UNIVERSITY OF CALIFORNIA,
IRVINE

# Demystifying the relationship between DNA sequence features and regulatory function

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Lily Li

Dissertation Committee:
Assistant Professor Zeba Wunderlich, Chair
Professor Ali Mortazavi
Associate Professor Kevin Thornton
Associate Professor Rahul Warrior
Professor Kyoko Yokomori

2021

# DEDICATION

To the me seven years ago
who started down this road
not knowing where it would take you.
You've grown so much and in such unexpected ways.
Let's look towards the future with anticipation!
(even if there's a tinge of anxiety mixed in there).

And to the countless friends, old and new,
who have made me the person I am today.
I wouldn't have been able to do this without you.
You remind me every day why this life is worth living.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

I want to thank my advisor, Zeba Wunderlich, who has always had her door open to us. Joining her lab when it was just starting was a leap of faith, but I believed that she was building a lab in which I would be able to drag myself out of bed and keep moving forward even when things weren't working. And my time in her lab has proved that to be true. These seven years have been tumultuous and hard in unexpected ways, but through it all, she has been supportive and understanding. She has given me room to grow and believed in me even when I didn't believe in myself. And she created a truly safe, nurturing environment in which we could be creative, express ourselves honestly and freely, and help each other grow. I am so thankful for this chance to learn and grow as part of this lab.

I am indebted to my labmates (Bryan Ramirez-corona, Marley Hilleger, Rachel Waymack, Kevin Cabrera, Lianne Cohen, Flo Ramirez, Arash Abiri, Subhapradha Rangarajan, Mario Gad, Ariana Lee, among others) for their feedback on my research over the years and for their friendship throughout grad school. Rachel and Mario's expertise were indispensable for the cloning of the many reporter constructs, and all of them created a warm and nurturing place in which to grow together.

I want to thank my committee members, Ali Mortazavi, Kevin Thornton, Rahul Warrior, and Kyoko Yokomori. Their probing questions about my research have been crucial to my development as a scientist. Specifically, Dr. Mortazavi was my first rotation advisor and has been a mentor throughout my time in graduate school. He was the first professor here that I felt comfortable chatting with, and he suggested that I rotate in Zebas lab.

I want to thank my parents who are always there for me and who always want the best for me. Thank you for working with me to try to understand each other better despite the difficulties!

I also want to thank my amazing MCB cohort. We went through so much together, and I'm so glad that I was able to be a part of this particular group. You've been like a safety net for me ever since we started bootcamp, and you've always there with a supportive word or some fun shenanigans.

Thank you to Debra Mauzy-Melitz and my fellow GAANN fellows, who taught me so much about the process of teaching and encouraging students to think critically.

I also want to extend my thanks to the MCSB administrative staff—Karen Martin, Naomi Carreon, Cely Dean, and Tina Rimal—and the Developmental and Cell Biology administrative staff—Lindsay Malter Simmons, Andrea Marie Wiley, and Grace Kuei—who were the first to greet us and create a welcoming environment and continued help us throughout our time in graduate school.

A thank you to all the other people I met in and because of grad school; you've made this experience so special, and now I can't imagine my life without you guys. And finally, all the old friends that have stuck around all these years—I'm truly grateful to have you in my life.

This work was supported in part by the Graduate Assistance in Areas of National Need (GAANN) fellowship from the US Department of Education, grant number P200A120207, as well as, the T32 training grant from the National Institute of Biomedical Imaging and Bioengineering (NIBIB), grant number EB009418.

Use of material in Chapter 2 has been granted with permission by *Frontiers in Genetics*.

# VITA

## Lily Li

### EDUCATION

**PhD in Biological Sciences**: UC Irvine, Irvine, CA 2021

- Mathematical and Computational Biology (MCB) Gateway Program

**BS in Bioengineering** with a **minor in English**: Caltech, Pasadena, CA 2012

### AWARDS

Graduate Assistance in Areas of National Need (GAANN) Fellowship 2017 - 18

Center for Complex Biological Systems Opportunity Award: Mechanisms of Disease 2016

- Dynamics of *Xic* lncRNAs in mouse embryonic stem cells and human cancer cells

National Science Foundation GRFP Honorable Mention 2016

NIH-NIBIB T32 Fellowship: Mathematical, Computational, & Systems Biology 2015 - 17

### RESEARCH EXPERIENCE

**Graduate Student Researcher, Wunderlich Lab** Apr 2015 - Jun 2021
UC Irvine, Irvine, California

- Combined statistical/physically-based models with imaging data to elucidate mechanisms of regulatory DNA function
- Integrated new scientific findings to update understanding and thus modify current projects of all team members
- Mentored 4 undergrad/grad students, resulting in research awards and acceptance into a highly competitive MD/PhD program
- Effectively presented research findings to diverse audiences at international conferences and in 1 first-author journal article

**Research Associate** Jul 2013 - Jul 2014
City of Hope National Medical Center, Duarte, CA

- Collaborated with PI and clinical fellows to write grants, IRB/IACUC protocols and revise papers
- Coordinated with surgeons from multiple institutions to collate, edit and illustrate *Surgery for Cancers of the Gastrointestinal Tract*, a textbook for surgical oncologists

## RESEARCH EXPERIENCE CTD

**Lab Technician:** Sep 2012 - Jun 2013
Harvey Mudd College, Claremont, CA

- Collaborated with the Dresch group in the Math Department to evaluate quantitative patterns in structure of regulatory DNA
- Worked in conjunction on multiple projects with undergraduates

**Junior Specialist:** Sep 2011 - Sep 2012
UC Irvine, Irvine, CA

- Studied the impact of HSV-1 latency-associated transcript fragments on apoptosis, resulting in 1 peer-reviewed publication
- Trained 5 students in molecular biology techniques and image processing software

## PUBLICATIONS

**Li L** and Wunderlich Z. (2017) An enhancer's length and composition are shaped by its regulatory task. *Front Genet.* doi: 10.3389/fgene.2017.00063

Lee S, Heinrich E, **Li L**, Lu J, Choi A, Levy R, Wagner J, Yip MLR, Vaidehi N, Kim J. (2015) CCR9-mediated Signaling through $\beta$-catenin and Identification of a Novel CCR9 Antagonist. *Mol Oncol.* 9(8): 1599 – 611.

Jiang X, Brown DJ, Osorio N, Hsiang C, **Li L**, Chan L, BenMohamed L, Wechsler SL. (2015) A herpes simplex virus type 1 mutant disrupted for microRNA H2 with increased neurovirulence and rate of reactivation. *J Neurovirol.* 21(2): 199 – 209.

*Illustrated:* Kim, Joseph, and Julio Garcia-Aguilar, eds. *Surgery for Cancers of the Gastrointestinal Tract: A Step-by-Step Approach.* New York: Springer, 2015. Print.

Luyimbazi D, Nelson R, Choi A, **Li L**, Chao J, Sun V, Hamner J, Kim J. (2014) Estimates of Conditional Survival in Gastric Cancer Reveal a Reduction of Racial Disparities with Long-Term Follow-up. *J Gastrointest Surg.* 19(2): 251 – 7.

Merchant SJ, **Li L**, Kim J. (2014) Racial and Ethnic Disparities in Gastric Cancer Outcomes: More Important than Surgical Technique? *World J Gastroenterol.* 20(33): 11546 – 11551.

Drewell RA, Nevarez MJ, Kurata JS, Winkler LN, **Li L**, Dresch JM. (2014) Deciphering the combinatorial architecture of a *Drosophila* homeotic gene enhancer. *Mech Dev.*131:68 – 77.

## RESEARCH PRESENTATIONS

**2020 EMBL: Transcription and Chromatin**, Virtual Aug 2020
**Li L**, Wunderlich Z.
Elucidating mechanisms by which the motif content of regulatory DNA tunes expression output.

## RESEARCH PRESENTATIONS CTD

**2018 Association for Biology Education (ABLE)**, Columbus, OH Jun 2018
**Li L**,Tarapore E, Chapman D, Mauzy-Melitz D. Barking up a storm: what dog DNA testing can tell us about statistical errors.

**2018 SABER West**, Irvine, CA Jan 2018
Karner H, **Li L**, Tormanen K, Mauzy-Melitz D. Strategies for Undergraduate Scientific Success: Resolving the Disconnect.

**58th *Drosophila* Research Conference**, San Diego, CA Apr 2017
**Li L**, Wunderlich Z. An enhancer's length and composition are shaped by its regulatory task.

**2016 NIBIB Training Grantees Meeting**, Bethesda, MD July 2016
**Li L**, Wunderlich Z. Dissecting Enhancer Architecture in the Developing *Drosophila*.

**2016 Winter q-bio Conference**, Oahu, HI Feb 2016
**Li L**, Wunderlich Z. Dissecting Enhancer Grammar in the Developing *Drosophila*.

## SERVICE

**Outcomes Officer:** Association for Women in Science (AWIS) Summer 2018 - 2020

- Evaluated our progress towards providing resources for women in science and helping faciliate cross-disciplinary interaction and use feedback to adjust approach
- Developed partnerships with industry to increase awareness and interest in diverse careers outside of academic research
- Coordinated with other officers to organize mentorship program for women in STEM with a mix of industry and academic mentors

**Fellow:** Graduate Assistance in Areas of National Need (GAANN) Sep 2017 - Jun 2018

- Collaborated with 10+ fellows to design and deliver a workshop to increase diversity of undergraduate researchers by teaching them how to find, join, and understand expectations in a research lab
- Conducted qualitative research on faculty's criteria for selecting student researchers and communicated this to students
- Worked with other fellows to develop and present a module for explaining introductory statistical analyses for DNA testing to college and university biology educators

**Website Manager & Camp Counselor:** AAUW NMI Aug 2016 - Feb 2018

- Consolidated feedback from members to update the chapter website to reflect upcoming events and encourage community involvement
- *Tech Trek:* Encouraged underprivileged middle school girls to pursue math, science and higher education at a science summer camp

## TEACHING EXPERIENCE

**Developmental and Cell Biology Lab** Spring 2017, Spring 2018

- **Head TA**: Developed quizzes and rubrics
- Worked with students to develop their independent projects, analyze their data in R, and craft a thorough lab report
- Graded weekly lab notebooks and lab reports

**Biology and Chemistry of Food and Cooking** Winter 2016

- Held weekly office hours and discussions for exam preparation
- Assisted with final project design

**Systems Biology Short Course** Jan 2016

- Assisted in teaching the module "Shadow Enhancers and Robustness in the *Drosophila* embryo"

**Mathematical & Computational Biology (MCB) Bootcamp** Sep 2015, Sep 2016

- Helped teach incoming MCB graduate students Matlab and Mathematica and work with them on critical thinking and reading of biology papers

**California State Summer School for Mathematics and Science** Jun - Jul 2015

- Taught Matlab and hands-on mathematical and computational modeling labs as a teaching assistant in a month-long summer enrichment program for high school students across California

# ABSTRACT OF THE DISSERTATION

Demystifying the relationship between DNA sequence features and regulatory function

By

Lily Li

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2021

Assistant Professor Zeba Wunderlich, Chair

All cellular processes from development to homeostasis depend on precise spatiotemporal gene expression. This precision is mediated by two pieces of regulatory DNA, enhancers and promoters, which integrate signals from activating and repressive transcription factors (TFs). Understanding how gene expression is encoded in these pieces of regulatory DNA is the larger question in the field that we hope to better understand. Here we approach this goal by tackling two questions.

First, we consider the complexity of a regulatory task as a potential organizing principle for how expression is encoded in enhancers. We define task complexity as the number of fates specified in a set of cells at once. We hypothesized that more complex regulatory tasks would be encoded in longer enhancers with more binding sites, as more binding sites can be rearranged within an enhancer in more ways. This allows for the specification of a wider variety of expression patterns, and therefore, more complex tasks. To test this hypothesis, we compared 100 enhancers that specify the complex anterior-posterior (AP) and the simpler dorsal-ventral (DV) axis patterning system. We also validated this hypothesis using a larger dataset of enhancers active across development, where we would expect task complexity to decrease over time. In both cases, we found that increased decision complexity is encoded in longer enhancers with more TF binding sites.

Second, we consider the role of multiple promoters in the context of a gene locus. Many genes have multiple enhancers and promoters. However, while the role of multiple enhancers in a gene locus has been studied, little work has been done to explicate the roles of multiple promoters for a single gene, especially when these promoters lead to the production of similar or identical isoforms. Here, we propose that like multiple enhancers, multiple promoters can provide redundancy like shadow enhancers or specificity by preferentially engaging with specific enhancers. To distinguish between these two roles, we chose a case study gene, knirps, that has multiple enhancers and promoters that each has different motif content and thus recruits different sets of proteins. As we expect that specificity is mediated by the proteins recruited to the enhancer and promoter, this set of enhancers and promoters allows us to test whether these promoters provide redundancy or specificity.

Using synthetic reporter constructs, we found that some, but not all, enhancers in the locus show a preference for one promoter. By analyzing the dynamics of these reporters, we identified specific burst properties during the transcription process, namely burst frequency and size, that are most strongly tuned by the specific combination of promoter and enhancer. Using locus-sized reporters, we discovered that even enhancers that show no promoter preference in a synthetic setting have a preference in the locus context. Our results suggest that the presence of multiple promoters in a locus is both due to enhancer preference and a need for redundancy and that broad promoters with dispersed transcription start sites are common among developmental genes. Our results also imply that it can be difficult to extrapolate expression measurements from synthetic reporters to the locus context, where many variables shape a gene's overall expression pattern.

# Chapter 1

# Introduction

Diverse processes in biology, from development to the maintenance of homeostasis, rely on the regulation of gene expression. Gene expression programs are largely encoded in the genome by two main pieces of regulatory DNA, enhancers and promoters, and depend on the proteins that bind them. To encode such diverse processes of differing complexity, enhancers and promoters have been found to vary widely in size, shape, and motif content, including the affinity, number, orientation, and spacing of transcription factor (TF) binding sites. By recruiting different combinations of TFs in different orientations and spacings, they act as platforms for signal integration, allowing organisms to respond dynamically to developmental and environmental cues. Thus, they are both able to respond to cues at the appropriate time and place as well as produce the desired expression levels and noise for a particular situation.

Sequence alone can affect expression output, but in the context of the organism, higher-order organization also plays a role. The distance between the enhancer and promoter will affect their frequency of interaction [161], and this interaction frequency is further modulated by the 3D chromatin architecture, as most enhancer-promoter interactions are contained within topologically-associated domains (TADs)[31, 68]. In addition, proteins like insulators can

further limit and/or change the frequency and stability of enhancer-promoter interactions, resulting in different expression output [83, 112].

While there is substantial evidence that chromatin architecture affects gene expression [42, 60, 66, 104, 146], it is not evident whether transcription leads to the formation of chromatin loops or vice versa. At least in *Drosophila*, experiments evaluating chromatin architecture using genome-wide chromosome capture (Hi-C) in ventralized embryos that only produce dorsal ectoderm, neuroectoderm, and mesoderm demonstrated that TADs and chromatin loops are maintained across tissue types in blastoderm embryos despite differences in gene expression and chromatin state [67]. Even at the small number of regions that do change in chromatin organization, these changes do not appear to be associated with changes in expression. While there may exist some relatively stable long-range enhancer-promoter interactions [49], a lack of widespread enrichment of these interactions suggests that these loops are not the primary mechanism by which enhancers and promoters interact during *Drosophila* development, especially as most enhancers are located near their target promoters [67].

Thus, higher-order organization may not play a large role during the blastoderm stage of *Drosophila* development [67] even though it may influence gene expression by modulating the ability of the enhancer and promoter to interact and be important in mammalian systems and later in development [11, 27, 64, 68, 78, 93, 115, 135]. In fact, independent of higher organization, the DNA sequence of enhancers and promoters is sufficient to encode expression patterns and levels in a non-modular fashion. As demonstrated by Gehrig et al., non-modular changes in expression output were driven by different combinations of enhancer-promoter pairs even though the distance between them was kept constant and the reporter construct was not integrated into the chromosome [46]. Thus, we are mainly focused on how sequence changes affect expression output.

Despite extensive study of transcriptional regulation, the design principles that underlie how

regulatory DNA is put together are still largely a mystery. The magic of DNA is that it seems so simple, just four nucleotides in varying order. And yet, it is responsible for instructing the complex and precise choreography of the cell. Here, we explore two questions that add to our understanding of how function is encoded in regulatory DNA. In Chapter 2, I address how regulatory function can shape the structure of enhancers, and in Chapter 3, the role of multiple promoters for a single gene. To provide a larger context for these two questions, I will discuss how enhancers work, how promoters work, how we measure and model the process of transcription, and how we can use this model to make sense of the underlying molecular mechanisms tuned by enhancers and promoters.

## 1.1 To control transcription in a broad array of settings, enhancers vary widely in structure

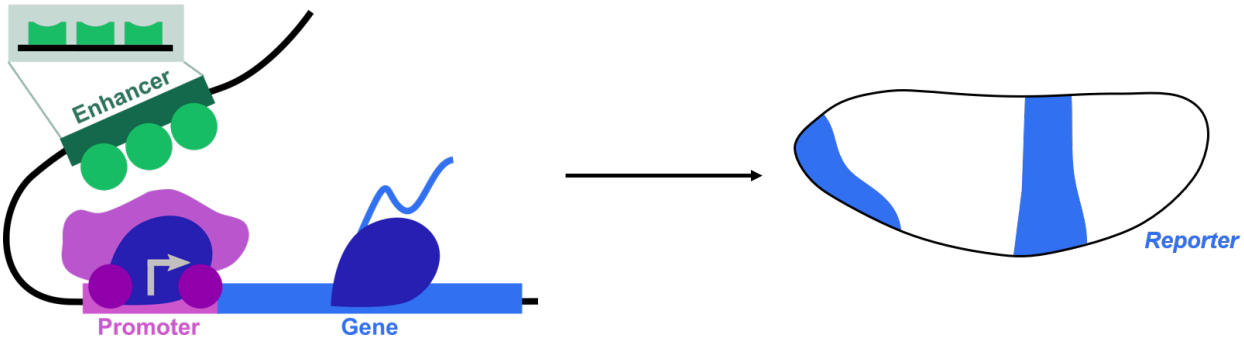


**Figure 1.1:** Enhancers consist of clusters of TF binding sites (dented green rectangles) for activators and repressors (green circles). The integration of these signals determines when and where genes get expressed.

Enhancers are *cis*-acting DNA sequences that direct when and where transcription occurs (**Figure 1.1**). They can consist of a few to tens of transcription factor (TF) binding sites for both activating and repressive factors [9, 175]. By acting as a platform of signal integration, enhancers can precisely control when and where genes are expressed, making possible processes as disparate as immune response and development.

Enhancers regulating these two processes represent two extremes of the spectrum of how enhancers can encode function. On one end of the spectrum are immune-response enhancers, like that of the human interferon-beta gene [150], which nucleates the cooperative assembly of a nucleoprotein complex called the enhanceosome [107]. These enhancers represent the most rigid model of how function is encoded in sequence, as the arrangement of TF binding sites is precisely defined such that nearly every nucleotide is bound [118]. On the other end of the spectrum are developmental enhancers, which are better explained by the billboard model [79]. This model suggests that as long as the TF binding sites are present, their arrangement is flexible. While no enhancer has been shown to have this level of flexibility, most enhancers likely fall somewhere in the middle of this spectrum, with spacing and arrangement important for some TFs and more flexible for others. By changing the number, orientation, affinity, and arrangement of the TF binding sites, enhancers can handle disparate regulatory tasks.

### 1.1.1 Tweaking the affinity and arrangement of TF binding sites is important for tissue-specific expression

The precise spatiotemporal expression patterns observed during development require the usage of TF binding sites of varying affinities. While the difficulty of identifying low-affinity degenerate sites has led to a focus on high-affinity sites, low-affinity sites are important in regulating gene expression in multiple organ systems from flies to humans [69]. As increasing the affinity of these sites leads to ectopic expression, enhancers can produce both tissue-specificity and robust expression by either optimizing the arrangement of clusters of low-affinity sites or by having multiple weak enhancers, each consisting of sets of suboptimally arranged TF binding sites [39, 40].

The arrangement of TF binding sites is also important for achieving tissue-specific expression. The orientation and order of sites as well as the spacing between sites determine the possible

interactions between the bound TFs as well as higher-order interactions. Thus, changing the arrangement of sites may disrupt such protein-protein interactions, leading to changes in expression levels and/or pattern [19, 25, 76, 100, 110, 140, 142, 149, 150]. Some scenarios in which the arrangement might be important include the formation of homo- or heterodimers, protein-protein interactions made possible by the helical phasing of sites, and short- or long-range repression depending on the repressors in question.

## 1.2 Biological and evolutionary constraints shape enhancer structure

Enhancers can vary greatly in structure, but that structure is constrained by multiple factors. As previously noted, an enhancers regulatory task, whether it is to respond to infection or to produce precise patterns of expression, can affect enhancer structure. In the case of the subset of immune response enhancers that are platforms for the assembly of enhanceosomes, the tightness of the structure of the assembled protein complex provides rigid limits to the enhancer structure [118].

In the case of tissue-specific expression, the need for specificity may constrain the enhancer structure by requiring suboptimization of TF binding site affinity and/or arrangement, as previously described. In addition, enhancer structure may be affected by the abundance and availability of TFs, the temporal dynamics of TFs, the need to couple multiple developmental processes [89], cell type, and developmental time point [69]. In fact, the evolutionary constraints on gene regulation exhibit an hourglass shape, with the bottleneck occurring mid-embryogenesis [101, 113], suggesting that this period has the most rigid evolutionary constraints on enhancer structure. In contrast to these constraints, the presence of shadow enhancers, enhancers of the same gene that drive overlapping expression patterns [82], may

relax these constraints while producing interesting interdependencies between the changes in the shadow enhancers. We explore how regulatory task can shape enhancer structure in Chapter 2.

The mechanisms by which transcription occurs may also constrain enhancer structure. Encoding enhancer-promoter specificity constrains the motif content and structure of enhancers, given evidence that particular TFs depend on some promoter motifs over others [16, 70]. The ability of an enhancer to activate a promoter may be affected by its proximity, with a proximal enhancer experiencing more rigid constraints to achieve assembly of a functional pre-initiation complex [108, 131]. More distal enhancers may interact with promoters through a variety of mechanisms, including looping, transcriptional hubs, topologically associating domains (TADs), etc. Which mechanisms are at play likely shape and constrain the enhancer structure in different ways. For example, formation of a transcriptional hub may depend on a particular enhancer structure, type of TF, or even the collection of enhancers in a TAD. On the other hand, the presence of a hub may make it possible for an enhancer to rely on lower affinity sites [97, 157, 158]. Chromatin dynamics, which are affected by the recruitment of pioneer TFs or TFs with chromatin remodeling capabilities, may also shape enhancer structure [3, 48, 103].

## 1.3 Promoters also vary in structure and exhibit distinctive structural signatures based on their regulatory role

Core promoters, defined as ±40bp around the transcription start site (TSS) at which RNA PolII binds and initiates transcription, are mainly considered modulators of expression levels [71]. Like enhancers, promoters act as platforms of signal integration. This is encoded in

the form of binding motifs for proteins required to initiate transcription, including RNA polymerase II (RNA Pol II), various general transcription factors (gTFs), and Mediator (**Figure 1.4**). Each of these proteins is a multisubunit complex, and the subunits present in these complexes are probabilistically determined by the motifs present in the promoter. In addition, the abundances of these subunits may also be cell-type-specific [77, 168, 167, 181]. In this way, the motif content of promoters can encode and tune expression output with exquisite precision in response to cellular and environmental signals.

While this is a reasonable framework to understand how promoter structure is shaped by their regulatory task, our understanding of how the motif content of promoters affects recruitment of transcriptional machinery and thus transcription is still an active field of study. Here, we will discuss how promoters can vary in structure and what we know about how that structure affects transcription.

### 1.3.1 Promoter shape

Promoters are characterized by a spectrum of shapes, or patterns of initiation (**Figure 1.3**) [162]. A focused or sharp promoter initiates transcription at a single site, or a narrow cluster of sites within five nucleotides, and is usually associated with developmental genes. This likely allows for more precise control of transcription initiation. In contrast, a dispersed or broad promoter initiates from a series of weak sites across 50-100 nucleotides and tends to be associated with ubiquitously expressed genes (often housekeeping genes). Sharp and broad promoters represent the two ends of the spectrum of promoter shape, with mixed promoters, consisting of a series of weak initiation sites and one strong dominant site, falling in between. (Analysis of nascent transcripts rather than steady-state RNA may further define the distribution of promoter shapes found in organisms). The existence of a spectrum of promoter shapes and the association of broader promoters with housekeeping genes and

sharper promoters with developmental genes suggests that promoter shape is influenced by the regulatory task of the promoter. Do housekeeping promoters tend to be broader than developmental promoters because they must successfully recruit transcriptional machinery without much reliance on enhancers? Do broader promoters form a transcriptional hub of transcriptional machinery? A better understanding of how shape influences a promoters ability to recruit transcriptional machinery and interact with enhancers will allow us to elucidate the principles of promoter design.

### 1.3.2 Promoter motifs

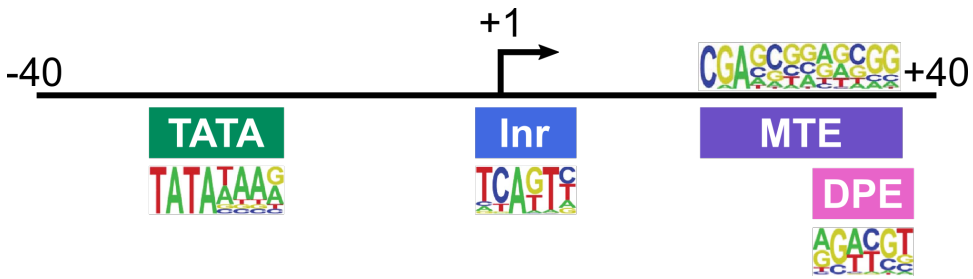


**Figure 1.2:** Some promoter motifs involved in transcription by RNA polymerase II and their sequence logos. Their locations with respect to the transcription start site (TSS) are drawn to scale. The motifs illustrated here have been found in both *Drosophila* and humans and are more typically found in developmental promoters.

The shape of a promoter is determined by the arrangement of the cluster of motifs present at a promoter. These motifs encode activity by acting as recognition sites for a series of general TFs that assemble at and around the transcription start site (TSS) to guide RNA PolII into position. Altogether about 100 proteins necessary for transcription assemble at the initiation site. This pre-initiation complex (PIC) minimally consists of RNA polymerase II and six general TFs (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH at TATA Box-dependent promoters), with each of these proteins consisting of multiple subunits. In particular, TFIID, which comprises TATA-binding protein (TBP) in some form and 13-15 TBP-associated factors (TAFs), plays a key role in promoter motif recognition [156]. TAFs 1 to 13 are evolutionarily conserved between yeast, *Drosophila*, and humans [156]. As its

various subunits interact with specific promoter motifs, the motifs present help determine which subunits are necessary and their configuration.

Note that there are no promoter motifs that are universally found in every promoter. In fact, some promoters contain no known promoter motifs, which suggests that there are unknown motifs still to be discovered. As core promoter motifs have been best studied in sharp promoters, the focus here will be on those.

**Initiator** The initiator motif (Inr) is the most common core promoter motif in bilateria, with estimates of 46% and 63% in human and *Drosophila* promoters, respectively [47, 176]. It encompasses the transcription start site (TSS) and is recognized by TFIID, specifically its subunits TAF1 and TAF2. These subunits form a heterodimer, with TAF1 binding and TAF2 performing an accessory role [20, 170]. Interestingly TAF1 appears to recognize features of the DNA structure rather than a strict DNA sequence [20]. Several other core promoter motifs, including the downstream promoter motif (DPE) and motif ten element (MTE), recognize other TFIID subunits and work in conjunction with Inr to position the transcription machinery at the TSS. In fact, these motifs exist at strict distances from the Inr motif, with minor shifts severely disrupting their ability to facilitate TFIID assembly at the initiation site.

**TATA Box** The most well-known promoter motif is probably the TATA Box motif, which is named after its AT-rich sequence (**Figure 1.2**); it is located at a strict distance upstream of the initiation site and binds the TATA-binding protein (TBP), another subunit of TFIID [75, 73]. Binding of TBP produces a kink in the DNA, allowing for other components of the transcriptional machinery to bind [75, 73]. Interestingly, biochemical studies have shown that TBP does not bind with high orientation specificity, indicating that other promoter motifs may determine the orientation of proteins binding at the promoter [24].

The TBP subunit seems to activate TATA-dependent transcription while repressing DPE-dependent transcription [65]. Alternatively, TATA-binding protein-related factor 1 (TRF1) has been shown to replace TBP in vitro and shares functions with TBP [54, 62]. Both TBP and TRF1 recruit the general TFs TFIIA and TFIIB to the promoter; however, TRF1 may be expressed in a tissue-specific manner and recruits a multiprotein complex that is distinct from TFIID [26, 54]. This suggests that while functionally similar, TRF1 may play a role in tissue-specific activation of a subset of genes.

The TATA Box motif can work in conjunction with Inr or in the absence of Inr. In the absence of Inr, TFIIA stabilizes TBP binding to the TATA Box motif. In the presence of Inr, the strict spacing between the TATA Box and Inr motifs leads to TFIID assembly in a particular conformation; in this conformation, TFIIA binding to TBP induces a conformational change in TFIID that improves binding to the core promoter and increases transcription initiation at the promoter [37].

In mammals, this strong synergistic stimulation of promoter activity by Inr and TATA Box also requires the architectural factor HMGA1 (High Mobility Group AT-Hook1) and Mediator. While Mediator tends to be considered impartial to promoters, the interaction of HMGA1, Mediator, and TFIID seems to specifically stimulate transcription at TATA + Inr core promoters and is mediated by the C-tail domain of HMGA1 [174]. Note that the protein kinase CK2, which is necessary for DPE-dependent transcription [92], phosphorylates the C-tail domain of HMGA1 [133], opening up the possibility that post-translational modifications can be used to modify interactions between these proteins and affect its promoter-specificity. In addition, HMGA1 seems to selectively interact with a Mediator complex lacking the CDK8 subunit, suggesting that the composition of Mediator may also affect promoter activation and enhancer-promoter interactions. HMGA1 and Mediator also work together to inhibit the repressive effects of negative cofactor 2 (NC2) and Topoisomerase I on TATA-dependent transcription [174].

**DPE** The downstream promoter element (DPE) exists at a precise distance downstream of the Inr motif with which it cooperatively recruits TFIID to the initiation site (**Figure 1.2**) [15]. In fact, TFIID appears to interact closely with the DNA from the Inr through DPE, as evidenced by TFIID footprints showing a periodic 10-bp DNaseI digestion pattern [15, 80]. Photocrosslinking indicates that it is TAF6-TAF9 complexes that interact with DPE [143]. DPE-dependent transcription also appears to depend on the presence of casein kinase II (CK2) and positive coactivator 4 (PC4) in addition to the gTFs and Mediator [92]. The effect of these cofactors appears to be promoter motif-specific, as CK2 has been shown to diminish TFIID-specific binding to other motifs like the downstream core element (DCE) [92].

As previously noted, TBP appears to activate TATA-dependent transcription and repress DPE-dependent transcription; thus, negative cofactor 2 (NC2 a.k.a Dr1-Drap1) and modifier of transcription 1 (Mot1 a.k.a. Hel89B, BTAF1) oppose TBPs functions by targeting TBP. NC2 prevents association of TBP with factors like TFIIA and TFIIB [151], and Mot1 is an ATPase that interacts with NC2 and the DNA to displace TBP-NC2 in an ATP-dependent manner [17]. Both target TBP, thus repressing TATA-dependent transcription and relieving TBPs inhibition of DPE-dependent transcription.

In place of TBP, it appears that TATA-binding protein-related factor 2 (TRF2), which evolved from a duplicated TBP gene, lost its ability to bind TATA Boxes and thus developed the ability to regulate TATA-less promoters [34]. Specifically, TRF2 preferentially activates DPE-dependent promoters and TCT (or ribosomal gene) promoters [72, 164]. TRF2 maintains TBPs ability to recruit TFIIA and TFIIB.

While a significant portion of research has focused on comparing TATA- vs. DPE-dependent transcription, not many have considered the molecular interactions at promoters that have both TATA Box and DPE motifs. Interestingly, however, the *ftz* promoter, which contains both motifs, has been used multiple times in studies attempting to elucidate TF-promoter

motif preferences [70] or the effect of promoter motifs on transcriptional bursts [177]. Since promoters containing both TATA Box and DPE motifs are not uncommon, understanding them could be a valuable area of study. We explore some aspects of their molecular mechanisms in Chapter 3.

### 1.3.3 Promoter motifs are enriched depending on the promoters regulatory role

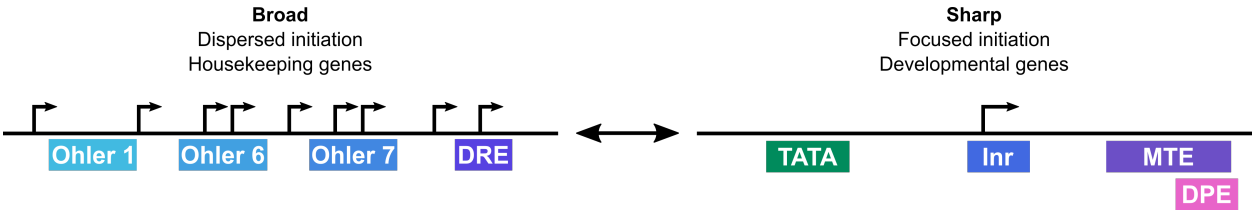


**Figure 1.3:** Promoter shape and motif content are shaped by the promoter's regulatory role. Promoter shape falls on a spectrum from broad (series of initiation sites over a 50-100bp region) to sharp (one or a few strong initiation sites over a 5bp region). Housekeeping genes tend to be associated with broad promoters, and are enriched for Ohler motifs 1, 6, and 7 and the DNA replication-related element (DRE). Developmental or regulated genes tend to be associated with sharp promoters and are enriched for Inr, TATA box, DPE, and MTE motifs.

The promoters regulatory role determines which promoter motifs are present. One example of a regulatory role is whether the gene needs to be expressed constitutively for the maintenance of cellular function (housekeeping gene) or whether the gene needs to be expressed at a precise point in time and space (developmental gene). As previously stated, housekeeping genes tend to have broader promoters, whereas developmental genes have sharper promoters (**Figure 1.3**). These two classes of promoters are also enriched for different motifs. In flies, housekeeping genes are mainly associated with Ohler motifs 1,6 and 7 and the DNA replication-related element (DRE), whereas, developmental genes are associated with Inr, TATA Box, DPE, and MTE, among others. The Ohler motifs were identified by using expressed sequence tag (EST) clusters to identify promoters and then looking for enrichment of motifs in these promoters. While many of the identified motifs were well-known, e.g. Ohler

motifs 2  4 and 9 are DRE, TATA Box, Inr, and DPE, the others were new and generally found to be associated with housekeeping genes [114].

Even within development, core promoters show preferential enrichment. Using a high-throughput TSS mapping experiment to profile promoter activity throughout development [6], researchers found that early embryogenesis involves the activation of promoters enriched for DRE, and Ohler 1 and 5-7, which notably are those associated with housekeeping genes. This likely reflects the minimal zygotic transcription that occurs early on in embryogenesis and the fact that housekeeping genes may not require the assistance of an enhancer to activate transcription. Intermediate embryogenesis is associated with Inr, DPE, and MTE, which are commonly found in developmental promoters. Interestingly, late embryogenesis is mainly associated with TATA Box motifs. The selective activation of these distinctive sets of core promoter motifs during different phases of embryogenesis suggests that core promoters are built to achieve specific regulatory tasks, in this case, broadly defining a window in which these genes can be activated by enhancers.

Just as a gene can have multiple enhancers, a gene can have multiple promoters. In fact, in flies, 40% of genes have more than one promoter [5]. While these promoters can drive the expression of distinct isoforms, there are many cases when they produce similar or identical transcripts. In this context, the role of multiple promoters has been relatively unstudied, but like multiple enhancers, they likely either provide some form of redundancy or an extra point of control. In either case, having multiple promoters may loosen the constraints on their evolution. We explore these roles in Chapter 3.

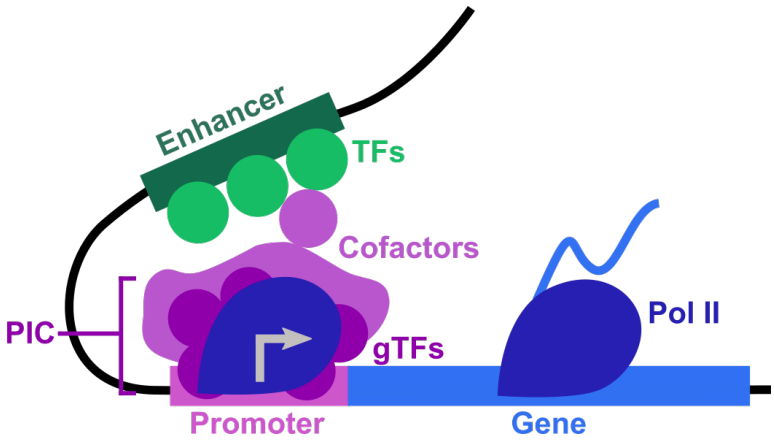


**Figure 1.4:** The combination of proteins recruited to the enhancer and promoter together determines the expression output of a gene.

## 1.4 Enhancers and promoters come together to modulate transcription

The combination of signal integration at both enhancers and promoters culminates in the activation of the transcriptional machinery recruited to the promoter (**Figure 1.4**). Activation may occur through enhancer-promoter looping or transcriptional hub formation. Recent experiments probing multi-way chromatin interactions at the $\alpha$-globin locus suggest that rather than forming mutually exclusive interactions with enhancers, the promoters and enhancers form a regulatory hub, allowing for simultaneous interaction without competition [116]. However, this does not preclude differential access to specific enhancers or promoters. A possible explanation of previous reports of promoter competition [4, 22, 30, 169] could be explained by a preponderance of transcriptional machinery localized to a proximal promoter possibly barring loop extrusion at a more distal promoter.

Both enhancers and promoters are clusters of TF binding sites, and we previously discussed how enhancer and promoter motif content and structure are shaped by regulatory roles. Given the recruitment of different TFs at housekeeping vs developmental enhancers and promoters [179], it should not be surprising that housekeeping and developmental promoters

interact with different enhancers. In fact, distinct sets of TFs were found to be enriched near core promoters depending on the stage of embryogenesis in which they were active [6]. Analysis of these preferred pairings shows that some TFs are strongly associated with particular sets of promoter motifs. The broad preferences of promoters regulating different phases of embryogenesis with specific TFs suggest that not only do their motif content encode enhancer-promoter specificity but they also allow for promoters to broadly define windows of transcription during embryogenesis with enhancers fine-tuning these windows.

In other bioinformatic studies that were functionally validated in cells, preferences of TFs for specific promoter motifs have also been shown, with enhancers that bind Caudal and Dorsal preferentially interacting with promoters containing DPE [70, 180]. Thus, a combination of the regulatory sequence of enhancers and promoters can encode in different transcriptional bursting patterns or expression output.

## 1.5 Measuring and modeling properties of the process of transcription allow us to decipher the molecular processes modulated by enhancers and promoters

Recent advances in visualizing transcription have been made possible due to the development of improved speed and sensitivity in fluorescence microscopy and RNA labeling approaches. These advances have made possible the ability to track transcription at a single locus in single cells in a real physiological context. Application of the RNA bacteriophage MS2 stem loop as an RNA tag such that nascent RNA is visualized with the binding of the MS2 coat protein (MCP)-GFP fusion protein has revealed the dynamics of transcription. Rather than a process that occurs at a continual and steady rate, transcription appears to be a pulsatile process, characterized by discontinuous bursts of activity.

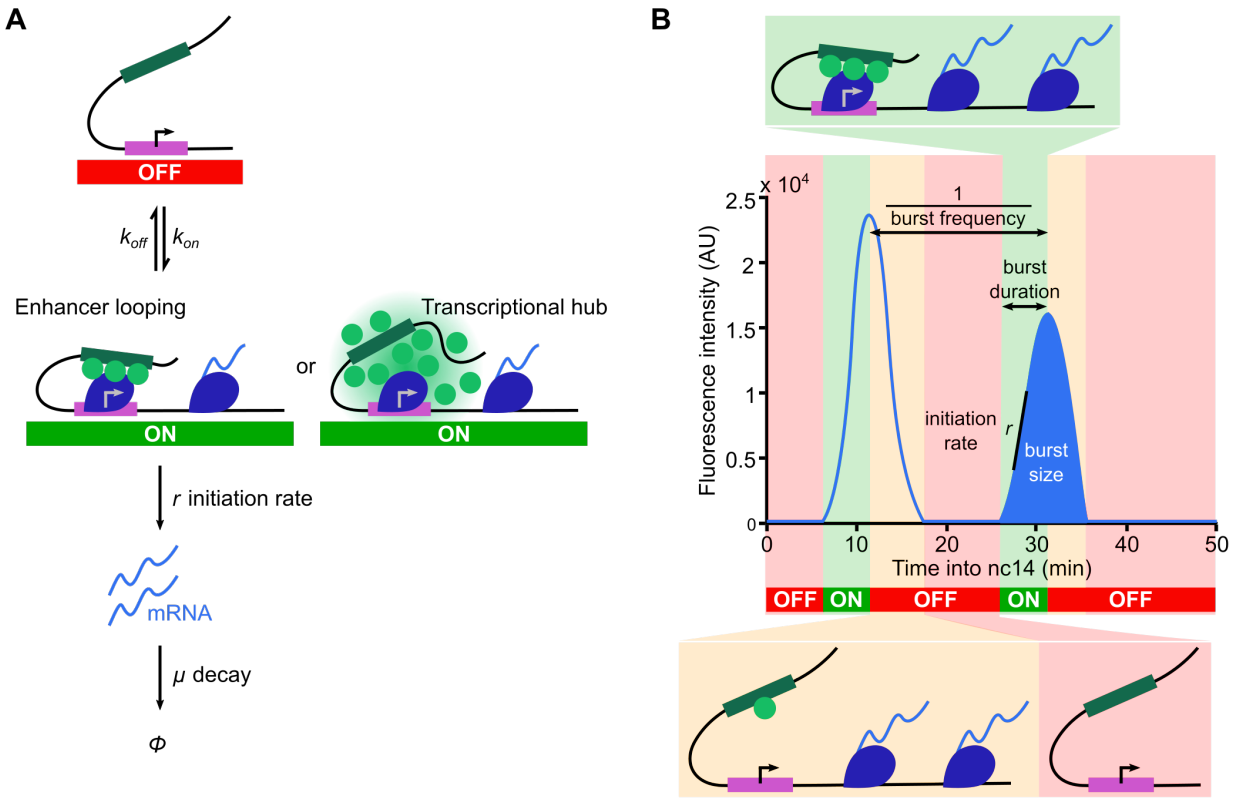


**Figure 1.5:** Two-state model of transcription and how it relates to molecular events of transcription. (A) Here, we represent the two-state model of transcription, in which the promoter is either (1) in the inactive state (OFF), in which RNA polymerase cannot bind and initiate transcription, or (2) in the active state (ON), during which it can. The promoter transitions between these two states with rates $k_{\text{on}}$ and $k_{\text{off}}$, with promoter activation involving both the interaction of the enhancer and promoter and the assembly of all the necessary transcription machinery for transcription initiation to occur. This may occur through enhancer looping or the formation of a transcriptional hub. In its active state, the promoter produces mRNA at rate $r$, and the mRNA decays by diffusing away from the gene locus at rate $\mu$. (B) MS2-tagging RNA allows us to track nascent transcription, and the resulting fluorescence trace (in light blue) is proportional to the number of nascent RNA produced over time. The graph is split into sections, representing different molecular states and how they correspond to fluctuating transcription over time. These states are represented by different colors—red when the promoter is OFF, green when it is ON, and yellow when transcription continues but the promoter is no longer ON, as no new polymerases are being loaded. The dynamics of these fluctuations or bursts can be characterized by quantifying various properties, including burst frequency (how often a burst occurs), burst size (number of RNA produced per burst), and burst duration (the period of active transcription during which mRNA is produced at rate $r$).

Live imaging cells of cells with the MS2 reporter system produces fluorescence traces that are proportional to the amount of nascent RNA produced over time. Properties of these

traces can be quantified and related back to the relevant molecular events that are occurring at the gene locus. However, interpreting these traces requires a model of transcriptional bursting. These bursts of transcription have been explored with multiple models, which are distinguished by the number of states, or levels of activity at which genes can be transcribed.

Ideally, one works with the simplest model that captures the behavior of the system of interest. To unravel the molecular events that shape transcription dynamics, the simplest model that accounts for discontinuous transcription is a two-state model of transcription (**Figure 1.5A**) [123, 159]. Here, a promoter can be in two statesinactive (OFF) or active (ON)and the gene is only transcribed when the promoter is active. The promoter transitions between these two states with rates kon and koff, with the transitions involving both the interaction of the enhancer and promoter and the assembly of the necessary transcriptional machinery. This interaction may be through direct enhancer-promoter looping or the formation of a transcriptional hub, a nuclear region with a high concentration of TFs, co-factors, and RNA polymerase [98]. In its active state, the promoter produces mRNA at rate $r$, and given our ability to observe only nascent transcripts, the mRNA decay rate $\mu$ denotes the diffusion of mRNA away from the gene locus.

Given this model, we can take the properties of the transcription dynamics and map them back to molecular events that are impacting them (**Figure 1.5B**). One property of a transcriptional burst is the burst duration. This is the period of active transcription and is dependent on $k_{off}$, the rate of promoter inactivation, which is related to the dissociation of enhancer-promoter looping. Another property is the burst size, or the number of transcripts produced per burst, which depends on the burst duration and the RNA Pol II initiation rate. Here, polymerase initiation rate is the cumulative output of both polymerase loading and pausing, but as short, aborted transcripts and paused PolII are not visible in MS2 measurements, this is likely representative of the polymerase loading rate. Finally, the burst frequency, or the inverse of the time between two bursts, depends on both the rate of pro-

moter activation and inactivation, $k_{on}$ and $k_{off}$, respectively. Previous work in the early embryo suggests that burst duration (and thus $k_{off}$) are reasonably consistent regardless of enhancer and promoter [166, 177]. Thus, within this regime, burst frequency is mainly dependent on the rate of promoter activation.

### 1.5.1 Many molecular players can modulate transcription dynamics

The most commonly characterized properties of bursting are burst frequency and size, which can be modulated by many mechanisms including the number and strength of TF binding sites [144] , presence and location of repressive sites [18], affinity and competition for general TFs [130], histone modifications [171], and the presence of nucleosome dis-favoring sequences [144]. In fact, the specific promoter motifs themselves can modulate different aspects of transcriptional bursting.

**Promoter motifs influence transcription dynamics** The diversity of shape and motif content of promoters can encode expression output, levels and noise, (by tweaking different burst parameters) by affecting different molecular mechanisms. Some experiments directly interfering with promoter motifs have shown that individual motifs affect different aspects of bursting patterns depending on the context. Specifically, experiments with the major histocompatibility complex I (MHC I) class gene PD1 show that Inr mainly affects burst size, whereas a TATA-like motif affects both burst frequency and size. However, transcriptional activation by $\gamma$-interferon changes their contributions to both burst parameters [56]. This suggests that the promoter plays a key role in allowing the gene to respond dynamically to tissue-specific and environmental conditions, which is particularly necessary for this immune surveillance gene.

How can motifs affect different bursting parameters? For example, TATA Box has been associated with high levels of expression noise [10, 90, 153]. By integrating datasets at different levels of resolution, scale, and type, researchers found that TBP can exist in four microstates(1) monomeric TBP, (2) TFIID:TBP complex, (3) SAGA:TBP complex, (4) Mot1p:TBP complex. Depending on the bendability of the DNA sequence, TBP may be more or less likely to bind as a monomer (microstate 1) or as part of TFIID (microstate 2). Binding of TBP as part of TFIID (microstate 2) leads to activation of the promoter and transcription initiation (ON state). If monomeric TBP binds (microstate 1), SAGA may assemble (microstate 3) and initiate transcription (ON state), or Mot1p (or similar remodeling factors, e.g. NC2) may remove TBP (microstate 4) leading to an OFF state. The length of time the promoter spends in each of these states affects the expression output. Notably, weaker TATA motifs have weaker affinities for monomeric TBP, increasing the likelihood that TBP will bind as part of the TFIID complex and precluding the access of other remodeling factors to TBP. This means that promoters containing weak TATA motifs are more stably occupied and show less noisy expression than those with strong TATA motifs. Thus, it is the combination of motif strength and the abundance of all potential factors and cofactors involved in transcription that determines possible complexes that assemble at a promoter and their residence times.

Promoter motifs also play a role in modulating other aspects of transcription dynamics. However, the role of each motif can vary from one locus to the next. In the TATA-only *Drosophila snail* promoter, the TATA Box affects burst size by tuning burst duration [127]. In the mouse PD1 proximal promoter, which consists of a CAAT Box, TATA Box, Sp1, and Inr motif, the TATA box may tune burst size and frequency [56]. A study of a synthetic Drosophila core promoter and the *ftz* promoter found that the TATA box tunes burst size by modulating burst amplitude and that Inr, MTE, and DPE tune burst frequency [177]. TATA Box also appears to be associated with increased expression noise, as TATA-containing promoters tend to drive larger, but less frequent transcriptional bursts [129]. In contrast to

TATA Box, Inr appears to be associated with promoter pausing, e.g. by adding a paused promoter state in the Inr-containing *Drosophila Kr* and *Ilp4* promoters [127]. In fact, a Pol II ChIP-seq study indicates that paused developmental genes appear to be enriched for GAGA, Inr, DPE, and PB motifs [129].

# Chapter 2

# An Enhancer's Length and Composition Are Shaped by Its Regulatory Task

The contents of this chapter appear in the journal *Frontiers in Genetics* [94].

## 2.1 Abstract

Enhancers drive the gene expression patterns required for virtually every process in metazoans. We propose that enhancer length and transcription factor (TF) binding site compositionthe number and identity of TF binding sites—reflect the complexity of the enhancer's regulatory task. In development, we define regulatory task complexity as the number of fates specified in a set of cells at once. We hypothesize that enhancers with more complex regulatory tasks will be longer, with more, but less specific, TF binding sites. Larger numbers of binding sites can be arranged in more ways, allowing enhancers to drive many

distinct expression patterns, and therefore cell fates, using a finite number of TF inputs. We compare ~100 enhancers patterning the more complex anterior-posterior (AP) axis and the simpler dorsal-ventral (DV) axis in *Drosophila* and find that the AP enhancers are longer with more, but less specific binding sites than the (DV) enhancers. Using a set of ~3,500 enhancers, we find enhancer length and TF binding site number again increase with increasing regulatory task complexity. Therefore, to be broadly applicable, computational tools to study enhancers must account for differences in regulatory task.

## 2.2 Introduction

Nearly every aspect of an organism, from its development to its immune response, is dependent on precise spatiotemporal control of gene expression. This control is mediated by the binding of transcription factor (TF) activators and repressors to stretches of regulatory DNA called enhancers.

Given their role in diverse biological processes, it is not surprising that enhancers vary widely in architecture—length, number of TF binding sites, and the average binding specificity of the TFs that bind them. Enhancers can be ~10 – 1,000 bps long, with a couple to tens of TF binding sites [9, 175]. Several theories have been put forth to explain why enhancers are built so differently. For example, differences in evolutionary pressures and TF cooperativity are invoked to explain why many developmental enhancers are robust to rearrangements of TF binding sites within them while some immune-responsive enhancers are intolerant to even point mutations [150, 74, 109, 2].

Although enhancers vary in architecture, some constraints apply to all enhancers, e.g. an enhancer's need to be distinguishable from the rest of the genome. Because eukaryotic TFs are highly degenerate, TF binding sites litter the genome, and an enhancer can only

achieve distinguishability if it consists of a cluster of TF binding sites within a short distance [165, 8, 43, 52, 106, 132, 95, 173, 55]. Enhancer length, number of TF binding sites, and average specificity of TFs binding an enhancer can be combined in different ways to achieve distinguishability. For example, an enhancer with higher average TF specificity requires fewer TF binding sites than one with lower average TF specificity to be distinguishable from the genomic background.

We propose that the complexity of an enhancer's "regulatory task" — the process that it controls — is one force that shapes enhancer architecture. In development, task complexity can be defined as the number of cell fates being specified in a set of roughly homogeneous cells at one time. When a cell can be driven to one of many cell fates, the task complexity is high; when a cell is making binary decisions between cell fates, the task complexity is low (**Figure 2.1A**). Since cell fate is largely specified by gene expression patterns, the more cell fates being specified, the more distinct expression patterns are needed. To accommodate this need using a limited set of TFs, these enhancers need to contain a larger number of TF binding sites, which allow for more rearrangements and, presumably, more expression patterns. Thus, we propose that enhancers with more binding sites can accommodate higher task complexity. Though intuitive, this proposal has never been verified systematically.

To evaluate this hypothesis, we characterize two sets of enhancers in *Drosophila melanogaster* and analyze the correlation between regulatory task complexity and enhancer architecture. In a set of ~100 early embryonic enhancers, those that pattern the more complex anterior-posterior (AP) axis are longer, have more binding sites, and have lower average TF specificity compared to those patterning the simpler dorsal-ventral (DV) axis. In a set of ~3,500 enhancers active throughout embryogenesis, we find enhancers active early are longer and have more binding sites than those active late, reflecting the general trend that task complexity decreases with developmental time. We conclude that the complexity of an enhancer's regulatory task is one of many forces shaping its architecture.

## 2.3 Results

To understand the properties required for an enhancer to be distinguishable from the genomic background, we calculate the probability of finding an enhancer with a particular length, number of TF binding sites, and average TF binding specificity [173] (Supplementary Material within). As a proxy for TF binding specificity, we use $p$, the probability of finding a "hit" or match to the TF binding motif in the genomic background (see Materials and Methods). Note that a larger $p$ corresponds to a lower binding specificity. The probability of finding an enhancer of length $w$, with $k$ TF binding sites is:

$$P(k) = \binom{w}{k} p^k (1-p)^{w-k}$$

To achieve distinguishability, $P(k)$ must be less than $1/N$, where $N$ is the genome size accessible for TF binding. Thus, the number of required binding sites increases with enhancer length and motif hit probability (Figure 2.1B). Considering the median, first and third quartiles of all *Drosophila* TF binding specificities, the corresponding number of TF binding sites required in a 1 kb enhancer decreases from 16 to 7 to 5 as TF binding specificity increases (or motif hit probability decreases).

To take into account the compaction of the genome, we consider different values of $N$. We use DNase I hypersensitivity profiles to estimate the accessible regions [152]. Whether we use a conservative estimate of accessible regions during development (4.1 Mb), a more relaxed estimate (19.4 Mb), or the entire genome (175.5 Mb) [152, 36], the same trends are seen (**Figure 2.1C**). For a 1 kb enhancer with binding sites for a TF with relatively low binding specificity, $p = 2 \times 10^{-3}$, the number of required binding sites increases from 13 to 15 as $N$ increases from 4.1 to 175.5 Mb, and thus the number of required binding sites is only weakly dependent on accessible genome size.

To test whether task complexity shapes the characteristics of enhancer architecture, we need

a set of enhancers that drive regulatory tasks of different complexities and knowledge of the transcription factors (TFs) that regulate them. The *Drosophila* embryonic AP and DV patterning systems neatly fit these criteria. The AP axis is more complex than the DV axis, with the AP axis consisting of 14 parasegments [111] and the DV axis consisting of six germ layers and sublayers [91], and therefore the patterning of the AP axis requires enhancers that drive more unique gene expression patterns. Years of work from many groups have identified ~40 principal TFs [41, 105] whose binding to ~100 characterized enhancers [119] drives AP and DV patterning.

To identify the TF binding sites within these enhancers, we use a computational approach. Though ChIP can experimentally identify TF-bound regions, existing data sets in the *Drosophila* embryo are low resolution, with ~100 base pair peaks [95, 105, 134], which are longer than the ~10 bp TF binding sites [182]. Therefore, we predict TF binding sites using experimentally measured binding motifs [147]. To select a threshold above which a sequence is deemed a true binding site, we develop a principled approach, scoring the aligned sequences used to create the motifs and setting a threshold such that 75% of these aligned sequences are predicted as true (see Materials and Methods).

We analyze 60 AP and 39 DV enhancers, identifying binding sites for 24 AP and 10 DV TFs. Consistent with our predictions, AP enhancers (median length = 1.3 kb) are longer than DV enhancers (median = 0.8 kb; **Figure 2.2A**; $p = 4.5 \times 10^{-4}$; Mann Whitney rank-sum test). AP enhancers also have a larger number of TF binding sites (median = 47) than DV enhancers (median = 9; **Figure 2.2B**; $p = 1.5 \times 10^{-13}$; Mann Whitney rank-sum test). To ensure that the difference is not due to the larger number of AP TFs, we also calculated the number of TF binding sites per enhancer, normalized by the number of TF motifs used to search the enhancer, and find the difference holds (AP median = 2.0, DV median = 0.9; **Figure 2.2C**; $p = 6.1 \times 10^{-6}$; Mann Whitney rank-sum test). AP enhancers also have a higher average motif hit probability (AP median = $4.9 \times 10^{-3}$, DV median =

$3.0 \times 10^{-3}$; **Figure 2.2E**; p = $7.5 \times 10^{-9}$; Mann Whitney rank-sum test), which is a result of differential binding rather than the TFs considered, as the specificity of AP and DV TFs have a similar distribution (**Figure 2.2D**, $p = 0.247$; Mann Whitney rank-sum test). This difference is likely driven by the fact that the key TFs that act as morphogens for these axes show markedly different binding specificities, with the key AP axis TFs having low binding specificities and the key DV axis TF having high binding specificity (**Figure S2.1**). In summary, we find that the enhancers that encode the lower complexity task of specifying the DV axis are composed of fewer binding sites, as predicted by our hypothesis. The DV enhancers require fewer binding sites because they are both shorter and use more specific TFs than the AP enhancers. Our hypothesis does not require that both enhancer length and motif specificity both differ, though in this case they do.

To determine whether these tradeoffs in enhancer architecture apply to a larger, if less well-characterized dataset, we analyze the Vienna Tile enhancers [81], which drive expression throughout *Drosophila* embryogenesis. To produce this dataset, the Stark lab measured the expression patterns driven by 7,705 enhancer candidates and found 4,480 enhancers that were active during development. Of these enhancers, we consider the 3,580 enhancer candidates that were successfully refined to the putative minimal enhancers using functional genomics [81]. To determine the relevant TFs, we match the stages of the active enhancers with the concurrently expressed TFs [154, 155, 53].

We assume that as development progresses, the task complexity decreases, approaching binary decisions between two cell fates. We found that enhancer length monotonically decreases over development (**Figure 2.3A; Table 2.1**). While stages 46 and 78, and stages 910 and 1112, have very similar length distributions for active enhancers ($p = 1$, $p = 0.1$, respectively; Mann Whitney rank-sum test), all other intervals have significantly different distributions of enhancer length (**Figure 2.3D**, $p < 0.05$; Mann-Whitney rank sum test with Bonferroni correction applied). Number of TF binding sites and average motif hit probabil-

ity, in contrast, do not show a clear trend (**Figures 3B and E**). However, there is a large increase in the number of TFs expressed in the final two time intervals (see Table 2.1), and when the number of binding sites is normalized by the number of binding motifs used to search the enhancer, the binding site trend mirrors the enhancer length trend (**Figures 3C and F**). We also verified that these trends were not unduly influenced by enhancers driving ubiquitous expression patterns (**Figure S2.2, Table 2.2**). Thus, decreasing complexity again is associated with decreasing enhancer length and with decreasing TF binding site number, when normalized appropriately. In this case, there is no clear trend with regards to motif specificity, which, as we note above, is not necessarily at odds with our hypothesis.

## 2.4 Discussion

We hypothesize that an enhancer's regulatory task complexity shapes its architecture. In the case of *Drosophila* axis patterning, the AP axis has higher task complexity than the DV axis, and accordingly, enhancer length, number of TF binding sites, and average motif hit probability increase with task complexity. In the case of *Drosophila* embryogenesis, where we posit that task complexity decreases over time, enhancer length and binding site number decrease accordingly.

Though the well-characterized *Drosophila* axis patterning systems are ideal for studying how an enhancer's regulatory tasks shape its design [106, 121, 96, 120, 51], the systems still have limitations. For example, autoregulatory enhancers like ftz_up, ftz_zebra, and gt_minus1 [59, 58, 61] have lower task complexity because they reinforce the expression patterns determined by other enhancers, and therefore, may not be consistent with the observed trends. However, we find these autoregulatory enhancers have parameters that generally fall within the bulk of the distribution. In addition, the enhancer boundaries in this dataset were determined one-at-a-time. However, in the Vienna Tile enhancer set, in

which boundaries are determined in a uniform manner, we still find that enhancer length decreases with developmental time and regulatory task complexity.

In contrast to the axis patterning data set, choosing principal TFs for the Vienna Tile enhancers is challenging because there is no consistent annotation of the expression patterns of TFs and enhancers. We match stage-specific expression of enhancers and TFs without considering tissue-specific expression, which impacts both the number of TF binding sites and the average motif hit probability and undoubtedly obscures the clarity of those trends.

We expect that there are many other forces shaping enhancer architecture, like protein-protein interactions between TFs or between TFs and cofactors, and therefore do not expect that regulatory task complexity alone can explain enhancer architecture. Additionally, a particular TF may be employed in an enhancer because it is expressed in the right place at the right time, and not because of its TF binding specificity, though an analysis of the binding specificities of the TFs encoded in the *Drosophila* genome shows that there is a wide distribution of TF specificities that is relatively independent of developmental stage (**Figures S3 – S6**). However, we can make educated guesses about the ways that enhancer architecture may vary depending on regulatory task and use this information to improve our ability to predict and design putative enhancers. As increasingly large sets of enhancers are identified in a variety of biological settings, we will undoubtedly uncover other forces impacting why an enhancer is built in a particular way.

## 2.5 Materials and Methods

### 2.5.1 Datasets Used in this Study

The 60 AP and 39 DV patterning enhancers were collected by [119] and provided here as Supplementary Data Sheet 1, and in this dataset, we considered the binding of 33 early patterning TFs [41, 105]. The Vienna Tile enhancer project tested the activity of 7,705 potential enhancers [81]. In this analysis, we considered the 3,580 enhancer candidates that were active during embryogenesis and whose boundaries had been refined using DHS regions or CBP/P300-bound and H3K4me1-marked regions. Here, we analyzed the binding of TFs that were concurrently expressed with active enhancers based on the Berkeley *Drosophila* Genome Project *in situ* annotations [154, 155, 53]. These TF lists are available in Supplementary Data Sheets 2, 3. 51.7% of the AP enhancers and 30.8% of the DV enhancers at least partially overlapped with the Vienna Tile enhancers. The five completely overlapping enhancers showed expression in the same stage (i.e., stages 4  6), except for the DV enhancers pnr and rho (Table 2.3). Outside of those two, the greatest amount of overlap that did not result in expression at the same stage was 60.5%.

### 2.5.2 Transcription Factor Binding Site Prediction

Transcription factor (TF) binding sites were computationally predicted using Patser [57] with the position weight matrices (PWMs) from FlyFactor Survey [182]. Pseudocounts were added to each element in the PWM in proportion to the intergenic frequency of the corresponding base to a total of 0.01. For those TFs with more than one PWM, the PWM derived from the largest set of aligned sequences was used, except in the case of giant and daughterless.

As no TF binding sites were identified in the set of AP enhancers when using the giant PWM selected using the previous criteria, a switch was made to another available giant PWM in FlyFactor Survey with which binding sites could be predicted. In the case of daughterless, which often binds as a heterodimer and has been identified as one of the key TFs in DV patterning, the PWM that was created using only daughterless and not any heterodimeric partners was used.

For the early patterning dataset, 33 of the 37 TFs [41, 105] determined to be principal regulatory factors for AP and DV patterning had available PWMs and were used. PWMs were not available for croc, Stat92E, tsh, or Dad. For the Vienna Tile dataset, a total of 292 TFs that had available PWMs and were expressed during embryogenesis were used.

To determine ln ($p$-value) cutoffs in a systematic manner, the aligned sequences from which the PWMs are derived are scored by Patser, and a 75th percentile ln ($p$-value) was chosen as a cutoff such that 75% of the aligned sequences are considered true binding sites. Cutoffs at multiple percentiles were considered, but the overall trends for relative numbers of putative TF binding sites identified remained constant regardless of the chosen cutoff (**Figure S2.7**).

Some PWMs were generated from DNase I footprints curated in the FlyReg database [51]; aligned versions of these footprints were not directly provided by FlyFactor Survey. The raw sequences were retrieved from FlyReg v2.0 and were aligned when possible. Note that Patser can only score sequences that are the same length or longer than the PWM, so some sequences used to create the PWM have been omitted when determining the percentile cutoffs.

### 2.5.3 Evaluating Transcription Factor Specificity

Information content is a measure of TF specificity. To measure information in a motif, we calculated the Kullback-Leibler distance [137, 148] between the motif and the composition of the intergenic regions of the genome

$$I = \sum_{i=1}^{L} \sum_{b \in A,C,G,T} p_i(b) \log_2 \frac{p_i(b)}{q(b)}$$

where $L$ is the length of the motif, $p_i(b)$ is the frequency of base $b$ at position $i$ in the motif, and $q(b)$ is the frequency of base $b$ in the intergenic regions of the genome. Note that $p = 2 - I$ is roughly the probability of a motif hit in the genome for a TF of information content $I$ [7]. For a set of TF binding sites in an enhancer, an average motif hit probability

$$p_{av} = \frac{\sum_j n_j 2^{-I_j}}{n_{total}}$$

is calculated, where $n_j$ = number of binding sites for TF $j$, $I_j$ = information content of TF $j$, and $n_{total}$ = the total number of TF binding sites in a particular enhancer. If a cluster is composed of sites of $m$ different TFs with identical motif hit probability $p$, the probability of finding a cluster of $k$ binding sites within $w$ bps is

$$P(k) = \binom{w}{k} (mp)^k (1 - mp)^{w-k}$$

Therefore, to characterize the average specificity of TFs employed in a specific enhancer, we choose to compute the average motif hit probability $p$, as opposed to the average information content $I$.

### 2.5.4 Quantitation and Statistical Analysis

Mann-Whitney rank tests were performed to compare all distributions, and the $p$-values were reported. The Mann-Whitney test was chosen because it does not require the assumption that the distributions to be compared are normally distributed. When multiple comparisons were made, the Bonferroni correction was applied.

### 2.5.5 Data and Software Availability

Python code for enhancer architecture analysis is available at the Wunderlich Lab GitHub (`https://github.com/WunderlichLab/Info_Content`).

The axis patterning enhancers originally collected by Papatsenko et al. are available as supplemental data file Supplementary Data 1 [119], which can be found at `https://www.frontiersin.org/article/10.3389/fgene.2017.00063/full#supplementary-material`.

## 2.6 Author Contributions

LL and ZW conceived of the study. LL carried out the analysis, and LL and ZW wrote the paper.

## 2.7 Funding

This work is supported by NIH grants R00HD073191 (to ZW) and T32EB009418 (to LL).

## 2.8 Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## 2.9 Acknowledgments

We thank Ali Mortazavi and Rahul Warrior for useful discussions.

## 2.10 Figures

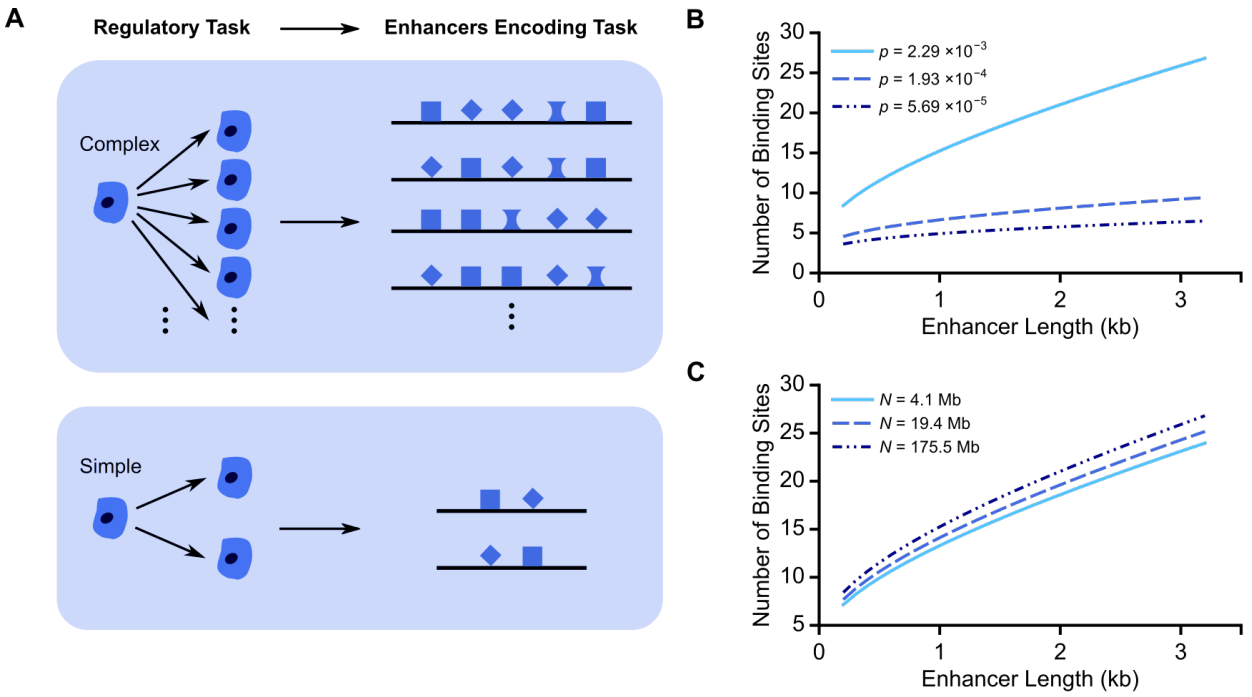


**Figure 2.1:** Regulatory task complexity can shape enhancer length and binding site composition. **(A)** We propose that more complex regulatory tasks, e.g., cell patterning decisions, are associated with longer enhancers with more binding sites. More binding sites can be arranged within an enhancer in more ways, allowing for the specification of a wider variety of expression patterns and, therefore, more complex tasks. **(B)** We plot the minimum number of TF binding sites required for enhancers of varying lengths to achieve distinguishability from the genomic background. We show the results for three motif hit probabilities, corresponding to the median, first and third quartiles of *Drosophila* TF binding specificities. As motif hit probability p decreases from ~2 in 1 kb ($2 \times 10^{-3}$) to ~6 in 100 kb ($6 \times 10^{-5}$), an enhancer of the same length requires fewer binding sites to be distinguishable from the background. **(C)** To test the effect of genome accessibility, we plot the minimum number of TF binding sites required for enhancers of varying length in the context of different accessible genome sizes ($N$). Varying the accessible regions of the genome has a minor impact on the trend of numbers of TF binding sites increasing with enhancer length.

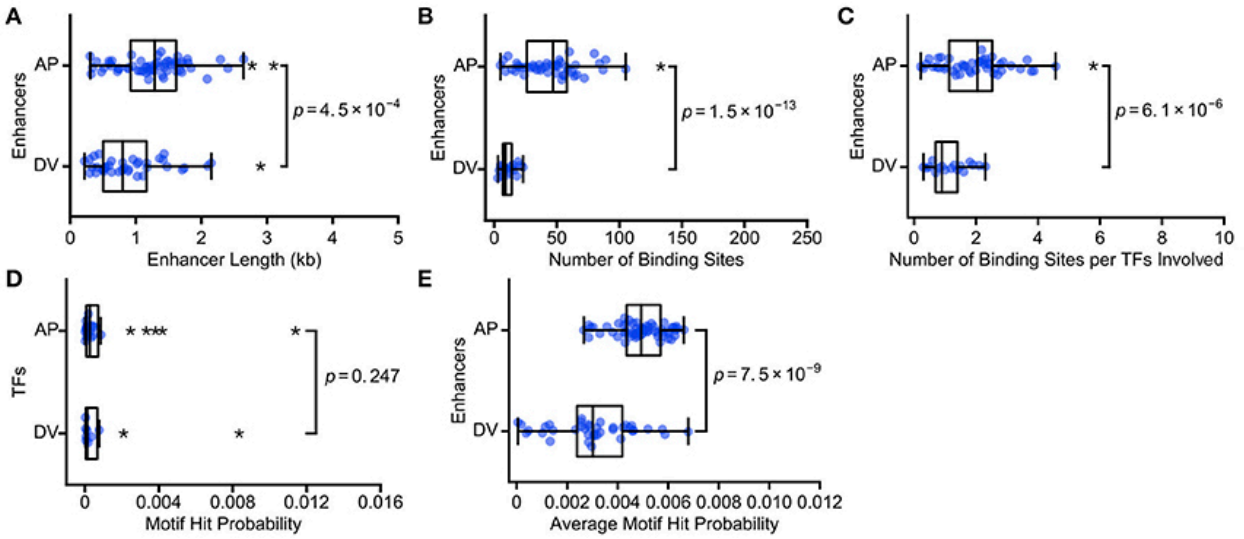


**Figure 2.2:** The more complex AP axis is patterned by enhancers with more TF binding sites. We show the scatterplots and associated boxplots of **(A)** the length of AP and DV enhancers, **(B)** the number of TF binding sites predicted in AP and DV enhancers, **(C)** the number of TF binding sites normalized by the number of TFs involved, **(D)** the motif hit probability of TFs involved in AP and DV patterning, and **(E)** the average motif hit probability of AP and DV enhancers. These data are consistent with our hypothesis that enhancers carrying out more complex regulatory tasks will have more binding sites, in this case because AP enhancers are both longer and have lower average TF binding specificity. In all box plots, the boxes indicate the lower and upper quartiles, with the line within the box indicating the median. Whiskers extend to 1.5*IQR (interquartile range) plus or minus the upper and lower quartile, respectively, and the stars indicate outliers that fall outside the whiskers. $p-$values from Mann-Whitney rank tests are shown.

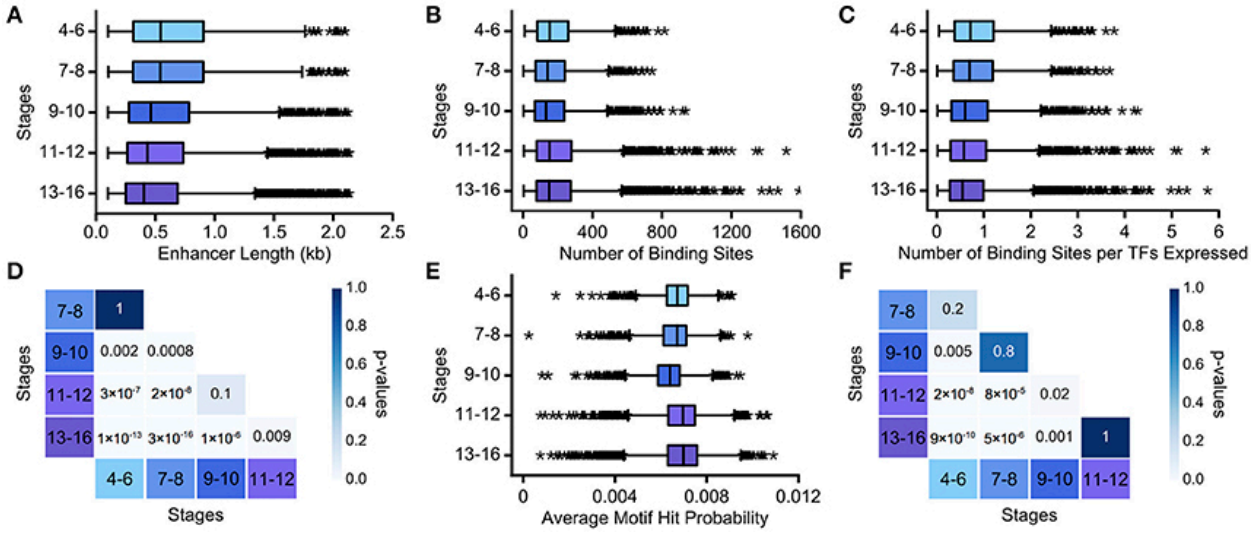


**Figure 2.3:** Decreasing regulatory task complexity over embryogenesis is associated with decreasing enhancer length. We show boxplots of **(A)** the length of minimal Vienna Tile enhancers, **(B)** the number of TF binding sites predicted in minimal Vienna Tile enhancers, **(C)** the number of TF binding sites predicted in minimal Vienna Tile enhancers normalized by TFs concurrently expressed, and **(E)** the average motif hit probability of minimal Vienna Tile enhancers over developmental stages 4 – 16. The heatmaps display the Bonferroni-adjusted $p$-values from the Mann-Whitney rank test between **(D)** pairwise distributions of Vienna Tile enhancer length and between **(F)** pairwise distributions of the number of TF binding sites predicted in Vienna Tile enhancers per TFs concurrently expressed. In all box plots, the boxes indicate the lower and upper quartiles, with the line within the box indicating the median. Whiskers extend to 1.5*IQR plus or minus the upper and lower quartile, respectively, and the stars indicate outliers that fall outside the whiskers.

## 2.11 Supplementary Material

### 2.11.1 Supplementary Data

The Supplementary Data for this article can be found at `https://www.frontiersin.org/article/10.3389/fgene.2017.00063/full#supplementary-material`.

**Supplementary Data 1 Axis patterning enhancers curated by Papatsenko, et al.** The axis patterning enhancers curated by Papatsenko, et al. are no longer available on the site where they were originally hosted [119]; to facilitate the use of this dataset, we have included the FastA file here.

**Supplementary Data 2 Transcription factors involved in axis specification.** The first row is a header that lists the axes, anterior-posterior (AP) and dorsal-ventral (DV). Each column lists the TFs that are involved in specification of that axis.

**Supplementary Data 3 Transcription factors expressed by stage during embryogenesis.** The first row is a header that lists the stages. Each column lists the TFs that are expressed during that stage based on BDGP *in situ* data.

### 2.11.2 Supplementary Figures

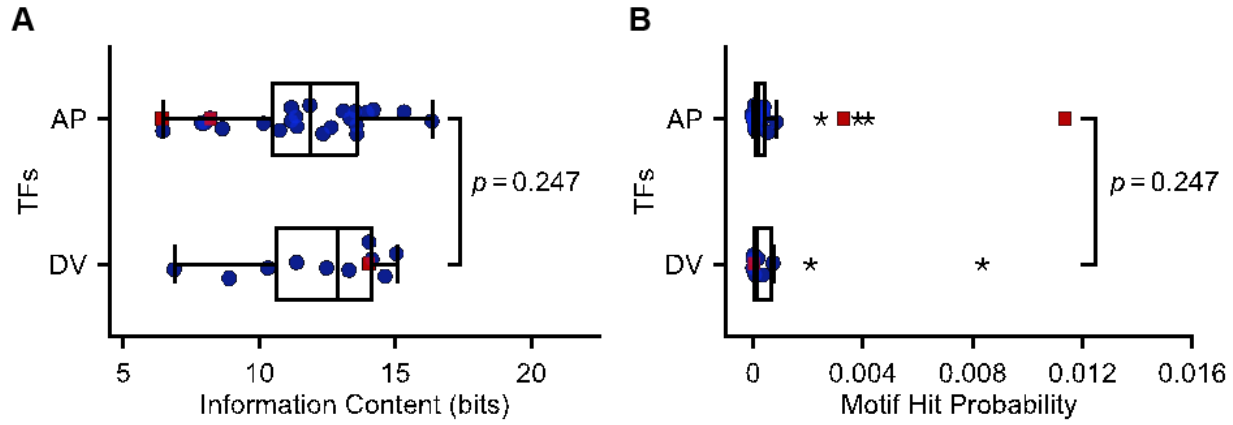


**Figure S2.1:** The key transcription factors whose gradients set up the anterior-posterior and dorsal-ventral axes fall towards the tails of the information content distribution. We show boxplots of **(A)** the information content and **(B)** the motif hit probability of TFs important in axis patterning. The red squares indicate the information content and motif hit probability of the key TFs that act as morphogens to set up the AP and DV axes. These key AP TFs—caudal (6.46 bits; $3.32 \times 10^{-3}$) and bicoid (8.23 bits; 0.0114)—have low information content and high motif hit probability, respectively; whereas, the key DV TF—dorsal (14.05 bits; $5.91 \times 10^{-5}$) has relatively high information content and low motif hit probability, respectively. In all box plots, the boxes indicate the lower and upper quartiles, with the line within the box indicating the median. Whiskers extend to 1.5*IQR plus or minus the upper and lower quartile, respectively, and the stars indicate outliers that fall outside the whiskers. $p$-values from Mann-Whitney rank tests are shown.

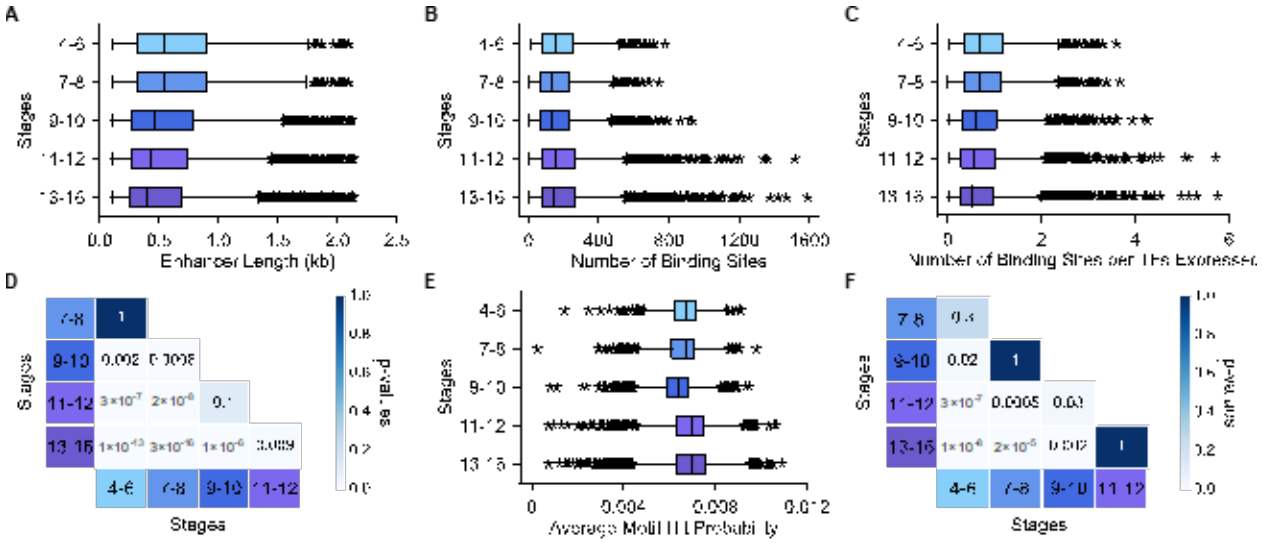


**Figure S2.2:** When Vienna Tile enhancers that drive ubiquitous expression are removed, the same trends in enhancer architecture are seen. We show boxplots of **(A)**the length of minimal Vienna Tile enhancers, **(B)** the number of TF binding sites predicted in minimal Vienna Tile enhancers, **(C)** the number of TF binding sites predicted in minimal Vienna Tile enhancers per TFs concurrently expressed, and **(E)** the average motif hit probability of minimal Vienna Tile enhancers over developmental stages 4-13. The heatmaps display the Bonferroni-adjusted $p$-values from the Mann-Whitney rank test between **(D)** pairwise distributions of Vienna Tile enhancer length and between **(F)** pairwise distributions of the number of TF binding sites predicted in Vienna Tile enhancers per TFs concurrently expressed. In all box plots, the boxes indicate the lower and upper quartiles, with the line within the box indicating the median. Whiskers extend to 1.5*IQR plus or minus the upper and lower quartile, respectively, and the stars indicate outliers that fall outside the whiskers.

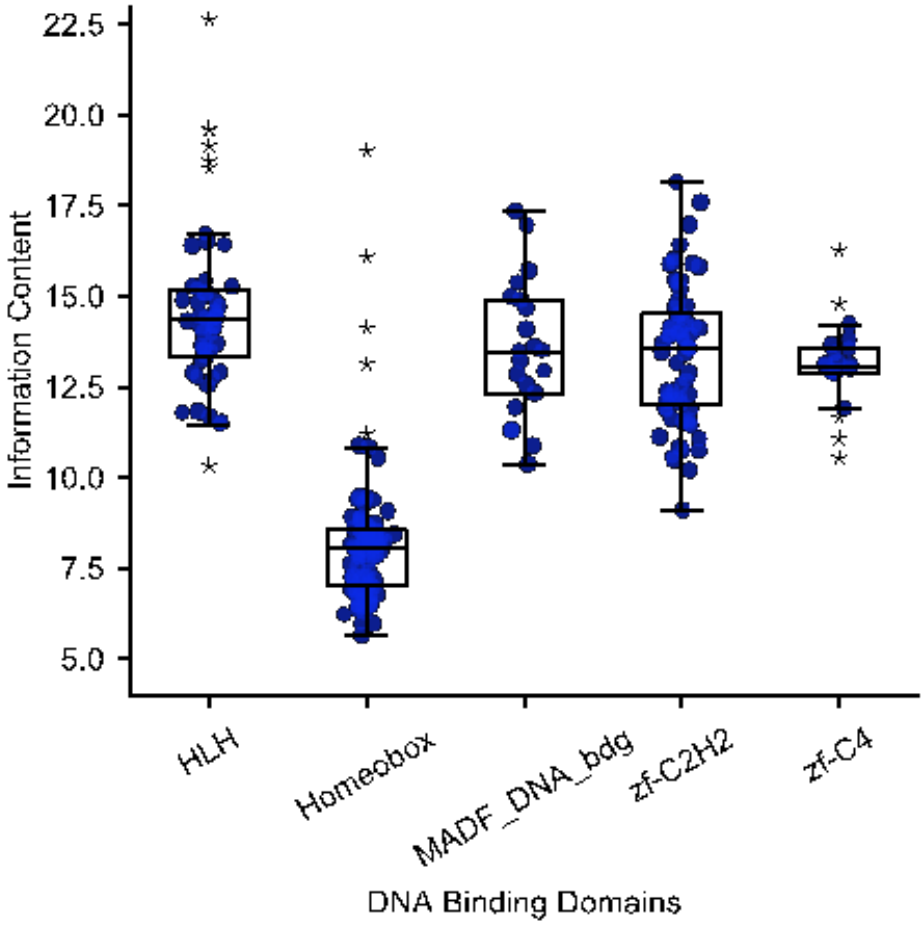


**Figure S2.3:** The information content of homeobox proteins is generally lower than that of transcription factors with other common DNA-binding domains. We show boxplots of the information content of the TFs by DNA-binding domain, with boxes indicating the lower and upper quartiles and the line within the box indicating the median. Whiskers extend to 1.5*IQR plus or minus the upper and lower quartile, respectively. The stars indicate outliers that fall outside the whiskers. Only those DNA-binding domains that are represented by more than 10 TFs in our sample have been included.

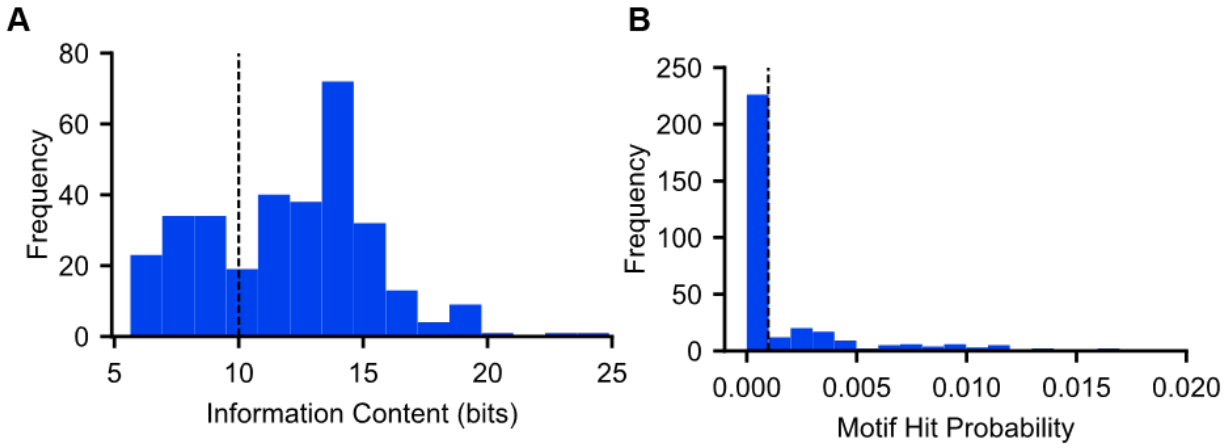


**Figure S2.4:** The information content of transcription factors fall into a bimodal distribution. **(A)** We show a histogram of the information content of the *Drosophila* transcription factors with a dashed line indicating a natural separation point between the two peaks. **(B)** We show a histogram of the motif hit probability of the *Drosophila* TFs with a dashed lined indicating the same point at which the two peaks of information content have been separated. Note that the smaller peak located to the left of the dashed line in **(A)** corresponds to the smaller peak located to right of the line in this motif hit probability distribution.

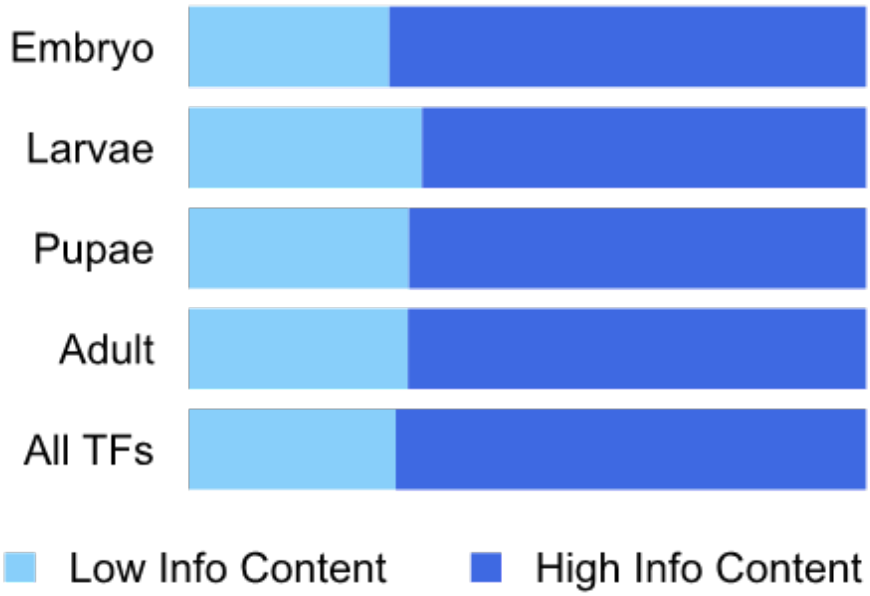


**Figure S2.5:** The distribution of low and high information content TFs being expressed remains relatively consistent over time. We show stacked bar graphs of the information content of TFs expressed while the flies are embryos, larvae, pupae, and adults, with a bar graph of all TFs for comparison. Each bar has been normalized for the number of TFs expressed at that life stage. Using 10 bits as the natural separation point between the peaks in the bimodal distribution of TF information contents from Supplementary Figure 4, we have divided our TFs into "Low Info Content" and "High Info Content" categories with information contents $> 10$ bits considered "High." Comparing the fraction of low versus high information content of TFs expressed at any stage compared to that of all TFs shows very little difference, suggesting that any differences we see in the distribution of average information content in enhancers is due to usage of TFs of certain information contents over others.

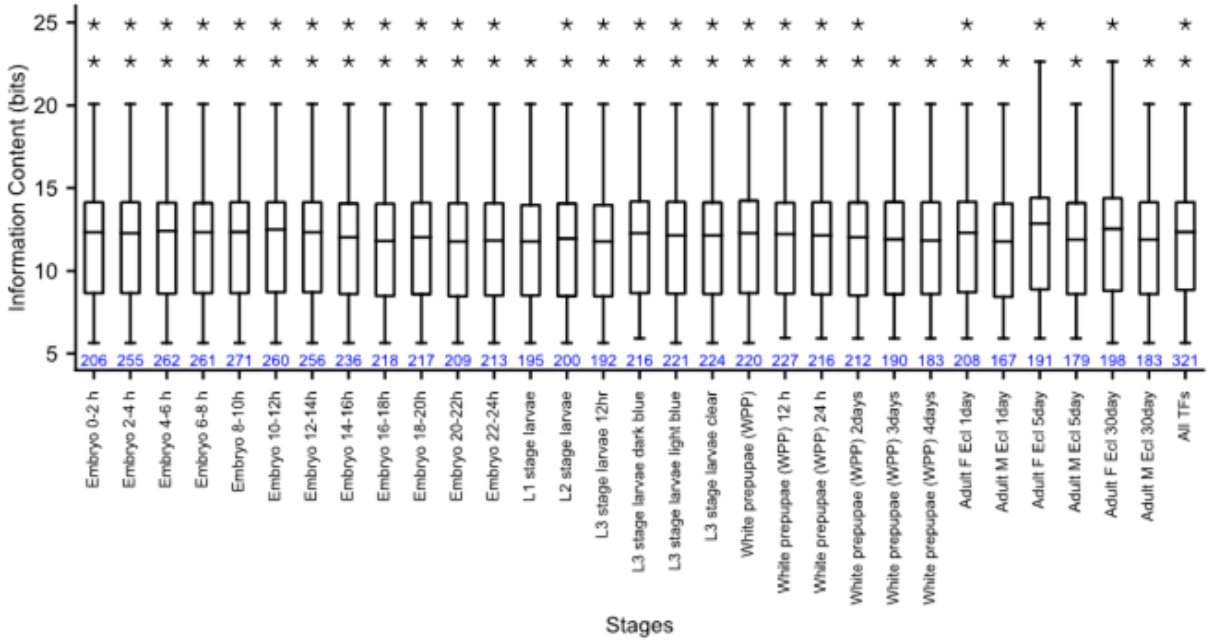


**Figure S2.6:** The distribution of TF information content remains consistent over time. We show boxplots of the information content of transcription factors being expressed over the lifespan of *Drosophila*, with the last boxplot of the information content of all TFs given for comparison. The boxes indicate the lower and upper quartiles and the line within the box indicating the median. Whiskers extend to 1.5*IQR plus or minus the upper and lower quartile, respectively. The stars indicate outliers that fall outside the whiskers. The numbers in blue above the x-axis indicate the number of TFs expressed at that stage and included in that boxplot.

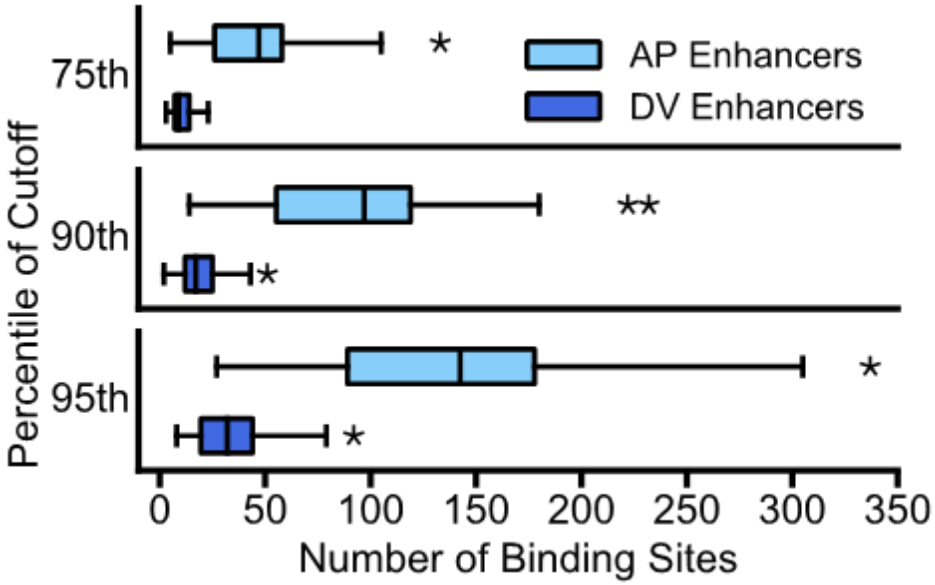


**Figure S2.7:** Trends in number of predicted binding sites in axis patterning enhancers remain unchanged when cutoffs for "true" binding sites are varied. We show boxplots of the number of TF binding sites predicted in AP and DV enhancers using percentile cutoffs of 75, 90, and 95, with boxes indicating the lower and upper quartiles and the line within the box indicating the median. Whiskers extend to 1.5*IQR plus or minus the upper and lower quartile, respectively. The stars indicate outliers that fall outside the whiskers. P-values calculated by performing Mann-Whitney rank tests on each pair of distributions are less than or equal to $1.5 \times 10^{-13}$.

### 2.11.3 Supplementary Tables

**Table 2.1:** Related to Figure 2; Summary data of minimal Vienna Tile enhancers. The second and third column list the number of active enhancers and TFs expressed during the stages between 4 and 16 listed in column 1. Columns 4-8 list the median values for enhancer length, number of TF binding sites, number of TF binding sites per TF expressed, and average motif hit probability of minimal Vienna Tile enhancers for stages between 4 and 16.

| **Stage** | **# Active Enhancers** | **# TFs Expressed** | **Medians** | | | |
|---|---|---|---|---|---|---|
| | | | **Enhancer Length** (bp) | **# TFBSs** | **# TFBSs per TF Expressed** | **Average Motif Hit Probability** |
| 4-6 | 713 | 216 | 545 | 154 | 0.713 | 0.00673 |
| 7-8 | 878 | 202 | 543 | 141 | 0.696 | 0.00671 |
| 9-10 | 1530 | 219 | 462 | 132 | 0.603 | 0.00641 |
| 11-12 | 2307 | 265 | 432 | 153 | 0.577 | 0.00697 |
| 13-16 | 3334 | 276 | 404 | 151 | 0.547 | 0.00699 |

**Table 2.2:** Related to Supplementary Figure 2; Summary data of minimal Vienna Tile enhancers when those driving ubiquitous expression are removed. The second and third column list the number of active enhancers and TFs expressed during the stages between 4 and 16 listed in column 1. Columns 4-8 list the median values for enhancer length, number of TF binding sites, number of TF binding sites per TF expressed, and average motif hit probability of minimal Vienna Tile enhancers for stages between 4 and 16.

| **Stage** | **# Active Enhancers** | **# TFs Expressed** | **Medians** | | | |
|---|---|---|---|---|---|---|
| | | | **Enhancer Length** (bp) | **# TFBSs** | **# TFBSs per TF Expressed** | **Average Motif Hit Probability** |
| 4-6 | 670 | 216 | 545 | 150 | 0.694 | 0.00671 |
| 7-8 | 8830 | 202 | 543 | 136 | 0.671 | 0.00670 |
| 9-10 | 1467 | 219 | 462 | 128 | 0.584 | 0.00639 |
| 11-12 | 2235 | 265 | 432 | 150 | 0.566 | 0.00696 |
| 13-16 | 3191 | 276 | 404 | 149 | 0.540 | 0.00700 |

**Table 2.3:** Related to Figure 2; Overlap between axis patterning and the minimal Vienna Tile enhancers. The first column is a list of the 99 axis patterning enhancers, ordered by the percent overlap (third column) with the minimal Vienna Tile enhancer in the second column. The fourth column indicates whether the minimal Vienna Tile enhancer was expressed concurrently with the axis patterning enhancer during stages 4-6.

| AP/DV | Minimal Vienna Tile | Overlap (%) | Minimal Vienna Tile Enhancer Active in Stages 4-6? |
|---|---|---|---|
| eve-37 | VT14362.1 | 100 | yes |
| pnr | VT42370.1 | 100 | no |
| rho | VT24025.1 | 100 | no |
| tup | VT9666.2 | 100 | yes |
| vnd | VT54910.1 | 100 | yes |
| tld | VT46946.1 | 97 | yes |
| eve-46 | VT14367.1 | 93.9 | yes |
| h-34 | VT27677.1 | 90.8 | yes |
| slp-ecto | VT1967.1 | 85.6 | yes |
| kni-cis | VT33934.2 | 79.7 | yes |
| zen | VT37509.2 | 79.5 | yes |
| eve-37 | VT14361.2 | 78.7 | yes |
| gt-P | VT55790.2 | 78.4 | yes |
| gt-1 | VT55795.2 | 75.5 | yes |
| slp-B | VT1971.2 | 75.2 | yes |
| nub-blst | VT6450.1 | 75 | yes |
| gt-P | VT55791.1 | 74.6 | yes |
| eve-46 | VT14366.1 | 74.3 | yes |
| slp1-head | VT1967.2 | 72.5 | yes |
| pdm2-plus1 | VT6483.1 | 71.4 | yes |
| eve-15 | VT14368.1 | 70.6 | yes |
| prd-1 | VT6169.2 | 64.7 | yes |
| otd-E | VT58874.1 | 63.1 | yes |
| mes3 | VT28266.1 | 62 | yes |

# Chapter 3

# Two promoters integrate multiple enhancer inputs to drive wild-type *knirps* expression in the *D. melanogaster* embryo

Thank you to Rachel Waymack and Mario Gad for their help in making this work possible!

## 3.1 Abstract

Proper development depends on precise spatiotemporal gene expression patterns. Most genes are regulated by multiple enhancers and often by multiple core promoters that generate similar transcripts. We hypothesize that these multiple promoters may be required either because enhancers prefer a specific promoter or because multiple promoters serve as a redundancy mechanism. To test these hypotheses, we studied the expression of the *knirps*

locus in the early *Drosophila melanogaster* embryo, which is mediated by multiple enhancers and core promoters. We found that one of these promoters resembles a typical "sharp" developmental promoter, while the other resembles a "broad" promoter usually associated with housekeeping genes. Using synthetic reporter constructs, we found that some, but not all, enhancers in the locus show a preference for one promoter. By analyzing the dynamics of these reporters, we identified specific burst properties during the transcription process, namely burst size and frequency, that are most strongly tuned by the specific combination of promoter and enhancer. Using locus-sized reporters, we discovered that even enhancers that show no promoter preference in a synthetic setting have a preference in the locus context. Our results suggest that the presence of multiple promoters in a locus is both due to enhancer preference and a need for redundancy and that "broad" promoters with dispersed transcription start sites are common among developmental genes. Our results also imply that it can be difficult to extrapolate expression measurements from synthetic reporters to the locus context, where many variables shape a gene's overall expression pattern.

## 3.2 Introduction

Diverse processes in biology, from early development to the maintenance of homeostasis, rely on the regulation of gene expression. Enhancers and promoters are the primary regions of the genome that encode these gene regulatory programs. Both enhancers and promoters are characterized by clusters of sequence motifs that act as platforms for protein binding, allowing for the integration of a spectrum of signals in the cellular environment. The majority of studies that dissect enhancer or promoter function typically investigate each in isolation, which assumes that their function is largely modular. In practice, this means that we assume an enhancer drives generally the same pattern, regardless of promoter, and that promoter strength is independent of the interacting enhancer. However, there is evidence that there can

be significant "interaction terms" between promoters and enhancers, with enhancer pattern shaped by promoter sequence, and promoter strength influenced by an enhancer [46, 63, 128].

Therefore, a key question is precisely how the sequences of an enhancer and a promoter combine to dictate overall expression output. Adding to the complexity of this question, developmental genes often have multiple enhancers, and many metazoan genes have alternative promoters [14, 86, 136, 138]. In a locus, multiple enhancers exist either because they drive distinct expression patterns or, in the case of seemingly redundant shadow enhancers, because they buffer noise in the system [82]. Though RAMPAGE data shows that >40% of developmentally expressed genes have more than one promoter [5], the role of multiple promoters has been relatively less explored. In some cases, alternative promoters drive distinct transcripts, but *hunchback* is a notable example of a gene with two highly conserved promoters that produce identical transcripts [99, 138].

This suggests there may be additional explanations for the prevalence of multiple promoters. One possibility is molecular compatibility—promoters can preferentially engage with different enhancers depending on the motif composition and proteins recruited to each [160, 163]. For example, enhancers bound by either the transcription factors (TFs) Caudal or Dorsal tend to interact with Downstream Promoter Element (DPE)-containing promoters [70, 180] and Bicoid-dependent *hunchback* transcription seems to depend on the presence of a TATA box and Zelda site at one promoter [99]. Another possibility is that having multiple promoters provides redundancy needed for robust gene expression, much like shadow enhancers.

To distinguish between these hypotheses, an ideal model is a gene with (1) multiple promoters that contain different promoter motifs and drive similar transcripts and with (2) multiple enhancers bound by different TFs. The *Drosophila* developmental gene *knirps* (*kni*) fits these criteria. It is a key developmental TF that acts in concert with other gap genes to direct anterior-posterior axis patterning of the early embryo. *kni* has two core promoters that drive nearly identical transcripts (only differing by five amino acids at the N-terminus)

and that are both used during the blastoderm stage (**Figure 3.1A − C**). Here, we define the core promoter as the region encompassing the transcription start site (TSS) and the 40bp upstream and downstream of the TSS [162]. Also, like many early developmental genes, its precise pattern of expression in the blastoderm is coordinated by multiple enhancers (**Figure 3.11A**). These characteristics make the *kni* locus a good system in which to examine the roles of multiple promoters in a single gene locus.

We used several approaches to delineate the roles of these two promoters. To examine the molecular compatibility of different *kni* enhancer-promoter pairs in a controlled setting, we created reporter constructs of eight *kni* enhancer-promoter pairs driving expression of an MS2 reporter. We found that some *kni* enhancers are able to interact with multiple promoters similarly, while others have a strong preference for one. By using the MS2 system to measure the transcription dynamics, we also determined the molecular events that lead to these preferences. Next, analysis of a *kni* locus reporter demonstrated that locus context can affect promoter-enhancer preferences and indicates that promoters both have different jobs and provide some amount of redundancy. Finally, we explored the role of different promoter motifs in specifying expression dynamics by using constructs with promoter mutations. Examining the *kni* locus has allowed us to (1) determine how transcription dynamics are impacted by molecular compatibility, (2) determine the roles of multiple promoters in a locus, and (3) probe how the motif content of promoters produces a particular expression output.

## 3.3 Results

### 3.3.1 Selection of enhancers and promoter pairs tested

*knirps* has two conserved promoters that drive very similar transcripts (**Figure 3.1A; Figure S3.1A and B**). Most previous studies discuss the role of a single *kni* promoter (promoter 1), though in practice, many of the constructs used in these studies actually contained both promoters, since promoter 2 is located in a *kni* intron [13, 35, 117, 124]. While more transcripts initiate from promoter 1 throughout most of development (**Figure 3.1B**), based on two different measures of transcript abundance, both promoters appear to be active during nuclear cycle 14, 2 - 3 hours after fertilization (**Figure 3.1B and 3.1C**) [6, 102]. These two promoters are distinguished by their motif content and by their "shape" (**Figure 3.1E**). Promoter 1 is composed of multiple Initiator (Inr) motifs, each of which can specify a transcription start site. These Inr motifs enable promoter 1 to drive transcription initiation in a 124 bp window, characteristic of a "broad" or "dispersed" promoter typically associated with housekeeping genes [70, 145]. There is a single DPE element in promoter 1; however, its significance is somewhat unclear, as it is only the canonical distance from a single, somewhat weak, Inr motif within the initiation window. Promoter 2 is composed of Inr, TATA Box and DPE motifs. This motif structure leads promoter 2 to initiate transcription in a 3bp region, which is characteristic of the "sharp" or "focused" promoter shape typically associated with developmental genes (**Figure S3.1C**).

To select key early embryonic *kni* enhancers, we took into account the expression patterns driven by the enhancers and their overlap in the locus. We split the enhancers into three groups based on their expression patterns and selected one representative enhancer per group—enhancers driving a diffuse posterior stripe (kni_proximal_minimal), enhancers driving a sharp posterior stripe (kni_KD, the "classic" *kni* posterior stripe enhancer), and enhancers driving the anterior band (kni_-5) (**Figure 3.1A**). Among the enhancers driving

a sharp posterior stripe, we decided to examine another enhancer, VT33935, in addition to kni_KD [117]. VT33935 was identified in a high-throughput screen for enhancer activity [81] and has only minimal overlap with the kni_KD enhancer but drives the same posterior stripe of expression. This suggests it may be an important contributor to *kni* regulation.

To determine the TF inputs to these enhancers, we scanned each enhancer using the motifs of TFs regulating early axis specification and calculated an overall binding capacity for each enhancer-TF pair (**Figure 3.2A and S3.2**). We found that kni_KD and VT33935 seem to be regulated by similar TFs, which suggests that together they comprise one larger enhancer. Here, we studied them separately, as historically kni_KD has been considered the canonical enhancer driving posterior stripe expression [117]. Since kni_KD, VT33935, and kni_proximal_minimal drive overlapping expression patterns, they can be considered a set of shadow enhancers. Despite their similar expression output, kni_proximal_minimal has different TF inputs than the other two, including different repressors and autoregulation by Kni itself [125]. kni_-5 is the only enhancer that controls expression of a ventral, anterior band. Accordingly, this is the only enhancer of the four that has dorsal-ventral TF inputs (Dorsal and Twist) (**Figure 3.2A**) [139]. In sum, analyses of the total binding capacity of these enhancers demonstrate that they are bound by different TFs (**Figure 3.2A**).

By using this set of endogenously interacting enhancers and promoters with varied motif content, we can elucidate the functional value of having multiple promoters. In particular, we can determine whether multiple promoters exist because different enhancers work with different promoters, or whether having multiple promoters provides necessary redundancy in the system, or some combination of the two.

### 3.3.2 Some enhancers tolerate promoters of different shapes and composition

To characterize the inherent ability of promoters and enhancers to drive expression, without complicating factors like enhancer competition, promoter competition, or variable enhancer-promoter distances, we created a series of eight transgenic enhancer-promoter reporter lines. Each reporter contains one enhancer and one promoter directly adjacent to each other, followed by MS2 stem loops inserted in the 5' UTR of the yellow gene (**Figure 3.1D**, see Methods for details). These tagged transcripts are bound by MCP-GFP fusion proteins, yielding fluorescent puncta at the site of nascent transcription. The fluorescence intensity of each spot is proportional to the number of transcripts in production at a given moment [45].

When considering the expression output driven by these enhancer-promoter combinations, several outcomes are possible. One possible outcome is that one promoter is simply stronger than the other – consistently driving higher expression, regardless of which enhancer it is paired with. Another possibility is that each enhancer drives higher expression with one promoter than with the other, but this preferred promoter differs between enhancers. This would suggest that promoter motifs and shape affect their ability to successfully interact with enhancers with different bound TFs to drive expression. Lastly, it is possible that some enhancers drive similar expression with either promoter, this suggests that the particular set (and orientation) of the TFs recruited to those enhancers allow them to transcend the differences in promoter architecture.

When comparing the mean expression levels, we found that some enhancers (kni_-5 and kni_proximal_minimal) have relatively mild preferences for one promoter over the other (**Figure 3.2B**; two-sided $t$-test comparing kni_-5-promoter1 vs. kni_-5-promoter2, $p = 0.12$ and kni_proximal_minimal-promoter1 vs. kni_proximal_minimal-promoter2 $p = 9.8 \times 10^{-5}$). Despite the significant differences between these enhancer-promoter constructs, the

effect size is relatively small, with the largest difference in mean expression being 1.2-fold. This suggests that the TFs recruited to these enhancers can interact with very different promoters more or less equally well. On the other hand, kni_KD and VT33935 respectively drive 2.9-fold and 3.2-fold higher expression with promoter 2 than promoter 1 at 62.5% embryo length (**Figure 3.2C**; one-sided $t$-test $p < 2.2 \times 10^{-16}$ for both). This suggests that the TFs recruited to kni_KD and VT33935, which are similar, (**Figure 3.2A**) limit their ability to successfully drive expression with promoter 1, which is a dispersed promoter. Taken together, this implies a simple model of promoter strength is not sufficient to account for these results. Instead, it is the combination of the proteins recruited to both enhancers and promoters that set expression levels, with some enhancers interacting equally well with both promoters and others having a preference.

These differences in enhancer preference or lack thereof may be mediated by the particular TFs recruited to them and the motifs present in the promoters. Previous researchers have found that the developmental TFs, Caudal (Cad) and Dorsal (Dl), tend to regulate genes with DPE motifs and drive lower expression when DPE has been eliminated [70, 180]. In addition, computational analysis of TF-promoter motif co-occurrence patterns indicates that Bcd shows a similar enrichment for DPE-containing promoters and a depletion for Inr- and TATA box-containing promoters when DPE is absent (**Figure S3.2**). A study also indicated that Bcd can work in conjunction with Zelda to activate a TATA Box-containing promoter, but this combination does not appear to be widely generalizable [99]. In accordance with that, we find that all four *kni* enhancers, which bind Cad and Bcd, drive relatively high expression with the DPE-containing promoter 2. Interestingly, in the case of kni_-5 and kni_pm, we find that they can also drive similarly high expression with the series of weak Inr sites that composes promoter 1. This indicates that while some factors mediating enhancer-promoter preference have been identified, there are additional factors we have yet to discover that are playing a role.

We also calculated the expression noise associated with each construct and plotted it against the expression output of each. Previous studies have suggested that TATA-containing promoters generally drive more noisy expression [129, 130]. Among our constructs, expression noise is generally inversely correlated with mean expression (**Figure 3.2C and 3.2D**), and the TATA-containing promoter 2 does not have uniformly higher noise than the TATA-less promoter 1. However, some constructs, notably those containing kni_-5, have higher noise than others with similar output levels, suggesting that, in this case, promoters alone do not determine expression noise.

### 3.3.3 Simple model of transcription and molecular basis of burst properties

To unravel the molecular events that result in these expression differences, we consider our results in the context of the two-state model of transcription [123, 159]. Here, the promoter is either (1) in the inactive state ("OFF"), in which RNA polymerase cannot initiate transcription or (2) in the active state ("ON"), in which it can (**Figure 3.3A**). The promoter transitions between these two states with rates $k_{\text{on}}$ and $k_{\text{off}}$, with the transitions involving both the interaction of the enhancer and promoter and the assembly of the necessary transcriptional machinery. This interaction may be through direct enhancer-promoter looping or through the formation of a transcriptional hub, a nuclear region with a high concentration of TFs, co-factors, and RNA polymerase [98]. For simplicity, we will use looping as a shorthand to include both scenarios. In its active state, the promoter produces mRNA at rate $r$, and given our ability to observe only nascent transcripts, the mRNA decay rate $\mu$ denotes the diffusion of mRNA away from the gene locus.

We track these molecular events by analyzing the transcription dynamics driven by each reporter and quantifying several properties. Total expression is simply the integrated signal

driven by each reporter. The burst duration is the period of active transcription, and is dependent on $k_{\text{off}}$, the rate of dissociation of enhancer and promoter looping (**Figure 3.3B**). The burst size, or number of transcripts produced per burst, depends on the burst duration and the RNA Pol II initiation rate. (Short, aborted transcripts and paused PolII are not visible in MS2 measurements). The burst frequency, or the inverse of the time between two bursts, depends on both $k_{\text{on}}$ and $k_{\text{off}}$. Previous work in the early embryo has shown burst duration (and thus $k_{\text{off}}$) to be reasonably consistent regardless of enhancer and promoter [166, 177]. Within this regime, burst frequency is mainly dependent on $k_{\text{on}}$. We used this model to characterize how the transcription output produced is affected by different combinations of the *kni* enhancers and promoters.

### 3.3.4 Using GLMs to parse the role of enhancers, promoters, and their interactions

To parse the role of enhancers, promoters, and their interactions more clearly in determining expression levels in these reporters, we built separate generalized linear models (GLMs) to describe each transcriptional property. We visually represented the model using a bar graph (**Figure 3.4A**) in which the contributions of enhancer, promoter, and their interactions are represented in bars of green, purple, and brown, respectively (**Figure 3.4B**). Since the relative differences in expression driven by different enhancer-promoter pairs are generally consistent across the AP axis, we used the expression levels at the location of maximum expression along the AP axis (22% and 63% for the anterior band and posterior stripe, respectively, **Figure 3.2C**).

If the molecular compatibility of the proteins recruited to the enhancer and promoter are important in determining a particular property, then we should find the interaction terms (in brown) to be sizeable in comparison with those of the enhancers (in green) and promoter

(in purple). If not, the interaction terms will be relatively small. To develop an intuition for this formalism, we first built a GLM to describe total expression output. Using the GLM, we can see that enhancer, promoter, and interactions terms each play an important role in determining the expression output (**Figure 3.4C**), consistent with our qualitative interpretation above.

To determine which molecular events are modulated by molecular compatibility, we then applied this same GLM structure to each burst property. For example, molecular compatibility could increase the probability of enhancer-promoter loop formation, hence increasing the burst frequency. Alternatively, molecular compatibility could increase the rate at which RNA PolII initiates transcription, increasing burst size.

### 3.3.5 Burst frequency and initiation rate are the primary determinants of expression levels

We found that the differences in total expression output are primarily mediated through differences in burst size (**Figure 3.4E**) and burst frequency (**Figure 3.4D**). Burst duration is very consistent across all constructs (**Figure 3.4F**). While the enhancer, promoter, and interaction terms all have a significant impact on duration (multivariate ANOVA; enhancer: $p = 4.4 \times 10^{-10}$; promoter: $p = 4.1 \times 10^{-5}$; interaction: $p = 4.6 \times 10^{-5}$), the effect size is small, with the largest difference being only 1.3-fold. Since burst size can be modulated by initiation rate and burst duration, and burst duration is relatively constant, this suggests that initiation rate and burst frequency are the primary dials used to tune transcription in these synthetic constructs.

Burst size is strongly dependent on both the enhancer and interaction terms; the interaction terms are a proxy for molecular compatibility. Of the variability in burst size explained by this model, enhancers and interaction terms account for 67.6% and 23.7% of the variance, re-

spectively (**Figure 3.4E**). The differences in burst size were mainly achieved by tuning PolII initiation rate (**Figure 3.4G**). Conversely, burst frequency is dependent on promoter and enhancer identity, with negligible interaction terms (**Figure 3.4D**). Since burst frequency mainly depends on association rate ($k_{\text{on}}$), this suggests that both enhancers and promoters play a large role in determining the likelihood of promoter activation, with molecular compatibility only minimally affecting this likelihood.

It is somewhat surprising that molecular compatibility plays only a small role in determining $k_{\text{on}}$, since one might expect the interactions between the proteins recruited to promoters and enhancers would determine the likelihood of promoter-enhancer looping. This may be the result of the design of these constructs, with promoters and enhancer immediately adjacent to each other, and this may differ in a more natural context (see below). However, we do observe that molecular compatibility is important in determining the PolII initiation rate. This suggests that the TFs and cofactors recruited to each reporter may act synergistically to both recruit RNA PolII to the promoter and promote its successful initiation. In sum, these results indicate that not only do enhancer, promoter, and their molecular compatibility affect expression output, but they do so by tuning different burst properties in this synthetic setting.

### 3.3.6 Despite promoter 2's compatibility with kni_-5, promoter 1 primarily drives anterior expression in the locus context

The constructs measured thus far only contain a single enhancer and promoter, and therefore measure the inherent ability of a promoter and enhancer to drive expression. However, in the native locus, other complications like differing enhancer-promoter distances, enhancer competition, or promoter competition may impact expression output. To measure the effect of these complicating factors, we cloned the entire *kni* locus into a reporter construct and

measured the expression patterns and dynamics of the wildtype locus reporter (wt) and reporters with either promoter 1 or 2 knocked out (Δp1 and Δp2) (**Figure 3.5A**). Due to the large number of Inr motifs, we made the Δ*p*1 construct by replacing promoter 1 with a piece of lambda phage DNA. To make the Δp2 construct, we inactivated the TATA, Inr, and DPE motifs by making several mutations (see Methods for additional details).

In the anterior, the kni_-5 enhancer is solely responsible for driving expression. Therefore, by comparing the expression output from the wildtype locus reporter and the kni_-5-promoter reporters in the anterior, we can measure the effect of the locus context, i.e. multiple promoters, differing promoter-enhancer distance, or other DNA sequence features. If the kni_-5-promoter reporters capture their ability to drive expression in the locus context, we would expect the locus reporter to drive expression equal to the sum of the kni_-5-p1 and kni_-5-p2 reporters. In contrast to this expectation, in the anterior band, the locus reporter drives a much lower level of expression than the sum of the two kni_-5 reporters (**Figure 3.5B**, dark purple vs black bar). In fact, the level is similar to the expression output of kni_-5 paired with either individual promoter, suggesting that kni_-5's expression output is altered by the locus context.

The observed sub-additive behavior may arise in several ways. It may be that promoter competition similarly reduces the expression output of both p1 and p2 in the anterior. In this case, knocking out either promoter would produce wildtype levels of expression, as competition would be eliminated. Alternatively, the ability to drive expression in the locus context could be uneven between the promoters. If this is the case, we would expect the promoter knockouts to have different effects on expression.

Consistent with the second scenario, we find that when promoter 2 is eliminated in the *kni* locus construct, the expression in the anterior remains essentially the same (two-sided $t$-test comparing mean expression levels of wt vs. Δp2, $p = 0.62$), while a promoter 1 knockout has a significant impact on expression levels (one-sided $t$-test comparing mean expression levels

of wt vs. $\Delta$p1, $p < 2.2 \times 10^{-16}$; **Figure 3.5B**). Thus, promoter 1 is sufficient to produce wildtype expression levels and patterns in the locus. The noise and the burst properties of the WT *kni* locus construct and the promoter 2 knockout are also nearly identical to the wildtype locus, further supporting the claim of promoter 1 sufficiency in the anterior (**Figure 3.5C − G**). Notably, even in a locus that contains promoters with and without a canonically placed DPE element (promoter 2 vs promoter 1), a Cad- and Dl-binding enhancer like kni_-5, can still primarily rely on the DPE-less promoter 1 to drive transcription.

When promoter 1 is eliminated from the locus, expression is cut to about one third of that of the wildtype locus construct, which is also lower than the expression output of the kni_-5-p2 construct. Thus, unlike promoter 1, promoter 2 loses its ability to drive wildtype levels of expression in the context of the locus. As promoter 2 is ~650bp upstream of promoter 1, this extra distance between kni_-5 and promoter 2 may be sufficient to reduce promoter 2's ability to drive expression. Alternatively, other features of the *kni* locus, such as the binding of other proteins or topological constraints, may interfere with the ability of the kni_-5 enhancer to effectively interact with promoter 2. The drop in expression is mediated by a tuning down of all burst properties (**Figure 3.5D − G**). In sum, the kni_-5 enhancer preferentially drives expression via promoter 1 in the locus, even though enhancer-promoter constructs indicate that it is equally capable of driving expression with promoter 2. When promoter 1 is absent from the locus, promoter 2 is able to drive a smaller amount of expression, suggesting that it can serve as a backup, albeit an imperfect one.

### 3.3.7 In the posterior, both promoters are required for wildtype expression levels

The posterior stripe is controlled by three enhancers, with kni_proximal_minimal producing similar levels of transcription with either promoter, and the other two enhancers strongly

preferring promoter 2 and driving lower expression overall (**Figure 3.2B**). Therefore, when considering the posterior stripe, the expression output of the locus reporter may differ from the individual enhancer-reporter constructs due to promoter competition, enhancer competition, different promoter-enhancer distances, or other DNA features. By comparing the sum of the six relevant enhancer-promoter reporters to the output of the locus reporter, we can see that the locus construct drives considerably lower expression levels than the additive prediction (**Figure 3.5B**, dark purple vs black bar). In fact, the locus reporter output levels are similar to the sum of the enhancer-promoter 2 reporters, suggesting that promoter 2 could be solely responsible for expression in the posterior, despite kni_proximal_minimal's ability to effectively drive expression with promoter 1. If promoter 2 is sufficient for posterior stripe expression, we would predict that the promoter 1 knockout would have a relatively small effect, while a promoter 2 knockout would greatly decrease expression in the posterior.

In contrast to this expectation, both promoter 1 and promoter 2 knockouts have a sizable effect on expression output, indicating that both are required for wildtype expression levels in the posterior (**Figure 3.5B**, light gray and gray bars). Specifically, knocking out promoter 2 severely reduces expression in the posterior stripe, producing about half the expression of the summed outputs of the enhancer-promoter1 constructs (**Figure 3.5B**, light gray vs light purple bars). Knocking out promoter 1 also reduces expression in the posterior stripe but not as severely as knocking out promoter 2 (**Figure 3.5B**, gray vs light gray bars). The promoter 1 knockout generates about half the expression of the summed expression output of the enhancer-promoter2 constructs (**Figure 3.5B**, gray vs purple bars). In both cases, the results indicate that the differences in locus context cause the enhancers to act sub-additively, even when only one promoter is present. The promoter knockouts also allow us to examine how they tune expression output. Knocking out either promoter impacts burst size (and thus initiation rate) and burst frequency, though knocking out promoter 2 has a more severe impact (**Figure 3.5D, 3.5E and 3.5G**). These results show that, in the posterior, both promoters are required to produce WT expression levels when considered in the endogenous

locus setting (**Figure 3.5B**, light and dark gray vs black bars). This is despite the fact that enhancer-promoter reporters indicate that, in the absence of competition, promoter 2 alone would suffice (**Figure 3.5B**, purple vs black bars).

### 3.3.8 PolII initiation rate is a key burst property that is tuned by promoter motif

Studying these enhancers and promoters in the locus context demonstrated that distance and competition affect a promoter's ability to drive expression, but now we narrow our focus to promoter 2's remarkable compatibility with enhancers that bind very different sets of TFs. To dissect how its promoter motifs enable promoter 2 to be so broadly compatible, we again made enhancer-promoter reporter lines in which one enhancer and one promoter are directly adjacent to each other, but this time the promoter is a mutated promoter 2 in which the TATA Box and DPE motifs have been eliminated (**Figure 3.6A**, see Methods for details). This allows us to determine whether a single, strong Inr site (mutated promoter 2) can perform similarly to a series of weak Inr sites (promoter 1) and to clarify the role of TATA Box and DPE motifs in tuning burst properties.

Promoter 2 is characterized by two TATA Boxes, an Inr motif, and a DPE motif. Previously, much research has focused on comparing TATA-dependent with DPE-dependent promoters; however, many promoters contain both. Here, we consider how the presence of both may impact transcription. We know that each of these motifs recruits subunits of TFIID, with TATA Box recruiting TBP or TRF1 [54, 62, 73], Inr recruiting TAF1 and 2 [20, 170], and DPE recruiting TAF6 and 9 [143], as well as other co-factors like CK2 and Mot1 [65, 92]. Strict spacing between TATA-Inr and Inr-DPE both facilitate assembly of all these factors and others into a pre-initiation complex [15, 37]. It is likely that a promoter with all three motifs will behave similarly, with the addition of each motif further tuning the composition,

configuration, or flexibility of the transcriptional complex. Given this, elimination of the TATA Box and DPE motifs may weaken the promoter severely through loss of cooperative interactions, especially for kni_KD and VT33935, which are significantly more compatible with promoter 2 than promoter 1. Alternatively, the single strong Inr site may be sufficient to recruit the necessary transcription machinery, especially in the case of kni_-5 and kni_proximal_minimal, which work well with the series of weak Inr sites that composes promoter 1.

When compared to promoter 1, we see that promoter 1-compatible enhancers (kni_-5 and kni_proximal_minimal) drive lower expression with a single Inr than with a series of weak Inr sites (**Figure 3.6B**, light purple bars). In contrast, enhancers less compatible with promoter 1 (kni_KD and VT33935) drive higher expression with a single Inr site than promoter 1 even without the TATA Box and DPE sites (**Figure 3.6B**, light purple bars), suggesting that the strong Inr is the key to better expression output with these enhancers. For all enhancers, the resulting expression change appears to be mediated mainly through a decrease in burst size due to a reduction in initiation rates (**Figure 3.6D – F**).

Given that all four enhancers are compatible with promoter 2, and promoter 2 appears to achieve higher expression by tuning PolII initiation rates, we posit that TATA Box and DPE are what help promoter 2 drive high initiation rates. When comparing p2ΔTATAΔDPE with promoter 2, we see that all enhancers produce lower expression (**Figure 3.6B**, dark purple bars), and this is mediated mainly through tuning burst size (**Figure 3.6D**) and, for some enhancers, also burst frequency (**Figure 3.6C**). Notably, burst size (and thus polymerase initiation rate), which were most dependent on molecular compatibility, are affected the most by the elimination of the TATA Box and DPE motifs (**Figure 3.6D and 3.6E**), indicating that molecular compatibility plays an important role mediating high expression output. Interestingly, even in the absence of the TATA Box and DPE motifs, the one strong Inr site is sufficient to produce higher expression with the enhancers less compatible

with promoter 1 (kni_KD and VT33935), and this increased expression is also mediated by higher polymerase initiation rates (**Figure 3.6B and 3.6F**, light purple bars). In conclusion, enhancers seem to fall into classes, which behave in similar ways with particular promoters, and the molecular compatibility that appears to tune PolII initiation rates seems to be mediated by the promoter motifs present in an enhancer-specific manner.

## 3.4 Discussion

We dissected the *kni* gene locus as a case study of the role of multiple promoters in controlling a single gene's transcription dynamics. Synthetic enhancer-promoter reporters allowed us to measure the ability of *kni* enhancer-promoter pairs to drive expression in the absence of complicating factors like promoter or enhancer competition. Using these reporters, we found that some promoters are broadly compatible with many enhancers, whereas others only drive high levels of expression with some enhancers. A detailed analysis of the transcription dynamics of these reporters indicates that the molecular compatibility of the proteins recruited to the enhancer and promoter tune expression levels by altering the initiation rate of transcriptional bursts. In the context of the whole locus, we found that some enhancer-promoter pairs drive lower expression than their corresponding synthetic reporters, due to the effects of promoter and enhancer competition, distance, or other factors. In fact, while the synthetic reporters indicate that both promoters can drive similarly high levels of expression in the anterior, in the locus, promoter 1 drives most of the expression, with promoter 2 supporting some low levels of expression in the absence of promoter 1. In the posterior, both promoters appear to be necessary to achieve wildtype levels of expression with enhancer competition leading to sub-additive expression. By mutating promoter motifs in the synthetic enhancer-reporter constructs, we found that the effects of promoter motif mutations fall into two different classes, depending on the enhancer that is paired with the

promoter. This suggests that there may be several discrete ways that a promoter can be activated by an enhancer, depending on the proteins recruited to each. Returning to our original hypotheses to explain the presence of two promoters in a single locus, we find that both differing enhancer-promoter preferences and a need for expression robustness in the face of promoter mutation may play a role.

Our work has highlighted the importance of both of *kni*'s promoters. Previous studies have almost exclusively focused on *kni*'s promoter 1 [117, 124], which unexpectedly looks like a typical housekeeping gene promoter, with a dispersed shape and series of weak Inr sites [162]. It is *kni*'s promoter 2, with its focused site of initiation and composition of TATA Box, Inr, and DPE motifs, that looks like a canonical developmental promoter [162].

Interestingly, despite only discussing promoter 1, in practice, studies interrogating the behavior of multiple *kni* enhancers often included both promoters, as promoter 2 is found in a *kni* intron [13, 35]. Our analysis clearly demonstrates both promoters' vital role in normal *kni* expression.

With these observations in mind, we wanted to determine the prevalence of a two promoter structure, with one broad and one sharp. To do so, we used the RAMPAGE data set, which includes a genome-wide survey of promoter usage during the 24 hours of *Drosophila* embryonic development [6] and cross-referenced these promoters with those in the Eukaryotic Promoter Database, which is a collection of experimentally validated promoters [33]. We found that 13% of embryonically expressed genes have at least two promoters. When we considered the two most commonly-used promoters, there is a clear trend of a broader primary (most used) promoter (median = 91bp) and a sharper secondary promoter (median = 42bp) (**Figure S3.1C**). This trend is still present if the genes are split into developmental and housekeeping genes, with developmental promoters (median = 43bp) generally more focused than housekeeping promoters (median = 90bp), as expected (**Figure S3.1D and E**). Among the primary promoters of developmental genes, 58% consist of a series of weak Inr

sites, much like *kni* promoter 1. This suggests that this promoter shape and motif content in developmental promoters may be more common than previously expected and should be explored.

There is growing evidence that promoter motifs play a role in modulating different aspects of transcription dynamics. However, the role of each motif can vary from one locus to the next. In the "TATA-only" *Drosophila snail* promoter, the TATA Box affects burst size by tuning burst duration [127]. In the mouse PD1 proximal promoter, which consists of a CAAT Box, TATA Box, Sp1, and Inr motif, the TATA box may tune burst size and frequency [56]. A study of a synthetic *Drosophila* core promoter and the *ftz* promoter found that the TATA box tunes burst size by modulating burst amplitude and that Inr, MTE, and DPE tune burst frequency [177]. TATA Box also appears to be associated with increased expression noise, as TATA-containing promoters tend to drive larger, but less frequent transcriptional bursts [129]. In contrast to TATA Box, Inr appears to be associated with promoter pausing, e.g. by adding a paused promoter state in the Inr-containing *Drosophila Kr* and *Ilp4* promoters [127]. In fact, a Pol II ChIP-seq study indicates that paused developmental genes appear to be enriched for GAGA, Inr, DPE, and PB motifs [129].

Similarly, the TFs bound at enhancers can affect transcription dynamics in diverse ways. Exploration of the role of TFs in modulating burst properties has indicated that BMP and Notch can tune burst frequency and duration, respectively [38, 63, 88]. Work that considers both the promoters and enhancer simultaneously have come to differing conclusions. Work in human Jurkat cells, wherein 8000 genomic loci were integrated with one of three promoters, showed that burst frequency is modulated at weakly expressed loci and burst size modulated at strongly expressed loci [28]. Work in *Drosophila* embryos and in mouse fibroblasts and stem cells suggest that stronger enhancers produce more bursts, and promoters tune burst size [44, 87]. On the whole, this work indicates that promoter motifs and the TFs binding enhancers can act to tune burst properties in a myriad of ways. Given the wide range

of possibilities, it is likely that setting, i.e. the combination of promoter motifs and the interacting enhancers, is particularly important in determining the resulting transcription dynamics.

Our work supports this notion. Notably, eliminating the TATA Box and DPE from promoter 2 seems to reinforce the idea that we have two classes of enhancers that behave in distinct ways with these promoters due to the different TFs bound at these enhancers. We find that polymerase initiation rate is a key property tuned by the molecular compatibility of the proteins recruited to the enhancer and promoter. Our observation is in contrast to previous studies in which PolII initiation rate seems constant despite swapping two promoters with different motif content or altering BMP levels or the strength of TF's activation domains [63, 141] and is tightly constrained for gap genes [183]. We suggest that the differences we see in our work, where initiation rate depends on molecular compatibility, versus other work, where initiation rate is controlled by other factors, again reinforces the idea that the role of any particular promoter motif or TF binding site can be highly context dependent.

Together, ours and previous work demonstrate that deriving a general set of rules to predict transcription dynamics from sequence is a challenge because the space of promoter motif content and enhancer TF binding site arrangements is large. The proteins recruited to both promoters and enhancer can combine to make transcriptional complexes with different constituent proteins, post-translational modifications, and conformations, that may even vary as a function of time. Due to the vast possibility space and context-dependent rules, most work has only scratched the surface of how promoter motifs or enhancers can modulate burst properties, suggesting a field rich for future investigation.

## 3.5 Acknowledgements

The authors wish to thank Leonila Lagunes, Srikiran Chandrasekaran, and all the Wunderlich lab members for helpful comments on the manuscript and Ali Mortazavi, Kyoko Yokomori, Kevin Thornton, and Rahul Warrior for useful discussion on the project. The authors thank Flo Ramirez for data analysis that inspired some of this work.

## 3.6 Funding

This work is supported by NIH-NICHD R00 HD073191 and NIH-NICHD R01 HD095246 (to ZW) and the US DoE P200A120207 and NIH-NIBIB T32 EB009418 (to LL).

## 3.7 Conflicts of Interest

None to report.

## 3.8 Materials and Methods

### 3.8.1 Datasets used in this study

The experimentally validated promoters and their experimentally determined transcription start sites (TSSs) were obtained from the Eukaryotic Promoter Database (EPD) New [33]. They were cross-referenced with the RNA Annotation and Mapping of Promoters for Analysis of Gene Expression (RAMPAGE) data obtained from five species of *Drosophila* [6] to form a high-confidence set of promoters for which promoter usage during development could be

evaluated. Single embryo RNA-seq obtained by Lott, et al. was indexed (with a $k$ of 17 for an average mapping rate of 96%) and quantified using Salmon v0.12.01. The resulting transcript-specific data was used to further resolve *kni* promoter usage during nuclear cycle 14 [102, 122]. Housekeeping genes were defined as in Corrales, et al. where genes were defined as housekeeping if their expression exceeded the 40th percentile of expression in each of 30 time points and conditions using RNA-seq data collected by modEncode [23] and a list of these can be found in the Supplementary Materials (File S1).

To study TF-promoter motif co-occurrence, we collected a total of ~1000 enhancer-gene pairs expressed during development in *Drosophila*. The majority were identified by traditional enhancer trapping (REDfly & CRM Activity Database 2, or CAD2) and consist of non-redundant experimentally characterized enhancers [12, 51]. About 15% were identified through functional characterization of 7000 enhancer candidates using high throughput *in situ* hybridization (Vienna Tile, or VT); these VT enhancers have been limited to those expressed during stages 4 – 6. The remaining 1% of enhancer-gene pairs have been identified through 4C-seq [49] and are active 3 – 4 hours after egg laying (stages 6 – 7). A list of these enhancer-promoter pairs and their coordinates can be found in the Supplementary Materials (File S2).

### 3.8.2 Motif prediction in promoters and enhancers

For enhancers, TF binding site prediction was performed using Patser [57] with position weight matrices (PWMs) from the FlyFactor Survey [182] and a GC content of 0.406. Each element in the PWM was adjusted with a pseudocount relative to the intergenic frequency of the corresponding base totaling 0.01. For TFs that had multiple PWMs available, PWMs built from the largest number of aligned sequences were chosen; that of Stat92E was taken from an older version of the FlyFactor Survey. For promoters, the transcription start clusters

(TSCs) [6] and the adjoining ±40bp were scanned for Inr, TATA Box, DPE, MTE, and TCT motifs using ElemeNT and the PWMs from [145].

### 3.8.3 Evaluation of total binding capacity of enhancers

Total binding capacity is a measure of the cumulative ability of an enhancer to bind a TF, and thus it takes into account the binding affinity of every $w$-mer in the enhancer for a TF binding site of length $w$ [172]. To calculate the total binding capacity, we start by computationally scoring each possible site in the enhancer for the motifs of TFs regulating early axis specification. Taking the exponential of the score, normalizing this exponential by the enhancer length $l$, and summing these values gives us an overall binding capacity for each enhancer and TF combination, which is roughly equal to the sum of the probabilities that a TF is bound to each potential site in the enhancer. Hence, we use the following formula

$$c(s,z) = \sum_{i=1}^{l-w+1} \frac{e^{\sum_{j=1}^{w} \ln \frac{p_j(b(j))}{q(b(j))}}}{l}$$

to calculate the total binding capacity $c$ of a given sequence $s$ for a given TF $z$ [172]. Here, $l$ is the length of the sequence being considered, $w$ is the width of the PWM of the TF, $b(i)$ is the base at position $i$ of the sequence, $p_j(b)$ is the frequency of seeing base $b$ at position $j$ of the PWM, and $q(b)$ is the background frequency of base $b$. Note that $\sum_{j=1}^{w} \ln \frac{p_j(b(j))}{q(b(j))}$ is equivalent to the score given to the $w$-mer at position $i$ in the sequence calculated using Patser, as described above [57].

### 3.8.4 Selection of enhancers to study

*knirps* enhancers expressed in the blastoderm were identified using REDfly [51], and the shortest, non-overlapping subset of enhancers was obtained using SelectSmallestFeature.py

available at the Halfon Lab GitHub `https://github.com/HalfonLab/UtilityPrograms`. The enhancers in this subset were categorized by the expression patterns they drove, and a representative enhancer was picked from each of these categories.

### 3.8.5 Generation of transgenic reporter fly lines

As described in Fukaya, et al., the four *kni* enhancers were each cloned into the pBphi vector, directly upstream of *kni* promoter 1, 2 or 2ΔTATAΔDPE; 24 MS2 repeats; and a yellow reporter gene [44]. Similarly, the *kni* locus and its promoter knockouts (Δp1 and Δp2) were each cloned into the pBphi vector, directly upstream of 24 MS2 repeats and a yellow reporter gene by Applied Biological Materials (Richmond, BC, Canada). We defined kni_-5 as chr3L:20699503-20700905(-), kni_proximal_minimal as chr3L:20694587-20695245(-), kni_KD as chr3L:20696543-20697412(-), VT33935 as chr3L:20697271-20699384(-), promoter 1 as chr3L:20695324-20695479(-), promoter 2 as chr3L:20694506-20694631(-), and the *kni* locus as chr3L:20693955-20701078(), using the *Drosophila melanogaster* dm6 release coordinates. Promoter motif knockouts (for p2ΔTATAΔDPE and locus Δp2) involved making the minimal number of mutations that would both inactivate the motif and introduce the fewest new motifs or TF binding sites (TATA: TATATATATC > TAGATGTATC, Inr: TCAGTT > TCGGTT, and DPE: AGATCA > ATACCA). The locus Δp1 construct involved replacing promoter 1 with a region of the lambda genome predicted to have the minimal number of relevant TF binding sites. The precise sequences for each reporter construct are given in a series of GenBank files included in the Supplementary Materials (File S3 – 18).

Using phiC31-mediated integration, each reporter construct was integrated into the same site on chr2L by injection into yw; PBac{y[+]-attP-3B}VK00002 (BDRC stock # 9723) embryos by BestGene Inc (Chino Hills, CA). To visualize MS2 expression, female flies expressing RFP-tagged histones and GFP-tagged MCP (yw; His-RFP/Cyo; MCP-GFP/TM3.Sb) were

crossed with males containing one of the MS2 reporter constructs.

### 3.8.6 Sample preparation and image acquisition

As in Garcia et al., live embryos were collected prior to nuclear cycle 14 (nc14), dechorionated, mounted with glue on a permeable membrane, immersed in Halocarbon 27 oil, and put under a glass coverslip [45]. Individual embryos were then imaged on a Nikon A1R point scanning confocal microscope using a 60X/1.4 N.A. oil immersion objective and laser settings of 40uW for 488nm and 35$mu$W for 561nm. To track transcription, 21 slice Z-stacks, at 0.5 um steps, were taken throughout nc14 at roughly 30sec intervals. To identify the Z-stack's position in the embryo, the whole embryo was imaged at the end of nc14 at 20X using the same laser power settings. To quantify expression along the AP axis, each transcriptional spot's location was placed in 2.5% anterior-posterior (AP) bins across the length of the embryo, with the first bin at the anterior of the embryo. Embryos were imaged at ambient temperature, which was on average 26.5°C.

### 3.8.7 Burst calling and calculation of transcription parameters

Tracking of nuclei and transcriptional puncta was done using a version of the image analysis MATLAB pipeline downloaded from the Garcia lab GitHub repository on January 8, 2020 and described in Garcia et al [45]. For every spot of transcription imaged, background fluorescence at each time point is estimated as the offset of fitting the 2D maximum projection of the Z-stack image centered around the transcriptional spot to a gaussian curve, using MATLAB *lsqnonlin*. This background estimate is subtracted from the raw spot fluorescence intensity. The resulting fluorescence traces across nc14 are then smoothed by the LOWESS method with a span of 10%. These smoothed traces are then used to quantify transcriptional properties and noise. Traces consisting of fewer than three timeframes are not included in

the calculations.

To quantify the transcription properties of interest, we used the smoothed traces to determine at which time points the promoter was "on" or "off" [166]. A promoter was considered "on" if the slope of its trace, i.e. the change in fluorescence, between one point and the next was greater than or equal to the instantaneous fluorescence value calculated for one mRNA molecule (FRNAP, described below). Once called "on", the promoter is considered active until the slope of the fluorescence trace becomes less than or equal to the negative instantaneous fluorescence value of one mRNA molecule, at which point it is considered inactive until the next time point it is called "on". The instantaneous fluorescence of a single mRNA was chosen as the threshold because we reasoned that an increase in fluorescence greater than or equal to that of a single transcript is indicative of an actively producing promoter, just as a decrease in fluorescence greater than that associated with a single transcript indicates that transcripts are primarily dissociating from, not being newly initiated at, this locus. Visual inspection of fluorescence traces agreed well with the burst calling produced by this method (**Figure S3.4**) [166].

Using these smoothed traces and "on" and "off" time points of promoters, we measured burst size, burst frequency, burst duration, polymerase initiation rate, and noise. Burst size is defined as the integrated area under the curve of each transcriptional burst, from one "on" frame to the next "on" frame, with the value of 0 set to the floor of the background-subtracted fluorescence trace (**Figure S3.4C**). Frequency is defined as the number of bursts in nc14 divided by time between the first time the promoter is called active and 50 min into nc14 or the movie ends, whichever is first (**Figure S3.4E**). The time of first activity was used for frequency calculations because the different enhancer constructs showed different characteristic times to first transcriptional burst during nc14. Duration is defined as the amount of time occurring between the frame a promoter is considered "on" and the frame it is next considered "off" (**Figure S3.4F**). Polymerase initiation rate is defined as the

slope at the midpoint between the frame a promoter is considered "on" and the frame it is next considered "off" (**Figure S3.4G**). The temporal coefficient of variation of each transcriptional spot i, was calculated using the formula:

$$CV(i) = \frac{\text{standard deviation}(m_i(t))}{\text{mean}(m_i(t))}$$

where $m_i(t)$ is the fluorescence of spot $i$ at time $t$. For these, and all other measurements, we control for the embryo position of the fluorescence trace by first individually analyzing the trace and then using all the traces in each AP bin (anterior-posterior; the embryo is divided into 41 bins each containing 2.5% of the embryo's length) to calculate summary statistics of the transcriptional dynamics and noise values at that AP position.

All original MATLAB code used for burst calling, noise measurements, and other image processing are available at the Wunderlich Lab GitHub [166] with a copy archived at `https://github.com/elifesciences-publications/KrShadowEnhancerCode`. Updates to include calculations of polymerase initiation rate are also available at the Wunderlich Lab GitHub (`https://github.com/WunderlichLab`).

### 3.8.8 Conversion of integrated fluorescence to mRNA molecules

To convert arbitrary fluorescence units into physiologically relevant units, we calibrated our fluorescence measurements in terms of mRNA molecules. As in Lammers et al., for our microscope, we determined a calibration factor, , between our MS2 signal integrated over nc13, FMS2, and the number of mRNAs generated by a single allele from the same reporter construct in the same time interval, NFISH, using the *hunchback* P2 enhancer reporter construct [45, 85]. Using this conversion factor, we calculated the integrated fluorescence of a single mRNA (F1) as well as the instantaneous fluorescence of an mRNA molecule (FRNAP). For our microscope, FRNAP is 379 AU/RNAP, and F1 is 1338 AU/RNAPmin.

We can use this values to convert both integrated and instantaneous fluorescence into total mRNAs produced and number of nascent mRNAs present at a single time point, by dividing by F1 and FRNAP, respectively.

### 3.8.9 Regression modeling and statistical analysis

To quantify the effect of enhancer, promoter, and interaction terms on burst parameters, we considered models of the form

$$g(Y) = \text{enhancer} + \text{promoter} + (\text{enhancer} \times \text{promoter})$$

where $Y$ is the burst property of interest and $g$ is the link function (**Figure 3.4A**). Model selection involved considering (1) the type of model, (2) the distribution that best fit the burst property data and (3) the appropriate predictors to include. We approached model selection with no specific expectations, opting to use generalized linear models (GLMs) because they were not much improved upon by adding random effects (GLMMs) and because they fit the data better than linear models (LMs).

Similarly, the appropriate distribution for each burst property was determined by fitting various distributions to the data and comparing their goodness-of-fit. As expected, total RNA produced and burst size (in transcripts per burst) were best described by a negative binomial distribution, as has been commonly used to describe count data. For the other burst properties, for which the appropriate distribution was less clear, we found that burst frequency was best fit by the Weibull distribution and burst duration and initiation rate were best fit by the gamma distribution. These choices were supported by the lower AIC values produced when comparing them to models using alternative distributions. They also seem reasonable given examples of other applications of these distributions. To keep the

interpretation consistent across models, we chose to use an identity link function for all models (**Figure 3.4B**); using the canonical link functions associated with each of these distributions produced the same trends (**Figure S3.5**).

The predictors we included were the enhancer and promoter and any interaction terms between the enhancer and promoter. In each case, dropping the interaction terms produced higher AIC values, suggesting that the interaction terms are important and should not be dropped by the model.

### 3.8.10 Data Availability Statement

Transgenic fly strains and plasmids are available upon request. Supplementary File S3.1 contains the gene names, the dm6 release coordinates, and the FlyBase numbers (FBgns) that matched to the gene names and coordinates [23]. File S3.2 contains DNA sequences of the enhancers and promoters used in the computational analysis presented in **Figure S3.2**. Files S3.3 – 3.18 contain GenBank files describing the plasmids used to make all the transgenic fly strains produced for this work.

## 3.9 Figures

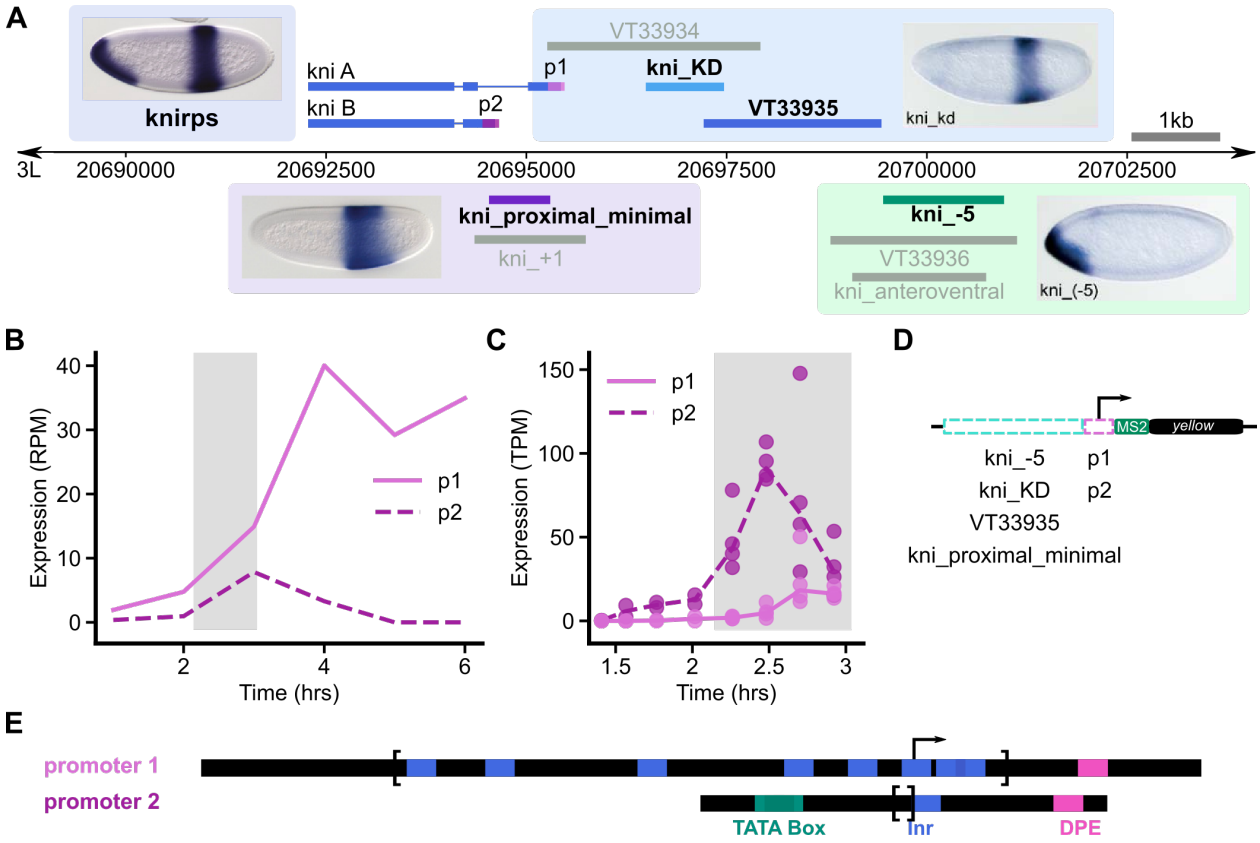


**Figure 3.1:** *knirps* as a case study. The *knirps* (*kni*) locus was chosen to study how the motif content of endogenous enhancers and promoters affects transcription dynamics. This locus was selected because it comprises multiple enhancers that bind different TFs and multiple core promoters that contain different promoter motifs. These enhancers and promoters are all active during the blastoderm stage. **(A)** The *kni* locus comprises multiple enhancers that together drive expression of a ventral, anterior band and a posterior stripe, as shown in the in situ at the top left. Enhancers that drive similar expression patterns have been displayed together in boxes with a representative in situ hybridization [125, 139]. The four enhancers selected for study are in color and labeled in bold text; the others are in gray. *kni* also has two promoters represented in two shades of purple, which drive slightly different transcripts (differing by only five amino acids). Expression data for the two *kni* promoters is shown, with RAMPAGE data [6] in **(B)** and RNA-seq data [102] in **(C)**; the time period corresponding to the blastoderm stage is highlighted in gray. Based on these two sets of data, the two *kni* promoters are both used during nuclear cycle 14 though which one is more active is less clear. Note that for the rest of development, promoter 1 is the more active one. **(D)** A total of eight MS2 reporter constructs containing pairs of each of the four enhancers matched with each of the two *kni* promoters were made. **(E)** The two *kni* promoters are shown here in black, consisting of the RAMPAGE-defined transcription start clusters (TSCs)between the brackets and an additional ±40bp from the TSCs. The two *kni* promoters can be

**Figure 3.1 (cont.):** distinguished by their motif content (with promoter 1 consisting of a series of Inr motifs and a DPE motif and promoter 2 consisting of an Inr, two overlapping TATA Boxes and a DPE motif). They also differ in the "sharpness" of their region of transcription initiation (shown between the brackets), with promoter 1 (125bp) being significantly broader than promoter 2 (4bp) based on RAMPAGE tag data [6].

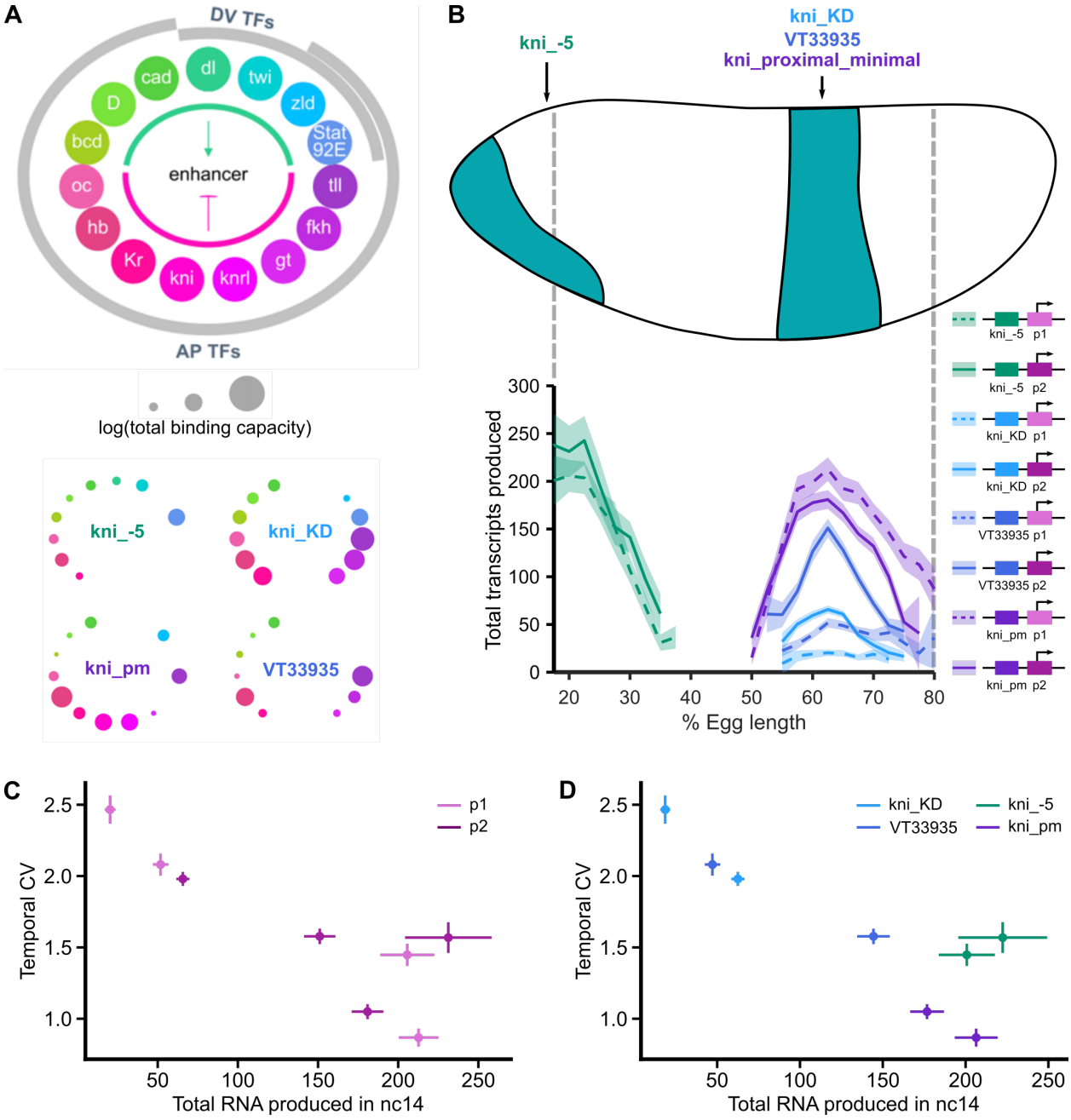


**Figure 3.2:** The *kni* enhancers differ in their capacity to bind different transcription factors and drive transcription with different promoters. The enhancers can be separated into two classes—those that produce high expression with either promoter (kni_-5 and kni_proximal_minimal) and those that produce much higher expression with promoter 2 (kni_KD and VT33935). Note that for simplicity, kni_proximal_minimal has been shortened to kni_pm in the figures. **(A)** Here ability of the *kni* enhancers to bind early axis-patterning TFs is quantified and represented visually. The logarithm of the predicted TF binding capacity of each of the *kni* enhancers is plotted as circles around the enhancer, with the color indicating the TF and the circle size increasing with higher binding capacity. The TFs are categorized by their role in regulating anterior-posterior (AP) or dorsal-ventral (DV) patterning and broadly by their roles as activators (indicated by the green arc) and repressors (indicated by the pink arc). Note that kni_KD and VT33935, which drive the

**Figure 3.2 (cont.):** same posterior stripe of expression, share very similar TFs and that kni_-5, the only enhancer with a DV component, is the only one bound by DV TFs. kni_proximal_minimal drives a similar expression pattern to kni_KD and VT33935, but notably has different predicted TF binding capacities. **(B)** The *Drosophila* embryo with the *kni* expression pattern at nuclear cycle 14 is shown; kni_-5 drives the expression of the anterior, ventral band, while the other three enhancers drive the expression of the posterior stripe. We made enhancer-promoter reporters containing each of the four enhancers matched with either promoter 1 or 2. Using measurements from these enhancer-promoter reporters (shown at the right), the total RNA produced by each construct during nuclear cycle 14 is plotted against position along the embryo length (AP axis). The error bands around the lines are 95% confidence intervals. The constructs containing promoter 1 are denoted with a dashed line and those containing promoter 2 with a solid line. Some, but not all, enhancers show a strong promoter preference. kni_KD and VT33935, which are bound by similar TFs, drive 2.9-fold and 3.4-fold higher expression with promoter 2 at 62.5% embryo length, respectively (one-sided $t$-test $p < 2.2\times10^{-16}$ for both), whereas, kni_-5 and kni_proximal_minimal show similar expression regardless of promoter with the largest difference only 1.2-fold at the anterior-posterior bin of maximum expression (22% and 63%, respectively) (two-sided $t$-test comparing kni_-5-promoter1 vs. kni_-5-promoter2, $p = 0.12$ and kni_proximal_minimal-promoter1 vs. kni_proximal_minimal-promoter2 $p = 9.8 \times 10^{-5}$). In panels **(C − D)**, the temporal coefficient of variation (CV) is plotted against the total RNA produced in nc14 at the anterior-posterior bin of maximum expression (22% and 63%) for the anterior band and the posterior stripe, respectively, with the error bars representing 95% confidence intervals. There is a general trend of mean expression levels being anti-correlated with CV, or noise. **(C)** Here, the data points are colored by the construct's promoter, with promoter 1 in light purple and promoter 2 in purple. Despite the general trend, there are cases when the same promoter (promoter 2) shows higher CV and total expression when paired with different enhancers (kni_proximal_minimal vs kni_-5). **(D)** Here, the data points are colored by the construct's enhancer. Again, despite the general trend, there are cases when the same enhancer (kni_-5) shows higher CV and total expression when paired with different promoters (promoter 1 vs 2).

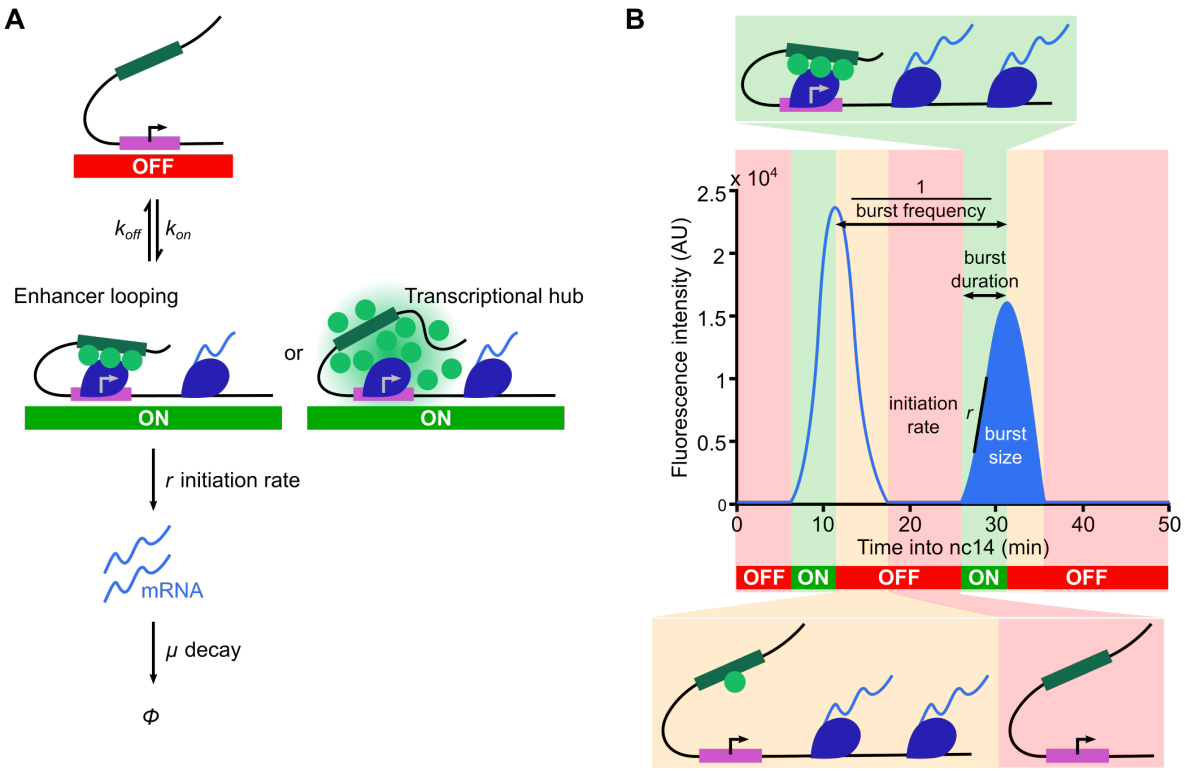


**Figure 3.3:** Two-state model of transcription in the context of tracking transcription dynamics. **(A)** Here, we represent the two-state model of transcription, in which the promoter is either (1) in the inactive state (OFF), in which RNA polymerase cannot bind and initiate transcription or (2) in the active state (ON), during which it can. The promoter transitions between these two states with rates $k_{\text{on}}$ and $k_{\text{off}}$, with promoter activation involving both the interaction of the enhancer and promoter and the assembly of all the necessary transcription machinery for transcription initiation to occur. This may occur through enhancer looping or through the formation of a transcriptional hub. In its active state, the promoter produces mRNA at rate $r$, and the mRNA decays by diffusing away from the gene locus at rate $\mu$. **(B)** MS2-tagging RNA allows us to track nascent transcription, and the resulting fluorescence trace (in light blue) is proportional to the number of nascent RNA produced over time. The graph is split into sections, representing different molecular states and how they correspond to fluctuating transcription over time. These states are represented by different colors—red when the promoter is OFF, green when it is ON, and yellow when transcription continues but the promoter is no longer ON, as no new polymerases are being loaded. The dynamics of these fluctuations or bursts can be characterized by quantifying various properties, including burst frequency (how often a burst a occurs), burst size (number of RNA produced per burst), and burst duration (the period of active transcription during which mRNA is produced at rate $r$).

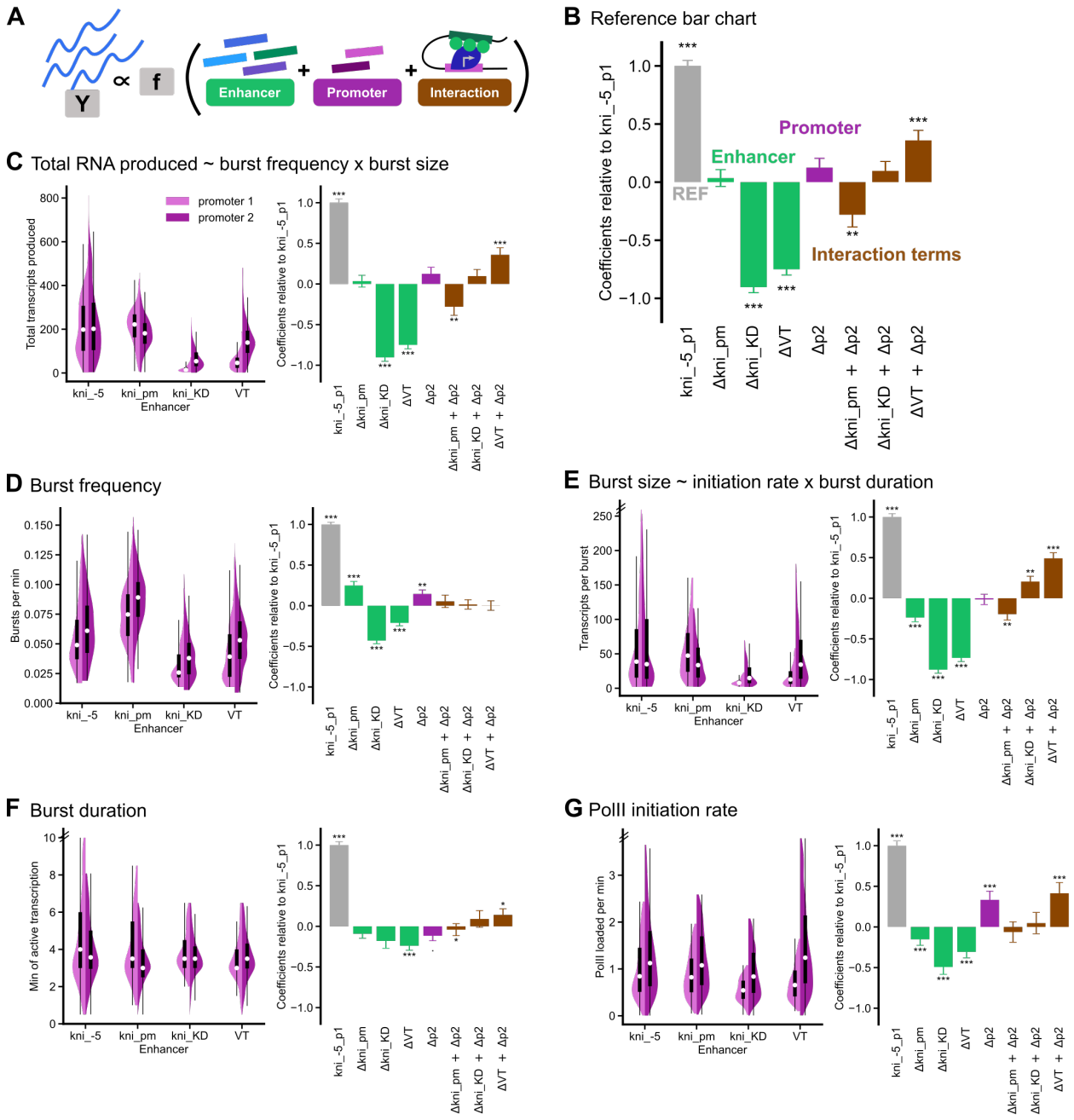


**Figure 3.4: Expression levels are mainly determined by burst frequency and initiation rate.** **(A)** To parse the effects of the enhancer, the promoter, and their interactions on all burst properties, we built generalized linear models (GLMs). $Y$ represents the burst property under study, $g$ is the identity link function, and the enhancers, promoters, and their interaction terms are the explanatory variables. The coefficients of each of these explanatory variables is representative of that variable's contribution to the total value of the burst property. **(B)** All burst property data was taken from the anterior-posterior bin of maximum expression (22% and 63%) for the anterior band and the posterior stripe, respectively. The coefficients and the 95% confidence intervals for each independent variable relative to that of a reference construct (kni_-5-p1) are plotted as a bar graph; * $p < 0.05$, ** $p < 0.01$,

**Figure 3.4 (cont.):** *** $p < 0.001$. The reference construct is represented in gray, and the effects of enhancer, promoter, and their interactions are represented in green, purple, and brown, respectively. Summing the relevant coefficients gives you the average value of the burst property for a particular construct relative to the reference construct. Thus, as the reference construct, kni_-5-p1 coefficient will always be 1. The average value of the burst property for a particular construct, e.g. VT-p2, relative to the reference construct, would be 0.75, which is the sum of the reference bar = 1, $\Delta$VT = -0.78, $\Delta$p2 = 0.17, and $\Delta$VT + $\Delta$p2 = 0.36. Note that for simplicity, kni_proximal_minimal and VT33935 has been shortened to kni_pm and VT, respectively, in the following graphs. In panels **(C – G)**, (left) split violin plots (and their associated box plots) of burst properties for all eight constructs will be plotted with promoter 1 in light purple and promoter 2 in purple. The black boxes span the lower to upper quartiles, with the white dot within the box indicating the median. Whiskers extend to 1.5*IQR (interquartile range) ś the upper and lower quartile, respectively. (right) Bar graphs representing the relative contributions of enhancer, promoter, and their interactions to each burst property are plotted as described in **(B)**. The double hash marks on the axes indicate that 90% of the data is being shown. **(C)** Expression levels are mainly determined by the enhancer and the interaction terms. Some enhancers (kni_-5 and kni_proximal_minimal) appear to work well with both promoters; whereas, kni_KD and VT, which are bound by similar TFs, show much higher expression with promoter 2. **(D)** Burst frequency is dominated by the enhancer and promoter terms, with promoter 2 consistently producing higher burst frequencies regardless of enhancer. **(E)** Burst size, which is determined by both initiation rate and burst duration, is dominated by the enhancer and interaction terms, with interaction terms representing the role of molecular compatibility. As **(F)** burst duration is reasonably consistent regardless of enhancer or promoter, differences in burst size are mainly dependent on differences in **(G)** PolII initiation rate.

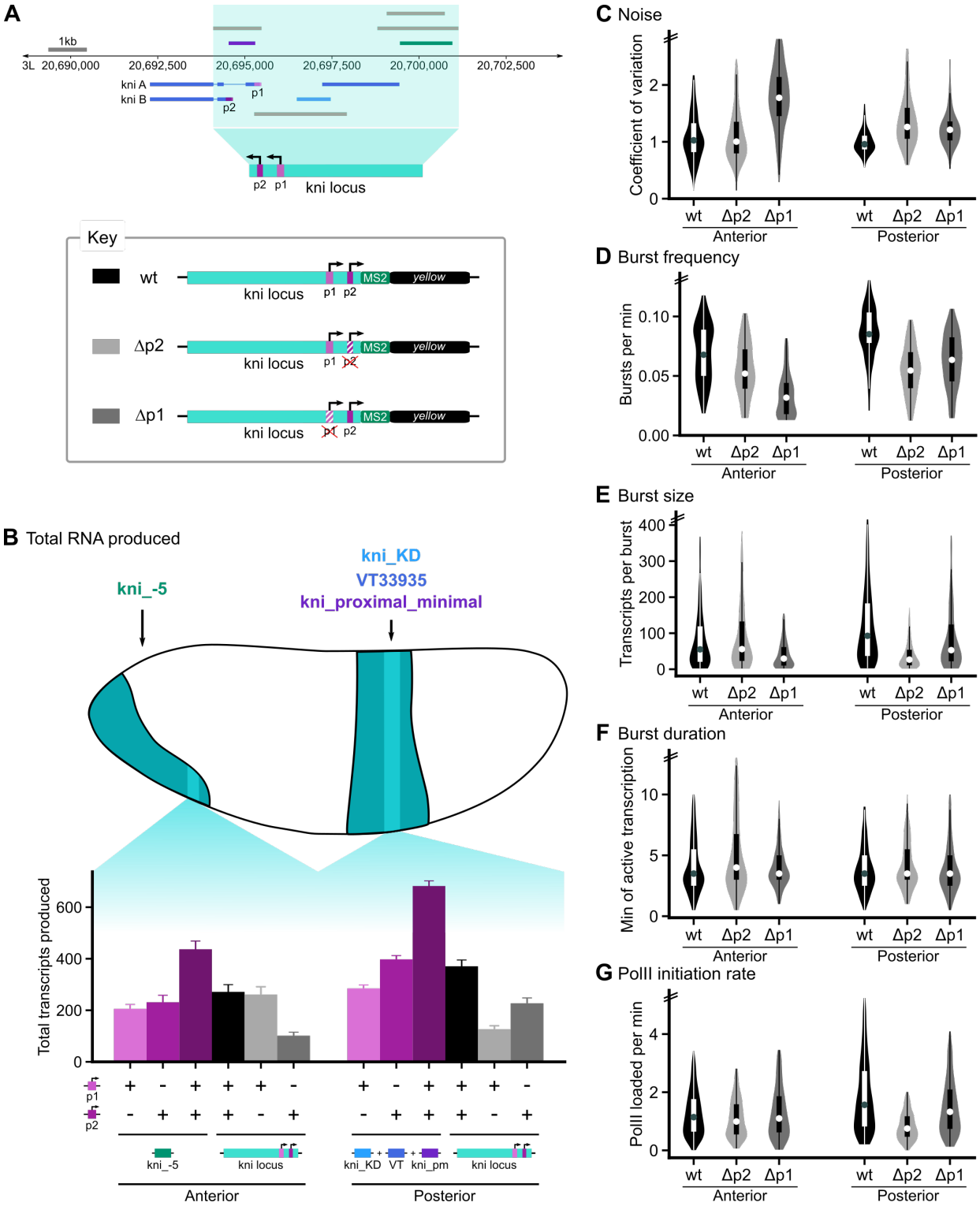


**Figure 3.5: The synthetic enhancer-promoter constructs are insufficient to capture the behavior of the *knirps* promoters within the endogenous locus. (A)** We cloned the entire *kni* locus into an MS2 reporter construct and measured the expression levels and dynamics of the wildtype (wt) locus reporter, and reporters with either promoter 1 or 2 knocked out (Δp1 and Δp2). To make the Δp1 reporter, we replaced promoter 1 with

**Figure 3.5 (cont.):** a piece of lambda phage DNA, due to the large number of Inr motifs. To make the Δp2 construct, we removed the TATA, Inr, and DPE motifs by making several mutations (see Methods for additional details). In panels **(B − G)**, all burst property data was taken from the anterior-posterior bin of maximum expression (22% and 63%) for the anterior band and the posterior stripe, respectively. **(B)** The *Drosophila* embryo with the *kni* expression pattern at nuclear cycle 14 is shown; kni_-5 drives the expression of the anterior, ventral band, while the other three enhancers drive the expression of the posterior stripe. The bin of maximum expression is highlighted in light teal. To compare the expression produced by the synthetic enhancer-promoter reporters with the locus reporters, we plotted bar graphs of the summed total RNA produced at the location of maximum expression in the anterior (left) and posterior (right) for six cases—just enhancer-promoter1 reporters (light purple), just enhancer-promoter2 reporters (purple), both enhancer-promoter1 and -promoter2 reporters (dark purple), the wt locus reporter (black), the locus Δp2 reporter (light gray), and the locus Δp1 reporter (dark gray). In panels **(C − F)** violin plots (and their associated box plots) of burst properties for all three reporters are plotted with the wt, Δp1, and Δp2 reporters in black, light gray, and dark gray, respectively. The internal boxes span the lower to upper quartiles, with the dot within the box indicating the median. Whiskers extend to 1.5*IQR (interquartile range) ś the upper and lower quartile, respectively. The double hash marks on the axes indicate that 95% of the data is being shown. **(C)** The coefficient of variation is inversely correlated with total RNA produced shown in **(B)**. In the anterior, the Δp2 reporter, which produces the same total RNA as the wt reporter, also produces the same amount of noise. **(D)** In the anterior of the embryo, burst frequency of the Δp2 reporter is less than the wt reporter even though they produce the same expression levels and noise. In the posterior, knocking out promoter 2 has a larger impact on burst frequency than knocking out promoter 1. **(E)** In both the anterior and posterior, burst size is directly correlated with total RNA produced. Note that in the posterior of the embryo, knocking out promoter 2 has a much larger impact on burst size than knocking out promoter 1. Burst size is dependent on PolII initiation rate and burst duration. While **(F)** burst duration is reasonably consistent regardless of promoter knockout, **(G)** PolII initiation rate is directly correlated with burst size. This suggests that differences in burst size are mainly mediated by differences in PolII initiation rate.

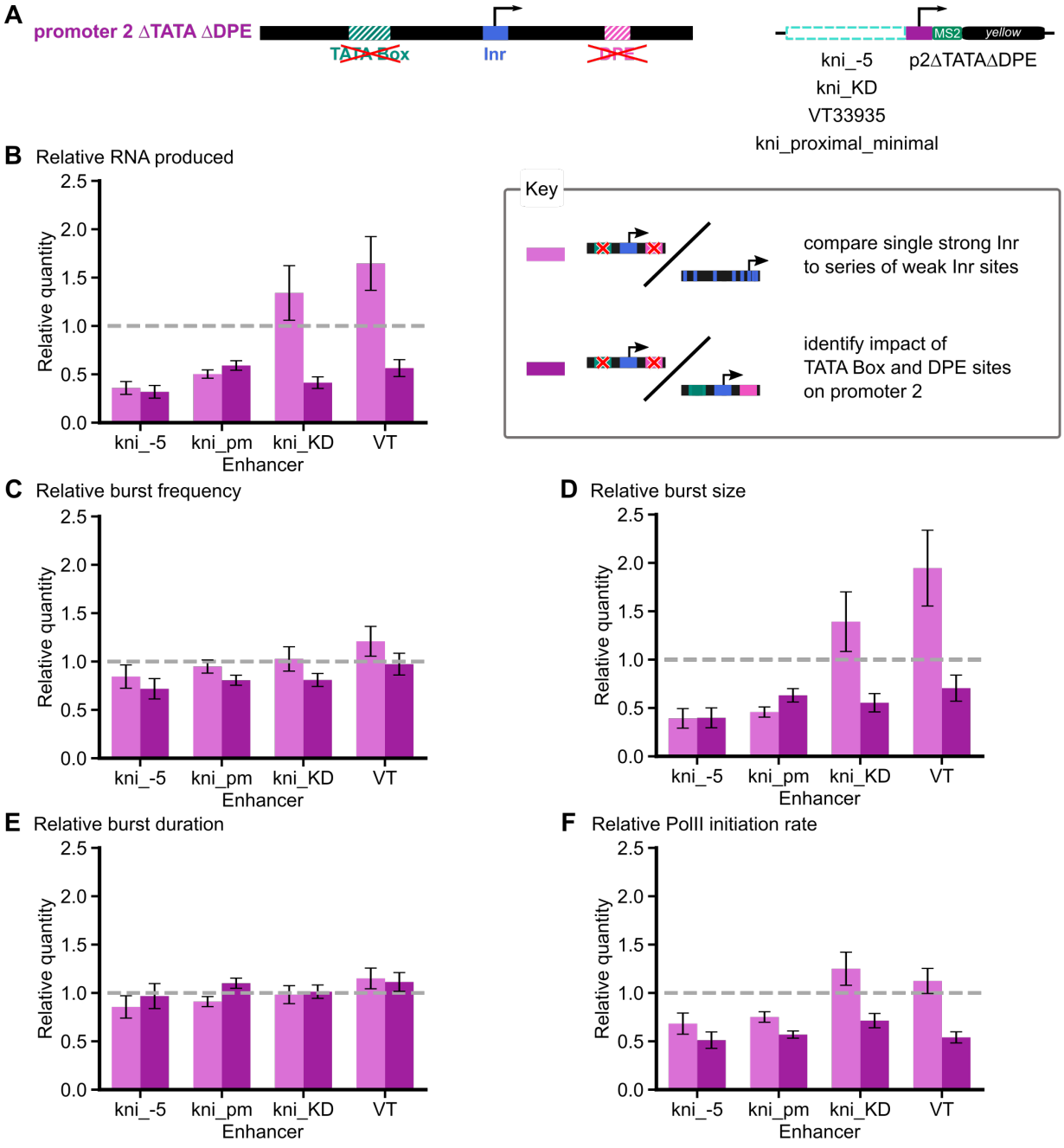


**Figure 3.6:** PolII initiation rate is a key burst property that is tuned by promoter motif. **(A)** We made enhancer-promoter reporters containing each of the four enhancers matched with a mutated promoter 2 (p2ΔTATAΔDPE) in which the TATA Box and DPE motifs have been eliminated by making several mutations (see Methods for details). In panels **(B – F)**, bar graphs of the burst properties produced by p2ΔTATAΔDPE relative to promoter 1 (in light purple) and to promoter 2 (in purple) are shown. By comparing p2ΔTATAΔDPE with promoter 1, we can determine whether a single, strong Inr site (mutated promoter 2) can perform similarly to a series of weak Inr sites (promoter 1), and by comparing p2ΔTATAΔDPE with promoter 2, we can clarify the role of TATA Box and DPE motifs in tuning burst

**Figure 3.6 (cont.):** properties. The error bars show the 95% confidence intervals. The gray dashed line at 1 acts as a reference—if there is no difference between the burst properties produced by p2ΔTATAΔDPE and either promoter 1 or 2, the bar should reach this line. All burst property data was taken from the anterior-posterior bin of maximum expression (22% and 63%) for the anterior band and the posterior stripe, respectively. Note that for simplicity, kni_proximal_minimal and VT33935 have been shortened to kni_pm and VT, respectively, in the following graphs. **(B)** When comparing p2ΔTATAΔDPE with promoter 1, we can see that the enhancers fall into two classes—those that drive less expression or more expression with a single strong Inr site than with a series of weak Inr sites. The enhancers (kni_-5 and kni_proximal_minimal) that drive less expression are the same ones that were similarly compatible with both promoters 1 and 2, whereas the enhancers that drive more expression (kni_KD and VT33935) are the ones that strongly preferred promoter 2. When comparing p2ΔTATAΔDPE with promoter 2, we see that eliminating TATA Box and DPE motifs reduces expression output for all enhancers. **(C)** When comparing p2ΔTATAΔDPE with either promoter 1 or promoter 2, we see that burst frequency is not substantially affected though, compared to promoter 2, there is a moderate decrease upon motif disruption. **(D)** When comparing the burst size of p2ΔTATAΔDPE reporters with either that of promoter 1 or promoter 2 reporters, we see the same behavior as with total RNA (shown in panel **(B)**). This suggests that burst size is the main mediator of the increase or decrease in total RNA produced. Burst size is dependent on PolII initiation rate and burst duration. As **(E)** burst duration is reasonably consistent regardless of promoter, it appears that **(F)** changes in burst size are mainly mediated by tuning PolII initiation rate. Together, this suggests that enhancers fall into two classes, based on their response to different promoters; however, regardless of class, PolII initiation rate is what underlies differences in expression output.

## 3.10 Supplementary Figures

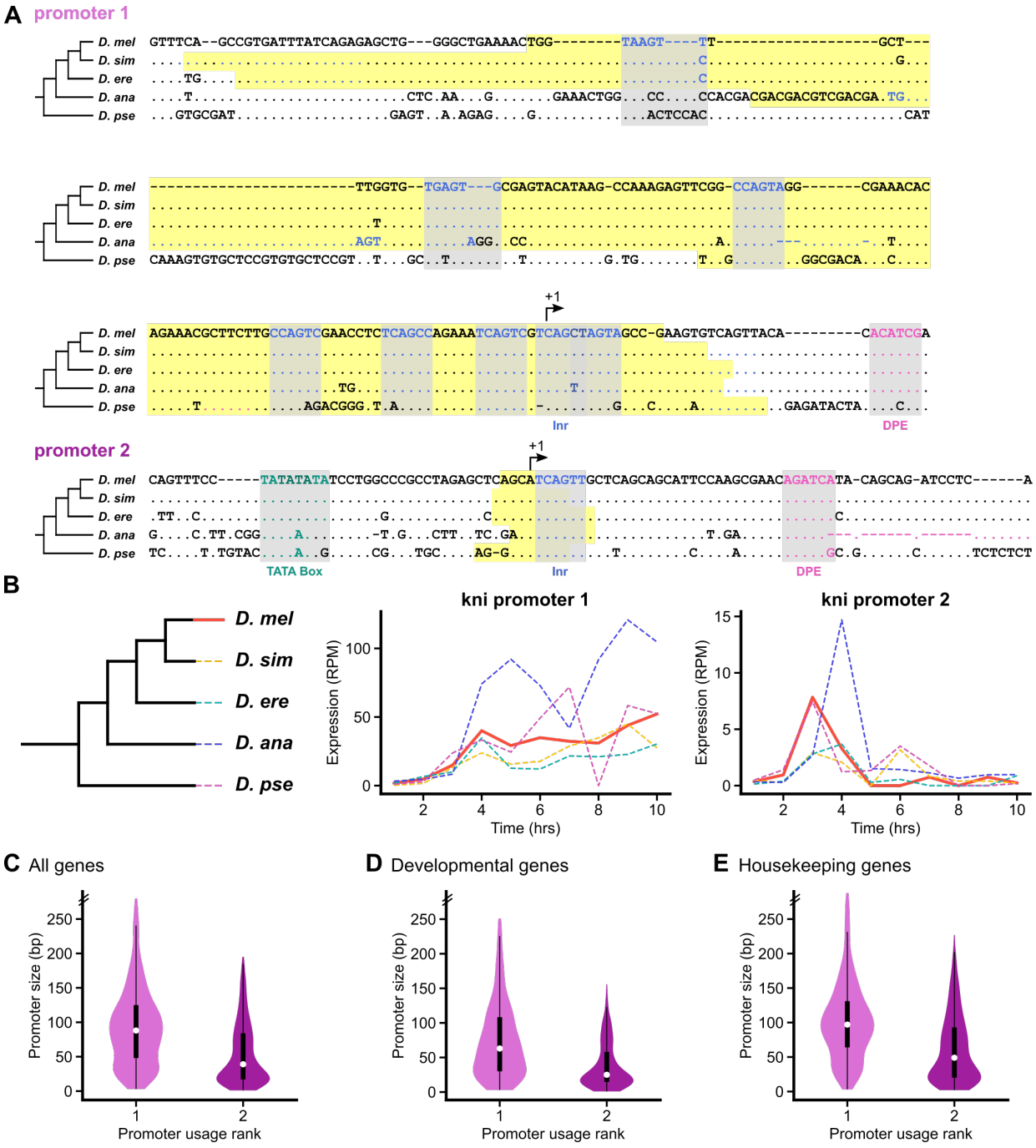


**Figure S3.1:** The *knirps* promoters show sequence and functional conservation, and this two-promoter structure is prevalent among genes expressed during development. **(A)** Both *kni* promoters are aligned with the orthologous sequences in four other *Drosophila* species, with dashes (-) representing unaligned sequence and dots (.) indicating matching base pairs. There is remarkable sequence conservation, with the core promoter motifs preserved across

**Figure S3.1 (cont.):** all five species. The highlighted regions represent transcription start clusters (TSCs), identified by Batut, et al [6] as regions of statistically significant clustering of cDNA 5' ends. **(B)** *kni* promoter activity over the first 10 hours of development is reasonably consistent across five species of *Drosophila*, with promoter 1 generally being used more than promoter 2. Specifically, note that both promoters are used in nuclear cycle 14 (2 – 3 hours) in all five species. **(C – E)** For developmentally expressed genes with multiple promoters that are represented in both the Eukaryotic Promoter Database and the Batut et al. RAMPAGE data [6, 32], violin plots of the two most used promoters, with the primary promoter (most used) in light purple and the secondary promoter (second most used) in dark purple. The black boxes span the lower to upper quartiles, with the white dot within the box indicating the median. Whiskers extend to 1.5*IQR (interquartile range) ś the upper and lower quartile, respectively. The double hash marks on the axes indicate that 95% of the data is being shown. **(C)** When the two most used promoters of genes expressed in embryogenesis ($n = 1177$) are plotted, the size of primary promoters is significantly larger than that of the secondary promoter. **(D)** When limited to promoters of developmentally controlled genes – genes whose expression pattern varies considerably as a function of developmental time — ($n = 387$) this trend of larger primary promoters is maintained, though on average, these promoters are sharper that those of the whole gene set in panel C. **(E)** When limited to promoters of housekeeping genes ($n = 790$), this trend of larger primary than secondary promoters is also maintained, though on average, these promoters are still broader than those of developmentally controlled genes.

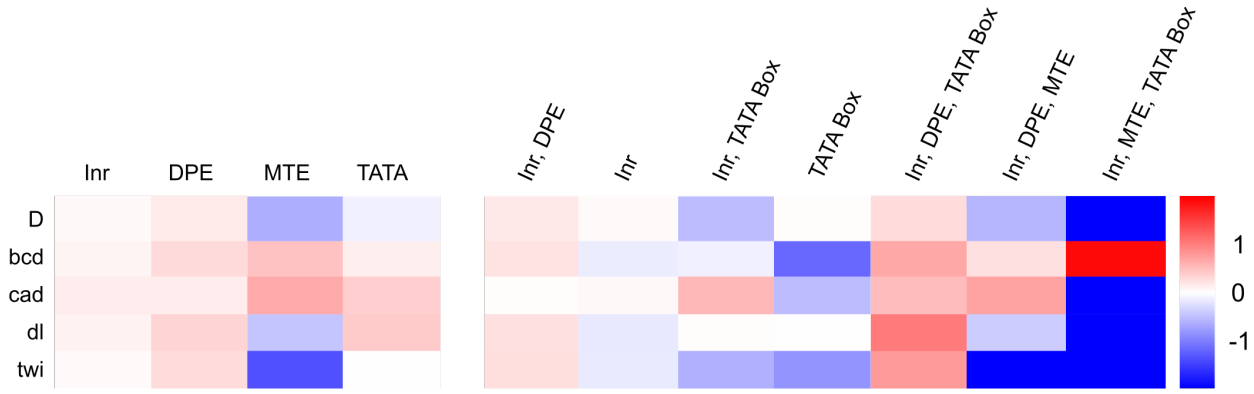


**Figure S3.2:** TFs show preferences for certain core promoter motifs. To identify patterns of TF-core promoter motif co-occurrence, we calculated the fold enrichment of core promoter elements associated with TF-target genes. The left heatmap shows the log fold-enrichment over background of the frequency of the core promoter motif (columns) for the set of promoters associated with enhancers controlled by the TF (rows). The right heatmap shows the log fold-enrichment over background of the frequency of the motif combination (columns) for the set of promoters associated with enhancers controlled by the TF (rows). For example, this means that column 1 (Inr) in the left heatmap shows enrichment of any promoters that contain Inr regardless of any other promoter motifs they might contain, whereas column 2 (Inr) in the right heatmap shows enrichment of promoters with only Inr and no other core promoter motifs.

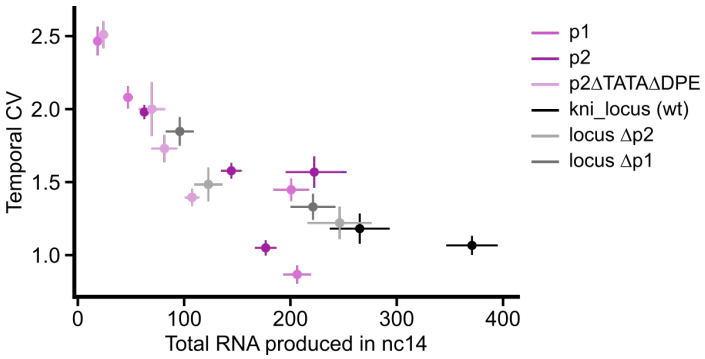


**Figure S3.3:** Noise is inversely correlated with total RNA produced. To examine the relationship between temporal coefficient of variation (CV) and activity of each construct, we plotted the mean temporal CV against the total RNA produced in nc14 at the anterior-posterior bin of maximum expression (22% and 63%) for the anterior band and the posterior stripe, respectively, with the error bars representing 95% confidence intervals. There is a clear trend of CV decreasing with increased total RNA produced though there are examples where constructs with the same promoter can produce higher noise than others with similar output levels, suggesting that promoters do not solely dictate noise levels.

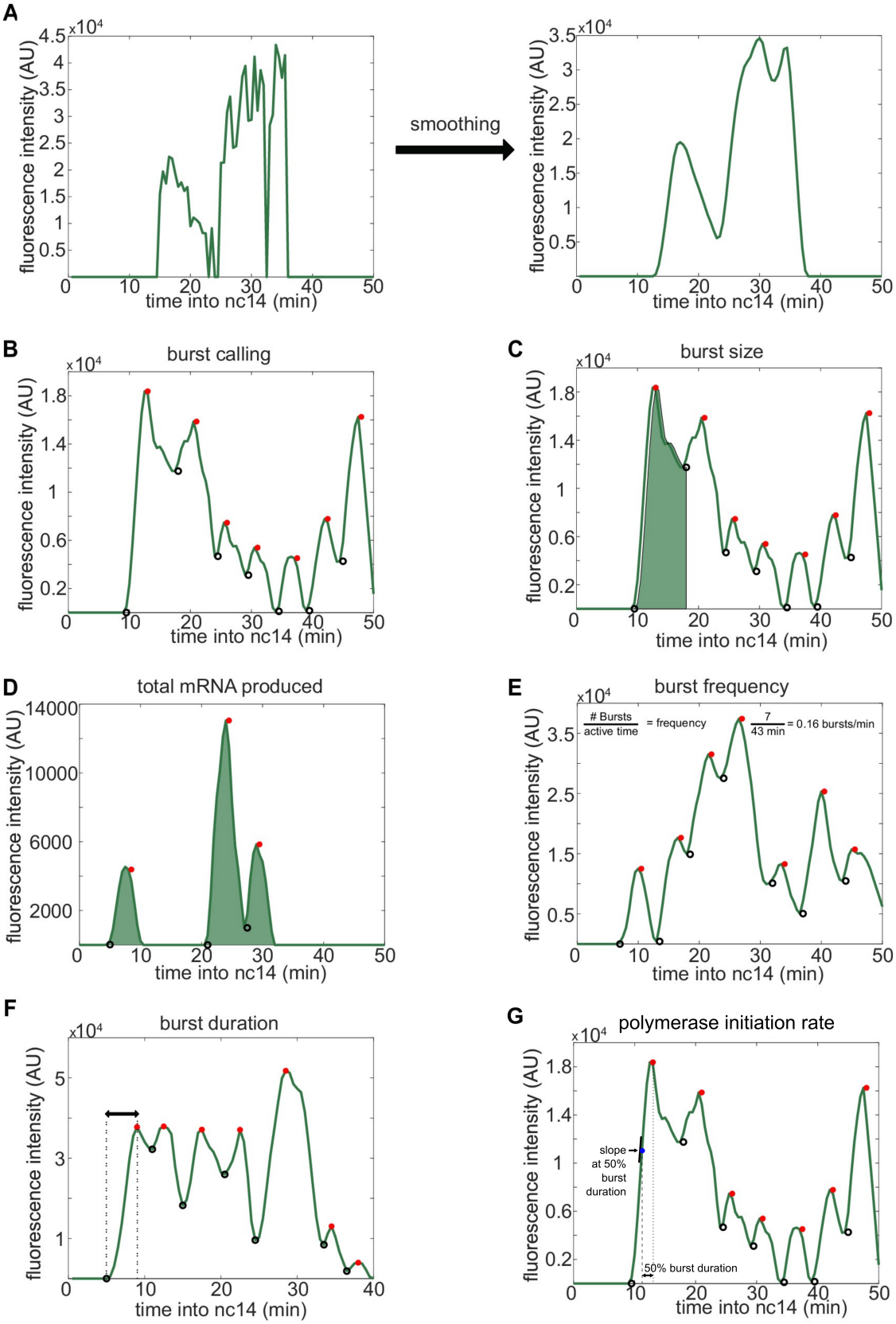


**Figure S3.4:** Visual inspection of burst calling algorithm. This figure is adapted from Waymack, et al. with one additional panel **(G)** added [166]. To quantify the burst properties

**Figure S3.4 (cont.):** of interest (burst size, burst frequency, burst duration, and polymerase initiation rate), we began by smoothing individual fluorescence traces using the LOWESS method with a span of 10%. Periods of promoter activity or inactivity were then determined based on the slope of the fluorescence trace. **(A)** Example of smoothing transcriptional traces. **(B)** Fluorescence trace of a single punctum during nc14. Open black circles indicate time points where the promoter has turned "on", filled red circles indicate time points where the promoter is identified as turning "off". **(C)** Transcriptional trace with the green shaded region under the curve used to calculate the size of the first burst. This area of this region is calculated using the trapz function in MATLAB and extends from the time point the promoter is called "on" until the next time it is called "on". Panels **(D - F)** show additional representative fluorescence traces of single transcriptional puncta during nc14. **(D)** A trace with the entire region under the curved shaded green represents the area used to calculate the total amount of mRNA produced. This area is calculated using the trapz function in MATLAB extends from the time the promoter is first called "on" until 50 min into nc14 or the movie ends, whichever comes first. **(E)** Burst frequency is calculated by dividing the number of bursts that occur during nc14 by the length of time from the first time the promoter is called "on" until 50 min into nc14 or the movie ends, whichever comes first. **(F)** Burst duration is calculated by taking the amount of time between when the promoter is called "on" and it is next called "off". **(G)** Polymerase initiation rate is calculated by taking the slope of the smoothed fluorescence race at the midpoint between when the promoter is called "on" and it is next called "off".

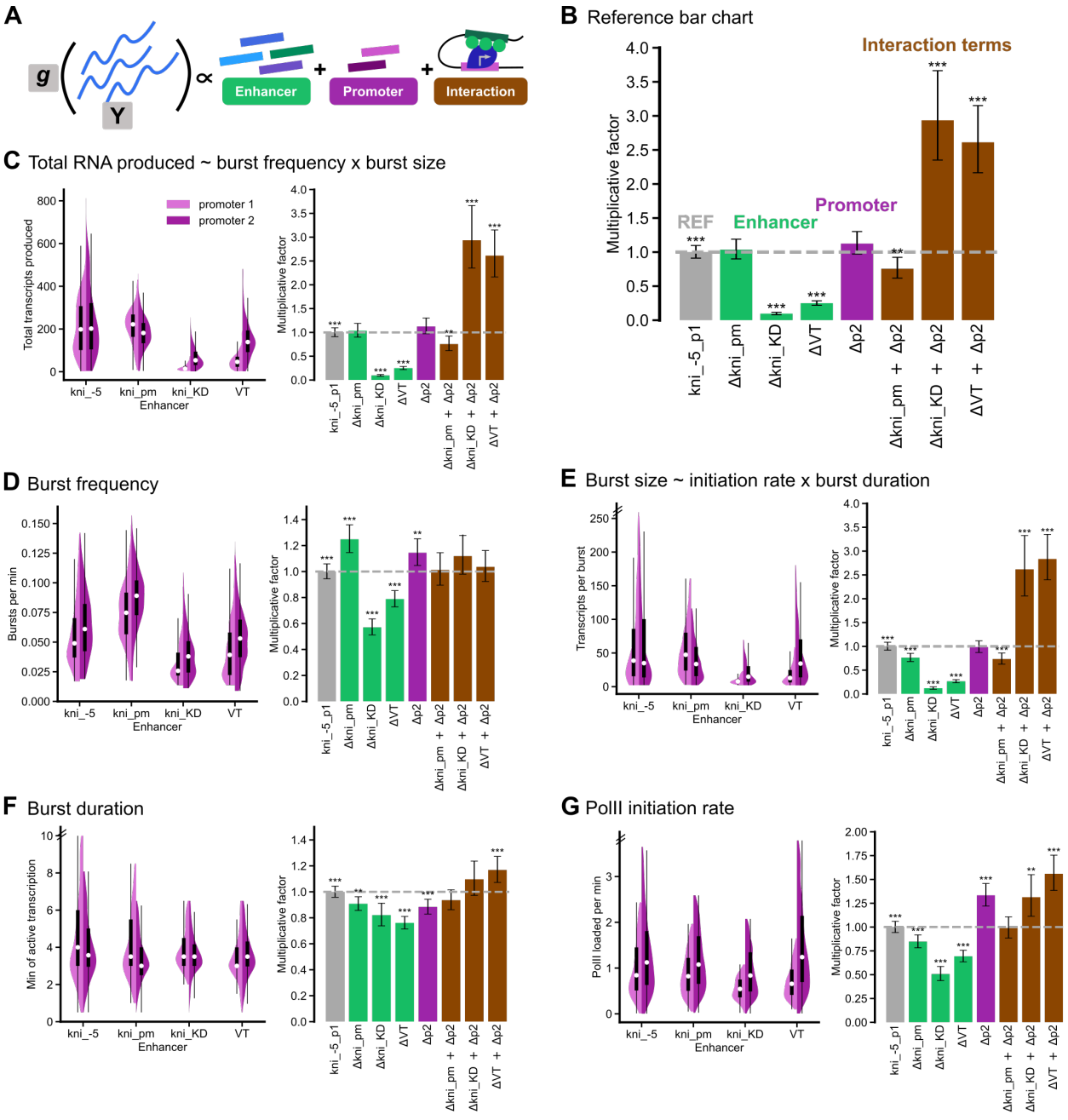


**Figure S3.5:** Using canonical link functions gives the same results. Here, we show the results from the generalized linear models (GLMs) when using the log link function instead of the identity link function, which was used in Figure 3.4. **(A)** To parse the effects of the enhancer, the promoter, and their interactions on all burst properties, we built GLMs. $Y$ represents the burst property under study, $g$ is the link function, and the enhancers, promoters, and their interaction terms are the explanatory variables. The coefficients of each of these explanatory variables is representative of that variable's contribution to the total value of the burst property. **(B)** All burst property data was taken from the anterior-posterior bin of maximum expression (22% and 63%) for the anterior band and the posterior stripe, respectively. We exponentiate the coefficients and the 95% confidence intervals for

**Figure S3.5 (cont.):** each independent variable to invert the log link function and call these quantities the "multiplicative factors." Performing this conversion yields a multiplicative relationship between our response variable (the burst property) and our explanatory variables. The reference construct (kni_-5-p1) has been set to 1 such that multiplying the relevant multiplicative factors gives you the value that, if multiplied by the reference construct value, will gives you the average value of the burst property for a particular construct. These factors are plotted as a bar graph; $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. The reference construct is represented in gray, and the effects of enhancer, promoter, and their interactions are represented in green, purple, and brown, respectively. Thus, the average value of the burst property for a particular construct, e.g. VT-p2, relative to the reference construct would be 0.73, which is the product of $\Delta$VT = 0.25, $\Delta$p2 = 1.1, and $\Delta$VT + $\Delta$p2 = 2.6. The average value of the burst property for VT-p2 would then be $0.73 \times 205 = 150$. Note that for simplicity, kni_proximal_minimal and VT33935 has been shortened to kni_pm and VT, respectively, in the following graphs. In panels **(C - G),** (left) split violin plots (and their associated box plots) of burst properties for all eight constructs will be plotted with promoter 1 in light purple and promoter 2 in purple. The black boxes span the lower to upper quartiles, with the white dot within the box indicating the median. Whiskers extend to 1.5*IQR (interquartile range) ś the upper and lower quartile, respectively. (right) Bar graphs representing the relative contributions of enhancer, promoter, and their interactions to each burst property are plotted as described in **(B)**. The double hash marks on the axes indicate that 90% of the data is being shown. **(C)** Expression levels are mainly determined by the enhancer and the interaction terms. Some enhancers (kni_-5 and kni_proximal_minimal) appear to work well with both promoters; whereas, kni_KD and VT, which are bound by similar TFs, show much higher expression with promoter 2. **(D)** Burst frequency is dominated by the enhancer and promoter terms, with promoter 2 consistently producing higher burst frequencies regardless of enhancer. **(E)** Burst size, which is determined by both initiation rate and burst duration, is dominated by the enhancer and interaction terms. As **(F)** burst duration is reasonably consistent regardless of enhancer or promoter, differences in burst size are mainly dependent on differences in **(G)** PolII initiation rate, with this burst property as the main molecular knob affected by molecular compatibility.

# Chapter 4

# Discussion

The ability to unlock the rules that govern how functions are encoded in regulatory sequences has been an elusive goal despite several decades of research. As with many other topics, further study has revealed layer after layer of added complexity to what originally seemed like a simple problem—decoding a language that consists of four simple nucleotides in varying order. Here, we have explored how one constraint (the regulatory function of an enhancer) can shape that enhancer's structure (Chapter 2) and how multiple promoters of a single gene can serve multiple regulatory roles (Chapter 3). These studies have expanded our understanding of how function is encoded in regulatory DNA and the molecular mechanisms underpinning the collective activation of transcription by enhancers and promoters.

However, to build a fuller understanding of how to interpret or "read" the regulatory sequence to determine its function requires the ability to identify 1) the relevant players (enhancer boundaries and what other sequences or proteins may affect enhancer/promoter function), 2) the sequence features (TF binding sites and their affinity and arrangement), and 3) the interdependent relationships between sequence features due to various mechanistic, biological, and evolutionary constraints. Importantly, identifying these features and the relationships

between them will involve elucidating the molecular mechanisms that facilitate the execution of the DNA's regulatory task.

First, it feels appropriate to address our motivations in trying to understand how regulatory DNA encodes instructions for gene expression. Being able to predict the function or expression output of enhancers is often invoked as a motivation for exploring this field. This is a problem that can likely be solved by obtaining more data and developing a machine learning algorithm to make predictions of expression output. It may even be the case that there are so many dependencies between sequence features and other players in the system that designing a synthetic enhancer with a desired function without the help of a computer algorithm is improbable. However, gaining an understanding of the molecular mechanisms that are affected by the sequence features in regulatory DNA may be a useful objective in itself. Even if this does not allow us to predict expression output directly, understanding these molecular underpinnings will help us identify disease variants and understand what molecular players we can target to alleviate the symptoms of disease.

## 4.1 How regulatory task constrains promoter shape

Here, we will consider some future directions for the projects covered in this document. Our work in Chapter 3 revealed that developmental genes often have multiple promoters, and there is a general trend of more active promoters also being broader, with a series of initiation sites spanning a range typically considered characteristic of housekeeping promoters. To uncover the evolutionary constraints on these sets of broad and highly active promoters versus the sharper and less active promoters, we can consider the natural variation in these regulatory elements and look for signs of selection pressure. Given the fact that the sharper promoters specify a narrow band of DNA in which initiation can occur, we expect that these promoters have more highly conserved motif content than the broader promoters which

consist of many sites of initiation. We can test this hypothesis by using the *Drosophila* Genome Nexus data [84], a compilation of ~1000 *Drosophila melanogaster* genomes, and comparing the frequency of SNPs within promoter motifs in sharper vs broader promoters. Alternatively, the constraint may be on a bulk feature like the number of Inr sites rather than strict conservation of each site; in that case, we can compare how the numbers of these sites compare across *Drosophila* species using the Batut RAMPAGE data [6]. This could help us understand which feature of promoters is crucial to delivering the desired expression output. Is it which motifs are present? Or is it the number of initiation sites, or even the pattern of initiation sites? Sequence conservation of any of these features could indicate that they are the key to successful transcription at promoters.

## 4.2 Does the molecular compatibility of proteins recruited to the enhancer and promoter generally affect polymerase loading?

In Chapter 3, we found that RNA polymerase initiation rate to be the molecular mechanism mostly acutely impacted by the molecular compatibility of the proteins recruited to the enhancer and promoter. Untangling how exactly polymerase loading is tuned and what specific molecular factors are at play will involve improving our understanding of TF activating domains, how these activating domains nucleate transcriptional hubs, and how Mediator plays a role in this process, among other things. However, we should start by testing how generalizable tuning polymerase loading is. We can do so by 1) identifying other gap genes and early developmental genes that are regulated by multiple enhancers regulated by similar and dissimilar sets of TFs and multiple promoters with motif content similar and dissimilar to those of the *knirps* promoters and 2) tracking the transcriptional dynamics of MS2-tagged

combinations of their enhancers and promoters. It was notable that all our reporters, including those with the mutated promoter 2, showed that molecular compatibilitys main effect was on polymerase initiation rate. Thus, it is likely that while the specific effect (tuning initiation rate up or down) is dependent on the proteins recruited to the enhancers and promoters and thus on the motif content of these regulatory elements, the fact that polymerase initiation rate is the key point of regulation is not. Accordingly, we would expect that these experiments would support this fact. Results indicating otherwise would suggest that there is something about the TFs recruited to the *knirps* enhancers that leads to this specific relationship between molecular compatibility and initiation rate.

## 4.3 Evaluating global TF-promoter motif preferences and their impact on expression output

While we took a detailed look at the impact of different combinations of endogenously paired enhancers and promoters in a developing *Drosophila* embryo in Chapter 3, we would like to build on the idea that TFs have preferences for core promoter motifs and identify these preferences by characterizing global TF-promoter motif co-occurrence patterns and validating these preferences in cell culture. As discussed in Chapter 3, our effort to identify these preferences reproduced similar results as previous bioinformatic analyses and experiments validating the preference between the TFs, Caudal and Dorsal [70, 180]. We also identified a preference between bicoid and DPE in combination with subsets of Inr, TATA Box, and MTE. Overall, we showed that TF and core promoter motifs show non-random co-occurrence patterns.

There are two hypotheses for why TFs and promoter motifs co-occur in a non-random manner. As suggested by previous work on the compatibility of certain TFs and core promoter

motifs, co-occurrence may be indicative of the dependency of that TF on the particular core promoter motif, leading to decreased expression levels when the motif is removed. Alternatively, increased co-occurrence may be the result of selection for robust expression output. While I favor the second hypothesis, it is possible that these preferred interactions lead to increased stability of the interaction between enhancer and promoter, both decreasing expression noise and increasing expression levels, as appears to be the general trend [29].

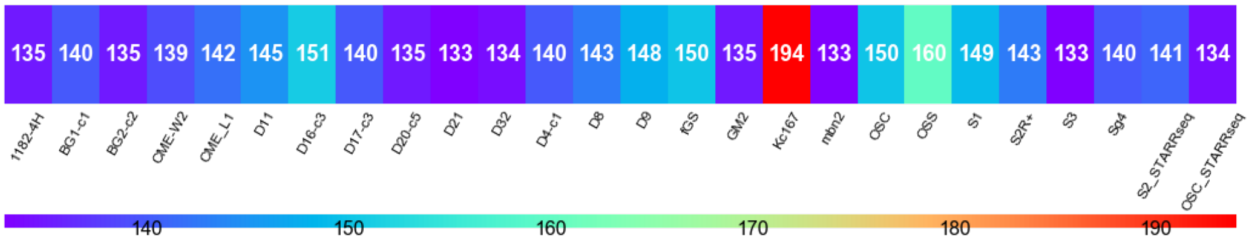


**Figure 4.1:** Heatmap of the number of TFs expressed in a set of cell lines based on modEncode RNA-seq data [50]. An RPKM of 1 is used as a cutoff for calling a TF expressed.

We can test these hypotheses by doing a global assay in *Drosophila* Kc167 cells [21], in which more of our early key TFs are expressed compared to the other commonly used S2 cells (**Figure 4.1**). We can use the pSTARR-seq_fly plasmid [1], which contains the *Drosophila* Synthetic Core Promoter (DSCP) [126]. Like the super core promoter (SCP), the DSCP contains four important core promoter motifs—TATA box, Inr, MTE, and DPE—and has been used in large-scale characterizations of enhancers and enhancer-promoter interactions, indicating that it works well with many enhancers [178]. We can make seven versions of the DSCP, removing the motifs as necessary to create only the motif combinations observed in my dataset of developmental promoters. We can then pair these seven promoters with the set of developmental enhancers previously collated when looking for patterns of TF-promoter motif co-occurrence. To obviate the need to add an additional marker for the promoter, the constructs containing each promoter can be transfected into separate pools of Kc167 cells. We can then perform STARR-seq on the seven pools of constructs (one for each promoter) to measure expression levels driven by each enhancer-promoter pair and do replicates of STARR-seq to measure expression noise. This strategy can be assessed by doing

a pilot study in which we identify a few enhancers that are predicted to work well with very different sets of promoter motifs, combinatorially pairing them with the seven promoters to create 21 reporter constructs and determining expression levels by performing qPCR.

These experiments will allow me to see various outcomes. It may be that higher expression levels are produced by the preferred pairing of TFs and core promoter motifs, suggesting that frequently co-occurring TF and promoter motif pairs may depend on each other to drive expression. Alternatively, expression levels may vary irrespective of the preferred pairing, suggesting that other features of the promoter or enhancer, e.g. unknown motifs or TF binding sites, may be modulating aspects of transcriptional bursting and affecting expression level and noise. It may also be possible that higher expression levels are driven by a particular set of promoter motifs, implying that particular promoter motif combinations may modulate aspects of transcriptional bursting, e.g. polymerase loading or rate of promoter switching to an inactive state, such that expression level is increased. I expect that the first outcome is most likely, with preferred pairings producing increased expression levels correlated with decreased expression noise. One mechanism by which this could occur is an increased stability of the interaction between enhancer and promoter, which would lead to increased burst frequency and decreased noise in addition to increased expression levels. However, noise could also be uncoupled from expression levels, with particular promoter motif combinations producing different levels of noise. This would suggest that the combination of promoter motifs intrinsically produces a certain level of noise that is unrelated to or relatively unaffected by the pairing with any particular enhancer or TF. These experiments will allow me to test the hypothesis that increased co-occurrence of TFs and core promoter motifs is a result of more stable interactions, leading to a decrease in expression noise and likely an increase in expression level.

Here, we explored some extensions to the work discussed in the previous chapters. This work probed questions of gene regulatory network function by pairing imaging-based and genomic

measurements of gene expression with statistical and physically-based computational models. Moving forward research methods that integrate the tools and ideas from multiple disciplines will be necessary to approach fully understanding how function is encoded in regulatory sequence. While my work here only scratched the surface of some questions about gene regulatory function, it's only left me feeling more excited by how much there is to still explore.

# Bibliography


[1] C. D. Arnold, D. Gerlach, C. Stelzer, Ł. M. Boryń, M. Rath, A. Stark, L. M. Boryn, M. Rath, A. Stark, Ł. M. Boryń, M. Rath, and A. Stark. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (New York, N.Y.)*, 339(6123):1074–7, Mar 2013.

[2] D. N. Arnosti and M. M. Kulkarni. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry*, 94(5):890–898, 2005.

[3] Ž. Avsec, M. Weilert, A. Shrikumar, S. Krueger, A. Alexandari, K. Dalal, R. Fropf, C. McAnany, J. Gagneur, A. Kundaje, and J. Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3):354–366, 2021.

[4] C. R. Bartman, S. C. Hsu, C. C. S. Hsiung, A. Raj, and G. A. Blobel. Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. *Molecular Cell*, 62(2):237–247, 2016.

[5] P. Batut, A. Dobin, C. Plessy, P. Carninci, and T. R. Gingeras. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Research*, 23(1):169–180, 2013.

[6] P. J. Batut and T. R. Gingeras. Conserved noncoding transcription and core promoter regulatory code in early *Drosophila* development. *eLife*, 6:156596, 2017.

[7] O. G. Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *Journal of molecular biology*, 193(4):723–50, Feb 1987.

[8] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen. Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):757–62, 2002.

[9] E. M. Blackwood and J. T. Kadonaga. Going the distance: a current view of enhancer action. *Science (New York, N.Y.)*, 281(5373):60–3, Jul 1998.

[10] W. J. Blake, M. Kærn, C. R. Cantor, and J. J. Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, 2003.

[11] B. Bonev, N. Mendelson Cohen, Q. Szabo, L. Fritsch, G. L. Papadopoulos, Y. Lubling, X. Xu, X. Lv, J. P. Hugnot, A. Tanay, and G. Cavalli. Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, 171(3):557–572.e24, 2017.

[12] S. Bonn, R. P. Zinzen, C. Girardot, E. H. Gustafson, A. Perez-Gonzalez, N. Delhomme, Y. Ghavi-Helm, B. Wilczyński, A. Riddell, and E. E. M. Furlong. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, 44(2):148–156, Feb 2012.

[13] J. P. Bothma, H. G. Garcia, S. Ng, M. W. Perry, T. Gregor, and M. Levine. Enhancer additivity and non-additivity are determined by enhancer strength in the *Drosophila* embryo. *eLife*, 4(AUGUST2015), 2015.

[14] J. B. Brown, N. Boley, R. Eisman, G. E. May, M. H. Stoiber, M. O. Duff, B. W. Booth, J. Wen, S. Park, A. M. Suzuki, K. H. Wan, C. Yu, D. Zhang, J. W. Carlson, L. Cherbas, B. D. Eads, D. Miller, K. Mockaitis, J. Roberts, C. A. Davis, E. Frise, A. S. Hammonds, S. Olson, S. Shenker, D. Sturgill, A. A. Samsonova, R. Weiszmann, G. Robinson, J. Hernandez, J. Andrews, P. J. Bickel, P. Carninci, P. Cherbas, T. R. Gingeras, R. A. Hoskins, T. C. Kaufman, E. C. Lai, B. Oliver, N. Perrimon, B. R. Graveley, and S. E. Celniker. Diversity and dynamics of the *Drosophila* transcriptome. *Nature*, 512(7515):393–399, Aug 2014.

[15] T. W. Burke and J. T. Kadonaga. *Drosophila* TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes and Development*, 10(6):711–724, 1996.

[16] J. E. Butler and J. T. Kadonaga. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes and Development*, 15(19):2515–2519, 2001.

[17] A. Butryn, J. M. Schuller, G. Stoehr, P. Runge-Wollmann, F. Förster, D. T. Auble, and K. P. Hopfner. Structural basis for recognition and remodeling of the TBP:DNA:NC2 complex by Mot1. *eLife*, 4(AUGUST2015):1–22, 2015.

[18] L. B. Carey, D. van Dijk, P. M. Sloot, J. A. Kaandorp, and E. Segal. Promoter Sequence Determines the Relationship between Expression Level and Noise. *PLoS Biology*, 11(4), 2013.

[19] J. W. Cave, F. Loh, J. W. Surpris, L. Xia, and M. A. Caudy. A DNA transcription code for cell-specific gene activation by Notch signaling. *Current biology : CB*, 15(2):94–104, Jan 2005.

[20] G. E. Chalkley and C. P. Verrijzer. DNA binding site selection by RNA polymerase II TAFs: A TAF(II)250-TAF(II)150 complex recognizes the initiator. *EMBO Journal*, 18(17):4835–4845, 1999.

[21] L. Cherbas, A. Willingham, D. Zhang, L. Yang, Y. Zou, B. D. Eads, J. W. Carlson, J. M. Landolin, P. Kapranov, J. Dumais, A. Samsonova, J. H. Choi, J. Roberts, C. A. Davis, H. Tang, M. J. Van Baren, S. Ghosh, A. Dobin, K. Bell, W. Lin, L. Langton, M. O. Duff, A. E. Tenney, C. Zaleski, M. R. Brent, R. A. Hoskins, T. C. Kaufman, J. Andrews, B. R. Graveley, N. Perrimon, S. E. Celniker, T. R. Gingeras, and P. Cherbas. The transcriptional diversity of 25 *Drosophila* cell lines. *Genome Research*, 21(2):301–314, 2011.

[22] S. W. Cho, J. Xu, R. Sun, M. R. Mumbach, A. C. Carter, Y. G. Chen, K. E. Yost, J. Kim, J. He, S. A. Nevins, S. F. Chin, C. Caldas, S. J. Liu, M. A. Horlbeck, D. A. Lim, J. S. Weissman, C. Curtis, and H. Y. Chang. Promoter of lncRNA Gene PVT1 Is a Tumor-Suppressor DNA Boundary Element. *Cell*, 173(6):1398–1412.e22, 2018.

[23] M. Corrales, A. Rosado, R. Cortini, J. van Arensbergen, B. van Steensel, and G. J. Filion. Clustering of *Drosophila* housekeeping promoters facilitates their expression. *Genome research*, 27(7):1153–1161, 2017.

[24] J. M. Cox, M. M. Hayward, J. F. Sanchez, L. D. Gegnas, S. Van Der Zee, J. H. Dennis, P. B. Sigler, and A. Schepartz. Bidirectional binding of the TATA box binding protein to the TATA box. *Proceedings of the National Academy of Sciences of the United States of America*, 94(25):13475–13480, 1997.

[25] J. Crocker, Y. Tamori, and A. Erives. Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS biology*, 6(11):e263, Nov 2008.

[26] T. E. Crowley, T. Hoey, J. K. Liu, Y. N. Jan, L. Y. Jan, and R. Tjian. A new factor related to TATA-binding protein has highly restricted expression patterns in *Drosophila*. *Nature*, 361(6412):557–561, 1993.

[27] S. Cruz-Molina, P. Respuela, C. Tebartz, P. Kolovos, M. Nikolic, R. Fueyo, W. F. van Ijcken, F. Grosveld, P. Frommolt, H. Bazzi, and A. Rada-Iglesias. PRC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation. *Cell Stem Cell*, 20(5):689–705.e9, 2017.

[28] R. D. Dar, B. S. Razooky, A. Singh, T. V. Trimeloni, J. M. McCollum, C. D. Cox, M. L. Simpson, and L. S. Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, 109(43):17454–17459, 2012.

[29] R. D. Dar, S. M. Shaffer, A. Singh, B. S. Razooky, M. L. Simpson, A. Raj, and L. S. Weinberger. Transcriptional Bursting Explains the Noise-Versus-Mean Relationship in mRNA and Protein Levels. *PloS one*, 11(7):e0158298, 2016.

[30] M. De Gobbi, V. Viprakasit, J. R. Hughes, C. Fisher, V. J. Buckle, H. Ayyub, R. J. Gibbons, D. Vernimmen, Y. Yoshinaga, P. De Jong, J. F. Cheng, E. M. Rubin, W. G. Wood, D. Bowden, and D. R. Higgs. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science*, 312(5777):1215–1217, 2006.

[31] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[32] R. Dreos, G. Ambrosini, R. Cavin Périer, and P. Bucher. EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic acids research*, 41(Database issue):D157–64, Jan 2013.

[33] R. Dreos, G. Ambrosini, R. Groux, R. Cavin Périer, and P. Bucher. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Research*, 45(November 2016):gkw1069, 2016.

[34] S. H. Duttke, R. F. Doolittle, Y. L. Wang, and J. T. Kadonaga. TRF2 and the evolution of the bilateria. *Genes and Development*, 28(19):2071–2076, 2014.

[35] E. El-Sherif and M. Levine. Shadow Enhancers Mediate Dynamic Shifts of Gap Gene Expression in the *Drosophila* Embryo. *Current biology : CB*, 26(9):1164–9, May 2016.

[36] L. L. Ellis, W. Huang, A. M. Quinn, A. Ahuja, B. Alfrejd, F. E. Gomez, C. E. Hjelmen, K. L. Moore, T. F. C. Mackay, J. S. Johnston, and A. M. Tarone. Intrapopulation Genome Size Variation in *D. melanogaster* Reflects Life History Variation and Plasticity. *PLoS Genetics*, 10(7), 2014.

[37] K. H. Emami, A. Jain, and S. T. Smale. Mechanism of synergy between TATA and initiator: Synergistic binding of TFIID following a putative TFIIA-induced isomerization. *Genes and Development*, 11(22):3007–3019, 1997.

[38] J. Falo-Sanjuan, N. C. Lammers, H. G. Garcia, and S. J. Bray. Enhancer Priming Enables Fast and Sustained Transcriptional Responses to Notch Signaling. *Developmental Cell*, 50(4):411–425.e8, 2019.

[39] E. K. Farley, K. M. Olson, W. Zhang, A. J. Brandt, D. S. Rokhsar, and M. S. Levine. Suboptimization of developmental enhancers. *Science (New York, N.Y.)*, 350(6258):325–8, 2015.

[40] E. K. Farley, K. M. Olson, W. Zhang, D. S. Rokhsar, and M. S. Levine. Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proceedings of the National Academy of Sciences of the United States of America*, 113(23):6508–13, Jun 2016.

[41] C. C. Fowlkes, C. L. L. Hendriks, S. V. E. Keränen, G. H. Weber, O. Rübel, M.-Y. 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, J. Malik, C. L. Luengo Hendriks, S. V. E. Keränen, G. H. Weber, O. Rübel, M.-Y. 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. A Quantitative Spatiotemporal Atlas of Gene Expression in the *Drosophila* Blastoderm. *Cell*, 133(2):364–374, Apr 2008.

[42] M. Franke, D. M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jerković, W. L. Chan, M. Spielmann, B. Timmermann, L. Wittler, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz, and S. Mundlos. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624):265–269, 2016.

[43] M. C. Frith, J. L. Spouge, U. Hansen, and Z. P. Weng. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Research*, 30(14):3214–3224, 2002.

[44] T. Fukaya, B. Lim, and M. Levine. Enhancer Control of Transcriptional Bursting. *Cell*, 166(2):358–68, Jul 2016.

[45] H. G. Garcia, M. Tikhonov, A. Lin, and T. Gregor. Quantitative Imaging of Transcription in Living *Drosophila* Embryos Links Polymerase Activity to Patterning, Nov 2013.

[46] J. Gehrig, M. Reischl, E. Kalmár, M. Ferg, Y. Hadzhiev, A. Zaucker, C. Song, S. Schindler, U. Liebel, and F. Müller. Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. *Nature methods*, 6(12):911–916, 2009.

[47] N. I. Gershenzon, E. N. Trifonov, and I. P. Ioshikhes. The features of *Drosophila* core promoters revealed by statistical analysis. *BMC Genomics*, 7:1–13, 2006.

[48] R. J. Geusz, A. Wang, J. Chiou, J. J. Lancman, N. Wetton, S. Kefalopoulou, J. Wang, Y. Qiu, J. Yan, A. Aylward, B. Ren, P. D. S. Dong, K. J. Gaulton, and M. Sander. Pancreatic progenitor epigenome maps prioritize type 2 diabetes risk genes with roles in development. *eLife*, 10:1–39, 2021.

[49] Y. Ghavi-Helm, F. A. Klein, T. Pakozdi, L. Ciglar, D. Noordermeer, W. Huber, and E. E. M. Furlong. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, 512(7512):96–100, 2014.

[50] B. R. Graveley, A. N. Brooks, J. W. Carlson, M. O. Duff, J. M. Landolin, L. Yang, C. G. Artieri, M. J. Van Baren, N. Boley, B. W. Booth, J. B. Brown, L. Cherbas, C. A. Davis, A. Dobin, R. Li, W. Lin, J. H. Malone, N. R. Mattiuzzo, D. Miller, D. Sturgill, B. B. Tuch, C. Zaleski, D. Zhang, M. Blanchette, S. Dudoit, B. Eads, R. E. Green, A. Hammonds, L. Jiang, P. Kapranov, L. Langton, N. Perrimon, J. E. Sandler, K. H. Wan, A. Willingham, Y. Zhang, Y. Zou, J. Andrews, P. J. Bickel, S. E. Brenner, M. R. Brent, P. Cherbas, T. R. Gingeras, R. A. Hoskins, T. C. Kaufman, B. Oliver, and S. E. Celniker. The developmental transcriptome of *Drosophila melanogaster*. *Nature*, 471(7339):473–479, 2011.

[51] M. S. Halfon, S. M. Gallo, and C. M. Bergman. REDfly 2.0: an integrated database of *cis*-regulatory modules and transcription factor binding sites in *Drosophila*. *Nucleic acids research*, 36(Database issue):D594–D598, 2008.

[52] M. S. Halfon, Y. Grad, G. M. Church, and A. M. Michelson. Computation-based discovery of related transcriptional regulatory modules and motifs using a experimentally validated combinatorial model. *Genome Research*, 12(7):1019–1028, 2002.

[53] A. A. S. Hammonds, C. C. a. Bristow, W. W. Fisher, R. Weiszmann, S. Wu, V. Hartenstein, M. Kellis, B. Yu, E. Frise, and S. E. Celniker. Spatial expression of transcription factors in *Drosophila* embryonic organ development. *Genome biology*, 14(12):R140, 2013.

[54] S. K. Hansen, S. Takada, R. H. Jacobson, J. T. Lis, and R. Tjian. Transcription properties of a cell type-specific TATA-binding protein, TRF. *Cell*, 91(1):71–83, 1997.

[55] R. C. Hardison and J. Taylor. Genomic approaches towards finding *cis*-regulatory modules in animals. *Nature Reviews Genetics*, 13(7):469–483, 2012.

[56] O. Hendy, J. Campbell, J. D. Weissman, D. R. Larson, and D. S. Singer. Differential context-specific impact of individual core promoter elements on transcriptional dynamics. *Molecular biology of the cell*, 28(23):3360–3370, Nov 2017.

[57] G. Z. Hertz and G. D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)*, 15(7-8):563–77, 1999.

[58] Y. Hiromi and W. J. Gehring. Regulation and function of the *Drosophila* segmentation gene fushi tarazu. *Cell*, 50(6):963–74, Sep 1987.

[59] Y. Hiromi, A. Kuroiwa, and W. J. Gehring. Control elements of the *Drosophila* segmentation gene *fushi tarazu*. *Cell*, 43(3 Pt 2):603–13, Dec 1985.

[60] D. Hnisz, A. S. Weintrau, D. S. Day, A. L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, and R. A. Young. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351(6280):1454–1458, 2016.

[61] A. Hoermann, D. Cicin-Sain, and J. Jaeger. A quantitative validated model reveals two phases of transcriptional regulation for the gap gene *giant* in *Drosophila*. *Developmental Biology*, 411(2):325–338, 2016.

[62] M. C. Holmes and R. Tjian. Promoter-selective properties of the TBP-related factor TRF1. *Science*, 288(5467):867–870, 2000.

[63] C. Hoppe, J. R. Bowles, T. G. Minchington, C. Sutcliffe, P. Upadhyai, M. Rattray, and H. L. Ashe. Modulation of the Promoter Activation Rate Dictates the Transcriptional Response to Graded BMP Signaling Levels in the *Drosophila* Embryo. *Developmental Cell*, 54(6):727–741.e7, 2020.

[64] T. H. S. Hsieh, C. Cattoglio, E. Slobodyanyuk, A. S. Hansen, O. J. Rando, R. Tjian, and X. Darzacq. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Molecular Cell*, 78(3):539–553.e8, 2020.

[65] J. Y. Hsu, T. Juven-Gershon, M. T. Marr, K. J. Wright, R. Tjian, and J. T. Kadonaga. TBP, Mot1, and NC2 establish a regulatory circuit that controls DPE-dependent versus TATA-dependent transcription. *Genes and Development*, 22(17):2353–2358, 2008.

[66] J. Ibn-Salem, S. Köhler, M. I. Love, H. R. Chung, N. Huang, M. E. Hurles, M. Haendel, N. L. Washington, D. Smedley, C. J. Mungall, S. E. Lewis, C. E. Ott, S. Bauer, P. N. Schofield, S. Mundlos, M. Spielmann, and P. N. Robinson. Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome biology*, 15(9):423, 2014.

[67] E. Ing-Simmons, R. Vaid, X. Y. Bing, M. Levine, M. Mannervik, and J. M. Vaquerizas. Independence of chromatin conformation and gene regulation during *Drosophila* dorsoventral patterning. *Nature Genetics*, 53(4):487–499, 2021.

[68] F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C.-A. A. Yen, A. D. Schmitt, C. A. Espinoza, and B. Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290–294, 2013.

[69] G. A. Jindal and E. K. Farley. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Developmental Cell*, 56(5):575–587, 2021.

[70] T. Juven-Gershon, J. Y. Hsu, J. W. Theisen, and J. T. Kadonaga. The RNA polymerase II core promoter - the gateway to transcription. *Current Opinion in Cell Biology*, 20(3):253–259, 2008.

[71] J. T. Kadonaga. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol.*, 1(1):40–51, 2012.

[72] A. Kedmi, Y. Zehavi, Y. Glick, Y. Orenstein, D. Ideses, C. Wachtel, T. Doniger, H. W. Ben-Asher, N. Muster, J. Thompson, S. Anderson, D. Avrahami, J. R. Yates, R. Shamir, D. Gerber, and T. Juven-Gershon. *Drosophila* TRF2 is a preferential core promoter regulator. *Genes and Development*, 28(19):2163–2174, 2014.

[73] J. L. Kim, D. B. Nikolov, and S. K. Burley. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, 365(6446):520–527, 1993.

[74] T. K. Kim and T. Maniatis. The mechanism of transcriptional synergy of an *in vitro* assembled interferon-beta enhanceosome. *Molecular cell*, 1:119–129, 1997.

[75] Y. Kim, J. H. Geiger, S. Hahn, and P. B. Sigler. Crystal structure of a yeast TBP/TATA-box complex. *Nature*, 365(6446):512–20, Oct 1993.

[76] D. M. King, C. K. Y. Hong, J. L. Shepherdson, D. M. Granas, B. B. Maricque, and B. A. Cohen. Synthetic and genomic regulatory elements reveal aspects of Cis-regulatory grammar in mouse embryonic stem cells. *eLife*, 9:1–24, 2020.

[77] D. V. Kopytova, A. N. Krasnov, M. R. Kopantceva, E. N. Nabirochkina, J. V. Nikolenko, O. Maksimenko, M. M. Kurshakova, L. A. Lebedeva, M. M. Yerokhin, O. B. Simonova, L. I. Korochkin, L. Tora, P. G. Georgiev, and S. G. Georgieva. Two

Isoforms of *Drosophila* TRF2 Are Involved in Embryonic Development, Premeiotic Chromatin Condensation, and Proper Differentiation of Germ Cells of Both Sexes. *Molecular and Cellular Biology*, 26(20):7492–7505, 2006.

[78] N. Krietenstein, S. Abraham, S. V. Venev, N. Abdennur, J. Gibcus, T. H. S. Hsieh, K. M. Parsi, L. Yang, R. Maehr, L. A. Mirny, J. Dekker, and O. J. Rando. Ultrastructural Details of Mammalian Chromosome Architecture. *Molecular Cell*, 78(3):554–565.e7, 2020.

[79] M. M. Kulkarni and D. N. Arnosti. Information display by transcriptional enhancers. *Development (Cambridge, England)*, 130(26):6569–75, Dec 2003.

[80] A. K. Kutach and J. T. Kadonaga. The Downstream Promoter Element DPE Appears To Be as Widely Used as the TATA Box in *Drosophila* Core Promoters. *Molecular and Cellular Biology*, 20(13):4754–4764, 2000.

[81] E. Z. Kvon, T. Kazmar, G. Stampfel, J. O. Yáñez-Cuna, M. Pagani, K. Schernhuber, B. J. Dickson, and A. Stark. Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo. *Nature*, 512(7512):91–5, 2014.

[82] E. Z. Kvon, R. Waymack, M. G. Elabd, and Z. Wunderlich. Enhancer redundancy in development and disease. *Nature Reviews Genetics*, 2021.

[83] O. Kyrchanova and P. Georgiev. Chromatin insulators and long-distance interactions in *Drosophila*. *FEBS Letters*, 588(1):8–14, 2014.

[84] J. B. Lack, J. D. Lange, A. D. Tang, R. B. Corbett-Detig, and J. E. Pool. A thousand fly genomes: An expanded *Drosophila* genome nexus. *Molecular Biology and Evolution*, 33(12):3308–3313, 2016.

[85] N. C. Lammers, V. Galstyan, A. Reimer, S. A. Medin, C. H. Wiggins, and H. G. Garcia. Multimodal transcriptional control of pattern formation in embryonic development. *Proceedings of the National Academy of Sciences of the United States of America*, 117(2):836–847, Dec 2020.

[86] J. R. Landry, D. L. Mager, and B. T. Wilhelm. Complex controls: The role of alternative promoters in mammalian genomes. *Trends in Genetics*, 19(11):640–648, 2003.

[87] A. J. Larsson, P. Johnsson, M. Hagemann-Jensen, L. Hartmanis, O. R. Faridani, B. Reinius, Å. Segerstolpe, C. M. Rivera, B. Ren, and R. Sandberg. Genomic encoding of transcriptional burst kinetics. *Nature*, 565(7738):251–254, Jan 2019.

[88] C. H. Lee, H. Shin, and J. Kimble. Dynamics of Notch-Dependent Transcriptional Bursting in Its Native Context. *Developmental Cell*, 50(4):426–435.e4, 2019.

[89] S. J. Lee, T. Sekimoto, E. Yamashita, E. Nagoshi, A. Nakagawa, N. Imamoto, M. Yoshimura, H. Sakai, K. T. Chong, T. Tsukihara, and Y. Yoneda. The structure of importin-beta bound to SREBP-2: nuclear import of a transcription factor. *Science (New York, N.Y.)*, 302(5650):1571–5, Nov 2003.

[90] B. Lehner. Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Molecular Systems Biology*, 4(170), 2008.

[91] M. Levine and E. H. Davidson. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences*, 102(14):4936–4942, 2005.

[92] B. A. Lewis, R. J. Sims, W. S. Lane, and D. Reinberg. Functional characterization of core promoter elements: DPE-specific transcription requires the protein kinase CK2 and the PC4 coactivator. *Molecular Cell*, 18(4):471–481, 2005.

[93] G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. K. Sung, M. Snyder, and Y. Ruan. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84–98, 2012.

[94] L. Li and Z. Wunderlich. An Enhancer's Length and Composition Are Shaped by Its Regulatory Task. *Frontiers in Genetics*, 8(May), May 2017.

[95] X.-y. Li, S. MacArthur, R. Bourgon, D. Nix, D. a. Pollard, V. N. Iyer, A. Hechmer, L. Simirenko, M. Stapleton, C. L. Luengo Hendriks, H. C. Chu, N. Ogawa, W. Inwood, V. Sementchenko, A. Beaton, R. Weiszmann, S. E. Celniker, D. W. Knowles, T. Gingeras, T. P. Speed, M. B. Eisen, and M. D. Biggin. Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS biology*, 6(2):e27, Feb 2008.

[96] A. P. Lifanov, V. J. Makeev, A. G. Nazina, and D. A. Papatsenko. Homotypic regulatory clusters in *Drosophila*. *Genome research*, 13(4):579–88, 2003.

[97] B. Lim, T. Heist, M. Levine, and T. Fukaya. Visualization of Transvection in Living *Drosophila* Embryos. *Molecular Cell*, 70(2):287–296.e6, 2018.

[98] B. Lim and M. S. Levine. Enhancer-promoter communication: hubs or loops? *Current Opinion in Genetics and Development*, 67:5–9, 2021.

[99] J. Ling, K. Y. Umezawa, T. Scott, and S. Small. Bicoid-Dependent Activation of the Target Gene hunchback Requires a Two-Motif Sequence Code in a Specific Basal Promoter. *Molecular Cell*, pages 1–10, 2019.

[100] F. Liu and J. W. Posakony. Role of architecture in the function and specificity of two Notch-regulated transcriptional enhancer modules. *PLoS genetics*, 8(7):e1002796, Jul 2012.

[101] J. Liu, R. R. Viales, P. Khoueiry, J. P. Reddington, C. Girardot, E. E. Furlong, and M. Robinson-Rechavi. The hourglass model of evolutionary conservation during embryogenesis extends to developmental enhancers with signatures of positive selection. *bioRxiv*, 2020.

[102] S. E. Lott, J. E. Villalta, G. P. Schroth, S. Luo, L. A. Tonkin, and M. B. Eisen. Noncanonical compensation of zygotic X transcription in early *Drosophila melanogaster* development revealed through single-embryo RNA-Seq. *PLoS Biology*, 9(2), 2011.

[103] L. Luna-Zurita, C. U. Stirnimann, S. Glatt, B. L. Kaynak, S. Thomas, F. Baudin, M. A. H. Samee, D. He, E. M. Small, M. Mileikovsky, A. Nagy, A. K. Holloway, K. S. Pollard, C. W. Müller, and B. G. Bruneau. Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. *Cell*, 164(5):999–1014, 2016.

[104] D. G. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, 2015.

[105] S. MacArthur, X.-Y. Li, J. Li, J. B. Brown, H. C. Chu, L. Zeng, B. P. Grondona, A. Hechmer, L. Simirenko, S. V. E. Keränen, D. W. Knowles, M. Stapleton, P. Bickel, M. D. Biggin, and M. B. Eisen. Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome biology*, 10(7):R80, Jan 2009.

[106] M. Markstein, P. Markstein, V. Markstein, and M. S. Levine. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc Natl Acad Sci U S A*, 99(2):763–768, 2002.

[107] M. Merika and D. Thanos. Enhanceosomes. *Current Opinion in Genetics and Development*, 11(2):205–208, 2001.

[108] L. Minnoye, I. I. Taskiran, D. Mauduit, M. Fazio, L. van Aerschot, G. Hulselmans, V. Christiaens, S. Makhzami, M. Seltenhammer, P. Karras, A. Primot, E. Cadieu, E. van Rooijen, J. C. Marine, G. Egidy, G. E. Ghanem, L. Zon, J. Wouters, and S. Aerts. Cross-species analysis of enhancer logic using deep learning. *Genome Research*, 31(12):1815–1834, 2020.

[109] N. Munshi, T. Agalioti, S. Lomvardas, M. Merika, G. Chen, and D. Thanos. Coordination of a transcriptional switch by HMGI(Y) acetylation. *Science*, 293(5532):1133–1136, 2001.

[110] Y. Nam, P. Sliz, W. S. Pear, J. C. Aster, and S. C. Blacklow. Cooperative assembly of higher-order Notch complexes functions as a switch to induce transcription. *Proceedings of the National Academy of Sciences of the United States of America*, 104(7):2103–2108, 2007.

[111] A. Nasiadka, B. H. Dietrich, and H. M. Krause. Anterior-posterior patterning in the *Drosophila* embryo. *Advances in Developmental Biology and Biochemistry*, 12:155–204, 2002.

[112] N. Nègre, C. D. Brown, P. K. Shah, P. Kheradpour, C. a. Morrison, J. G. Henikoff, X. Feng, K. Ahmad, S. Russell, R. a. H. White, L. Stein, S. Henikoff, M. Kellis, and K. P. White. A comprehensive map of insulator elements for the *Drosophila* genome. *PLoS genetics*, 6(1):e1000814, Jan 2010.

[113] A. S. Nord, M. J. Blow, C. Attanasio, J. A. Akiyama, A. Holt, R. Hosseini, S. Phouanenavong, I. Plajzer-Frick, M. Shoukry, V. Afzal, J. L. Rubenstein, E. M. Rubin, L. A. Pennacchio, and A. Visel. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell*, 155(7):1521–1531, 2013.

[114] U. Ohler, G.-c. Liao, H. Niemann, and G. M. Rubin. Computational analysis of core promoters in the *Drosophila* genome. *Genome biology*, 3(12):RESEARCH0087, 2002.

[115] A. M. Oudelaar, R. A. Beagrie, M. Gosden, S. de Ornellas, E. Georgiades, J. Kerry, D. Hidalgo, J. Carrelha, A. Shivalingam, A. H. El-Sagheer, J. M. Telenius, T. Brown, V. J. Buckle, M. Socolovsky, D. R. Higgs, and J. R. Hughes. Dynamics of the 4D genome during in vivo lineage specification and differentiation. *Nature Communications*, 11(1), 2020.

[116] A. M. Oudelaar, C. L. Harrold, L. L. Hanssen, J. M. Telenius, D. R. Higgs, and J. R. Hughes. A revised model for promoter competition based on multi-way chromatin interactions at the $\alpha$-globin locus. *Nature Communications*, 10(1), 2019.

[117] M. J. Pankratz, M. Busch, M. Hoch, E. Seifert, M. J. Pankratz, M. Busch, M. Hoch, E. Seifert, and H. Jackle. Spatial Control of the Gap Gene *knirps* in the *Drosophila* Embryo by Posterior Morphogen System. *Science*, 255(5047):986–989, 1992.

[118] D. Panne. The enhanceosome. *Current Opinion in Structural Biology*, 18(2):236–242, 2008.

[119] D. Papatsenko, Y. Goltsev, and M. Levine. Organization of developmental enhancers in the *Drosophila* embryo. *Nucleic Acids Research*, 37(17):5665–5677, 2009.

[120] D. Papatsenko and M. Levine. Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the *Drosophila* embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14):4966–4971, 2005.

[121] D. A. Papatsenko, V. J. Makeev, A. P. Lifanov, M. Régnier, A. G. Nazina, and C. Desplan. Extraction of functional binding sites from unique regulatory regions: the *Drosophila* early developmental enhancers. *Genome research*, 12(3):470–81, Mar 2002.

[122] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, 2017.

[123] J. Peccoud and B. Ycart. Markovian modeling of gene-product synthesis, 1995.

[124] F. Pelegri and R. Lehmann. A role of polycomb group genes in the regulation of gap gene expression in *Drosophila*. *Genetics*, 136(4):1341–1353, 1994.

[125] M. W. Perry, A. N. Boettiger, and M. Levine. Multiple enhancers ensure precision of gap gene-expression patterns in the *Drosophila* embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 108(33):13570–13575, Aug 2011.

[126] B. D. Pfeiffer, A. Jenett, A. S. Hammonds, T.-T. B. Ngo, S. Misra, C. Murphy, A. Scully, J. W. Carlson, K. H. Wan, T. R. Laverty, C. Mungall, R. Svirskas, J. T. Kadonaga, C. Q. Doe, M. B. Eisen, S. E. Celniker, and G. M. Rubin. Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America*, 105(28):9715–20, Jul 2008.

[127] V. Pimmett, M. Dejean, C. Fernandez, A. Trullo, E. Betrand, O. Radulescu, and M. Lagha. Quantitative imaging of transcription in living *Drosophila* embryos reveals the impact of core promoter motifs on promoter state dynamics. *bioRxiv*, 2021.

[128] J. Y. Qin, L. Zhang, K. L. Clift, I. Hulur, A. P. Xiang, B. Z. Ren, and B. T. Lahn. Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS ONE*, 5(5):3–6, 2010.

[129] V. Ramalingam, M. Natarajan, J. Johnston, and J. Zeitlinger. TATA and paused promoters active in differentiated tissues have distinct expression characteristics. *Molecular Systems Biology*, 17(2):1–12, 2021.

[130] C. N. J. Ravarani, G. Chalancon, M. Breker, N. S. de Groot, and M. M. Babu. Affinity and competition for TBP are molecular determinants of gene expression noise. *Nature communications*, 7:10417, Feb 2016.

[131] M. Rebeiz, B. Castro, F. Liu, F. Yue, and J. W. Posakony. Ancestral and conserved cis-regulatory architectures in developmental control genes. *Developmental Biology*, 362(2):282–294, 2012.

[132] M. Rebeiz, N. L. Reeves, and J. W. Posakony. SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. *Proc Natl Acad Sci U S A*, 99(15):9888–9893, 2002.

[133] R. Reeves. HMGA proteins: Flexibility finds a nuclear niche? *Biochemistry and Cell Biology*, 81(3):185–195, 2003.

[134] S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, S. Washietl, B. I. Arshinoff, F. Ay, P. E. Meyer, N. Robine, N. L. Washington, L. Di Stefano, E. Berezikov, C. D. Brown, R. Candeias, J. W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M. Y. Tolstorukov, S. Will, A. A. Alekseyenko, C. Artieri, B. W. Booth, A. N. Brooks, Q. Dai, C. A. Davis, M. O. Duff, X. Feng, A. A. Gorchakov, T. Gu, J. G. Henikoff, P. Kapranov, R. Li, H. K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S. K. Powell, N. C. Riddle, A. Sakai, A. Samsonova, J. E. Sandler, Y. B.

Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K. H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger, S. E. Brenner, M. R. Brent, L. Cherbas, S. C. R. Elgin, T. R. Gingeras, R. Grossman, R. A. Hoskins, T. C. Kaufman, W. Kent, M. I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J. W. Posakony, B. Ren, S. Russell, P. Cherbas, B. R. Graveley, S. Lewis, G. Micklem, B. Oliver, P. J. Park, S. E. Celniker, S. Henikoff, G. H. Karpen, E. C. Lai, D. M. MacAlpine, L. D. Stein, K. P. White, and M. Kellis. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science (New York, N.Y.)*, 330(6012):1787–97, Dec 2010.

[135] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113, Sep 2012.

[136] U. Schibler and F. Sierra. Alternative Promoters in Developmental Gene Expression. *Annual Review of Genetics*, 21(1):237–257, Dec 1987.

[137] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *Journal of molecular biology*, 188(3):415–431, 1986.

[138] C. Schröder, D. Tautz, E. Seifert, and H. Jäckle. Differential regulation of the two transcripts from the *Drosophila* gap segmentation gene hunchback. *The EMBO journal*, 7(9):2881–7, 1988.

[139] M. D. Schroeder, M. Pearce, J. Fak, H. Q. Fan, U. Unnerstall, E. Emberly, N. Rajewsky, E. D. Siggia, and U. Gaul. Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biology*, 2(9):E271, Sep 2004.

[140] K. H. Scully, E. M. Jacobson, K. Jepsen, V. Lunyak, H. Viadiu, C. Carriere, D. W. Rose, F. Hooshmand, A. K. Aggarwal, and M. G. Rosenfeld. Allosteric effects of Pit-1 DNA sites on long-term repression in cell type specification. *Science*, 290(5494):1127–1131, 2000.

[141] A. Senecal, B. Munsky, F. Proux, N. Ly, F. E. Braye, C. Zimmer, F. Mueller, and X. Darzacq. Transcription factors modulate c-Fos transcriptional bursts. *Cell Reports*, 8(1):75–83, 2014.

[142] K. Senger, G. W. Armstrong, W. J. Rowell, J. M. Kwan, M. Markstein, and M. Levine. Immunity Regulatory DNAs Share Common Organizational Features in *Drosophila*. *Molecular Cell*, 13(1):19–32, 2004.

[143] H. Shao, M. Revach, S. Moshonov, Y. Tzuman, K. Gazit, S. Albeck, T. Unger, and R. Dikstein. Core Promoter Binding by Histone-Like TAF Complexes. *Molecular and Cellular Biology*, 25(1):206–219, 2005.

[144] E. Sharon, D. Van Dijk, Y. Kalma, L. Keren, O. Manor, Z. Yakhini, and E. Segal. Probing the effect of promoters on noise in gene expression using thousands of designed sequences. *Genome Research*, 24(10):1698–1706, 2014.

[145] A. Sloutskin, Y. M. Danino, Y. Orenstein, Y. Zehavi, T. Doniger, R. Shamir, and T. Juven-Gershon. ElemeNT: a computational tool for detecting core promoter elements. *Transcription*, 6(3):41–50, 2015.

[146] M. Spielmann, D. G. Lupiáñez, and S. Mundlos. Structural variation in the 3D genome. *Nature Reviews Genetics*, 19(7):453–467, 2018.

[147] G. D. Stormo. DNA Motif Databases and Their Uses. In *Current Protocols in Bioinformatics*, pages 2.15.1–2.15.6. John Wiley & Sons, Inc., Hoboken, NJ, USA, Sep 2015.

[148] G. D. Stormo and D. S. Fields. Specificity, free energy and information content in protein-DNA interactions. *Trends in Biochemical Sciences*, 23(3):109–113, 1998.

[149] C. I. Swanson, N. C. Evans, and S. Barolo. Structural Rules and Complex Regulatory Circuitry Constrain Expression of a Notch- and EGFR-Regulated Eye Enhancer. *Developmental Cell*, 18(3):359–370, Mar 2010.

[150] D. Thanos and T. Maniatis. Virus induction of human IFN-beta gene expression requires the assembly of an enhanceosome. *Cell*, 83(7):1091–1100, 1995.

[151] M. C. Thomas and C. M. Chiang. The general transcription machinery and general cofactors. *Critical reviews in biochemistry and molecular biology*, 41(3):105–178, 2006.

[152] S. Thomas, X.-Y. Li, P. J. Sabo, R. Sandstrom, R. E. Thurman, T. K. Canfield, E. Giste, W. Fisher, A. Hammonds, S. E. Celniker, M. D. Biggin, and J. a. Stamatoyannopoulos. Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. *Genome Biology*, 12(5):R43, 2011.

[153] I. Tirosh and N. Barkai. Two strategies for gene regulation by promoter nucleosomes. *Genome Research*, 18(7):1084–1091, 2008.

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

[155] P. Tomancak, B. P. Berman, A. Beaton, R. Weiszmann, E. Kwan, V. Hartenstein, S. E. Celniker, and G. M. Rubin. Global analysis of patterns of gene expression during *Drosophila* embryogenesis. *Genome biology*, 8(7):R145, 2007.

[156] L. Tora. A unified nomenclature for TATA box binding protein (TBP)-associated factors (TAFs) involved in RNA polymerase II transcription. *Genes and Development*, 16(6):673–675, Mar 2002.

[157] A. Tsai, A. K. Muthusamy, M. R. Alves, L. D. Lavis, R. H. Singer, D. L. Stern, and J. Crocker. Nuclear microenvironments modulate transcription from low-affinity enhancers. *eLife*, 6:e28975, Nov 2017.

[158] P. F. Tsai, S. Dell'Orso, J. Rodriguez, K. O. Vivanco, K. D. Ko, K. Jiang, A. H. Juan, A. A. Sarshad, L. Vian, M. Tran, D. Wangsa, A. H. Wang, J. Perovanovic, D. Anastasakis, E. Ralston, T. Ried, H. W. Sun, M. Hafner, D. R. Larson, and V. Sartorelli. A Muscle-Specific Enhancer RNA Mediates Cohesin Recruitment and Regulates Transcription In *trans*. *Molecular Cell*, 71(1):129–141.e8, 2018.

[159] E. Tunnacliffe and J. R. Chubb. What Is a Transcriptional Burst? *Trends in Genetics*, 36(4):288–297, 2020.

[160] J. van Arensbergen, B. van Steensel, and H. J. Bussemaker. In search of the determinants of enhancer-promoter interaction specificity. *Trends in Cell Biology*, 24(11):695–702, 2014.

[161] G. van den Engh, R. Sachs, and B. J. Trask. Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. *Science (New York, N.Y.)*, 257(5075):1410–2, Sep 1992.

[162] L. Vo Ngoc, Y.-L. Wang, G. A. Kassavetis, and J. T. Kadonaga. The punctilious RNA polymerase II core promoter. *Genes & development*, 31(13):1289–1301, 2017.

[163] X. Wang, J. Hou, C. Quedenau, and W. Chen. Pervasive isoformspecific translational regulation via alternative transcription start sites in mammals. *Molecular Systems Biology*, 12(7):875, 2016.

[164] Y. L. Wang, S. H. Duttke, K. Chen, J. Johnston, G. A. Kassavetis, J. Zeitlinger, and J. T. Kadonaga. TRF2, but not TBP, mediates the transcription of ribosomal protein genes. *Genes and Development*, 28(14):1550–1555, 2014.

[165] W. W. Wasserman and J. W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *Journal of molecular biology*, 278(1):167–181, 1998.

[166] R. Waymack, A. Fletcher, G. Enciso, and Z. Wunderlich. Shadow enhancers can suppress input transcription factor noise through distinct regulatory logic. *eLife*, 9:1–57, 2020.

[167] J. wen Yin, Y. Liang, J. Y. Park, D. Chen, X. Yao, Q. Xiao, Z. Liu, B. Jiang, Y. Fu, M. Bao, Y. Huang, Y. Liu, J. Yan, M. sheng Zhu, Z. Yang, P. Gao, B. Tian, D. Li, and G. Wang. Mediator MED23 plays opposing roles in directing smooth muscle cell and adipocyte differentiation. *Genes and Development*, 26(19):2192–2205, 2012.

[168] D. Wichadakul, J. McDermott, and R. Samudrala. Prediction and integration of regulatory and protein-protein interactions. *Methods in molecular biology (Clifton, N.J.)*, 541:101–43, Jan 2009.

[169] M. Wijgerde, F. Grosveld, and P. Fraser. Transcription complex stability and chromatin dynamics *in vivo*. *Nature*, 377(6546):209–213, 1995.

[170] C.-H. H. Wu, L. Madabusi, H. Nishioka, P. Emanuel, M. Sypes, I. Arkhipova, and D. S. Gilmour. Analysis of Core Promoter Sequences Located Downstream from the TATA Element in the. *Society*, 21(5):1593–1602, 2001.

[171] S. Wu, K. Li, Y. Li, T. Zhao, T. Li, Y. F. Yang, and W. Qian. Independent regulation of gene expression level and noise by histone modifications. *PLoS Computational Biology*, 13(6):1–27, 2017.

[172] Z. Wunderlich, M. D. Bragdon, K. B. Eckenrode, T. Lydiard-Martin, S. Pearl-Waserman, and A. H. Depace. Dissecting sources of quantitative gene expression pattern divergence between *Drosophila* species. *Molecular Systems Biology*, 8(604):604, 2012.

[173] Z. Wunderlich and L. A. Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in genetics : TIG*, 25(10):434–440, 2009.

[174] M. Xu, P. Sharma, S. Pan, S. Malik, R. G. Roeder, and E. Martinez. Core promoter-selective function of HMGA1 and Mediator in Initiator-dependent transcription. *Genes and Development*, 25(23):2513–2524, 2011.

[175] J. O. Yáñez-Cuna, E. Z. Kvon, and A. Stark. Deciphering the transcriptional *cis*-regulatory code. *Trends in genetics : TIG*, 29(1):11–22, Jan 2013.

[176] C. Yang, E. Bolotin, T. Jiang, F. M. Sladek, and E. Martinez. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, 389(1):52–65, 2007.

[177] M. Yokoshi, M. Cambón, and T. Fukaya. Regulation of transcriptional bursting by core promoter elements in the *Drosophila* embryo. *bioRxiv*, 2021.

[178] M. a. Zabidi, C. D. Arnold, K. Schernhuber, M. Pagani, M. Rath, O. Frank, and A. Stark. Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, 518(7540):556–559, 2014.

[179] M. A. Zabidi and A. Stark. Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors. *Trends in Genetics*, 32(12):801–814, 2016.

[180] Y. Zehavi, O. Kuznetsov, A. Ovadia-Shochat, and T. Juven-Gershon. Core Promoter Functions in the Regulation of Gene Expression of *Drosophila* Dorsal Target Genes. *Journal of Biological Chemistry*, 289(17):11993–12004, 2014.

[181] R. Zhou. SOX9 interacts with a component of the human thyroid hormone receptor-associated protein complex. *Nucleic Acids Research*, 30(14):3245–3252, 2002.

[182] L. J. Zhu, R. G. Christensen, M. Kazemian, C. J. Hull, M. S. Enuameh, M. D. Basciotta, J. a. Brasefield, C. Zhu, Y. Asriyan, D. S. Lapointe, S. Sinha, S. a. Wolfe, and M. H. Brodsky. FlyFactorSurvey: A database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Research*, 39(SUPPL. 1):111–117, 2011.

[183] B. Zoller, S. C. Little, and T. Gregor. Diverse Spatial Expression Patterns Emerge from Unified Kinetics of Transcriptional Bursting. *Cell*, 175(3):835–847.e25, 2018.