). In our previous paper, we analyzed the fusion constructs we present here using

computational models at the scale of individual enhancers and the entire locus (Samee et al., 2017). Our results indicate that different rules govern the sequence-to-function relationship within enhancers and over the entire locus, and that the scale of molecular interactions is larger than that mapped for linear sequence space. Deciphering the underlying molecular mechanisms at both scales will be critical to predict the consequences of regulatory mutations, which occur at the scale of TF binding sites (such as single nucleotide polymorphisms) and the locus scale (such as deletions, inversions and duplications (Haraksingh and Snyder, 2013; Maurano et al., 2015; Stankiewicz and Lupski, 2010). Future studies will decipher the principles by which enhancers functionally interact by continuing to integrate models of gene expression with experimental measurements, which will likely include rules at different length scales and probabilistic rather than rigid enhancer boundaries.

Materials and Methods

Transgenic fly lines

We used RedFly to identify coordinates of the *D. mel even-skipped* stripe 3/7 and stripe 4/6 enhancers (Gallo et al., 2011). The eve_stripe_3+7 element is 510bp (Release 5 coordinates 2R: 5863006-5863516) (Small et al., 1996), while the eve_stripe4_6 element is 800bp (Release 5 coordinates 2R:5871404-5872203) (Fujioka et al., 1999). Note that the stripe 4/6 enhancer coordinates from REDfly contain an extra 208bp on the 3' end compared to the construct tested in Fujioka et al., 1999 (Fujioka et al., 1999). Enhancers were PCR amplified from genomic DNA from w118 *Drosophila melanogaster* flies, sequence verified and inserted into the multiple cloning site of the pBOY vector (Hare et al., 2008) using isothermal assembly (Gibson et al., 2009). The spacer sequence in these constructs was computationally screened to remove binding sites for transcription factors known to regulate *eve* using the algorithm SiteOut (Estrada et al., 2016). A link to a list of all enhancers and spacer sequences used can be found in the Supplemental Information. pBOY contains an *eve* core promoter 20bp downstream of the multiple cloning site that drives an *eve*/*LacZ* fusion transcript. The vector also contains an attB site for phiC31 site specific integration (Fish et al., 2007) and the mini-white gene for selection of transformants. Each plasmid was injected into attP2 flies (Markstein et al., 2008) by BestGene and homozygosed using the mini-white eye color marker.

Embryo collection and in situ hybridization

Embryo collection and whole mount *in situ* hybridization was performed as previously described (Luengo Hendriks et al., 2006). Briefly, 0-4hr embryos (25°C) were collected, dechorionated in 50% bleach, fixed in a 1:4 mixture of 10% formaldehyde to heptane, and devitellinized in heptane and methanol by shaking. Embryos were post-fixed in formaldehyde and a formaldehyde-based hybridization buffer. Hybridizations were performed at 56°C with three full length cDNA probes: a DIG-labeled probe for fushi tarazu (ftz), a DNP-labeled *LacZ* or *eve* probe and a DNP-labeled probe against *huckebein* (*hkb*). The probes were detected by

successive antibody staining using anti-DIG-HRP (anti-DIG-POD; Roche, Basel, Switzerland) and anti-DNP-HRP (Perkin-Elmer TSA-kit, Waltham, MA, USA), and labeled by reactions with coumarin- and Cy3-tyramide (Perkin-Elmer). Embryos were treated with RNase and incubated with Sytox Green (Invitrogen, Carlsbad, CA, USA) to stain nuclei. Finally, embryos were dehydrated in ethanol and mounted in DePex (Electron Microscopy Sciences, Hatfield, PA, USA), using #1 coverslips to form a bridge to preserve 3D embryo morphology.

Imaging and image processing

Embryos were imaged and computationally segmented for further analysis (Fowlkes et al., 2008). A three-dimensional image stack of each embryo was acquired on a Zeiss LSM Z10 with a plan-apochromat 20x0.8 NA objective using 2-photon microscopy. Embryos were binned into six time points of approximately 10 minute windows using the extent of membrane invagination under phase-microscopy as a morphological marker. Time points correspond to 0-3%, 4-8%, 9-25%, 26-50%, 51-75% and 76-100% membrane invagination along the side of the embryo that has progressed most. We imaged embryos from all age ranges and, except where noted, display data from the early blastoderm (4% - 31% membrane invagination) when *hkb* normalization is used and from 4% - 51% membrane invagination when expression is normalized to stripe 1 (Figure 3.5 only). Image files were processed into PointCloud representations containing the coordinates and fluorescence levels for each nucleus.

hkb normalization

Normalization to a *hkb* co-stain was performed to test the variation in absolute levels of expression across reporter lines (Wunderlich et al., 2014). Embryos were stained with a mixture of *LacZ*- or *eve*-DNP and *hkb*-DNP probe. For each embryo, background was calculated as the mode of the fluorescence distribution. After subtracting background, mean *hkb* fluorescence was calculated as the geometric mean of the anterior and posterior expression domains. We noted that eve stripe 7 overlaps slightly with the posterior expression domain of hkb, and so chose to

use the geometric mean of anterior and posterior rather than solely the posterior domain as in (Wunderlich et al., 2014) to limit the impact of overlapping expression. The fluorescence in each nucleus was then divided by the mean *hkb* fluorescence to yield a normalized expression level.

Endogenous enhancer deletions

We expressed guide RNAs that target regions just outside the annotated eve37 or eve46 enhancer boundaries under Dm-snRNA:U6:96Ab promoter (Addgene plasmid # 45946, (Gratz et al., 2013)) and injected these constructs into *D. mel* lines expressing genetically encoded Cas9 under *vasa* (Bloomington Stock 51324) or *nanos* regulatory sequences (information for this line yw;;nos-Cas9(III-attP2) is available on BestGene website). gRNA sequences are from the UCSC genome browser release dm6 and are listed in Figure S3.5B. The regions around the PAM sites were sequenced in these flies to make sure there were no mutations within the PAM. Injections were performed by BestGene.

Males were collected from each injection and crossed with double balancer virgins w[*]; Kr[If-1]/CyO; D[1]/TM3, Ser[1] (Bloomington Stock 7198). From this cross, CyO balancer containing males were collected and crossed with virgins from the double balancer line. From the male progeny of this cross, we purified genomic DNA and sequenced the region surrounding the enhancers and identified lines carrying a deletion. CyO males from these lines were selected and crossed with double balancer virgins and the resulting deletion/CyO males and virgins were propagated. The deletion of eve37 is lethal and therefore we carry the deletion allele in a balanced line. The deletion of eve46 results in decreased viability, but a small number of homozygotes survive. To compare these two deletions, we also carry the eve46 deletion allele in a balanced line.

In the resulting population, there are embryos homozygous for the deletion (expected ¼), heterozygous for the deletion (expected ½) and wildtype embryos (expected ¼). We collected fixed embryos from these balanced lines and stained them for endogenous *eve* mRNA and another pair-rule gene, *ftz*. We curated embryos into WT/heterozygous and homozygous

groups based on their expression pattern in processed images; embryos containing homozygous deletions express fewer than 7 clear stripes. 24 of 133 (18%) embryos and 19 of 108 (17.6%) processed embryos contained an eve46 or eve37 homozygous deletion, respectively. This number is lower than the expected 1 in 4, and may be due to pattern defects that cause drop-out during image processing. To compare levels of expression in homozygous deletions to wildtype *eve*, in each individual embryo, we normalized to the level of stripe 1 expression; any changes to stripe 1 expression level are therefore obscured, but changes in position are still visible.

Data analysis and visualization

Extraction of lateral line traces, and detection of stripe boundaries, and *hkb* normalization were performed in MATLAB using the PointCloud Toolbox (http://bdtnp.lbl.gov/Fly- Net/bioimaging.jsp?w=analysis) and custom scripts. Briefly, lateral line traces are a smoothed moving window average over a 1/16th DV strip (about 5 nuclei wide) along the left side of the embryo and the normalized mean expression level across lines is displayed. We tested multiple measures of peak level (95th and 99th percentile expression, maximum expression level of the stripe) and are reporting the most conservative metric. To identify the 95th percentile, we ranked normalized expression values from all cells in the stripe (0.45 to 0.55 for stripe 3 and 0.8 to 0.9 for stripe 7) and report the 95th percentile value. To quantify the effect of deleting enhancers, we normalized cellular expression measurements to stripe 1 expression.

References

Arnosti, D.N., Barolo, S., Levine, M., and Small, S. (1996). The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development *122*, 205–214.

Bahr, C., von Paleske, L., Uslu, V.V., Remeseiro, S., Takayama, N., Ng, S.W., Murison, A., Langenfeld, K., Petretich, M., Scognamiglio, R., et al. (2018). A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. Nature *553*, 515–520.

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

Barolo, S., and Posakony, J.W. (2002). Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. Genes Dev. *16*, 1167–1181.

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., de Santiago, I., Lavitas, L.-M., Branco, M.R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature *543*, 519–524.

Berrocal, A., Lammers, N.C., Garcia, H.G., and Eisen, M.B. (2018). Kinetic sculpting of the seven stripes of the Drosophila even-skipped gene.

Bothma, J.P., Garcia, H.G., Ng, S., Perry, M.W., Gregor, T., and Levine, M. (2015). Enhancer additivity and non-additivity are determined by enhancer strength in the Drosophila embryo. Elife *4*.

Bulger, M., and Groudine, M. (2010). Enhancers: the abundance and function of regulatory sequences beyond promoters. Dev. Biol. *339*, 250–257.

Bussemaker, H.J., Li, H., and Siggia, E.D. (2001). Regulatory element detection using correlation with expression. Nat. Genet. *27*, 167–171.

Carleton, J.B., Berrett, K.C., and Gertz, J. (2017). Multiplex Enhancer Interference Reveals Collaborative Control of Gene Regulation by Estrogen Receptor α-Bound Enhancers. Cell Syst *5*, 333–344.e5.

Carroll, S.B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell *134*, 25–36.

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., and Gregor, T. (2018). Dynamic interplay between enhancer-promoter topology and gene activity. Nat. Genet.

Chetverina, D., Fujioka, M., Erokhin, M., Georgiev, P., Jaynes, J.B., and Schedl, P. (2017). Boundaries of loop domains (insulators): Determinants of chromosome form and function in multicellular eukaryotes. Bioessays *39*.

Cho, W.-K., Spille, J.-H., Hecht, M., Lee, C., Li, C., Grube, V., and Cisse, I.I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science eaar4199.

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G.M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X., et al. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. Science *361*.

Clyde, D.E., Corado, M.S.G., Wu, X., Paré, A., Papatsenko, D., and Small, S. (2003). A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. Nature *426*, 849–853.

Courey, A.J., and Jia, S. (2001). Transcriptional repression: the long and the short of it. Genes Dev. *15*, 2786–2796.

Crocker, J., Tsai, A., Muthusamy, A.K., Lavis, L.D., Singer, R.H., and Stern, D.L. (2017). Nuclear Microenvironments Modulate Transcription From Low-Affinity Enhancers.

Cubeñas-Potts, C., Rowley, M.J., Lyu, X., Li, G., Lei, E.P., and Corces, V.G. (2016). Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. Nucleic Acids Res.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387.

Dunipace, L., Ozdemir, A., and Stathopoulos, A. (2011). Complex interactions between cis-regulatory modules in native conformation are critical for Drosophila snail expression. Development *138*, 4075–4084.

El-Sherif, E., and Levine, M. Shadow Enhancers Mediate Dynamic Shifts of Gap Gene Expression in the Drosophila Embryo. Curr. Biol.

Estrada, J., Ruiz-Herrero, T., Scholes, C., Wunderlich, Z., and DePace, A.H. (2016). SiteOut: An Online Tool to Design Binding Site-Free DNA Sequences. PLoS One *11*, e0151740.

Fakhouri, W.D., Ay, A., Sayal, R., Dresch, J., Dayringer, E., and Arnosti, D.N. (2010). Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo. Mol. Syst. Biol. *6*, 341.

Fish, M.P., Groth, A.C., Calos, M.P., and Nusse, R. (2007). Creating transgenic Drosophila by microinjecting the site-specific phiC31 integrase mRNA and a transgene-containing donor plasmid. Nat. Protoc. *2*, 2325–2331.

Fowlkes, C.C., Hendriks, C.L.L., Keränen, S.V.E., Weber, G.H., Rübel, O., Huang, M.-Y., Chatoor, S., DePace, A.H., Simirenko, L., Henriquez, C., et al. (2008). A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm. Cell *133*, 364–374.

Fujioka, M., Emi-Sarker, Y., Yusibova, G.L., Goto, T., and Jaynes, J.B. (1999). 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.

Fujioka, M., Wu, X., and Jaynes, J.B. (2009). A chromatin insulator mediates transgene homing and very long-range enhancer-promoter communication. Development *136*, 3077–3087.

Fujioka, M., Sun, G., and Jaynes, J.B. (2013). The Drosophila eve insulator Homie promotes eve expression and protects the adjacent gene from repression by polycomb spreading. PLoS Genet. *9*, e1003883.

Fujioka, M., Mistry, H., Schedl, P., and Jaynes, J.B. (2016). Determinants of Chromosome Architecture: Insulator Pairing in cis and in trans. PLoS Genet. *12*, e1005889.

Gallo, S.M., Gerrard, D.T., Miner, D., Simich, M., Des Soye, B., Bergman, C.M., and Halfon, M.S. (2011). REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. Nucleic Acids Res. *39*, D118–D123.

Ghavi-Helm, Y., Klein, F.A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E.E.M. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. Nature *512*, 96–100.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., 3rd, and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods *6*, 343–345.

Goto, T., Macdonald, P., and Maniatis, T. (1989). 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.

Gratz, S.J., Cummings, A.M., Nguyen, J.N., Hamm, D.C., Donohue, L.K., Harrison, M.M., Wildonger, J., and O'Connor-Giles, K.M. (2013). Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. Genetics *194*, 1029–1035.

Gratz, S.J., Ukken, F.P., Rubinstein, C.D., Thiede, G., Donohue, L.K., Cummings, A.M., and O'Connor-Giles, K.M. (2014). Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in Drosophila. Genetics *196*, 961–971.

Gray, S., and Levine, M. (1996). Short-range transcriptional repressors mediate both quenching and direct repression within complex loci in Drosophila. Genes Dev. *10*, 700–710.

Haraksingh, R.R., and Snyder, M.P. (2013). Impacts of variation in the human genome on gene regulation. J. Mol. Biol. *425*, 3970–3977.

Harding, K., Hoey, T., Warrior, R., and Levine, M. (1989). Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila. EMBO J. *8*, 1205–1212.

Hare, E.E., Peterson, B.K., Iyer, V.N., Meier, R., and Eisen, M.B. (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet. *4*, e1000106.

Hay, D., Hughes, J.R., Babbs, C., Davies, J.O.J., Graham, B.J., Hanssen, L.L.P., Kassouf, M.T., Marieke Oudelaar, A., Sharpe, J.A., Suciu, M.C., et al. (2016). Genetic dissection of the [alpha]-globin super-enhancer in vivo. Nat. Genet. *48*, 895–903.

He, X., Samee, M.A.H., Blatti, C., and Sinha, S. (2010). Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput. Biol. *6*.

Henriques, T., Scruggs, B.S., Inouye, M.O., Muse, G.W., Williams, L.H., Burkholder, A.B., Lavender, C.A., Fargo, D.C., and Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. Genes Dev.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Huang, J., Li, K., Cai, W., Liu, X., Zhang, Y., Orkin, S.H., Xu, J., and Yuan, G.-C. (2018). Dissecting super-enhancer hierarchy based on chromatin interactions. Nat. Commun. *9*, 943.

Ilsley, G.R., Fisher, J., Apweiler, R., De Pace, A.H., and Luscombe, N.M. (2013). Cellular resolution models for even skipped regulation in the entire Drosophila embryo. Elife *2*, e00522.

Ing-Simmons, E., Seitan, V.C., Faure, A.J., Flicek, P., Carroll, T., Dekker, J., Fisher, A.G., Lenhard, B., and Merkenschlager, M. (2015). Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. Genome Res. *25*, 504–513.

Janssens, H., Hou, S., Jaeger, J., Kim, A.-R., Myasnikova, E., Sharp, D., and Reinitz, J. (2006). Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. Nat. Genet. *38*, 1159–1165.

Kaye, E.G., Kurbidaeva, A., Wolle, D., Aoki, T., Schedl, P., and Larschan, E. (2017). Drosophila Dosage Compensation Loci Associate with a Boundary-Forming Insulator Complex. Molecular and Cellular Biology *37*.

Kazemian, M., Blatti, C., Richards, A., McCutchan, M., Wakabayashi-Ito, N., Hammonds, A.S., Celniker, S.E., Kumar, S., Wolfe, S.A., Brodsky, M.H., et al. (2010). Quantitative analysis of the Drosophila segmentation regulatory network using pattern generating potentials. PLoS Biol. *8*.

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

Kirschner, M., and Gerhart, J.C. (2006). The Plausibility of Life: Resolving Darwin's Dilemma (Yale University Press).

Le Noir, S., Boyer, F., Lecardeur, S., Brousse, M., Oruc, Z., Cook-Moreau, J., Denizot, Y., and Cogné, M. (2017). Functional anatomy of the immunoglobulin heavy chain 3 ' super-enhancer needs not only core enhancer elements but also their unique DNA context. Nucleic Acids Res. *45*, 5829–5837.

Li, L.M., and Arnosti, D.N. (2011). Long- and short-range transcriptional repressors induce distinct chromatin states on repressed genes. Curr. Biol. *21*, 406–412.

Li, X.-., and Eisen, M.B. Mutation of sequences flanking and separating transcription factor binding sites in a Drosophila enhancer significantly alter its output.

Lim, B., Fukaya, T., Heist, T., and Levine, M. (2018). Temporal dynamics of pair-rule stripes in living Drosophila embryos. Proc. Natl. Acad. Sci. U. S. A.

Lovén, J., Hoke, H.A., Lin, C.Y., Lau, A., Orlando, D.A., Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R.A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320–334.

Ludwig, M.Z., Manu, Kittler, R., White, K.P., and Kreitman, M. (2011). Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. PLoS Genet. *7*, e1002364.

Luengo Hendriks, C.L., Keränen, S.V.E., Fowlkes, C.C., Simirenko, L., Weber, G.H., DePace, A.H., Henriquez, C., Kaszuba, D.W., Hamann, B., Eisen, M.B., et al. (2006). Three-dimensional

morphology and gene expression in the Drosophila blastoderm at cellular resolution I: data acquisition pipeline. Genome Biol. *7*, R123.

Maksimenko, O., and Georgiev, P. (2014). Mechanisms and proteins involved in long-distance interactions. Front. Genet. *5*, 28.

Markstein, M., Pitsouli, C., Villalta, C., Celniker, S.E., and Perrimon, N. (2008). Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. Nat. Genet. *40*, 476–483.

Maurano, M.T., Wang, H., John, S., Shafer, A., Canfield, T., Lee, K., and Stamatoyannopoulos, J.A. (2015). Role of DNA Methylation in Modulating Transcription Factor Occupancy. Cell Rep. *12*, 1184–1195.

Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I.E., Males, M., Viales, R.R., and Furlong, E.E.M. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev.

Moorthy, S.D., Davidson, S., Shchuka, V.M., Singh, G., Malek-Gilani, N., Langroudi, L., Martchenko, A., So, V., Macpherson, N.N., and Mitchell, J.A. (2017). Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. Genome Res. *27*, 246–258.

Ni, J.-Q., Zhou, R., Czech, B., Liu, L.-P., Holderbaum, L., Yang-Zhou, D., Shim, H.-S., Tao, R., Handler, D., Karpowicz, P., et al. (2011). A genome-scale shRNA resource for transgenic RNAi in Drosophila. Nat. Methods *8*, 405–407.

Perry, M.W., Boettiger, A.N., and Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. Proc. Natl. Acad. Sci. U. S. A. *108*, 13570–13575.

Postika, N., Metzler, M., Affolter, M., Müller, M., Schedl, P., Georgiev, P., and Kyrchanova, O. (2018). Boundaries mediate long-distance interactions between enhancers and promoters in the Drosophila Bithorax complex. PLoS Genet. *14*, e1007702.

Pott, S., and Lieb, J.D. (2015). What are super-enhancers? Nat. Genet. *47*, 8–12.

Prazak, L., Fujioka, M., and Gergen, J.P. (2010). Non-additive interactions involving two distinct elements mediate sloppy-paired regulation by pair-rule transcription factors. Dev. Biol. *344*, 1048–1059.

Proudhon, C., Snetkova, V., Raviram, R., Lobry, C., Badri, S., Jiang, T., Hao, B., Trimarchi, T., Kluger, Y., Aifantis, I., et al. (2016). Active and Inactive Enhancers Cooperate to Exert Localized and Long-Range Control of Gene Regulation. Cell Rep. *15*, 2159–2169.

Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J., Hannett, N.M., Zamudio, A.V., Manteiga, J.C., et al. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. Science.

Samee, M.A.H., and Sinha, S. (2014). Quantitative modeling of a gene's expression from its intergenic sequence. PLoS Comput. Biol. *10*, e1003467.

Samee, M.A.H., Lydiard-Martin, T., Biette, K.M., Vincent, B.J., Bragdon, M.D., Eckenrode, K.B., Wunderlich, Z., Estrada, J., Sinha, S., and DePace, A.H. (2017). Quantitative Measurement and Thermodynamic Modeling of Fused Enhancers Support a Two-Tiered Mechanism for Interpreting Regulatory DNA. Cell Rep. *21*, 236–245.

Scholes, C., Biette, K.M., Harden, T.T., and DePace, A.H. (2018). Computations performed by shadow enhancers and enhancer duplications vary across the Drosophila embryo.

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

Siersbæk, R., Rabiee, A., Nielsen, R., Sidoli, S., Traynor, S., Loft, A., Poulsen, L.L.C., Rogowska-Wrzesinska, A., Jensen, O.N., and Mandrup, S. (2014). Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. Cell Rep. *7*, 1443–1455.

Small, S., Kraut, R., Hoey, T., Warrior, R., and Levine, M. (1991). Transcriptional regulation of a pair-rule stripe in Drosophila. Genes Dev. *5*, 827–839.

Small, S., Blair, A., and Levine, M. (1992). Regulation of even-skipped stripe 2 in the Drosophila embryo. EMBO J. *11*, 4047–4057.

Small, S., Blair, A., and Levine, M. (1996). Regulation of Two Pair-Rule Stripes by a Single Enhancer in theDrosophilaEmbryo. Dev. Biol. *175*, 314–324.

Spitz, F. (2016). Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. Semin. Cell Dev. Biol. *57*, 57–67.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Staller, M.V., Yan, D., Randklev, S., Bragdon, M.D., Wunderlich, Z.B., Tao, R., Perkins, L.A., DePace, A.H., and Perrimon, N. (2013). Depleting Gene Activities in Early Drosophila Embryos with the "Maternal-Gal4–shRNA" System. Genetics *193*, 51–61.

Staller, M.V., Vincent, B.J., Bragdon, M.D.J., Lydiard-Martin, T., Wunderlich, Z., Estrada, J., and DePace, A.H. (2015). Shadow enhancers enable Hunchback bifunctionality in the Drosophila embryo. Proc. Natl. Acad. Sci. U. S. A. *112*, 785–790.

Stankiewicz, P., and Lupski, J.R. (2010). Structural variation in the human genome and its role in disease. Annu. Rev. Med. *61*, 437–455.

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

Suzuki, H.I., Young, R.A., and Sharp, P.A. (2017). Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis. Cell *168*, 1000–1014.e15.

Swanson, C.I., Evans, N.C., and Barolo, S. (2010). Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. Dev. Cell *18*, 359–370.

Tsai, A., Muthusamy, A.K., Alves, M.R., Lavis, L.D., Singer, R.H., Stern, D.L., and Crocker, J. (2017). Nuclear microenvironments modulate transcription from low-affinity enhancers. Elife *6*.

Vincent, B.J., Staller, M.V., Lopez-Rivera, F., Bragdon, M.D.J., Pym, E.C.G., Biette, K.M., Wunderlich, Z., Harden, T.T., Estrada, J., and DePace, A.H. (2018). Hunchback is counter-repressed to regulate even-skipped stripe 2 expression in Drosophila embryos. PLoS Genet. *14*, e1007644.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Wunderlich, Z., Bragdon, M.D., Eckenrode, K.B., Lydiard-Martin, T., Pearl-Waserman, S., and DePace, A.H. (2012). Dissecting sources of quantitative gene expression pattern divergence between Drosophila species. Mol. Syst. Biol. *8*, 604.

Wunderlich, Z., Bragdon, M.D., and DePace, A.H. (2014). Comparing mRNA levels using in situ hybridization of a target gene and co-stain. Methods *68*, 233–241.

Zinzen, R.P., Senger, K., Levine, M., and Papatsenko, D. (2006). Computational models for neurogenic gene expression in the Drosophila embryo. Curr. Biol. *16*, 1358–1365.

Chapter 4: A fully synthetic platform for interrogating combinatorial control of gene regulation by mammalian transcription factors

Kelly M. Biette*, Minhee Park*, Timothy T. Harden, Clarissa Scholes, Ahmad Khalil and Angela H. DePace

Author Contributions to Chapter 4:

K.M.B., M.P., T.T.H., C.S., A.K., A.H.D. designed experiments; K.M.B., M.P. and C.S. performed experiments, K.M.B., M.P., and T.T.H. analyzed the data; K.M.B. and A.H.D wrote results and discussion. *denotes equal contribution

Introduction

Gene expression in animals is widely considered to be regulated by combinatorial action of multiple transcription factors (TFs) (Carey, 1998; Latchman, 1997; Struhl, 1991). Animal enhancers, DNA sequences that direct gene expression in space and time, contain many binding sites for multiple TFs. By working in combinations, a relatively small number of factors can produce a large range of transcriptional outputs (McKenna and O'Malley, 2002; Smith et al., 2005; Spitz and Furlong, 2012; Wagner, 1999). This type of system is useful for generating diverse cell types and enabling dynamic responses to a myriad of different environmental signals using a small number of components. However, two fundamental questions about how combinatorial control is implemented remain. First, is every TF functionally unique, or are they organized into functional classes? Second, how do TFs of different functions combine and how does this depend on the arrangement of TF binding sites in regulatory DNA? These problems have been central to the field of transcription for decades, and a large number of careful case studies have yielded insights into particular TFs and regulation of their cognate genes (Fishburn et al., 2005; Guertin et al., 2010; Staller et al., 2018; Xiao et al., 1991) A major challenge for systems biology is to build on these insights using empirical approaches to determine if there are general principles of combinatorial control; and if so, to use them to predict the expression output of regulatory DNA.

While there are great advantages to an empirical platform, interpreting this type of data is challenging. What should we expect? What form will rules take? What type of platform will allow us to best interrogate combinatorial control and produce generalizable principles? One approach is to train a statistical model on very large datasets, which would need to be collected

for many combinations of TFs at a given enhancer (Kinney and Atwal, 2014). This type of work at the lac operon in bacteria (Kinney et al., 2010) and interferon-beta enhancer in human cells (Melnikov et al., 2012) produced a map of critical binding sites in a given regulatory sequence, but the results are difficult to generalize to any piece of sequence. Another option is a model based on physical principles, where once the rules are known, the output of new sequences and TF combinations can be predicted (discussed in (Bentovim et al., 2017) . We are aiming for this second type of experimental system, which requires us to understand the activity of the individual components and the rules that govern their interaction.

With this approach in mind, how should we think about the range of output we might achieve from recruiting engineered TF combinations? Physical models of TF interactions exist, and are largely classified into three types based on experimental evidence: cooperative binding, collaborative cooperativity, and kinetic synergy. The predominant mode of cooperative interaction that has been studied thus far is binding interactions, both physical interactions between individual TFs, and between TFs and DNA. There are many examples of TFs influencing the DNA binding of other proteins, such as RNA polymerase, through direct protein-protein interactions (Adams and Workman, 1995; Gertz et al., 2009; Veitia, 2003) as well as through collaborative cooperativity, where TFs influence each other indirectly by changing nucleosome occupancy or recruiting other factors (McKenna and O'Malley, 2002; Mirny, 2010). Experimental evidence for direct and indirect combinatorial control led to a body of experimental evidence focused on how binding site position, arrangement, and spacing produces precision in gene expression patterns (Gisselbrecht et al., 2013; Melnikov et al., 2012; Sharon et al., 2012; Smith et al., 2013; Weingarten-Gabbay et al.). However, there are some observations from animal regulatory DNA that does not fit with this view of combinatorial control, such as the degenerate nature of binding motifs (Wunderlich and Mirny, 2009), the extremely fast binding kinetics of TFs to DNA (Stavreva et al., 2015; Voss and Hager, 2014) and the finding that binding sites come and go over evolutionary time (Arnold et al., 2014; Hare et

al., 2008; Tsong et al., 2006), suggesting that direct physical interactions between TFs are not the only source of specificity in transcript regulation.

Another way TFs could function to cooperatively control transcription is by working on different biochemical steps of the transcription cycle. Eukaryotic transcription is known to be controlled cyclically, with multiple steps that may be regulated at a given gene, including chromatin remodeling, recruitment of the pre-initiation complex, promoter-proximal pause release, and others (Fuda et al., 2009; Nechaev and Adelman, 2011; Scholes et al., 2016). Gene-specific activation usually requires more than one type of TF, suggesting that combinatorial or synergistic effects between them are possible (Beagrie and Pombo, 2016; Herschlag and Johnson, 1993; Spitz and Furlong, 2012). Such effects likely depends on sequence and promoter context and has only been tested directly in a small number of studies (Blau et al., 1996; Chi et al., 1995; Keung et al., 2014; Porter et al., 1997). A recent study in flies demonstrated that TFs can be classified into functional groups, and that in some cases TFs from different groups can function synergistically to co-activate an endogenous gene (Stampfel et al., 2015). Another study in flies showed that the activator GAGA factor (GAF) works upstream of RNAP pausing at many genes, presumably by opening chromatin, and heat-shock factor (HSF) acts later at the same genes to release paused RNAP into productive elongation (Duarte et al., 2016). These examples motivate a medium-throughput screening assay to measure how many TF behave alone and in combination at a given promoter.

It is unlikely that these modes of combinatorial control are mutually exclusive in eukaryotes. Our goal is to interrogate their limitations with models and experiments to improve our understanding of how regulatory information is encoded and processed. An ideal experimental platform to interrogate combinatorial control would enable measurement of expression driven by TFs individually and in pairwise or greater combinations with the genomic context rigorously controlled. This is challenging at endogenous genes, where the effect of a single TF is difficult to isolate, as recruiting factors interferes with the natural regulation of that gene, and any conclusions may not be generalizable. We would like to control the affinity and

specificity of TF binding sites, which is difficult with endogenous TFs where DNA recognition is often weak and non-specific (Crocker et al., 2015; Farley et al., 2015, 2016; GuhaThakurta and Stormo, 2001). Ultimately, we would like to be able to compare our data to existing models of transcriptional regulation, which will be more rigorous if we can control exactly the identity of the TFs, when they are present, and in what concentrations. From this perspective, a synthetic platform gives us the most control over our experimental conditions and enables us to test specific, mechanistic hypotheses about how TFs work together.

We built a modular, synthetic platform to recruit TFs to a particular site in specific numbers and combinations and measure the resulting expression quantitatively from a GFP reporter. The technology to do so has only recently become available with new advances in protein engineering (Khalil et al., 2012; Park et al., 2018), DNA design (Estrada et al., 2016) and synthesis. We engineered zinc finger binding domains to recruit activation domains to a precisely defined genomic location upstream of a single promoter driving a reporter gene in mammalian cells. In this system, we can tune a number of "knobs" that mimic endogenous gene activation, including the activation domain recruited, the binding site arrangement and affinity, and the promoter used. We find that changes in each of these features impacts expression driven by a small pilot set of TFs. Our results demonstrate some evidence for TFs working together by promoting different steps of the transcription cycle, and this platform generates results that are particularly amenable to use in computational modeling, which is currently underway. The platform we describe can be used for additional screening of many pairs of TFs in a systematic and rigorously controlled way.

Results

A highly controlled and fully synthetic platform for interrogating TF function in mammalian cells

In order to investigate how TFs work together, we recruited specific combinations of TFs to a synthetic promoter and measured the resulting changes in gene expression. This is feasible with synthetic fusion proteins, "synTFs" (Khalil et al., 2012), which consist of an activation domain fused to an engineered zinc-finger binding domain that exclusively recognizes a binding site that is orthogonal to the mammalian genome (Figure 4.1A). The synTFs, which we call "116" and "127" drive similar levels of expression from a single binding site, suggesting that their affinity is similar (Figure S4.1A). The short (20bp) binding sites are similar in length to many mammalian motifs and can be added into reporters in any array size or combination, making them a flexible tool for studies of TF behavior.

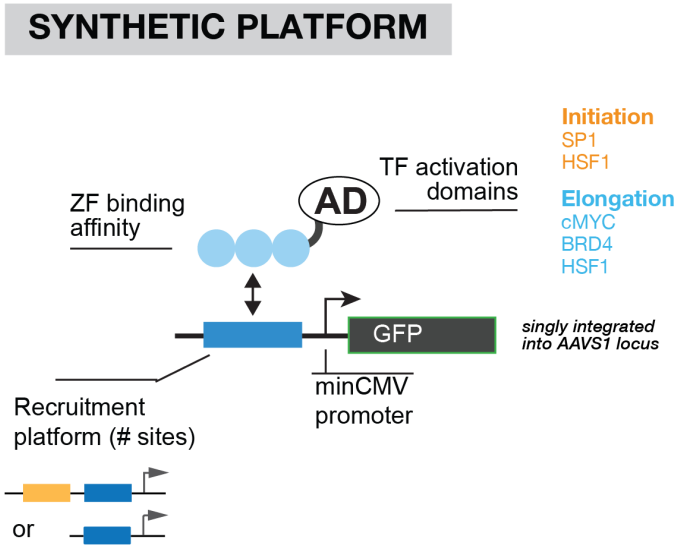


Figure 4.1 : A fully synthetic system for interrogating combinatorial control in mammalian gene regulation. Our modular platform allows us to systematically vary the identify of the attached activation domain, affinity of the DNA binding domain, number of TF binding sites and promoter driving GFP expression, as highlighted in this schematic, and measure the resulting effects on GFP expression. synTFs are custom designed zinc finger (ZF) binding domains fused to any activation domain of interest. In this pilot study, we tested how four TFs, canonically thought to work on initiation or elongation, active expression in our minimal system. Initiation factors : SP1 and HSF1. Elongation factors : cMYC, BRD4 and HSF1. Note that HSF1 is thought to work on both initiation and elongation steps. GFP reporters containing single or two synTF binding sites upstream of a minCMV promoter were singly integrated into the AAVS1 locus of HEK293T cells and synTFs are transfected into these stable lines alone and in combination. The resulting expression is measured by flow cytometry 2 days post-transfection unless otherwise noted.

We created a synthetic reporter construct with either a single synTF binding site or two sites (one for each ZF-116 and ZF-127) upstream of a minimal cytomegalovirus (minCMV) promoter driving GFP expression (Figure 4.1) and singly integrated these reporters into the AAVS1 locus of HEK293T cells (described in (Park et al., 2018). We selected the minCMV promoter for these experiments because it drove a leaky basal expression level and was activated best by the strong viral activator VP16 compared the other promoters we tried (Figure S4.2). The mammalian cMYC promoter was also activated in our early experiments, but since we use cMYC as an activator in our assay we chose to avoid the confounding effects of endogenous auto-regulation in our otherwise synthetic experiments and work exclusively with the minCMV promoter.

The synTF binding sites in our promoter are separated by 14bp computationally designed to be devoid of binding sites for representative TFs from most mammalian families (see Methods, (Estrada et al., 2016)). We removed these binding sites to ensure that the promoter is activated by the factors we transfect, rather than by endogenous proteins. We transfect one or two synTFs into cell lines containing a single or two synTF binding sites and measure GFP expression driven by each combination using flow cytometry. We show that TF choice, the number and type of binding sites in the promoter, and the affinity of the DNA binding domain impact expression by individual and pairs of TFs.

Characterizing a panel of mammalian synTFs

Because we are varying multiple parameters and testing how TFs work in combination, we initially characterized a small set of four mammalian TFs based on descriptions of their activity in the literature. We screened 9 synTFs for activity at the minCMV promoter and selected 4 of them to use going forward based on (1) they drove expression at the minCMV promoter and (2) are found endogenously in mammalian cells (as opposed to being viral proteins). The factors we selected are SP1, cMYC, BRD4 and HSF1.

SP1 is thought to work on initiation (i.e. recruitment of the basal transcriptional machinery), cMYC and BRD4 work downstream of this step to facilitate elongation of RNA Polymerase, and HSF1 is thought to work on both initiation and elongation steps. From here, we refer to SP1 as an "initiation factor," to cMYC and BRD4 as "elongation factors" and HSF1 as working on "both" steps.

The initiation factors SP1 and HSF1 interact with general transcription factors in the pre-initiation complex (Brown et al., 1998; Emili et al., 1994; Gill et al., 1994). The elongation factors cMYC and BRD4 are thought to act downstream of initiation by recruiting the kinase p-TEFb, which stimulates the release of RNAP from promoter-proximal pausing into productive elongation (Eberhardy and Farnham, 2001, 2002; Itzen et al., 2014). Notably, HSF1 can also increase the release of RNAP from the promoter into productive elongation (Blau et al., 1996; Yankulov et al., 1994), but has a mutant form that has been demonstrated to be elongation-deficient (Brown et al., 1998). We included this mutant in early iterations of our study, but it did not behave much differently than wild-type HSF1 individually or in combination with other synTFs, so we dropped it to reduce the combinatorial space of our screen.

A panel of mammalian synTFs drives a range of outputs at at the minCMV promoter

Our first step was to determine the level of expression driven by individual synTFs to use as a baseline for how they activate expression together. We transfected 10ng of each activation domain fused to the ZF-127 binding domain into cells containing a single ZF-127 site upstream of the minCMV promoter and measured the level of GFP expression using FACS (Figure 4.2A). These four synTFs all activated this promoter, but expression levels varied with HSF1 activating ~4 fold over baseline and cMYC driving ~1.5 fold increase in expression. Both factors thought to work on initiation (HSF1 and SP1) activated to a greater extent than either elongation factor (cMYC and BRD4) in this assay.

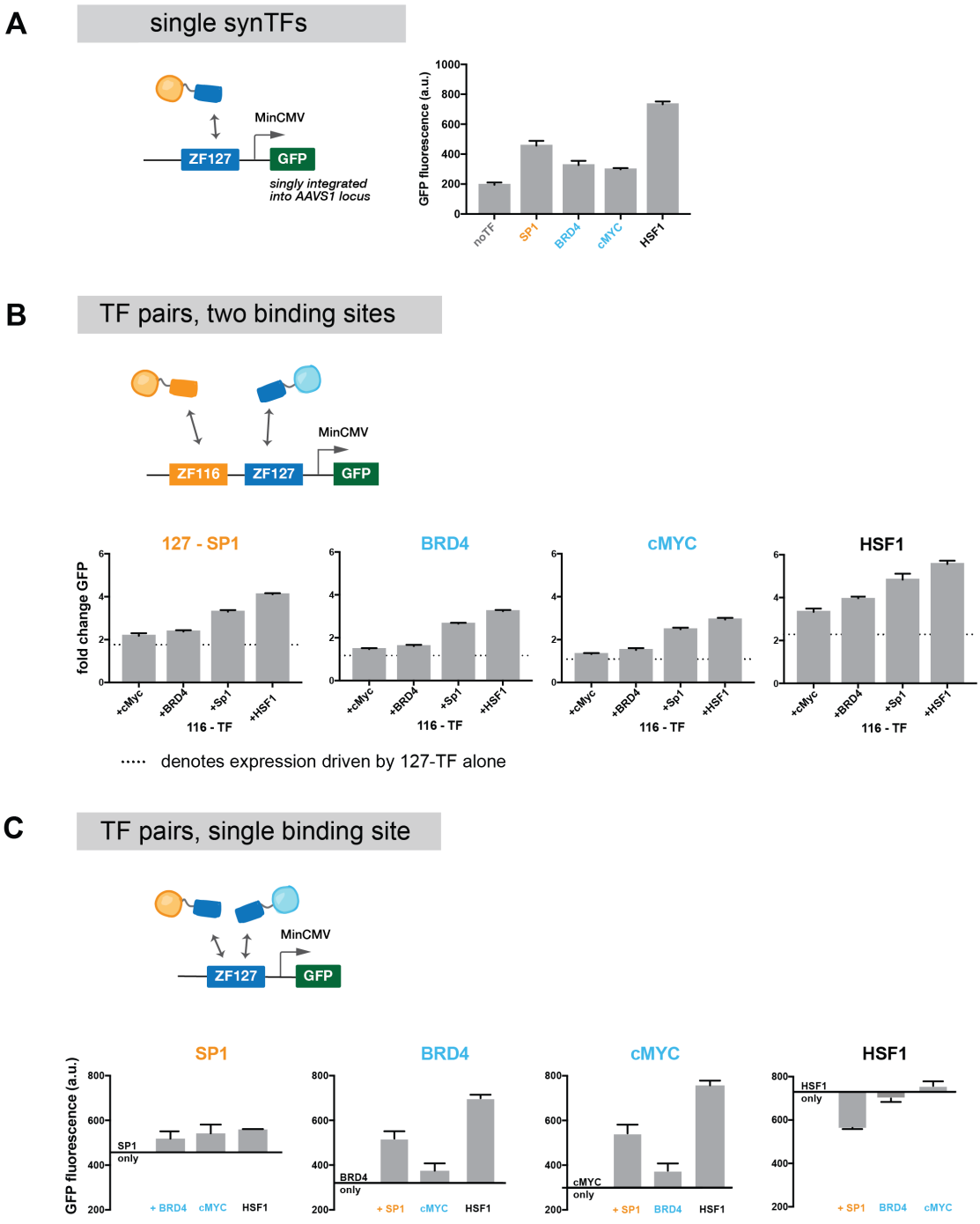


Figure 4.2. Single synTFs activate the minCMV promoter and produce a range of combinatorial behavior. (A) We measured expression driven by single synTF activation domains fused to the ZF-127 binding domain and display the mean GFP fluorescence measured by FACS. Initiation factors in orange, elongation factors in blue, HSF1 (both initiation and elongation) in black. (B) We recruited all pairwise combinations of synTFs to a single promoter

Figure 4.2 (continued).
containing one binding site for ZF -116 located slightly (14bp) upstream of a single binding site for ZF-127. We transfected 20ng total synTF DNA and plot the resulting fold-change GFP expression over baseline (i.e. no TFs transfected) from three technical replicates. For each TF, the dotted line represents the amount of expression measured by from transfection of the 127-TF alone. Note that these values are plotted as fold-change because they were collected from a different flow cytometry machine than all other experiments in this chapter, and so the raw values cannot be directly compared. (C) We co-transfected pairs of synTFs that have identical DNA binding domains and forced the TF pairs to compete for a single genomic binding site. Resulting expression (mean GFP signal) is plotted. For each TF, the baseline of the graph is the amount of expression the synTF drives alone (taken from Figure 4.2A) so that the effect of adding a second TF can be quickly evaluated (i.e. increased expression if the bar is above the line and decreased expression if it drops below). For all TFs except HSF1, adding another TF increases expression. All experiments were performed in triplicate and the average of 3 technical replicates is plotted. Error bars denote standard deviation (SD).

A clear limitation of the following experiments is that we have not yet demonstrated that the cellular concentration of each synTF is the same. While they are expressed under the same promoter and transfected in the same concentration, the activations domains that we use range in size from 54 (BRD4) to 236 (SP1) amino acids so it is possible that they are expressed in the cell with some variability, which would affect our interpretation of these results. Experiments to confirm the cellular expression of these proteins by western blot are underway.

Two synTFs are better than one: recruiting pairs drives more expression in all cases

The simplest first test of how TFs work together is to compare the expression driven by a single TF to the expression output achieved when they are both present at the same promoter. We used two different synTF binding sites to recruit synTFs in all pairwise combinations and measured expression in each instance. Values are plotted as fold-change GFP over baseline in this experiment because the data was collected on a different flow cytometer than all other experiments, so the raw values cannot be directly compared.

In all cases, expression driven by two synTFs was higher than that driven the single 127-TF fusion alone (compare bars to dotted line in Figure 4.2B). In this experiment, expression seems to scale by the strength of the individual TFs recruited. For example, cMyc is the weakest

activator on its own (Figure 4.2A) and recruiting cMYC to the 116 site in any of these experiments causes a slight increase in expression over what the 127-TF was driving on its own. On the other hand, HSF1 is the strongest activator in our pilot set, and in all cases, recruiting HSF1 to the promoter raises expression dramatically over what the single TF was able to do on its own. The design of this experiment simply tells us that expression is higher with two synTFs than with one, which suggests they may work cooperatively but cannot distinguish how.

Recruiting TF pairs to a single genomic binding site reveals functional interactions between TFs

One experiment that may allow us to distinguish between binding cooperativity and "kinetic" cooperativity at the level of function is to ask whether activators interact when they are not physically present on the DNA at the same time. The most direct way to show that TFs can work on different steps in transcript regulation is to show that it can occur in a time-separated manner. Our synthetic platform allows us to recruit both TFs to the same single site upstream of the promoter, such that only one can be bound to the DNA at once, and measure expression driven by each alone and both transfected together.

If we consider using this assay as a diagnostic for synergistic activation, we needed a null hypothesis to which we can compare our results. Thinking of TFs as "time-sharing" in this situation, i.e. one or the other can be bound, but never both, a relatively naive null is that TFs each occupy the binding site 50% of the time, given that they have exactly the same DNA binding domain and only one site is available. In this case, if TFs work completely independently, we would expect expression output from both transfected together to be the average of expression driven by either synTF alone. If expression is higher with two synTFs than the output driven by either one alone, that argues for two (or more) regulated steps in the process of creating an mRNA. If there was only one regulated step, expression should never increase above that driven by the strongest synTF, as adding another TF that works on the same step should not drive more expression from the same promoter.

In most instances, GFP expression driven by pairs of synTFs is higher than expression driven by a single synTF, except in the case of HSF1 (Figure 4.2C). For each TF, the baseline of the graph is the amount of expression the synTF drives alone (taken from Figure 4.2A) so that the effect of adding a second TF can be quickly evaluated (i.e. increased expression if the bar is above the line and decreased expression if it drops below). Interestingly, adding initiation factors to BRD4 and cMYC increased expression much more than adding the other elongation factor. It is notable that expression driven by HSF1 cannot be significantly increased with the addition of any one of our activators, suggesting that if two regulated steps exist, HSF1 is capable of acting on both of them.

Tuning expression output by rationally mutating the affinity of synTF binding domains

The results in Figure 4.2C demonstrate limited evidence for interactions between TFs that cannot be explained simply by cooperative binding to the DNA, as both TFs cannot be physically bound at the same time. We wanted to test how expression driven by pairs of synTFs "time-sharing" changed when their binding domains were mutated to decrease the affinity of the synTF to DNA. Because our results in Figure 4.2C used the same DNA binding domain, we were concerned that a single TF may bind, but never unbind, thus preventing a second synTF from accessing the site in the promoter. This is a particular issue given that our promoters are singly integrated into the genome, so we are not averaging signal across many binding sites in a given cell.

To test the hypothesis that binding affinity might alter interactions between synTFs, we mutated five, seven, or twelve of the arginine residues in the zinc finger array, changing them to alanines (Figure 4.3A). These residues form contacts between the protein and the phosphate backbone of the DNA, enhancing the interaction strength of the synTF (Khalil et al., 2012). When mutant binding domains fused to an HSF1 activation domain were transfected into the

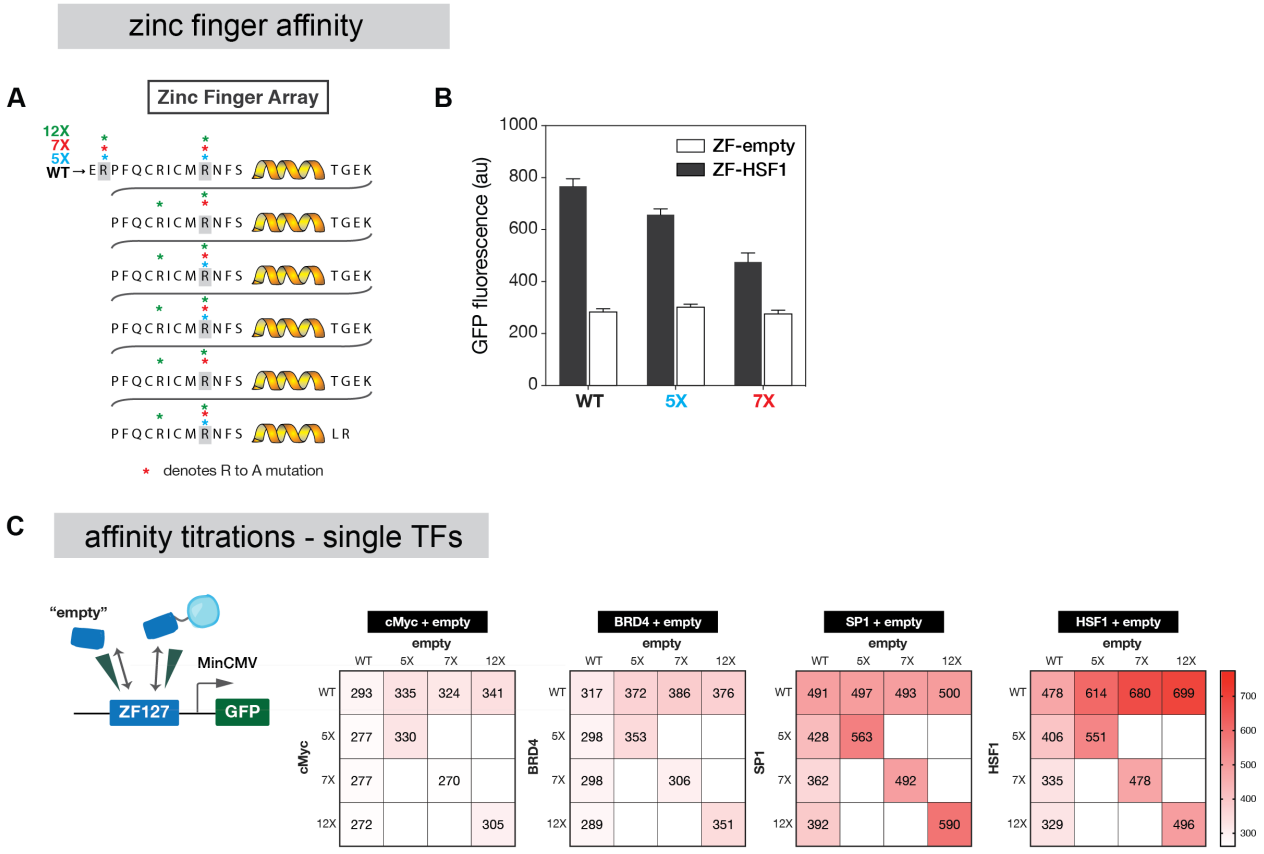


Figure 4.3. Tuning expression output by rationally mutating the affinity of synTF binding domains. (A) Schematic of zinc finger binding domain sequence. We mutated 5 or 7 of the arginine residues on the synTF that contact the phosphate backbone of the DNA at the synTF binding site. Mutated residues are denoted with an asterisk; 12X in green, 7X in red, 5X in blue. (B) We measured GFP expression driven by HSF1 fused to a WT, 5X, or 7X mutant binding domain and an "empty" (i.e. no activation) binding domain contained the same mutations. The mean GFP fluorescence from three technical replicates is displayed. (C) We explored how affinity alters the "time-sharing" behavior of single synTFs by transfecting equal amounts of the individual synTF and an "empty" zinc finger without an activation domain into the cell line containing only a single genomic binding site for both factors. Mean GFP expression from three technical replicates is plotted as a heatmap with raw values in each cell and scaled by color from white (lowest) to red (highest). In all plots, the binding domain of the synTF of interest changes on the vertical axis and the "empty" binding domain affinity changes along the horizontal axis. Along the diagonal, TF and empty have equivalent binding domains and along the vertical or horizontal, one of the pair has a weaker affinity while the other remains wild-type (no arginines mutated).

line containing a single zinc finger binding site upstream of the minCMV promoter, the HSF1 fusions containing mutations drove less expression than the WT (Figure 4.3B), and scaled with the number of mutated arginines (12X data not shown). In this experiment, we use expression

driven by the binding domain fused to a strong activator as a proxy for affinity and confirm that identical activation domains drive less expression with the mutations whereas expression driven by an "empty" binding domain does not change. We have not yet confirmed that all synTFs are expressed to the same extent in cells post-transfection. We are currently confirming cellular expression levels of the mutants post-transfection and are expressing and purifying these proteins to make in vitro measurements of their binding affinities using single molecule imaging.

Titrating the effect of single synTFs with an empty competitor reveals unpredictable binding preferences

We explored how affinity alters the "time-sharing" behavior of single synTFs by transfecting equal amounts of the individual synTF and the "empty" zinc finger (binding domain but no activation domain) into the cell line containing only a single ZF-127 binding site upstream of the minCMV promoter (Figure 4.3C). In all cases, the binding domain of the synTF changes along the vertical axis and the binding domain of the empty competitor changes along the horizontal axis. Along the diagonal, the two have the same affinity DNA binding domain but along either axis one of the pair has a weaker affinity and the other remains wild-type.

Each of the synTFs behaved slightly differently in this competitor titration experiment. cMYC drove slightly more expression when the affinity of the competitor was reduced, but the effect is very small. BRD4 drives the highest level of expression when it has a wild-type domain and its competitor has the lowest affinity, but expression in this condition is quite comparable to when both BRD4 and the empty TF have 12X mutant domains. SP1 on the other hand, drives the most expression when both it and its competitor have weakened binding domains, even more than when SP1 is wild-type and the empty protein is highly mutant. Finally, HSF1 behaves as we might expect for a strong activator : it drives the greatest amount of GFP expression when it has a strong affinity binding domain and the competitor domain is weak. This experiment suggests

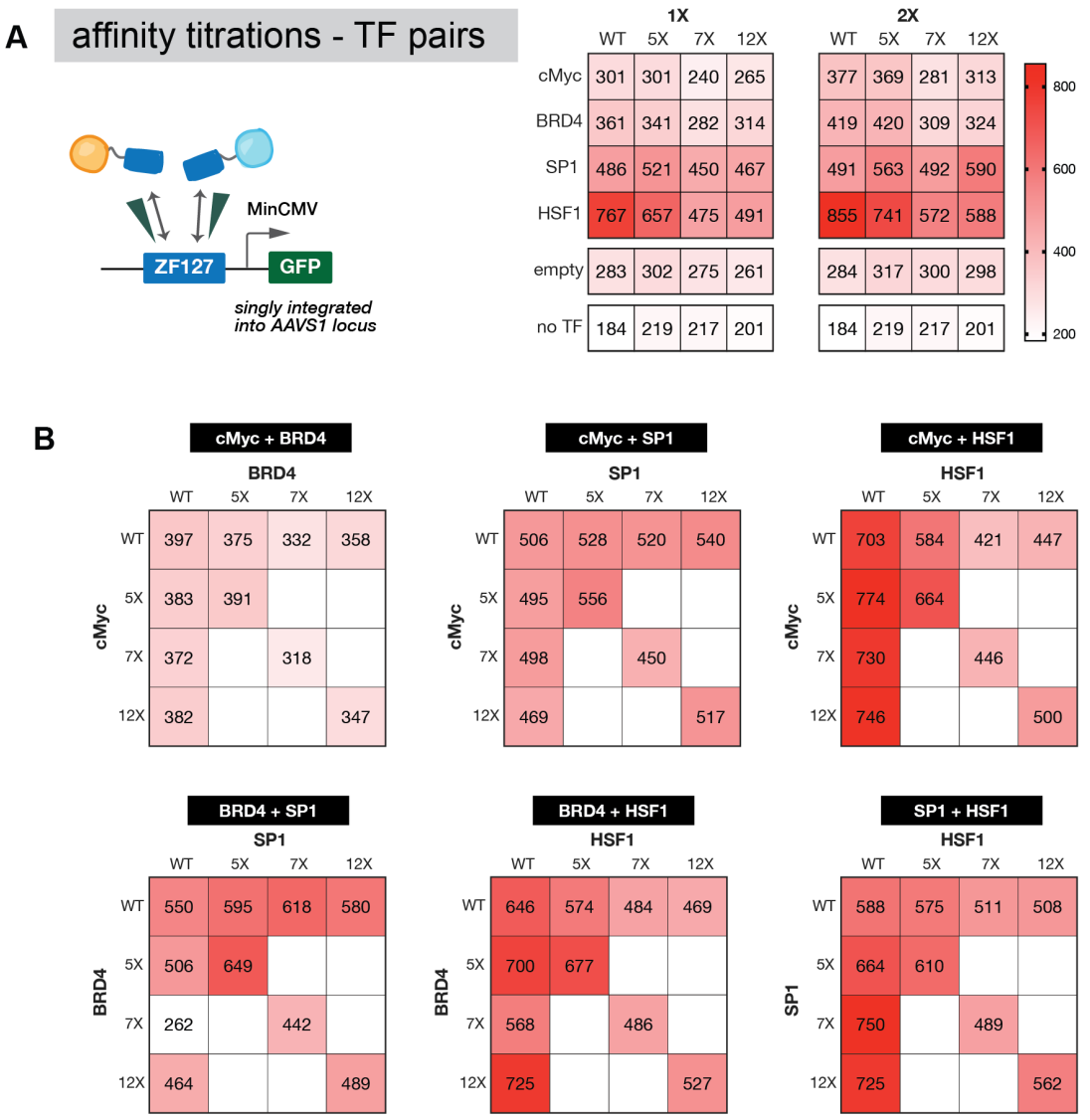


Figure 4.4: Combinatorial behavior changes with affinity when TFs are forced to compete for a single genomic binding site. We then measured how "time-sharing" by two different synTFs changes when their binding affinities are varied. (A) We transfected all single synTFs into the stable cell line containing a single ZF-127 binding site upstream of the minCMV promoter, as denoted by the schematic on the left. On the right, expression output driven by 10ng of each synTF is shown as a heatmap scaled from lowest (white) to highest (red) with mean values from three technical replicates in each cell. The right panel (2X) is the same experiment with the amount of transfected DNA doubled. We use this as a control for doubling the amount of synTF DNA transfected when looking at pairs rather than single synTF controls. In all cases, doubling the amount of synTF transfected increases expression slightly. (B) GFP expression for all pairwise synTF combinations. In each graph, affinity for one synTF varies along the vertical axis and the other is varied along the horizontal axis. Along the diagonal, both TFs have the same binding domain. The number of arginine to alanine mutations in the DNA binding domain is denoted along the top horizontal or vertical axis and all combinations of TF pairs are shown. Interestingly, BRD4 and HSF1 seem to prefer high affinity binding, while SP1 drives more expression in combination with a weak binding domain and affinity does not appear to matter for cMYC.

that each TF has slightly different binding domain preferences that impact how much expression it drives in combination.

"Tuning" TF contributions in competition experiments using affinity mutations

Finally, we asked whether we can use affinity mutations to tune expression output by altering the "time-sharing" behavior of synTF combinations at a single promoter. It is difficult to interpret the synTF interactions we observe in Figure 4.3C; since there is only a single binding site, the synTFs repress each other as only one can be bound at a time. We attempted to relax this constraint by exploring how expression changed when we mixed binding domains of different affinities. For example, can we drive more expression by HSF1 and elongation factors when HSF1 can move quickly (assuming our affinity mutations actually alter the off-rate of TFs) on and off the DNA? We were also interested in deciphering the binding preferences of particular activation domains as a proxy for their biochemical activity. Elongation factors, including BRD4 and cMYC, may not need to bind DNA and are perhaps more effective if recruited to an mRNA molecule or downstream of the transcription start site, where they could effect promoter-proximal pause release (Eberhardy and Farnham, 2002; Itzen et al., 2014). In contrast, if SP1 and HSF1 function at the initiation step, we hypothesized they would drive more expression when strongly bound. We varied the binding domains used in pairwise combinations of TFs and measured the resulting GFP expression driven by each.

We first transfected all single synTFs with zero (WT), five (5X), seven (7X) or 12 (12X) arginine (R) to alanine (A) mutations into the stable cell line containing a single ZF-127 binding site upstream of the minCMV promoter and measured GFP expression driven by 10ng (1X) or 20ng (2X) transfected DNA (Figure 4.4A ; cMyc promoter Figure S4.3). We use this 2X measurement as a control for doubling the amount of synTF DNA transfected when looking at pairs of TFs and note that in all cases, including the empty TF, expression rises slightly when the amount of DNA is doubled.

For each combination, we plot one synTF with affinity decreasing from left to right on the top horizontal and the second synTF with affinity decreasing from top to bottom along the vertical axis (Figure 4.4B). Along the diagonal, both synTFs have the same DNA binding domain. cMYC drove more expression with SP1 and HSF1 than with BRD4, but these TFs drive higher levels of expression than BRD4 on their own. cMYC also does not appear to have an affinity preference, as it drives similar levels with a WT or mutant DNA binding domain. BRD4 and HSF1, on the other hand, appear to drive the most expression with high affinity binding domains. SP1, surprisingly, seems to prefer weak binding, driving more expression as a 5, 7, or 12X mutant than with a WT domain. This assay highlights the utility of remaining agnostic in regards to traditional interpretations of TF function, and may enable us to class TFs by binding preferences as we use a similar assay for more activation domains.

Discussion

We built a platform to systematically interrogate combinatorial control by mammalian TFs in cell culture. The synthetic design of our experiments enabled us to recruit a pilot set of four TFs to a defined genomic location with high specificity and limited off-target effects. Our results suggest that TFs drive a range of expression from a single binding site at the minCMV promoter, and that when recruited in combination, pairs always increase expression compared to individual synTFs. We focused much of our efforts on experiments to dissect cooperativity at the level of function, using a single binding site, which revealed some support for kinetic synergy between TFs that work on initiation and elongation. Our results also suggest that different TFs have different affinity preferences for their binding sites, which may be a design principle of combinatorial control that warrants further study.

Importantly, these results do not rule out binding cooperativity, and one challenge looking forward is to use this empirical platform to look for combinatorial control both at the level of binding and the level of function. This could be done using alternative zinc-finger site organizations and by interrogating the behavior of many more TFs, including those that work on

steps other than initiation and elongation. In order to more rigorously interpret our results, we are currently pursuing several lines of experimentation, such as cellular measurements of synTF expression level and in vitro measurements of ZF affinity. With this pilot, we learned a great deal about the features of this experimental platform to consider when designing future studies, discussed below.

Design considerations for a medium-throughput screening platform

The small size of our pilot study (4 synTFs) was a deliberate decision. In order to investigate the multiple variables in the experimental design of this screening platform, we needed to test a small number of TFs individually and in pairs under multiple conditions: the number and spacing of binding sites needed to activate expression, different promoters driving the reporter gene, and the affinity of the binding domain for the synthetic site. These experiments, which require testing all pairwise combinations of TFs in triplicate, explode in combinatorial space quite quickly. With our four synTFs, we can test all pairwise combinations in six conditions (not including single TFs or other controls), which allowed us to systematically test other variables and compare the results within a single experiment. Our earliest experiments used 9 different synTFs, which required 45 wells of cells (135 wells to test in triplicate) to measure every pairwise combination. This matrix was too large to use for optimization, and testing different binding domain affinities and site arrangements was much simpler with a smaller panel.

These experiments provided a baseline set of expectations and logistical concerns for considering a medium-throughput screen. Scaling up, to say 50 synTFs, and testing them in all pairwise combinations would require experiments on the order of ~1225 combinations (not including individual TFs or controls), in triplicate. To test these pairs at two different affinities (WT-WT, WT-mutant, mutant-WT, mutant-mutant) adds three additional conditions per pair, making the number of cells and wells required to quantify interactions even higher. This type of

throughput is certainly possible with liquid handling and robotics, but it was important to first define the variables that matter for this type of screening platform.

Single molecule measurements of TF affinity

Our experiments using altered binding domains were designed to test the hypothesis that proteins fused to the "wild-type" synTF binding domain were not "time-sharing" efficiently because the off-rate of these proteins was too high. While we rationally mutated the binding domain of the synTFs in this study and show that reporter expression driven by transfected HSF1 decreased, we used expression as a proxy for affinity rather than measuring the binding kinetics of the synTF. This type of experiment is now possible with single molecule microscopy (Friedman and Gelles, 2012).

We have designed experiments to express and purify these DNA binding domains and experimentally determine their binding kinetics on labelled DNA. These experiments will address several outstanding questions, including whether the wild-type and mutated binding domains have different binding kinetics and if the on-rate is constant for all mutant proteins. These experiments are underway thanks to the hard work of Tim Harden and Qiu Wu.

Comparing results to existing models of TF behavior

One goal with these experiments was to be able to fit our results to existing computational models of gene regulation. The classical thermodynamic model, which assumes that TFs work together by facilitating a single step in transcription regulation (the recruitment of RNA polymerase), has been critical for understanding gene regulation in bacteria (Bintu et al., 2005; Ptashne and Gann, 1997). More recently, work from our lab and others put forward a kinetic model of transcription, whereby more than one step is required to produce a transcript, and activators work on distinct "slow" steps in this process (Scholes et al., 2016). Our experiments are uniquely suited to compare these models of transcriptional control because we have measurements of expression driven by each individual TF and all TF combinations from a

system where each input and output can be precisely manipulated and quantitatively measured. We are optimistic that this line of work, being carried out by Tim Harden, will identify pieces of data that are consistent with each of these models, as well as some that demonstrate where our existing models fail; highlighting the value of comparing models to experimental data and developing alternative approaches.

Moving into flies: considerations for future experiments

Ultimately, our goal is to use this experimental platform to test and compare the regulatory logic across different species, including yeast, flies, and mammalian cells. There are several advantages to moving into flies: we know a great deal about the activators that bind enhancers to regulate endogenous genes during early development, and there are a number of activator pairs that have been shown to work together in this process (Duarte et al., 2016; Tsurumi et al., 2011), Harden unpublished data). Single enhancer studies in flies have yielded a great deal of knowledge about how activators and repressors maintain the precision of gene expression during development (Jaeger, 2011; Janssens et al., 2006; Stanojevic et al., 1991) and the biochemical activities of many TFs are known (Arnosti et al., 1996; Bothma et al., 2011; Clyde et al., 2003; Liu and Ma, 2013; Schulz et al., 2015). In this system, we can screen TF pairs for synergistic expression effects in cell culture and move into the embryo with promising hits; thereby enabling us to ask how combinatorial gene regulation occurs under spatial and temporal constraints in an intact, developing organism. Finally, a great deal of technology exists in the blastoderm to titrate and label TF inputs and quantitatively measure the resulting expression with fixed or live imaging, making the *Drosophila* blastoderm an exciting future direction for this line of work (Bentovim et al., 2017; Bothma et al., 2014; Gregor et al., 2014).

Beyond "activator" and "repressor": towards a mechanistic view of TF function

At the outset, we wanted to ask whether TFs are organized into functional classes and if we can use these groups to learn about how their activities combine. The approach we developed

is a first step in answering these questions, as we distill animal transcription to a minimal system, and interrogate how transcriptional programs are combinatorially controlled. Our results from this study suggest that all four TFs activated the minCMV promoter, but did so in different ways and to different extents. We believe this platform will enable us to screen for functional interactions between many TFs, both that have been well-studied and that remain to be characterized. By learning about their biochemical activity alone and in combination, we hope to be able to elucidate design principles of regulatory circuits, and one day use this insight to engineer new ones by drawing on TFs with different behaviors.

Materials and Methods

Reporter design

Arrays of 1 or 2 binding sites were cloned upstream of the minimal cytomegalovirus (minCMV) promoter driving expression of EGFP with a rabbit globin 3'UTR. Two different zinc finger synTF domains were used in this study, zinc finger "A" (labelled '116' in our vector maps) recognizes a 20bp sequence (gGTCGTTGCGGTAGTCGAAg) and zinc finger "B" (labelled '127' in our vector maps) recognizes a separate 20bp sequence (cGGCGTAGCCGATGTCGCGc). These sequences were screened against the mammalian genome and designed to be most orthogonal, such that the synTFs should not bind anywhere else in the genome (see (Keung et al., 2014; Park et al., 2018) for details).

When required between binding sites, spacer sequences were computationally designed using the SiteOut tool (http://depace.med.harvard.edu/siteout) to not contain binding sites for 42 mammalian transcription factors. We selected the TFs to screen against based on motif data, pwm availability, and chose at least 1 representative member from 11 major classes of mammalian TFs included in the JASPAR database (Supplementary Table, Appendix C; http://jaspar.genereg.net/). SiteOut was run using standard parameters for spacer design, including a background GC content of 40.6% and a p-value cutoff of 0.003. To measure activation from a single binding site at various distances from the promoter, the yeast upstream activating sequence (UAS) was substituted for 20bp zinc-finger binding sequence. All plasmids were constructed using standard molecular biology techniques and gibson isothermal assembly. Constructs were transformed into E. coli alpha-select silver efficiency competent cells (Bioline) and sequence verified.

synTF design

SynTF proteins containing an activation domain of interest fused to an N-terminal zinc-finger binding domain with a GGGGS flexible linker were driven under control of a ubiquitin promoter and contain a 5' sv40 nuclear localization sequence, C-terminal HA or MYC tag (for

ZF-127 and 116, respectively) and rabbit globin polyA 3' UTR. ZF-127 and ZF-116 binding domains were selected after bioinformatically screening synTF target sequences against the mammalian genome. To tune the affinity of the zinc finger, 3, 5, or 7 arginine residues in the zinc finger array were mutated to alanine as denoted in Figure 4.3. The sequences used to generate the synTF fusions are listed in the Supplemental Information (Appendix C).

Genomic integration

Reporter lines were generated by site-specific integration of reporter constructs into 293T cells using CRISPR/Cas9 mediated homologous recombination into the AAVS1 (PPP1R2C) locus as previously described (Park et al., 2019). Briefly, 60,000 cells were plated in a 48-well plate and transfected the following day by PEI with a mixture of the following : 70ng of gRNA_AAVS1-T2 plasmid (Addgene 41820), 70 ng of VP12 humanSpCas9-Hf1 plasmid (Addgene 72247), and 175 ng of donor reporter plasmid. Donor reporter plasmids contain flanking arms homologous to the AAVS1 locus, a puromycin resistance cassette, and constitutive mCherry expression. After transfection, cells were cultured in 2 mg/mL puromycin selection for at least 2 weeks with splitting 1:10 every 3 days. Monoclonal populations for each reporter cell line were isolated by sorting single cells from this population into a 96 well plate and growing cell lines from each well. A minimum of 6 cell lines were transfected with a strong synTF activator (HSF1, SP1 or VP16) and a monoclonal cell line to be used going forward was based on the percentage of mCherry-positive cells and strength of GFP expression.

Cell culture and transfection

HEK293T cells stably expressing the reporter of interest were cultured in DMEM with L-glutamine, 4.5g/L Glucose and Sodium Pyruvate (DMEM, Thermo Fisher Scientific) supplemented with 10% FBS (Clontech), 1% GlutaMAX supplement (Thermo Fisher Scientific), 1% MEM Non-Essential Amino Acids solution (Thermo Fisher Scientific) and 1% penicillin-

streptomycin (Thermo Fischer Scientific) as described previously (Park et al., 2018). Cells were split every 3 days, and maintained at 37'C and 5% CO2 in a humidified incubator.

Stable cell lines containing the promoter of interest were transfected with synTF plasmid constructs using polyethylenimine (PEI) + NaCl as described in Park et al. (2019). 60,000 cells/ well were plated in 48-well plates and transfected the following day with a total of 10ng per activator + single stranded filler DNA to 200ng total. Media was changed after 4 hours. Cells were collected and prepared for flow cytometry two days post transfection, unless otherwise noted.

Flow cytometry and data analysis

For each measurement, cells were harvested and run on a BD LSRFortessa equipped with a high throughput auto sampler (BD Biosciences). A minimum of 10,000 events were collected for each well and were gated by forward and side scatter, as described in Park et al. (2019). The geometric mean of each fluorescence distribution was calculated in FlowJo. When necessary, fold-change GFP expression was determined from the fluorescence ratio of values from the reporter alone to the reporter transfected with the synTFs. Flow cytometer laser/filter configurations used in this study were: EGFP (488 nm, 510/10), mCherry (561 nm, 615/25), iRFP-720 (638 nm, 720/30). Data were plotted using Prism GraphPad software. All experiments performed in technical triplicate.

References

Adams, C.C., and Workman, J.L. (1995). Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. Mol. Cell. Biol. *15*, 1405–1421.

Arnold, C.D., Gerlach, D., Spies, D., Matts, J.A., Sytnikova, Y.A., Pagani, M., Lau, N.C., and Stark, A. (2014). Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nat. Genet. *46*, 685–692.

Arnosti, D.N., Gray, S., Barolo, S., Zhou, J., and Levine, M. (1996). The gap protein knirps mediates both quenching and direct repression in the Drosophila embryo. EMBO J. *15*, 3659–3666.

Beagrie, R.A., and Pombo, A. (2016). Gene activation by metazoan enhancers: Diverse mechanisms stimulate distinct steps of transcription. Bioessays.

Bentovim, L., Harden, T.T., and DePace, A.H. (2017). Transcriptional precision and accuracy in development: from measurements to models and mechanisms. Development *144*, 3855–3866.

Bintu, L., Buchler, N.E., Garcia, H.G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005). Transcriptional regulation by the numbers: models. Curr. Opin. Genet. Dev. *15*, 116–124.

Blau, J., Xiao, H., McCracken, S., O'Hare, P., Greenblatt, J., and Bentley, D. (1996). Three functional classes of transcriptional activation domain. Mol. Cell. Biol. *16*, 2044–2055.

Bothma, J.P., Magliocco, J., and Levine, M. (2011). The snail repressor inhibits release, not elongation, of paused Pol II in the Drosophila embryo. Curr. Biol. *21*, 1571–1577.

Bothma, J.P., Garcia, H.G., Esposito, E., Schlissel, G., Gregor, T., and Levine, M. (2014). Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living Drosophila embryos. Proc. Natl. Acad. Sci. U. S. A. *111*, 10598–10603.

Brown, S.A., Weirich, C.S., Newton, E.M., and Kingston, R.E. (1998). Transcriptional activation domains stimulate initiation and elongation at different times and via different residues. EMBO J. *17*, 3146–3154.

Carey, M. (1998). The enhanceosome and transcriptional synergy. Cell *92*, 5–8.

Chi, T., Lieberman, P., Ellwood, K., and Carey, M. (1995). A general mechanism for transcriptional synergy by eukaryotic activators. Nature *377*, 254–257.

Clyde, D.E., Corado, M.S.G., Wu, X., Paré, A., Papatsenko, D., and Small, S. (2003). A self-organizing system of repressor gradients establishes segmental complexity in Drosophila. Nature *426*, 849–853.

Crocker, J., Abe, N., Rinaldi, L., McGregor, A.P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., et al. (2015). Low affinity binding site clusters confer hox specificity and regulatory robustness. Cell *160*, 191–203.

Duarte, F.M., Fuda, N.J., Mahat, D.B., Core, L.J., Guertin, M.J., and Lis, J.T. (2016). Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation. Genes Dev. *30*, 1731–1746.

Eberhardy, S.R., and Farnham, P.J. (2001). c-Myc mediates activation of the cad promoter via a post-RNA polymerase II recruitment mechanism. J. Biol. Chem. *276*, 48562–48571.

Eberhardy, S.R., and Farnham, P.J. (2002). Myc recruits P-TEFb to mediate the final step in the transcriptional activation of the cad promoter. J. Biol. Chem. *277*, 40156–40162.

Emili, A., Greenblatt, J., and Ingles, C.J. (1994). Species-specific interaction of the glutamine-rich activation domains of Sp1 with the TATA box-binding protein. Mol. Cell. Biol. *14*, 1582–1593.

Estrada, J., Ruiz-Herrero, T., Scholes, C., Wunderlich, Z., and DePace, A.H. (2016). SiteOut: An Online Tool to Design Binding Site-Free DNA Sequences. PLoS One *11*, e0151740.

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

Farley, E.K., Olson, K.M., Zhang, W., Rokhsar, D.S., and Levine, M.S. (2016). Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proc. Natl. Acad. Sci. U. S. A. *113*, 6508–6513.

Fishburn, J., Mohibullah, N., and Hahn, S. (2005). Function of a eukaryotic transcription activator during the transcription cycle. Mol. Cell *18*, 369–378.

Friedman, L.J., and Gelles, J. (2012). Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation. Cell *148*, 679–689.

Fuda, N.J., Ardehali, M.B., and Lis, J.T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature *461*, 186–192.

Gertz, J., Siggia, E.D., and Cohen, B.A. (2009). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature *457*, 215–218.

Gill, G., Pascal, E., Tseng, Z.H., and Tjian, R. (1994). A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the Drosophila TFIID complex and mediates transcriptional activation. Proceedings of the National Academy of Sciences *91*, 192–196.

Gisselbrecht, S.S., Barrera, L.A., Porsch, M., Aboukhalil, A., Estep, P.W., 3rd, Vedenko, A., Palagi, A., Kim, Y., Zhu, X., Busser, B.W., et al. (2013). Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. Nat. Methods *10*, 774–780.

Gregor, T., Garcia, H.G., and Little, S.C. (2014). The embryo as a laboratory: quantifying transcription in Drosophila. Trends Genet. *30*, 364–375.

Guertin, M.J., Petesch, S.J., Zobeck, K.L., Min, I.M., and Lis, J.T. (2010). Drosophila heat shock system as a general model to investigate transcriptional regulation. Cold Spring Harb. Symp. Quant. Biol. *75*, 1–9.

GuhaThakurta, D., and Stormo, G.D. (2001). Identifying target sites for cooperatively binding factors. Bioinformatics *17*, 608–621.

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

Herschlag, D., and Johnson, F.B. (1993). Synergism in transcriptional activation: a kinetic view. Genes Dev. *7*, 173–179.

Itzen, F., Greifenberg, A.K., Bösken, C.A., and Geyer, M. (2014). Brd4 activates P-TEFb for RNA polymerase II CTD phosphorylation. Nucleic Acids Res. *42*, 7577–7590.

Jaeger, J. (2011). The gap gene network. Cell. Mol. Life Sci. *68*, 243–274.

Janssens, H., Hou, S., Jaeger, J., Kim, A.-R., Myasnikova, E., Sharp, D., and Reinitz, J. (2006). Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. Nat. Genet. *38*, 1159–1165.

Keung, A.J., Bashor, C.J., Kiriakov, S., Collins, J.J., and Khalil, A.S. (2014). Using targeted chromatin regulators to engineer combinatorial and spatial transcriptional regulation. Cell *158*, 110–120.

Khalil, A.S., Lu, T.K., Bashor, C.J., Ramirez, C.L., Pyenson, N.C., Joung, J.K., and Collins, J.J. (2012). A synthetic biology framework for programming eukaryotic transcription functions. Cell *150*, 647–658.

Kinney, J.B., and Atwal, G.S. (2014). Parametric inference in the large data limit using maximally informative models. Neural Comput. *26*, 637–653.

Kinney, J.B., Murugan, A., Callan, C.G., Jr, and Cox, E.C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc. Natl. Acad. Sci. U. S. A. *107*, 9158–9163.

Latchman, D.S. (1997). Transcription factors: an overview. Int. J. Biochem. Cell Biol. *29*, 1305–1312.

Liu, J., and Ma, J. (2013). Dampened regulates the activating potency of Bicoid and the embryonic patterning outcome in Drosophila. Nat. Commun. *4*, 2968.

McKenna, N.J., and O'Malley, B.W. (2002). Combinatorial control of gene expression by nuclear receptors and coregulators. Cell *108*, 465–474.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr, Kinney, J.B., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat. Biotechnol. *30*, 271.

Mirny, L.A. (2010). Nucleosome-mediated cooperativity between transcription factors. Proc. Natl. Acad. Sci. U. S. A. *107*, 22534–22539.

Nechaev, S., and Adelman, K. (2011). Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. Biochim. Biophys. Acta *1809*, 34–45.

Park, M., Patel, N., Keung, A.J., and Khalil, A.S. (2018). Engineering Epigenetic Regulation Using Synthetic Read-Write Modules. Cell.

Porter, W., Saville, B., Hoivik, D., and Safe, S. (1997). Functional synergy between the transcription factor Sp1 and the estrogen receptor. Mol. Endocrinol. *11*, 1569–1580.

Ptashne, M., and Gann, A. (1997). Transcriptional activation by recruitment. Nature *386*, 569–577.

Scholes, C., DePace, A.H., and Sánchez, Á. (2016). Combinatorial Gene Regulation through Kinetic Control of the Transcription Cycle. Cell Syst.

Schulz, K.N., Bondra, E.R., Moshe, A., Villalta, J.E., Lieb, J.D., Kaplan, T., McKay, D.J., and Harrison, M.M. (2015). Zelda is differentially required for chromatin accessibility, transcription factor binding, and gene expression in the early Drosophila embryo. Genome Res.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530.

Smith, A.D., Sumazin, P., and Zhang, M.Q. (2005). Identifying tissue-selective transcription factor binding sites in vertebrate promoters. Proc. Natl. Acad. Sci. U. S. A. *102*, 1560–1565.

Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet. *45*, 1021–1028.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Staller, M.V., Holehouse, A.S., Swain-Lenz, D., Das, R.K., Pappu, R.V., and Cohen, B.A. (2018). A High-Throughput Mutational Scan of an Intrinsically Disordered Acidic Transcriptional Activation Domain. Cell Syst *6*, 444–455.e6.

Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature.

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

Stavreva, D.A., Coulon, A., Baek, S., Sung, M.-H., John, S., Stixova, L., Tesikova, M., Hakim, O., Miranda, T., Hawkins, M., et al. (2015). Dynamics of chromatin accessibility and long-range interactions in response to glucocorticoid pulsing. Genome Res. *25*, 845–857.

Struhl, K. (1991). Mechanisms for diversity in gene expression patterns. Neuron *7*, 177–181.

Tsong, A.E., Tuch, B.B., Li, H., and Johnson, A.D. (2006). Evolution of alternative transcriptional circuits with identical logic. Nature *443*, 415–420.

Tsurumi, A., Xia, F., Li, J., Larson, K., LaFrance, R., and Li, W.X. (2011). STAT is an essential activator of the zygotic genome in the early Drosophila embryo. PLoS Genet. *7*, e1002086.

Veitia, R.A. (2003). A sigmoidal transcriptional response: cooperativity, synergy and dosage effects. Biol. Rev. Camb. Philos. Soc. *78*, 149–170.

Voss, T.C., and Hager, G.L. (2014). Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat. Rev. Genet. *15*, 69–81.

Wagner, A. (1999). Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. Bioinformatics *15*, 776–784.

Weingarten-Gabbay, S., Nir, R., Lubliner, S., Sharon, E., Kalma, Y., Weinberger, A., and Segal, E. Deciphering Transcriptional Regulation of Human Core Promoters.

Wunderlich, Z., and Mirny, L.A. (2009). Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet. *25*, 434–440.

Xiao, H., Perisic, O., and Lis, J.T. (1991). Cooperative binding of Drosophila heat shock factor to arrays of a conserved 5 bp unit. Cell *64*, 585–593.

Yankulov, K., Blau, J., Purton, T., Roberts, S., and Bentley, D.L. (1994). Transcriptional elongation by RNA polymerase II is stimulated by transactivators. Cell *77*, 749–759.

Chapter 5: Discussion

Overview

In this dissertation, I described our work using the enhancers of *eve* to challenge the canonical definition of enhancers as modular units using computational models and quantitative experiments in Drosophila embryos. In Chapter 2, we found that the pattern driven by fusions of eve37 and eve46 suggests that interactions between enhancers occur, and that these regulatory sequences are read as two modules and then integrated by the promoter. In Chapter 3, we showed that rearrangements and deletions of eve enhancers demonstrate that there are functional interactions between enhancers and with other regulatory sequences that are important for the precision of eve expression during Drosophila development. Finally, in Chapter 4, we investigated combinatorial gene regulation by multiple activators using a fully synthetic system in mammalian cell culture.

One long-term goal of this work is to be able to accurately predict gene expression from sequence. Our results argue that to do this well, we need to consider the full locus context surrounding a developmental gene. This may include interactions between enhancers, the effect of modifying sequences, the role of DNA topology, and interactions between TFs. A mechanistically motivated model, based on the *eve* case study, will complement statistical models of gene expression based on single enhancers or genome wide data. We are optimistic that things we learn from studies with *eve* will motivate the types of sequences and interactions people go looking for in future genome wide work. Here, I describe specific experimental directions that arise from this thesis, using the *eve* locus (Part I) and our synthetic platform (Part II) to make progress on the phenomenology and mechanism of locus scale gene regulation.

Part I : What is the nature of enhancer interactions?

While people have long thought of enhancers as acting modularly, my results from Chapters 2 and 3 show that this simple framework does not explain the results of moving eve37 and eve46 around relative to each other or the effect of deleting eve37 from the endogenous

locus. Based on our results, there is likely a spectrum of enhancer behavior, ranging from fully modular to involved in complex interactions, where some enhancers are more critical for the overall output of a locus or interact with modifying sequences to greater extent.

In the most common cartoon used to illustrate enhancer function, an "active" enhancer is depicted as looped to the promoter while "inactive" enhancers (that do not drive expression in these cells) are not considered a part of this set of functional interactions (Figure 5.1, left). Data from my thesis and other labs argues that this view is too simple. One important question then is, given what we know now, how do we draw a better cartoon? Can we write down the physical picture we have in our heads so that we can dissect how it does and doesn't fit with our data? While all cartoons are inherently flawed, given that they are static, two dimensional representations of a dynamic, three dimensional process, the experiments below bring us closer to a useful cartoon in two ways. First, what sequences should we include? And second, what can we learn about the proteins involved that define the features of these interactions?

Recent work from the field is coalescing around the framework of transcriptional "hubs," which incorporates many of the observations of non-modular enhancer behavior, such as enhancer interactions and clusters of low-affinity binding sites (reviewed in (Furlong and Levine, 2018). In this cartoon, clusters of highly active enhancers are co-localized in the cell with high levels of RNAP, cofactors, and TFs at sites of active transcription (Figure 5.1, right). This model gives a different function to "inactive" enhancers that are accessible at the same developmental stage; one hypothesis is that they increase the local concentration of relevant TFs that other enhancers may use. There is some limited support for this hypothesis in the literature (Chong et al., 2018; Crocker et al., 2015; Mir et al., 2017; Tsai et al., 2017) but the evidence thus far has been from single molecule microscopy of tagged TFs and sites of active transcription.

When we consider the *eve* locus, our data suggests that these interactions occur over long range, and are enhancer specific; a cartoon where there are arrows between each of the enhancers is also too simple. We know from our deletion experiments that the eve46 deletion doesn't seem to matter for the precision of stripes driven by other enhancers, but deletion of

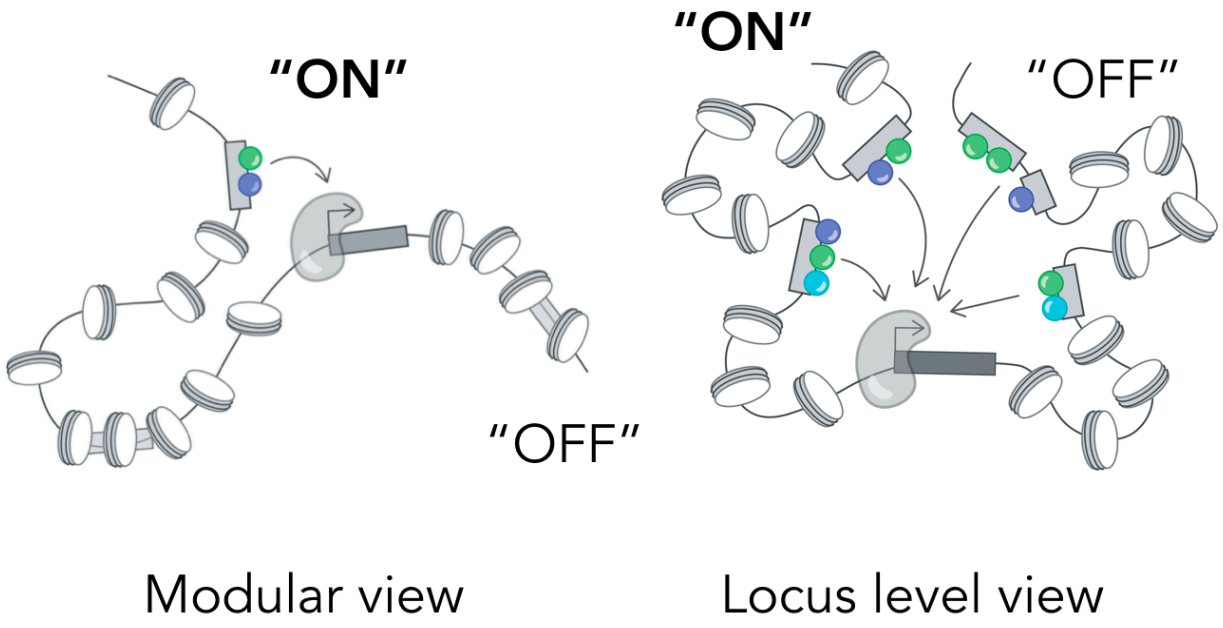


Figure 5.1 : Cartoons of enhancer function. (Left) In this common cartoon illustrating modular enhancer function, the "active" enhancer is looped to the promoter through a large chromatin loop and inactive enhancers are not considered as part of the set of functional interactions. (Right) In considering "locus-level" gene regulation, "inactive" enhancers may contribute to the precision of gene expression, and are likely in close proximity to other enhancers and the promoter. While this cartoon might be closer, we still don't have enough information about specific interactions between enhancers; one of the goals of our future work is to draw and validate arrows between them for eve. Cartoons by Olivia Foster Rhodes.

eve37 affects stripes 2, 4, and 5, to different degrees at early time points (Figure 3.5). Two immediate questions arise : What sequences besides canonical enhancers are important for the precision of *eve* expression? And second, what are the functional consequences of these interactions, and how do they change when an enhancer is deleted? Below, I outline the hypotheses we are considering and some experiments to test them. These experiments were designed in collaboration with Olivia Foster Rhodes and Drs. Angela DePace and Ed Pym.

Which sequences outside the annotated enhancers affect their function?

Our results from reporters suggest that there are sequences outside of the minimized blastoderm enhancers that play a role in the precision of gene expression in the *eve* locus. A great deal of progress has been made in understanding how individual enhancers contribute to the precision and robustness of the overall *eve* pattern, but much less work has been done to understand the role of sequences between and around enhancers that make up the full developmental locus. These sequences may have modifying activities that contribute to the overall pattern, but have been historically missed in studies using reporter assays because they may not drive strong expression on their own. We can first add candidate sequences back to our existing reporter constructs and ask whether these regions modulate the pattern driven by "active" enhancers. Then, using a complementary approach, we can then go back to the whole locus and delete regions that seem like they may have a modifying effect.

In Chapter 3, we showed that the eve46 enhancer drove much higher levels of expression when placed 5' of the reporter gene compared to the 3' position, despite the 3' end being the endogenous orientation of the enhancer relative to the promoter (Figure 3.4). When we place the eve37 enhancer in the 5' position and keep eve46 3' (again representing the endogenous situation), expression driven by eve46 is further reduced. One explanation for this observation is that the eve46 enhancer requires more regulatory information to drive the proper pattern when located 3' of the gene, which may not have been appreciated in qualitative reporter experiments. This situation provides a nice test case for experiments to study the role of "modifying" sequences in eve expression.

We have identified 3 candidate sequences that may contain this missing regulatory information: 1) the region immediately 5' of the eve46 enhancer, 2) the region immediately 5' of the eve TSS, and 3) the two insulator sequences (Homie and Nhomie) that flank the eve locus. In early characterizations of eve, the region upstream of the eve46 enhancer was identified as "stripes 4 and 6 enhancement" (Sackerson et al., 1999) but wasn't studied after eve46 was minimized in (Fujioka et al., 1999). The region between the eve TSS and the stripe 2 enhancer

(approximately 1.1kb) contains a great deal of potential regulatory information that is often left out of studies using the 42bp minimized eve promoter (Small et al., 1992). Specifically, this region contains many footprinted binding sites, sites for architectural and looping factors (Chetverina et al., 2017; Cubeñas-Potts et al., 2017; Postika et al., 2018) and is significantly conserved across fly species (Berman et al., 2004). The two insulators that flank *eve*, Homie and Nhomie, have only recently been studied (Fujioka et al., 2009, 2013, 2016). Previous work in *eve* showed that deletion of Homie from a large transgene resulted in reduced expression of *eve* stripes 4 and 6 (Fujioka et al., 2013), and one could imagine that these sequences, which are thought to be important for the overall topology of the locus, are necessary for bringing 3' sequences in to contact with the promoter.

We can identify whether these sequences improve eve46 expression from the 3' position by adding them back to existing reporters containing eve46 alone and with the eve37 enhancer on the 5' side. Quantifying the effects on pattern by LacZ in situ and comparing them to existing reporters lacking these sequences will demonstrate any "modifying" behavior that these sequences might have on expression of eve46. It would also be interesting to consider if other 3' enhancers (stripe 1 and stripe 5) have similar requirements for these extra sequences to function from a downstream position.

Can we identify modifying sequences with endogenous deletions?

The complementary approach to the experiments suggested above in reporter constructs is deleting candidate sequences from the endogenous locus and measuring their effects on *eve* expression. While the throughput is lower, using this approach for *eve* enhancers provided critical information that experiments in reporters could not: specifically, the presence of the eve46 enhancer impacted the function of eve37 in reporter constructs, but stripes 3 and 7 were fine when eve46 was deleted from the endogenous locus. We have a set of candidate sequences that would be interesting for deletion experiments, as evidence from the literature suggests they may be important for the overall precision of the *eve* pattern during blastoderm embryogenesis.

There are two ways to go about this experiment: in the endogenous locus with balancer chromosomes (as in Chapter 3) or in a line containing a bacterial artificial chromosome (BAC) from *Drosophila willistoni* (*D. wil*), which can rescue homozygous eve null flies with 50% survival. Doing these experiments in the line rescued by the BAC reduces possible network effects from changes in eve expression and sequence divergence between these orthologs allows for specific genome engineering and mapping of reads in true breeding lines, which is critical for genomic techniques. We can replace each region of interest using neutral DNA of an equivalent length designed using SiteOut (Estrada et al., 2016) and scarless genome engineering with dsRed as a selectable marker (Gratz et al., 2015; Lamb et al., 2017). The effect of replacements can be studied using in situ against the endogenous eve expression pattern, or using MS2 with loops added to the eve 3' UTR (flies from (Lim et al., 2018) are in the lab).

We identified a set of initial candidate regions to delete from the *eve* locus: 1) the late enhancer, 2) the region 5' of the eve2 enhancer, 3) the region 5' of the eve46 enhancer (described above) and 4) Homie and Nhomie, the insulator sequences flanking *eve* (also described above). We are particularly interested in deleting the late enhancer because it drives *eve* expression on its own in the late blastoderm and during gastrulation, and may functionally interact with the stripe enhancers during earlier temporal windows of eve expression (Fujioka et al., 1995; Sackerson et al., 1999). It is thought to narrow the stripes driven by stripe enhancers (Fujioka et al., 2002; Schroeder et al., 2011), suggesting this element has modifying activity, but we do not understand how (or if) this "handoff" occurs. The late enhancer may also explain results from our enhancer deletions, as we do not see embryos corresponding to late stage 5 that are missing *eve* stripes, even though ¼ of embryos in the experiment should have this genotype, suggesting that the late enhancer rescues the stripe deletion phenotype. We expect that a deletion of the late enhancer will not impact the formation of 7 stripes in the early blastoderm, but will have overall or stripe-specific effects on the *eve* pattern by not narrowing stripes in late stage 5, and disappearing during gastrulation (based on (Fujioka et al., 1995; Manoukian and Krause, 1993; Sackerson et al., 1999).

The sequence immediately 5' of the eve2 enhancer is interesting for a couple of reasons. First, Steve Small showed that this region drives slightly higher expression of *eve* stripe 2, suggesting it may be another "enhancer of an enhancer" modifying region (Small et al., 1992) and Michael Ludwig showed that this region is important for robustness of stripe 2 to environmental perturbation (Ludwig et al., 2011). Second, work from our lab and others showed that chromatin in the 500bp upstream of eve2 is accessible (Haines and Eisen, 2018) and bound by the pioneer factor Zelda (Harrison et al., 2011) and KB, unpublished data). Given what is known about the role of this sequence in stripe 2 regulation, we expect the level of stripe 2 expression to decrease when it is deleted. If this sequence has a role in locus-level *eve* regulation, such as in establishing a favorable chromatin state, the position, level or onset timing of stripes driven by other enhancers may also change.

As discussed above, the region between eve46 and the coding sequence is thought to improve eve46 expression, which we would expect to see in this experiment as decreased expression of stripes 4 and 6 when this sequence is deleted. It will be interesting to see if other stripes also change with this deletion, which we would not expect if the function of this sequence is as a modifier of eve46, but would expect if this sequence is important in some overall feature of the locus, such as establishing a favorable topology for the downstream enhancers to contact the *eve* promoter region. This is a large region (~3.3kb), so smaller deletions may be needed to follow-up if any effects are observed.

Finally, the eve insulators Homie and Nhomie, discussed above, have been shown to interact in cis and separate the *eve* regulatory sequence into a single topologically associated domain (Fujioka et al., 2013, 2016). Homie has been deleted from a rescue transgene, which reduced expression of all stripes driven by 3' enhancers: stripe 1, 4, 5 and 6 (Fujioka et al., 2013) but was missing Nhomie, motivating individual and combined deletions of these sequences to determine how they work together to facilitate proper *eve* expression.

Do other eve enhancers behave similarly to deletion of eve37 or eve46?

Finally, it would be useful to delete additional stripe enhancers to map the functional relationships between them. For example, we see that deletion of eve37 significantly altered stripes 2, 4, and 5 but eve46 did not matter for the other enhancers. Work from other labs has deleted *eve* stripe 1 (Lim et al., 2018) and saw a change in the position and onset timing of *eve* stripe 2. Michael Ludwig and Marty Kreitman replaced the eve2 enhancer with mini-white, and when we stained this line for *eve* mRNA, we saw effects on the position of stripes 3 and 7 and the level of stripe 6 ((Ludwig et al., 2011) Appendix B, Figure S3.5). The Eisen lab also replaced eve2 with another piece of sequence, in this case a fragment of bacterial DNA from an ampicillin resistance cassette (Li and Eisen), and did not observe effects on the position or level of other *eve* stripes. However, in this experiment they also replaced eve1, and show representative images from in situ experiments without quantitative analysis, so there may be subtle effects.

It would be useful to delete the remaining stripe enhancers from the endogenous locus (rescued by the BAC, as above) and measure the effects on the overall *eve* pattern with high temporal resolution. While the effects of deleting the enhancers that drive stripes 1 and 2 have been measured, doing a clean deletion (rather than bacterial sequence replacement) of eve2, as well as eve5 and the late element would uncover relationships between the blastoderm enhancers and flesh out the spectrum of interactive behavior we observe in this locus (i.e. eve37 seems to interact with other enhancers while eve46 does not). Finally, it would be interesting to delete *eve* enhancers that are active at different times, such as the late element (discussed above), muscle enhancer (between eve46 and eve1) or one of the neuronal enhancers (Fujioka et al., 1999) and look for effects on patterning in the blastoderm. If deleting one of these enhancers changes the stripes driven by canonical stripe enhancers, this would be additional evidence of non-modular behavior and locus-level control of *eve* expression. Additional information about what enhancers interact with one another will inform the underlying mechanism of these interactions. For example, interactions may occur between enhancers that are regulated by the same TFs, or between enhancers on the same side of the locus, implicating topology. Our goal is

to be able to draw arrows between enhancers in the locus, much like we've drawn arrows between TFs in regulatory networks, such that we can better understand the rules that govern these interactions in eve and other developmental loci.

Can we use genomic techniques to investigate the mechanism of enhancer interactions?

When considering this cartoon of locus-level gene regulation in *eve*, one question we thought about for a long time was how the picture of many enhancers interacting with each other and the promoter might change when an enhancer is deleted. In the case of the eve46 deletion, it seems that the only cells that are affected are the ones where eve46 is looped to the promoter and driving expression, because expression pattern or level does not seem to change in cells where *eve* is driven by other enhancers (Figure 3.5). However, when eve37 is lost, cells that normally do not express *eve* at all are turned on (stripe boundaries are shifted) and the level of *eve* expression in stripes 2, 4, 5, and 6 is also affected. We have considered several classes of mechanism to explain the functional interactions that we observe (Figure 5.2).

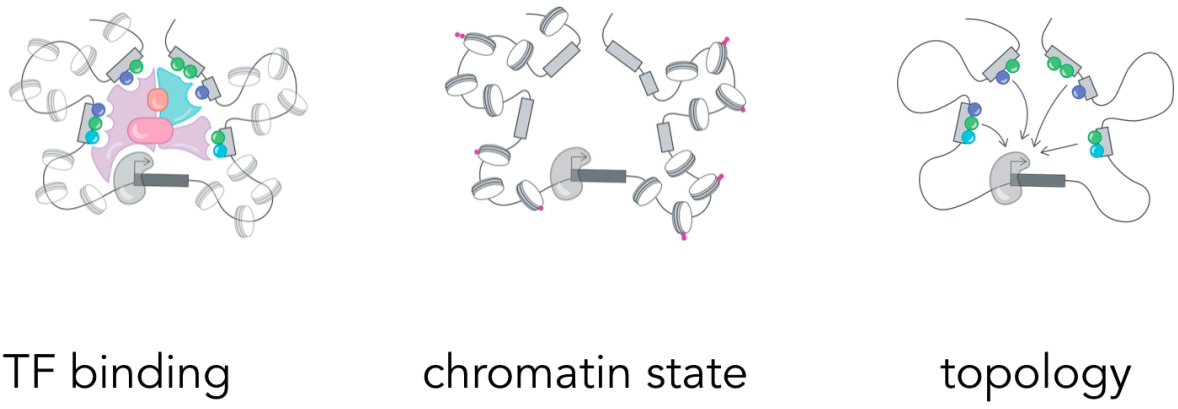


Figure 5.2 : Classes of hypotheses that could explain enhancer interactions in the eve locus. It is possible that eve enhancers interact to share regulators, maintain a favorable chromatin state, or to remain in close physical proximity to the promoter. Importantly, these hypotheses are not mutually exclusive. Cartoon by the talented Olivia Foster Rhodes.

First, *eve* enhancers may interact with other enhancers by sharing TFs. This interaction could be at the level of specific TFs or cofactors, which are shared by some, but not all enhancers. This mechanism could explain differences in the position or level of the stripes changing when eve37 is deleted, depending on if the factors are activators or repressors. Second, eve37 could be important for establishing a favorable chromatin conformation. This hypothesis is especially appealing given that this enhancer deletion primarily drives changes in stripe level, and is the 5' most enhancer in the locus. Finally, eve37 deletion could alter the physical location of other enhancer sequences, and change the topology of enhancer-enhancer and enhancer-promoter interactions.

Several years ago, we attempted to investigate the first two of these hypotheses in the eve46 deletion line using ChIP-seq and ATAC-seq. We chose to do this with eve46 for two reasons. First, we had done an initial characterization of the effects of deleting each of these enhancers and had preliminarily observed effects on the stripe 5 boundary in eve46 deletion flies. We hadn't imaged enough embryos to break our results down by time point, and so the shift we saw was because a disproportionate number of our embryos lacking stripes 4 and 6 were from the earliest time points, skewing the pattern towards the posterior of the embryo. Second, the eve46 enhancer deletion was not lethal, whereas the eve37 deletion was (eve46 flies survived without the balancer chromosomes). We thought that we could generate a homozygous eve46 deletion line from this resulting population so that we could collect embryos in which every progeny lacked both copies of the eve46 enhancer, rather than needing to genotype progeny from a balanced line (GFP balancers were not yet available). In our pilot study, we assayed PolII and Zelda occupancy using ChIP-seq and ChIP-nexus, and did not find any changes in the *eve* locus with the eve46 deletion. We also checked for any changes in chromatin accessibility with ATAC-seq and H3K27ac ChIP-seq, and did not find any evidence that chromatin state was changing.

These experiments were flawed for many reasons, but forced us to think about the challenges of looking for small effects in a small number of cells. The eve46 homozygotes were sickly and did not lay well, meaning that we could not get enough embryos from them to do genomics (which require a large amount of starting material). We attempted to mitigate this by crossing these lines to the line containing the *D. wil* BAC, which rescued the viability defects by including a second copy of the gene locus but made downstream analyses difficult as the genetic background of the control and eve46 deletion lines was now different. Second, only after returning from the Stowers Institute and carefully imaging many more embryos, did we realize that the effect of deleting eve46 is really only on stripes 4 and 6, rather than on the positioning of stripes driven by other enhancers. Ed Pym and Lauren Bush are developing GFP-marked balancer chromosomes, so that we will know the genotype of each embryo on the slide in future imaging experiments, rather than using the *eve* stripe pattern to determine which embryos are homozygous for the enhancer deletion.

This is a long winded justification of why we need a spatially resolved genomics assay in the blastoderm : to be able to make stripe-by-stripe comparisons of TF binding, chromatin accessibility, and physical interactions in the eve locus. Other groups have made progress towards this goal, work from the Eisen lab sliced embryos in half and compared chromatin accessibility from the anterior and posterior ends (Haines and Eisen, 2018). They saw that chromatin state is similar, but some subtle differences can be observed that correlate with known expression differences.

Can spatially resolved genomics improve our picture?

We are implementing the INTACT (Isolation of Nuclei TAgged in specific Cell Types) method (Deal and Henikoff, 2010; Steiner et al., 2012) in the blastoderm to isolate nuclei from stripes 1, 2, and 5 by expressing an nuclear tag under control of the specific stripe enhancers. The INTACT cassette contains a biotin ligase recognition peptide (BLRP) and a FLAG tag for affinity purification with the *eve* 3'UTR for spatial localization (Davis and Ish-Horowicz, 1991;

Macdonald et al., 1986). We will express stripe specific cassettes, dissociate the embryos and purify nuclei, and determine spatial enrichment by RNA-seq, as each stripe should have a unique spatial expression profile (Karaiskos et al., 2017). This work is being done in collaboration with Ed Pym, who has designed the INTACT cassette and shown that it drives patterned expression of *eve* stripes 1, 2, or 5, depending on the enhancer.

With this approach, we can measure the chromatin accessibility, topological organization, or TF occupancy in single *eve* stripes. This allows us to test each of the hypotheses about the functional consequences of enhancer interactions in turn. Current ChIP-seq datasets are from whole embryo experiments, which average gap gene binding across all *eve* enhancers; this binding does not match their spatial expression profile (ex. Stripe 1 enhancer is bound by Kruppel, but Kruppel is not expressed in the anterior of the embryo, (Li et al., 2008). We can also ask whether the chromatin accessibility in the *eve* locus is different in cells expressing a particular *eve* stripe compared to regions expressing a different stripe using ATAC-seq.

Finally, the experiment I am most excited about is measuring enhancer-enhancer and enhancer-promoter interactions in single *eve* stripes. Measuring DNA topology has led to numerous insights about gene regulation, and is a cornerstone feature of the idea of modular enhancers: the active ones are close to or physically touching the promoter while "inactive" enhancers are not (Benabdallah and Bickmore, 2015; Chen et al., 2018). We have hesitated to measure DNA topology in the *eve* locus because it is approximately 16kb from end to end, and the resolution of these approaches has not been high enough for enhancers that are only 1-2kb apart. However, new technology including micro-C or careful design of 4-cutter enzyme combinations suggests resolution in the 100-500bp range should be attainable (Hsieh et al., 2015; Singh and Hampsey, 2014). We have also designed 100bp capture oligos tiling the 37 TFs expressed in the blastoderm with 70bp overlaps (look in the -20 for a silver envelope labelled "Custom Array") to focus our sequencing depth on regions of interest at this developmental stage (Davies et al., 2016; Tyssowski et al., 2018). We are especially interested in the

conformation of the locus in each stripe, and whether the “inactive” stripe enhancers are in close proximity to the promoter or the “active” enhancer.

Part II : Using synthetic enhancers to investigate combinatorial control by multiple TFs

The experiments described in Chapter IV were a pilot for using synTFs to investigate combinatorial control in mammalian cells. We learned a great deal from these experiments about which variables matter, how they can be tuned, and what we would change going forward. Several exciting directions have emerged from this work, including moving into the Drosophila blastoderm and testing additional combinations of TFs known to co-regulate endogenous genes. Importantly, we don’t need to distinguish between the thermodynamic and kinetic world views, but aim to use an empirical platform to learn more about how transcription is controlled by more than one TF and to be able to contextualize these results with existing (or new!) models of transcript regulation.

A significant challenge to this line of experiments is the combinatorial space of TF-TF and TF-promoter interactions. Our pilot experiments suggest that a targeted characterization of more promoters and more TFs will help us determine the context dependence of our assays. For example, if we are looking for evidence that TFs interact by working on different functional steps, we need to be measuring expression driven by a promoter that has more than one slow rate. Further, investigating combinatorial control by TFs that are known to work together to co-regulate endogenous genes might yield insights into this process faster than by combining TFs that aren’t normally present together at endogenous promoters. Our experiments will define the critical role of the promoter in gene regulation and the biochemical role of TFs at different promoters. These experiments were designed in collaboration with Drs. Angela DePace and Clarissa Scholes.

Can we detect evidence that promoters are regulated at more than one step?

One challenge to understanding how TFs work together is that the sequence elements present in any promoter likely dictate the steps required to produce an mRNA and the relative rates of those steps (Kwak et al., 2013; Lubliner et al., 2015; Ohler and Wassarman, 2010). Ideally, we could just measure the rates of a complete transcription cycle for a given promoter, but the technology to make those measurements is only just becoming available in bacteria (Friedman and Gelles, 2012, 2015) and are not yet tractable in animal cells. However, it is possible to characterize the transcription dynamics of a panel of promoters using live imaging of nascent transcription (Choubey et al., 2015) and use the mean and variance of mRNA production to distinguish between promoters that have only one, or more than one, regulated step in transcript production (Choubey et al., 2015; Corrigan et al., 2016; Larson et al., 2011).

Given that we are excited about moving this work into flies, and a great deal is known about Drosophila developmental genes and their promoters (FitzGerald et al., 2006; Haberle and Lenhard, 2016; Ohler et al., 2002; Zabidi et al., 2015), we will take advantage of published RNAP occupancy data (Core et al., 2012; Gaertner et al., 2012; Kwak et al., 2013; Muse et al., 2007) to identify a set of candidate promoters that may have complex regulation for additional study. We will select promoters that drive patterned expression (Fowlkes et al., 2008), and that have RNAP build-up at the transcription start site or 5' region, indicating initiation is rate-limiting, or ~30-50bp downstream of the TSS, indicating promoter-proximal pausing. Importantly, these experiments require synTFs for the Drosophila genome, which the Khalil lab is successfully developing.

We can use CRISPR/Cas9 to integrate promoters of interest (~3-5 seems to be the right number to start with) driving an 5' MS2- 3' PP7 reporter into S2 cells (Figure 5.3). Including both MS2 and PP7 on opposite ends of the reporter allows us to measure the elongation rate of the polymerase (Day et al., 2016). Then, we can transfect synTFs of interest,

beginning with HSF1 (a highly conserved, moderate-strength activator) and measure nascent transcription in single cells. From these data, we can calculate the Fano factor (ratio of

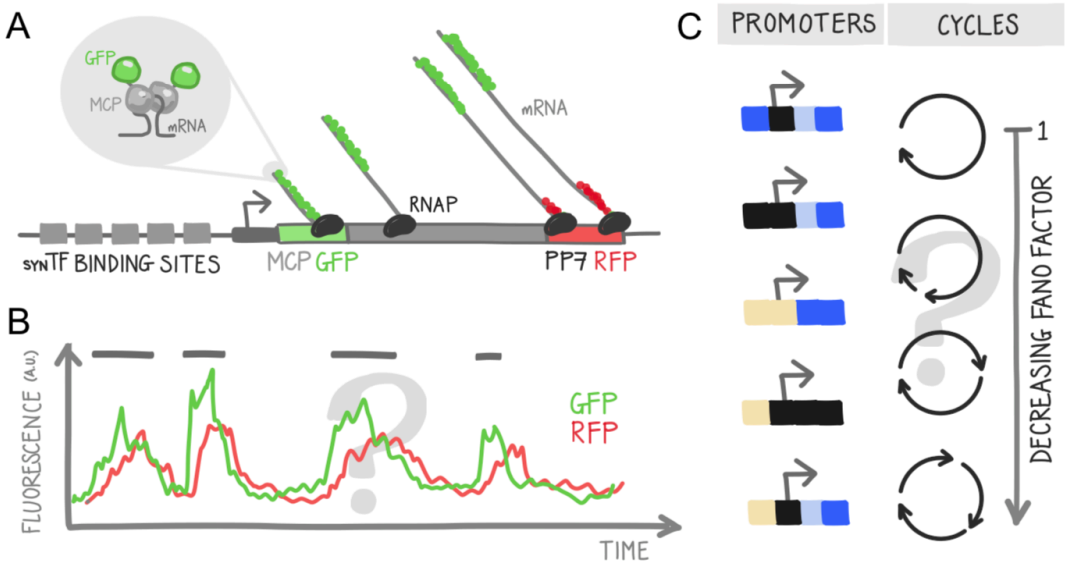


Figure 5.3 : Identify promoters with more than one slow step using live imaging of nascent transcription. (A) MS2 imaging. Stem loops incorporated into the 5' (MS2) and 3' (PP7) ends of a reporter gene are bound by fluorescent fusion proteins, creating a spot of fluorescent when transcribed. **(B)** Hypothetical data from live imaging in a single cell over time. We measure the size and frequency of transcription bursts (gray lines) and extract the mean and variance. **(C)** We hypothesize that promoters with different combinations of motifs (blue and yellow) differ the rates in their transcription cycles. We will calculate the Fano factor (ratio of variance to mean) for each promoter from MS2 traces. Figure by Clarissa Scholes.

the variance to the mean), and select promoters with a Fano Factor of less than 1 (Choubey et al., 2015). While these experiments tell us whether there is more than one slow step, they do not tell us how many steps there or, or which steps are slow, which could be assayed using inhibitors of particular steps. These experiments will identify an initial set of promoters we can use for a more in-depth characterization of combinatorial control, as both binding and kinetic cooperativity likely depends on the promoter to which they are recruited.

Do we see evidence for combinatorial control at endogenous genes?

We know a great deal about activators and repressors that work together to pattern the Drosophila blastoderm, despite not knowing the specific biochemical mechanisms by which these proteins function. Rather than just looking for synergy between pairs of TFs, we should screen activators that work together to co-regulate endogenous genes. For example, the eve2 enhancer is activated by Bicoid, Hunchback and Zelda (reviewed in (Vincent et al., 2016), and preliminary data from the lab suggests that Bicoid and Zelda activate eve2 in functionally distinct ways (Tim Harden, unpublished data). Combining what we know about *eve* and the activators that regulate it with the synthetic platform described in Chapter 4 would enable us to test combinations of TFs that are known to work together in our assay. Importantly, using pairs of TFs that co-regulate endogenous genes is a useful baseline for our synthetic experiments, and it will be informative to understand the expression range driven by these pairs and compare them to viral, yeast, and mammalian activators. The biochemical and genetic tools available to precisely control and label protein inputs and measure mRNA outputs are unparalleled in Drosophila (reviewed in Gregor et al., 2014); precise quantification of these features of our assay is one great advantage to working in the blastoderm.

Realistically, even just testing the combinatorial behavior of *eve* activators with the *eve* promoter would be an interesting place to start. Again, this type of experiment gets large and unruly quickly, so testing ~5 activators (approximately covering the TFs that regulate *eve*) in pairwise combinations would be an appropriate scale for the first study in flies. This approach is easy to adopt for other developmental genes, and given that the regulatory networks at this developmental stage contain a relatively small number of factors, we could learn a great deal about how these TFs combine rather quickly.

Do synTFs have similar effects on RNA polymerase as their endogenous counterparts?

Finally, there has been limited functional characterization of many TFs. While we have a handful of examples of proteins that work on chromatin remodeling, initiation, or elongation, these are often drawn from diverse cell types, assays, or model systems. Ideally, we would like to know more about how TFs work on the steps of transcription, and whether working on different steps is a mode of combinatorial control that is actually used by cells at promoters and enhancers to execute precise gene expression. One well-studied example is GAF and HSF1, which work together to accomplish different steps in regulation at the Drosophila heat-shock genes (Duarte et al., 2016). We focus extensively on "initiation" and "elongation" factors in Chapter 4, and one strategy to identify TFs that actually work on these steps is to perturb their levels by RNAi and measure how RNA polymerase distribution changes across the genes they regulate (Figure 5.4, Fuda et al., 2015).

We have designed experiments to knock down TF levels with RNAi in Drosophila S2 cell culture and measure the effects on RNA polymerase occupancy using NET-seq (Churchman and Weissman, 2011) and on transcription activation using RNA-seq. We will do this in a cell line where we have integrated a "synthetic enhancer": an array of synTF sites upstream of a well-studied promoter (*hsp70* or *eve* are good candidates) driving a reporter gene. This allows us to compare how synTFs behave in this synthetic context to how the same TFs operate at endogenous genes. These contexts are quite different, as in endogenous genes TFs work with many other factors, but at the synthetic enhancers their effects are isolated unless additional synTFs are transfected and recruited. Thus, comparing TF behavior in both contexts will be informative for this line of experimentation moving forward.

In order to conduct NET-seq and RNA-seq in multiple conditions for each TF studied, we need to narrow our list of TFs to a reasonable scope (~5). We will choose TFs that are activators expressed in the Drosophila blastoderm, are expressed in S2 cells, and co-occur in endogenous enhancers by analyzing existing ChIP-seq data (Li et al., 2008), beginning with the activators of *eve*. We will knock down the TF of interest using RNAi, and compare gene expression and polymerase occupancy to "control" cells that have expression of the synTF but

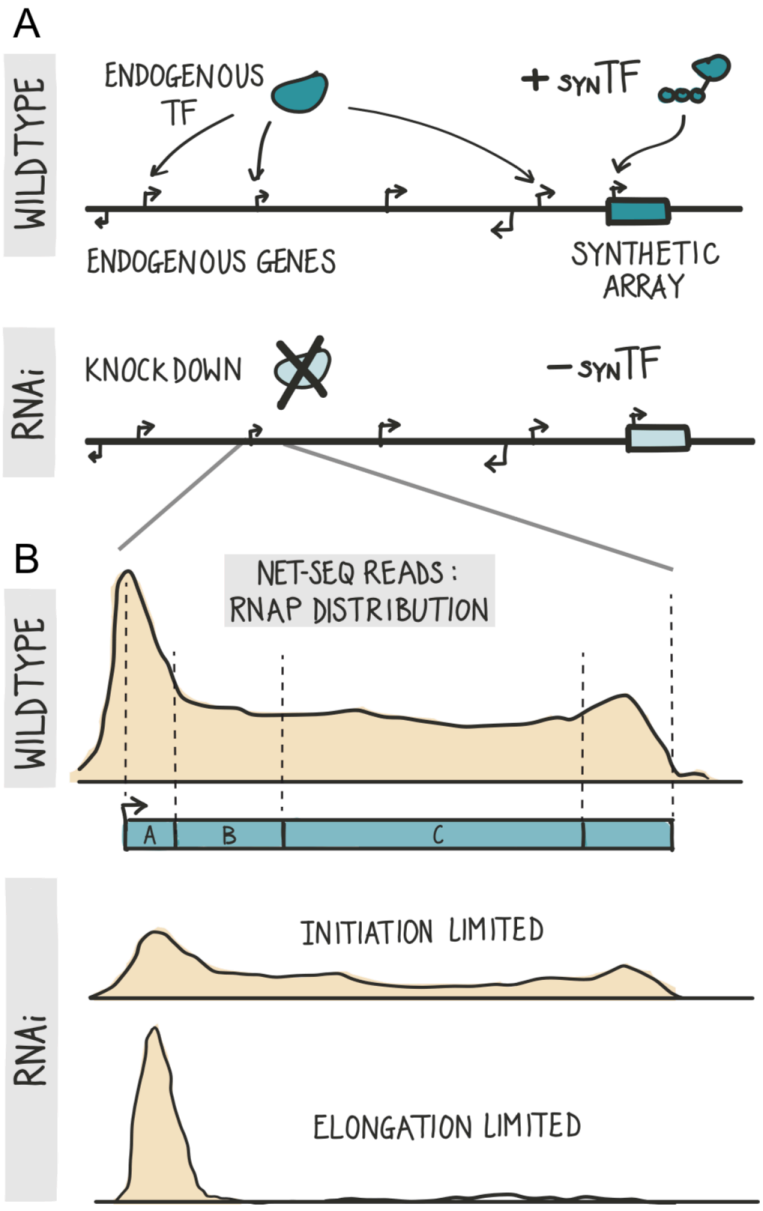


Figure 5.4 : Investigate TF function at endogenous loci and synthetic arrays with NET-seq. (A) We will integrate a reporter driven by an array of 5 synTF binding sites upstream of the *eve* promoter into mammalian 293T and Drosophila S2 cells. We will compare NET-seq profiles in two conditions: "wildtype", where the endogenous TF functions at its target promoters and transfected synTF is recruited to the synthetic array; and "RNAi", where the endogenous TF is depleted using RNAi and no synTF is transfected. **(B)** To interpret the NET-seq data we will split target genes into three regions (A, B, C) and define initiation factors as TFs where the overall number of reads decreases and the ratio of B/A increases towards 1 after RNAi. We will define elongation factors as TFs where the reads in region C decrease after RNAi or the ratio of A/B increases, indicating more paused RNAP. Figure by Clarissa Scholes.

are otherwise wild-type. The genome wide NET-seq and RNA-seq readouts allow us to see any off-target effects of the synTF activators and ensure that synTFs change RNA polymerase dynamics at the synthetic promoter in a way that is similar to their effects at endogenous genes.

We can identify how TFs change RNA polymerase occupancy by subdividing the gene into regions and defining TF function by the changes in the relative number of NET-seq reads in each region before and after knockdown (Larschan et al., 2011; Fuda et al., 2015). Reduced RNA polymerase in the gene body following knockdown indicates that elongation (or prior steps) are regulated by the TF, while reduction of binding over the promoter suggests that initiation is affected in the knockdown (Figure 5.4, (Larschan et al., 2011). To compare synTFs to their endogenous counterparts, we will perform a similar analysis at the reporter gene with and without synTF expression. We are specifically interested in comparing how the TF acts in isolation at the integrated promoter with how it acts in its endogenous context with many other factors at a variety of distances from the promoter. It may not be possible to identify TFs that work exclusively on initiation or elongation, but these experiments will illuminate how single activators affect polymerase dynamics at endogenous and synthetic genes. We are optimistic that more accurate information about single TFs will improve our future studies that work to decipher mechanisms of combinatorial control.

Outlook

Animal transcription is a highly dynamic process by which many pieces of regulatory sequence work together to drive highly precise and robust patterns of gene expression. The results presented in this thesis argue that enhancers are not simply modular, and that we must consider interactions multiple scales to understand the function of regulatory DNA. I am optimistic that we have a great deal to learn using these two case studies, both in the *eve* locus and with a minimal synthetic system. The combination of these two approaches will define the rules that govern interactions in regulatory sequence at multiple scales, both between TFs and between enhancers. By continuing to integrate quantitative experimental measurements and

computational models of gene expression, we will continue to advance our understanding of how animal transcription is controlled at the locus scale.

References

Benabdallah, N.S., and Bickmore, W.A. (2015). Regulatory Domains and Their Mechanisms. Cold Spring Harb. Symp. Quant. Biol. *80*, 45–51.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., and Celniker, S.E. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol. *5*, R61.

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., and Gregor, T. (2018). Dynamic interplay between enhancer-promoter topology and gene activity. Nat. Genet.

Chetverina, D., Fujioka, M., Erokhin, M., Georgiev, P., Jaynes, J.B., and Schedl, P. (2017). Boundaries of loop domains (insulators): Determinants of chromosome form and function in multicellular eukaryotes. Bioessays *39*.

Chong, S., Dugast-Darzacq, C., Liu, Z., Dong, P., Dailey, G.M., Cattoglio, C., Heckert, A., Banala, S., Lavis, L., Darzacq, X., et al. (2018). Imaging dynamic and selective low-complexity domain interactions that control gene transcription. Science *361*.

Choubey, S., Kondev, J., and Sanchez, A. (2015). Deciphering Transcriptional Dynamics In Vivo by Counting Nascent RNA Molecules. PLoS Comput. Biol. *11*, e1004345.

Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368–373.

Core, L.J., Waterfall, J.J., Gilchrist, D.A., Fargo, D.C., Kwak, H., Adelman, K., and Lis, J.T. (2012). Defining the status of RNA polymerase at promoters. Cell Rep. *2*, 1025–1035.

Corrigan, A.M., Tunnacliffe, E., Cannon, D., and Chubb, J.R. (2016). A continuum model of transcriptional bursting. Elife *5*.

Crocker, J., Abe, N., Rinaldi, L., McGregor, A.P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., et al. (2015). Low affinity binding site clusters confer hox specificity and regulatory robustness. Cell *160*, 191–203.

Cubeñas-Potts, C., Rowley, M.J., Lyu, X., Li, G., Lei, E.P., and Corces, V.G. (2017). Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. Nucleic Acids Res. *45*, 1714–1730.

Davies, J.O.J., Telenius, J.M., McGowan, S.J., Roberts, N.A., Taylor, S., Higgs, D.R., and Hughes, J.R. (2016). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. Nat. Methods *13*, 74–80.

Davis, I., and Ish-Horowicz, D. (1991). Apical localization of pair-rule transcripts requires 3' sequences and limits protein diffusion in the Drosophila blastoderm embryo. Cell *67*, 927–940.

Day, C.R., Chen, H., Coulon, A., Meier, J.L., and Larson, D.R. (2016). High-throughput single-molecule screen for small-molecule perturbation of splicing and transcription kinetics. Methods *96*, 59–68.

Deal, R.B., and Henikoff, S. (2010). The INTACT method for cell type–specific gene expression and chromatin profiling in Arabidopsis thaliana. Nat. Protoc. *6*, 56.

Duarte, F.M., Fuda, N.J., Mahat, D.B., Core, L.J., Guertin, M.J., and Lis, J.T. (2016). Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation. Genes Dev. *30*, 1731–1746.

Estrada, J., Ruiz-Herrero, T., Scholes, C., Wunderlich, Z., and DePace, A.H. (2016). SiteOut: An Online Tool to Design Binding Site-Free DNA Sequences. PLoS One *11*, e0151740.

FitzGerald, P.C., Sturgill, D., Shyakhtenko, A., Oliver, B., and Vinson, C. (2006). Comparative genomics of Drosophila and human core promoters. Genome Biol. *7*, R53.

Friedman, L.J., and Gelles, J. (2012). Mechanism of transcription initiation at an activator-dependent promoter defined by single-molecule observation. Cell *148*, 679–689.

Friedman, L.J., and Gelles, J. (2015). Multi-wavelength single-molecule fluorescence analysis of transcription mechanisms. Methods *86*, 27–36.

Fuda, N. J., Guertin, M. J., Sharma, S., Danko, C. G., Martins, A. L., Siepel, A. & Lis, J. T. (2015). GAGA factor maintains nucleosome free regions and has a role in RNA polymerase II recruitment to promoters. PLoS Genet. 11, e1005108.

Fujioka, M., Jaynes, J.B., and Goto, T. (1995). Early even-skipped stripes act as morphogenetic gradients at the single cell level to establish engrailed expression. Development *121*, 4371–4382.

Fujioka, M., Emi-Sarker, Y., Yusibova, G.L., Goto, T., and Jaynes, J.B. (1999). 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.

Fujioka, M., Yusibova, G.L., Patel, N.H., Brown, S.J., and Jaynes, J.B. (2002). The repressor activity of Even-skipped is highly conserved, and is sufficient to activate engrailed and to regulate both the spacing and stability of parasegment boundaries. Development *129*, 4411–4421.

Fujioka, M., Wu, X., and Jaynes, J.B. (2009). A chromatin insulator mediates transgene homing and very long-range enhancer-promoter communication. Development *136*, 3077–3087.

Fujioka, M., Sun, G., and Jaynes, J.B. (2013). The Drosophila eve insulator Homie promotes eve expression and protects the adjacent gene from repression by polycomb spreading. PLoS Genet. *9*, e1003883.

Fujioka, M., Mistry, H., Schedl, P., and Jaynes, J.B. (2016). Determinants of Chromosome Architecture: Insulator Pairing in cis and in trans. PLoS Genet. *12*, e1005889.

Furlong, E.E.M., and Levine, M. (2018). Developmental enhancers and chromosome topology. Science *361*, 1341–1345.

Gaertner, B., Johnston, J., Chen, K., Wallaschek, N., Paulson, A., Garruss, A.S., Gaudenz, K., De Kumar, B., Krumlauf, R., and Zeitlinger, J. (2012). Poised RNA polymerase II changes over developmental time and prepares genes for future expression. Cell Rep. *2*, 1670–1683.

Gratz, S.J., Rubinstein, C.D., Harrison, M.M., Wildonger, J., and O'Connor-Giles, K.M. (2015). CRISPR-Cas9 Genome Editing in Drosophila. Curr. Protoc. Mol. Biol. *111*, 31.2.1–20.

Gregor, T., Garcia H.G., Little, S.C. (2014) The embryo as a laboratory: quantifying transcription in Drosophila. Trends Genet. *30*, 364–375.

Haberle, V., and Lenhard, B. (2016). Promoter architectures and developmental gene regulation. Semin. Cell Dev. Biol. *57*, 11–23.

Haines, J.E., and Eisen, M.B. (2018). Patterns of chromatin accessibility along the anterior-posterior axis in the early Drosophila embryo. PLoS Genet. *14*, e1007367.

Harrison, M.M., Li, X.-Y., Kaplan, T., Botchan, M.R., and Eisen, M.B. (2011). Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. PLoS Genet. *7*, e1002266.

Hsieh, T.-H.S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O.J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell *162*, 108–119.

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., Kocks, C., Rajewsky, N., and Zinzen, R.P. (2017). The Drosophila embryo at single-cell transcriptome resolution. Science *358*, 194–199.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950–953.

Lamb, A.M., Walker, E.A., and Wittkopp, P.J. (2017). Tools and strategies for scarless allele replacement in Drosophila using CRISPR/Cas9. Fly *11*, 53–64.

Larschan, E., Bishop, E.P., Kharchenko, P.V., Core, L.J., Lis, J.T., Park, P.J., and Kuroda, M.I. (2011). X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. Nature *471*, 115–118.

Larson, D.R., Zenklusen, D., Wu, B., Chao, J.A., and Singer, R.H. (2011). Real-time observation of transcription initiation and elongation on an endogenous yeast gene. Science *332*, 475–478.

Li, X.-., and Eisen, M.B. Mutation of sequences flanking and separating transcription factor binding sites in a Drosophila enhancer significantly alter its output.

Li, X.-Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C.L., et al. (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol. *6*, e27.

Lim, B., Fukaya, T., Heist, T., and Levine, M. (2018). Temporal dynamics of pair-rule stripes in living Drosophila embryos. Proc. Natl. Acad. Sci. U. S. A.

Lubliner, S., Regev, I., Lotan-Pompan, M., Edelheit, S., Weinberger, A., and Segal, E. (2015). Core promoter sequence in yeast is a major determinant of expression level. Genome Res. *25*, 1008–1017.

Ludwig, M.Z., Manu, Kittler, R., White, K.P., and Kreitman, M. (2011). Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. PLoS Genet. *7*, e1002364.

Macdonald, P.M., Ingham, P., and Struhl, G. (1986). Isolation, structure, and expression of even-skipped: a second pair-rule gene of Drosophila containing a homeo box. Cell *47*, 721–734.

Manoukian, A.S., and Krause, H.M. (1993). Control of segmental asymmetry in Drosophila embryos. Development *118*, 785–796.

Mir, M., Reimer, A., Haines, J.E., Li, X.-Y., Stadler, M., Garcia, H., Eisen, M.B., and Darzacq, X. (2017). Dense Bicoid hubs accentuate binding along the morphogen gradient. Genes Dev. *31*, 1784–1794.

Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nat. Genet. *39*, 1507–1511.

Ohler, U., and Wassarman, D.A. (2010). Promoting developmental transcription. Development *137*, 15–26.

Ohler, U., Liao, G.-C., Niemann, H., and Rubin, G.M. (2002). Computational analysis of core promoters in the Drosophila genome. Genome Biol. *3*, RESEARCH0087.

Postika, N., Metzler, M., Affolter, M., Müller, M., Schedl, P., Georgiev, P., and Kyrchanova, O. (2018). Boundaries mediate long-distance interactions between enhancers and promoters in the Drosophila Bithorax complex. PLoS Genet. *14*, e1007702.

Sackerson, C., Fujioka, M., and Goto, T. (1999). The even-skipped locus is contained in a 16-kb chromatin domain. Dev. Biol. *211*, 39–52.

Schroeder, M.D., Greer, C., and Gaul, U. (2011). How to make stripes: deciphering the transition from non-periodic to periodic patterns in Drosophila segmentation. Development *138*, 3067–3078.

Singh, B.N., and Hampsey, M. (2014). Detection of short-range chromatin interactions by chromosome conformation capture (3C) in yeast. Methods Mol. Biol. *1205*, 209–218.

Small, S., Blair, A., and Levine, M. (1992). Regulation of even-skipped stripe 2 in the Drosophila embryo. EMBO J. *11*, 4047–4057.

Steiner, F.A., Talbert, P.B., Kasinathan, S., Deal, R.B., and Henikoff, S. (2012). Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. Genome Res. *22*, 766–777.

Tsai, A., Muthusamy, A.K., Alves, M.R., Lavis, L.D., Singer, R.H., Stern, D.L., and Crocker, J. (2017). Nuclear microenvironments modulate transcription from low-affinity enhancers. Elife *6*.

Tyssowski, K.M., DeStefino, N.R., Cho, J.-H., Dunn, C.J., Poston, R.G., Carty, C.E., Jones, R.D., Chang, S.M., Romeo, P., Wurzelmann, M.K., et al. (2018). Different Neuronal Activity Patterns Induce Different Gene Expression Programs. Neuron.

Vincent, B.J., Estrada, J., and DePace, A.H. (2016). The appeasement of Doug: a synthetic approach to enhancer biology. Integr. Biol. *8*, 475–484.

Zabidi, M.A., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature *518*, 556–559.

Appendix A : Supplemental Material for Chapter 2

Sequences and Expression Readouts

Enhancer sequences controlling different stripes (or pairs of stripes) of eve expression along the anterior-posterior (A/P) axis were collected from the REDfly database (Gallo, Gerrard et al. 2011). All enhancer and spacer sequences are given below. The details of synthesizing the sequence constructs and obtaining their readouts are given in Methods. Briefly, fused enhancers were placed upstream of the eve core promoter driving expression of LacZ. LacZ expression was measured at cellular resolution over 6 time points in blastoderm embryos, using fluorescent in situ hybridization. Image stacks were analyzed to produce rotationally aligned pointclouds (Fowlkes, Hendriks et al. 2008), and lateral traces were extracted. Readouts of the enhancers eve_3/7 and eve_4/6 were also obtained from (Lydiard-Martin, Bragdon et al. 2014). Since eve stripes 3 to 7 span ~40—95% of the A/P axis of the Drosophila blastoderm, for the purposes of model fitting described below, we have utilized ~30 A/P enhancers whose readouts fall within that domain. The relevant sequence- and expression-data were obtained from REDfly and FlyEx (Pisarev, Poustelnikova et al. 2009, Gallo, Gerrard et al. 2011) as has been described in our previous works (He, Samee et al. 2010, Samee and Sinha 2014).

Enhancer sequences

The following sequences correspond to the versions of eve37, eve46 and spacers that were assayed in reporter constructs. Lowercase letters indicate wild-type sequence, while uppercase letters indicate designed spacer sequences. 5' to 3' sequence of top strand is listed for each sequence.

eve46
aggatccctgggctctgggctctggactatccgccgaccctccatatccatgatttacaattctcgtttttttcgcgttatttttttaggggctttaa
tgaccgtcgtaaagccgcaggaggaccaggaccaggactctgctcacatttcgcgcactgattctaaaaaatgaaatcatttttttcttgaatttt
cacggcgcgcctcgagcaggactctttgttctcggccaggcaattgtccttttttgcgctcagctctcagtttttttcgtccagcgggcattaccta
cacggcgttttatggcggagatgatattcgcctgggatcggttccgtttttttaggccataaaaattaggcggcataaaaaaaactgcattggaat
tctagttctagtttcaagtttttaggtttccaggtttctgccagcccgcctagattcgcatttcgcggaattcggaagcggaacagaatgccaga
atggtcagaatcctggctgaccttgccttttggccaggggccgtaaaaaaaattgactcgctgcggtgcgcggaatatttttttaaatctgactttc
caacaatctctgatctgggttcgaatcgtaaaaaaaaaagcagaacaaaaagcgggcattttcgtcggcaaatgatctgttaatgggccgggc
taaaaaactaagtcacaaagtcacaaggttgtccggtaaattgacccggttaagaatgtctgtctgtaccgagaaggatgcaggacattcag
cacttcaaagctcccaccgctcgaaggattcccccgaagattcac

eve37
ggatcctcgaaatcgagagcgacctcgctgcattagaaaactagatcagttttttgttttggccgaccgatttttgtgcccggtgctctctttacg
gtttatggccgcgttcccatttcccagcttctttgttccgggctcagaaatctgtatggaattatggtatatgcagatttttatgggtcccggcgat
ccggttcgcggaacgggagtgtcctgccgcgagaggtcctcgccggcgatccttgtcgcccgtattaggaaagtagatcacgttttttgttccc
attgtgcgctttttttcgctgcgctagttttttttccccgaacccagcgaactgctctaatttttttaattcttcacggcttttcattgggctcctggaaa
aacgcggacaaggttataacgctctacttacctgcaattgtggccataactcgcactgctctcgtttttaagatccgtttgtttgtgtttgtttgtc
cgcgatggcattcacgtttttacgagctc

eve46 reverse orientation
gtgaatcttcgggggaatccttcgagcggtgggagctttgaagtgctgaatgtcctgcatccttctcggtacagacagacattcttaaccgggt
caatttaccggacaaccttgtgactttgtgacttagtttttttagcccggcccattaacagatcatttgccgacgaaaatgcccgctttttgttctgc
ttttttttttacgattcgaacccagatcagagattgttggaaagtcagatttaaaaaaatattccgcgcaccgcagcgagtcaatttttttacggcc
cctggccaaaaggcaaggtcagccaggattctgaccattctggcattctgttccgcttccgaattccgcgaaatgcgaatctaggcgggctgg
cagaaacctggaaacctaaaaacttgaaactagaactagaattccaatgcagttttttttatgccgcctaatttttatggcctaaaaaacggaac
cgatcccaggcgaatatcatctccgccataaaacgccgtgtaggtaatgcccgctggacgaaaaaactgagagctgagcgcaaaaaagga
caattgcctggccgagaacaaagagtcctgctcgaggcgcgccgtgaaattcaagaaaaaatgatttcatttttttagaatcagtgcgcgaaat
gtgagcagagtcctggtcctggtcctcctgcggctttacgacggtcattaaagcccctaaaaaaataacgcgaaaaaaacgagaattgtaaa
tcatggatatggagggtcggcggatagtccagagcccagagcccagggatcct

eve37 reverse orientation
gagctcgtaaaaacgtgaatgccatcgcggacaaacaaacacaaacaaacggatcttaaaaacgagagcagtgcgagttatggccacaat
tgcaggtaagtagagcgttataaccttgtccgcgtttttccaggagcccaatgaaaagccgtgaagaattaaaaaattagagcagttcgctgg
gttcggggaaaaaaaactagcgcagcgaaaaaagcgcacaatgggaacaaaaaacgtgatctactttcctaatacgggcgacaaggatcgc
cggcgaggacctctcgcggcaggacactcccgttccgcgaaccggatcgccgggacccataaaaatctgcatataccataattccatacag

atttctgagcccggaacaaagaagctgggaaatgggaacgcggccataaaccgtaaagagagcaccgggcacaaaaatcggtcggccaa
aacaaaaaactgatctagttttctaatgcagcgaggtcgctctcgatttcgaggatcc

spacer_1000
GGAAAGCTGGCTGGAGTGCGATCTTCCTGAGGCCGATACTGTCGTCGTCCCCTCAAACTGGCA
GATGCACGGTTACGATGCGCCCATCTACACCAACGTGACCTATCCCATTACGGTCAATCCGCCG
TTTGTTCCCACGGAGAATCCGACGGGTTGTTACTCGCTCACATTTAATGTTGATGAAAGCTGGC
TACAGGAAGGCCAGACGCGAATTATTTTTGATGGCGTTAACTCGGCGTTTCATCTGTGGTGCA
ACGGGCGCTGGGTCGGTTACGGCCAGGACAGTCGTTTGCCGTCTGAATTTGACCTGAGCGCAT
TTTTACGCGCCGGAGAAAACCGCCTCGCGGTGATGGTGCTGCGCTGGAGTGACGGCAGTTATC
TGGAAGATCAGGATATGTGGCGGATGAGCGGCATTTTCCGTGACGTCTCGTTGCTGCATAAAC
CGACTACACAAATCAGCGATTTCCATGTTGCCACTCGCTTTATGATGATTTCAGCCGCGCTGTA
CTGGAGGCTGAAGTTCAGATGTGCGGCGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTA
TGGCAGGGTGAAACGCAGGTCGCCAGCGGCACCGCGCCTTTCGGCGGTGAAATTATCGATGAG
CGTGGTGGTTATGCCGATCGCGTCACACTACGTCTGAACGTCGAAAACCCGAAACTGTGGAGC
GCCGAAATCCCGAATCTCTATCGTGCGGTGGTTGAACTGCACACCGCCGACGGCACGCTGATT
GAAGCAGAAGCCTGCGATGTCGGTTTCCGCGAGGTGCGGATTGAAAATGGTCTGCTGCTGCTG
AACGGCAAGCCGTTGCTGATTCGAGGCGTTAACCGTCACGAGCATCATCCTCTGCATGGTCAG
GTCATGGATGAGCAGACGATGGTGCAGGATATCCTGCTGATGAAGCAGAACAACTTTAACGCC
GTGCGC TGTTCGCATTATCCGAACCATCCGCTGTGGTACACGCTGTGCGACC

spacer_200
GCAGGGTGAAACGCAGGTCGCCAGCGGCACCGCGCCTTTCGGCGGTGAAATTATCGATGAGCG
TGGTGGTTATGCCGATCGCGTCACACTACGTCTGAACGTCGAAAACCCGAAACTGTGGAGCGC
CGAAATCCCGAATCTCTATCGTGCGGTGGTTGAACTGCACACCGCCGACGGCACGCTGATTGA
AGCAGAAGCCT

Supplemental Methods

Transcription Factor (TF) data

We used nine TFs – Bicoid (BCD), Caudal (CAD), Zelda (ZLD), Stat92E (STAT), Giant (GT), Hunchback (HB), Knirps (KNI), Kruppel (KR), and Tailless (TLL) – as the relevant inputs to the GEMSTAT model. Position weight matrices (PWMs) of these TFs were taken from the Fly Factor Survey database (Noyes, Meng et al. 2008), and their protein concentration profiles along the A/P axis (in the same temporal stage as the genes) were obtained from the FlyEx database (Pisarev, Poustelnikova et al. 2009). To annotate a TF's binding sites in a sequence, we first compute the log likelihood ratio (LLR) score of each k-bp window in the sequence, where k denotes the length of the TF's PWM and the two likelihoods in the ratio are computed from the PWM and a uniform background distribution. A window is then annotated as a binding site for the TF if the window's LLR score is above a pre-determined fraction of the LLR score of the TF's optimal site. There were several motifs available for each TF. We therefore needed to select the motif and the threshold on the LLR-score for each TF. To this end, we first listed all the footprinted binding sites in the enhancers for eve from the Redfly database (Gallo, Gerrard et al. 2011). For each TF, we selected the motif and the threshold that could identify all the footprinted sites along with the minimum number of additional sites (most of which had weak affinity).

Weighted pattern generating potential

The weighted pattern generating potential (w-PGP) (Samee and Sinha 2014) score computes two numbers for each datapoint (value of expression readout along the A/P axis): (1) a reward for the correctly predicted level of expression and (2) a penalty for over- or under- prediction. The w-PGP score is a linear combination of the reward and penalty terms across all the bins, and is normalized to be bound between 0 and 1. Briefly, let r_i and p_i denote the real and the predicted expression values at bin i, respectively. The amount of correctly predicted expression

in bin *i* can then be defined as min (p_i, r_i), and the reward at bin *i* is computed as $r_i X \min(p_i, r_i)$. Essentially this scheme weighs the amount of correctly predicted expression by the real expression level in that bin so that bins with greater expression levels contribute more to the reward term. The penalty at the *i*-th bin is defined as $(1 - r_i) \times \text{abs}(p_i - r_i)$. The factor abs ($p_i - r_i$), which represents the amount of wrong (over- or under-) prediction, is thus weighted by the extent of non-expression ($1 - r_i$) in that bin – this ensures that wrong predictions at the low expression bins are penalized more strongly. We have shown in our earlier work that w-PGP is a better objective for optimizing models against expression readouts in comparison to the two other common objective functions namely RMSE (root mean-squared-error) and CC (correlation coefficient).

Supplemental Figures

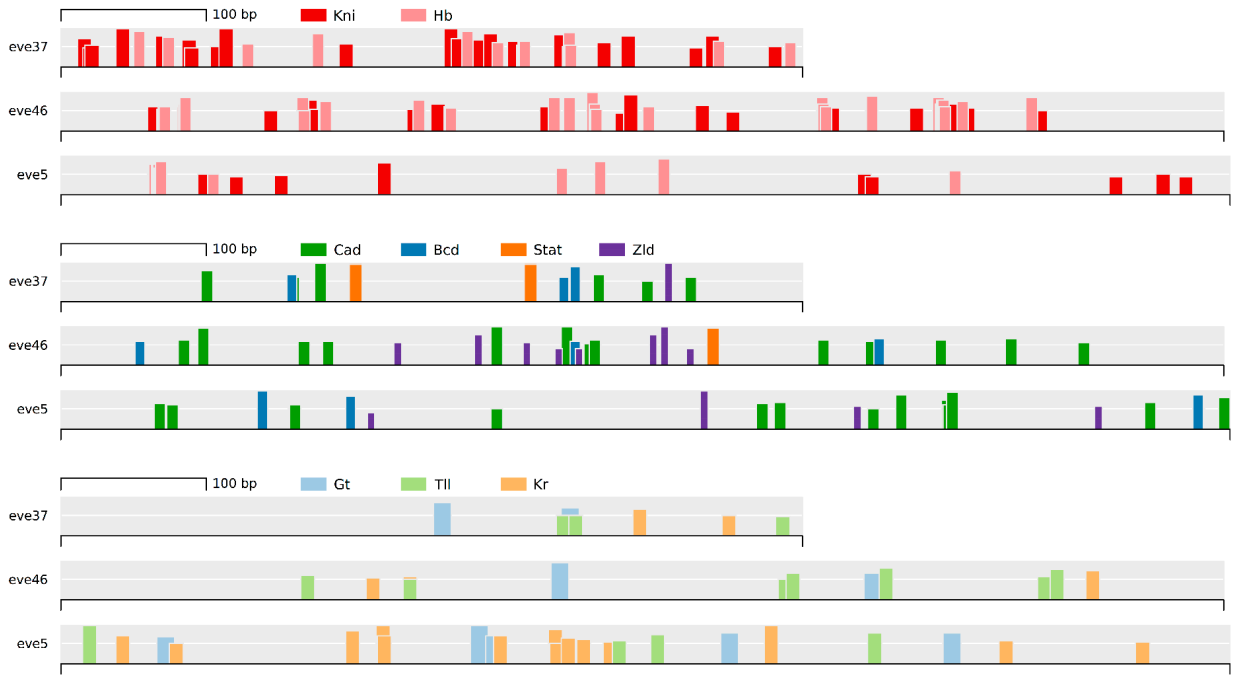


Figure S2.1. Computationally predicted binding sites of the nine TFs in eve_37, eve_46, and eve_5 enhancers, Related to Figure 2.1. For clarity we show sites of smaller groups of TFs into separate panels: Hb and Kni sites (top), Bcd, Cad, STAT, and Zld sites (middle), and Gt, Kr, and Tll sites (bottom). Enhancers are drawn to scale. Each colored box denotes a binding site of the corresponding TF; the width and the height of a box denote the length of the binding site (bp) and the ratio of LLR and maxLLR (explained in Supplementary Text) for that TF at that site, respectively. Plots created using inSite (https://www.cs.utah.edu/~miriah/insite/).

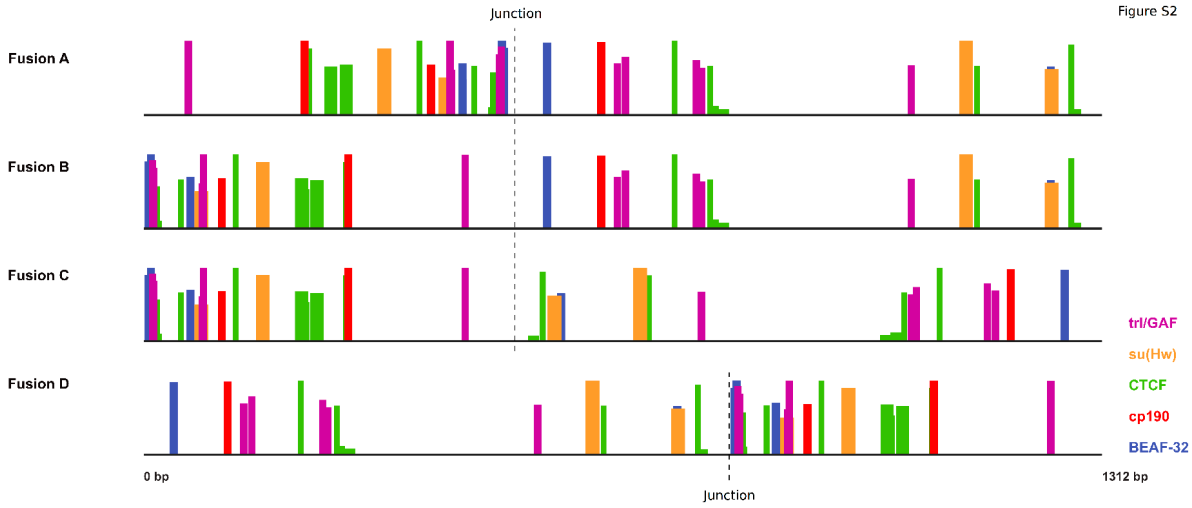


Figure S2.2. Computationally predicted binding sites of insulator and architectural proteins in the fusion constructs, Related to Figure 2.2. No new sites for any of these TFs was created at the junction of our synthetic constructs. Semantics are the same as in Figure S2.1.

Supplemental Tables

Table S2.1: Description of parameters in different models, related to Figures 2.3, 2.4, 2.5.

Model	Parameters per TF		Global parameters	Parameter per enhancer
GEMSTAT	Relative strength of DNA binding	Transcriptional effect (activating/repressing)	Basal transcription rate	-
GEMSTAT-SRR	Relative strength of DNA binding	Transcriptional effect (activating or short-range repressing)	Basal transcription rate	-
GEMSTAT-GL	Relative strength of DNA binding	Transcriptional effect (activating/repressing)	Basal transcription rate	Relative strength

Table S2.2 : w-PGP scores of different models for each synthetic construct, related to Figures 2.3, 2.4, 2.5.

	Direct-Interaction, constrained optimization	Direct-Interaction, unconstrained optimization	GEMSTAT-GL (locus level model)	GEMSTAT-SRR (short-range repression)
Fusion A	0.622888	0.673042	0.879516	0.925779
Fusion B	0.735868	0.711058	0.934725	0.920015
Fusion C	0.729233	0.679346	0.929167	0.906202
Fusion D	0.722374	0.797772	0.941819	0.919115
SpacerC 200		0.753688	0.930726	0.821166
SpacerC 1000		0.813482	0.930749	0.929190

Table S2.3. Windows selected by GEMSTAT-GL from each synthetic construct, related to Figure 2.4. For each window, the tuple shows the offset from the 5' end of the corresponding enhancer and the length of the window.

	Window from eve_3/7	Window from eve_4/6
Fusion A	(30, 440)	(320, 410)
Fusion B	(0, 400)	(320, 400)
Fusion C	(0, 500)	(20, 710)
Fusion D	(230, 270)	(80, 660)
SpacerC 200	(0, 510)	(20, 700)
SpacerC 1000	(0, 500)	(100, 690)

Supplemental References

Fowlkes, C. C., C. L. Hendriks, S. V. Keranen, G. H. Weber, O. Rubel, M. Y. Huang, S. Chatoor, A. H. DePace, L. Simirenko, C. Henriquez, A. Beaton, R. Weiszmann, S. Celniker, B. Hamann, D. W. Knowles, M. D. Biggin, M. B. Eisen and J. Malik (2008). "A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm." Cell 133(2): 364-374.

Gallo, S. M., D. T. Gerrard, D. Miner, M. Simich, B. Des Soye, C. M. Bergman and M. S. Halfon (2011). "REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila." Nucleic Acids Res 39(Database issue): D118-123.

He, X., M. A. Samee, C. Blatti and S. Sinha (2010). "Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression." PLoS Comput Biol 6(9).

Noyes, M. B., X. Meng, A. Wakabayashi, S. Sinha, M. H. Brodsky and S. A. Wolfe (2008). "A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system." Nucleic Acids Res 36(8): 2547-2560.

Pisarev, A., E. Poustelnikova, M. Samsonova and J. Reinitz (2009). "FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution." Nucleic Acids Res 37(Database issue): D560-566.

Samee, M. A. and S. Sinha (2014). "Quantitative modeling of a gene's expression from its intergenic sequence." PLoS Comput Biol 10(3): e1003467.

Appendix B : Supplemental Information for Chapter 3

Supplemental Figures

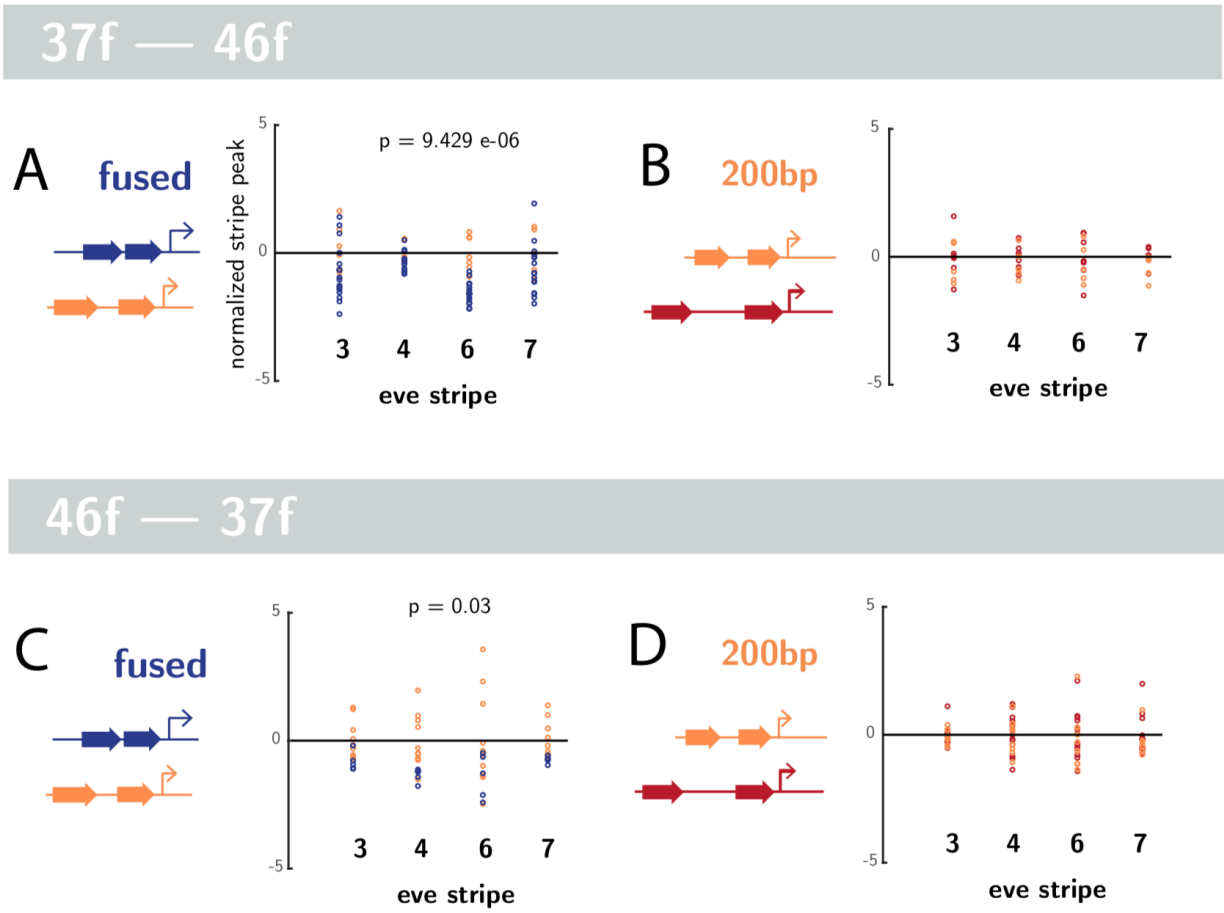


Figure S3.1: Individual points for enhancer rearrangements. We compared the peak expression (95% percentile) of each stripe in individual embryos. The mean peak values for each control are centered on zero to facilitate comparison. (Top) Constructs with eve37 upstream of eve46. (A) eve37 and eve46 separated by a 1kb spacer (red) and 200bp spacer (orange). (B) eve37 and eve46 separated by a 200bp spacer (orange) and fused together (dark blue). (Bottom) Constructs with eve46 upstream of eve37. (C) eve46 and eve37 separated by a 1kb spacer (red) and 200bp spacer (orange). (D) eve46 and eve37 separated by a 200bp spacer (orange) and fused together (dark blue). Constructs containing enhancer fusions drive lower expression levels than constructs with enhancers separated by a 200bp spacer ($p < 0.05$ by ANOVA). $n \geq 5$ for all lines.

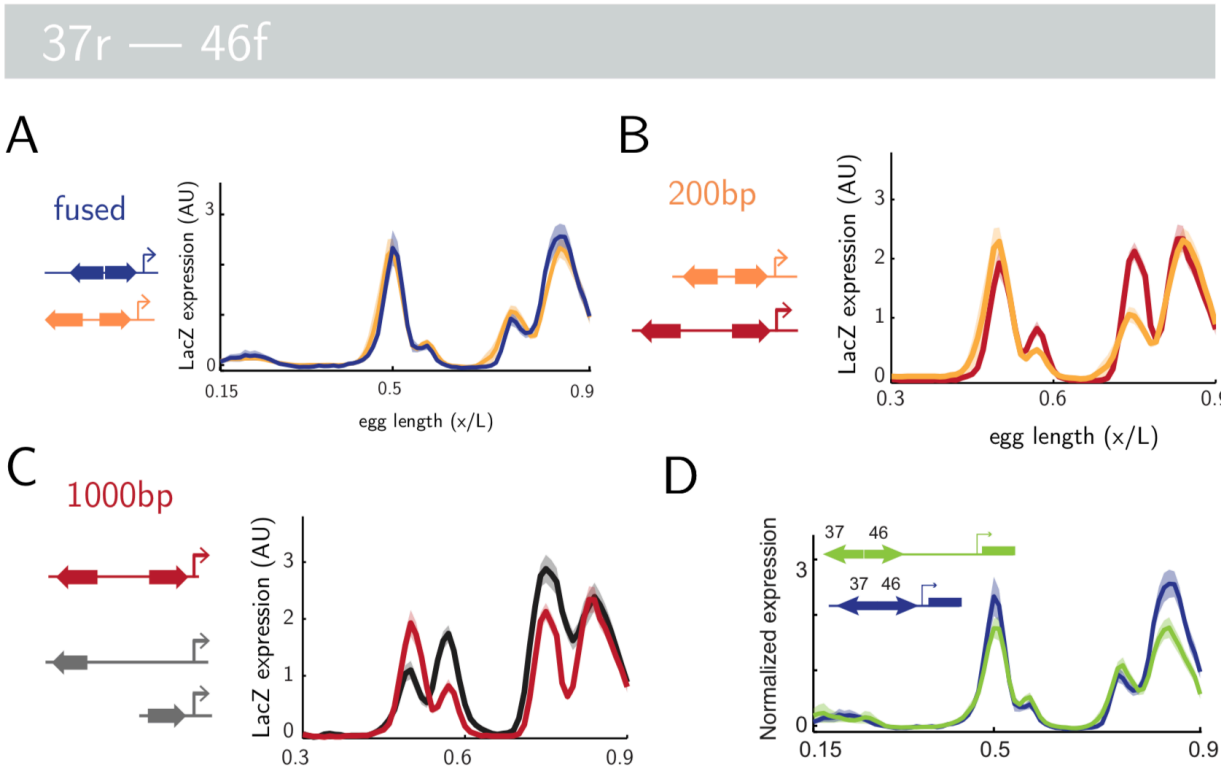


Figure S3. 2: Additional configuration of rearranged enhancers also drives non-additive patterns. We created an additional construct containing eve37 in the reverse orientation upstream of eve46 in the forward orientation. (A - C) We compared expression driven by constructs with a 0bp (purple), 200bp (orange), and 1kb (red) LacZ spacer sequence as described in Figure 3.2. Single enhancer controls are shown in grey. hkb normalized expression for each transgenic line is displayed as a function of AP position with the shadow indicating standard error of the mean (SEM). n ≥ 6 for all lines. (D) We also moved the 37r-46f fusion (blue) upstream of the promoter by introducing a 1000bp LacZ spacer (green). The expression pattern is similar for both configurations, with a modest decrease in stripes 3 and 7 when the enhancers are moved away. Data displayed as in A-C.

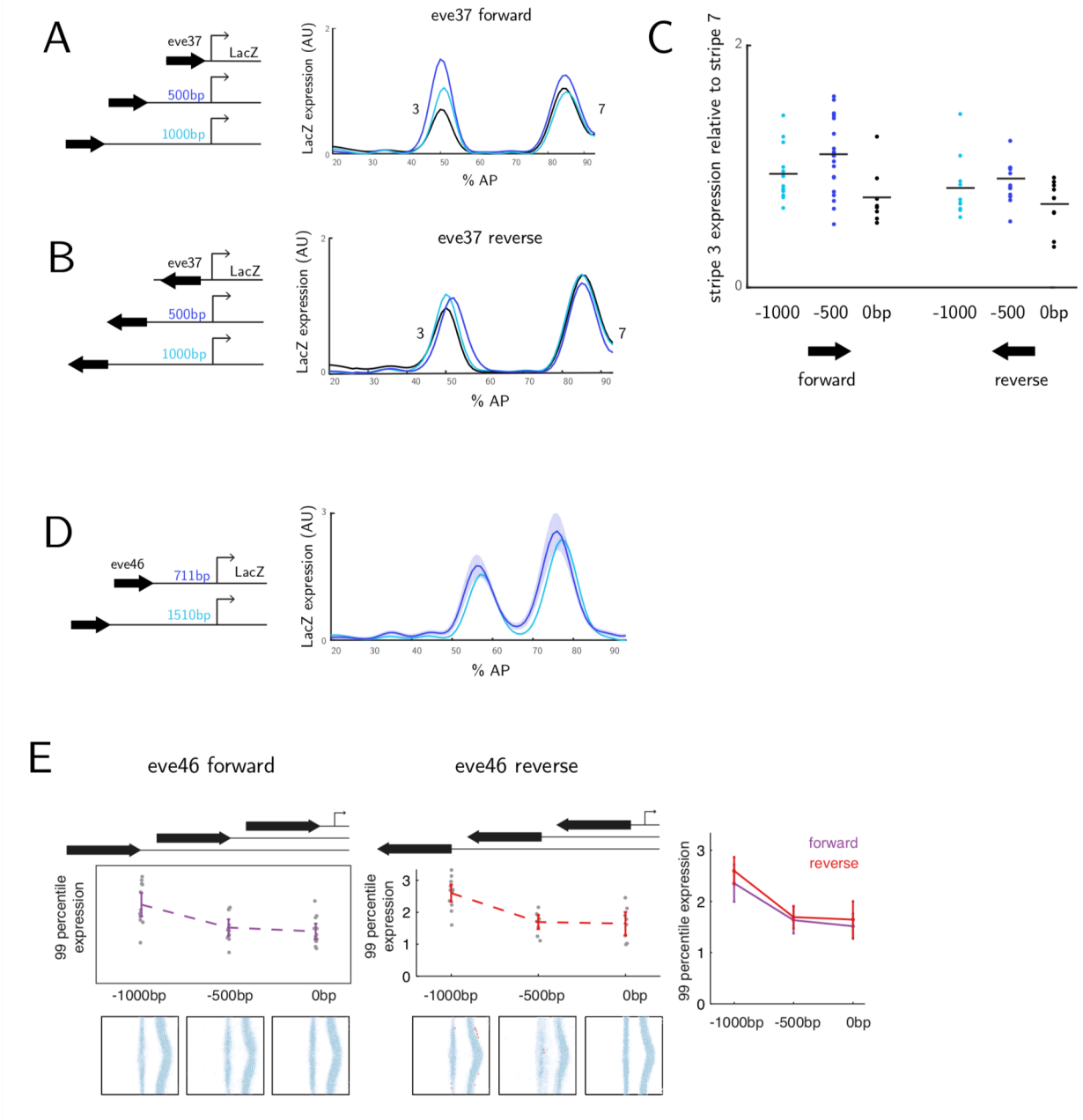
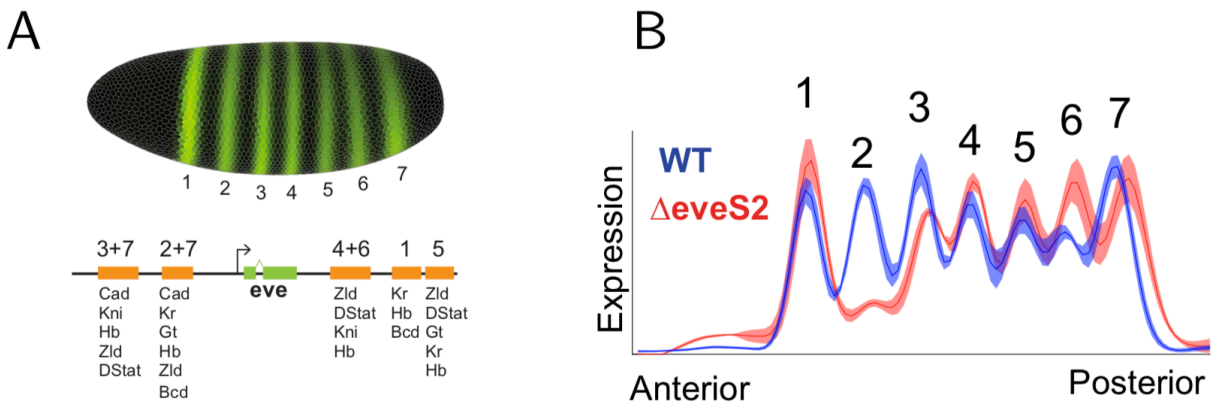


Figure S3.3 : eve37 and eve46 are sensitive to distance and orientation relative to the promoter. (A - B) We measured expression driven by the eve37 enhancer at three distances from the promoter and two orientations, as indicated in the schematics on the left side of each panel. eve37 at the promoter (black), eve37 with a 500bp spacer (dark blue), and eve37 with a 1kb spacer (light blue). Expression values were normalized by co-staining with endogenous *hkb* (see Methods) to enable comparison across transgenic lines. AU, arbitrary units. n ≥ 8 for all lines. (C) The ratio of peak stripe 3 expression (0.45-0.55 egg length) to peak stripe 7 expression (0.8 to 0.9 egg length) for individual embryos is plotted for each configuration of distance and orientation. The mean is marked with a black bar. We observe significant differences in expression level from varying position and orientation (p-value = 0.0015 by 6-way ANOVA). The effect of varying distance from the promoter was significant for

Figure S3.3 (continued).
the forward orientation but not the reverse ($p = 0.02$ and $p = 0.1028$ by ANOVA, respectively) and the effect of orientation was significant at the -500bp position ($p = 0.05$ by ranksum test) but not at the other distances. (D) eve46 at two distances from the promoter with a 711bp (dark blue) and 1510 bp (light blue) neutral spacer. Expression is displayed as a line trace along the AP axis. Shadow represents standard error of the mean (SEM). $n \geq 5$ for all lines. (E) eve46 at three distances and two orientations using a LacZ spacer sequence. We measured expression driven by the eve 4/6 enhancer at three distances from the promoter and two orientations as indicated in schematics at top of each panel. Individual embryos are shown as grey dots; bars indicate the mean and 95% confidence interval of the standard error of the mean (SEM). We also overlay the measurements for both orientations in order to see the influence of orientation. We use 99 percentile expression in the trunk (0.2-0.9 egg length) to estimate the level of expression driven by each construct. $n \geq 6$ for all lines. We observe significant differences in expression dependent on distance and orientation. We also thresholded gene expression in the embryos to test whether the position of expression changed. We show an unrolled embryo view for each distance with the percentage of embryos in which a cell expresses the reporter plotted in blue. Cells that were significantly different from the reference line (0bp from promoter in forward orientation) are plotted in red ($p<0.05$, Fisher's Exact Test with permutation to control for multiple hypothesis testing). Position does not change for most lines. The most extreme position shift is a narrowing of the stripes in reverse orientation eve46 at 1000bp from promoter.

Figure S3.4: Single embryo traces for enhancer deletions. Eve expression across the AP axis is displayed as a line trace normalized to stripe 1 expression. Data is separated by time point through developmental stage 5, with WT lines is shown in black, Δeve46 in blue, and Δeve37 in orange. Shadow indicates SEM. (Top) WT and eve46 deletion single embryo traces. (Bottom) WT and eve37 deletion single embryo traces. n ≥ 5 embryos for all time points.



C

gRNA	gRNA sequence	Genomic coordinates
eve37 5'	3' - GGAATCGAGGGCTGCGAGAC**AGG** - 5'	2R: 9975235..9975257
eve37 3'	5' - GTGCACAGTTCTGTCTTAAG**TGG** - 3'	2R: 9976333..9976355
eve46 5'	5' - GATTGAGGCAATCACTGGTG**TGG** - 3'	2R:9983803..998325
eve46 3'	5' - CAAGGTCAGGCCGCCCTTAA**TGG** - 3'	2R: 9984802..9984824

Figure S3.5: Additional experimental details. (A) Even-skipped (eve) is expressed in 7 stripes driven by 5 enhancers that are spread over ~16kb of regulatory DNA. Known genetic regulators of *eve* stripes are listed below their corresponding enhancers. (B) Ludwig et al. replaced the minimal eve stripe 2 enhancer with a fragment of the *mini-white* gene in the endogenous locus. We stained this line for *eve* mRNA by *in situ* hybridization and present average peak expression as a line trace along the AP axis normalized to stripe 1 expression. Shadow indicates SEM. Data is not separated by time point and is from the whole of developmental stage 5. n ≥ 5 embryos with WT shown in blue and Δeve2 in red. (C) gRNAs used in CRISPR/Cas9 stripe enhancer deletions. The 23bp listed includes the target sequence and PAM site (bold). Genomic coordinates are from UCSC genome browser dm6 release.

Enhancer Sequences

Sequence is listed 5’ to 3’; lower case indicates endogenous enhancer sequence, upper case is spacer sequence.

eve46
aggatccctgggctctgggctctggactatccgccgaccctccatatccatgatttacaattctcgtttttttcgcgttatttttttaggggctttaa
tgaccgtcgtaaagccgcaggaggaccaggaccaggactctgctcacatttcgcgcactgattctaaaaaatgaaatcatttttttcttgaatttt
cacggcgcgcctcgagcaggactctttgttctcggccaggcaattgtccttttttgcgctcagctctcagtttttttcgtccagcgggcattaccta
cacggcgttttatggcggagatgatattcgcctgggatcggttccgtttttttaggccataaaaattaggcggcataaaaaaaactgcattggaat
tctagttctagtttcaagtttttaggtttccaggtttctgccagcccgcctagattcgcatttcgcggaattcggaagcggaacagaatgccaga
atggtcagaatcctggctgaccttgccttttggccaggggccgtaaaaaaaattgactcgctgcggtgcgcggaatatttttttaaatctgactttc
caacaatctctgatctgggttcgaatcgtaaaaaaaaaagcagaacaaaaagcgggcattttcgtcggcaaatgatctgttaatgggccgggc
taaaaaactaagtcacaaagtcacaaggttgtccggtaaattgacccggttaagaatgtctgtctgtaccgagaaggatgcaggacattcag
cacttcaaagctcccaccgctcgaaggattcccccgaagattcac

eve37
ggatcctcgaaatcgagagcgacctcgctgcattagaaaactagatcagttttttgttttggccgaccgatttttgtgcccggtgctctctttacg
gtttatggccgcgttcccatttcccagcttctttgttccgggctcagaaatctgtatggaattatggtatatgcagatttttatgggtcccggcgat
ccggttcgcggaacgggagtgtcctgccgcgagaggtcctcgccggcgatccttgtcgcccgtattaggaaagtagatcacgttttttgttccc
attgtgcgctttttttcgctgcgctagttttttttccccgaacccagcgaactgctctaatttttttaattcttcacggcttttcattgggctcctggaaa
aacgcggacaaggttataacgctctacttacctgcaattgtggccataactcgcactgctctcgtttttaagatccgtt
tgtttgtgtttgtttgtccgcgatggcattcacgtttttacgagctc

eve46 reverse orientation
gtgaatcttcgggggaatccttcgagcggtgggagctttgaagtgctgaatgtcctgcatccttctcggtacagacagacattcttaaccgggt
caatttaccggacaaccttgtgactttgtgacttagttttttagcccggcccattaacagatcatttgccgacgaaaatgcccgctttttgttctgc
ttttttttttacgattcgaacccagatcagagattgttggaaagtcagatttaaaaaatattccgcgcaccgcagcgagtcaatttttttacggcc
cctggccaaaaggcaaggtcagccaggattctgaccattctggcattctgttccgcttccgaattccgcgaaatgcgaatctaggcgggctgg
cagaaacctggaaacctaaaaacttgaaactagaactagaattccaatgcagttttttatgccgcctaatttttatggcctaaaaaacggaac
cgatcccaggcgaatatcatctccgccataaaacgccgtgtaggtaatgcccgctggacgaaaaaactgagagctgagcgcaaaaaagga
caattgcctggccgagaacaaagagtcctgctcgaggcgcgccgtgaaattcaagaaaaaatgatttcatttttttagaatcagtgcgcgaaat
gtgagcagagtcctggtcctggtcctcctgcggctttacgacggtcattaaagcccctaaaaaaaataacgcgaaaaaaacgagaattgtaaa
tcatggatatggagggtcggcggatagtccagagcccagagcccagggatcct

eve37 reverse orientation
gagctcgtaaaaacgtgaatgccatcgcggacaaacaaacacaaacaaacggatcttaaaaacgagagcagtgcgagttatggccacaat
tgcaggtaagtagagcgttataaccttgtccgcgtttttccaggagcccaatgaaaagccgtgaagaattaaaaaattagagcagttcgctgg
gttcggggaaaaaaaactagcgcagcgaaaaaagcgcacaatgggaacaaaaaacgtgatctactttcctaatacgggcgacaaggatcgc
cggcgaggacctctcgcggcaggacactcccgttccgcgaaccggatcgccgggacccataaaaatctgcatataccataattccatacag
atttctgagcccggaacaaagaagctgggaaatgggaacgcggccataaaccgtaaagagagcaccgggcacaaaaatcggtcggccaa
aacaaaaaactgatctagttttctaatgcagcgaggtcgctctcgatttcgaggatcc

1000bp neutral spacer
CCTCACATTCTATGATAGTCTAAGAGCAAGGTATGAGGAATATAGGAAGTTTTAGGAATGTTA
GAAGCGCATACCGAGATACATCGTGCTGGATCTATGTGAGTACCTTTGCTATACTAGGCTGTCT
CAAGTTACTATTACAATGCTTTGAAGTGAGTATATCTAATACATCGCAGTGTTGGTTGTTGGAG
ACTAGTATACTGGGTGGAAGTGCGAGTTAACTCGAACTCGCAACTATTACGTTTGGAACATGG
TATAGCAAGAATCACTACATCGAAGAATCTACGATCTCACACTAAGAGTACTATATCTCAAAAC
GGTCTATACTTGAAACTTGCAACTCACAGCGTATTAGATTGCTGTAGAGAGACTACATTGGTTG
GAGTCGGTTGAGGGATCACTGCAAGATATTAAGTCTCGAGGATAGTCTCAAGTCTATCAAGAG

TCTGAGGCCCGATCCTCTATCGTACGTTCAAGTATCTTCACCTCGTACACTACTATACTTTAAGT
CCTCATACTAGTAGATTGTCTAGGACAAGATATAGCGTTTCTAGTGCCAAAATTCTGAGTCTCC
CTATAATATTCCAGCCTTAGAACAGATCTACGAGGGAGATATCTCTAGAGATATTTGAAGCGAA
TAATAGTTCCTGAATCACAGATTCTATAGAGGTATTGATTCCGAACATATAATTTTAGAGATCT
CTCCGTACAAACCGTTGTGTCAAACTCTAACAGGACACGACACTCACCATAGTGCATTTGAGAA
ACTATCATGATTTGTCGTATGAGGAACTCACAGAGCTTCGTGATCGTGGTCCTCGTCTAGACTG
TAATAGTTCTCGTATAGGCCCCTAGTATGATAGAATCGCAGTATTACATTGGTACAACGATGAT
TCAGACTACAAGTATTAGACATAGAGTTTCTAGCTCAAAGCACTAGAAACCAAGACTCTCGTTG
TACGATCGCTGGGCATAGAACTAATAGCTCGCAAGCTACGAG

200bp neutral spacer
CCTCACATTCTATGATAGTCTAAGAGCAAGGTATGAGGAATATAGGAAGTTTTAGGAATGTTA
GAAGCGCATACCGAGATACATCGTGCTGGATCTATGTGAGTACCTTTGCTATACTAGGCTGTCT
CAAGTTACTATTACAATGCTTTGAAGTGAGTATATCTAATACATCGCAGTGTTGGTTGTTGGAG
ACTAGTATA

1511bp neutral spacer
CCTCACATTCTATGATAGTCTAAGAGCAAGGTATGAGGAATATAGGAAGTTTTAGGAATGTTA
GAAGCGCATACCGAGATACATCGTGCTGGATCTATGTGAGTACCTTTGCTATACTAGGCTGTCT
CAAGTTACTATTACAATGCTTTGAAGTGAGTATATCTAATACATCGCAGTGTTGGTTGTTGGAG
ACTAGTATACTGGGTGGAAGTGCGAGTTAACTCGAACTCGCAACTATTACGTTTGGAACATGG
TATAGCAAGAATCACTACATCGAAGAATCTACGATCTCACACTAAGAGTACTATATCTCAAAAC
GGTCTATACTTGAAACTTGCAACTCACAGCGTATTAGATTGCTGTAGAGAGACTACATTGGTTG
GAGTCGGTTGAGGGATCACTGCAAGATATTAAGTCTCGAGGATAGTCTCAAGTCTATCAAGAG
TCTGAGGCCCGATCCTCTATCGTACGTTCAAGTATCTTCACCTCGTACACTACTATACTTTAAGT
CCTCATACTAGTAGATTGTCTAGGACAAGATATAGCGTTTCTAGTGCCAAAATTCTGAGTCTCC
CTATAATATTCCAGCCTTAGAACAGATCTACGAGGGAGATATCTCTAGAGATATTTGAAGCGAA
TAATAGTTCCTGAATCACAGATTCTATAGAGGTATTGATTCCGAACATATAATTTTAGAGATCT
CTCCGTACAAACCGTTGTGTCAAACTCTAACAGGACACGACACTCACCATAGTGCATTTGAGAA
ACTATCATGATTTGTCGTATGAGGAACTCACAGAGCTTCGTGATCGTGGTCCTCGTCTAGACTG
TAATAGTTCTCGTATAGGCCCCTAGTATGATAGAATCGCAGTATTACATTGGTACAACGATGAT
TCAGACTACAAGTATTAGACATAGAGTTTCTAGCTCAAAGCACTAGAAACCAAGACTCTCGTTG
TACGATCGCTGGGCATAGAACTAATAGCTCGCAAGCTACGAGACTCGGATGTATAGTAAGACT
ATTCGTCACCACTATGTGAGTATCTTGGTGAAACGATATATCCAGCACTCCACAGAATATCGAA
TCAAGCACTAACTCATAGCAACGTTGTAAGTCCTATGGGGAGAATAGAGTCTGTCTTCGCGTG
GATCCTAGTATATCTGTAGTGCGTTGAACACTACGAGACATGAGTAAGCGAGTATACCGTCTTG
TGGAGTGTCAGAGCTAGTATCACGAAGAATTCTACGACGTACAAGTTTCCTAGGCTGGAACAA
TATCTCATGTTCAGATCAAGGCATCATCGTCCAGCCAAGAATCTTTCAGAGGACTCTTGGATCC
ATACTAGGCTCAAATACGTCTCAATTTGAAACTCTATTTGGAAGACTATCAAGTATAGACGGTT
CTTGCTATGTCCGATACATGATTCGAGTCTTGCGAGTCTCACCCTAGACATTGAATCTAATACG
CGCTCAGACACTGGTGAACTACGAGTCTATCTCTTTCTCGTCCA

1801bp neutral spacer
CCTCACATTCTATGATAGTCTAAGAGCAAGGTATGAGGAATATAGGAAGTTTTAGGAATGTTA
GAAGCGCATACCGAGATACATCGTGCTGGATCTATGTGAGTACCTTTGCTATACTAGGCTGTCT
CAAGTTACTATTACAATGCTTTGAAGTGAGTATATCTAATACATCGCAGTGTTGGTTGTTGGAG
ACTAGTATACTGGGTGGAAGTGCGAGTTAACTCGAACTCGCAACTATTACGTTTGGAACATGG
TATAGCAAGAATCACTACATCGAAGAATCTACGATCTCACACTAAGAGTACTATATCTCAAAAC
GGTCTATACTTGAAACTTGCAACTCACAGCGTATTAGATTGCTGTAGAGAGACTACATTGGTTG
GAGTCGGTTGAGGGATCACTGCAAGATATTAAGTCTCGAGGATAGTCTCAAGTCTATCAAGAG
TCTGAGGCCCGATCCTCTATCGTACGTTCAAGTATCTTCACCTCGTACACTACTATACTTTAAGT
CCTCATACTAGTAGATTGTCTAGGACAAGATATAGCGTTTCTAGTGCCAAAATTCTGAGTCTCC
CTATAATATTCCAGCCTTAGAACAGATCTACGAGGGAGATATCTCTAGAGATATTTGAAGCGAA
TAATAGTTCCTGAATCACAGATTCTATAGAGGTATTGATTCCGAACATATAATTTTAGAGATCT

CTCCGTACAAACCGTTGTGTCAAACTCTAACAGGACACGACACTCACCATAGTGCATTTGAGAA
ACTATCATGATTTGTCGTATGAGGAACTCACAGAGCTTCGTGATCGTGGTCCTCGTCTAGACTG
TAATAGTTCTCGTATAGGCCCCTAGTATGATAGAATCGCAGTATTACATTGGTACAACGATGAT
TCAGACTACAAGTATTAGACATAGAGTTTCTAGCTCAAAGCACTAGAAACCAAGACTCTCGTTG
TACGATCGCTGGGCATAGAACTAATAGCTCGCAAGCTACGAGACTCGGATGTATAGTAAGACT
ATTCGTCACCACTATGTGAGTATCTTGGTGAAACGATATATCCAGCACTCCACAGAATATCGAA
TCAAGCACTAACTCATAGCAACGTTGTAAGTCCTATGGGGAGAATAGAGTCTGTCTTCGCGTG
GATCCTAGTATATCTGTAGTGCGTTGAACACTACGAGACATGAGTAAGCGAGTATACCGTCTTG
TGGAGTGTCAGAGCTAGTATCACGAAGAATTCTACGACGTACAAGTTTCCTAGGCTGGAACAA
TATCTCATGTTCAGATCAAGGCATCATCGTCCAGCCAAGAATCTTTCAGAGGACTCTTGGATCC
ATACTAGGCTCAAATACGTCTCAATTTGAAACTCTATTTGGAAGACTATCAAGTATAGACGGTT
CTTGCTATGTCCGATACATGATTCGAGTCTTGCGAGTCTCACCCTAGACATTGAATCTAATACG
CGCTCAGACACTGGTGAACTACGAGTCTATCTCTTTCTCGTCCAAGATTCAAGACTATACTGAC
GATTGATACATCTAACATAGTATGATAGACCCTAGCGTACAATCTCGTAGTTCAACTAGGAACT
ACTCGACGCTTCAACTCACTACTCGTACAGAGATAGACTCCTAGCAGTGGTGAGACAGTATGA
GAAACGCTAGATATAATATTGTAAGTTTTCAGACTCCTACGATTCTAGCTCGCTTGAGAATCTT
GGTCTGTAAGTTCTAGTCGATCTCGTAGGCATCATGCGAGAATGTACTCGTAGTGCATAGAGA
CGAATTCACAAGTACT

lacZ 1000bp spacer
GGAAAGCTGGCTGGAGTGCGATCTTCCTGAGGCCGATACTGTCGTCGTCCCCTCAAACTGGCA
GATGCACGGTTACGATGCGCCCATCTACACCAACGTGACCTATCCCATTACGGTCAATCCGCCG
TTTGTTCCCACGGAGAATCCGACGGGTTGTTACTCGCTCACATTTAATGTTGATGAAAGCTGGC
TACAGGAAGGCCAGACGCGAATTATTTTTGATGGCGTTAACTCGGCGTTTCATCTGTGGTGCA
ACGGGCGCTGGGTCGGTTACGGCCAGGACAGTCGTTTGCCGTCTGAATTTGACCTGAGCGCAT
TTTTACGCGCCGGAGAAAACCGCCTCGCGGTGATGGTGCTGCGCTGGAGTGACGGCAGTTATC
TGGAAGATCAGGATATGTGGCGGATGAGCGGCATTTTCCGTGACGTCTCGTTGCTGCATAAAC
CGACTACACAAATCAGCGATTTCCATGTTGCCACTCGCTTTATGATGATTTCAGCCGCGCTGTA
CTGGAGGCTGAAGTTCAGATGTGCGGCGAGTTGCGTGACTACCTACGGGTAACAGTTTCTTTA
TGGCAGGGTGAAACGCAGGTCGCCAGCGGCACCGCGCCTTTCGGCGGTGAAATTATCGATGAG
CGTGGTGGTTATGCCGATCGCGTCACACTACGTCTGAACGTCGAAAACCCGAAACTGTGGAGC
GCCGAAATCCCGAATCTCTATCGTGCGGTGGTTGAACTGCACACCGCCGACGGCACGCTGATT
GAAGCAGAAGCCTGCGATGTCGGTTTCCGCGAGGTGCGGATTGAAAATGGTCTGCTGCTGCTG
AACGGCAAGCCGTTGCTGATTCGAGGCGTTAACCGTCACGAGCATCATCCTCTGCATGGTCAG
GTCATGGATGAGCAGACGATGGTGCAGGATATCCTGCTGATGAAGCAGAACAACTTTAACGCC
GTGCGC TGTTCGCATTATCCGAACCATCCGCTGTGGTACACGCTGTGCGACC

lacZ 200bp spacer
GCAGGGTGAAACGCAGGTCGCCAGCGGCACCGCGCCTTTCGGCGGTGAAATTATCGATGAGCG
TGGTGGTTATGCCGATCGCGTCACACTACGTCTGAACGTCGAAAACCCGAAACTGTGGAGCGC
CGAAATCCCGAATCTCTATCGTGCGGTGGTTGAACTGCACACCGCCGACGGCACGCTGATTGA
AGCAGAAGCCT

Appendix C: Supplemental Material for Chapter 4

Supplemental Figures

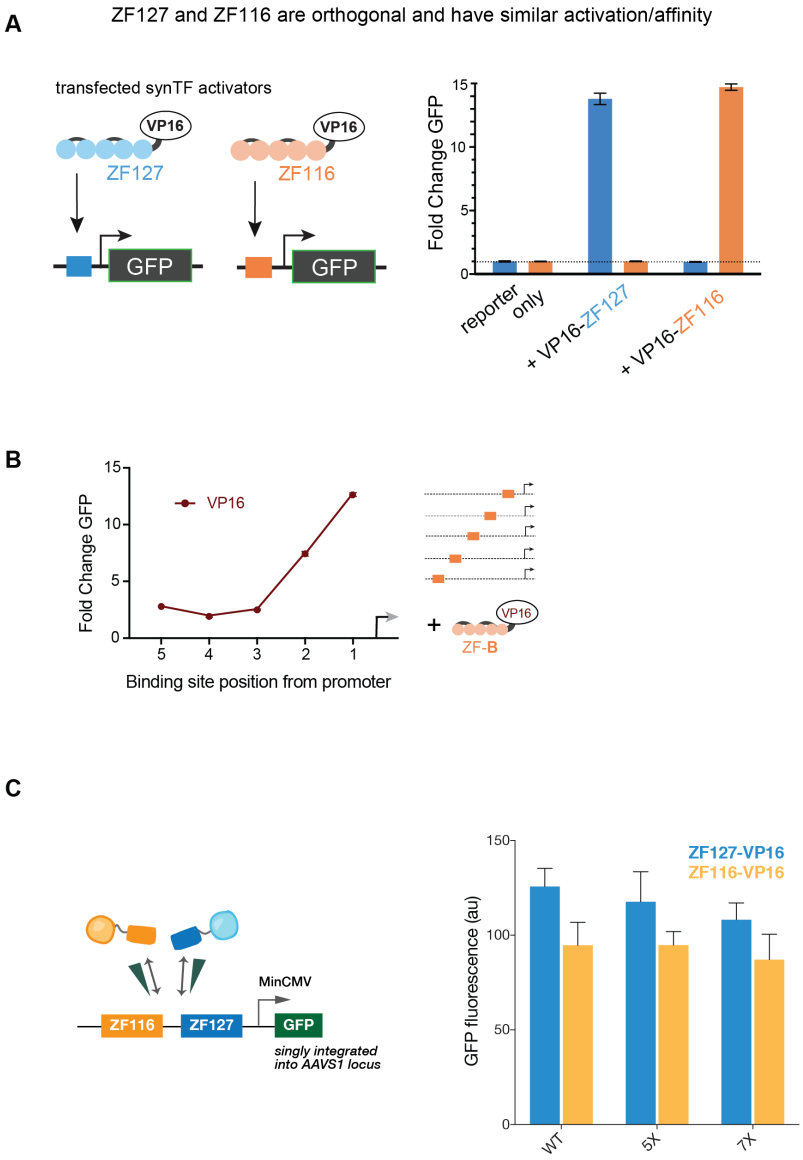


Figure S4.2: Characterizing ZF-116 and ZF-127. *(*A) We fused two ZF binding domains (ZF-116 and ZF-127) to the VP16 activation domain and transfected them into HEK293T cells where they bind with high specificity to their cognate sequences, which we placed upstream of the minimal CMV promoter driving expression of a GFP reporter. Experiments carried out in triplicate; error bars are standard deviation of the mean. (B) Fold change expression driven from a single site bound by VP16 synTF decreases as the site is moved away from the promoter. (C) The site closest to the promoter drives slightly more expression. We measured activation driven by VP16 fused to ZF-116 (yellow, upstream) or ZF-127 (blue, promoter-proximal) and see that for all mutants, expression is slightly higher when VP16 is bound to the closer site. Experiments carried out and plotted as in (A).

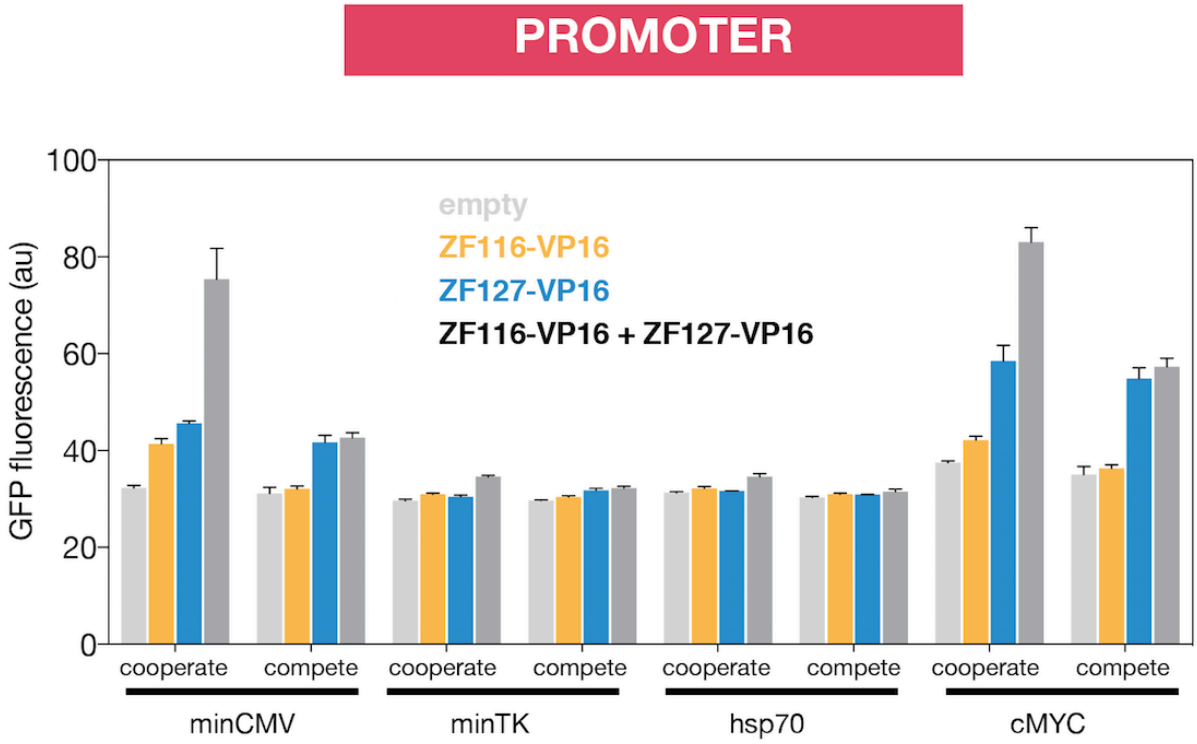


Figure S4.1 : Not all promoters are activated by a VP16 synTF. We tested the minCMV, minimal thymidine kinase (minTK), mammalian hsp70 and mammalian cMYC promoters for activity in our assay. In each case, we cloned the promoter upstream of either a site for ZF-116 and ZF-127 (1 site each, labelled "cooperate") or a single site for ZF-127 (labelled "compete), transfected a synTF fused to the strong viral activation domain VP16, and measured reporter (GFP) expression using flow cytometry. Both minCMV and cMYC responded in this assay, no activation of the minTK or hsp70 promoter was detected. Experiments performed in triplicate, average of three technical replicates is plotted with error bars indicating standard deviation (SD).

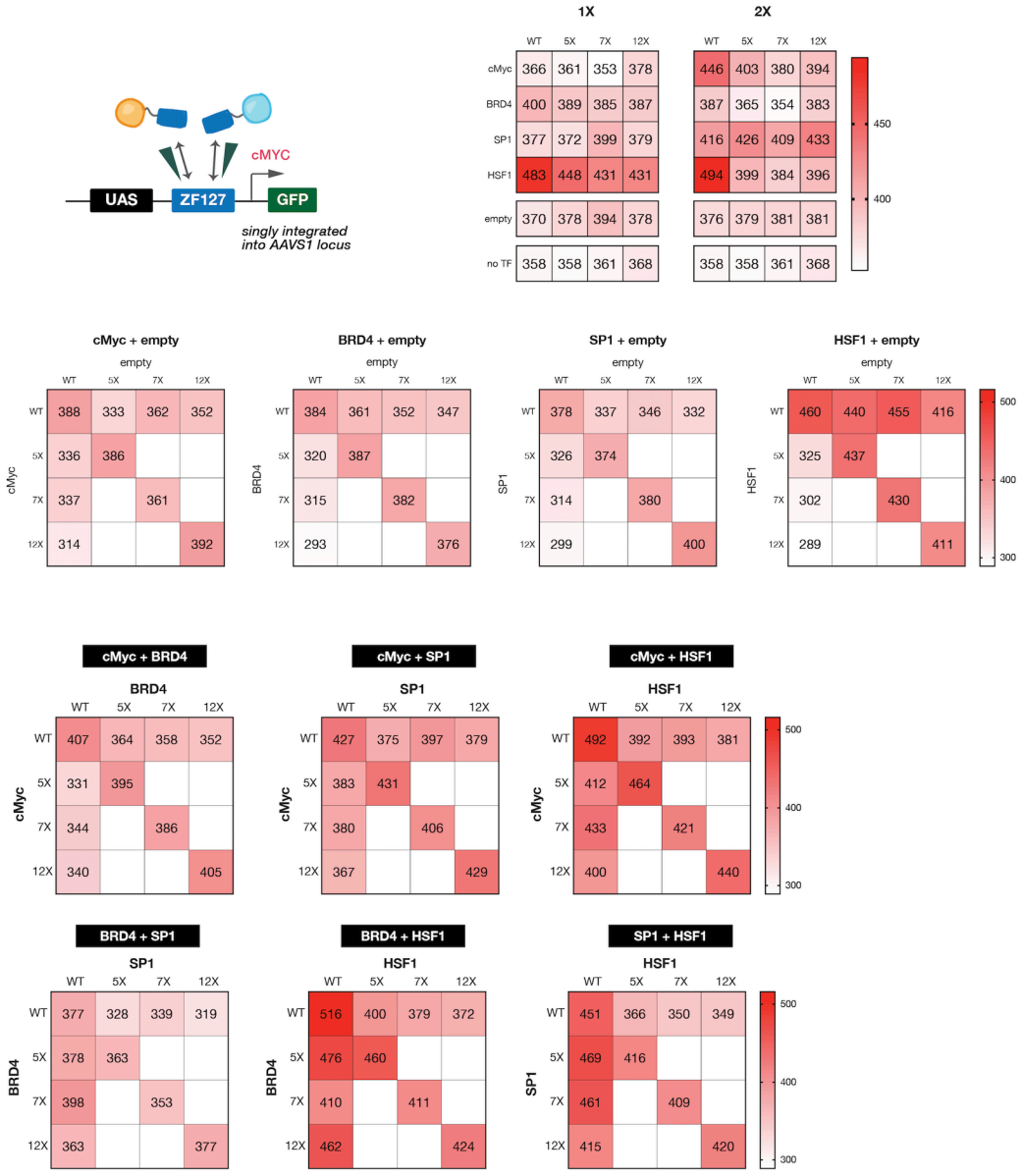


Figure S4.3 : Combinatorial behavior changes with affinity at the cMYC promoter. (Top)We also measured how TFs "time-share" at the cMYC promoter using a single ZF-127 site genetically integrated into HEK293T cells. Expression output driven by 10ng single synTFs at this promoter is shown as a heatmap scaled from lowest (white) to highest (red) with mean values from three technical replicates in each cell. The right panel (2X) is the same experiment with the amount of transfected DNA doubled. We use this as a control for doubling the amount of synTF DNA transfected when looking at pairs rather than single synTF controls. In all cases, doubling the amount of synTF transfected increases expression slightly. (Middle) GFP expression for all single synTFs and "empty" competitor, with varying affinities. The number of arginine to alanine mutations in the DNA binding domain is denoted along the top horizontal or vertical axis and all combinations of TF pairs are shown. (Bottom) GFP expression

Figure S4.3 (continued).
for all pairwise synTF combinations. In each graph, affinity for one synTF varies along the vertical axis and the other is varied along the horizontal axis. Along the diagonal, both TFs have the same binding domain. The number of arginine to alanine mutations in the DNA binding domain is denoted along the top horizontal or vertical axis and all combinations of TF pairs are shown.

Sequences used for synTF activation domains

SP1. Residues 263 – 499.

AACATCACCTTGCTACCTGTCAACAGCGTTTCTGCAGCTACCTTGACTCCCAGCTCTCAGGCAG
TCACGATCAGCAGCTCTGGGTCCCAGGAGAGTGGCTCACAGCCTGTCACCTCAGGGACTACCA
TCAGTTCTGCCAGCTTGGTATCATCACAAGCCAGTTCCAGCTCCTTTTTCACCAATGCCAATAG
CTACTCAACTACTACTACCACCAGCAACATGGGAATTATGAACTTTACTACCAGTGGATCATCA
GGGACCAACTCTCAAGGCCAGACACCCCAGAGGGTCAGTGGGCTACAGGGGTCTGATGCTCTG
AACATCCAGCAAAACCAGACATCTGGAGGCTCATTGCAAGCAGGCCAGCAAAAAGAAGGAGAG
CAAAACCAGCAGACACAGCAGCAACAAATTCTTATCCAGCCTCAGCTAGTTCAAGGGGGACAG
GCCCTCCAGGCCCTCCAAGCAGCACCATTGTCAGGGCAGACCTTTACAACTCAAGCCATCTCCC
AGGAAACCCTCCAGAACCTCCAGCTTCAGGCTGTTCCAAACTCTGGTCCCATCATCATCCGGAC
ACCAACAGTGGGGCCCAATGGACAGGTCAGTTGGCAGACTCTACAGCTGCAGAACCTCCAAGT
TCAGAACCCACAAGCCCAAACAATCACCTTAGCCCCAATGCAGGGTGTTTCCTTGGGGCAGACC
AGCAGCAGCAAC

VP16. Residues 410-490.

CTGTCGACGGCTCCACCAACTGATGTTTCATTGGGTGATGAATTGCATTTGGATGGAGAAGAT
GTTGCTATGGCTCATGCTGATGCATTAGATGATTTCGATTTAGACATGTTGGGAGATGGTGATT
CTCCAGGTCCAGGTTTCACTCCACATGATTCTGCTCCTTACGGTGCATTGGATATGGCTGATTT
TGAATTCGAACAAATGTTCACTGATGCTTTAGGAATTGATGAATATGGTGGTTAA

HSF1. Residues 370-529.

CCTGAAAAGTGCCTCAGCGTAGCCTGCCTGGACAAGAATGAGCTCAGTGACCACTTGGATGCT
ATGGACTCCAACCTGGATAACCTGCAGACCATGCTGAGCAGCCACGGCTTCAGCGTGGACACC
AGTGCCCTGCTGGACCTGTTCAGCCCCTCGGTGACCGTGCCCGACATGAGCCTGCCTGACCTTG
ACAGCAGCCTGGCCAGTATCCAAGAGCTCCTGTCTCCCCAGGAGCCCCCCAGGCCTCCCGAGG
CAGAGAACAGCAGCCCGGATTCAGGGAAGCAGCTGGTGCACTACACAGCGCAGCCGCTGTTCC
TGCTGGACCCCGGCTCCGTGGACACCGGGAGCAACGACCTGCCGGTGCTGTTTGAGCTGGGAG
AGGGCTCCTACTTCTCCGAAGGGGACGGCTTCGCCGAGGACCCCACCATCTCCCTGCTGACAG
GCTCGGAGCCTCCCAAAGCCAAGGACCCCACTGTCTCCTAG

cMyc. Residues 1-70

ATGGATTTTTTTCGGGTAGTGGAAAACCAGCAGCCTCCCGCGACGATGCCCCTCAACGTTAGCT
TCACCAACAGGAACTATGACCTCGACTACGACTCGGTGCAGCCGTATTTCTACTGCGACGAGGA
GGAGAACTTCTACCAGCAGCAGCAGCAGAGCGAGCTGCAGCCCCCGGCGCCCAGCGAGGATAT
CTGGAAGAAATTCGAGCTG

BRD4. Residues 1308-1362

CCACAGGCCCAGAGCTCCCAGCCCCAGTCCATGCTGGACCAGCAGAGGGAGTTGGCCCGGAAG
CGGGAGCAGGAGCGAAGACGCCGGGAAGCCATGGCAGCTACCATTGACATGAATTTCCAGAGT
GATCTATTGTCAATATTTGAAGAAAATCTTTTCTGA