Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by

M.Sc. Gilberto Alejandro Alvarez Canales

Born in: Guanajuato, Mexico

Oral examination: 12th of September of 2024

Developmental constraints and the
regulatory logic of *Drosophila* enhancers

Referees:

Prof. Dr. Michael Knop

Dr. Alexander Aulehla

Summary

Enhancers can generate specific patterns of gene expression, which are essential for organismal development. Despite decades of research, how enhancers encode the information necessary to drive these precise patterns remains unclear. To address this question, in this project, I focused on the interplay of two different approaches: 1) Using well-known endogenous enhancers, I searched for putative missing regulatory elements by integrating information from different mutational screenings, and 2) Using the combination of synthetic enhancers expression profiles and thermodynamic models, I tested the possibility of several mechanisms of their associated Transcription Factors (TFs). In the first approach, I focused on two of the best-characterized animal enhancers, the minimal stripe 2 enhancer (MS2E) and the E3N enhancer. The MS2E enhancer can generate the second of the seven stripes of the *Even-skipped* gene expression pattern. The E3N enhancer regulates the expression of the *Shavenbaby* gene, which has an eight-striped pattern. For the MS2E enhancer, a systematic and targeted mutagenesis screening was performed and analyzed, while for the E3N enhancer, I analyzed data from a randomized mutagenesis approach. For both enhancers, I estimated and associated affinity changes of multiple TF binding sites (TFBSs) with the observed expression phenotypes. A dense encoded architecture was observed for the two enhancers where almost each mutated section generates a phenotype that differs from the wild-type. Affinity changes and motif turnover analyses of known TFBSs could explain only a small fraction of these phenotypes. Finally, I explored experimental alternatives for finding additional associated TFs and the evolutionary implications of a dense encoded enhancer architecture. These results suggest the need to find additional regulators and improve current versions of the binding profiles of already known TFs. In the second approach, I evaluated the roles of early patterning TFs through synthetic enhancers and mechanistic modeling. I observed that early embryonic enhancers are associated with higher information at the sequence level composition than later and synthetic random enhancers. Additionally, the synthetic system could encode different sharp gene expression patterns with different combinations of known TFBSs. The results from implementing generalized thermodynamic models suggested that these TFs are highly context-dependent. This analysis suggested multiple equally performing mechanisms for different possibilities for TF-TF interactions, TF function, and modes of regulation. These models could predict observed expression patterns, such as a broad stripe in the center of the embryo, but other expression patterns, such as the presence of additional anterior expression, could not be explained. Finally, I propose alternative mechanisms by which known TFs work in this context and new directions for enhancer design to expand the understanding of encoding developmental patterns.

Zusammenfassung

Enhancer können spezifische Muster der Genexpression erzeugen, die für die Entwicklung von Organismen unerlässlich sind. Trotz jahrzehntelanger Forschung ist noch immer unklar, wie Enhancer die Informationen kodieren, die für die Erzeugung dieser präzisen Muster erforderlich sind. Um diese Frage zu beantworten, habe ich mich in diesem Projekt auf das Zusammenspiel zweier unterschiedlicher Ansätze konzentriert: 1) Unter Verwendung bekannter endogener Enhancer habe ich nach mutmaßlich fehlenden regulatorischen Elementen gesucht, indem ich Informationen aus verschiedenen Mutationsscreenings integriert habe, und 2) Unter Verwendung der Kombination von Expressionsprofilen synthetischer Enhancer und thermodynamischen Modellen habe ich die zugehörigen Transkriptionsfaktoren (TFs) auf die Möglichkeit mehrerer Mechanismen hin getestet. Im ersten Ansatz habe ich mich auf zwei der am besten charakterisierten tierischen Enhancer konzentriert, den Minimal Stripe 2 Enhancer (MS2E) und den E3N Enhancer. Der MS2E Enhancer kann den zweiten der sieben Streifen des *Even-skipped*-Genexpressionsmusters erzeugen. Der E3N Enhancer reguliert die Expression des *Shavenbaby*-Gens, das ein acht-Streifen-Muster aufweist. Für den MS2E-Enhancer wurde ein systematisches und gezieltes Mutagenese-Screening durchgeführt und analysiert, während ich für den E3N-Enhancer Daten aus einem randomisierten Mutagenese-Ansatz analysierte. Für beide Enhancer schätzte ich Affinitätsänderungen mehrerer TF-Bindungsstellen (TFBSs) und assoziierte sie mit den beobachteten Expressionsphänotypen. Für die beiden Enhancer wurde eine dichte codierte Architektur beobachtet, bei der fast jeder mutierte Abschnitt einen Phänotyp erzeugt, der sich vom Wildtyp unterscheidet. Affinitätsänderungen und Motivumsatzanalysen bekannter TFBSs konnten nur einen kleinen Teil dieser Phänotypen erklären. Schließlich untersuchte ich experimentelle Alternativen zum Auffinden zusätzlicher assoziierter TFs und die evolutionären Auswirkungen einer dichten codierten Enhancer-Architektur. Diese Ergebnisse legen die Notwendigkeit nahe, zusätzliche Regulatoren zu finden und aktuelle Versionen der Bindungsprofile bereits bekannter TFs zu verbessern. Im zweiten Ansatz habe ich die Rolle früher musterbildender TFs durch synthetische Enhancer und mechanistische Modellierung untersucht. Ich habe beobachtet, dass frühe embryonale Enhancer mit mehr Informationen auf Sequenzebene verbunden sind als spätere oder zufällige synthetische Enhancer. Darüber hinaus konnte das synthetische System verschiedene scharfe Genexpressionsmuster mit verschiedenen Kombinationen bekannter TFBSs kodieren. Die Ergebnisse aus der Implementierung verallgemeinerter thermodynamischer Modelle deuteten darauf hin, dass diese TFs stark kontextabhängig

sind. Diese Analyse deutete auf mehrere gleich funktionierende Mechanismen für verschiedene Möglichkeiten von TF-TF-Interaktionen, TF-Funktionen und Regulierungsarten hin. Diese Modelle konnten beobachtete Expressionsmuster vorhersagen, wie etwa einen breiten Streifen in der Mitte des Embryos, aber andere Expressionsmuster, wie etwa das Vorhandensein zusätzlicher anteriorer Expression, konnten nicht erklärt werden. Abschließend schlage ich alternative Mechanismen vor, mit denen bekannte TFs in diesem Zusammenhang funktionieren, und neue Richtungen für das Enhancer-Design, um das Verständnis der Kodierung von Entwicklungsmustern zu erweitern.

Acknowledgments

First, I would like to thank you, Justin, for allowing me to do very exciting scientific projects in your lab at the intersection of theory and experimental biology. Your lab is one of those places where the imagination is the only limit, always bringing inspiration to go beyond details. Also, this experience allowed me to forge a more complete mindset on landing ideas into the experimental reality. Thanks for always finding ways to keep an exciting and motivating atmosphere in the lab.

During my long lab stay, I have not only met great labmates and colleagues, but also I have made great friends. Thank you, Rafael, for being a great mentor and friend. My stay here wouldn't have been the same without you. Sharing our spark of scientific curiosity and complementarity in our approaches was a great combination that has shaped my way of doing science. You are a very inspiring scientist, and I´m happy we worked together during our time here. Thank you, Mariana; you made a very welcoming start to my PhD. Your advice and support were always helpful, and I enjoyed all our cultural adventures a lot. Thanks, Albert, Tim, and Kerstin, for all the academic and non-academic support; I learned much from you about our Drosophila friends and microscopy. The projects we had together were very enriching. Thank you so much, Xueying, Lautaro, Noa, Natalia, Blanca, Paco, and Anna. You have been an amazing support and company for the last few years. It is always nice to talk about science and society with you. Our common interests always bring very nice conversation topics and ideas for projects. I am looking forward to hearing about your future projects, too. Mindy, thanks a lot for all the advice and supervision on the mathematical models; this was an essential part of my project. I really enjoyed talking with you about the extent of mathematical applications in Biology. Thanks, Tin and Marlize. We had a great time during my PhD start, and I will never forget our adventures. Esther, thanks also for the time we shared in our collaboration. It was great to work with you and with the amazing expression patterns you found; I hope you had a great time working with Rafael and me. Vani, Matteo, Gulina, and Johaness, thanks a lot for being a great company in the lab, and I hope you had a great time in the group.

Thank you, to my inspiring family, who have always supported me in life and provided us with a curiosity rich environment for science. Gracias a mi papá,José de Jesús, por transferirnos ese amor a la naturaleza que tienes. Definitivamente, todas esas experiencias para encontrar especies de aves y fósiles, forjaron mi pasión por la Biología. Gracias a mi mamá, Luz María, que siempre nos apoya y apoyó desde chicos para tener las herramientas y conocimientos necesarios en la escuela

y en la vida. A mis hermanos, Pepe, Gerardo y Mildred, que al crecer en este ambiente familiar, desarrollamos diversas curiosidades, ya sea por el universo, la música y cultura. Siendo yo el mas joven, definitavemente mucho lo aprendí de ustedes. I am so thankful to count with you in all circumstances. I would like to thank my friends that I made here in Heidelberg for being always supportive and all our amazing experiences we had together. We have been an amazing group of friends and I will always carry you in my heart. Thanks a lot Jesus, Sebas, Anna Lippert, Beto, Karo, Anna Mathioudaki, Anita, Lucia, Felix, Andrea, Hendrik, Aline, Max, Fynn, Carolina, Wolfy, Fergus, Maxime, Alex, Javi, Agata, Kevin, Marina, Elisa and the rest of our predoc friends. It was an incredible cohort. I hope we will enjoy more conferences, dancing parties, mario karty and running events together. Thanks a lot also Olga, for all the great adventures and support in all aspects, definitely my period in Heidelberg was very nice thanks to you.

I want to thank my TAC members, Michael Knop, James Sharpe, and Alexander Aulehla. All of you have been very supportive of my project and my life. I also enjoyed our meetings with the mixture of different areas of expertise. Thanks a lot for providing feedback on my projects. Also, I want to thank Hernan and Eva Geissen for all the advice and feedback on the mathematical models.

Contents

List of Figures

1 Introduction

1.1 The search for the mechanisms behind pattern formation

The emergence of form during animal embryo development was a question that puzzled early naturalists from different ancient cultures around the world. Embryological descriptions can be found in religious and medical texts from ancient India and China from centuries BCE. The most accurate ancient description of human embryonic development is written in an ancient Indian text named *Garbhāvakrāntisūtra* whose explanations match current standards of stage classification. The first representations of an embryo seem to be located in the western hemisphere where several Mesoamerican cultures have sculpted human embryos, some dating back from before the 1st millennium BCE in the valley of Oaxaca (Marcus 1998) (Tate 1999) (Wallingford 2021). In ancient Greece, Aristotle's work named *On the Generation of Animals* established the main current of thinking in Europe for organismal development for more than a millennium (Barresi and Gilbert 2020).

Aristotle's work at the Bay of Kalloni precisely described the form and the order of appearance of organs in different animal embryos. These observations led him to question whether developmental processes happen simultaneously or sequentially. One of his most remarkable ideas, which I will cover in my thesis, is about a mechanism behind the information that determines the form, which he called Eidos. Back then, there were no tools to explore the biochemical nature behind the forms of animals, and he made several metaphysical inferences of this mechanism based on elements and movements (Aristotle ca. 350 BCE).

During the next centuries, anatomical descriptions became more precise and included more animal species. The invention of the microscope allowed more detailed observations of the anatomical structures, but this knowledge still lacked the capacity to dissect mechanisms for development. It was not until the second half of the XIX century that combining knowledge in Biochemistry and Experimental Embryology made searching for the molecules responsible for structure formation possible.

From the last part of the 1800s, Roux, Driesch, and Morgan's experiments showed the capacity to manipulate specific cells inside frog and sea urchin embryos. Then, it became clear that one of the next tasks was identifying which embryonic material could determine form. This was achieved through embryo transplantation experiments by Hans Spemmann and Hilde Mangold, who found

a structure capable of determining form, which they named "organizer". An organizer can change the fate of other neighboring cells when transplanted (Witkowski 1985) (J. Green 2002).

In the subsequent decades, multiple molecular signals for Developmental processes were identified, shifting the field towards a biochemical nature of pathways as the causal elements for morphogenesis. This was complemented by the idea of a regulatory logic that controls gene expression, and that this logic would produce different expression programs in different tissues (Jacob and Monod 1961) (Britten and E H Davidson 1969) Thanks to decades of molecular characterization of these regulatory programs today it is known that multicellular organisms rely on the differential regulation of multiple genes in space and time. These regulatory programs execute essential morphogenetic processes such as proliferation and tissue differentiation. Gene regulation works in a network-like manner where multiple interactions can happen simultaneously, and feedback plays a role. Gene regulatory networks have been characterized for different examples of animal morphogenesis, such as segmentation, organogenesis, and skin and coat patterning (Eric H Davidson 2006) (McGinnis 2005).

Nowadays, the signals responsible for asymmetries in different developmental processes are still being identified, isolated, and manipulated. More evidence is being gathered that the signals are not just biochemical but also they can be mechanical. Changes in morphology are inherently related to changes in the mechanical forces of a system. These forces can be internal and can be generated by the cytoskeleton. Hydric forces can be related to changes in the osmotic pressure and volume. Additionally, external forces come from neighboring tissues and other environmental factors. All these observations highlight the need to study morphogenesis in a multiscale manner (Maroudas-Sacks and Keren 2021).

Besides the substantial characterization of the causal agents of pattern formation, some scientists were worried that the mere presence of chemical and biophysical agents does not guarantee the appearance of a pattern. To create a pattern, uniformity needs to be affected by certain necessary information. In 1952, departing from a mathematical perspective, Alan Turing coined the term "Morphogen" to refer to these molecules that could generate a pattern. According to his mathematical model, the condition for a pattern emergence implies the existence of a mathematical instability that arises through reaction and diffusion in the chemical system he proposed. Not long after this proposition of structure formation, known as self-organization, Lewis Wolpert formulated a different mechanism for pattern formation called Positional Information. This mechanism was created as a unify-

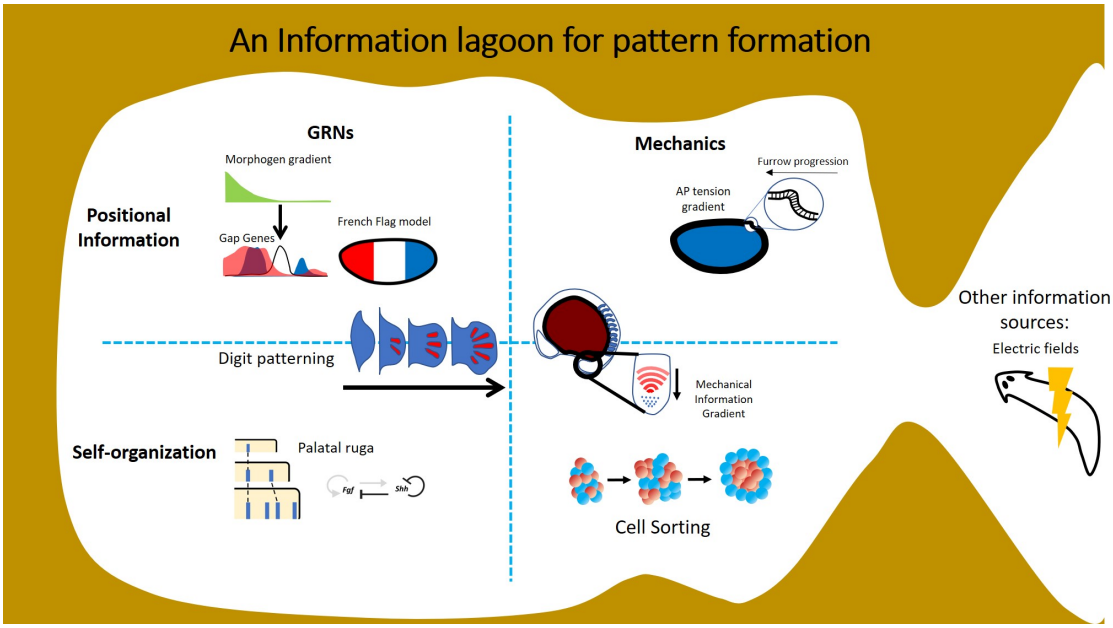


Figure 1: The nature of Information for Pattern formation. Taking inspiration from Aristotle's ideas from his work on the Bay of Kalloni, I represented the mechanisms for pattern formation as a Lagoon. The inside of the lagoon represents our current knowledge of mechanisms, such as the ones with a biochemical and mechanical nature. Coastal lagoons are often open and connected to the sea and oceans; this openness also applies to getting new knowledge; for example, electromagnetism seems to include important information for positioning in developmental processes based on observations in planaria (Levin, Pietak, and Bischof 2019). I created this schematic based on figures from (Cerchiari et al. 2015) (J. B. A. Green and Sharpe 2015), (Genuth and Holley 2020)

ing approach for the notions of gradients from the early embryonic experiments with Organizers and the early 1900s' ideas of morphogenetic fields. Positional information establishes that cells can get their position specified by a spatial field of chemicals within a reference system. (J. B. A. Green and Sharpe 2015) (De Robertis 2006) (Turing 1952).

In summary, the signals that control animal development can emerge under special circumstances from 2 main mechanisms: 1) the time evolution of the components of a system by a process known as Self-Organization, and 2) these signals can be specified in determined positions of the embryo by the mechanism of Positional information. It is important to note that these mechanisms are not mutually exclusive and can occur simultaneously in a system (J. B. A. Green and Sharpe

2015), as I show in Figure 1.

1.2 The complexity of gene regulation across the domains of life

Bacteria, Archaea, and Eukaryotic genes are expressed through the processes of transcription and translation. In this work, I will focus on the initiation of the transcriptional process, which controls the rate of mRNA synthesis. The RNA polymerase synthesizes an mRNA molecule departing from a DNA sequence. The initiation of this process is regulated by cis-regulatory elements such as promoters. Promoters have sequences where proteins can bind, such as the TATA-binding protein, which can recruit the RNA polymerase for transcription. In these cis-regulatory modules, a class of proteins called Transcription Factors (TFs) can bind certain sequences and regulate gene expression. Cis-regulatory elements that control gene expression can be near the promoter where the polymerase binds or be thousands of base pairs away (Craig et al. 2021).

TFs can control gene expression by activating or inhibiting gene expression. These roles are determined by how a TF allows the Polymerase and the basal transcriptional machinery to be assembled and affect the residence time, among other mechanisms. TFs can regulate a gene through direct chemical interactions with the eukaryotic polymerase, Pol II. However, indirect interactions also exist, such as affecting neighboring TFs or modifying the surrounding DNA landscape to avoid the binding of other molecules (Chen and Pugh 2021). Interestingly, TFs binding profiles can have different specificities that depend on the domain of life. In bacteria, a TF has, on average, 23 bits of information to recognize a site, which is enough for finding a unique region in their genomes. In Eukaryotes, the average TF recognition is 12 bits of information, which is insufficient to direct a unique location in their larger genomes. TFs can control gene expression for different tasks, such as housekeeping processes, or as a way to respond to environmental changes and developmental processes (Wunderlich and Mirny 2009).

Archaea and Eukaryotic genomes are covered by chromatin, which is made of packed DNA by proteins called histones. Compacted chromatin regions avoid genome-wide gene expression by limiting the binding of the transcriptional machinery. Recently, it has been shown that some species of Bacteria also have histones with a different interaction mechanism with DNA, although their function is still under research. The regulation of the compaction of chromatin plays a significant role in the control of gene expression in eukaryotes. Some of these

functions are associated with activating or repressing specific pathways under different environmental and cellular conditions. Post-translational modifications in the histones and chemical marks in DNA, such as methylation, can modulate the chromatin compaction. As a result, chromatin accessibility landscapes are very different across tissues (Craig et al. 2021) (Hocher et al. 2023).

Different proteins can regulate the accessibility of compacted chromatin regions. One class of these regulators is called Pioneer factors, which are Transcription factors that can interact with the heterochromatin and make the regulatory DNA accessible. This process is known as chromatin remodeling and is mediated by different enzymes. Once the chromatin is accessible, other Transcription Factors can bind to the cis-regulatory modules and regulate gene expression. The function of Pioneer factors is essential for developmental processes, regeneration, and mechanisms of diseases where differentiation processes play a role (González-Sastre et al. 2017) (Balsalobre and Drouin 2022).

Once the Pol-II has synthesized the mRNA, it can degrade in different ways depending on the presence of specific features in that context; for example, the length of the polyA tail is linked with mRNA stability. Another example is the stabilization of mRNA mediated by the Hu proteins during the development of the nervous system. mRNA processing includes differential splicing, which makes different isoforms of an mRNA. The translation process can also be regulated; for example, some small RNAs, such as miRNAs, can inhibit translation. After translation, proteins can be degraded differentially in several contexts based on post-translational modifications or the stability of a determined protein (Barresi and Gilbert 2020) (Oliveto et al. 2017).

1.3 Gene regulation in Multicellular systems

All the previously mentioned regulatory mechanisms happen inside a multicellular organism where differential gene expression happens in space and time. Differential gene expression is essential for the development and physiology of organisms, and it is known that different tissues exhibit particular transcriptional profiles that are highly evolutionarily conserved (Brawand et al. 2011). These gene expression patterns can determine the position and time where physical processes such as segmentation, folding, organogenesis, differentiation, and tissue functionality can happen. In addition to the promoters, there are cis-regulatory sequences, known as Enhancers, that provide the necessary information to express a gene in a determined space and time (Barresi and Gilbert 2020).

The importance of cis-regulatory elements is highlighted through an evolutionary scope. The cis-regulatory hypothesis establishes that cis-regulatory elements are more likely to be under evolutionary pressure because mutational effects on these sequences will have fewer consequences. This is supported by the high conservation of proteins, in which mutational effects can have larger undesired effects. In contrast, most of the genome variation comes from intergenic regions. Additionally, enhancers are a main source of evolutionary change since they are more variable between species than promoters (Carroll 2008) (Wittkopp and Kalay 2011)

Long-range interactions in animals have been documented as essential to control gene expression. These events can happen thanks to the 4D arrangement of the genome and the creation of transcriptional hubs. There are different models in which enhancers can interact with promoters depending on the distance, such as the looping, tracking, and linking mechanisms. This diversity of mechanisms is context-dependent and adds another layer of complexity to the biology of gene regulation (Furlong and Michael Levine 2018) (Schoenfelder and Fraser 2019).

Different processes, such as diffusion and direct transport, affect the spatial distribution of regulatory signals. These processes are relevant in a multicellular context since they can break or generate symmetry. Diffusion can either homogenize the concentration of a given signaling molecule or, if coupled with chemical reactions in a network, can lead to patterning. Many currently known examples of patterning can be explained by the diffusion of morphogens. Still, there is increasing evidence that transport through structures like cytonemes can be important for morphogenesis, too. These signals will regulate gene expression in different ways depending on their concentrations in different parts of an organism (Turing 1952) (Muller et al. 2013) (Durrieu et al. 2018) (Hall et al. 2023).

More recently, evidence shows that mechanical forces also influence development. Certain groups of cells and their proliferation can exert forces in different tissue regions that can activate different functional programs. Inside the cell, active stresses can direct different behaviors that collectively influence tissue morphogenesis. This means mechanical processes can interplay and influence gene expression programs regulated during morphogenesis. Properties such as elasticity, friction, and viscosity can generate gradients or self-organize, generating the driving force for patterning. Thus, morphogenesis can be encoded by genetic material and mechanical forces (Maroudas-Sacks and Keren 2021).

1.4 The elements behind Positional information and Self-organization.

During the second half of the XX century, works in the Fruitfly, Sea Urchin, Frog, and Hydra started to uncover the interactions among morphogens and genes essential for development. These discoveries allowed the proposal and mapping of developmental gene regulatory networks. With this evidence, positional information and early heterogeneity mechanisms have been prevalent as the main drivers of pattern formation. Nowadays, there are many beautiful examples of biological patterning where their gene networks have been at least partially characterized, such as the hydra body plan, the eyespots for the butterfly wings, somitogenesis in vertebrates, pigmentation in fishes and wildcats, and body segmentation, wing and eye development in the Fruitfly (Driever and Nüsslein-Volhard 1988) (McGinnis 2005) (Carroll 2008).

When Alan Turing formulated his theory, he proposed a simple 2-node chemical network that could produce instability in a homogeneous chemical system. This phenomenon could be interpreted as the generation of a pattern. Certain parameters and conditions must be fulfilled for such a system to generate a pattern. Even though the parameters regime has been criticized for being biologically implausible, biological systems with morphogen gradients and GRNs in multicellular organisms have complex architectures that could make a Turing-like mechanism plausible. Larger networks allow a broader range of dynamic behaviors that could drive morphogenetic processes. Some TFs are interconnected in complex networks involving feedback. This feedback provides interesting dynamical behaviors such as robustness, oscillations, and spatial instabilities such as the ones required for the patterns proposed by Turing. Several biological systems have been suggested to be controlled by Turing-like self-organization processes, such as the palatal ridge and digit patterning development in mice (Turing 1952) (J. B. A. Green and Sharpe 2015).

In *Drosophila melanogaster*, the Gap genes network is one of the most studied examples of how a biochemical system drives pattern formation. Different sections from these networks have been suggested to have different dynamical behaviors, such as bi-stability and oscillations. For many years, this system inspired the search for self-organization and positional information mechanisms for development (Jaeger 2011) (J. B. A. Green and Sharpe 2015).

Different network architectures can be manipulated in a multicellular system, which can be used to identify the mechanism behind a pattern. The work from Raspopovic et al. is an example of the successful identification of components from

a putative Turing network involved in digit patterning. These components include Sox9, Bmp, and Wnt regulatory pathways identified in mouse limb buds, which are responsible for the position where digits will appear. Additional examples have shown the role of mutual repression in sharp boundaries in the Dorsal-Ventral neural tube development and the anterior-posterior development of Drosophila. Other examples show the possibility of genetic oscillators for gene patterning in the early embryo from Tribolium, Drosophila melanogaster, and in the presomitic mesoderm of vertebrates (Tsiairis and Aulehla 2016) (Verd et al. 2018).

This historical overview I have put here shows that theoretical approaches can often propose the mechanisms for pattern formation (Britten and E H Davidson 1969) (Turing 1952). The complex regulatory network knowledge nowadays makes it difficult to pinpoint a mechanism for a given pattern due to many parameters lacking experimental measurements. Additionally, the system's delimitation is often difficult since many elements are still unknown in a regulatory network. From the mathematical perspective, using differential equations of variables that represent chemical concentrations has helped to predict certain regulatory systems' dynamics (Jaeger et al. 2004). These models can be approached by including spatial variables like the reaction-diffusion physical models. But even theoretically, given their complexity, many of these models are limited in their analytical solutions and can only be approached with simulations that will depend on more assumptions. Nonetheless, this hasn´t limited scientists, and many of these approaches keep inspiring the identification of mechanisms with new experimental data.

Applying these approaches in a biological context has many experimental limitations, too. For example, identifying certain parameters might require broad mutagenesis screenings for each system's components. This will require implementing high throughput approaches for multicellular systems. Another difficulty includes the lack of the right level of coarse-graining since some molecular events that were not considered relevant in reality have an influence, and current techniques might not be able to observe them.

One of the solutions to reduce the complexity of endogenous systems would be to encode pattern formation through synthetic networks. These networks can be controlled inside a multicellular organism, and the chances of interaction with other endogenous elements can be reduced. This will be done first by understanding how to make targeted synthetic enhancers and synthetic TFs. Then, using this approach, it will be possible to integrate synthetic enhancers with their respective synthetic Transcription Factors. This coupling can help to achieve and test

pattern formation questions in a specific time and space with the right regulatory parameters. In summary, future advances in synthetic pattern formation will be accomplished by combining knowledge from theoretical and endogenous network architectures and a characterization of cis-regulatory modules.

1.5 The Drosophila embryo as a model for Pattern formation

Drosophila melanogaster is one of the best-understood model organisms in Biology. Its developmental processes and genetics have been deeply described for over a century. In this model, as in most insects, there is superficial cleavage and nuclear divisions before cell formation. During the early embryogenesis of *Drosophila melanogaster*, a series of nuclear division cycles occur first with a short uniform duration of 8 min per cycle per cell. After this, the nuclei start moving to the surface, and their division becomes slower. The new conformation receives the name of the syncytial blastoderm, and additional nuclear divisions still occur. Cellularization occurs after the 13th nuclear cycle, giving origin to the cellular blastoderm (Scott F 2006). In the main part of this work, I will focus on the 14th nuclear cycle since it shows a beautiful example of precise patterning at the molecular level, which I will detail in the next paragraphs.

At the molecular level, during the very early stages of development, maternally deposited morphogens carry a positional cue that will allow other genes to be activated. These morphogens are called maternal effect genes and are composed of mRNAs that code for Transcription Factors. Two of the systems that are initially regulated by different maternal effect genes are the Dorsal-Ventral (DV) and Anterior-Posterior (AP) axis determination processes (Alberts et al. 2002).

Bicoid and *Nanos* are maternal effect genes involved in the AP-axis determination. These mRNAs are inserted in each of the poles of the egg during its maternal deposition The genes that are subsequentially activated by the maternal effect genes are called Gap genes. The Gap genes are conformed by *Krüppel*, *Giant*, *Knirps*, and *Hunchback*, among others. The Gap genes will activate the segmentation pathway of the fruitfly represented by the Pair-rule genes. After approximately 3 hours, a precise expression pattern of these pair-rule genes will appear, forming different stripe patterns known as parasegments. Examples of the pair-rule genes are *even-skipped (eve)*, *fushi-tarazu (ftz)*, *hairy (h)*, among others. From these genes I mentioned above, I will focus on the expression pattern of the gap genes and *eve* (Nüsslein-Volhard and Wieschaus 1980) (Alberts et al. 2002).

From the pair-rule genes, other classes of TFs are activated, like the segment polarity genes. These genes do not affect the number of segments, but each segment gets reduced and modified (Nüsslein-Volhard and Wieschaus 1980). The segment polarity genes will drive the expression of the homeotic genes. Homeotic genes are more directly connected with segmentation, organogenesis, and defining specific parts of the fruitfly's body(Alberts et al. 2002).

The DV-axis determination is another early process in *Drosophila*'s embryo development that has been explored for connecting its gene regulatory elements to pattern formation and morphogenesis. In this process, maternally deposited morphogens such as *Dorsal* will define different regions through gene regulation of its target TFs. In addition, mechanical forces have been deeply explored in this system by the role of gene regulatory networks in the ventral furrow formation. The DV-axis determination system gives origin to the mesoderm and neuroectoderm that later will differentiate into different tissues from the heart, gut, and nervous system (Leptin 1991) (Alberts et al. 2002).

The early embryogenesis in *Drosophila melanogaster* presents several interesting puzzles. For example, in its development, information has to be encoded and interpreted to generate such precise patterns in a limited time. The cephalic furrow, or the boundary where the head will form, is an example of how evolution can push a system to the limits of information processing. Although the mechanism is still in debate, it is clear that the effective time for gene regulation to originate a precise cephalic furrow pattern is very limited (Tkacik, Callan, and Bialek 2008) (Tran et al. 2018).

After the maternal and gap genes have been expressed, the seven stripes of the pair-rule genes appear. These stripes get a defined expression pattern around the Nuclear Cycle 14 (NC14) (see Figure 2). This is an excellent model for segmentation patterning since several genes and the general mechanism that controls this pattern have been identified (Akam 1989). Moreover, several cis-regulatory modules of the *Even-Skipped* gene, which encode for a single stripe, have been isolated. Nonetheless, even though this pattern is controlled by positional information, it is still unknown how it can be encoded in a sequence (Small et al. 1991a) (Vincent, Estrada, and Angela H DePace 2016a).

The regulatory elements that control specific stripes from *eve* have already been well described. These TF networks reach a descriptive level that has been experimentally tested in different ways, such as TF perturbation assays and reporter lines with binding sequences for these TFs. In the context of the enhancer

that encodes for the second stripe pattern of *eve*, the TFs *Kruppel* and *Giant* have been proposed to be repressors. *Hunchback*, *Caudal*, and *Bicoid* have been proposed to have an activator role for this specific enhancer (Small et al. 1991a).

Besides understanding the role of the TFs, it is important to understand how they change in space and time. Many of the essential patterning elements in the early embryo have been spatially tracked using chemical and fluorescent labels for mRNA and proteins. Additionally, reporter lines have allowed us to identify essential cis-regulatory elements for a given pattern and how they operate. Combining these techniques has helped identify that the role of different TFs can depend on the enhancer context.

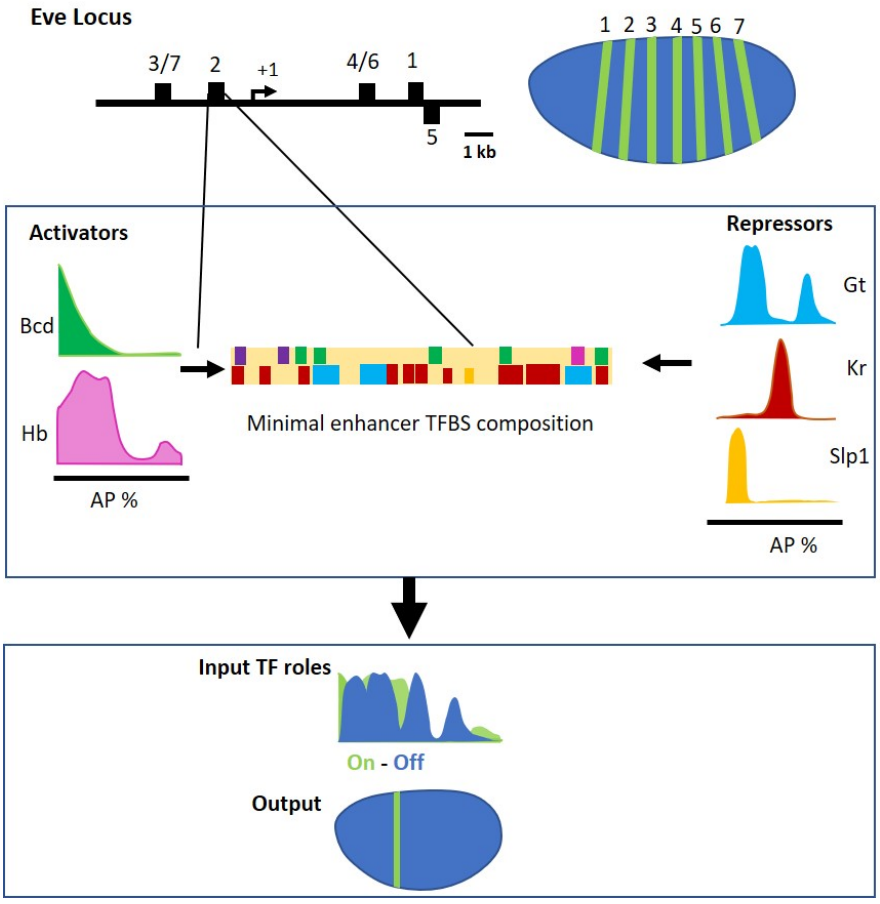


Figure 2: Gene regulation of a Drosophila's stripe pattern. In the upper part, the locus for the gene *even-skipped* is shown in a linear coordinate system, where the start of the gene is located in the +1 position. Different regulatory elements can generate their respective stripe pattern, represented by numbers in this plot. Focusing on the second stripe pattern, going to the middle panel, one can observe that this regulatory region is controlled by different activators and repressors with binding sites distributed along the sequence. In the lower panel, the spatial overlap of these activators and repressors has a net area where transcription can occur, corresponding to the second stripe's location. This scheme was created based on figures from (Vincent, Estrada, and Angela H DePace 2016a) and (Stephen Small and David N Arnosti 2020)

So far, I have been talking about gene regulation during early embryogenesis. The subsequent processes after cellularization will involve spatial reorganization of the cells and tissue formation. Different tissue layers will arise during gastrulation, which will give origin to more specialized tissues later on. The tissue type divergence, additional expression of more TFs, the different activities in regulatory regions of such tissues, mechanical forces, and the dynamic spatial reordering of the cells add much more complexity to the analysis of the system. This mixture of processes complicates understanding the mechanisms behind the observed patterns.

1.6 Measuring gene regulation on patterns across development

There are different ways to track the effects of gene regulation in developmental processes. Classical genetics uses approaches such as mutagenesis, knock-outs, and overexpression approaches to determine the role of a genetic factor in development. Besides the phenotypes at the anatomical level, gene expression patterns can only be observable at the molecular level using different techniques, such as fluorescent labels. These approaches have been essential to identify the input TFs of a gene regulatory system (Alberts et al. 2002).

On the other hand, to visualize the output of a gene regulatory system, several cis-regulatory modules in *Drosophila* development have already been well described to the point of narrowing them to sequences named minimal enhancers. Minimal enhancers serve as a simplified model to decode the mechanisms for synthesizing such patterns (S. Small, Blair, and M. Levine 1992). Using these regulatory regions with reporter lines, it is possible to explore which parts of the sequence have certain functions in a determined developmental stage. Synthetic biology allows the simplification of these systems and tests the effect of the presence, arrangement, and evolvability of certain sequences in these regulatory models (Crocker and Ilsley 2017).

A molecular phenotype of an organism can be described by labeling the components involved in a determined pathway; in the case of gene regulation, messenger RNAs and proteins can be labeled. Additionally, some systems can incorporate a cis-regulatory module with a reporter gene where the output activity can be evaluated isolatedly. A reporter gene can be observed with different staining processes, such as fluorescence and colorimetric assays. One can label the protein or the RNA to see the effects of a determined sequence presence, perturbation, or

context in gene expression (Levsky and Singer 2003). Combining classical genetics approaches with reporter lines and biochemical assays for TF-DNA interaction has been a major force in the building of gene regulatory networks.

Different molecules in certain stages of development can be studied using chemically fixed samples. Then, one can get a population of individuals with a molecular phenotype of interest. mRNAs and proteins can be detected with RNA and DNA probes and antibodies designed to bind specific features of their targets. Additionally, it is possible to track the concentration and position of determined molecules in the same individual across different stages using live imaging techniques (Garcia, Tikhonov, et al. 2013).

New techniques allow the measurement of dynamic phenotypes focused on parameters such as polymerase loading, pausing, and transcription initiation rates to understand which kinetic parameters are important for morphogenesis. These parameters have already been measured for different patterns in Drosophila. For example, for the second stripe pattern of *eve*, controlling the transcriptional window in which a regulatory region is active is essential and not just the modulation of the bursting rate (Lammers et al. 2020). Other dynamic features, such as diffusion, transport, and energy dissipation processes, are also being studied for developmental systems. However, the experimental techniques are still quite indirect and under development.

Complementary to the previously mentioned approaches where the observations are directly associated with visual phenotypes, genomics, structural biology, and evolutionary biology tools can also address the mechanisms by which the regulatory elements work.

Genomic tools can provide alternative evidence for identifying essential regulatory regions and the role of TF binding in some of them. Genome-wide explorations also provide different sources of information for generating maps that explore the specificity and affinity of different DNA sequences. With single-cell omics and spatial transcriptomics, it is possible to get information about the regulatory processes during differentiation and development. Different cell types can exhibit different regulatory programs that can be tracked in different stages of the process, allowing the building of more global GRNs across development (Badia-I-Mompel et al. 2023). Sometimes, it is easier to associate these regulatory programs with developmental processes, but there are still challenges in cell type assignation. For example, when gradients of morphogens and the cell types are not fully differentiated, the signal can be lost without a spatial input. Additionally, evolutionary

biology has helped to determine important regulatory regions based on natural selection and conservation signatures. Using selection signatures, it is possible to formulate hypotheses and predict mechanisms that allow exploring the phenotypic space in development.

Synthetic biology allows the measurement and control of different activities of regulatory elements, reducing the system's complexity. Different contributions for each element can be measured using reporter lines, optogenetic approaches, and synthetic gene regulatory elements. Transcription factors from other organisms, such as Gal4, can be used with specific promoters to drive the expression of this TF under specific developmental times and locations and activate specific targets in those contexts. Transcription factors can also be designed with protein domains such as zinc fingers and TALEs. These elements can be designed to bind specific DNA sequences to regulate a target gene. Synthetic cis-regulatory modules can be designed to introduce arrangements of binding sites for TFs, either endogenous or synthetic. Together, these approaches allow us to understand the cis-regulatory grammar and how TFs work in specific contexts to generate a pattern (Garcia, Brewster, and Phillips 2016a) (Crocker, Tsai, and Stern 2017a).

1.7 Mechanistic sequence to Expression models

One of the current goals in Developmental Biology is to understand how, when, and where a gene is regulated in a multicellular organism. The XX-century success of statistical mechanics in understanding chemical reactions inspired biochemists to test this approach in living organisms. In 1982, Shea and Ackers and, later on, Berg and Von Hippel showed that similar principles can be applied to understanding gene regulation. These models have successfully grasped the logic based on the binding energies of transcription factors, cooperativities, and other interaction energies with the transcriptional machinery, at least in some systems, such as the phage lambda operon and Lac-operon (Ackers, Johnson, and Shea 1982) (Berg and Hippel 1987).

The architecture of the Lac Operon system allows different parameters to be controlled. In this system, parameter-free models were created where it was possible to capture the experimental data accurately on how the effects of different perturbations, such as the combination of different binding affinities with different concentrations of the inducer molecule, could be predicted (Garcia and Phillips 2011).

These thermodynamic gene regulatory functions are built based on the prob-

ability of a gene being active. One of the goals of thermodynamic modeling is to account for the mechanisms behind each specific gene regulatory sequence. These systems can consider cooperativities among TFs, different mechanisms for repression and activation, different roles for each TF, and chromatin accessibility, among other features. This probability of an active gene is determined by a quotient between the active states and the partition function in the denominator, which is the sum of all the possible states of the transcriptional machinery (see cartoon in the first principles panel in Fig. 3). The energy corresponding to each state can be derived from the Boltzmann distribution, considering the chemical concentrations of each component. With the energies at hand, weights can be estimated for each of these states to determine the probability of the active gene scenario (Bintu et al. 2005).

1.8 Mechanistic Sequence to Expression models in Multicellular Systems

With the characterization of cis-regulatory elements in animals and molecular gene expression patterns, a similar goal was set to understand gene regulation in developmental systems. In this scenario, the position and intensity of gene expression are measured in different stages of development. These models can be applied to developmental systems by microscopy imaging gene expression and single-cell omics.

Thermodynamic modeling approaches have already been implemented in developmental systems with a limited extent of success. The efforts to achieve this understanding have been hampered by the high complexity of the eukaryotic gene regulation and the spatio-temporal lack of characterization of multicellular systems. Animal regulatory regions are highly complex, with many regulatory sites. Some of them operate with low-affinity sites, and some steps in the process might involve energy dissipation (Crocker, Abe, et al. 2015a) (Estrada, Wong, et al. 2016) (Fuqua et al. 2020a). On top of that, the spatiotemporal information of the input concentration gradients and inherent structure of the enhancers is required. With this in mind, it is necessary to pick a starting point, and from the positive experience in synthetic systems in bacteria and eukaryotes, together with the current knowledge from native enhancers, the thermodynamic models are a good approach to uncovering the regulatory logic in animals.

1.8.1 Data-driven Sequence to Expression models

With the advent of multiomics, single-cell omics, and tissue-specific datasets, it is now possible to generate predictive models for developmental systems using data-driven approaches such as Deep Learning. These models integrate genomic data from chromatin accessibility, histone marks, binding of Transcription factors, and gene expression in different tissues and stages. The integration of these datasets allows the finding of specific features from the genomic sequences regulated and controlled by the TFs, including long-range interactions. These approaches are also successful in predicting the effects of gene expression in different perturbations, such as mutations in reporter assays and natural variants obtained from populations. Even though these approaches are promising, their success depends on specific features, and not all tissues and TF binding sites perform well in predictability (Avsec, Agarwal, et al. 2021a) (Taskiran et al. 2023) (Almeida et al. 2023).

1.9 An experimental and theoretical setup for studying pattern formation in Drosophila

In this work, I will be focusing on reporter lines as a tool to study pattern formation. These genetic lines allow the exploration of expression patterns of a given enhancer inside a chromosomal context. They include an enhancer sequence, a promoter, and a reporter gene. The reporter gene will be used to understand how cis-regulatory sequences generate a pattern.

2 different kinds of systems will be evaluated with these reporter lines:

1) Endogenous enhancers from *Drosophila melanogaster* using reporter lines with different mutations in their enhancer sequence.

2) A semi-synthetic system where the cis-regulatory regions are artificial but include binding sites of endogenous transcription factors from *Drosophila*.

1.9.1 Endogenous systems

Native systems were explored in light of their gene regulatory elements. 2 different minimal enhancers were mutated, and the resulting phenotypes were explored to see if different features would allow me to understand the regulatory logic based on the input transcription factors and enhancer sequence. This was done using

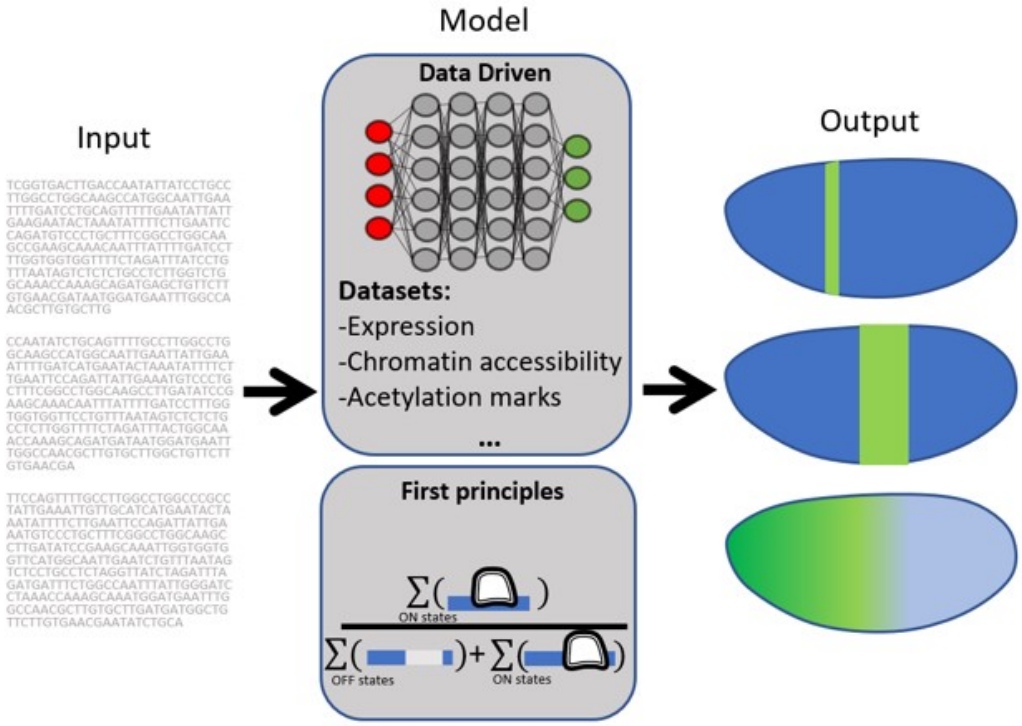


Figure 3: Sequence to expression models schematic, where only the input sequence can predict an expression pattern. The models can be trained in a data-driven approach, with neural networks, for example. On the other hand, one can build first principle models, such as thermodynamic models, that estimate the probability of a gene being in the ON or OFF state. The trained models can include additional information, such as the presence or absence of certain TFs in different parts of the embryo.

two different sorts of mutagenesis. For the minimal stripe 2 enhancer (MS2E), the sequence was mutated systematically in the regions between already described TFBSs, which I call "spacers" in this work. Rafael Galupa and Esther Karumbi synthesized, crossed, and acquired microscopy images for this systematic mutation screen. The second enhancer dataset I studied is the E3N enhancer, which encodes a stripe pattern related to the trichome appearance in larvae later in development. The E3N enhancer sequences were generated and mutated by Tim Fuqua and Noa Borst, in a randomized fashion with different amounts of point mutations. In this work, I analyzed phenotypes from different mutated versions of these minimal enhancers to find additional regulatory elements essential for this pattern. Additionally, these mutants allowed me to test if our current understanding of these

enhancers would allow me to build a predictive model of their expression patterns after a mutation.

1.9.2 Semi-synthetic systems: Random and Tailored enhancers

The complexity of the previously mentioned endogenous systems can be harnessed using simplified synthetic enhancers that will allow me to ask how currently known TFBSs can generate a pattern. Together, Rafael Galupa, Mindy Perkins, and I, focused on disentangling the ways that known binding sites of endogenous TFs can be coordinated to generate an output of a gene expression pattern in a determined developmental stage. This was contrasted with a randomized set of enhancers designed and measured by Justin Crocker, Kerstin Richter, Natalia Misounou, and Rafael Galupa. Comparisons of both sets of enhancers reveal how the sequence space can be shaped by developmental constraints, given that these enhancers behave differently at different developmental stages.

The series of targeted design enhancers was designed by Garth Ilsley, Rafael Galupa, Justin Crocker, and me. This exploratory approach consisted of different binding site arrangements for the Gap genes and maternal effect genes transcription factors. *Bicoid* (Bcd), *Hunchback* (Hb), *Giant* (Gt), *Kruppel* (Kr), *Caudal* (Cad), and *Zelda* (Zld). These sequences include synthetic spacer regions depleted of these TFs binding sites and can vary in length. Different versions of binding sites were also explored in their affinity and sequence overlapping context. Using different thermodynamic models, I used this dataset to evaluate our current understanding of these TFs.

2 Part I. Decoding pattern formation in Endogenous systems

2.1 Introduction

2.1.1 Understanding the language that controls minimal enhancers

The amount of enhancer sequences present in a single Metazoa is vast. For example, for a human, it is estimated that the number of enhancers is in the order of millions. In *Drosophila melanogaster*, this number is estimated to be something in the order of 10^4 to 10^5 enhancers (Jindal and Farley 2021) (Kvon, Kazmar, et al. 2014a). Analyses of these sequences have provided a broad range of interpretations of their operation mechanisms where affinity, orientation, type, and arrangement of Transcription Factors Binding Sites (TFBSs) are important (Fig. 4A). These complex rules of an enhancer regulatory logic are called "Regulatory Grammar."

The regulatory grammar can behave in contrasting ways. On the one hand, there is a special case named the 'enhanceosome' model, where the grammar is highly constrained (Fig. 4B, right). In this system, the regulatory elements are tightly dependent on each other, and any perturbation can affect the output expression. On the other hand, there are enhancers where the grammar allows a high level of flexibility; this system is known as the 'billboard' model. The billboard model allows for modifications in the system, such as rearrangements of the elements, without affecting the output expression (Fig. 4B, left).

One example of the enhanceosome is the mammalian β-interferon gene enhancer. This enhancer is regulated by 3 protein complexes that form a specific interaction arrangement. A crystal structure of this enhancer shows that every base pair of its sequence seems to interact with the protein assembly (Fig. 4B, right) (Panne, Tom Maniatis, and Harrison 2007). These structural observations, together with the highly evolutionary conservation of the sequence, served as an indication of a very constrained grammar for this enhancer. Examples of enhancers that follow the billboard model can have a broad range of behaviors since the mere definition of an enhancer implies information structure in the sequence. One example of an enhancer that is highly flexible is the ASE5, which tolerates shuffling of binding sites, change in the spacer sequences outside the known binding sites, and even the reduction of those spaces, making shorter versions of the enhancer; in most cases, the expected expression pattern was observed (Fig. 4B, left) (Mike Levine 2010) (Liu and Posakony 2012) (Jindal and Farley 2021).

Jindal and Farley suggest one can see the enhancers within a 'Dependency

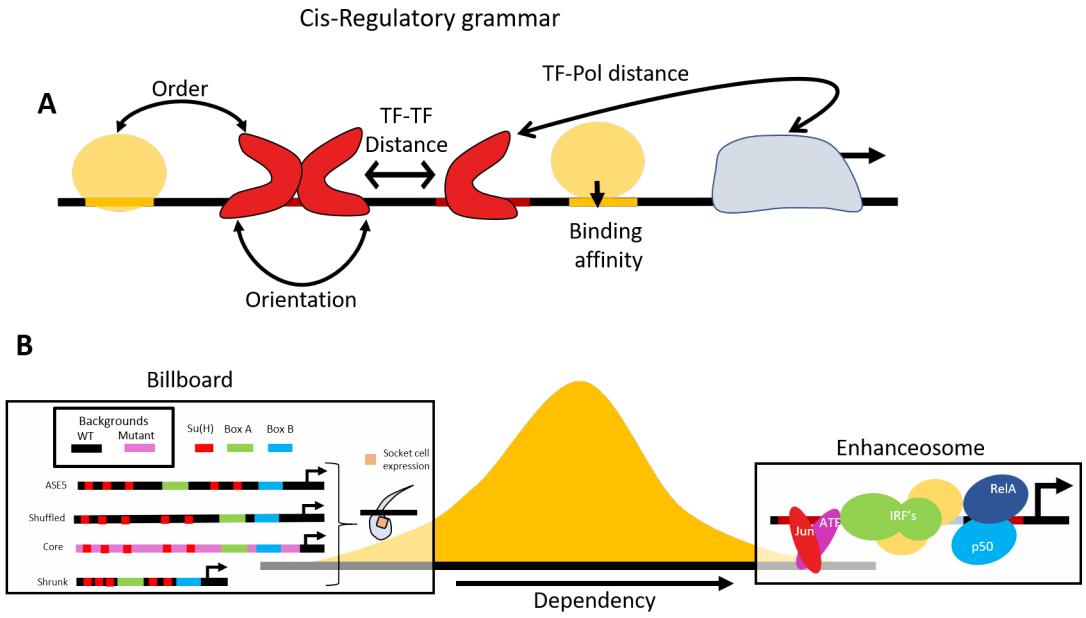


Figure 4: Enhancer grammar features. A) Enhancer grammar parameters that have been described to influence gene expression. B) Dependency grammar and its extreme examples of the Billboard and Enhanceosome models. This scheme was created based on figures from (Mike Levine 2010) (Liu and Posakony 2012) (Jindal and Farley 2021)

Grammar' spectrum. The first studies that focused on the importance of grammar estimated the conservation of a sequence in the endogenous locus (Erives and Michael Levine 2004). This approach proved useful, but for some cases, it generated several artifacts since it created a perception of structure that can also emerge by inherent mutational processes (Richard W. Lusk and Michael B. Eisen 2010b). Additionally, other comparative approaches wouldn't find conservation, and this can be due to compensatory mutations or redundant enhancers, which are hidden features not accounted for by current approaches. With the use of mutagenesis approaches in reporter lines, it is possible to evaluate the effect of specific mutations on the functionality of a single sequence, allowing a more independent evaluation of grammar. The caveat of using reporter lines is that the biological genomic context is lost at the expense of dissecting and understanding the grammar in a reduced scenario.

Drosophila melanogaster is an excellent system to explore these rules in the context of developmental biology because it has some of the most characterized developmental enhancer sequences where the position, timing, and regulators of a

pattern are very well known. Some of its minimal enhancers have been mutated deeply to understand their regulatory logic, such as the stripe 2 of *eve*, the E3N enhancer of *Shavenbaby*, enhancers of *Sparkling* involved in the eye development, or enhancers involved in the wing development. Nonetheless, these efforts have found lots of difficulties in understanding how these elements can drive patterning because these sequences are not as modular as a billboard model (D N Arnosti et al. 1996) (Swanson, Evans, and Barolo 2010) (Fuqua et al. 2020b).

2.1.2 Dissection of the Minimal Stripe 2 enhancer

The Minimal Stripe 2 enhancer (MS2E) consists of 484 bp and generates a pattern that resembles the second stripe of *eve*. It has a series of binding sites for some of the Gap genes, the Pioneer Factor Zelda, and maternal genes, such as Bicoid. The mechanism of this enhancer has been widely studied for more than 30 years with multiple lines of experimental evidence, making it one of the best-characterized cis-regulatory sequences in the animal kingdom (Goto, P. Macdonald, and Maniatis 1989) (Frasch and Levine 1987) (Struffi et al. 2011).

Trans-regulatory molecules such as TFs can be mutated, allowing the identification of the role of possible regulators of a pattern. For example, for the MS2E pattern, when *Giant* and *Krüppel* are knocked down, the second stripe expands, indicating a repression role by these TFs. On the other hand, when *Bicoid* and *Hunchback* are knocked down, the stripe expression level decreases (Frasch and Levine 1987) (Stanojevic, Small, and Levine 1991) (Small et al. 1991b). The previously mentioned assays are based on perturbations and visual inspections of the stripes pattern. They also validated the binding and activity using Protein-DNA binding assays and cotransfection assays, respectively. More recently, genomics experiments like Chip/Chip, DNAseI, and Chip-Seq allow the identification of TFs that are binding to the DNA regulatory regions inside *Drosophila* (Bergman, Carlson, and Celniker 2005) (X.-Y. Li, MacArthur, et al. 2008a) (Bradley et al. 2010).

Mutagenesis experiments in the DNA sequence of MS2E enhancers using reporter lines help to reveal the roles of Transcription factor binding in the expression of this pattern. When mutated, binding sites for *Giant* repressor generate an expansion in the anterior part of the pattern. Similarly, high-affinity sites for *Krüppel* repressor have been mutated, but a posterior expansion of the pattern was unexpectedly not observed. One possibility is that the decaying levels of activator TFs control the posterior part of stripe 2. Mutating *Bicoid* binding sites confirm its activation role proposed with TF knockdown experiments. When *Bicoid* sites are mutated, the stripe pattern is reduced significantly. The mutation of

a high-affinity *Hunchback* site has shown reductions in specific parts of the stripe. Moreover, when *Hunchback* is coupled with *Bicoid* binding sites mutations, the stripe completely disappears. (S. Small, Blair, and M. Levine 1992).

Besides establishing the directionality of a TF gene regulation with mutagenesis approaches, other explanations can be approached based on the same phenotypes. For example, there are variable levels of expression based on which *Bicoid* binding sites are mutated. Certain *Bicoid* binding sites mutations have dramatic effects, while in other sites, there are just moderate effects. These observations can be explained by the existence of *Bicoid-Bicoid* cooperativity. The facilitation of monomers binding in some sites is due to neighboring high-affinity sites. *Bicoid-Bicoid* cooperativity has also been suggested and tested for one of the *Hunchback* enhancers (S. Small, Blair, and M. Levine 1992) (Wolfgang Driever and Nusslein-Volhard 1989) (G. Struhl, K. Struhl, and P. M. Macdonald 1989).

The experimental evidence mentioned above explains only partially the Stripe 2 pattern. For example, the lack of expression in the anteriormost part of this pattern when *Giant* sites are mutated or when *Giant* is depleted indicates that there could be other repressors at play, heterotypic interactions or that there is missing activation due to modifications in *Bicoid* and *Hunchback* (S. Small, Blair, and M. Levine 1992).

Given the complexity of the rules by which regulatory elements in the MS2E play, experimental approaches have been combined with quantitative models of gene expression to test the working hypotheses for *Bicoid* mentioned above. A good model would predict this pattern's behavior under different genetic perturbations. For example, in 2013, Ilsley *et al.*, using previously known experimental data, explained better the lack of expression in the anteriormost part of the embryo for the MS2E by considering that *Bicoid* also has a repressor role in this enhancer (Ilsley et al. 2013). Other working hypotheses for the MS2E have been tested with several quantitative and mechanistic models. Some of these models have been trained with enhancer fusions. These models could predict the outputs of experiments not included in the model training, such as enhancer patterns from other Drosophilid and Sepsid species (A.-R. Kim et al. 2013).

Even though several experiments and endogenous enhancers have helped the field to make a working ground truth for this enhancer, there is contradictory evidence for these assumptions coming from Synthetic Biology. Different enhancer versions of the MS2E sequence have been generated. In these enhancer versions, the spacer sequences between the known TFBSs necessary for the 2nd stripe pat-

tern were exchanged for putatively neutral designed sequences. This experiment was done with 2 different versions of neutral sequences, and in both cases, the synthetic enhancers failed to drive the 2nd stripe expression despite containing the presumably necessary binding sites. These observations confirm the lack of knowledge of the mechanisms behind the expression of the Stripe 2 pattern. (Vincent, Estrada, and Angela H DePace 2016b).

2.1.3 Experimental dissections of other canonical minimal enhancers

Shavenbaby is a gene that will dictate where the trichomes of larvae will appear. The trichomes are structures that emerge in a stripe-like pattern to provide traction to the larvae of *Drosophila melanogaster*. One of *shavenbaby*'s minimal enhancers is named E3N. The E3N enhancer has a length of 292 bp, and it has binding sites for *Ultrabithorax (Ubx)*, *Extradenticle (Exd)*, *Pangolin (Pan)*, and *Pointed (Ets)*, among other TFs. This enhancer has a set of low affinity binding sites for *Ubx* that can work in an additive manner (Crocker, Abe, et al. 2015b) (Fuqua et al. 2020b).

Recently, in the Crocker lab, a randomized mutagenesis experiment on the E3N enhancer was performed, and each of the hundreds of sequence variants was introduced in a reporter line. A phenotypic score was generated for the possible contribution of each region along the sequence to a change in expression. The main observation was that almost every section of the enhancer had an effect, even in regions with no known binding sites mapped for the canonical TFs. Moreover, sequence conservation was a poor predictor of the expected phenotypic effect. Another study using a *Drosophila* enhancer for the *yellow spot* gene active in a different stage of development reached the same conclusions. For the *yellow spot* enhancer, instead of random point mutations, the enhancer was divided by blocks, and each block was mutated systematically. This experiment revealed that the enhancer for the *yellow spot* gene is densely encoded as well (Fuqua et al. 2020b) (Le Poul et al. 2020a)

.

Understanding how a cis-regulatory module generates a pattern often will require integrating different indirect processes that are not yet characterized. For example, the 3D architecture of the genome or the presence of microenvironments are known examples that can affect transcription (Tsai, Singer, and Crocker 2018). For this reason, a deep exploration of the regulatory logic of minimal enhancers can help reduce unknown parameters and separate the specific tasks of regulatory elements involved in embryo patterning.

2.1.4 Estimation of functional features in Minimal enhancers

The internal composition of enhancer sequences can be interpreted based on their types and numbers of TFBSs, the affinity of these sites, the orientation of binding sites, DNA shape, modularity of the enhancers, nucleotide composition, and homology-based sequence patterns, among other features.

One of the most used tools for predicting the internal TFBS compositions of enhancers is the Position Weight Matrix (PWM). A PWM can extract the statistical features from a set of sequences with a determined functional role, for example, sequences from a TF-DNA interaction assay (Stormo 2013). These statistical features can be summarized in a characteristic sequence motif representing the probability of seeing a certain nucleotide in a determined position. Once a PWM from a TF is available, one can look for binding sites for that TF in a sequence and estimate their affinity. For *Drosophila melanogaster* there are different sources of sequences from which these motifs have been extracted, for example, from Chip-Seq, SELEX, DNAse I, or bacterial 1-hybrid experiments (L. J. Zhu et al. 2011a) (Hammal et al. 2022).

Current TF-DNA interaction approaches are generally biased for high-affinity sites since detecting a binding site depends on the strongest signal to avoid experimental errors. Recently, a tool called NRLB (No Reads Left Behind), which uses data from SELEX-seq experiments coupled with a biophysical model of TF-DNA interaction, allows the inclusion of additional data that has been shown to find already characterized low affinity binding sites (Rastogi et al. 2018) (Crocker, Abe, et al. 2015b). This tool is useful for enhancers active in the late embryo and can map Hox-genes binding sites .

For decades, scientists have tried to identify DNA sequence features that can differentiate enhancers from other non-coding DNA regions. Although different methods have been implemented, none have been sufficient to find enhancers based on general sequence principles like nucleotides or TFBSs composition. Multiomic approaches seem more promising for this task, integrating DNA sequence features and experimental data from chromatin accessibility assays, polymerase binding, and epigenetic information such as acetylation marks (Avsec, Agarwal, et al. 2021b).

2.1.5 Identifying elements for enhancer grammar in Endogenous systems through systematic and randomized mutations

In a joint effort, Rafael, Tim, Noa, Mariana, Esther, and I evaluated the effect of mutations of different endogenous enhancers in different embryonic stages. This set of enhancers is relevant to the embryogenesis of *Drosophila*. For this approach, I focused on minimal enhancers that were mutagenized and introduced inside reporter lines. The resultant expression patterns for each mutant were imaged using confocal microscopy.

For the minimal enhancer of the second stripe (MS2E), here, Rafael, Esther, and I synthetically generated different systematic mutated versions of MS2E. This systematic approach was done by mutating one spacer sequence at a time to see which missing sequences are required for the stripe pattern. This knowledge makes searching for new regulators and the importance of flanking sequences of already-known motifs possible. These results reveal that additional elements across the enhancer are necessary for this expression pattern. According to this experiment, these critical sequences are distributed across the entire enhancer length.

For the E3N minimal enhancer, Tim, Noa, and Marlize generated a series of reporter lines with different randomly mutated versions of this enhancer. Then, I analyzed how these different mutations affected the affinity of already-known TFBSs. Additionally, I predicted putative novel regulators for this enhancer for different regions.

2.2 Results

2.2.1 Eve Stripe 2 enhancer: Extended binding sites for activators are not sufficient to generate MS2E expression

In the work of Vincent et al. from 2016, there are two alternative versions of the endogenous MS2E named spDP1 and spDP2 in this work. All the known binding sites were preserved in these 2 versions of the MS2E, and the spacers were mutated (Fig. 5A). This mutagenesis was controlled to avoid creating new binding sites (Estrada, Ruiz-Herrero, Scholes, Wunderlich, and Angela H. DePace 2016b) (Vincent, Estrada, and Angela H DePace 2016b). Rafael Galupa commercially synthesized these 2 enhancers and cloned them into a *lacZ*-reporter plasmid. This genetic construct was integrated into the fly genome at a specific genomic location. This approach of reporter lines was used for all the enhancer sequences in this section.

In Figure 5B and Figure 6A, both spDP1 and spDP2 enhancers drove anterior expression, and these observations are consistent with the work from Vincent et al. in 2016. This expression pattern doesn't correspond to the expected MS2E expression region, which is located at 35-45% of the AP axis (Fig. 5C).

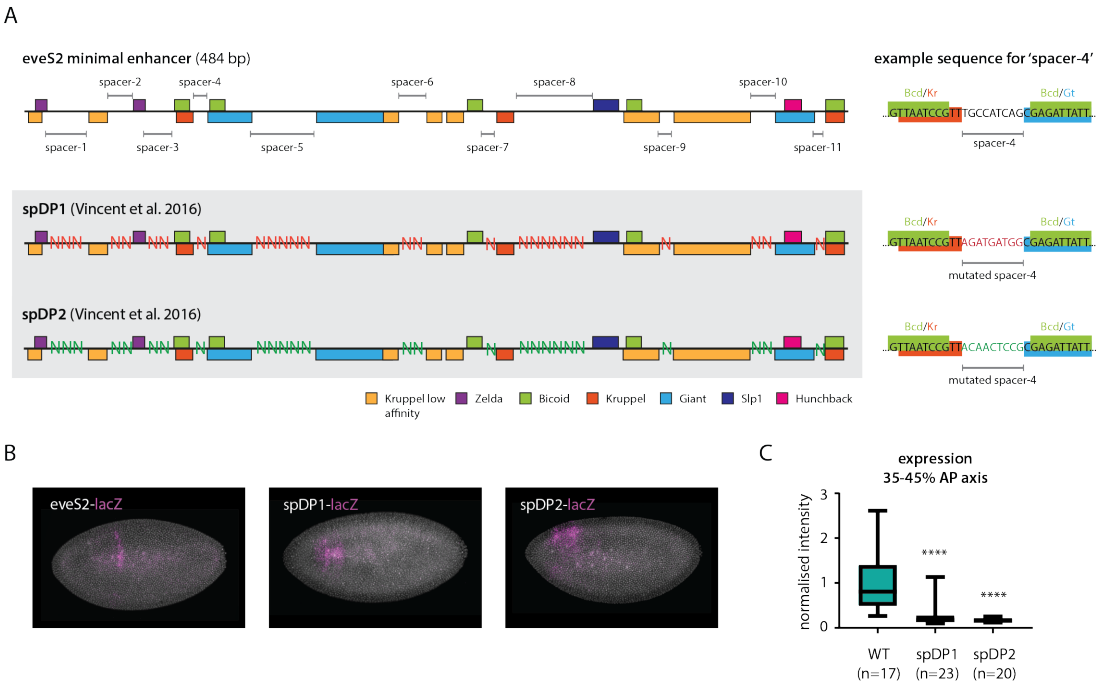


Figure 5: Two spacer mutant versions of the MS2E enhancer do not drive the expected stripe pattern. A) MS2E enhancer schematic (top) that represents its TFBSs and "spacer." In the middle and bottom are the sequences of the two mutant versions of this enhancer designed by Vincent et al. 2016. On the right is an example of mutated versions of spacer-4. B) RNA *in situ* hybridization for lacZ from endogenous MS2E and the spDP1 and spDP2 enhancers. C) Distributions of fluorescent intensities per embryo for each line at 35-45% of the anteroposterior (AP) axis. The intensities are normalized to the average WT intensity. Number of embryos used for each sample (n). Mann-Whitney tests were done for this statistical analysis (**** $p < 0.0001$). The figure was done by Rafael Galupa and Gilberto Alvarez. Gilberto Alvarez processed the sequences and analyzed the images, and Rafael Galupa did the experiments and plotted the data.

The mutation of the spacer sequences could have interfered specifically with transcriptional activation since these enhancers fail to generate expression in the MS2E region. This can be caused by affecting the known binding sites for activators or the existence of binding sites of activators that haven´t been previously suggested for this enhancer.

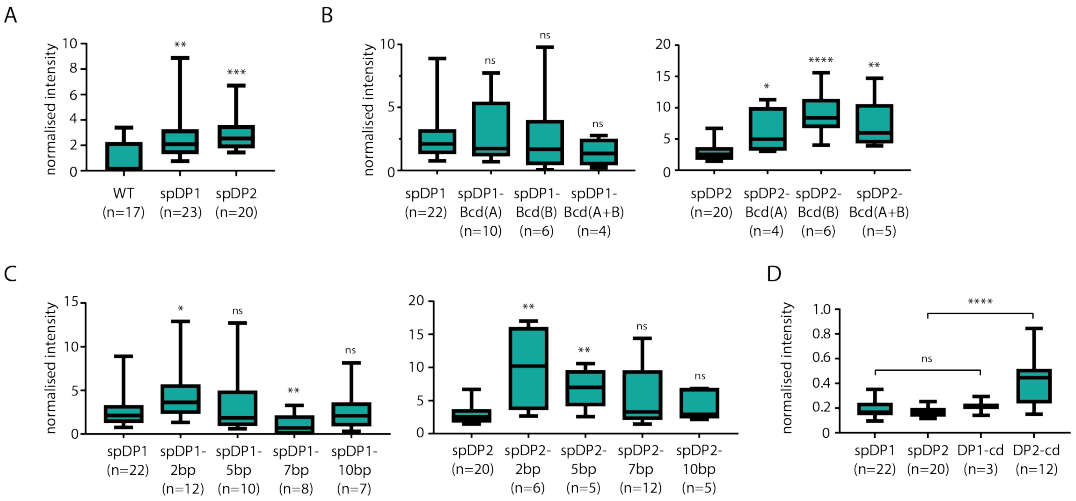


Figure 6: Anterior expression intensities for the different mutant versions of the MS2E. A) There is a gain of anterior expression on both mutant versions. B) The spDP1 mutant line did not show a gain of expression when extending *Bcd* motifs but the spDP2 mutant line showed a gain of activity when extending these *Bcd* motifs. C) Motif extension increases expression levels for several conditions of extension length. D) The addition of *Caudal* doesn't generate a stripe pattern, but in one of the mutant versions, it generates a gain of anterior expression.

Rafael Galupa first hypothesized that the known motifs for activators in spDP1 and spDP2 might lack flanking sequences to perform their adequate function, as was shown to be the case in the context of other enhancers (Gordân et al. 2013) (Farley et al. 2015) (X.-Y. Li and Michael B Eisen 2018) (Park et al. 2019a). The Bicoid motif sequence from Ronch et al. 1993 was selected and mapped in the MS2E. Two annotated Bcd sites were found to be affected in spDP1 and spDP2 (Fig. 7A). Rafael Galupa and Esther Karumbi reconstituted those "extended" Bcd motifs in the mutated lines. The reconstitution was done for each site alone or in combination (Fig. 7A). (Ronchi et al. 1993)

The extended Bcd motifs did not affect expression in the MS2E region in the spDP1 background. In contrast, the spDP2 background shows higher intensity

levels in the same region (Fig. 7B-C). However, these expression levels did not correspond to a stripe pattern based on visual inspection of the embryos. To quantify this observation, I made an image analysis pipeline coupled with a stripe detector algorithm (see Methods for further details). No stripe patterns were detected (Fig. 7D). I also observed higher intensity levels for the anterior region where the original spDP1 and spDP2 enhancers already show expression (Fig. 61B). This indicates that this Bicoid motif sequence can increase the expression but cannot rescue the MS2E pattern.

Rafael Galupa generated new enhancer sequences based on spDP1 and spDP2 in which he preserved varying lengths of endogenous flanking sequences (2bp, 5bp, 7bp, and 10bp) for each possible motif of activating TFs (*Bcd, Hb, Zld*) (Fig. 7E). No observable stripe was found in any of these lines (Fig. 7F). The quantification of the expression profiles in the MSE2 region revealed an increase in intensity levels for most of the extended motif mutant lines (Fig. 7G). Still, this increase in intensity did not correlate with the length of the endogenous flanking sites. This suggests that important sequences for known TF binding sites do not go beyond 2-5bp on each side of the core motif. Using the computational quantification and stripe detection method, I find that the extended activator binding sites do not rescue MS2E expression. These findings suggest that the failure to rescue a stripe in the MS2E position in these mutated enhancers is not due to the known activators TFBSs.

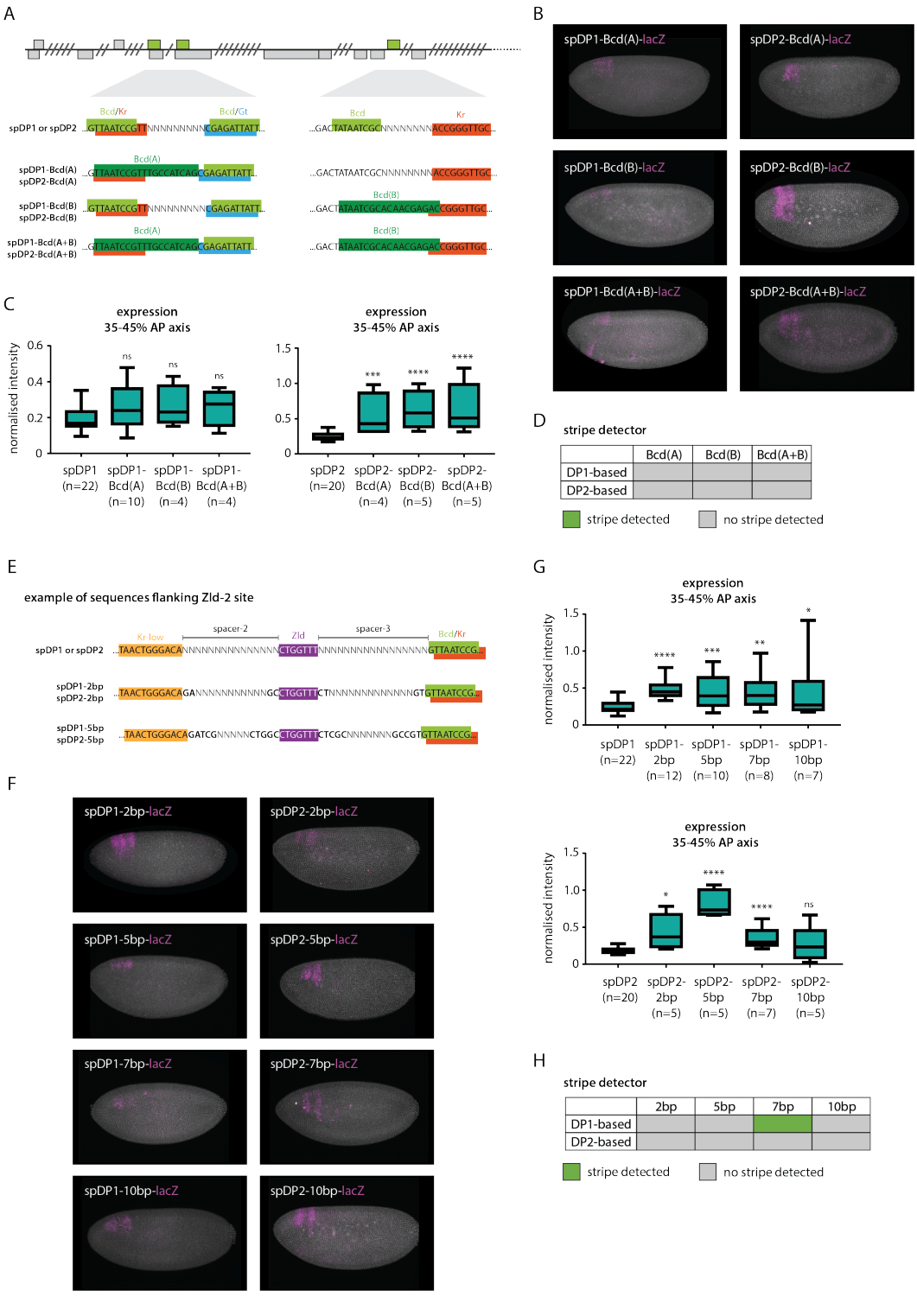

A
Bcd/Kr
Bcd/Gt
Bcd
Kr
spDP1 or spDP2
spDP1-Bcd(A)
spDP2-Bcd(A)
spDP1-Bcd(B)
spDP2-Bcd(B)
spDP1-Bcd(A+B)
spDP2-Bcd(A+B)
B
spDP1-Bcd(A)-lacZ
spDP2-Bcd(A)-lacZ
spDP1-Bcd(B)-lacZ
spDP2-Bcd(B)-lacZ
spDP1-Bcd(A+B)-lacZ
spDP2-Bcd(A+B)-lacZ
C
expression
35-45% AP axis
normalised intensity
spDP1 (n=22)
spDP1-Bcd(A) (n=10)
spDP1-Bcd(B) (n=4)
spDP1-Bcd(A+B) (n=4)
spDP2 (n=20)
spDP2-Bcd(A) (n=4)
spDP2-Bcd(B) (n=5)
spDP2-Bcd(A+B) (n=5)
D
stripe detector
Bcd(A)
Bcd(B)
Bcd(A+B)
DP1-based
DP2-based
stripe detected
no stripe detected
E
example of sequences flanking Zld-2 site
Kr-low
spacer-2
Zld
spacer-3
Bcd/Kr
spDP1 or spDP2
spDP1-2bp
spDP2-2bp
spDP1-5bp
spDP2-5bp
F
spDP1-2bp-lacZ
spDP2-2bp-lacZ
spDP1-5bp-lacZ
spDP2-5bp-lacZ
spDP1-7bp-lacZ
spDP2-7bp-lacZ
spDP1-10bp-lacZ
spDP2-10bp-lacZ
G
expression
35-45% AP axis
normalised intensity
spDP1 (n=22)
spDP1-2bp (n=12)
spDP1-5bp (n=10)
spDP1-7bp (n=8)
spDP1-10bp (n=7)
spDP2 (n=20)
spDP2-2bp (n=5)
spDP2-5bp (n=5)
spDP2-7bp (n=7)
spDP2-10bp (n=5)
H
stripe detector
2bp
5bp
7bp
10bp
DP1-based
DP2-based
stripe detected
no stripe detected

Figure 7: Expanded versions of canonical activator TFBSs are insufficient to generate MS2E expression. A) The two extended *Bicoid* motifs, Bcd(A) and Bcd(B), were added in spDP1 or spDP2 constructs. B) RNA *in situ* hybridization examples for lacZ from indicated transgenes. C) Distribution of fluorescence intensities per embryo for each line at 35-45% of the anteroposterior (AP) axis. The intensities are normalized to the average WT intensity. Number of embryos used for each sample (n). Mann-Whitney tests were done for this statistical analysis (*** $p<0.001$, **** $p<0.0001$). D) Mutants where a stripe was found with the detection algorithm are indicated in green. E) Activator motifs scheme where sites are extended by 2 or 5 bp. F) A set of embryos with RNA *in situ* hybridization for lacZ for the indicated mutant lines. G) Distribution of fluorescence intensities per embryo for each line at 35-45% of the anteroposterior (AP) axis. The intensities are normalized to the average WT intensity. Number of embryos used for each sample (n). Mann-Whitney tests were done for this statistical analysis (* $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$). H) Mutants where a stripe was found with the detection algorithm are indicated in green.

This figure was done by Rafael Galupa and Gilberto Alvarez. Gilberto Alvarez processed the sequences and analyzed the images and expression profiles; Rafael Galupa did the experiments and plotted the data.

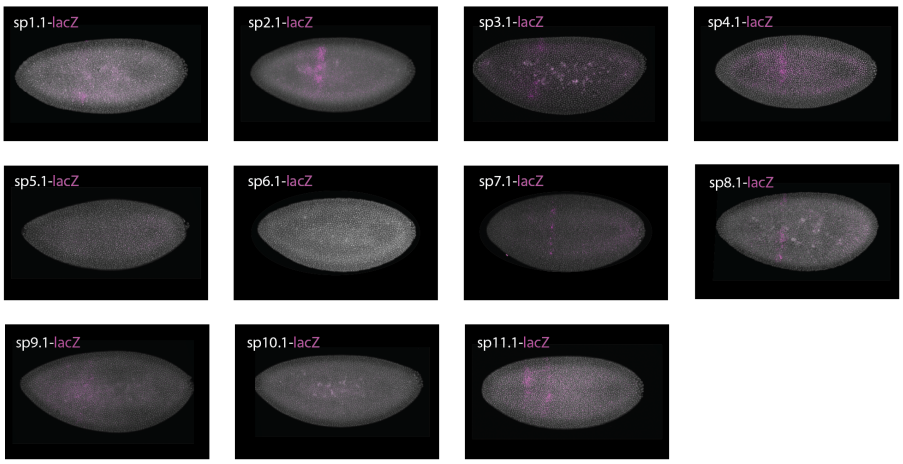
2.2.2 Most of the spacer sequences contain important information for the MS2E expression pattern

The next reasonable hypothesis is that missing regulatory elements (e.g., binding sites) could be located within the spacer sequences. To narrow down which are the critical spacers, Rafael mutated one spacer sequence at a time. Rafael synthesized two mutant versions for each of the eleven spacers based on the spDP1 and spDP2 enhancers (Fig. 5A).

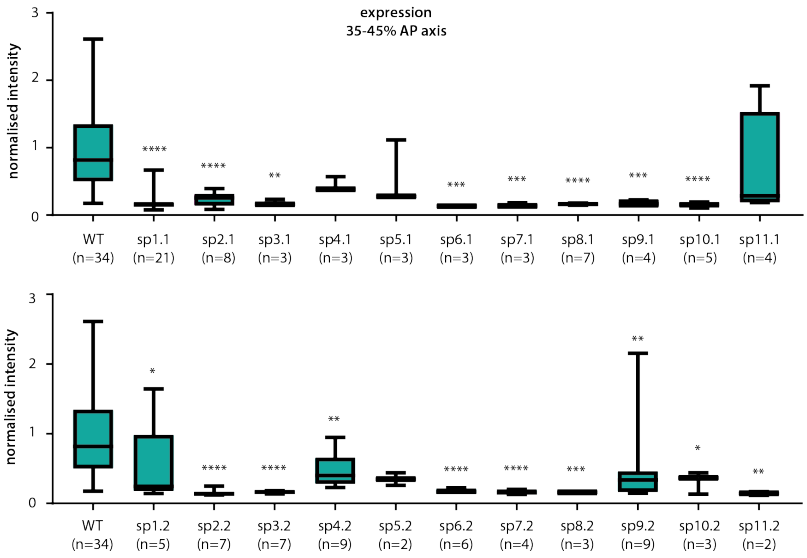
In general, mutating any of the spacers reduced the intensity levels of the MS2E stripe (Fig. 8A-B). Several cases showed no expression, while a vestigial stripe could still be observed in others. I extracted the signal from the images and processed it with the stripe detector algorithm. This method found that in (54%) of the cases, no stripe was detected in the right position, regardless of the mutant version (spDP1- or spDP2-based). In (27%) of the cases, a stripe was detected for only one of the mutant versions. For three spacer mutations (27%), a stripe is detected in both mutant versions.

Interestingly, mutating the same spacer with different sequence versions does not have the same effect. This suggests that these expression patterns are not only due to losing the original sequences but that different spacer contexts matter. In all the cases, except for sp11.1, MS2E expression was significantly affected. These results indicate that all spacer sequences could include elements that affect MS2E expression.

A



B



C

stripe detector

	spacer1	spacer2	spacer3	spacer4	spacer5	spacer6	spacer7	spacer8	spacer9	spacer10	spacer11
.1 version											
.2 version											

stripe detected no stripe detected

Figure 8: Most spacer sequences contain critical information for the MS2E expression pattern. A) Examples of embryos stained with RNA *in situ* for lacZ from indicated spacer mutants. B) Distribution of fluorescence intensities per embryo for each line at 35-45% of the anteroposterior (AP) axis. The intensities are normalized to the average WT intensity. Number of embryos used for each sample (n). Statistical analysis was performed using a Mann-Whitney test (* $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$). C) Mutants where a stripe was found with the detection algorithm are indicated in green. The figure was done by Rafael Galupa and Gilberto Alvarez. Gilberto Alvarez processed the sequences and analyzed the images and expression profiles; Esther Karumbi and Rafael Galupa did the experiments and plotted the data.

2.2.3 Identifying novel putative regulators of MS2E expression

One possibility to account for the lost information in the mutants is that the spacer sequences harbor binding sites for additional regulators. *Caudal*, an important maternal effect TF, has been proposed as a possible activator of MS2E. Additionally, there are predicted binding sites for *Caudal* in the MS2E sequence (Berman et al. 2002) (Janssens et al. 2006) (A.-R. Kim et al. 2013) (Vincent, Estrada, and Angela H DePace 2016b). To test this hypothesis, I mapped *caudal* motif across the MS2E sequence and found three binding sites (p-value<0.001; see Methods). One of these binding sites was already within preserved sequences in spDP1 and spDP2. Then, Rafael and Esther synthesized new versions of spDP1 and spDP2 enhancer sequences in which these caudal motifs were preserved. (Fig. 9A). One variant successfully generated expression in the MS2E region (Fig. 6D). However, this expression pattern doesn´t fulfill the conditions to be detected as a stripe (Fig. 9B-D). In conclusion, *Caudal* motifs alone cannot rescue MS2E expression.

Next, I wanted to identify new TFs as candidate regulators by scanning the MS2E sequence for motifs from 218 TFs present at the early embryonic stages (see Methods). Among the top 15 hits for each spacer, I detected that 83 of the 218 TFs (38%) have at least one motif in one of the spacer sequences (p-value < 0.01). In all the spacers, 89% of the motifs occur only once or twice (Fig. 9E). Now, by looking at each spacer 92% of the motifs occur in only one or two spacer sequences (Fig. 9F).

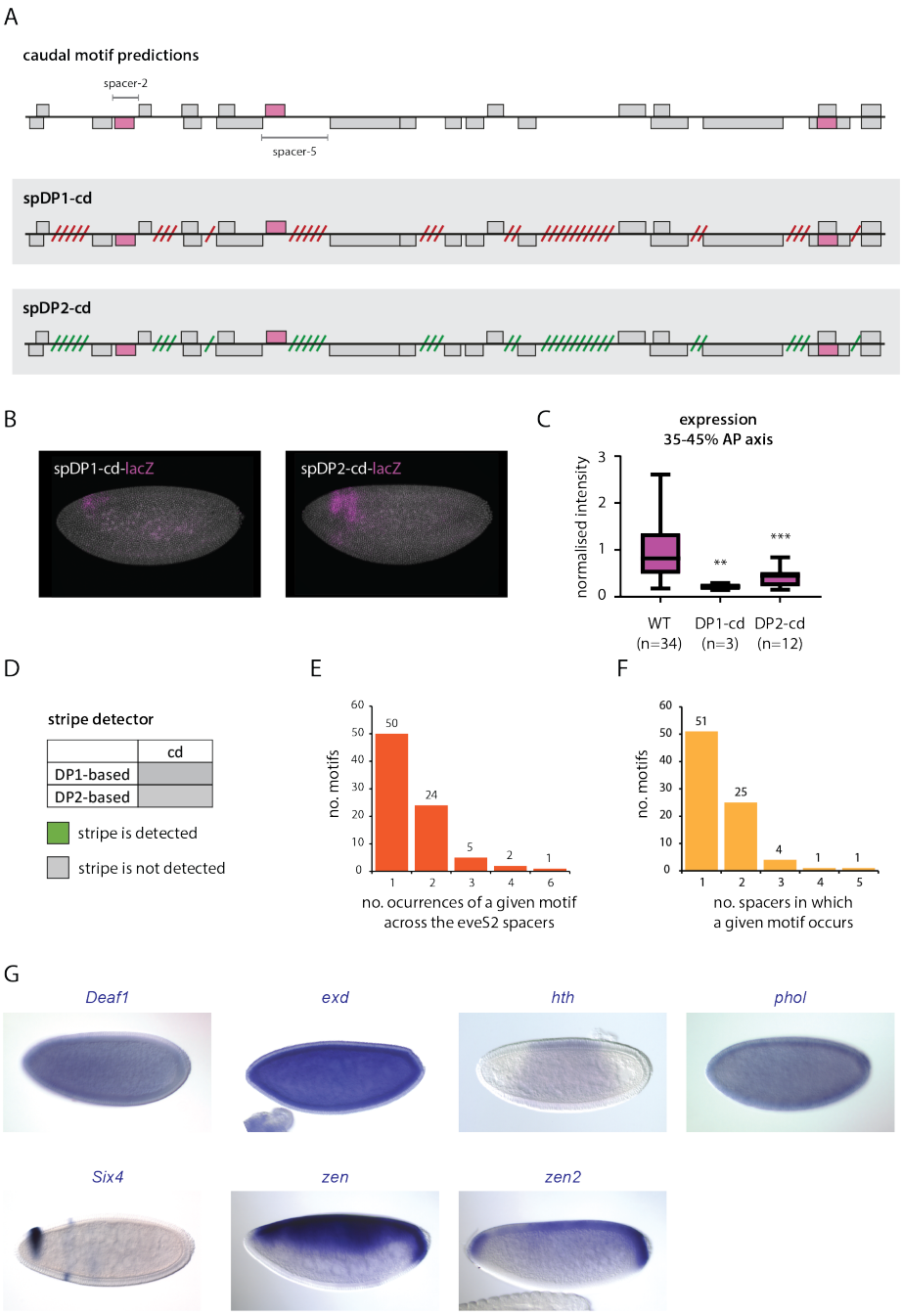

A
caudal motif predictions
spacer-2
spacer-5
spDP1-cd
spDP2-cd
B
spDP1-cd-lacZ
spDP2-cd-lacZ
C
expression
35-45% AP axis
normalised intensity
WT
(n=34)
DP1-cd
(n=3)
DP2-cd
(n=12)
D
stripe detector
cd
DP1-based
DP2-based
stripe is detected
stripe is not detected
E
no. motifs
no. ocurrences of a given motif
across the eveS2 spacers
F
no. motifs
no. spacers in which
a given motif occurs
G
Deaf1
exd
hth
phol
Six4
zen
zen2

Figure 9: Motifs of *Caudal* alone are insufficient to rescue the MS2E expression pattern. A) Synthetic sequences diagram where the motifs of *Caudal* were reconstituted. B) Embryos stained with RNA *In situ* for lacZ Embryos. C) Distribution of fluorescence intensities per embryo for each line at 35-45% of the anteroposterior (AP) axis. The intensities are normalized to the average WT intensity. Number of embryos used for each sample (n). Statistical analysis was performed using a Mann-Whitney test (* $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$). D) Mutants where a stripe was found with the detection algorithm are indicated in green. E) Distribution of frequency of motifs by spacer. F) Distribution of the number of spacers that contain a certain motif. G) Candidate TFs identified through motif analysis. The images from the expression domains of these TFs are from the Berkeley Drosophila Genome Project. The figure was done by Rafael Galupa and Gilberto Alvarez. Gilberto Alvarez processed the sequences and analyzed the images and expression profiles; Esther Karumbi and Rafael Galupa did the experiments and plotted the data

The next task was to generate a distribution of frequencies of motifs for each spacer. This analysis shows that there are six motifs are present in more than three spacers: *Deaf1* (3 spacers), *exd* (5 spacers), *Hth* (4 spacers), *phol* (3 spacers), *Six4* (3 spacers) and *zen* or *zen2* (3 spacers). From these genes, their expression patterns were corroborated to be in stage 5 (Fig. 9G). All these TFs are thus promising candidates for MS2E regulation; of note, *Exd* and *Hth* work together as a complex (Kurant et al. 1998) (Pai et al. 1998). and their motifs do appear adjacent to each other in three spacer sequences.

From the previous analysis, a list of candidate TFs was proposed. I performed TF depletions by RNAi (see Methods) to see if a list of 11 putative TFs could disrupt the early embryo's MS2E pattern at a molecular level. Additionally, I evaluated developmental defects in larvae. In none of my depletion lines I observed molecular phenotypes where the Stripe pattern for the endogenous *eve* was affected in the number of segments. These observations indicate that the depletion of this selection of TFs might not affect the endogenous pattern of eve. Additional validations need to be done since the RNAi depletion can be mild enough to allow still lower concentrations of TFs to work normally and generate a WT phenotype.

2.2.4 Analysis of the E3N enhancer reveals that dense encoded features constrain predictability from sequence

2.2.5 Affinity profiles of minimal embryonic enhancers

The E3N enhancer is positively regulated by several Hox genes such as *Ubx* and *AbdA* and other TFs such as *Pnt* and *Exd*. For each of the mutated versions of the E3N enhancer, I mapped the binding sites using PWMs for transcription factors present in *Drosophila melanogaster*. Affinities for the Hox genes binding sites were also mapped with NRLB to include low-affinity sites. Both approaches were used to estimate the effects of the mutations in each affinity of the existent sites and to explore the gains or losses of other binding sites.

I created an automatic pipeline for generating sequence profiles for the binding sites' identities and their affinities for each enhancer and mutant. Assuming binding sites are discrete entities, motif turnover was estimated based on the gains and losses of binding sites that fulfilled the statistical significance threshold and had at least 20 percent affinity from a WT affinity site.

2.2.6 Associating affinities to expression output using a β-galactosidse assay

To test the predictability of sequence features in this enhancer, I mapped the affinities for all these genes reported to contribute to its expression in the 274 mutant reporter line sequences. These reporter lines were stained with a β-galactosidse assay by Tim Fuqua and Marlize Van Breugel. The affinity for each TF for each mutant enhancer was summed up along the sequence length. Then, I performed a ratio of affinities for the mutant over the wildtype enhancer to determine if there was a total gain or loss of affinity for each TF (Fig. 10).

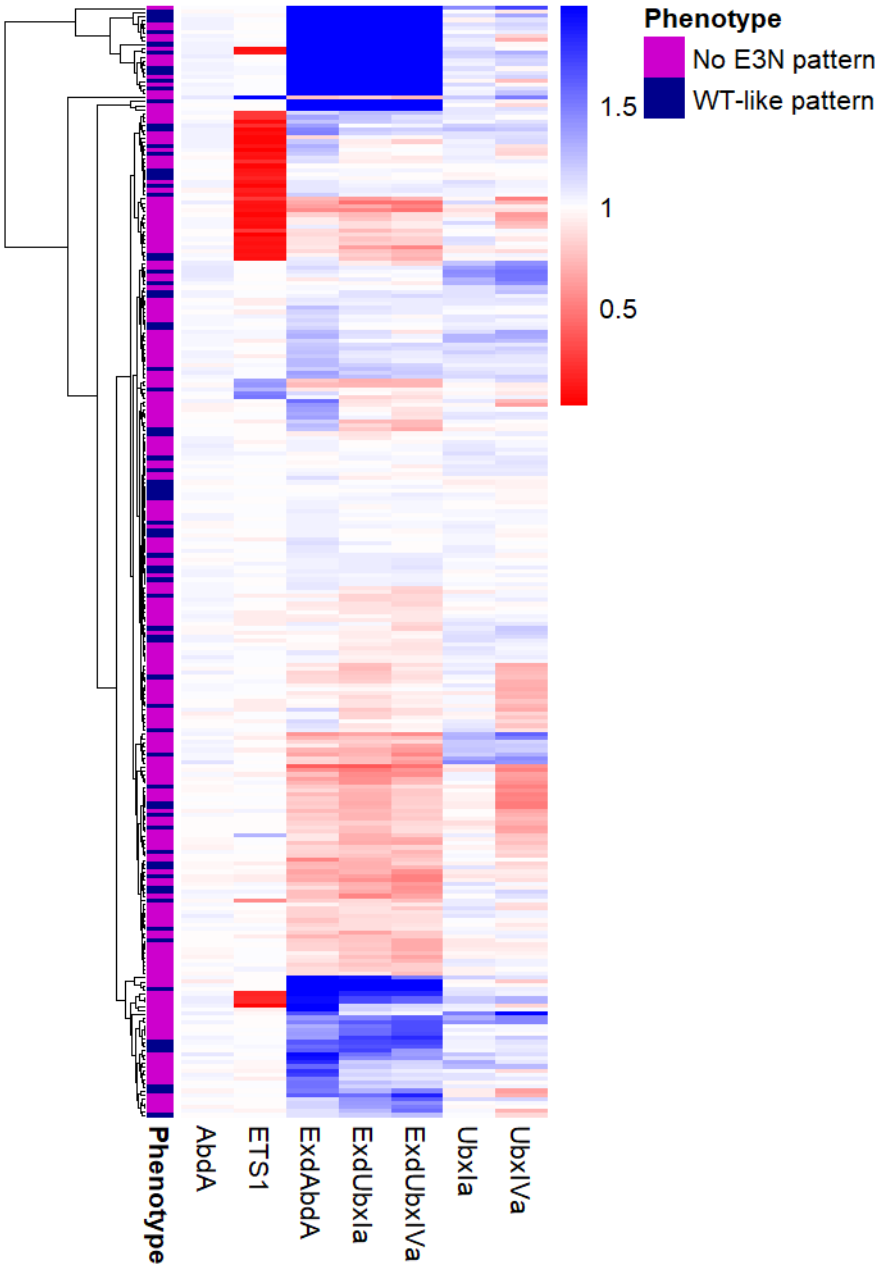

Phenotype
No E3N pattern
WT-like pattern
1.5
1
0.5
Phenotype
AbdA
ETS1
ExdAbdA
ExdUbxIa
ExdUbxIVa
Ubxla
UbxIVa

Figure 10: Heatmap of the total gain and loss of affinities for each mutant line's relevant TFs present in the NRLB model. Affinity values are normalized to the WT. A blue color represents a gain of affinity, while a red color represents a loss of affinity. The phenotype bar represents the experimental evaluation from the galactosidase assay.
The experimental expression data used for this plot was done by Timothy Fuqua and Marlize Van Breugel (Fuqua et al. 2020b).

Enrichment tests were done to assess the predictability of expression phenotypes on these affinities alone. I tested for the over-representation of an E3N-like phenotype for lines with high-affinity sites and loss of phenotype for lines with lower-affinity sites. Using different set sizes of highest and lowest affinities, it is shown that for the top 20 lines in the extremes of the affinity distribution, the hypergeometric tests are significant (Affinity gain and Affinity loss, p-value$<$0.05). Nonetheless, these significance values are non-robust to larger set sizes, indicating the presence of additional elements that could influence this enhancer. For example, the loss of phenotypes can be due to other unaccounted sites for most of the mutants.

Additionally, the mutant lines from this assay have, on average, more than 10 mutations, which makes it difficult to attribute the affinities of these TF sites as the only reason for the disruption or maintenance of phenotypes. Tim Fuqua identified putative binding sites for additional TFs with the statistical scores he obtained from this mutational screening. One of these TFs is Homothorax (*hth*), and another is Pangolin (*pan*), which behaves as an activator and a repressor, respectively. I generated profiles of affinities using PWMs this time to include these TFs, too.

I did a PWM mapping to estimate the affinities for each mutant, including the additional TFs proposed by Tim Fuqua. The results for each of the affinities for each TF in contrast to the wild-type are shown in Fig. 11.

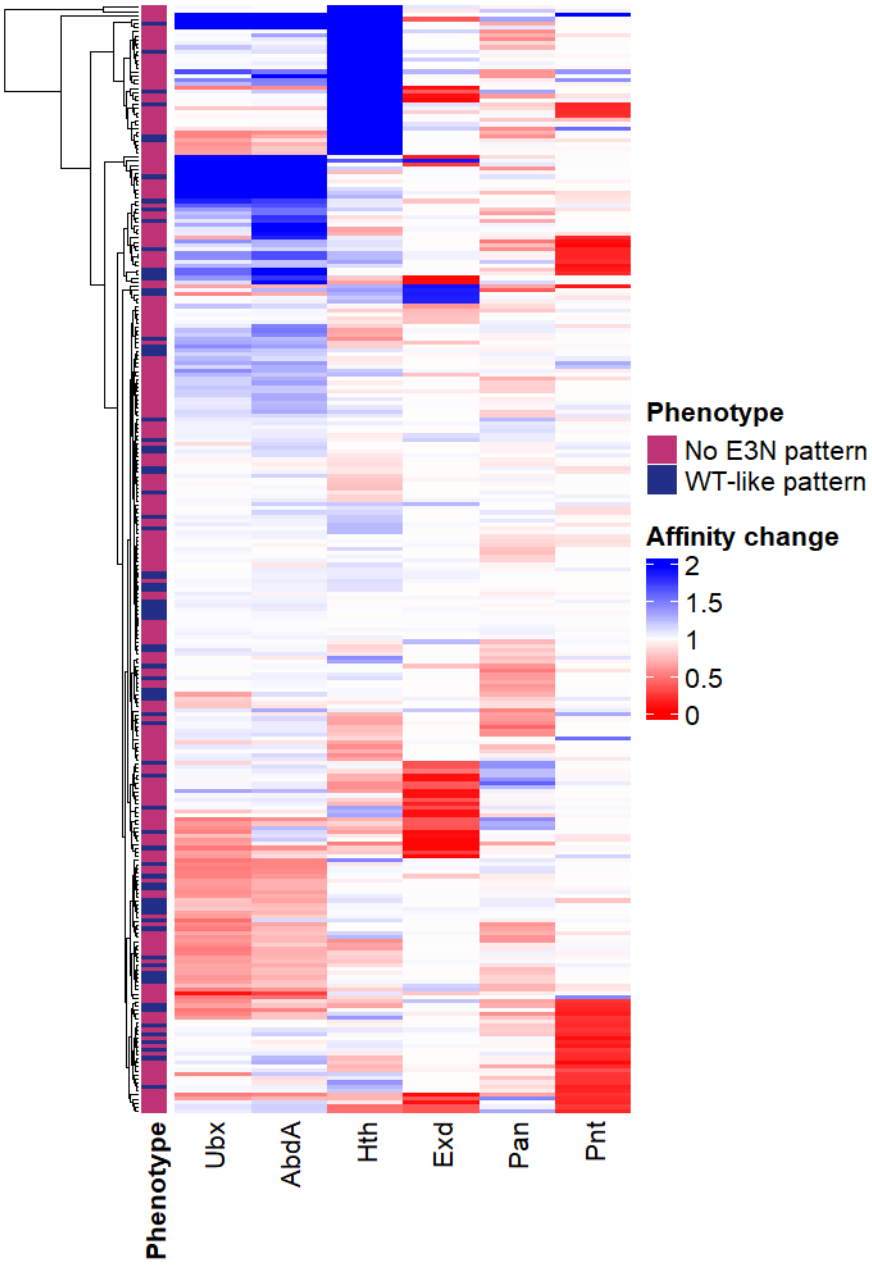

Phenotype
Ubx
AbdA
Hth
Exd
Pan
Pnt
Phenotype
No E3N pattern
WT-like pattern
Affinity change
2
1.5
1
0.5
0

Figure 11: Heatmap of the total gain and loss of affinities for each mutant line's relevant TFs present in a PWM-based model. Affinity values are normalized to the WT. A blue color represents a gain of affinity, while a red color represents a loss of affinity. The phenotype bar represents the experimental evaluation from the galactosidase assay.
The experimental expression data used for this plot was done by Timothy Fuqua and Marlize Van Breugel (Fuqua et al. 2020b).

This PWM-based method did not find an over-representation of expected phenotypes by changes in affinities. Since PWMs can only consider high-affinity binding sites, a subset of the lines with 5 mutations or less was chosen. Although for the previously tested TF sites, I see a robust over-representation of loss of phenotype in loss of affinity lines (p-value$<$0.05) for different set sizes, this is not observed for a gain of affinity. Additionally, including *Pan* and *Hth* does not improve the signal, and the direct effects of mutations on affinities were not possible to assign (Affinity loss, p-value$<$ NS; Affinity gain, p-value$<$ NS). These observations indicate different possible non-exclusive scenarios: a) There is a greater complexity in the E3N sequence, b) the galactosidase assay fails for the assignation of phenotypes, or c) our current models for transcription binding are incorrect.

2.2.7 Correlating affinities to expression output using antibody staining

Noa Borst and Tim Fuqua selected a different set of mutated E3N enhancer sequences. These sequences have different levels of mutations, from 1 to 10 point mutations and 10 lines per each amount of mutations. These reporter lines were immunostained and quantified by Tim Fuqua and Noa Borst. I mapped the binding affinities for the Hox genes in this set of sequences using the tool NRLB. Since different TF affinity changes can have different effect sizes on gene expression, I did a multiple regression to see which TF has a significant role in gene expression changes.

For the multiple regression, I used a model that combined the information from NRLB and PWM-predicted affinities. A model with the essential known TFs identifies the affinities from *Extradenticle-UbxIVa*, *Homothorax*, and *Extradenticle* as statistically significant in predicting the expected expression behavior upon mutation on this enhancer. Nonetheless, the R-squared is very low (0.16), which indicates that there are lots of unexplained expression values. For example, I selected the mutant lines with only one mutation and estimated the affinity change

for each TF (Fig. 12).

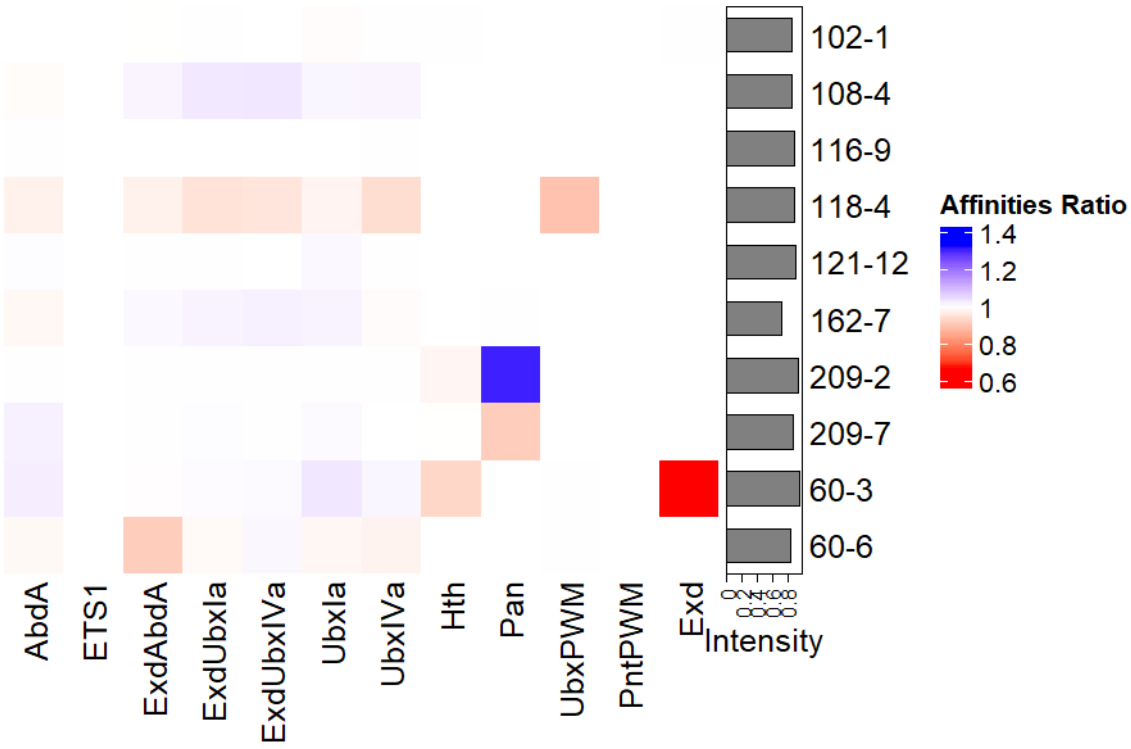


Figure 12: Heatmap of the total gain and loss of affinities for each mutant line's relevant TFs present in a mixed model with NRLB and PWM-based data. Affinity values are normalized to the WT. A blue color represents a gain of affinity, while a red color represents a loss of affinity. The phenotype bar represents the experimental fluorescence intensity from an antibody staining assay.
The experimental expression data used for this plot was generated by Timothy Fuqua and Noa Borst (Galupa et al. 2023).

As I observed with the previous analysis, it is impossible to explain the observed expression levels for several lines, even in cases of a single-point mutation. The mutant line 108-4 only has a gain of affinity on one of the well-characterized *Ubx* binding sites; however, it shows expression levels lower than the WT. The same goes for lines 162-7, where the only differences are the gains of the *Ubx* sites and a gain of a small *Pan* site.

2.3 Discussion

2.3.1 Stripe 2 pattern

With the systematic mutagenesis for the MS2E, the aim was to identify the "missing" sequence elements important for its expression. The results from these experiments suggest that essential information for the MS2E pattern is located across the entire enhancer. In previous studies, discrete deletions in this enhancer (Andrioli et al. 2002) or single mutations (Galupa et al. 2023) across the MS2E sequence led to reduced or no activity. Recent studies on other enhancers show that the information necessary to determine enhancer-driven spatial (and temporal) patterns is densely and broadly distributed across an enhancer sequence. (Fuqua et al. 2020b) (Kvon, Y. Zhu, et al. 2020) (Le Poul et al. 2020a) (Galupa et al. 2023) (X.-Y. Li and Michael B Eisen 2018).

The motif search analysis I performed suggested that MS2E might bind more transcription factors than previously known, and Rafael and I, proposed seven new regulators based on the motifs that occur more frequently and in their expression domains (Fig. 9). The expression domains from Fig. 9 come from the Berkely Drosophila Genome Project (Tomancak, Beaton, et al. 2002) (Tomancak, Berman, et al. 2007) (Hammonds et al. 2013). This approach does not exclude the fact that motifs that occur only once or twice might also be important for expression. Previous studies have also predicted binding sites for *Tailless*, *Knirps* and *Sloppy Paired 1* within MS2E (Berman et al. 2002) (Janssens et al. 2006) (X.-Y. Li, MacArthur, et al. 2008b). It will be interesting to test the potential contribution of these TFs and their motifs for MS2E-driven expression.

In summary, this study has allowed the exclusion of some hypotheses regarding the regulation of the MS2E enhancer and proposed new alternatives by identifying potential new regulators and highlighting the need to look beyond binding sites for transcription factors. Further functional studies combining extensive synthetic approaches and enhancer sequences in their native context will be necessary to get closer to cracking the regulatory code of developmental enhancers.

2.3.2 The role of *Caudal* in the reconstituted MS2E enhancers

The fact that *Caudal* improves the signal indicates that it might be a necessary element for the MS2E pattern, but it requires additional motifs or information layers. One possibility of the lack of activation for *Caudal* is that the motif version for Caudal I selected differs from that suggested by Vincent *et al.* in 2016. I used a PWM generated through DNAseI assays intrinsic from *Drosophila*, and Vincent *et al.* in 2016 suggested a PWM obtained with bacterial 1-hybrid experiments.

2.3.3 Novel regulators

Interestingly, all of the seven TFs identified as possible novel regulators are associated with homeotic genes: *Exd, Hth* and *Six4* are homeodomain-containing TFs (Rauskolb, Peifer, and Wieschaus 1993) (Rieckhof et al. 1997) (Seo et al. 1999), *Deaf1* and *Phol* bind homeotic response elements (Brown et al. 2003) (Gross and McGinnis 1996) (Brown et al., 2003; Gross and McGinnis, 1996) and *Zen* and *Zen2* are part of the Hox gene complex *ANT-C* (Rushlow et al. 1987). Maternal and early embryonic depletion of *Zen* and *Zen2* have been reported to have no early embryo phenotype based on larval cuticles (Staller et al. 2013), but the other factors have not been formally tested in the early embryo.

Recently, Rafael and I, have reported single-point mutations that decrease or abolish MS2E expression (Galupa et al. 2023). Half of those mutations occur within spacer sequences, and the other half within binding sites for repressors. 5 out of 6 significantly decreased MS2E expression. Using motif search analysis, I investigated whether those single-point mutations affected predicted motifs within the respective spacers (see Methods). Interestingly, I found that for each of those single-point mutations, certain motifs predicted in the wild-type sequence were "lost", i.e., they were either not predicted in the mutant sequences or their score was significantly decreased (Fig. 4H). These include motifs for TFs such as *Exd* and *Zen/Zen2*, identified in the analysis above, and *CG12155*.

Interestingly, I did not observe any molecular phenotype on the *eve* stripe pattern staining from all the candidates I tested for their regulatory impact using RNAi for each of the selected TFs. One of the possible explanations is that I am evaluating the phenotype at the level of the whole gene. At this scale, robustness mechanisms, such as the presence of redundant enhancers, can compensate for the TF perturbation. One solution is to repeat the experiment using a reporter line for the MS2E and see the effects only at the minimal enhancer level.

Additionally, the RNAi experiments partially deplete the expression levels of a given TF, and it is expected that the effects won´t be as strong as those of a TF knockout. Anyway, both kinds of perturbations can have indirect effects on other elements of the regulatory network. Protein-DNA binding assays can solve this problem and directly validate the interaction of a TF with a given enhancer.

2.3.4 E3N enhancer pattern

In the case of the enrichment tests with X-gal, I assumed that a gain of affinity could be directly associated with the maintenance of the typical stripe phenotype of the WT E3N sequence. This was decided based on the experimental observations that show that when there is an affinity gain for *Ubx*, the embryo keeps a stripe pattern. These observations could not be true for other TFs that their behavior upon affinity changes hasn´t been evaluated in a targeted manner. For the quantitative dataset with Antibody staining, linear relations were made between affinity and intensity changes in the observed fluorescence. This approach assumes that phenotypic changes can be directly associated with the binding kinetics of TFs; this approach has the caveat that it cannot take into account non-linear effects of the affinity such as loss of specificity, for example (Crocker, Abe, et al. 2015b). Additionally, binding sites could be created for other TFs that haven´t been described in the literature so far.

The E3N enhancer features a dense encoded regulatory region. This kind of system challenges the predictability of effects upon evolutionary and experimental effects. Quantitative models often account for separate features such as motif binding sites or DNA shape. Nowadays, Deep Learning models can account for any feature. Still, if scientists in the field want to understand its mechanisms, systematic regulatory dissections must be performed to understand the essential elements of the E3N expression.

Using the multilinear model, I observe that only *UbxIVa-Exd* and *Hth* are identified as significant variables. Interestingly, these are the only TFs that have been validated through protein-DNA binding for this enhancer. However, the role of *Hth* suggested by the model is a negative one instead of an activator. It could be that this role is a context-dependent one since, according to Fuqua et al. in 2021, it is shown that in one of the binding sites they mutated, *Hth* works as an activator when it is together to an *Ubx* binding site. On the other hand, a *Hth* repressor role has been suggested for *Drosophila*'s eye development (Pichaud and Casares 2000).

One problem is that the models, using correlation from sequence mutations to expression output, change their accuracy using different input TFs. I did not find other additional significant TFs, meaning that the proposed additional TFs, such as *Pan* and *Pnt*, might behave context-dependent or non-linearly. Additionally, validations of the binding of these TFs to the E3N enhancer would be necessary.

2.3.5 Possible evolutionary implications of densely encoded enhancers: E3N and MS2E

The difficulties in assessing predictability by affinities alone could be due to additional TF binding sites being created, destroyed, or uncovered. Moreover, since some minimal enhancers have a dense abundance of known TFBSs, this implies that the newly identified TFBSs can overlap known binding sites. Under these circumstances, densely encoded regions can generate different outputs. For example, a binding site mutation can be compensated by another binding site that overlaps its sequence (Figure 13A). On the other hand, it can happen that instead of buffering the mutational effects, one can see an amplification of the same. The buffering of mutational effects is an attractive question from the evolutionary perspective to get robust sequences to mutations (Figure 13A, lower panel).

I started to explore the possible neighboring sequences of these overlapping binding sites to see how point mutations will affect the affinities of both TFs sites. This approach gives me an idea of the expected predictability of the effects of such mutations in random mutagenesis experiments. The distribution of effects of having a mutation on the WT sequence was plotted for each transcription factor, and the expected net output resulted from having overlapped binding sites (Figure 13 B). If there is a signature of stabilizing selection at the sequence affinity level, one would expect that mutations will compensate for the effects of one binding site against the other. This future direction will test how certain binding sites could provide robustness. There are several missing aspects on which this question needs to be refined. For example, the non-proportionality of affinity changes between different TFs. Additionally, experiments will be needed to know at what precise affinity one TF can take over the other TFBSs, for example, when there is a concentration dependence for a given pattern. Another possible direction is to evaluate how many mutations away the WT phenotype is still maintained and if there are specific paths to robustness for a different starting sequence.

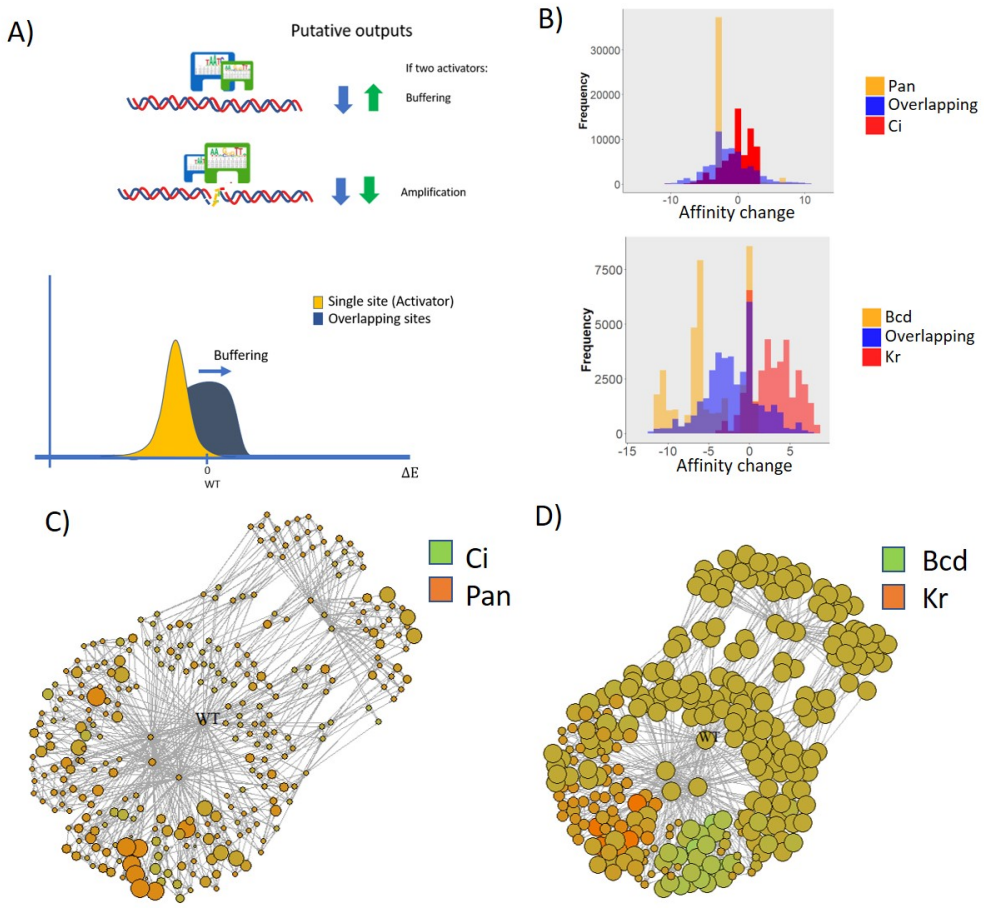


Figure 13: Overlapping sites can be a source of robustness to mutations. A) Upon mutation, when binding sites overlap, various effects can be expected, such as compensatory or amplifying effects. B) Predicted distribution of effects in affinities for overlapping sites of Pangolin and Ci (upper panel) and Bcd and Kr (Lower panel). C) and D) are graph representations of the mutational effects on these initial overlapping binding sites. The node size represents the maximum affinity for both sites, and the color represents the ratio between both affinities.

Previous works have highlighted a longer half-life for overlapping sites and that their emergence can happen just by evolutionary biases such as a dominance of deletions over insertions and selection of composition alone (Richard W. Lusk and Michael B. Eisen 2010b); one caveat of this approach is that the phenotype is still susceptible because changes in affinities can lead to diverse pattern outcomes even if the binding sites are still present. In this case, instead of focusing on the existence or absence of these sites, one can focus on the capacity of affinity compensation due to the role that each TF has. I think combining both approaches that consider the evolutionary trends towards shorter sequences, compositional biases, the half-life of binding sites, and the developmental implications of this architecture will provide us with a better idea of the emergence of cis-regulatory grammar.

Tim Fuqua, Noa Borst, and I looked for additional TFs that could explain specific phenotypes. Tim proposed that *Cubitus interruptus* a TF involved in the Hedgehog signaling pathway and could be involved in the width of the stripe. I mapped the PWM for this TF and found the putative binding sites. This binding site, in effect, is detected in a hotspot for mutations involved in stripe width phenotypes, and I observed that it overlaps one of the binding sites for *Pan*. As a contrast, I included the pair *Kr-Bcd*, a well-described model of overlapping binding sites for the early embryo in *Drosophila*.

Also, I explored if, given an initial wildtype sequence condition, how many steps are required to reach a maximal affinity or abolish a binding site. I found conflicting evidence in some cases, such as overlapping *Pan* and *Ci* sites in contrast to *Kr* and *Bcd*, since the starting point for relative affinity values are very different. Additional low-affinity sites and the importance of extended sites reflect the need to understand better the thermodynamics of transcription factor binding.

2.3.6 General remarks on our understanding of endogenous enhancers

It seems necessary to do an exhaustive screening to get an accurate enhancer map to manipulate them. For this, additional low-affinity binding site exploration in the minimal MS2E and E3N enhancer can be a fruitful direction to complete previous efforts already done with *Ubx* in the E3N enhancer. Similar to our approach, other variables, such as extended TF sites from other TFs and protein-DNA binding validations, will be necessary to complete an enhancer picture.

The PWMs available in current databases are quite limited since they can come from indirect evidence such as bacterial hybrids or *in vitro* assays. Additionally, most technologies that read *in vivo* binding have constraints due to the experi-

mental technique used. For example, Chip-Seq tends to get false positives because it detects non-specific binding. New technologies such as Chec-seq try to solve this problem by avoiding crosslinking and chromatin solubilization (Grünberg and Zentner 2017).

Besides testing the capacity of TFs to bind in a certain sequence, the temporal and spatial dynamics of a TF are essential to determine whether it is present in a given cell type. Antibody atlases efforts would be necessary to complement the observations of the effects of disrupting input TF binding sites.

However, the missing regulatory elements do not have to be TFBSs, they might instead be involved in nucleosome positioning, DNA shape or 3D organization of the chromatin, local features known to influence and contribute to enhancer function (Barozzi et al. 2014) (Fujioka et al. 2016) (Levo et al. 2015) (Rohs et al. 2009) (White et al. 2013) (Yáñez-Cuna et al. 2014). The fact that Rafael, Esther, and I often observed different results between the spDP1 and spDP2 backgrounds, both harboring different mutated sequences but chosen to be absent of TF binding sites, supports the idea that elements other than the known TF binding sites play a role in the enhancer-driven expression (Estrada, Wong, et al. 2016) (Vincent, Estrada, and Angela H DePace 2016b).

Rafael and I note that it is possible that the results observed here are exacerbated by the transgenic context in which the different MS2E variants were tested. At the endogenous locus, the effects of such mutations could be buffered, as was shown to be the case for a mutation in a single TF binding site within MS2E (López-Rivera et al. 2020a).

Finally, one of the consequences of having a dense encoded enhancer is that these systems seem fragile under perturbations like mutations on binding sites. Do these exact sequences are more robust in the endogenous genomic context? As mentioned, mutations on the minimal enhancer can be compensated by having additional information from redundant enhancers or the genomic context (López-Rivera et al. 2020a). It could be that local perturbations behave differently in the endogenous context since there could be microenvironments, and the recruiting of proteins won´t be affected by these mutations (Tsai, Alves, and Crocker 2019). It could be that this apparent fragility in a minimal enhancer in a reporter line is not even perceived as fragility in the endogenous context.

2.4 Contributions

2.4.1 Systematic mutations on the MS2E

Gilberto Alvarez (me) performed the sequence analysis in search of DNA motifs and participated in the sequence design of the lines with *Caudal* motif. I also developed the image analysis pipelines for the embryos from the MS2E systematic mutation. Additionally, I did the TF-RNAi depletion lines fly husbandry, embryo collections, and antibody staining.

Rafael Galupa conceived the MS2E systematic mutation assay departing from the synthetic sequences in the work from Vincent *et al.* 2016. He designed the sequences and was in charge of generating the fly lines. Additionally, he coordinated the project.

Rafael Galupa and Esther Karumbi did the fly husbandry, embryo collections, and the experiments for the RNA *in situ* hybridization assay. They also did the microscopy and acquisition of embryo images.

Mindy Perkins provided us with advice on developing the image analysis pipeline.

Justin Crocker supervised, contributed intellectually, and funded the project.

2.4.2 E3N enhancer

I performed the sequence analyses for all the mutants with NRLB and PWM approaches. I performed the statistical analyses to find associations of sequence affinity with the experimental phenotypes.

Timothy Fuqua and Marlize Van Breugel generated the experimental data with X-gal phenotyping. Timothy Fuqua helped to coordinate the directions for analyzing this data set.

Noa Borst and Timothy Fuqua performed the experiments and generated the quantitative phenotype information from the 100 mutant lines of E3N. Timothy Fuqua and Noa Borst suggested putative transcription factors binding sites from this data set, on which I performed sequence analyses.

Mindy Perkins and I explored the idea of robustness by overlapping binding sites. She developed a quantitative metric to learn the distribution of effects upon mutations. Additionally, I analyzed the mutational implications of overlapping TF

binding sites using motif predictions.

Justin Crocker conceived and contributed intellectually to the mutant screening design and interpretation of my sequence analyses. Justin Crocker supervised and funded the project.

2.5 Methods

2.5.1 Mutant lines libraries generation

Rafael Galupa designed the systematic mutagenesis for the MS2E enhancer. Genscript synthesized and cloned the sequences into a pLacZattB plasmid at the HindIII/XbaI site upstream of the Hsp70 promoter. Genetivision then integrated these sequences into the attP2 landing site. Rafael Galupa and Esther Karumbi homozygosed and genotyped the transgenic lines.

Timothy Fuqua was in charge of generating the mutant lines for the E3N enhancer random mutagenesis. For the antibody-stained library, Noa Borst and Timothy Fuqua randomly selected 10 different lines for each category encompassing 1 to 10 mutations, aiming for 100 lines in total. The enhancer library for the X-gal experiment was assembled with Genscript. The X-gal library used a degenerate PCR with a 2% mutational frequency (Fuqua et al. 2020b). The enhancer library for the Antibody staining experiment was assembled with Genscript, too. These constructs were cloned into a pLacZattB plasmid at HindIII/XbaI site. These lines were injected into a VK33 line by Genetivision. These transgenic lines were homozygosed by Timothy Fuqua and Noa Borst.

2.5.2 TF-RNAi lines

Rafael and I selected a set of 11 candidate TFs with a putative regulatory role for the second stripe for performing TF depletion with RNAi. Charalampos Galouzis provided us with *Exd* and *Phol* RNAi-TF lines. The rest of the RNi-TF fly lines were ordered through Bloomington. All these lines were crossed with a maternal Gal4 driver line at 25 degrees Celsius, and the F2 generation was selected to screen for phenotypes.

2.5.3 TF-RNAi lines: Fixation and Antibody staining

Embryos were collected from egg-laying chambers where the parental line was left for acclimatization for 2 days. I collected embryos 4 hours after swapping plates to get them to stage 5. Embryos were collected in baskets. A 50% bleach solution is used for removing the chorion for 90 seconds. A rinsing step removes the residuals, and then an additional wash with a buffer is performed (0.1 M NaCl,0.04% Triton X-100). Embryos were transferred with a brush to scintillation vials with a solution of 700 μl of 16% PFA, 3 ml of 100% Heptane, and 1.7 ml of a solution PBS/EGTA. Then the scintillation vials with the embryos were shaken at 250

rpms. The aqueous fraction was removed, methanol was added, and a vortexing step was performed for a minute. Embryos in the interphase were discarded, and the upper fraction containing heptane was discarded. Two additional washes were done with methanol, and the embryos were stored at -20C.

I used an *Even-skipped* antibody from DSHB for the antibody staining. I passed the samples that were in methanol to 1.5ml tubes and rehydrated them. For this, I put the embryos in a solution of PBT/MeOH (50%), and the samples were rocked in a nutator for 10 minutes. Then I did 4 washes with PBT, and in each wash, I put the samples to rock in the nutator. I prepared a solution of PBT/blocking reagent (diluted 1:5) and washed the samples with this solution. Another washing step was done with this solution, and I put the samples to rock for 30 minutes in the nutator. A primary antibody solution was prepared in a PBT/Block solution with 200 microliters per tube. The samples were left rocking for 2 hours at room temperature. I removed the primary antibody and the blocking reagent by doing 3 quick washes with PBT. I did 5 more washes with PBT and rocked them in a nutator for 10 minutes between each wash. Later, I made a wash with a PBT/blocking solution and an additional wash with PBT/Block for 1 hour rocking in a nutator. I prepared the secondary antibody with a PBT/blocking solution. The samples were rocked in a nutator overnight at 4°C and protected from the light. The next day, I removed the secondary antibody by rinsing twice with PBT and performed 7 additional washes with PBT by putting them to rock in a nutator for 10 minutes between each step. ProLong Gold was acclimated at room temperature, and in the meantime, I removed as much as possible of PBT from the samples. I added 150 microliters of ProLong Gold to each sample. The samples were resuspended and left to incubate for 10 minutes. Finally, I mounted the samples on slides, and sealed them with nail polish the next day.

2.5.4 E3N and MS2E affinities profiles

Affinity predictions for all the TFs sets used in this study were done through PWMs and for the Hox genes with NRLB. For the E3N enhancer, NRLB was used to get the affinity profile along the enhancer sequence for *Ubx*,*Abd-A*, *Pointed*, and their coupling with *Exd* (Rastogi et al. 2018). Then, a total affinity score was estimated by summing all the affinities for each mutant sequence. The total affinity for each TF for each sequence was normalized to the total affinity of the WT sequence for that TF.

A similar approach was taken with the PWM estimated affinities. These PWMs were obtained from the FlyFactorSurvey and MotifDB databases. For the E3N

enhancer mutant sequences, the software PWMenrich was used since it allows the estimation of affinities per base pair and can generate a total affinity score for each sequence (Robert Stojnic 2017). Similarly to the approach with NRLB, the total affinity for each TF for each sequence was normalized to the total affinity of the WT sequence for that TF.

2.5.5 Correlating affinities to Output expression for the E3N enhancer

To associate affinity changes with phenotypes, I took two approaches since I was provided two different classes of datasets: On the one hand, Timothy Fuqua and Marlize Van Breugel provided me with a qualitative dataset with binary phenotypes observed through an x-gal approach; on the other hand, Noa Borst and Timothy Fuqua provided me with a quantitative dataset from antibody staining.

For the qualitative dataset, I selected the extreme sets in the distribution of affinity changes. Then, I performed a Hypergeometric test in R to see if a loss of phenotype was associated with a loss of affinity and if a gain of affinity was associated with the maintenance of phenotype.

For the quantitative phenotype. Noa Borst and Timothy Fuqua gave me the expression values for each mutant reporter line. I performed a multilinear model in R to see if the affinity variables could explain the expression gain or loss. The affinity variables were joint affinities from PWMs and NRLB output values.

2.5.6 In situ hybridization protocol for the MS2E

In situ hybridization (probes): probes for lacZ (reporter) and snail (internal control) were generated from PCR products using the in vitro transcription (IVT) kit from Roche (#11175025910) and following manufacturer's instructions. A list of primer sequences for each PCR product can be found in Table S1 from (Galupa et al. 2023). For each gene, distinct PCR products were pooled before IVT reaction. Probes were diluted in hybridization buffer (Hyb; 50% formamide, 4X SSC, 100 μg/mL salmon DNA, 50 μg/mL heparin, 0.1% Tween-20) at 50ng/μL. Prior to hybridization, a probe solution was prepared (per sample, 50 ng of each probe in 100 μL), denatured at 80 ^{o}C for 5min, then immediately put on ice for 5min, and finally incubated at 56 ^{o}C for 10min before added to the embryos.

In situ hybridization (procedure): embryos stored in methanol were washed in methanol/ethanol (50:50), three-times in 100% ethanol and then permeabilized in xylenes (90% in ethanol) for 1h, after which embryos were washed six times in

ethanol and three times in methanol. Embryos were then washed three times in PBT (PBS + 0.1% Tween-20) before post-fixation for 25min in fixative solution (225 µl 16% PFA, 500 µl PBT). Embryos were then washed several times in PBT for 40 min, followed by a wash in PBT/Hyb (50:50) at room temperature and a 30min-wash in pre-warmed Hyb at 56 ºC. Embryos were then incubated with probe solution at 56ºC overnight. The next day, embryos were washed in Hyb (three quick washes followed by three 30-min washes), then in Hyb/PBT (50:50), then in PBT several times for one hour before incubated for 30 min in blocking solution (Roche #11921673001; diluted 1:5 in PBT). Embryos were then incubated in blocking + primary antibodies diluted 1:500 (anti-DIG, Roche #11333089001; anti-FITC, ThermoFisher #A889) at 4 ºC overnight. The next day, embryos were washed in PBT (three quick washes followed by four 15-min washes), and then incubated at room temperature in blocking solution + secondary antibodies diluted 1:500 (AlexaFluor 488 and 555, ThermoFisher #A21206 and #A21436, respectively). After 2 hours, embryos were washed in PBT (three quick washes followed by four 15-minute washes), mounted on Prolong Gold with DAPI (ThermoFisher, P36935), and left to curate overnight before imaging.

2.5.7 Image Acquisition and Analysis for the MS2E

Embryos were imaged using a confocal microscope Zeiss LSM 880 confocal. Images were processed using a combination of automated scripts with manual curation. For this analysis, images of stage 5 embryos were selected, and Rafael Galupa and Esther Karumbi performed a maximum-intensity projection in ImageJ. I developed an automatic pipeline in Matlab (version R2019b; The MathWorks, Inc.) where embryo rotations to their AP-axis are performed by using their Feret diameters after fitting an ellipsoid. From this image, I averaged all rows' intensity values for each pixel's horizontal position, excluding the ones in the background outside the embryo.

After obtaining an AP intensity profile for each embryo, I performed statistical processing for all of them in R. I used a Gaussian filter to smooth the signals. Then, I subtracted the general background by taking only the values higher than half of the maximum intensity detected for each embryo. After this, I performed a linear interpolation, which allowed me to associate equal AP coordinates to all embryos. Another background removal is done using the 10% quantile of intensities to obtain a net expression value from the background intensity within the embryo. Additionally, the signal is normalized to 50% quantile of the intensity within the embryo in the 60% to 80% coordinates from the AP axis. These *in situ* hybridization assays were done in different batches, and to control for each

batch effect, the threshold for background removal was based on the second stripe from a WT reference line. To get the interval confidence for each mutant set, I performed a bootstrapping for all the signals using 1000 replicates and a 95% interval confidence.

The automatic stripe pattern detection code I designed looks for the region within 15% to 75% of the AP axis. In this region, the detector looks for an intensity peak above a threshold; its minimal width is 5% of the AP axis, and a decrease below the threshold surrounds this peak. The defined threshold value was 5% of the maximum intensity using a stripe 2 pattern from the WT control experiments. The values of these parameters were chosen as the minimal criteria that allow the detection of all stripes in the WT control patterns. Finally, the collected data was plotted by Rafael Galupa, using GraphPad Prism version 10.2.3 for Windows, GraphPad Software, Boston, Massachusetts USA, www.graphpad.com.

3 Part II. Encoding synthetic patterns in Drosophila embryos

3.1 Abstract

The spatiotemporal evolution of chemical and mechanical agents provides the required information for the developmental processes of an organism. Gene regulatory networks modulate the concentration and identity of the required regulatory molecules. Nonetheless, the lack of knowledge of mechanisms and components from developmental gene regulatory networks limits the generation of predictable outcomes of pattern formation under genetic and environmental perturbations. Cis-regulatory elements are important components of these networks controlled by the input concentrations of different TFs in different locations of an organism, generating a gene expression pattern. Several efforts have been made to create predictive models for these expression Patterns based on DNA sequence alone. In this work, I systematically test the capacity of sequence-to-expression models using the well-known gene regulatory network of the maternal and Gap genes. These genes are involved in developing the *Drosophila* blastoderm and are responsible for creating precise stripe gene expression patterns in a limited time. For this purpose, I use a set of synthetic enhancers that include binding sites for repressors and activators essential for the Anterior-Posterior patterning in Drosophila´s early embryo.

3.2 Introduction

3.2.1 Synthetic biology for understanding gene regulation

The first identified enhancer sequences came from the SV40 virus DNA to see if additional pieces of DNA outside the promoter could affect gene expression. In these experiments, candidate viral sequences known to have essential activity for the virus were located upstream of a reporter gene in mammalian cell lines. These sequences were fused with a promoter containing a transcriptional start site and a reporter gene. Interestingly, these sequences showed a remarkable increase in the reporter expression (Banerji, Rusconi, and Schaffner 1981). After these first experiments, more enhancers started to be identified in more organisms and within the context of developmental biology. Some of these experiments include the mutagenesis experiments that started to uncover the role of TFs and the cis-regulatory grammar of different arrangements of TFBSs.

The understanding of endogenous systems has been improved with exploratory approaches using comparisons of sequences among different species. These ap-

proaches allowed the field to identify putative regulatory elements on a large scale. Using conservation as a proxy for functional role has allowed the mapping of important non-coding elements that help regulate expression. For example, before the whole genome sequencing era, for the gene *Hoxb-4*, which is known to be important for vertebrate development, distant comparisons between mice and putterfish DNA sequences allowed identifying some of its regulatory elements. Since the putterfish genome is compact, it was very useful for identifying conserved regulatory elements (Aparicio et al. 1995). When the whole genomes of mice and humans were available, the first comparisons showed that around 5% of them are conserved while only 1.5% of them encode for proteins. More importantly, when these possible important conserved non-coding sequences were functionally tested, the outputs showed roles for tissue and time-specific expression. (Visel, Bristow, and Pennacchio 2007).

Although comparative approaches allow the finding of important regulatory sequences, it is important to consider that most of them would be difficult to identify since these regions are more susceptible to evolutionary changes. Even more interesting is that when new sequences are added to the alignments, conserved arrangements of TFBSs can appear just by random effects and create the idea that certain organizational forms of TFBSs are being selected (Richard W Lusk and Michael B Eisen 2010a). This artifact can create the impression of the existence of certain regulatory grammars when there are not. Another difficulty can be that the TF mechanisms might have diverged among the different species even though they might have a similar DNA binding domain.

Methods like Chip-Seq, ATAC-seq, and RNA-seq allowed the search of statistical trends on the presence of regulatory molecules and their targets. These approaches can provide evidence for a regulatory role in specific segments of the non-coding genome. Nowadays, with single-cell omics, the sequence space that can be explored to understand the gene regulation grammar can be expanded since different sequences are active and accessible for different cell types that work with different subsets of transcription factors (Berest et al. 2019) (Bravo González-Blas, De Winter, et al. 2023).

The previously mentioned genomic approaches are helping to fill the gaps in the enhancer genomic map using the endogenous system. The integration of natural variation and its effects in the binding of TFs, chromatin accessibility, expression level, and timing of the gene regulation allows the association of genotype and phenotype. Additionally, the genome harbors millions of sequences that can be compared to find statistical signals that might represent the grammar. These ap-

proaches have successfully identified key elements for gene regulation, and new techniques are being developed to improve the extent of these approaches to finding grammar rules. One of them is Chip-Nexus, whose base-pair resolution allows for the search of grammar rules such as cooperativity (Avsec, Weilert, et al. 2021). Spatial assays provide an additional layer of information since gene regulation works differently depending on the context; for example, spatial multi-omics coupled with deep learning allowed decoding of how regulatory grammar controls cell identity in hepatocytes (Bravo González-Blas, Matetovici, et al. 2024).

Nonetheless, there may be a limit to the inference problem using endogenous sequences alone. Even with the use of multiomics and comparative approaches, the extent of the regulatory complexity indicates that it cannot be handled without adding synthetic sequences to the equation. For example, while using a simplified toy model of human TFs binding, De Boer and Taipale, show that to understand pairwise heterotypic cooperativities, one would require 220,0000,000 parameters to test an accurate picture of a regulatory model (Boer and Taipale 2024).

Additionally, the complexity increases if one considers the other layers of cis-regulation, such as higher-order cooperativities and low-affinity binding sites, which are often non-detectable until tested experimentally. Feedback and non-equilibrium processes have been suggested to work as mechanisms for some enhancers (Park et al. 2019b) (Sönmezer et al. 2021), and even if one can get parameters related to these phenomena, the current theories have difficulty predicting these systems' outcomes and time evolution.

As García and Phillips said in 2016, synthetic biology allows us to bend nature to test specific hypotheses instead of bending our models to explain our biological data (Garcia, Brewster, and Phillips 2016b). The genetic engineering of the first promoter fusion experiments in SV40 and the development and import of molecular tools that don not exist in endogenous systems have allowed us to simplify the complexity of biology to tackle specific questions. For example, using foreign TFs such as Gal4 from yeast and generating modified versions of TFs using different activation, cooperativity, and binding domains opened the doors to control gene expression in a specific spatial and temporal context. Synthetic transcription factors allow cleaner testing of specific hypotheses, such as the role of cooperativity, binding affinity, input concentration, and non-equilibrium in gene regulation. On the other hand, biological systems have many unknown molecular components, making the synthetic approach an essential method to uncover the principles of the known molecules of interest, in this case, Transcription factors and their binding sites.

Creating synthetic promoters and synthetic enhancers with features outside the natural variation allows us to test the effects of mutations on the mechanisms of the transcriptional machinery. Crocker and Ilsley suggest that a suitable synthetic enhancer that works as a null hypothesis is required to test fundamental enhancer hypotheses. A null hypothesis enhancer is normally outside of what is available with natural variation of endogenous systems,

Integrating both sources of knowledge, from endogenous and synthetic systems, is beneficial for understanding and guiding experimental tests of putative hypotheses. For example, training thermodynamic models with data from the endogenous stripe 2 enhancers under different experimental perturbations and its integration with comparative genomic sequences from these enhancers has allowed the selection of putative models to explain the experimental data better. Recently, Deep-learning methods have been trained using multi-omics datasets and/or *in vivo* enhancer activity assays with success but only for specific cell types and TFs contexts (Almeida et al. 2024) (Taskiran et al. 2024).

3.2.2 Building synthetic enhancers

In the previous section, I highlighted the importance of Synthetic Biology for dealing with the high complexity in cis-regulatory regions. For this chapter, I will refer to synthetic enhancers as sequences with features outside natural variation or with a synthetic edition distance larger than the one generated by natural variation by point mutation.

Synthetic enhancers can be generated by the edition or introduction of the sequence in endogenous loci or at independent genomic locations like landing sites and vectors outside the chromosomes. The synthetic enhancers I will refer to in this work are introduced in an independent locus, specifically in a landing site. In these circumstances, the grammar effects are less likely to be affected by the redundant enhancers, which can be present in their endogenous context or by additional information outside the minimal enhancer. In the endogenous loci, the genomic context has been through similar evolutionary processes and is likely to affect the minimal enhancer (López-Rivera et al. 2020b).

There are several difficulties with the design of synthetic enhancers. Sequences added in an independent locus can still be influenced by the information in that new genomic context. Another difficult task is the design of neutral sequences for filling spaces between binding sites in a synthetic enhancer. There is no such thing

as a neutral sequence model of DNA. There are methods to generate sequences that avoid specific TFs binding features, but by chance, these sequences can include the unexpected binding of other TFs. Some TF families have similar binding sites, making the sequence specificity difficult to contain (Estrada, Ruiz-Herrero, Scholes, Wunderlich, and Angela H DePace 2016a).

Another factor to consider when generating synthetic sequences is the cost of the synthesis. Overall, the ideal scenario is to explore the sequence space with millions of combinations to understand specific scenarios. According to De Boer and Taipale, one alternative is the generation of randomized sequences on a large scale, which can provide enough data to dissect cis-regulatory grammar mechanisms. Randomized DNA can be synthesized in a pool by random processes compared to a directed design. This approach has been successful in addressing questions regarding the enhancer grammar.

3.2.3 Exploring enhancer grammar through synthetic enhancers

In Chapter 1, I explained how to infer cis-regulatory grammar code using endogenous enhancers and their synthetic mutagenized variants. Large-scale mutagenesis allows the exploration of certain grammar features; for example, point mutations can likely generate affinity changes in existent binding sites or even disrupt their function. Point mutations can also shift the preferred position of a binding site. Depending on the context, these mutations could generate new binding sites by chance if there are already sequences with a predisposition for them. The caveat with this approach is that based on empirical evidence from previous works and my observations in Chapter 1, endogenous enhancers seem to have important information even between the known binding sites (Fuqua et al. 2020c) (Le Poul et al. 2020b). Additionally, binding sites can overlap with another TF binding site, often making a targeted mutation difficult to achieve since it will have unexpected effects.

One solution to this problem is to systematically design enhancers with specific binding sites while avoiding generating other important binding sites in that context (Estrada, Ruiz-Herrero, Scholes, Wunderlich, and Angela H DePace 2016a) (Crocker, Tsai, and Stern 2017b). Designed enhancers can evaluate different features, such as different versions of a binding site with different affinities, arrangements, and number of TFBBs. This systematic exploration could gradually include new TFBSs and start evaluating questions about heterotypic interactions, chromatin modifications, and synergistic activation. Using these systems, one can question TFs' role one step at a time. After these tests with specific, separate

binding sites, higher complexity sequence designs can evaluate the effect of overlapping and low-affinity binding sites. In summary, Synthetic enhancers can allow us to understand the endogenous system better and help us to choose from different hypotheses on grammar mechanisms. After finding mechanistic features on synthetic enhancers, one can look in the endogenous genome for these features or their absence due to evolutionary constraints.

In addition to a systematic enhancer design, a Null model has to be generated to see the expected behavior regarding expression patterns after inserting DNA sequences of any kind in an organism. For this purpose, I use a set of synthetic random sequences whose expression output was evaluated across different stages of development.

3.2.4 Quantitative modeling of Synthetic enhancers: First-principles

In this work, I gradually built several models, step by step, from the simplest set of enhancers to a more general model for early embryonic developmental enhancers. By following this approach, I could choose and discard among the models that can equally fit the data due to the large number of parameters in more complex enhancer architectures and avoid overfitting. Similar systematic approaches have also been applied in other *Drosophila* enhancers and in bacterial cis-regulation to avoid the simultaneous fitting of many parameters and ambiguous predictions (Samee et al. 2015) (Razo-Mejia et al. 2018) (Y. J. Kim et al. 2022).

The gradual fitting scheme works by choosing the best mechanistic models that can explain the simplest dataset, in this case only *Bcd* enhancers, and when adding new TFs, evaluate which models still can explain the expanded dataset, which role plays the newly added TF, and if new heterotypic interactions can happen (Fig. 14A). The models that I can test are the most general thermodynamic models, which take into account the affinity of the TFBSs and the concentration of a TF in a given A-P position, the role of the TF, whether it functions as an activator, as a repressor, or if it is bifunctional. For this purpose, I adapted a computational implementation of thermodynamic models, known as GEMSTAT, to my sequential fitting scheme. This implementation assumes that gene expression depends on the occupancy of the promoter by the polymerase, and this value can be estimated by the probability of this event (Fig. 14 A, second panel, cartoon representation of this probability). GEMSTAT is a very convenient implementation for exploring a broad extent of models since it can automatically find TFBSs for a set of TFs while using PWMs and assign their statistical weights based on their relative affinity. Likewise, this tool can consider different modes of regulation, for example,

whether the interactions are direct with the polymerase, if there is a synergistic multiplicative role by these TFs together, and if there is inhibition by modifying the chromatin in the neighborhood of a repressor. (He et al. 2010).

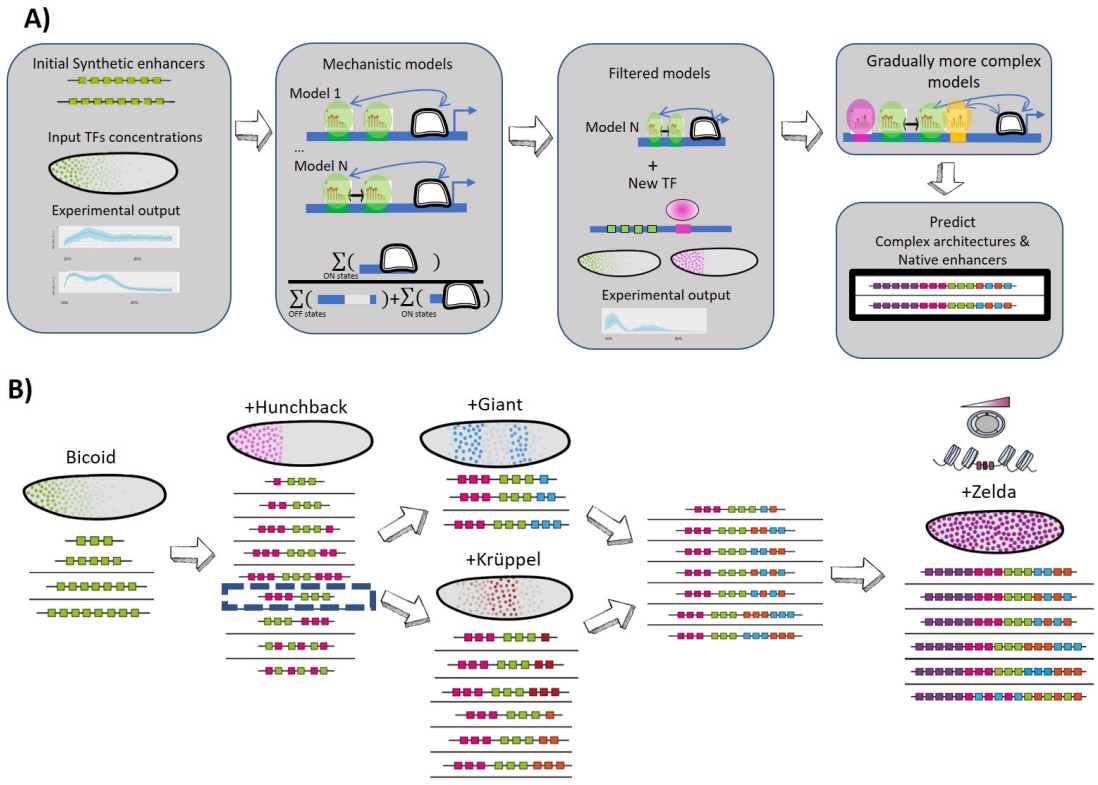


Figure 14: Model fitting schemes. A) Diagram of the iterative fitting approach for thermodynamic models that goes from the simplest enhancer architectures to more complex ones with several binding sites for different TFs. B) Sequence of enhancers sets that were fitted to focus on TF-specific parameters at each stage

For this approach, the sequential order goes from *Bicoid* binding sites only enhancers, then *Hunchback* binding sites are added, followed by either the addition of *Giant* or *Krüppel* binding sites separately and together. Finally, *Zelda* binding sites are added.

3.3 Results

3.3.1 Generating a Null model of expression patterns using synthetic enhancers made of random DNA

Kerstin Richter, Natalia Misonou, and Rafael Galupa synthesized enhancers with random DNA sequences that followed a uniform distribution in the composition of nucleotides. These enhancers were inserted upstream of a heat shock promoter (hsp70) and a *lacZ* reporter gene. The expression patterns from these enhancers were obtained by immunostaining the galactosidase protein. Gene expression activity was measured for these random enhancers (Fig 15A-C). The expression patterns from some of these enhancers matched the spatial distribution of some TFs, especially those known to have a role in chromatin accessibility (Fig 15F-H). These findings show that even for randomized sequences, there is an extent of predictability depending on the class of TFBSs that happened to be present in these sequences.

Then, I estimated and compared the information content of some of the known TFs with chromatin accessibility activity and an activation role to see how likely these TFBSs can occur by chance (Fig 15E). For the essential TFs in the early embryo, *Bcd* binding profile has a low information content while *Zelda* binding profile has a high information content. This means that *Bcd* sites are relatively easy to find by chance, but still, the embryos that had *Bcd* TFBSs did not show activity (Fig 15I). This is consistent with the idea that *Bcd* works in homotypic clusters (Lifanov et al. 2003).

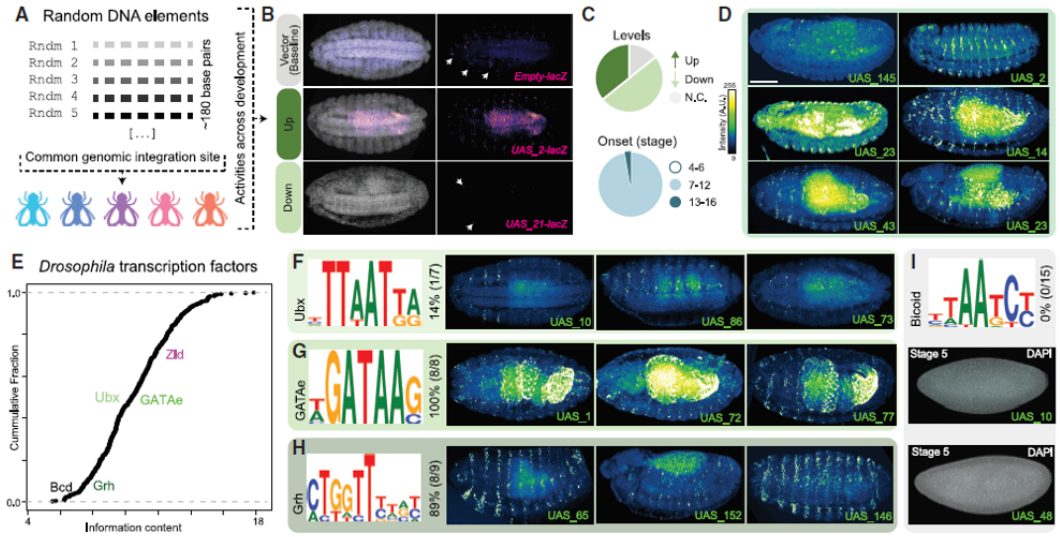


Figure 15: Predictability of expression patterns and its limitations on a synthetic randomized DNA set of Enhancers (Galupa et al. 2023). A) Randomized DNA sequences were synthesized and inserted in a landing site of *Drosophila melanogaster* with a reporter gene. B-D) Most of these sequences showed positive or negative activity, indicating its possible role in gene regulation. E) Different essential transcription factors were ranked based on their information content and role in development. F-H) Some of these sequences, when scanned for their TFBSs, show a characteristic pattern corresponding to their respective regulatory TFs.

The experiments used for this analysis were done by Rafael Galupa, Natalia Misunou, and Kerstin Richter. The statistical and motif analyses were performed by Gilberto Alvarez.

3.3.2 Constraints in early embryo enhancers

Interestingly, most of these enhancers had any activity in the late-stage embryos, while none had activity in the early embryo (Fig 16A). From the expression patterns from the active enhancers in the late embryo, I mapped significant TFBSs with high affinity. To understand why there is no observed early embryo expression with randomized sequences of DNA, I tested if this set of random sequence enhancers had a composition in the number of TFBSs favorable for TFs expressed in later stages (Fig 16B). Measuring the total number of binding sites, I observed more binding sites for late TFs than early ones (Wilcoxon test, p-val < 0.05). This is due to the presence of more TFs in later stages than in the early ones, and when I normalize the number of TFs in each stage, this difference disappears. More importantly, even though the number of late TFBSs is higher, this is not a plausible hypothesis to explain the activity differences between early and late stages since, in the lines that show a loss of activity in the late stage (named inactive), the number of TFBSs for late TFs in inactive enhancers is even most likely to be higher than in the ones that showed activity in the late stage (Fig. 16B) (Wilcoxon test, p-val < 0.001). Another feature that I tested was the enhancer lengths since it is known to be important in distinguishing enhancers by complexity (Lily Li and Wunderlich 2017). No differences were observed with enhancer length distribution between the active and inactive randomized DNA sequences.

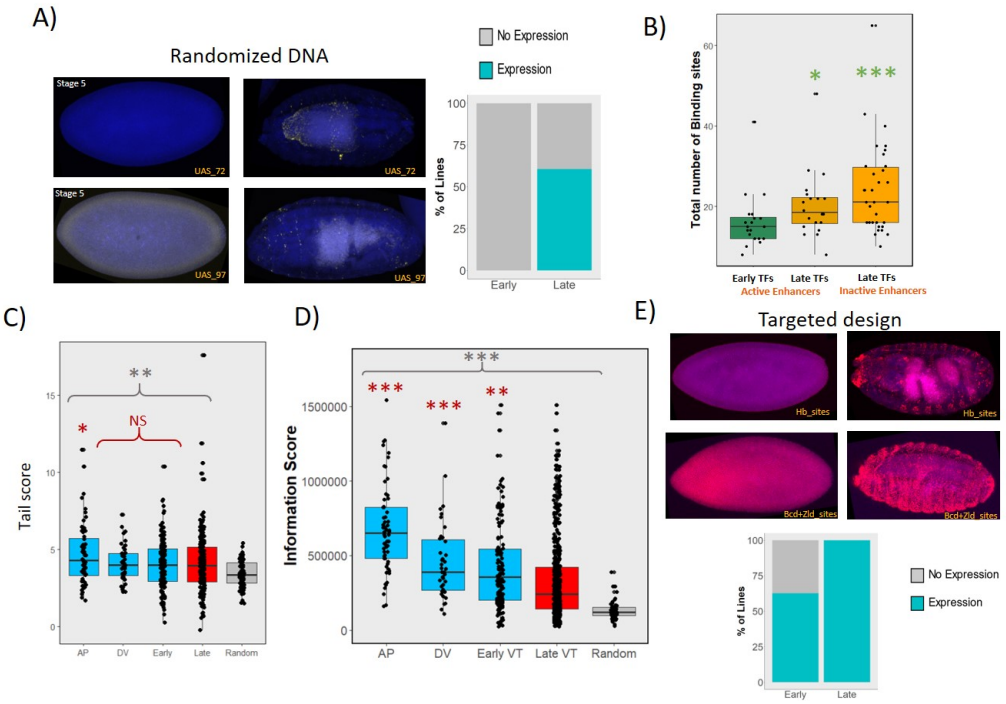


Figure 16: Comparison of different enhancer families with a randomized DNA set of enhancers. A) None of these randomized cis-regulatory sequences showed expression in the early embryo, compared with later stages, where more than a third showed a gain of activity. B) The number of TFBSs for early or late TFs is not enough to explain the observed activity patterns since the inactive enhancers in the later stage have a similar number to the active ones in the late stage. C) The homotypic tail score can differentiate endogenous enhancers from randomized enhancers but can only differentiate Anterior-Posterior developmental enhancers from Late enhancers. The rest of the early enhancers do not score differently from the Late enhancers. D) The information score allows me to distinguish early and late enhancers. E) All the enhancers that were designed targeted with specific known TFBSs show late expression activity, including those that could not get expression in the early embryo.
The experiments used for this analysis were done by Rafael Galupa, Natalia Misunou, Kerstin Richter, and Gilberto Alvarez.

I tested if endogenous early embryo enhancers count with specific features distinguishing them from late enhancers in *Drosophila melanogaster*. For example, since it is known that some TFs come in homotypic clusters, I tested the hypothesis that these random enhancers lack homotypic clusters since they are very unlikely to happen by chance and that they would be distinguishable from groups of endogenous enhancers known to have homotypic clusters (Fig 16C) (Long Li et al. 2007). I selected different sets of enhancers; for example, I used a set of enhancers responsible for Anterior-Posterior and Dorsal-ventral patterning for the early embryo (Papatsenko, Goltsev, and Michael Levine 2009) (Lily Li and Wunderlich 2017). Additionally, I used the set of enhancers of the Vienna Tiles that have been validated with reporter lines and checked for activity at different stages (Kvon, Kazmar, et al. 2014b) (Lily Li and Wunderlich 2017). I applied the Fluffy-tail test, which has been suggested as a measure for homotypic clusters since it looks for the overrepresentation of k-mers (Abnizova et al. 2005) (Long Li et al. 2007). Comparing the distributions of these sets of enhancers with the randomized DNA enhancer set, I see that the early enhancers, in this case, the AP and VT (early) enhancers, have a higher tail score than the randomized enhancers (Wilcoxon test, p-val < 0.001). I also tested if the VT late enhancers differed from the VT early enhancers' tail score. However, I did not observe differences (Wilcoxon test, p-val=NS).

The previous metric provided only a partial distinction between our random and endogenous enhancers and no difference between late and early enhancers composition. I attempted other metrics that could encompass more general features of TFBS composition in enhancers. One of these metrics uses the information content of a TF, which tells me the probability of observing a TFBS by chance. First, I compared the information content of late versus early TFs, but no differences were found (Fig. 17B) (Wilcoxon test, p-val=NS).

Recently, a modified version of an information content metric was implemented to get this score at the level of a whole enhancer, encompassing different numbers of binding sites for different TFs (Lily Li and Wunderlich 2017). Although the length of the sequence is part of the information of an enhancer, I am interested in the intrinsic sequence composition features that have shaped enhancers in different developmental contexts. For this reason, I modified this metric by normalizing the probability of observing this enhancer by each enhancer's length (See Methods).

Using the new information score per enhancer, it was possible to distinguish late and early enhancers from different subsets (Wilcoxon test, p-value<0.001) (See Fig. 16D). The information content depends on the technology used to get the

binding profile of a TF (See Fig. 17A). Statistical tests show that these differences among the different technologies were significant (Kruskal Wallis, p-val $<$ 2e-16). To see if these results are consistent based on a single technology, I selected only the PWMs obtained with a bacterial 1-hybrid experiment and Sanger sequencing since this subset was the most numerous one to avoid losing data. This analysis showed that for the PWMs from Sanger sequencing, I observed the same trend of distinguishability of late enhancers versus early enhancers. Also, late enhancers tend to decrease their information content, behaving more similarly to random enhancers, although they still have a significantly higher information content.

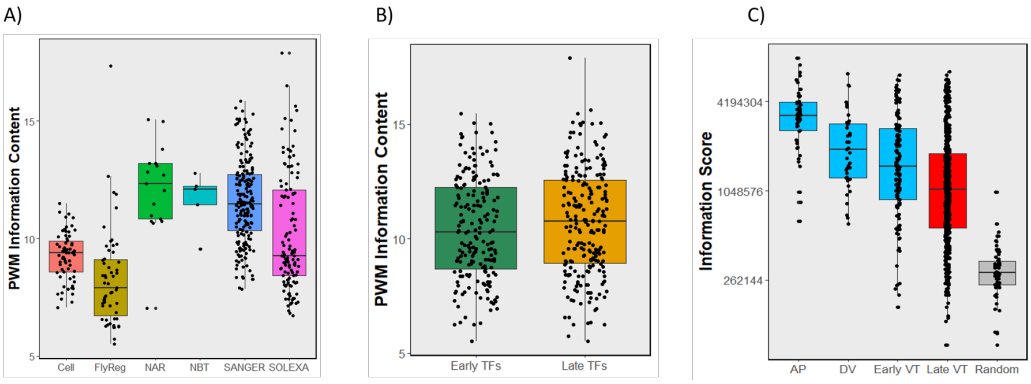


Figure 17: Technology used for TF Information content doesn´t affect the enhancer family trend. A) Information content for each TF depends on the technology that is used to infer a PWM. B) There are no differences between information content of TFs from different stages. C) The same trend was maintained when I delimited the PWM technology for bacterial 1-hybrid sequenced by Sanger. Early enhancers show a higher information score than late enhancers.

So far, these analyses have been done using only the early-stage TFs since the question concerns the absence of expression in the early stage. Alternatively, I compared the random enhancers using early and late TFs, but this did not affect the original results since the information scores of random enhancers were still significantly smaller relative to the endogenous enhancers (Wilcoxon test, p-value$<$0.001). Also, I tested if the random enhancers, active in the late stages, have a higher information content than the inactive ones, and no differences were found.

This set of randomized DNA enhancers opened a question of the constraints in the sequence space to achieve expression in the early developmental stages. To see how likely and predictable early expression patterns are, I compared our randomized DNA enhancer observations with a synthetic set of enhancers that contains

binding sites for TFs active in the early stages. From the set of randomized DNA sequences, none had expression activity in the early embryo, while around 35% had a gain of expression in the late embryo. On the other hand, from our targeted set of synthetic-designed enhancer sequences, I see that around 60% of them had activity in the early embryo. Interestingly, from a selected subset of 8 of these enhancers, some with early activity and some without, all of them had activity in the later stages. This is consistent with the observations from the random DNA enhancers, where later stages seem more permissive in generating gene expression (See Fig. 16E).

3.3.3 Encoding patterns using synthetic enhancers made of known motifs

Now that I see that these synthetic enhancers can drive early expression, new questions emerge with this dataset, for example, what kind of TFBSs are required to generate early expression? and How the concentration and position of the input TFs and the type and arrangement of TFBSs will shape the output expression?

A mechanistic point of view was selected for a quantitative understanding of the required logic to build different patterns from this set of TFBSs. For this purpose, I trained thermodynamic and statistical models to test their capacity to predict the position and intensity of gene expression in reporter lines. These models take the concentrations of TFs and different genomic architectures of synthetic enhancers as input. Different concentrations of the input TFs can play a role in the observed patterns in the multicellular context since each cell can have a different TF concentration. On top of that, these regulatory rules become complex since the regulatory regions can be densely encoded with many sites for different TFs. This can make difficult to interpret the experiments without a quantitative model.

An automatic implementation for thermodynamic models known as GEMSTAT, was used as a starting point to explore the different putative mechanisms behind well-known TFs and their binding sites. The set of synthetic enhancers I used, include different combinations for the known binding sites of *Hb*, *Bcd*, *Gt*, *Kr*, *Zld*. I made an automated iterative implementation from GEMSTAT, a tool that trains different thermodynamic models and evaluates possible mechanisms to help understand and predict complex synthetic expression patterns (He et al. 2010).

This set of synthetic enhancers has different levels of complexity in the arrangements of TFBSs. The simplest enhancers have only *Bicoid* binding sites, and I

chose this set as the starting point for the iterative approach. This set of enhancers contains 3, 5, 7, and 8 *Bcd* binding sites. These enhancers were used to train and select putative models for *Bicoid* mechanisms and modes of regulation. As expected, the expression level of the reporter increases with the increasing number of **Bcd** binding sites. In the case of 3 *Bcd* binding sites, the signal could not be detected from the background. Interestingly, even for this simpler architecture, GEMSTAT could not accurately fit these enhancers' expression boundaries since they seem quite sharp given their position, which is more anterior than expected.

I numerically measured the expression profile's steepness to evaluate the observed expression boundaries by plotting Bicoid concentration against the enhancer expression. The steepness values were also measured for all model fits and then ranked, but none of the models could reach the observed steepness (Figure 18B). Additionally, I explored if there was a tendency to have sharper boundaries depending on the mode of regulation, but no specific trends were observed. Another way to measure the steepness is through the Hill function, which has been used for molecular saturation through binding. The Hill function for a *Hb* enhancer, known as HbP2, has been reported to be around 5 (Park et al. 2019b). For the endogenous *Hb*, it has been reported to be 5 with protein immunostaining and 6 with RNA *In situ* hybridization (Gregor et al. 2007) (Park et al. 2019b). In this case, I find a large Hill-coefficient corresponding to approximately 9, which cannot be fitted by a classical thermodynamic model of *Bicoid* in equilibrium (Figure 18C). Since this enhancer has 8 *Bcd* binding sites, a Hill coefficient of 8 or 9 can imply the need to consider higher-order cooperativities and non-equilibrium mechanisms (Park et al. 2019b) (Martinez-Corral et al. 2024). Additionally, I fitted the number of binding sites using the same parameters from a known *Hb* thermodynamic model (Phillips 2020). This model suggested 11 *Bcd* binding sites to observe this level of sharpness.

To study the sharpness of our synthetic lines more deeply, I selected the line with 8 *Bcd* binding sites. I hypothesized that changing the promoter would affect the sharpness if the effect comes from higher-order cooperativities involving the promoter. For this reason, Rafael Galupa and I designed 2 alternative enhancers with the same 8 *Bicoid* binding sites with different promoters, the promoter for the *Armadillo* gene and the synthetic core promoter DSCP. Blanca Pijuan-Sala and I performed the experiments to get the RNA expression profiles for these reporter lines. Blanca and I only observed expression from one of these enhancers, specifically the one with the DSCP promoter. Blanca and I collected the images for this enhancer and the control with an HSP70 promoter.

From these different promoter lines, I compared their expression profiles and

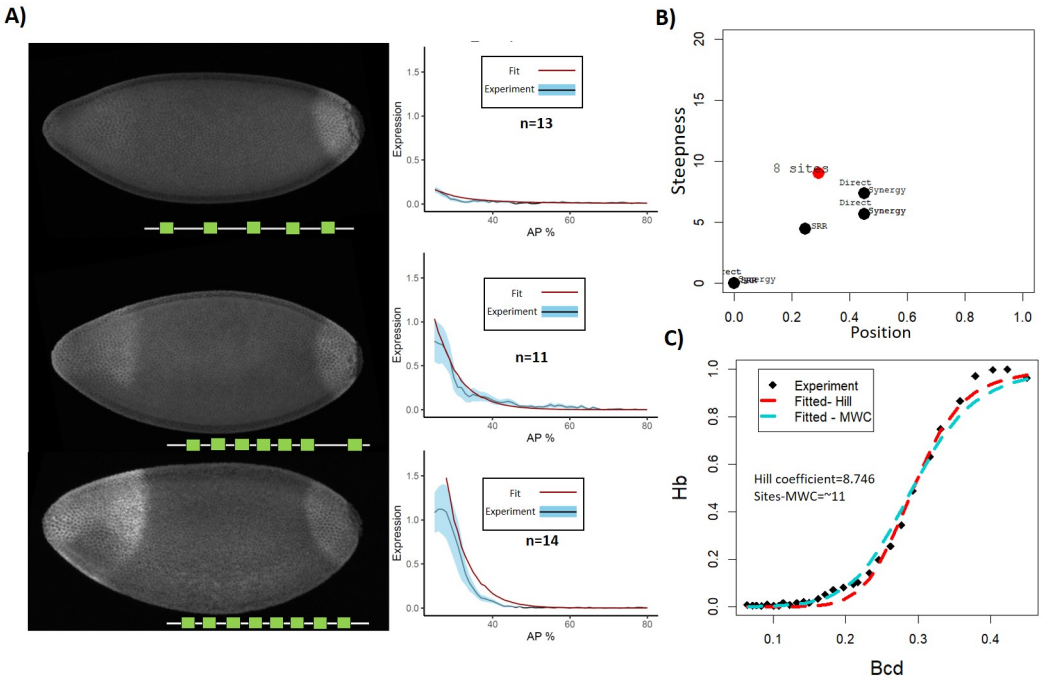


Figure 18: Fitting of Enhancers with different number of *Bicoid* binding sites. A) Enhancers that contain 8 binding sites for *Bicoid* show a sharp gradient decay that the generalized models cannot explain. B) Steepness and position cannot be reached by any tested models. C) A fitting with a Hill function shows a high Hill coefficient, which a classical thermodynamic model for this enhancer cannot fit for this number of binding sites.
The experiments used for this analysis were done by Rafael Galupa. Gilberto Alvarez performed the models, analyses, and plots.

steepness (Fig. 24A). I measured the sharpness by fitting a Hill function. The Hill coefficient for the lines with the HSP70 promoter was 1.66 larger than that of the DSCP promoter Hill coefficient. Additionally, the expression levels of both enhancers differ; the enhancer with a DSCP promoter has higher expression, and its variation is larger. This finding paves a new direction for questions regarding promoter involvement in higher-order cooperativities in sharp boundaries.

In order to continue with more complex enhancer architectures, from all the fitted models, the ones with the most lines with a high correlation coefficient (See Methods) were selected for the next fitting stage with a new TF. For the modeling, I added a new TF with its corresponding expression profiles and PWM, which was used to fit the new set of synthetic enhancers with additional TFBSs. This

approach was used for all the next steps, where I introduced a new TF to the model.

In the next stage of our fitting scheme, there is a series of enhancers with *Bcd* binding sites and different numbers and arrangements of *Hb* sites. Interestingly, the fittings with GEMSTAT for this enhancer series always showed high anterior expression for lines where no anterior expression was observed in the experiments (Fig. 19A). I implemented an anterior expression correlation score to quantify the experimental data's agreement with our thermodynamic models (Figure 19B). This score quantifies how the model correlates with the reporter lines lacking anterior expression (Figure 19C). While checking the lines that lacked anterior expression compared with the nest GEMSTAT predictions, it turns out that these lines had 2 isolated binding sites of *Hb*. I hypothesized that these 2 isolated binding sites of *Hb* could have a role of repression, which is consistent with propositions that *Hb* can have a dual role (Figure 20A). GEMSTAT can test the hypothesis that a TF can have a dual role, but it can do it only in the simplest assumption of having an alternative version of a TF with an opposite role. In this case, it seems that the dual role can only happen when 2 binding sites are present in isolation, and that's why GEMSTAT fails in doing the fitting.

A previously suggested hypothesis with a pair of endogenous enhancers proposes that *Hb*, when present in a palindromic array, can form dimers with an inhibitory effect (Papatsenko, Goltsev, and Michael Levine 2009). However, I observe that the condition of a palindromic site is unnecessary, at least for these cases, since our enhancer sequences do not include such palindromes. In both scenarios, dimers can form and have a repression activity. Additionally, one can observe that the repressor dimer loses its inhibitory effect when an additional site of *Hb* is present. In other words, the balance between activation and repression should ensure that repression only exists in the isolated pairs of binding sites without invoking special conditions. To test if this is possible under the most general thermodynamic model for a repressor *Hb* dimer in equilibrium with activator *Hb* monomers, I built the model for each of our enhancers (See Methods). I performed an MCMC parameter fitting to see if the suggested dimer would behave as a repressor. The results from the MCMC parameter fitting show that the dimer-polymerase interaction parameter (Crp, see functions in the Methods) reflects the possibility of this dimer as a repressor since its value is less than 1.

With this model, I can observe a significant alternation (Correlation Coefficient of 0.97, p-value$<$0.01) between high and lower anterior expression, depending on the isolation of *Hb* pairs (Fig. 20D). Nonetheless, even though this model can generate this behavior, it also shows a trend of higher activation in the anterior

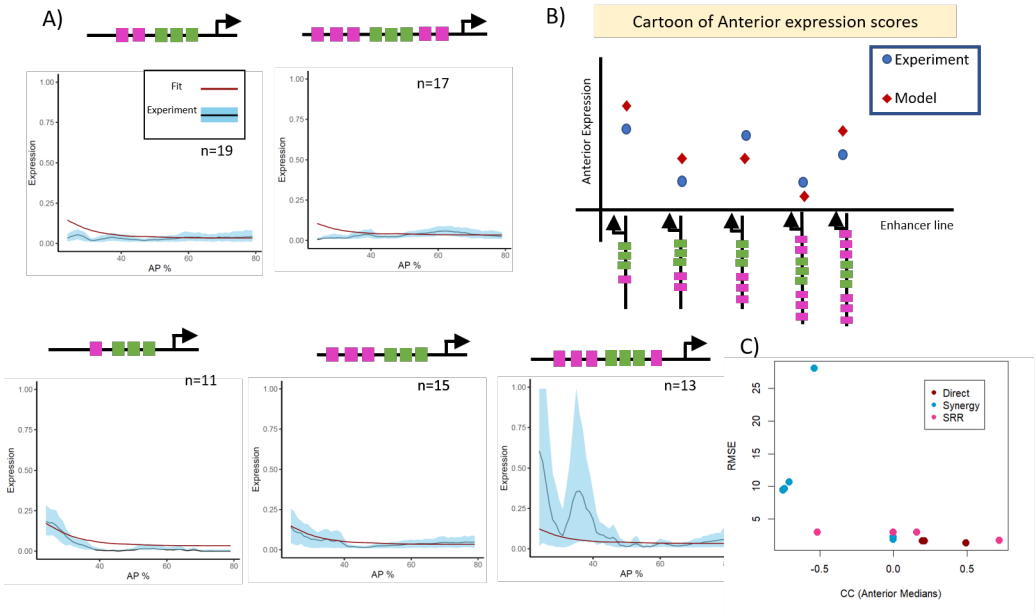


Figure 19: Generalized thermodynamic model fittings for *Bicoid* (green) enhancers with different arrangements and numbers of *Hunchback* binding sites (magenta). A) AP expression profiles for each tested enhancer and one of the best fits. B) Schematic of the expected anterior expression trend with this set of enhancers. C) The models with higher correlation coefficients and lowest root mean squared error include Direct and Range repression modes of regulation.
The experiments used for this analysis were done by Rafael Galupa. Gilberto Alvarez performed the models, analyses, and plots.

that decreases to the posterior. This trend is not observed in the experiments in the enhancers where there are isolated pairs of *Hunchback* binding sites. This means other mechanisms can be at play, such as the dimer being a more stable molecule where there is not an equilibrium exchange with its monomers or interacting heterotypically with activators like *Bicoid*. Including these additional mechanisms can be an interesting direction in which to test this hypothesis with newer experiments. The possibility that a dimer can have a repression role while having its activator monomers in equilibrium tells us that under the most general circumstances, without adding extra parameters or conditions, this mechanism could explain the experimental observations with our synthetic lines.

For the next stage of the proposed fitting scheme, I only fitted the parame-

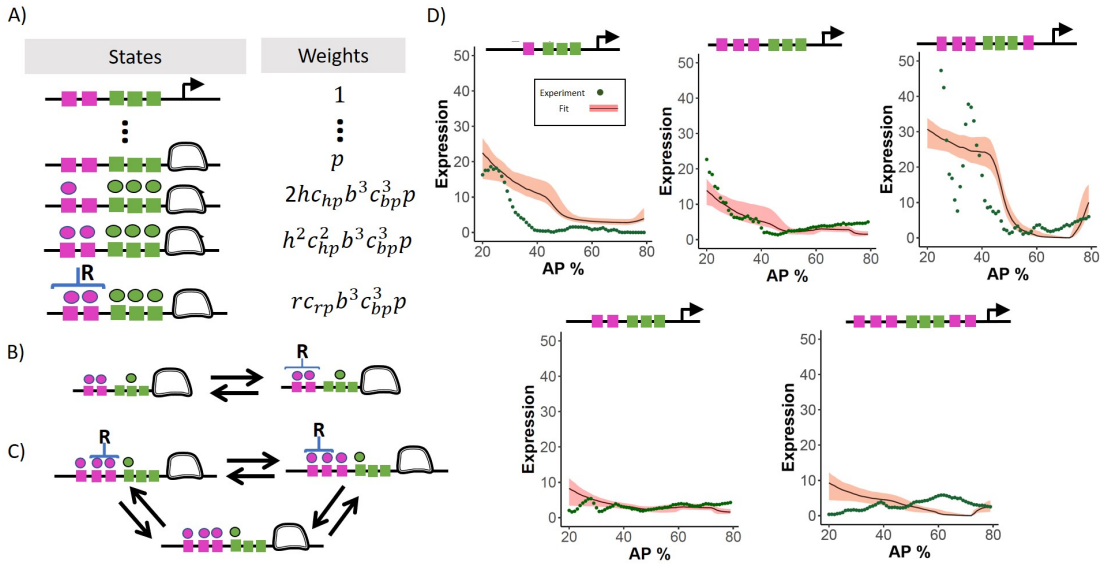


Figure 20: MCMC fittings for a simplified *Hb* Dimer model for a set of *Bicoid* binding sites with different arrangements and numbers of *Hunchback* binding sites enhancers. A) A repressor dimer can be formed when 2 *Hb* binding sites are contiguous. B) The simplest model assumes the repressor dimer can coexist in equilibrium with 2 bound *Hb* monomers. C) The system gets more complex with more binding sites. With 3 *Hb* binding sites, the simplest model assumes the coexistence of an activator with the repressor dimer or 3 *Hb* activator monomers. D) MCMC fitting results: in red is the model prediction with its variance, and in green are the experimental data points. Gilberto Alvarez performed the models, analyses, and plots.

ters for a baseline architecture of binding sites, including three *Hb* sites followed by three *Bcd* sites(See Figure 14B, second step). This baseline architecture will be present in most of our synthetic enhancers. As no more pairs of isolated *Hb* sites exist in these constructs, these specific conditions and parameters for the new *Hb* dimer model won´t be considered for the next fitting steps. This allowed me to keep exploring mechanisms through the generalized models' framework from GEMSTAT since I do not require additional special modes of regulation, such as a repressor dimer. The next fitting steps include separate binding sites for 2 binding site versions of *Giant* and *Krüppel*.

After fitting the baseline *Hb* and *Bcd* enhancer, the best models were chosen for fitting independent sets of enhancers that include additionally *Giant* or *Krüppel* binding sites. Interestingly, increasing the number of binding sites of either TF

increases the fluorescence signal; this can be explained by a) the bifunctional role of these TFs under specific contexts, b) the existence of a background expression that gets subtracted and generates an impression of a gain of activity and c) specific heterotypic interactions.

It is suggested that a central expression pattern can emerge with binding sites for activators such as *Bicoid* and *Hunchback* together with *Giant*, as it has been proposed for the *Krüppel* enhancer (Papatsenko, Goltsev, and Michael Levine 2009). This dataset helps to validate that hypothesis with these synthetic enhancers. Nonetheless, the fact that there are many equally good-performing models suggests that there are many putative mechanisms by which this can happen. On the other hand, for the anterior expression pattern that emerges when adding *Krüppel* binding sites, there are no reports besides the suggestion that *Krüppel* could work as an activator in lower concentrations (Sauer and Jäckle 1991). Nonetheless, these concentrations, if existent, require more sensitive quantification, and there is no data available that I could include in the model. I found a set of models that can predict the presence of anterior expression for this textitKrüppel's constructs, and these models were selected for the next fitting step.

In the next step, after selecting the top 20% quantile of putative models for *Gt* and *Kr*, the set of enhancers composed of *Hb*, *Bcd*, *Gt*, and *Kr* were fitted to get insights on possible heterodimeric interactions of *Gt* and *Kr*. All these enhancers had no reporter expression, which could imply that the role of the repressors is spatially complementary in the expression domain for these activators. Another explanation is that the repression overcomes the activation in this specific context.

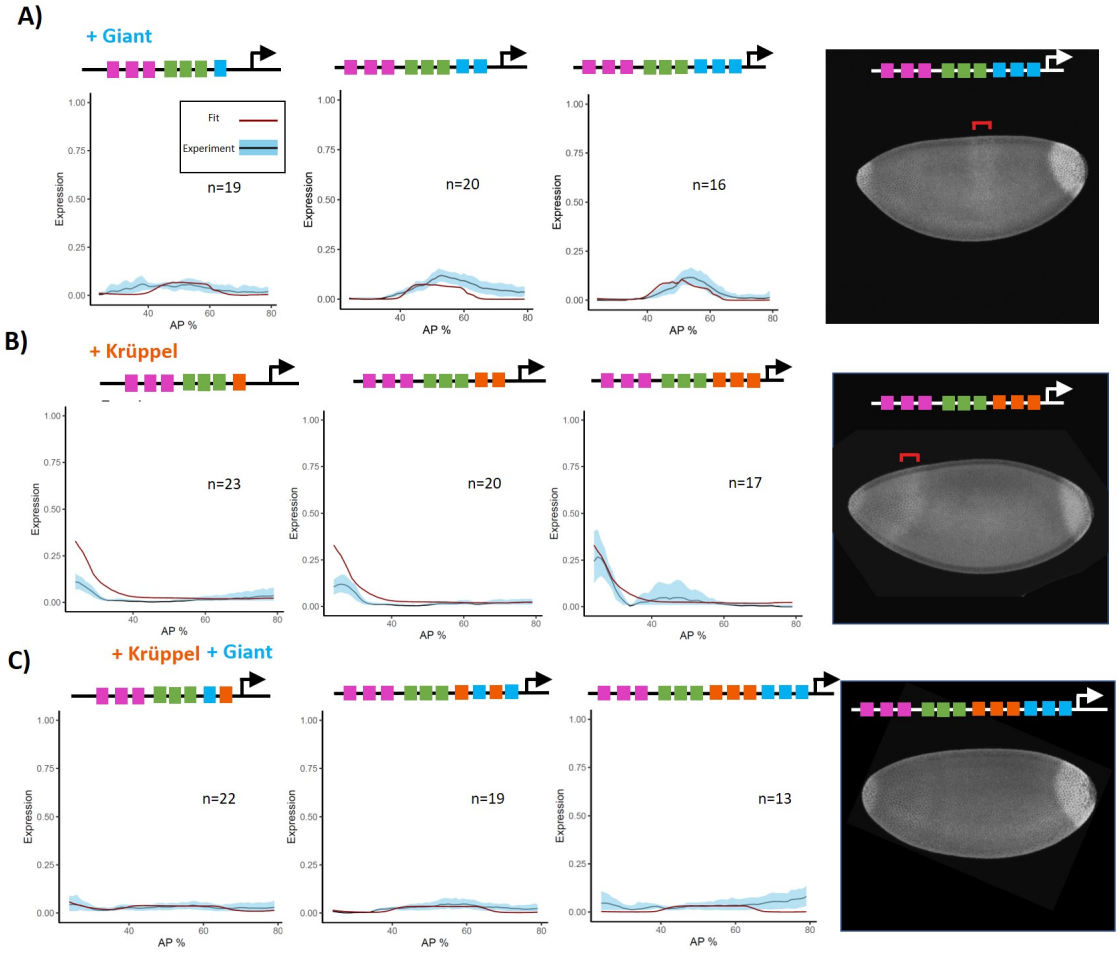


Figure 21: Fittings for enhancers with *Giant* and *Krüppel* binding sites. A) An example of one of the best fits of enhancers with *Giant* binding sites, as one can observe. There is a central expression domain, which can be predicted with a generalized thermodynamic model. This expression domain gets refined by adding *Giant* binding sites. B) An example of one of the best fits of enhancers that include *Krüuppel* binding sites. There is an anterior expression domain, which can be predicted with a generalized thermodynamic model. This expression domain gets higher by adding *Krüppel* binding sites. C) An example of one of the best fits for enhancers that include *Giant* and *Krüppel* binding sites. No expression is observed in these enhancers, and such expression levels can be fitted, fulfilling the condition of having expression when the different binding sites are added separately but lacking expression when the binding sites are together.
The experiments used for this analysis were done by Rafael Galupa. Gilberto Alvarez performed the models, analyses, and plots.

In the last step of the sequential model building, I included the enhancers with binding sites for *Bcd*, *Hb*, *Kr*, and *Gt*. As mentioned above, these enhancers did not show expression, but when adding binding sites for *Zld*, expression activity is acquired in most synthetic enhancers. From these enhancers, *Krüppel's* repression role may be surpassed by activation since the reporter expression is located in *Krüppel's* expression domain. With the results from the model fits, I can confirm that these observations are possible by the action of activators being stronger than *Krüppel's* repression or by heterotypic interactions; these patterns can be observed without invoking additional modes of regulation. However, the fact that few particular enhancers lack expression suggests the existence of specific binding site arrangement interactions that can inhibit the activation from the present TFs (Fig. 22B, left panel).

Another observation is that these models fail to fit the stripe intensity and tend to overestimate it. The right panel of Figure 22B and its corresponding embryo show that a central domain of expression is present, but the model fails to fit the intensity of the stripe. This problem can be due to the high variance in certain enhancers, which causes the mean intensity values in the bootstrapping to go very high. I am working on a new statistical analysis to solve this problem. However, it is important to highlight the relevance of including variation in fitting model schemes, like Bayesian and Monte Carlo approaches do. By adding variation in the fitting of models, one can learn more about a system since it is an additional source of information that can help distinguish equally performing mean-expression-based models.

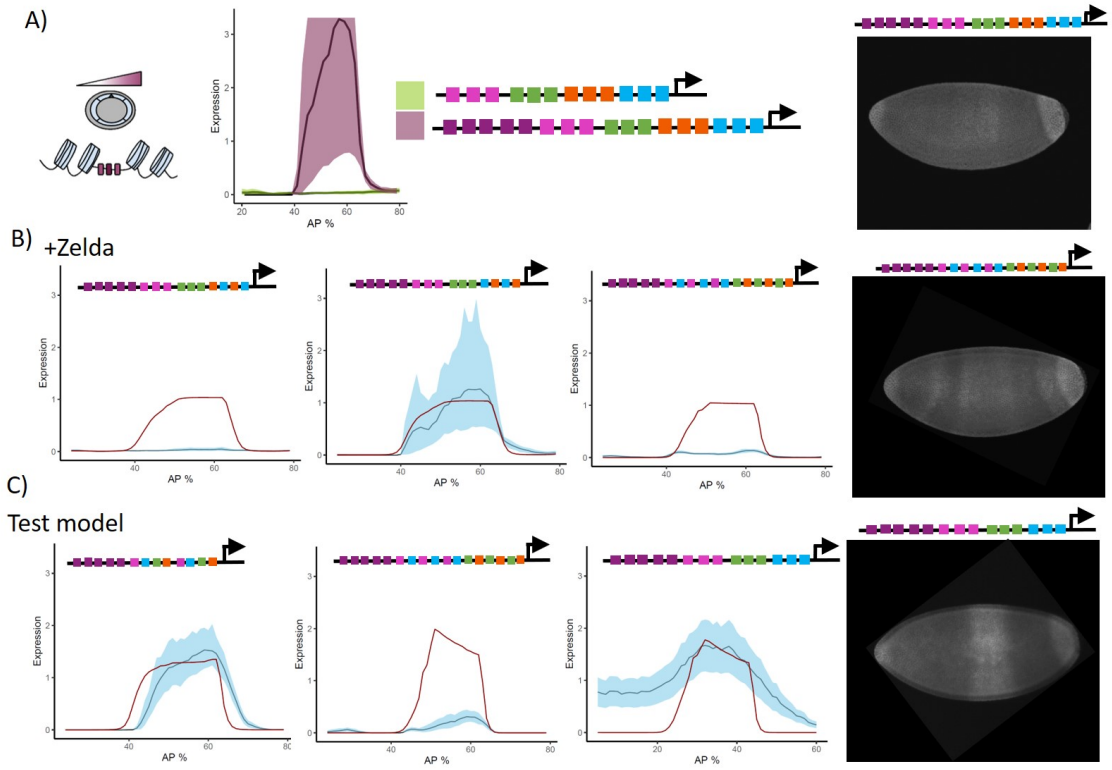


Figure 22: Fittings for enhancers when *Zelda* binding sites are added. A) Adding the pioneer activity from *Zelda* binding sites allows the expression gain in a central domain in comparison to the control line with the same architecture lacking *Zelda* binding sites. B) This central reporter expression domain can be explained by thermodynamic models, and it is observed in most of the lines with all the same TFBS but different architectures. There are some important exceptions; for example, in the left panel, no expression was observed for this architecture, and the model still suggests a central expression domain. In the right plot panel, there is a central expression domain, which can be observed in the embryo picture next to it, but the levels do not correspond to the fit. C) The testing of the best models with a different set of enhancers shows that most of the observed stripes can be predicted and that the shape of the stripe can change on different enhancer architectures. However, a similar problem is presented with some lines that behave exceptionally where it fails to predict the observed pattern.
The experiments used for this analysis were done by Rafael Galupa and Esther Karumbi. Gilberto Alvarez performed the models, analyses, and plots.

The addition of *Zelda* binding sites opened a new scope to explore the grammar since many of these interactions with the known binding sites, capable of generating stripe patterns alone, did not show any activity. After adding *Zelda* binding sites, most of these synthetic enhancers generated a broad central stripe pattern, which a large set of models can also explain. After training the last set of models with different enhancer architectures that include binding sites for all 5 TFs, I explored how the best models from this last fitting step would perform with another set of synthetic enhancers not included in the model training. The best-performing models predict the existence of a stripe when all the TFs binding sites are present, which matches most of the patterns. However, they fail for specific cases where anterior expression is strong (Figure 22C, right panel) or the stripe is narrowed due to overlapping binding sites (Figure 23B, right panel). Many of these lines with stripe patterns also present a high variability among embryos, making the mean expression level a misleading metric. In this case, the pattern position and shape could be a better metric for estimating the accuracy of the models.

Some of the most interesting patterns obtained with these transcription factors are narrow stripes, such as the ones from the pair-rule genes. The narrow stripe pattern is formed when an overlapping *Krüppel* is added on the *Bicoid* sites (Fig). As mentioned in the first chapter, this configuration occurs in the second stripe enhancer of *eve*. No models could reduce the stripe from the non-overlapping line to the size and position of the stripe in the overlapping sites. This can be explained by the fact that generalized thermodynamic models cannot consider the overlapping of sites since its effects will depend on different regulatory modes. These observations provide a new direction for the project, which I will deepen in the discussion since controlling these overlapping sites can be crucial for positioning narrow stripe patterns.

3.4 Discussion and Conclusions

3.4.1 Statistical features of endogenous enhancers vs Random DNA

Generating synthetic random enhancers allows the creation of a null model in different developmental contexts. This approach helped me to uncover specific features of early transcriptional enhancers since randomized sequences can only be expressed in later stages. This led me to hypothesize that late embryonic enhancers would behave more similarly to the randomized sequences in specific sequence features. An information content analysis showed a lower information density for late embryonic and randomized enhancers compared to early embryonic enhancers.

These informational metrics seem consistent even for different technologies that allow the inference of PWMs. Something that needs to be considered is that this is under the assumption that the PWM sampling of TFs has been enough and is representative of the actual extent of TFs, which haven't been described yet for any organism. On top of that, many TFs do not work in a way that makes current PWMs a good way to represent their function; for example, low-affinity sites can have important activity, and Non-DNA binding domains can lead the recruitment of the transcriptional machinery (Kribelbauer et al. 2019) (Kumar et al. 2023).

One possible future direction from these results is the evaluation of developmental constraints in the evolution of enhancers by comparing them with other *Drosophila* species. Early enhancers seem more constrained in their sequence space upon mutations, and their features require more information than later-expressed enhancers. This led me to question if the early enhancers are under selective pressure since the embryonic development of *Drosophila melanogaster* has a fast rate compared with other species of *Drosophila*. For example, *Drosophila virilis* is known to take as long as almost double the time to reach the late blastoderm stage (Ninova, Ronshaugen, and Griffiths-Jones 2014).

To better understand the constraints of gene expression during development, it would be necessary to explore how these randomized enhancers behave in more developmental stages and tissues. This is an ongoing project with Anna-Lena Vigil and Ian Laiker, who are testing these enhancers for transcriptional activity in adult testis and several larval tissues.

3.4.2 Learning grammar through synthetic enhancers

As Lewis Carroll wrote, "What do you consider the largest map that would be really useful?" this question has inspired novelists, philosophers, and plenty of discussions on creating mathematical models. In this work, using generalized thermodynamic models as a departure point has provided a way of mechanistically explaining how expression patterns can emerge from a simple set of rules of a few TFs. Nonetheless, the cases where these models fail to explain the experimental data are an opportunity to learn new mechanisms unknown to a given system. As I found at the beginning of this work, certain enhancers seem to contain higher information, implying that they are more constrained in their evolution, and probably their TF composition has a high parametric space. Complementary to these findings, the experiments with the synthetic enhancers with the 5 TFs, essential for embryo development, show that to achieve expression, these enhancers need a considerable amount of binding sites. Likewise, the expression patterns sometimes

showed specific context-dependent behaviors based on the particular arrangements of binding sites. Additionally, the generalized thermodynamic modeling approach helped me to associate putative explanations with expression patterns that could not have been easily predicted without models.

As it has been seen and suggested, in some contexts, animal enhancer grammar is a highly parametric system that will require extensive screenings to understand a single endogenous enhancer example. This is where uncovering general trends can be important in reducing the space to relevant features, such as specific constraints in sets of enhancers due to their biological context. These general constraints can teach us how evolutionary and developmental processes interact at the level of cis-regulation. If what I observe for these synthetic enhancers is a pervasive rule for constrained developmental enhancers, these systems could require millions of experiments and still miss some essential features since each spatial and temporal context will matter, too. The advantage of this work is that it was done with an extensive gradual exploration of a synthetic system that allows testing the role of different TF binding sites in isolation and combination.

What is the purpose of learning Enhancer grammar? Using the philosophy of a synthetic biologist, If one can synthesize patterns at will, this would mean that the system is understood. Nonetheless, for highly parametric systems, one can run into the problem of being unable to prove a general model. Similar to the problem posed by Chomsky on the existence of a Universal grammar in Linguistics. There is no way to prove the reduction of the problem to a generalizable set of grammar rules. Nonetheless, the impact of grasping some of these rules is big. For example, finding major controlling transcription factors can have broader medical and biotechnological applications, like controlling gene expression inside a specific tissue using genetic engineering. From the fundamental science point of view, for example, in Evolutionary biology and Developmental biology, the challenge is larger, and one way to delimit it is to learn the essential elements to generate a pattern and how they impact the evolution of cis-regulation.

Here, I want to point out a difference between having an enhanceosome model, which does not necessarily mean having a high number of parameters. In the enhanceosome, it could be that the parameter space is simple, and it's just a fragile enhancer because it happens to be constrained in its current state, for example, just by steric occupancy. For the AP enhancers, my findings indicate that specific sections of the enhancer should be maintained through evolution despite the assumption that compensatory evolution is the main driver of their evolution. One future direction for testing the essentiality of some of these modules in generating

expression patterns such as the ones from the Pair-Rule genes is testing overlapping binding sites, which generates precise, narrower stripe patterns (Fig. 23B).

3.4.3 The role of *Giant* and *Krüppel*

Interestingly, the observations with these TFs show that either the background signal decreases or these TFs work as activators in low concentrations. For *Krüppel* it has been shown in vitro that it can work as an activator in its monomeric form that can be present in higher quantities in low concentrations (Sauer and Jäckle 1991). On the other hand, assuming these TFs only work as repressors, The total background reduction hypothesis is possible under the assumption that low concentrations of these TF would be uniformly distributed along the embryo. Another possibility is that under specific interactions with *Hb* or *Bcd* these TFs can work as activators. So far, the propositions of *Krüppel* being an activator in low concentrations come from some *in vitro* assays and outside the second stripe enhancer context. Giant has been proposed as an auto-activator in its enhancer (Hoermann, Cicin-Sain, and Jaeger 2016). Hunchback has been proposed to have a dual activator and repressor role based on cooperativities regulating Kruppel enhancers (Papatsenko, Goltsev, and Michael Levine 2009).

The fact that I could not find models capable of explaining the specific enhancers, with all TFs present, that lacked expression in the central stripe area indicates that very specific interactions may be at play. These interactions can depend on the TF binding site arrangement order or more complex traits. It could be possible that other TFs that I am not aware of might be binding the enhancer sequence and causing this specific effect.

3.4.4 Generation of stripe patterns

Most models could achieve stripe patterns, even by the simple rules of generalized thermodynamic models. This tells us that it is relatively easy to generate stripe patterns in the center of the embryo using these 5 TFs. Given the spatial domain and role of some of these TFs, such as *Krüppel*, which works as a repressor, these central stripes might seem counterintuitive. Interestingly, when overlapping *Bcd* and *Kr* binding sites are present, the central stripe gets narrower towards its posterior part. However, GEMSTAT is not capable of taking into account overlapping binding sites, which can be seen in the lack of fitting for the lines where *Bcd* and *Kr* binding sites are overlapping (Figure 23 B, left panel).

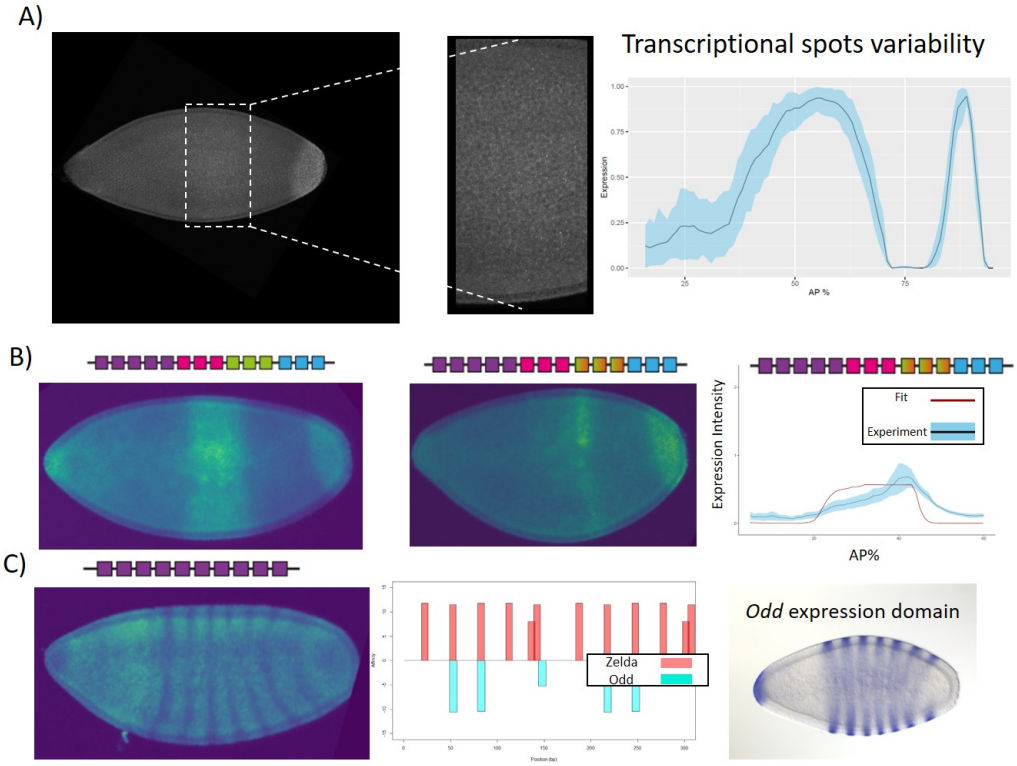


Figure 23: Putative directions and Ongoing work. A) Some of these enhancers show sparse, strong transcriptional spots that become a source of variation. Additional work must be done to include variation in an automated system for thermodynamic models and the role of opening the chromatin by *Zelda*. B) The overlapping of TFBSs of *Bcd* and *Kr*, can make a narrower stripe of expression, prompting to a candidate mechanism for endogenous stripe enhancers. C) A synthetic enhancer intended to be only made of *Zelda* binding sites, was found to overlap with binding sites for *Odd*. The image of the *Odd* expression domain is from the Berkeley Drosophila Genome Project.
The experiments used for this analysis were done by Rafael Galupa and Esther Karumbi. Gilberto Alvarez performed the models, analyses, and plots.

Overlapping sites could provide space for a more precise spatial activation domain since the repression mechanism of *Krüppel* can work by blocking the binding of the activators at high concentrations. Additionally, for other stripes, such as the second stripe enhancer of *eve*, a possible mechanism is that the overlapping of *Giant* with either activator is required to get the anterior stripe pattern since the enhancers that only include *Bicoid* and *Krüppel* overlapped binding sites cannot generate a second stripe. For example, in the second stripe of *eve*, there are *Giant* binding sites which are overlapping *Hunchback* sites. A combination of different types of overlapping sites and activation strength could play a role in positioning these narrow stripes, and it is an interesting direction for completing this work. For this, an expansion of the thermodynamic model scheme will be required to consider the possibility of overlapping binding sites. Additional experiments with different number types and overlapping sites will be required to test this model.

Another interesting stripe pattern was observed on enhancers designed to be composed of only *Zelda* binding sites. This pattern shows a negative stripe pattern, which can be explained by the presence of a TF, which affects the activation by *Zelda* binding sites. I performed a motif analysis for putative pair-rule genes that could bind this enhancer. Interestingly, I found that *odd-skipped* can overlap binding sites of *Zelda*, which can explain this negative stripe pattern due to the lack of activation by the overlapping *Zelda* or by the repressor role that this transcription factor can have (Figure 23C).

"A thing that doesn't change with time is a memory of younger days"
-Sheik, The Legend of Zelda.

3.4.5 Time dependent processes: Bistability, *Zelda* and Non-equilibrium mechanisms.

The experiments with different promoters indicate that a promoter won't only affect expression levels but also the sharpness of a pattern (Fig. 24A). It would be necessary to do further tests to understand if this can be explained only by a parameter associated with higher-order cooperativities or if there are mechanisms outside thermodynamic equilibrium playing a role here. The presence of higher-order cooperativities has been suggested in a systematic study to explain the role of *Runt* binding sites that have been inserted in the *Hunchback* promoter (Y. J. Kim et al. 2022). Higher order cooperativities have also been suggested as a putative candidate for explaining the sharp patterns from the *HbP2* promoter; nonetheless, in that work, the cooperativities failed to reach the observed experimental steepness, and mechanisms outside equilibrium were suggested to explain this phenomenon (Park et al. 2019b). Based on the experiments I report in this work, this doesn't exclude the fact that TF-TF-Pol interactions might happen.

Another putative mechanism for sharpness comes from the theory of Dynamical systems. When a system has autoactivation, bistability can emerge depending on the shape of a gene regulatory function (Fig. 24B, left panel) (Alon 2006). The intersection points from the degradation and activation function indicate the fixed points where the system will evolve towards or move away, depending on their nature. In yellow, I highlighted the stable fixed points that indicate where the system will move depending on its initial concentration (Fig. 24B, central panel). As each of the nuclei in the *Drosophila* embryo has a different starting concentration of the morphogen, depending on the shape of the activation function, the distance between the stable points can modify the sharpness of a pattern. For example, it has been suggested that bistability from *Hb* contributes to its sharp patterns by mutating the capacity of *Hb* to self-regulate. (Lopes et al. 2008). Additionally, *Hb*'s pattern dynamics can be explained when spatial bistability is considered (Perkins 2021).

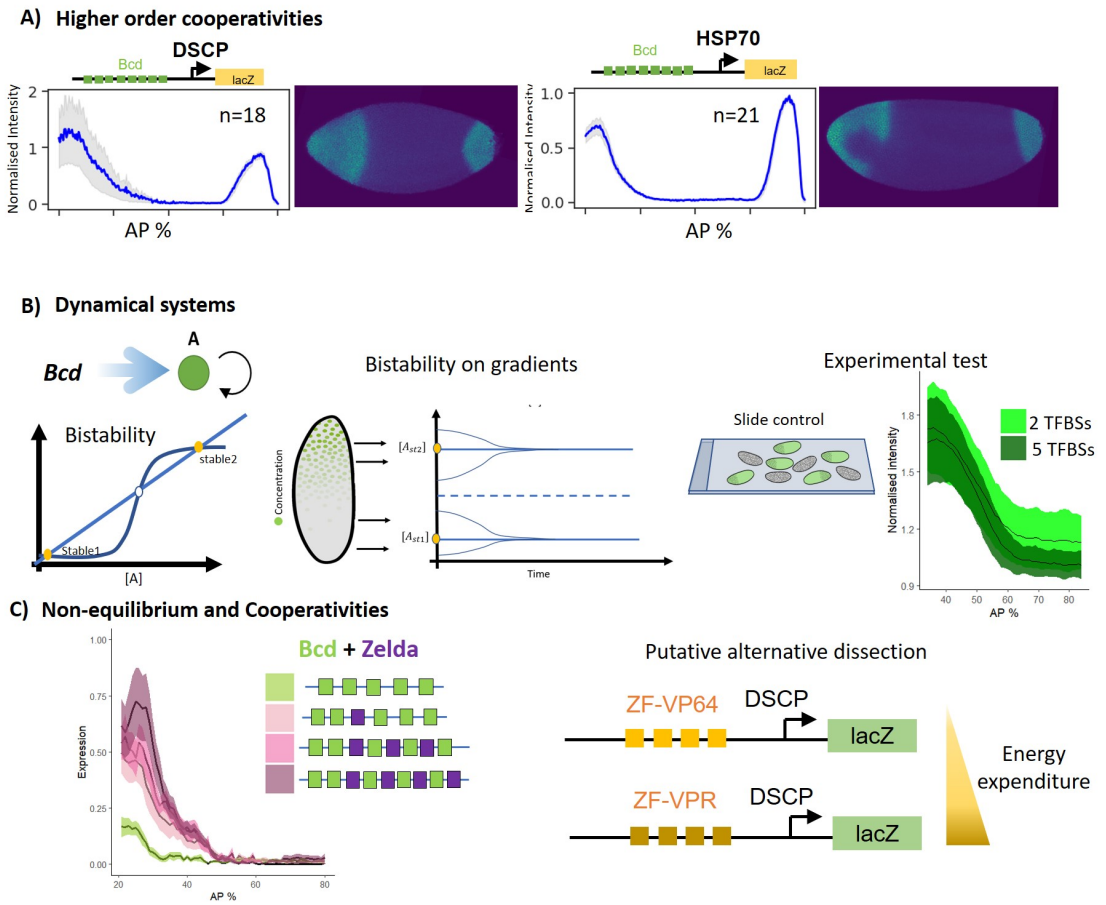


Figure 24: Future directions in the sources of sharpness A) The use of different promoters in *Bcd* enhancers produced different expression levels and different steepness values B) A self-regulating system can generate bistability and different steepness values in a multicellular context, where the initial concentrations are different in each nucleus. An experimental test was performed, and a higher steepness and lower posterior expression were detected in the enhancer with a higher number of self-activation binding sites, suggesting the role of spatial bistability in this system. C) Synthetic enhancers with an increasing number of *Zelda* binding sites show changes in steepness, which can be due to additional interactions or the pioneer activity of this TF. Additionally, a more targeted way to test the role of cooperativities and energy expenditure in patterning would be with a synthetic system that has activation domains with different extents of propensities in interaction formation.
The experiments used for this analysis were done by Gilberto Alvarez, Rafael Galupa, and Blanca Pijuan-Sala. Gilberto Alvarez performed the models, analyses, and plots.

To test the role of bistability in a synthetic system, Rafael and I designed enhancers composed of a Hunchback promoter, which receives the morphogen input, 0,2 and 5 UAS binding sites for Gal4, a DSCP promoter, and the coding sequence of Gal4, which will generate the self-regulation. Based on the hypothesis that different shapes of the activation sigmoid function can affect the sharpness of a pattern, the steepness was measured for this experiment. In the first experimental test, I focused only on steepness. As expected, the enhancer with 5 UAS sites generated a larger steepness than the control and the enhancer with 2 UAS sites line (Hill coefficient: 5 UAS= 5, Empty= 4.7, 0 UAS= 4.7 and 2 UAS= 4.5). Interestingly, the line with 2 UAS binding sites had a less sharp boundary than the control, contrary to what one could expect since there are additional activator binding sites. More interestingly, the 5 UAS sites line showed a lower intensity signal in the posterior half of the embryos.

Rafael and I performed an additional experiment with control embryos on the same slide to measure intensity levels at the whole embryo level and further tested if this is a consequence of the dynamics of the concentrations given the stable fixed points in the system (Figure 24B, right panel). As it is observed in the plot, the line with 2 UAS sites has a higher posterior expression (t-test, p-value < 0.02), meaning that either there is a gain of posterior expression from this specific enhancer or the line with 5 UAS binding sites pushes down a basal level of signal that is present in the second half of the control and 2 UAS sites enhancers. Both possible scenarios are consequences of spatial bistability playing a role in the observed steepness differences. This experiment also confirms a 0.5 higher Hill coefficient difference in the 5 UAS sites enhancer than in the 2 UAS sites enhancer line. Additional experiments would need to be done to see if there is a basal signal in the posterior of the embryos and if a different number of UAS sites would allow me to better map the parameters that control spatial bistability.

Finally, the role of *Zelda* was explored while adding different numbers of its binding sites in an enhancer with *Bicoid*. The gradual addition of *Zelda* binding sites showed increased intensity and expression activity in more posterior regions (Fig. 24C, left panel). This posterior shift is consistent with recent observations with a parallel synthetic system that uses *Bicoid* and *Zelda* binding sites enhancers (Fernandes et al. 2022). Adding a *Zelda* binding site drastically reduced the steepness since more expression was allowed in the posterior (from a Hill coefficient of 6.8 to 3.7). However, I also observed that increasing the number of *Zelda* binding sites increased the steepness, where the line with four *Zelda* binding sites has a Hill coefficient of 5. This increase in steepness from the addition of *Zelda* binding sites hasn't been reported yet, and it might require a temporal dissection to see if

this is due to the pioneer activity of this TF or if it's due to TF-TF interactions. Another way to explore the role of non-equilibrium processes in the activity from an enhancer and patterning would be to design synthetic TFs such as Zinc Fingers (ZF) with different activation domains. These activation domains can have different propensities to form interactions; for example, the activation domain VP16 has been observed to have lower interactions than the ones with a VPR activation domain. Additionally, the experiments with VPR suggested the existence of energy-dependent steps in the transcriptional process of its reporter (Trojanowski et al. 2022). Nevertheless, the role of these processes remains to be uncovered in the context of patterning.

3.5 Contributions

3.5.1 Randomized set of DNA

Gilberto Alvarez (me) did all the statistical analyses for the different sequences, including motifs, k-mers, and information content. I also did image analyses and microscopy for the antibody staining of synthetic enhancers.

Rafael Galupa, Kerstin Richter, and Natalia Misonou did the experimental *Drosophila* work for the randomized DNA, which included fly crosses, embryo collections, fixations, antibody staining and microscopy.

Justin Crocker supervised the project, contributed intellectually and funded the project.

3.5.2 Modeling the enhancer grammar of TFs in the early embryo

Gilberto Alvarez (me) wrote the scripts and performed the sequential modeling scheme and the exploratory pipeline that uses GEMSTAT for the selection of best-fitting models. I developed the image and statistical analysis pipeline for the embryos. I performed statistical analyses of the modeling results. I did sequence motif analyses of these synthetic lines to corroborate and explore possible phenotypes. I performed fly crosses for the different promoter lines and microscopy image acquisition of these *Drosophila* genetic lines.

Rafael Galupa supervised me with the directions for the project. He performed the sequence design and experimental work for most of these lines, including fly husbandry, embryo collections, RNA *in situ* hybridization, and microscopy image acquisition.

Mindy Perkins supervised me on the modeling and fitting scheme I performed with GEMSTAT and MATLAB. She also supervised us in the image analysis processing.

Esther Karumbi performed the experiments on several synthetic enhancer lines, including fly crosses, embryo collections, RNA *in situ* hybridization, and imaging in the microscope.

Blanca Pijuán-Sala did the embryo collection, fixations and RNA *in situ* hybridization, and microscopy image acquisition of the *Drosophila* genetic lines with *Bicoid* binding sites and different promoters.

Garth Ilsley participated in the design and motif analysis for the binding site selection on these sequences.

Justin Crocker supervised, participated in the design, and funded this project.

3.6 Methods

3.6.1 Random library and Enhancers with a targeted design library synthesis

Justin Crocker, Rafael Galupa, Natalia Misunou, and Kerstin Richter were responsible for the sequence design, synthesis, and generation of the fly lines of the enhancers with randomized DNA sequences. Justin Crocker, Garth Ilsley, Rafael Galupa, and Kerstin Richter were responsible for the sequence design, synthesis, and generation of the fly lines of the enhancers with a targeted design for known TFs.

Genscript synthesized both sets of sequences. The only selection requirement for the randomized DNA sequences was size, which was approximately 180 bp. These sequences were mixed and assembled using digestion enzymes (HindIII,XbaI, and BsaI) and inserted in a pLacZattB plasmid. The randomized DNA library was injected in a VK33 landing site line, and the targeted design enhancer library in an attP2 line. Genetivision made injections of both libraries. The homozygosing and genotyping of these lines were made by Natalia Misunou, Kerstin Richter, and Rafael Galupa.

3.6.2 Sequence analyses

I used PWMs for *Drosophila melanogaster* from FlyFactorSurvey(L. J. Zhu et al. 2011b). For the early TFs I restricted the search to motifs from TFs present from stage 1 to 6, and for the late TFs I used the 13 to 16 stage(Lily Li and Wunderlich 2017). Motif search analysis was done using FIMO (Grant, Bailey, and Noble 2011), setting a threshold p-value of 0.01.

For the Tail Fluffiness score, I estimated the distribution of all 5-mers for each enhancer sequence, and for each 5-mer, a list of its similar words is estimated (1 mismatch away). The maximal similar word list is obtained according to (Abnizova et al. 2005), and a null distribution is estimated by generating 50 shuffled sequences maintaining its nucleotide composition. The fluffiness coefficient is estimated by:

$$F = \frac{L_{max} - L_{random}}{\sigma_{random}}$$

Where L_{max} is the actual size of the maximal word list in the sequence, and L_{random} and σ_{random} are the mean and standard deviation estimations from the shuffled set of sequences. This score comes from (Abnizova et al. 2005).

The Enhancer Information score is a modification from the one proposed in (Lily Li and Wunderlich 2017). First, I estimated the information content per TF using the Kullback-Leibler distance (Schneider et al. 1986).

$$I_{TF} = \sum_{i=1}^{S} \sum_{j=A}^{T} p_{i,j} log2 \frac{p_{i,j}}{b_j}$$

Where $p_{i,j}$ is the probability of seeing the "j" nucleotide at the position "i". b_j is the estimated background frequency in the genome of "j". I performed the cumulative distribution of these values in R. Then, according to (Lily Li and Wunderlich 2017), one can estimate the probability of seeing an enhancer with a certain motif composition by:

$$P_{enh} = \frac{\sum_i n_i 2^{-I_i}}{n_{sites}}$$

Where P_{enh} is the average probability of seeing a given motif hit composition in an enhancer. n_i is the number of TFBSs for TF "i", I_i is the corresponding information content for that TF, and n_sites is the total number of TFBSs for the different TFs that have a hit on that enhancer. Since this score depends on the enhancer length, I normalized it by the length of each corresponding enhancer, and then I took the reciprocal value for visualization in my analysis.

I validated the synthetic sequences with a targeted design to see if they were depleted from known motifs. These motifs are PWMs selected for the early embryo, and they are included in SiteOut. SiteOut was used to deplete additional control sequences for Bicoid binding sites (Estrada, Ruiz-herrero, et al., 2016).

3.6.3 Embryo manipulation for synthetic enhancers

Embryos were collected and fixed after 4 hours, using the same protocol described in Chapter 1. Rafael Galupa and Esther Karumbi performed the *in situ* hybridization protocol for this set of designed enhancers according to the protocol in the Methods section in Chapter 1. To have an internal control, a co-staining was implemented. The staining was done for *LacZ*, which was used as a reporter gene, and for *Forkhead* as an internal control (Wunderlich, Bragdon, and Angela H DePace 2014).

3.6.4 Image processing for Synthetic enhancers

Embryo images were acquired with confocal microscopy (Zeiss LSM 880) at 20x. Images were processed using a combination of automated scripts with manual curation. Rafael Galupa, Esther Karumbi, and I performed a maximal projection for the channel with the output fluorescence. I developed an automated pipeline for embryo orientation and measuring AP expression profiles. I describe the details of this pipeline in Chapter 1. The profiles were smoothed for each of the embryos and rescaled to make composite profiles of the AP axis for all the embryos. The background was obtained from the first half of the embryo to acquire the signal. Then, as these embryos were co-stained with *Forkhead*, I normalized the intensity across the AP axis by the posterior 90% of the embryo. Once the data was acquired for all the embryos, bootstrapping was done to obtain confidence intervals for the expression profiles.

3.6.5 Automated thermodynamic model implementation.

An automated pipeline based on general thermodynamic models was implemented. The thermodynamic functions and fitting were done through GEMSTAT (He et al. 2010). The PWMs used were the standard for the Gap and maternal genes already included in the software. This new pipeline was adapted to explore different modes of regulation, different possibilities of TF roles, and possible pairwise cooperativities to uncover unexplored mechanisms for each of the factors or grammar rules that arise with them based on our experimental observations. In each step of the sequential process, the best models were selected based on how many lines can be fitted according to a correlation coefficient higher than 0.6 (following the same threshold as (He et al. 2010)). This value is flexible enough to avoid models that will only overfit. For each model, I quantify how many lines fullfill that criteria and I select the top 20% of models from that distribution.

For the Hill function and MWC model fits for the *Bcd* lines, expression profiles were smoothed using local regression in R. Then, differentiation was used to estimate steepness and position. The Hill Function and MWC model were fitted using non-linear least scores in R. The MWC model for an enhancer regulated by *Bcd* comes from (Phillips et al. 2012) (Phillips 2020).

For the Hill function I used this formula:

$$GRF_{Hill} = \frac{lacZ_{max} Bcd^n}{Bcd^n + Bcd_{half}^n}$$

Where $lacZ_{max}$ is the maximum intensity of the reporter expression profile. Bcd is the intensity value from the *Bicoid* protein and Bcd_{half} is the value of *Bi-*

coid concentration at half of the maximal expression of the reporter."n" represents the Hill coefficient that was fitted.

For a simple thermodynamic model (MWC) of *Bcd*

$$GRF_{MWC} = \frac{e^{-\beta\Delta\epsilon}(1+\frac{[Bcd]}{K_o})^n}{e^{-\beta\Delta\epsilon}(1+\frac{[Bcd]}{K_o})^n + 1}$$

Where $e^{-\beta\Delta\epsilon}$, it is known as the Boltzmann weight, which represents a probability that is associated with the energy difference for the active and inactive states. K_o is the dissociation constant for *Bcd*. "n" represents the number of sites that were fitted.

3.6.6 Dimer model implementation

Thermodynamic models were used for each of the lines with different number and arrangements of *Hb* binding sites (See below). Parameters were fitted with Markov Chain Monte Carlo using the Matlab package MCMCstat (Haario et al. 2006). The fitting was done simultaneously for all the lines.

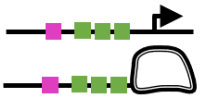
For each enhancer model, below, I show the states included in the function for estimating gene expression. To avoid redundancy, I only represent the unique states, the empty promoter, and all the possible combinations of TFs bound with a Polymerase bound. To make the partition function, one must consider the states without the polymerase bound.

For writing the thermodynamic models, I followed a similar approach from (Y. J. Kim et al. 2022). In this model, the concentrations and their dissociation constants are written as $b = [Bcd]/K_b$, $h = [Hb]/K_h$ and $r = [Hb]/(2K_r)$. The parameters for the interaction with the polymerase are written as "c" and their subindexes indicate if it is from *Bcd*, *Hunchback* or Repressor (*Hb-dimer*). The parameter values were chosen in the same range as in (Y. J. Kim et al. 2022), depending on the role of the TF. Concentrations for the TF in the AP axis were taken from (He et al. 2010).

Each line's number is the identifier assigned for each enhancer.

Line 425

1 *Hb* sites + 3 *Bcd* Sites 

States	Weights
	1
	p
	$b^3 c_{bp}^3 p$
	$3b^2 c_{bp}^2 p$
	$3bc_{bp}p$
	$hc_{hp}b^3 c_{bp}^3 p$
	$3hc_{hp}b^2 c_{bp}^2 p$
	$3hc_{hp}bc_{bp}p$
	$hc_{hp}p$

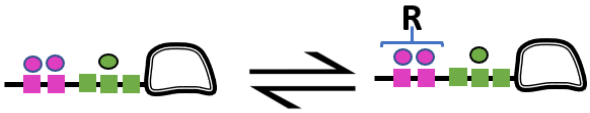
In order to avoid redundancy in the text, states without the Polymerase bound are not shown, however they were used for the estimation of the transcriptional rate. The same applies for the rest of the architectures.

Line 426

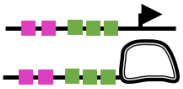
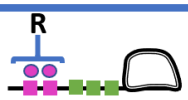
2 *Hb* sites + 3 *Bcd* Sites 

Assumption:

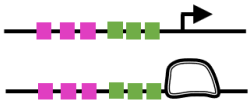
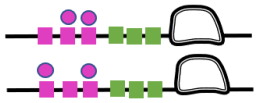
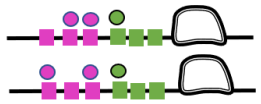
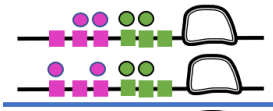
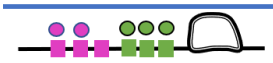
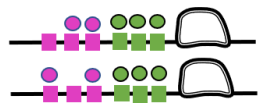
While bound to the DNA the dimer can be formed and it is in equilibrium with the 2 monomer activators from *Hunchback*



Continuation: Line 426

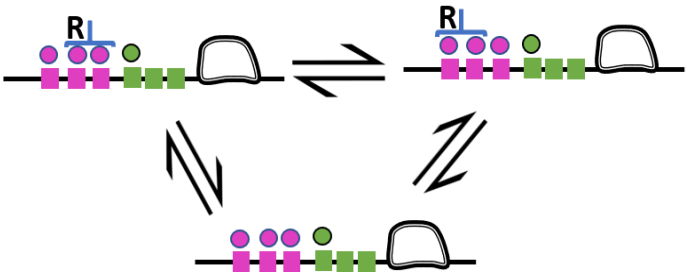
States	Weights
	1
	p
	$b^3c_{bp}^3p$
	$3b^2c_{bp}^2p$
	$3bc_{bp}p$
	$2hc_{hp}b^3c_{bp}^3p$
	$6hc_{hp}b^2c_{bp}^2p$
	$6hc_{hp}bc_{bp}p$
	$2hc_{hp}p$
	$rc_{rp}p$
	$3rc_{rp}bc_{bp}p$
	$3rc_{rp}b^2c_{bp}^2p$
	$rc_{rp}b^3c_{bp}^3p$
	$h^2c_{hp}^2p$
	$3\ h^2c_{hp}^2\ bc_{bp}p$
	$3h^2c_{hp}^2b^2c_{bp}^2p$
	$h^2c_{hp}^2\ b^3c_{bp}^3p$

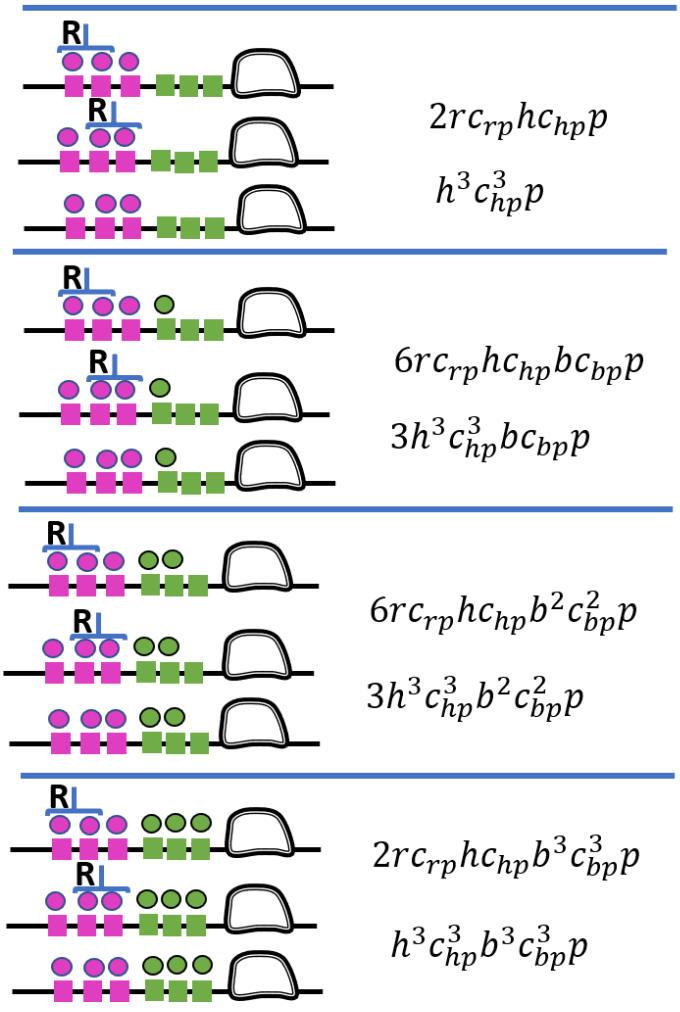
Line 420

States	Weights
	1
	p
	$3bc_{bp}p$
	$3b^2c_{bp}^2p$
	$b^3c_{bp}^3p$
	$3hc_{hp}bc_{bp}p$
	$9hc_{hp}bc_{bp}p$
	$9hc_{hp}b^2c_{bp}^2p$
	$3hc_{hp}b^3c_{bp}^3p$
	$2\,rc_{rp}p$
	$3h^2c_{hp}^2p$
	$6\,rc_{rp}bc_{bp}p$
	$9h^2c_{hp}^2bc_{bp}p$
	$6\,rc_{rp}\,b^2c_{bp}^2p$
	$9h^2c_{hp}^2b^2c_{bp}^2p$
	$2\,rw_{rp}b^3c_{bp}^3p$
	$3h^2c_{hp}^2b^3c_{bp}^3p$

Continuation: Line 420

Equilibrium states with 3 *Hb* sites occupying the 3 sites





$2rc_{rp}hc_{hp}p$

$h^3c_{hp}^3p$

$6rc_{rp}hc_{hp}bc_{bp}p$

$3h^3c_{hp}^3bc_{bp}p$

$6rc_{rp}hc_{hp}b^2c_{bp}^2p$

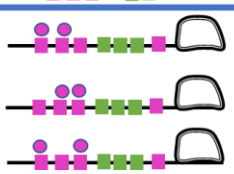
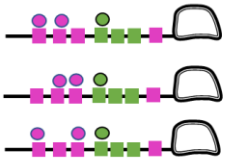
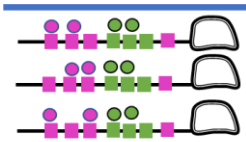
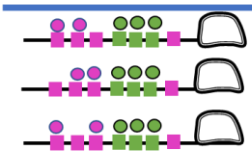
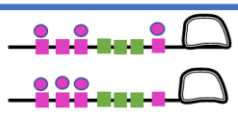
$3h^3c_{hp}^3b^2c_{bp}^2p$

$2rc_{rp}hc_{hp}b^3c_{bp}^3p$

$h^3c_{hp}^3b^3c_{bp}^3p$

Line 427

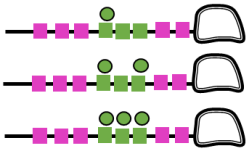
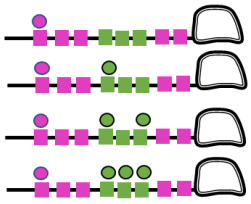
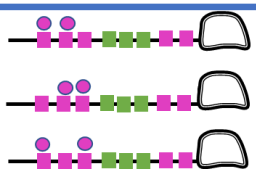
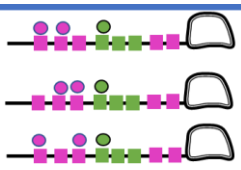
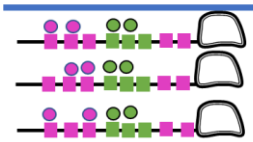
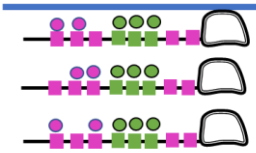
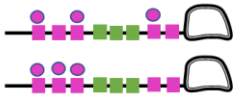
4 *Hb* sites + 3 *Bcd* Sites

States	Weights
	1
	$3bc_{bp}p$
	$3b^2c_{bp}^2p$
	$3b^3c_{bp}^3p$
	$4hc_{hp}p$
	$12hc_{hp}bc_{bp}p$
	$12hc_{hp}b^2c_{bp}^2p$
	$4hc_{hp}b^3c_{bp}^3p$
	$2\,rc_{rp}p$ $6h^2c_{hp}^2p$
	$6\,rc_{rp}bc_{bp}p$ $18h^2c_{hp}^2bc_{bp}p$
	$6\,rc_{rp}b^2c_{bp}^2p$ $18h^2c_{hp}^2b^2c_{bp}^2p$
	$2\,rc_{rp}b^3c_{bp}^3p$ $6h^2c_{hp}^2b^3c_{bp}^3p$
	$4\,h^3c_{hp}^3p$ $4\,rc_{rp}hc_{hp}p$

States	Weights
	$12\ h^3 c_{hp}^3 b c_{bp} p$
	$12\ r c_{rp} h c_{hp} b c_{bp} p$
	$12\ h^3 c_{hp}^3 b^2 c_{bp}^2 p$
	$12\ r c_{rp} h c_{hp} b^2 c_{bp}^2 p$
	$4\ h^3 c_{hp}^3 b^3 c_{bp}^3 p$
	$4\ r c_{rp} h c_{hp} b^3 c_{bp}^3 p$
	$2 r c_{rp} h^2 c_{hp}^2 p$
	$h^4 c_{hp}^4 p$
	$6 r c_{rp} h^2 c_{hp}^2 b c_{bp} p$
	$3 h^4 c_{hp}^4 p$
	$6 r c_{rp} h^2 c_{hp}^2 b^3 c_{bp}^3 p$
	$3 h^4 c_{hp}^4 p$
	$2 r c_{rp} h^2 c_{hp}^2 b^3 c_{bp}^3 p$
	$h^4 c_{hp}^4 p$

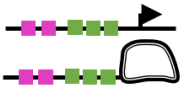
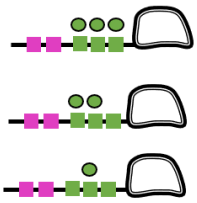
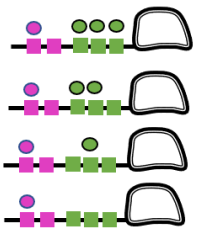
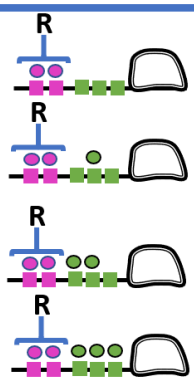
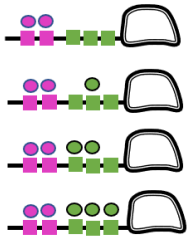
Line 428

5 *Hb* sites + 3 *Bcd* Sites

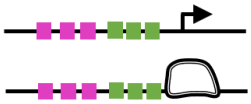
States	Weights
	1
	$3bc_{bp}p$ $3b^2c_{bp}^2p$ $3b^3c_{bp}^3p$
	$5hc_{hp}p$ $15hc_{hp}bc_{bp}p$ $15hc_{hp}b^2c_{bp}^2p$ $5hc_{hp}b^3c_{bp}^3p$
	$3\,rc_{rp}p$ $10h^2c_{hp}^2p$
	$9rc_{rp}bc_{bp}p$ $30h^2c_{hp}^2bc_{bp}p$
	$9\,rc_{rp}b^2c_{bp}^2p$ $30h^2c_{hp}^2b^2c_{bp}^2p$
	$3\,rc_{rp}b^3c_{bp}^3p$ $10h^2c_{hp}^2b^3c_{bp}^3p$
	$10\,h^3c_{hp}^3p$ $9\,rc_{rp}hc_{hp}p$

States	Weights
	$30\ h^3c_{hp}^3bc_{bp}p$
	$27\ rc_{rp}hw_{hp}bc_{bp}p$
	$30\ h^3c_{hp}^3b^2c_{bp}^2p$
	$27\ rc_{rp}hc_{hp}b^2c_{bp}^2p$
	$10\ h^3c_{hp}^3b^3c_{bp}^3p$
	$9\ rc_{rp}hc_{hp}b^3c_{bp}^3p$
	$5rc_{rp}h^2c_{hp}^2p$
	$2r^2{c_{rp}}^2p$
	$5h^4c_{hp}^4p$
	$15rc_{rp}h^2c_{hp}^2bc_{bp}p$
	$6r^2{c_{rp}}^2bc_{bp}p$
	$15h^4c_{hp}^4bc_{bp}p$
	$15rc_{rp}h^2c_{hp}^2b^2c_{bp}^2p$
	$6r^2{c_{rp}}^2b^2c_{bp}^2p$
	$15h^4c_{hp}^4b^2c_{bp}^2p$
	$5rc_{rp}h^2c_{hp}^2b^3c_{bp}^3p$
	$2r^2{c_{rp}}^2b^3c_{bp}^3p$
	$5h^4c_{hp}^4b^3c_{bp}^3p$
	$2r^2c_{rp}^2hc_{hp}p$
	$h^5c_{hp}^5p$
	$6r^2c_{rp}^2hc_{hp}bc_{bp}p$
	$3h^5c_{hp}^5bc_{bp}p$
	$6r^2c_{rp}^2hc_{hp}b^2c_{bp}^2p$
	$3h^5c_{hp}^5b^2c_{bp}^2p$
	$2r^2c_{rp}^2hc_{hp}b^3c_{bp}^3p$
	$h^5c_{hp}^5b^3c_{bp}^3p$

Continuation: Line 426

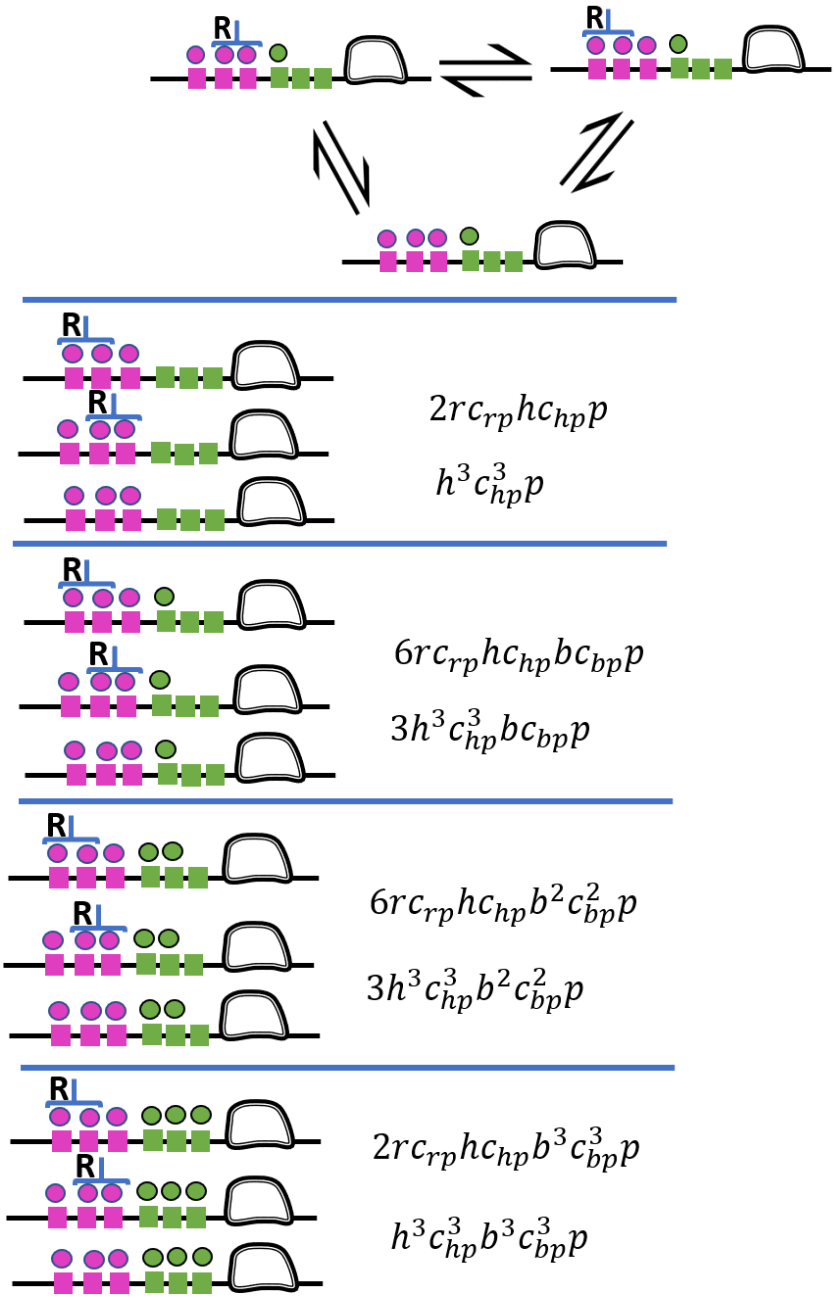
States	Weights
	1 p
	$b^3c_{bp}^3p$ $3b^2c_{bp}^2p$ $3bc_{bp}p$
	$2hc_{hp}b^3c_{bp}^3p$ $6hc_{hp}b^2c_{bp}^2p$ $6hc_{hp}bc_{bp}p$ $2hc_{hp}p$
	$rc_{rp}p$ $3rc_{rp}bc_{bp}p$ $3rc_{rp}b^2c_{bp}^2p$ $rc_{rp}b^3c_{bp}^3p$
	$h^2c_{hp}^2p$ $3\ h^2c_{hp}^2\ bc_{bp}p$ $3h^2c_{hp}^2b^2c_{bp}^2p$ $h^2c_{hp}^2\ b^3c_{bp}^3p$

Line 420

States	Weights
	1
	p
	$3bc_{bp}p$
	$3b^2c_{bp}^2p$
	$b^3c_{bp}^3p$
	$3hc_{hp}bc_{bp}p$
	$9hc_{hp}bc_{bp}p$
	$9hc_{hp}b^2c_{bp}^2p$
	$3hc_{hp}b^3c_{bp}^3p$
	$2\,rc_{rp}p$
	$3h^2c_{hp}^2p$
	$6\,rc_{rp}bc_{bp}p$
	$9h^2c_{hp}^2bc_{bp}p$
	$6\,rc_{rp}\,b^2c_{bp}^2p$
	$9h^2c_{hp}^2b^2c_{bp}^2p$
	$2\,rw_{rp}b^3c_{bp}^3p$
	$3h^2c_{hp}^2b^3c_{bp}^3p$

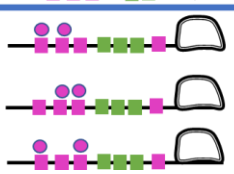
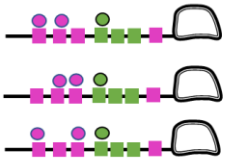
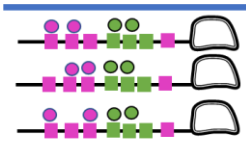
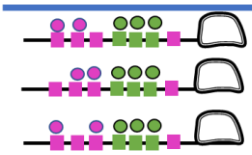
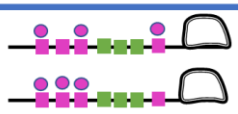
Continuation: Line 420

Equilibrium states with 3 *Hb* sites occupying the 3 sites



Line 427

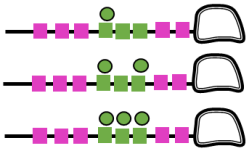
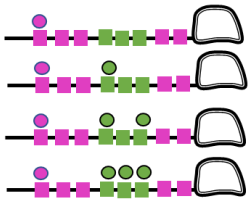
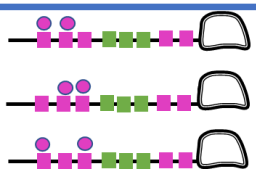
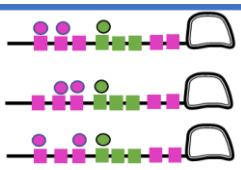
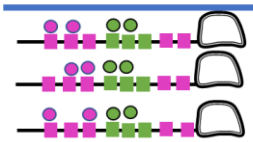
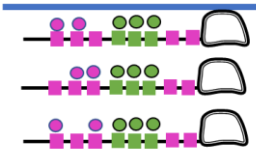
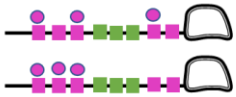
4 *Hb* sites + 3 *Bcd* Sites

States	Weights
	1
	$3bc_{bp}p$
	$3b^2c_{bp}^2p$
	$3b^3c_{bp}^3p$
	$4hc_{hp}p$
	$12hc_{hp}bc_{bp}p$
	$12hc_{hp}b^2c_{bp}^2p$
	$4hc_{hp}b^3c_{bp}^3p$
	$2\,rc_{rp}p$ $6h^2c_{hp}^2p$
	$6\,rc_{rp}bc_{bp}p$ $18h^2c_{hp}^2bc_{bp}p$
	$6\,rc_{rp}b^2c_{bp}^2p$ $18h^2c_{hp}^2b^2c_{bp}^2p$
	$2\,rc_{rp}b^3c_{bp}^3p$ $6h^2c_{hp}^2b^3c_{bp}^3p$
	$4\,h^3c_{hp}^3p$ $4\,rc_{rp}hc_{hp}p$

States	Weights
	$12\ h^3 c_{hp}^3 b c_{bp} p$
	$12\ r c_{rp} h c_{hp} b c_{bp} p$
	$12\ h^3 c_{hp}^3 b^2 c_{bp}^2 p$
	$12\ r c_{rp} h c_{hp} b^2 c_{bp}^2 p$
	$4\ h^3 c_{hp}^3 b^3 c_{bp}^3 p$
	$4\ r c_{rp} h c_{hp} b^3 c_{bp}^3 p$
	$2 r c_{rp} h^2 c_{hp}^2 p$
	$h^4 c_{hp}^4 p$
	$6 r c_{rp} h^2 c_{hp}^2 b c_{bp} p$
	$3 h^4 c_{hp}^4 p$
	$6 r c_{rp} h^2 c_{hp}^2 b^3 c_{bp}^3 p$
	$3 h^4 c_{hp}^4 p$
	$2 r c_{rp} h^2 c_{hp}^2 b^3 c_{bp}^3 p$
	$h^4 c_{hp}^4 p$

Line 428

5 *Hb* sites + 3 *Bcd* Sites

States	Weights
	1
	$3bc_{bp}p$ $3b^2c_{bp}^2p$ $3b^3c_{bp}^3p$
	$5hc_{hp}p$ $15hc_{hp}bc_{bp}p$ $15hc_{hp}b^2c_{bp}^2p$ $5hc_{hp}b^3c_{bp}^3p$
	$3\,rc_{rp}p$ $10h^2c_{hp}^2p$
	$9rc_{rp}bc_{bp}p$ $30h^2c_{hp}^2bc_{bp}p$
	$9\,rc_{rp}b^2c_{bp}^2p$ $30h^2c_{hp}^2b^2c_{bp}^2p$
	$3\,rc_{rp}b^3c_{bp}^3p$ $10h^2c_{hp}^2b^3c_{bp}^3p$
	$10\,h^3c_{hp}^3p$ $9\,rc_{rp}hc_{hp}p$

States	Weights
	$30\ h^3c_{hp}^3bc_{bp}p$
	$27\ rc_{rp}hw_{hp}bc_{bp}p$
	$30\ h^3c_{hp}^3b^2c_{bp}^2p$
	$27\ rc_{rp}hc_{hp}b^2c_{bp}^2p$
	$10\ h^3c_{hp}^3b^3c_{bp}^3p$
	$9\ rc_{rp}hc_{hp}b^3c_{bp}^3p$
	$5rc_{rp}h^2c_{hp}^2p$
	$2r^2{c_{rp}}^2p$
	$5h^4c_{hp}^4p$
	$15rc_{rp}h^2c_{hp}^2bc_{bp}p$
	$6r^2{c_{rp}}^2bc_{bp}p$
	$15h^4c_{hp}^4bc_{bp}p$
	$15rc_{rp}h^2c_{hp}^2b^2c_{bp}^2p$
	$6r^2{c_{rp}}^2b^2c_{bp}^2p$
	$15h^4c_{hp}^4b^2c_{bp}^2p$
	$5rc_{rp}h^2c_{hp}^2b^3c_{bp}^3p$
	$2r^2{c_{rp}}^2b^3c_{bp}^3p$
	$5h^4c_{hp}^4b^3c_{bp}^3p$
	$2r^2c_{rp}^2hc_{hp}p$
	$h^5c_{hp}^5p$
	$6r^2c_{rp}^2hc_{hp}bc_{bp}p$
	$3h^5c_{hp}^5bc_{bp}p$
	$6r^2c_{rp}^2hc_{hp}b^2c_{bp}^2p$
	$3h^5c_{hp}^5b^2c_{bp}^2p$
	$2r^2c_{rp}^2hc_{hp}b^3c_{bp}^3p$
	$h^5c_{hp}^5b^3c_{bp}^3p$

References

Abnizova, Irina et al. (Apr. 2005). "Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the Drosophila genome: the fluffy-tail test". en. In: *BMC Bioinformatics* 6.1, p. 109.

Ackers, G K, A D Johnson, and M A Shea (Feb. 1982). "Quantitative model for gene regulation by lambda phage repressor". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 79.4, pp. 1129–1133.

Akam, M (Sept. 1989). "Drosophila development: making stripes inelegantly". en. In: *Nature* 341.6240, pp. 282–283.

Alberts, B. et al. (2002). *Molecular Biology of the Cell 4th Edition: International Student Edition*. Routledge. ISBN: 9780815332886. URL: `https://books.google.de/books?id=ozigkQEACAAJ`.

Almeida, Bernardo P de et al. (Dec. 2023). "Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo". en. In: *Nature*.

— (Feb. 2024). "Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo". en. In: *Nature* 626.7997, pp. 207–211.

Alon, Uri (July 2006). *An introduction to systems biology*. en. Chapman & Hall/CRC mathematical and computational biology series. Philadelphia, PA: Chapman & Hall/CRC.

Andrioli, Luiz Paulo Moura et al. (Nov. 2002). "Anterior repression of a Drosophila stripe enhancer requires three position-specific mechanisms". en. In: *Development* 129.21, pp. 4931–4940.

Aparicio, S et al. (Feb. 1995). "Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 92.5, pp. 1684–1688.

Aristotle (ca. 350 BCE). *On the Generation of Animals*. The Electronic Scholarly Publishing Project.

Arnosti, D N et al. (Jan. 1996). "The eve stripe 2 enhancer employs multiple modes of transcriptional synergy". en. In: *Development* 122.1, pp. 205–214.

Avsec, Žiga, Vikram Agarwal, et al. (Oct. 2021a). "Effective gene expression prediction from sequence by integrating long-range interactions". en. In: *Nat. Methods* 18.10, pp. 1196–1203.

— (Oct. 2021b). "Effective gene expression prediction from sequence by integrating long-range interactions". en. In: *Nat. Methods* 18.10, pp. 1196–1203.

Avsec, Žiga, Melanie Weilert, et al. (Mar. 2021). "Base-resolution models of transcription-factor binding reveal soft motif syntax". en. In: *Nat. Genet.* 53.3, pp. 354–366.

Badia-I-Mompel, Pau et al. (Nov. 2023). "Gene regulatory network inference in the era of single-cell multi-omics". en. In: *Nat. Rev. Genet.* 24.11, pp. 739–754.

Balsalobre, Aurelio and Jacques Drouin (July 2022). "Pioneer factors as master regulators of the epigenome and cell fate". en. In: *Nat. Rev. Mol. Cell Biol.* 23.7, pp. 449–464.

Banerji, Julian, Sandro Rusconi, and Walter Schaffner (Dec. 1981). "Expression of a β-globin gene is enhanced by remote SV40 DNA sequences". en. In: *Cell* 27.2, pp. 299–308.

Barozzi, Iros et al. (June 2014). "Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers". en. In: *Mol. Cell* 54.5, pp. 844–857.

Barresi, M.J.F. and S.F. Gilbert (2020). *Developmental Biology*. Sinauer Series. Oxford University Press. ISBN: 9781605358222.

Berest, Ivan et al. (Dec. 2019). "Quantification of differential transcription factor activity and multiomics-based classification into activators and repressors: diffTF". en. In: *Cell Rep.* 29.10, 3147–3159.e12.

Berg, Otto G and Peter H von Hippel (Feb. 1987). "Selection of DNA binding sites by regulatory proteins". en. In: *J. Mol. Biol.* 193.4, pp. 723–743.

Bergman, Casey M, Joseph W Carlson, and Susan E Celniker (Apr. 2005). "Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster". en. In: *Bioinformatics* 21.8, pp. 1747–1749.

Berman, Benjamin P et al. (Jan. 2002). "Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 99.2, pp. 757–762.

Bintu, Lacramioara et al. (Apr. 2005). "Transcriptional regulation by the numbers: models". en. In: *Curr. Opin. Genet. Dev.* 15.2, pp. 116–124.

Boer, Carl G de and Jussi Taipale (Jan. 2024). "Hold out the genome: a roadmap to solving the cis-regulatory code". en. In: *Nature* 625.7993, pp. 41–50.

Bradley, Robert K et al. (Mar. 2010). "Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species". en. In: *PLoS Biol.* 8.3, e1000343.

Bravo González-Blas, Carmen, Seppe De Winter, et al. (Sept. 2023). "SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks". en. In: *Nat. Methods* 20.9, pp. 1355–1367.

Bravo González-Blas, Carmen, Irina Matetovici, et al. (Jan. 2024). "Single-cell spatial multi-omics and deep learning dissect enhancer-driven gene regulatory networks in liver zonation". en. In: *Nat. Cell Biol.* 26.1, pp. 153–167.

Brawand, David et al. (Oct. 2011). "The evolution of gene expression levels in mammalian organs". en. In: *Nature* 478.7369, pp. 343–348.

Britten, R J and E H Davidson (July 1969). "Gene regulation for higher cells: a theory". en. In: *Science* 165.3891, pp. 349–357.

Brown, J Lesley et al. (Jan. 2003). "The Drosophila pho-like gene encodes a YY1-related DNA binding protein that is redundant with pleiohomeotic in homeotic gene silencing". en. In: *Development* 130.2, pp. 285–294.

Carroll, Sean B (July 2008). "Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution". en. In: *Cell* 134.1, pp. 25–36.

Cerchiari, Alec E et al. (Feb. 2015). "A strategy for tissue self-organization that is robust to cellular heterogeneity and plasticity". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 112.7, pp. 2287–2292.

Chen, Haining and B Franklin Pugh (July 2021). "What do Transcription Factors Interact With?" en. In: *J. Mol. Biol.* 433.14, p. 166883.

Craig, N. et al. (2021). *Molecular Biology: Principles of Genome Function*. OUP Oxford. ISBN: 9780199658572.

Crocker, Justin, Namiko Abe, et al. (Jan. 2015a). "Low affinity binding site clusters confer hox specificity and regulatory robustness". en. In: *Cell* 160.1-2, pp. 191–203.

— (Jan. 2015b). "Low affinity binding site clusters confer hox specificity and regulatory robustness". en. In: *Cell* 160.1-2, pp. 191–203.

Crocker, Justin and Garth R Ilsley (Dec. 2017). "Using synthetic biology to study gene regulatory evolution". en. In: *Curr. Opin. Genet. Dev.* 47, pp. 91–101.

Crocker, Justin, Albert Tsai, and David L Stern (Jan. 2017a). "A fully synthetic transcriptional platform for a multicellular eukaryote". en. In: *Cell Rep.* 18.1, pp. 287–296.

— (Jan. 2017b). "A fully synthetic transcriptional platform for a multicellular eukaryote". en. In: *Cell Rep.* 18.1, pp. 287–296.

Davidson, Eric H (May 2006). *The regulatory genome*. en. 2nd ed. San Diego, CA: Academic Press.

De Robertis, Edward M (Apr. 2006). "Spemann's organizer and self-regulation in amphibian embryos". en. In: *Nat. Rev. Mol. Cell Biol.* 7.4, pp. 296–302.

Driever, W and C Nüsslein-Volhard (July 1988). "A gradient of bicoid protein in Drosophila embryos". en. In: *Cell* 54.1, pp. 83–93.

Driever, Wolfgang and Christiane Nusslein-Volhard (Jan. 1989). "The bicoid protein is a positive regulator of hunchback transcription in the early Drosophila embryo". In: *Nature* 337.6203, pp. 138–143. ISSN: 1476-4687. DOI: 10.1038/337138a0. URL: https://doi.org/10.1038/337138a0.

Durrieu, Lucia et al. (Sept. 2018). "Bicoid gradient formation mechanism and dynamics revealed by protein lifetime analysis". en. In: *Mol. Syst. Biol.* 14.9, e8355.

Erives, Albert and Michael Levine (Mar. 2004). "Coordinate enhancers share common organizational features in the Drosophila genome". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 101.11, pp. 3851–3856.

Estrada, Javier, Teresa Ruiz-Herrero, Clarissa Scholes, Zeba Wunderlich, and Angela H DePace (Mar. 2016a). "SiteOut: An online tool to design binding site-free DNA sequences". en. In: *PLoS One* 11.3, e0151740.

— (Mar. 2016b). "SiteOut: An Online Tool to Design Binding Site-Free DNA Sequences". In: *PLOS ONE* 11.3. Ed. by Arnar Palsson, e0151740. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0151740. URL: http://dx.doi.org/10.1371/journal.pone.0151740.

Estrada, Javier, Felix Wong, et al. (June 2016). "Information integration and energy expenditure in gene regulation". en. In: *Cell* 166.1, pp. 234–244.

Farley, Emma K et al. (Oct. 2015). "Suboptimization of developmental enhancers". en. In: *Science* 350.6258, pp. 325–328.

Fernandes, Gonçalo et al. (Apr. 2022). "Synthetic reconstruction of the hunchback promoter specifies the role of Bicoid, Zelda and Hunchback in the dynamics of its transcription". en. In: *Elife* 11.

Frasch, M and M Levine (Nov. 1987). "Complementary patterns of even-skipped and fushi tarazu expression involve their differential regulation by a common set of segmentation genes in Drosophila". en. In: *Genes Dev.* 1.9, pp. 981–995.

Fujioka, Miki et al. (Feb. 2016). "Determinants of Chromosome Architecture: Insulator Pairing in cis and in trans". en. In: *PLoS Genet.* 12.2, e1005889.

Fuqua, Timothy et al. (Nov. 2020a). "Dense and pleiotropic regulatory information in a developmental enhancer". en. In: *Nature* 587.7833, pp. 235–239.

— (Nov. 2020b). "Dense and pleiotropic regulatory information in a developmental enhancer". en. In: *Nature* 587.7833, pp. 235–239.

— (Nov. 2020c). "Dense and pleiotropic regulatory information in a developmental enhancer". en. In: *Nature* 587.7833, pp. 235–239.

Furlong, Eileen E M and Michael Levine (Sept. 2018). "Developmental enhancers and chromosome topology". en. In: *Science* 361.6409, pp. 1341–1345.

Galupa, Rafael et al. (Jan. 2023). "Enhancer architecture and chromatin accessibility constrain phenotypic space during Drosophila development". en. In: *Dev. Cell* 58.1, 51–62.e4.

Garcia, Hernan G, Robert C Brewster, and Rob Phillips (Apr. 2016a). "Using synthetic biology to make cells tomorrow's test tubes". en. In: *Integr. Biol. (Camb.)* 8.4, pp. 431–450.

— (Apr. 2016b). "Using synthetic biology to make cells tomorrow's test tubes". en. In: *Integr. Biol. (Camb.)* 8.4, pp. 431–450.

Garcia, Hernan G, Mikhail Tikhonov, et al. (Nov. 2013). "Quantitative imaging of transcription in living Drosophila embryos links polymerase activity to patterning". en. In: *Curr. Biol.* 23.21, pp. 2140–2145.

Genuth, Miriam A. and Scott A. Holley (Sept. 2020). "Mechanics as a Means of Information Propagation in Development". In: *BioEssays* 42.11. ISSN: 1521-1878. DOI: 10.1002/bies.202000121. URL: http://dx.doi.org/10.1002/bies.202000121.

González-Sastre, Alejandro et al. (2017). "The pioneer factor Smed-gata456-1 is required for gut cell differentiation and maintenance in planarians". en. In: *Int. J. Dev. Biol.* 61.1-2, pp. 53–63.

Gordân, Raluca et al. (Apr. 2013). "Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape". en. In: *Cell Rep.* 3.4, pp. 1093–1104.

Goto, T, P Macdonald, and T Maniatis (May 1989). "Early and late periodic patterns of even skipped expression are controlled by distinct regulatory elements that respond to different spatial cues". en. In: *Cell* 57.3, pp. 413–422.

Grant, Charles E, Timothy L Bailey, and William Stafford Noble (Apr. 2011). "FIMO: scanning for occurrences of a given motif". en. In: *Bioinformatics* 27.7, pp. 1017–1018.

Green, Jeremy (2002). "Morphogen gradients, positional information, and Xenopus: Interplay of theory and experiment". In: *Developmental Dynamics* 225.4, pp. 392–408. DOI: https://doi.org/10.1002/dvdy.10170. eprint: https://anatomypubs.onlinelibrary.wiley.com/doi/pdf/10.1002/dvdy.10170. URL: https://anatomypubs.onlinelibrary.wiley.com/doi/abs/10.1002/dvdy.10170.

Green, Jeremy B A and James Sharpe (Apr. 2015). "Positional information and reaction-diffusion: two big ideas in developmental biology combine". en. In: *Development* 142.7, pp. 1203–1211.

Gregor, Thomas et al. (July 2007). "Probing the limits to positional information". en. In: *Cell* 130.1, pp. 153–164.

Gross, C T and W McGinnis (Apr. 1996). "DEAF-1, a novel protein that binds an essential region in a Deformed response element". en. In: *EMBO J.* 15.8, pp. 1961–1970.

Grünberg, Sebastian and Gabriel E Zentner (June 2017). "Genome-wide Mapping of Protein-DNA Interactions with ChEC-seq in Saccharomyces cerevisiae". en. In: *J. Vis. Exp.* 124.

Haario, Heikki et al. (Dec. 2006). "DRAM: Efficient adaptive MCMC". en. In: *Stat. Comput.* 16.4, pp. 339–354.

Hall, Eric T et al. (Dec. 2023). "Cytoneme signaling provides essential contributions to mammalian tissue patterning". en. In: *Cell*.

Hammal, Fayrouz et al. (Jan. 2022). "ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments". en. In: *Nucleic Acids Res.* 50.D1, pp. D316–D325.

Hammonds, Ann S et al. (Dec. 2013). "Spatial expression of transcription factors in Drosophila embryonic organ development". en. In: *Genome Biol.* 14.12, R140.

He, Xin et al. (Sept. 2010). "Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression". en. In: *PLoS Comput. Biol.* 6.9, e1000935.

Hocher, Antoine et al. (Nov. 2023). "Histones with an unconventional DNA-binding mode in vitro are major chromatin constituents in the bacterium Bdellovibrio bacteriovorus". en. In: *Nat. Microbiol.* 8.11, pp. 2006–2019.

Hoermann, Astrid, Damjan Cicin-Sain, and Johannes Jaeger (Mar. 2016). "A quantitative validated model reveals two phases of transcriptional regulation for the gap gene giant in Drosophila". en. In: *Dev. Biol.* 411.2, pp. 325–338.

Ilsley, Garth R et al. (Aug. 2013). "Cellular resolution models for *even skipped* regulation in the entire *Drosophila* embryo". In: *eLife* 2. Ed. by Roderic Guigo, e00522. ISSN: 2050-084X. DOI: 10.7554/eLife.00522. URL: https://doi.org/10.7554/eLife.00522.

Jacob, François and Jacques Monod (June 1961). "Genetic regulatory mechanisms in the synthesis of proteins". en. In: *J. Mol. Biol.* 3.3, pp. 318–356.

Jaeger, Johannes (Jan. 2011). "The gap gene network". en. In: *Cell. Mol. Life Sci.* 68.2, pp. 243–274.

Jaeger, Johannes et al. (Aug. 2004). "Dynamical analysis of regulatory interactions in the gap gene system of Drosophila melanogaster". en. In: *Genetics* 167.4, pp. 1721–1737.

Janssens, Hilde et al. (Oct. 2006). "Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene". en. In: *Nat. Genet.* 38.10, pp. 1159–1165.

Jindal, Granton A and Emma K Farley (Mar. 2021). "Enhancer grammar in development, evolution, and disease: dependencies and interplay". en. In: *Dev. Cell* 56.5, pp. 575–587.

Kim, Ah-Ram et al. (Feb. 2013). "Rearrangements of 2.5 kilobases of noncoding DNA from the Drosophila even-skipped locus define predictive rules of genomic cis-regulatory logic". en. In: *PLoS Genet.* 9.2, e1003243.

Kim, Yang Joon et al. (Dec. 2022). "Predictive modeling reveals that higher-order cooperativity drives transcriptional repression in a synthetic developmental enhancer". en. In: *Elife* 11.

Kribelbauer, Judith F et al. (Oct. 2019). "Low-affinity binding sites and the transcription factor specificity paradox in eukaryotes". en. In: *Annu. Rev. Cell Dev. Biol.* 35.1, pp. 357–379.

Kumar, Divya Krishna et al. (May 2023). "Complementary strategies for directing in vivo transcription factor binding through DNA binding domains and intrinsically disordered regions". en. In: *Mol. Cell* 83.9, 1462–1473.e5.

Kurant, E et al. (Mar. 1998). "Dorsotonals/homothorax, the Drosophila homologue of meis1, interacts with extradenticle in patterning of the embryonic PNS". en. In: *Development* 125.6, pp. 1037–1048.

Kvon, Evgeny Z, Tomas Kazmar, et al. (Aug. 2014a). "Genome-scale functional characterization of Drosophila developmental enhancers in vivo". en. In: *Nature* 512.7512, pp. 91–95.

— (Aug. 2014b). "Genome-scale functional characterization of Drosophila developmental enhancers in vivo". en. In: *Nature* 512.7512, pp. 91–95.

Kvon, Evgeny Z, Yiwen Zhu, et al. (Mar. 2020). "Comprehensive in vivo interrogation reveals phenotypic impact of human enhancer variants". en. In: *Cell* 180.6, 1262–1271.e15.

Lammers, Nicholas C et al. (Jan. 2020). "Multimodal transcriptional control of pattern formation in embryonic development". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 117.2, pp. 836–847.

Le Poul, Yann et al. (Dec. 2020a). "Regulatory encoding of quantitative variation in spatial activity of a Drosophila enhancer". en. In: *Sci. Adv.* 6.49, eabe2955.

— (Dec. 2020b). "Regulatory encoding of quantitative variation in spatial activity of a Drosophila enhancer". en. In: *Sci. Adv.* 6.49, eabe2955.

Leptin, M (Sept. 1991). "twist and snail as positive and negative regulators during Drosophila mesoderm development". en. In: *Genes Dev.* 5.9, pp. 1568–1576.

Levin, Michael, Alexis M Pietak, and Johanna Bischof (Mar. 2019). "Planarian regeneration as a model of anatomical homeostasis: Recent progress in biophysical and computational approaches". en. In: *Semin. Cell Dev. Biol.* 87, pp. 125–144.

Levine, Mike (Sept. 2010). "Transcriptional enhancers in animal development and evolution". en. In: *Curr. Biol.* 20.17, R754–63.

Levo, Michal et al. (July 2015). "Unraveling determinants of transcription factor binding outside the core binding site". en. In: *Genome Res.* 25.7, pp. 1018–1029.

Levsky, Jeffrey M and Robert H Singer (Jan. 2003). "Gene expression and the myth of the average cell". en. In: *Trends Cell Biol.* 13.1, pp. 4–6.

Li, Lily and Zeba Wunderlich (May 2017). "An enhancer's length and composition are shaped by its regulatory task". In: *Front. Genet.* 8.

Li, Long et al. (2007). "Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses". en. In: *Genome Biol.* 8.6, R101.

Li, Xiao-Yong and Michael B Eisen (Aug. 2018). "Effects of the maternal factor Zelda on zygotic enhancer activity in the*Drosophila*embryo".

Li, Xiao-Yong, Stewart MacArthur, et al. (Feb. 2008a). "Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm". en. In: *PLoS Biol.* 6.2, e27.

— (Feb. 2008b). "Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm". en. In: *PLoS Biol.* 6.2, e27.

Lifanov, Alexander P et al. (Apr. 2003). "Homotypic regulatory clusters in Drosophila". en. In: *Genome Res.* 13.4, pp. 579–588.

Liu, Feng and James W Posakony (July 2012). "Role of architecture in the function and specificity of two Notch-regulated transcriptional enhancer modules". en. In: *PLoS Genet.* 8.7, e1002796.

Lopes, Francisco J P et al. (Sept. 2008). "Spatial bistability generates hunchback expression sharpness in the Drosophila embryo". en. In: *PLoS Comput. Biol.* 4.9, e1000184.

López-Rivera, Francheska et al. (Dec. 2020a). "A Mutation in the Drosophila melanogaster eve Stripe 2 Minimal Enhancer Is Buffered by Flanking Sequences". en. In: *G3 (Bethesda)* 10.12, pp. 4473–4482.

— (Dec. 2020b). "A Mutation in the Drosophila melanogaster eve Stripe 2 Minimal Enhancer Is Buffered by Flanking Sequences". en. In: *G3 (Bethesda)* 10.12, pp. 4473–4482.

Lusk, Richard W and Michael B Eisen (Jan. 2010a). "Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers". en. In: *PLoS Genet.* 6.1, e1000829.

— (Jan. 2010b). "Evolutionary Mirages: Selection on Binding Site Composition Creates the Illusion of Conserved Grammars in Drosophila Enhancers". In: *PLoS Genetics* 6.1. Ed. by Gregory S. Barsh, e1000829. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1000829. URL: http://dx.doi.org/10.1371/journal.pgen.1000829.

Marcus, J. (1998). *Women's Ritual in Formative Oaxaca: Figure-making, Divination, Death and the Ancestors*. Memoirs. University of Michigan Press. ISBN: 9780915703487. URL: https://books.google.de/books?id=S3LgDwAAQBAJ.

Maroudas-Sacks, Yonit and Kinneret Keren (Oct. 2021). "Mechanical patterning in animal morphogenesis". en. In: *Annu. Rev. Cell Dev. Biol.* 37.1, pp. 469–493.

Martinez-Corral, Rosa et al. (May 2024). "The Hill function is the universal Hopfield barrier for sharpness of input-output responses". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 121.22, e2318329121.

McGinnis, W (Nov. 2005). "From DNA to diversity, molecular genetics and the evolution of animal design, 2nd edition". In: *J. Hered.* 96.6, pp. 725–725.

Muller, P et al. (Apr. 2013). "Morphogen transport". In: *Development* 140.8, pp. 1621–1638.

Ninova, Maria, Matthew Ronshaugen, and Sam Griffiths-Jones (Aug. 2014). "Conserved temporal patterns of microRNA expression in Drosophila support a developmental hourglass model". en. In: *Genome Biol. Evol.* 6.9, pp. 2459–2467.

Nüsslein-Volhard, C and E Wieschaus (Oct. 1980). "Mutations affecting segment number and polarity in Drosophila". en. In: *Nature* 287.5785, pp. 795–801.

Oliveto, Stefania et al. (Feb. 2017). "Role of microRNAs in translation regulation and cancer". en. In: *World J. Biol. Chem.* 8.1, pp. 45–56.

Pai, C Y et al. (Feb. 1998). "The Homothorax homeoprotein activates the nuclear localization of another homeoprotein, extradenticle, and suppresses eye development in Drosophila". en. In: *Genes Dev.* 12.3, pp. 435–446.

Panne, Daniel, Tom Maniatis, and Stephen C Harrison (June 2007). "An atomic model of the interferon-beta enhanceosome". en. In: *Cell* 129.6, pp. 1111–1123.

Papatsenko, Dmitri, Yury Goltsev, and Michael Levine (Sept. 2009). "Organization of developmental enhancers in the Drosophila embryo". en. In: *Nucleic Acids Res.* 37.17, pp. 5665–5677.

Park, Jeehae et al. (June 2019a). "Dissecting the sharp response of a canonical developmental enhancer reveals multiple sources of cooperativity". en. In: *Elife* 8.

— (June 2019b). "Dissecting the sharp response of a canonical developmental enhancer reveals multiple sources of cooperativity". en. In: *Elife* 8.

Perkins, Melinda Liu (June 2021). "Implications of diffusion and time-varying morphogen gradients for the dynamic positioning and precision of bistable gene expression boundaries". en. In: *PLoS Comput. Biol.* 17.6, e1008589.

Phillips, Rob (Sept. 2020). *The molecular switch.* Princeton University Press.

Phillips, Rob et al. (Nov. 2012). *Physical biology of the cell.* 2nd ed. London, England: Garland Science.

Pichaud, F and F Casares (Aug. 2000). "homothorax and iroquois-C genes are required for the establishment of territories within the developing eye disc". en. In: *Mech. Dev.* 96.1, pp. 15–25.

Rastogi, Chaitanya et al. (Apr. 2018). "Accurate and sensitive quantification of protein-DNA binding affinity". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 115.16, E3692–E3701.

Rauskolb, C, M Peifer, and E Wieschaus (Sept. 1993). "Extradenticle, a regulator of homeotic gene activity, is a homolog of the homeobox-containing human proto-oncogene pbx1". en. In: *Cell* 74.6, pp. 1101–1112.

Razo-Mejia, Manuel et al. (Apr. 2018). "Tuning transcriptional regulation through signaling: A predictive theory of allosteric induction". en. In: *Cell Syst.* 6.4, 456–469.e10.

Rieckhof, G E et al. (Oct. 1997). "Nuclear translocation of extradenticle requires homothorax, which encodes an extradenticle-related homeodomain protein". en. In: *Cell* 91.2, pp. 171–183.

Robert Stojnic, Diego Diez (2017). *PWMEnrich*.

Rohs, Remo et al. (Oct. 2009). "The role of DNA shape in protein-DNA recognition". en. In: *Nature* 461.7268, pp. 1248–1253.

Ronchi, E et al. (July 1993). "Down-regulation of the Drosophila morphogen bicoid by the torso receptor-mediated signal transduction cascade". en. In: *Cell* 74.2, pp. 347–355.

Rushlow, C et al. (1987). "Maternal regulation of zerknüllt: a homoeobox gene controlling differentiation of dorsal tissues in Drosophila". en. In: *Nature* 330.6148, pp. 583–586.

Samee, Md Abul Hassan et al. (Dec. 2015). "A systematic ensemble approach to thermodynamic modeling of gene expression from sequence data". en. In: *Cell Syst.* 1.6, pp. 396–407.

Sauer, F and H Jäckle (Oct. 1991). "Concentration-dependent transcriptional activation or repression by Krüppel from a single binding site". en. In: *Nature* 353.6344, pp. 563–566.

Schneider, T D et al. (Apr. 1986). "Information content of binding sites on nucleotide sequences". en. In: *J. Mol. Biol.* 188.3, pp. 415–431.

Schoenfelder, Stefan and Peter Fraser (Aug. 2019). "Long-range enhancer-promoter contacts in gene expression control". en. In: *Nat. Rev. Genet.* 20.8, pp. 437–455.

Scott F, Gilbert (Mar. 2006). "Developmental biology. Sixth edition." en. In: *Am. J. Med. Genet.* 99.2, pp. 170–171.

Seo, H C et al. (May 1999). "Six class homeobox genes in drosophila belong to three distinct families and are involved in head development". en. In: *Mech. Dev.* 83.1-2, pp. 127–139.

Small, S et al. (May 1991a). "Transcriptional regulation of a pair-rule stripe in Drosophila". en. In: *Genes Dev.* 5.5, pp. 827–839.

— (May 1991b). "Transcriptional regulation of a pair-rule stripe in Drosophila". en. In: *Genes Dev.* 5.5, pp. 827–839.

Small, S., A. Blair, and M. Levine (1992). "Regulation of even-skipped stripe 2 in the Drosophila embryo." In: *The EMBO Journal* 11.11, pp. 4047–4057. DOI: https://doi.org/10.1002/j.1460-2075.1992.tb05498.x. eprint: https://www.embopress.org/doi/pdf/10.1002/j.1460-2075.1992.tb05498.x. URL: https://www.embopress.org/doi/abs/10.1002/j.1460-2075.1992.tb05498.x.

Small, Stephen and David N Arnosti (Sept. 2020). "Transcriptional enhancers in Drosophila". en. In: *Genetics* 216.1, pp. 1–26.

Sönmezer, Can et al. (Jan. 2021). "Molecular co-occupancy identifies transcription factor binding cooperativity in vivo". en. In: *Mol. Cell* 81.2, 255–267.e6.

Staller, Max V et al. (Jan. 2013). "Depleting gene activities in early Drosophila embryos with the "maternal-Gal4-shRNA" system". en. In: *Genetics* 193.1, pp. 51–61.

Stanojevic, D, S Small, and M Levine (Nov. 1991). "Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo". en. In: *Science* 254.5036, pp. 1385–1387.

Stormo, Gary D (June 2013). "Modeling the specificity of protein-DNA interactions". en. In: *Quant. Biol.* 1.2, pp. 115–130.

Struffi, Paolo et al. (Oct. 2011). "Combinatorial activation and concentration-dependent repression of the Drosophila even skipped stripe 3+7 enhancer". In: *Development* 138.19, pp. 4291–4299. ISSN: 0950-1991. DOI: 10.1242/dev.065987. eprint: https://journals.biologists.com/dev/article-pdf/138/19/4291/1158327/4291.pdf. URL: https://doi.org/10.1242/dev.065987.

Struhl, G, K Struhl, and P M Macdonald (June 1989). "The gradient morphogen bicoid is a concentration-dependent transcriptional activator". en. In: *Cell* 57.7, pp. 1259–1273.

Swanson, Christina I, Nicole C Evans, and Scott Barolo (Mar. 2010). "Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer". en. In: *Dev. Cell* 18.3, pp. 359–370.

Taskiran, Ibrahim I et al. (Dec. 2023). "Cell type directed design of synthetic enhancers". en. In: *Nature*.

— (Feb. 2024). "Cell-type-directed design of synthetic enhancers". en. In: *Nature* 626.7997, pp. 212–220.

Tate, Carolyn; Bendersky Gordon (1999). "Olmec Sculptures of the Human Fetus". In: *Perspectives in Biology and Medicine*. DOI: 10.1353/pbm.1999.0017.

Tkacik, Gasper, Curtis G Callan Jr, and William Bialek (Aug. 2008). "Information flow and optimization in transcriptional regulation". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 105.34, pp. 12265–12270.

Tomancak, Pavel, Amy Beaton, et al. (2002). In: *Genome Biol* 3.12, research0088.1.

Tomancak, Pavel, Benjamin P Berman, et al. (2007). "Global analysis of patterns of gene expression during Drosophila embryogenesis". en. In: *Genome Biol.* 8.7, R145.

Tran, Huy et al. (Oct. 2018). "Precision in a rush: Trade-offs between reproducibility and steepness of the hunchback expression pattern". en. In: *PLoS Comput. Biol.* 14.10, e1006513.

Trojanowski, Jorge et al. (May 2022). "Transcription activation is enhanced by multivalent interactions independent of phase separation". en. In: *Mol. Cell* 82.10, 1878–1893.e10.

Tsai, Albert, Mariana Rp Alves, and Justin Crocker (July 2019). "Multi-enhancer transcriptional hubs confer phenotypic robustness". en. In: *Elife* 8.

Tsai, Albert, Robert H Singer, and Justin Crocker (Apr. 2018). "Transvection goes live—visualizing enhancer-promoter communication between chromosomes". en. In: *Mol. Cell* 70.2, pp. 195–196.

Tsiairis, Charisios D and Alexander Aulehla (Feb. 2016). "Self-organization of embryonic genetic oscillators into spatiotemporal wave patterns". en. In: *Cell* 164.4, pp. 656–667.

Turing, A M (Aug. 1952). "The chemical basis of morphogenesis". en. In: *Philos. Trans. R. Soc. Lond.* 237.641, pp. 37–72.

Verd, Berta et al. (Feb. 2018). "A damped oscillator imposes temporal order on posterior gap gene expression in Drosophila". en. In: *PLoS Biol.* 16.2, e2003174.

Vincent, Ben J, Javier Estrada, and Angela H DePace (Apr. 2016a). "The appeasement of Doug: a synthetic approach to enhancer biology". en. In: *Integr. Biol. (Camb.)* 8.4, pp. 475–484.

— (Apr. 2016b). "The appeasement of Doug: a synthetic approach to enhancer biology". en. In: *Integr. Biol. (Camb.)* 8.4, pp. 475–484.

Visel, Axel, James Bristow, and Len A Pennacchio (Feb. 2007). "Enhancer identification through comparative genomics". en. In: *Semin. Cell Dev. Biol.* 18.1, pp. 140–152.

Wallingford, John B. (Feb. 2021). "Aristotle, Buddhist scripture and embryology in ancient Mexico: building inclusion by re-thinking what counts as the history of developmental biology". In: *Development* 148.3, dev192062. ISSN: 0950-1991. DOI: 10.1242/dev.192062. eprint: https://journals.biologists.com/dev/article-pdf/148/3/dev192062/1815452/dev192062.pdf. URL: https://doi.org/10.1242/dev.192062.

White, Michael A et al. (July 2013). "Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 110.29, pp. 11952–11957.

Witkowski, Jan (1985). "The hunting of the organizer: an episode in biochemical embryology". In: *Trends in Biochemical Sciences* 10.10, pp. 379–381. ISSN: 0968-0004. DOI: https://doi.org/10.1016/0968-0004(85)90058-1. URL: https://www.sciencedirect.com/science/article/pii/0968000485900581.

Wittkopp, Patricia J and Gizem Kalay (Dec. 2011). "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence". en. In: *Nat. Rev. Genet.* 13.1, pp. 59–69.

Wunderlich, Zeba, Meghan D Bragdon, and Angela H DePace (June 2014). "Comparing mRNA levels using in situ hybridization of a target gene and co-stain". en. In: *Methods* 68.1, pp. 233–241.

Wunderlich, Zeba and Leonid A Mirny (Oct. 2009). "Different gene regulation strategies revealed by analysis of binding motifs". en. In: *Trends Genet.* 25.10, pp. 434–440.

Yáñez-Cuna, J Omar et al. (July 2014). "Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features". en. In: *Genome Res.* 24.7, pp. 1147–1156.

Zhu, Lihua Julie et al. (Jan. 2011a). "FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system". en. In: *Nucleic Acids Res.* 39.Database issue, pp. D111–7.

— (Jan. 2011b). "FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system". en. In: *Nucleic Acids Res.* 39.Database issue, pp. D111–7.

I hereby declare that I, Gilberto Alvarez Canales, prepared this Ph.D. thesis: 'Developmental constraints and the regulatory logic of Drosophila enhancers' on my own and with no other sources and aids than quoted.