Regulation of Neurogenic Ectoderm Specification in *Drosophila melanogaster*

Thesis by

Louisa M. Liberman

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in Biology

California Institute of Technology

Pasadena, California

Defended May 13, 2009

To Ellen and Jim Liberman, for their love and their genes

and

my grandparents, Muriel and Saul Liberman

Acknowledgements

I am grateful for many people at Caltech and beyond, who have been great friends, collaborators, and mentors providing support and insight throughout this adventure.

First and foremost, I would like to thank my advisor, Angela Stathopoulos, who has guided my scientific interests. She provided much of the inspiration for this work, while allowing me the opportunity for scientific exploration and the freedom to develop my ideas. I have appreciated her sincere advice, and her wealth of knowledge along the way. This has been an incredible learning experience and I thank her for supporting me both academically and personally.

I am thankful to Eric Davidson, Scott Fraser, Paul Sternberg, and Kai Zinn for participating in my research as lively and helpful members of my thesis committee. They have suggested experiments that improved my research, and have been exceptionally encouraging. I am truly thankful for their understanding and expertise.

The Stathopoulos Lab has been a great place to work. I would particularly like to thank Greg Reeves for collaborating with me on the Dorsal quantification project, and for helping to see it to completion. I thank Sarah Payne for many helpful discussions, tea and a reliable futon.

I thank Dylan Morris, Alex Farley, Tristan Ursell, Sidney Cox and Puneet Sethi for their unnecessary kindness, inviting me to live in their home while writing this thesis.

Finally, I thank my husband, Ryan Baugh who is a wonderful companion: providing love, reassurance and playful distraction. I cannot imagine this experience without him.

Abstract

Creating a functional organism requires reproducible developmental patterning. A nuclear gradient of the NF-κB transcription factor, Dorsal, provides positional information necessary to specify the mesoderm, neurogenic ectoderm, dorsal ectoderm and amnioserosa along the dorsal-ventral axis in *Drosophila melanogaster* embryos. In this work we investigate the role that Dorsal and other transcription factors play in these crucial patterning events. We focus primarily on the gene regulation that controls patterning of the presumptive neurogenic ectoderm that is specified in lateral regions of the embryo. We investigate this early patterning event in two ways: first, by studying a known regulatory element for this region, and second, by examining the levels of Dorsal in the nuclei. We find that Dorsal can function with Zelda, a maternally deposited ubiquitous activator, to specify the neurogenic ectoderm. We then ask if the levels of Dorsal in wild type embryos are predictive of the gene expression outputs, as suggested by existing models. We measure the amount of Dorsal protein able to activate target gene expression in mutants, where the levels of Dorsal protein have been genetically manipulated. Our measurements indicate that Dorsal does not regulate gene expression in a concentration-dependent fashion. Instead, our data support the idea that Dorsal functions with other proteins to establish gene expression boundaries. These studies jointly suggest that regulation of differential gene expression requires combinatorial interactions between spatially localized and uniformly distributed transcription factors.

Contents

Chapter 1

Introduction

"Time flies like an arrow, fruit flies like a banana."

- Groucho Marx

The fruit fly, *Drosophila melanogaster,* has been used as a genetic system for close to a century. During this time, *Drosophila* has served as a model organism for studying many fundamental questions in biology. In the early 1900s Thomas H. Morgan started to work on *Drosophila*, by feeding the flies bananas, and assigning a student the task of attempting to induce mutations. They started keeping the flies in the dark, and then switched to radiation. After a couple of years without results, a mutant fly appeared in lab with white eyes. It is uncertain whether this fly was the result of spontaneously mutation or the radium rays. Regardless of the cause, this was the beginning of an exceptionally exciting time for Morgan and the members of his Fly Room. They had, in this fly and many other mutants that followed, powerful tools for studying inheritance and a novel approach for investigating development. The laboratory described and established many genetic practices that are now commonly used in fly labs around the world. Among other advances, Morgan and his students demonstrated that genes could be mapped to linear arrays, or complementation groups, that would prove to correspond cytologically to chromosomes. They also discovered sex-linked inheritance and explained abnormalities in the sex ratio of progeny, they showed that recombination occurs in females, but not males, and that nondisjunction increases with maternal age (for a more detailed account see Shine and Wrobel, 1976). These essential findings paved the way for *Drosophila* as a genetic system. Extensive research in the fruit fly since these studies has led to greater understanding of nearly every aspect of biology. Genetic analysis of molecular and cellular biology, development, physiology, neurobiology, evolution and behavior in flies has given us a greater appreciation for the fly and facilitated other research in slower growing and more complex animals.

One of the most beautiful aspects of *Drosophila* development is the remarkable expression patterns that form during embryogenesis. Stripes of gene expression along both the anterior-posterior and dorsal-ventral axes are easily visualized by *in situ* hybridization of RNA probes to the transcripts in the embryo. Expression domains predict the specification and eventual location

of germ layers that differentiate in these embryonic regions. Using multiplex *in situ* hybridization techniques, several transcripts can be observed simultaneously (Fig. 1).

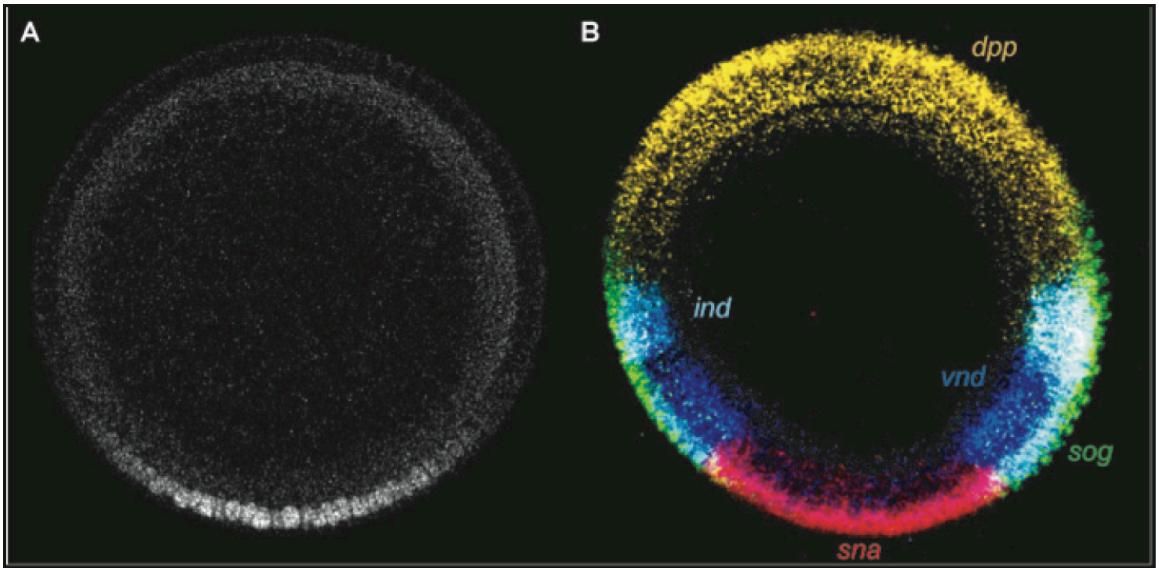


Fig. 1. Cross section of nuclear cycle 14 embryo depicts location of Dorsal protein in relation to target gene expression. A) Dorsal protein gradient. B) Expression domains for genes that require Dorsal for proper expression: *snail* (*sna*) expression is present in the presumptive mesoderm, *ventral neurogenic defective* (*vnd*) forms a domain in the ventral regions of the neurogenic ectoderm. *intermediate neuroblasts defective* (*ind*) is expressed in lateral regions of the presumptive neurogenic ectoderm, *short gastrulation* (*sog*) is expressed in a broad lateral domain of the presumptive neurogenic ectoderm. In dorsal regions of the embryo zerknullt (*zen)* and *decapentaplegic (dpp)* are expressed. (Figure by Greg Reeves).

How do those patterns form with such reproducible precision in each animal? This question motivates this thesis. Our investigation focuses on previously characterized transcription factors and the non-coding DNA sequences they bind in order to regulate target gene expression. Identification of such regulatory elements has proven nontrivial, and while numerous transcription factors and target genes have been implicated in patterning, relatively few of the regulatory elements mediating control have been characterized. We sought to identify *cis*-regulatory elements that orchestrate the specification of the neurogenic ectoderm and to dissect their function with genetic analysis inferring their mechanism of action.

A review of axis specification

Specification of both the anterior-posterior and dorsal-ventral axes has been well studied in the *Drosophila* embryo. The initial polarity of the egg is established in the ovary. Somatic follicle cells surround a germ cell, which divides into 16 interconnected cells. The 15 anterior cells become nurse cells to supply proteins and mRNA to one of the posterior cells that will become the egg (Nüsslein-Volhard et al., 1987). Similar to many insects, the *Drosophila* embryo is a syncytium, which undergoes 14 nuclear divisions prior to cellularization. As a single cell, maternal deposits are supplied to the embryo; subsequently they diffuse or are actively regulated, establishing spatial gradients of transcription factors. The anterior-posterior axis is established by three groups of maternal factors. The anterior group, including the transcription factor Bicoid and its target genes, is responsible for pattering the head and thoracic regions of the embryo. *bicoid* mRNA is deposited in the anterior of the oocyte and is translated after fertilization. Nanos, an RNA binding protein, is present in the posterior regions of the embryo, and is required for the regulation of genes that give rise to posterior body structures: the pole plasm and abdominal segments (Lehmann and Nüsslein-Volhard, 1991). These two groups work antagonistically to pattern the anterior-posterior axis. Additionally, terminal signaling activates the gene *torso* whose gene product establishes structures at both the anterior and posterior parts of the embryo including the acron and the telson (Klingler et al., 1988; Nüsslein-Volhard, 1991).

Dorsal-ventral axis specification

Twelve maternal genes, called the dorsal group establish dorsal-ventral axis asymmetry in the *Drosophila* embryo (Morisato and Anderson, 1995). Dorsal is the only identified maternal factor that provides positional information along this axis. Christiane Nüsslein-Volhard first described the Dorsal mutant phenotype in a detailed account of maternal effect mutations that

alter the spatial coordinates of the embryo (Nüsslein-Volhard, 1979). Similar to the anterior-posterior determinants, *dorsal* mRNA is maternally deposited in the oocyte. Unlike *bicoid*, *dorsal* is not tethered or constrained to a particular region of the early embryo. *dorsal* mRNA is distributed ubiquitously throughout the embryo. Once it is translated, it is differentially localized to the nucleus to form a nuclear gradient of Dorsal protein with highest amounts in ventral regions and lowest amounts in dorsal regions.

The dorsal-ventral asymmetries, which establish the Dorsal gradient, originate in the oocyte. During stage 8 of oogenesis, the oocyte nucleus moves to the dorso-anterior cortex of the oocyte. The *gurken* mRNA, associated with the nucleus, is translated and Gurken protein that activates EGFR signaling, repressing *pipe* in dorsal regions of the somatic follicle cells surrounding the oocyte. *pipe*-expressing follicle cells secrete some unknown factor which initiates a protease signaling cascade responsible for regulating the Toll ligand, Spätzle.

After fertilization, the asymmetries established in the oocyte are maintained in the embryo. In ventral regions of the embryo, the protease cascade results in Toll activation, degrading the I-κB homolog, Cactus. This degradation breaks apart the Cactus-Dorsal complex allowing Dorsal to enter the nucleus (Moussian and Roth, 2005). Dorsal nuclear localization occurs five times during each nuclear cycle from cycle 10 through 14 (DeLotto et al., 2007). This dynamic process is even more active as Dorsal protein shuttles in and out of the nucleus throughout the nuclear cycles (DeLotto et al., 2007). In *cactus* mutant embryos, Dorsal localizes to the all the nuclei, but there is still more Dorsal in ventral nuclei than dorsal nuclei, implicating an additional input for Dorsal nuclear localization that is *cactus*-independent (Bergmann et al., 1996). These proteins are all essential for creating the dorsal-ventral asymmetry seen in wild type embryos. Many of these pathway components are also conserved in mammals (reviewed in Belvin and Anderson, 1996)

Dorsal and NF-κB

The Dorsal protein is part of the Rel/NF-κB family of proteins which all share a 300 amino-acid Rel homology domain, DNA binding and dimerization domains as well as a nuclear translocation signal. This protein family is involved in multiple cellular processes including axis specification, as discussed (above), regulation of the innate immune response, muscle development and hematopoiesis. Rel/NF-κB proteins are sequestered in a cytoplasmic complex by inhibitory proteins that contain ankyrin repeats, required for protein-protein interactions (Inoue et al., 1992). Cactus binds to Dorsal in Drosophila embryos and I-κB inhibits NF-κB in mammalian cells likely by binding to the nuclear localization signal (Verma et al., 1995). NF-κB does not appear to play a role in mammalian embryonic development (Baeuerle and Baltimore, 1996). However, the Drosophila Rel protein, Dif (Dorsal related immunity factor) and the mammalian Rel protein, NF-κB, both regulate innate immune responses. In *Drosophila* it has been proposed that Dif and Dorsal may function as heterodimers to mount an immune response in the fat body when infected. These similarities in pathway components and function further support studying Dorsal in *Drosophila* so that we may provide further insights into human biology.

Morphogen hypothesis

Dorsal is required to pattern the dorsal-ventral axis in a concentration-dependent fashion, which is why it is widely accepted as a morphogen, although it may be more accurate to refer to it simply as a graded transcription factor. Probably the most recognized concept for explaining how morphogen gradients confer positional information to cells, the French Flag Model, was popularized by Lewis Wolpert (Wolpert, 1968). This model is attractive in its simplicity: a gradient of a morphogen diffuses from a source across a row of cells, each with the potential to develop into blue, white or red (see Fig. 2). The cells will form a French Flag pattern if the

morphogen provides positional information and varying concentration thresholds result in different gene expression states (e.g.: blue, white, and red).

The French Flag Model is useful as a framework to suggest additional complexities for explaining gene expression patterns. The morphogen gradient model requires that the patterning gradient provide a reproducible signal. Reproducibility is essential because cells in the morphogen field must read the concentration to determine their location in the gradient and then respond appropriately to this input (Ashe and Briscoe, 2006). Several models have been proposed which attempt to explain how analog inputs, such as morphogens, could produce discrete outputs.

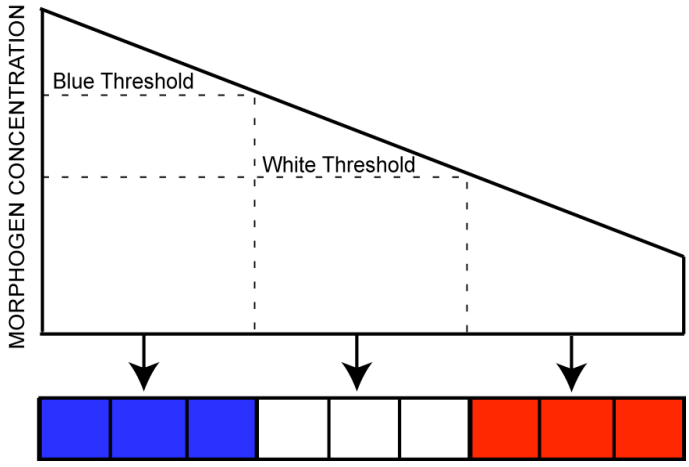


Fig. 2. The French Flag Model. Cells experience the concentration of a morphogen. At each morphogen concentration threshold, different genes would be expressed to create different gene expression states. In this case, highest amounts of morphogen produce a blue state, moderate amounts produce a white state and lower concentrations levels produce a red state. (Adapted from Wolpert, 1968).

Establishing germ layers

Dorsal is responsible for specifying at least four germ layers along the Dorsal ventral axis of the developing *Drosophila* embryo (Stathopoulos and Levine, 2002a). Dorsal activates the expression of genes in the presumptive mesoderm and the neurogenic ectoderm. In dorsal regions of the embryo, Dorsal has been shown to function as a repressor of genes that give rise to the dorsal ectoderm and amnioserosa. The presumptive mesoderm forms in ventral regions of the embryo where Dorsal is in high abundance. In this region, Dorsal activates the expression of Twist, which then functions in a cooperative fashion with Dorsal to regulate the expression of presumptive mesoderm genes, including *snail (sna)* (Ip et al., 1992b; Jiang and Levine, 1993).

Dorsal may also function cooperatively with other bHLH factors since it can also bind to the protein T4 bHLH (González-Crespo and Levine, 1993). It has been shown that the ventral neurogenic ectoderm, in ventral-lateral regions of the early embryo, is also patterned by combinatorial interactions between Dorsal and Twist (Ip et al., 1992b; Jiang and Levine, 1993).

Although Dorsal is required for expression of genes in lateral domains of the embryo, less is known about how the presumptive neurogenic ectoderm is patterned. This region can be subdivided into horizontal columns by the gene expression domains of three homeobox genes: *ventral nervous system defective (vnd), intermediate neuroblast defective (ind)* and *muscle segment homeobox (msh).* It has been demonstrated by a number of groups that Vnd acts to repress *ind* and that mis-expression of Vnd represses both *ind* and *msh* expression (Cowden and Levine, 2003; McDonald et al., 1998; Von Ohlen and Doe, 2000; Weiss et al., 1998). These genes are conserved in mammals and may represent conserved mechanism for central nervous system patterning.

Yet another group of genes expressed in the presumptive neurogenic ectoderm, encompassing a broad domain in lateral regions of the embryo, also requires Dorsal for proper patterning. Interestingly, the Dorsal gradient drops off quite steeply throughout the expression domain of these neurogenic ectoderm genes. In *dorsal* null mutants, genes that are normally expressed in the presumptive neurogenic ectoderm are no longer expressed. Although we know that Dorsal is required for the expression of genes such as the FGF ligand, *thisbe* (*ths*), the BMP antagonist, *short gastrulation* (*sog*), and the ADAM metalloprotease, *Neurogenic ectoderm 3* (*Neu3*), the mechanism used to activate these genes in a broad lateral expression domain is not fully understood.

The requirement for Dorsal to specify cell fate along the dorsal-ventral axis is well accepted, but a quantitative relationship between Dorsal levels and target gene expression has not been established. Other researchers have attempted to quantify levels of nuclear Dorsal and

subsequently correlate these levels with gene expression outputs (Zinzen et al., 2006), but their experiments used profiles of whole embryos and did not account for functional Dorsal levels which are present within the nuclei. Quantifying the functional amount of Dorsal in the nuclei is essential for determining if Dorsal acts as a classical morphogen. Experiments with the Bicoid gradient indicate that the morphogen hypothesis for pattern formation could not explain residual pattern formation in the absence of a wild type Bicoid gradient (Ochoa-Espinosa et al., 2009). These results suggest a more complex mechanism for pattern formation, where transcriptional inputs in the form of graded transcription factors coupled with ubiquitous activators form a regulatory network capable of controlling the expression of genes in precise locations (Davidson, 2006). We should not assume that the concentration of Dorsal alone in a particular nucleus is sufficient to determine target gene expression.

Gradient interpretation models

In an effort to incorporate combinatorial regulatory interactions into developmental patterning models, we will consider three proposals for reliable interpretation of a morphogen or graded input. The first is the differential sensitivity or affinity model; the second model considers combinatorial interactions between transcription factors; the third is a site-occupancy model. All of these models are useful for understanding the progression of thinking surrounding the interpretation of morphogens or graded inputs. Each of them makes testable predictions that should help to ascertain which models provide the best insight for future exploration.

The differential sensitivity model, developed for the Bicoid morphogen, suggests that the location of threshold boundaries within the gradient is regulated by the strength Bicoid binds to its respective *cis*-regulatory modules (Driever et al., 1989). Bicoid binds with high affinity to regulatory regions of genes, such as Hunchback, when Bicoid concentration is low and with low

affinity in regions where Bicoid concentration is high. Testable predictions result from this model. The first is that target gene expression should correlate with a wild type threshold level when the morphogen levels are altered, a second is that *cis*-regulatory elements that direct these targets should contain binding sites with variable morphogen affinity depending on their placement in the morphogen gradient.

A similar affinity model was proposed for Dorsal that differs from the Bicoid model by including a cooperativity component in addition to binding affinity. This model is discussed in a paper that demonstrates cooperative interactions between Dorsal and the bHLH transcription factor Twist (Jiang and Levine, 1993). This model predicts that if two transcription factors function cooperatively, and one transcription factor is removed, a non-linear decrease in expression should be observed.

A more recent hypothesis for how the Dorsal gradient provides spatial regulatory information to its target genes is the binding site occupancy model. In this model, the rate-limiting step for transcription is the binding of Dorsal, Twist and Snail, on regulatory regions of DNA in a cooperative fashion to produce sharp boundaries of expression (Zinzen et al., 2006). This hypothesis does not account for the target expression boundaries seen in the lateral regions of the embryo because Twist and Snail are not present in the nuclei in these regions. This model could explain the expression boundaries in the presumptive mesoderm and ventral neurogenic ectoderm. This model predicts that varying the concentrations of each of the transcription factors would result in reduced cooperativity and reduced expression of a given target gene.

In the case of the anterior-posterior determinant, Bicoid, it was recently found that threshold levels do not predict the gene expression outputs observed, and flattening the Bicoid protein levels in the embryo still produces embryos with some anterior-posterior positional information (Ochoa-Espinosa et al., 2009). Previous investigations found that Bicoid affinity for sites in regulatory regions along the axis did not correlate with the amount of Bicoid present in

that part of the axis (Ochoa-Espinosa et al., 2005). These results, and their implications for morphogen-dependent patterning of gene expression, suggest that the mechanism by which Dorsal regulates gene expression should be investigated more thoroughly.

Recent work, including this thesis, attempts to address the basic problem of patterned gene expression and cell specification. DeLotto and colleagues investigated the dynamics of Dorsal protein localization in transgenic flies which contained a Dorsal:GFP fusion. We initially were interested in using these flies to study Dorsal localization in relation to gene expression, but the construct is missing an export sequences near the N-terminus of Dorsal (Xylourgidis et al., 2006), adjacent to the GFP protein fusion and does not complement the Dorsal mutants (see Appendix B). Despite these deficiencies, the work revealed some useful insights into the dynamics of Dorsal localization. They found that Dorsal shuttles in and out of all the nuclei during each nuclear cycle, including the nuclei in Dorsal regions of the embryo. We investigated whether the presence of Dorsal in the lateral and Dorsal nuclei could account for the expression of genes in the presumptive neurogenic ectoderm.

The promise of genomics

In 2000, the first sequence of the *Drosophila melanogaster* genome was completed and with its completion held the promise of understanding the information encoded in the fly's ~120 mega base pair genome (Adams et al., 2000). Since then, the genomes of eleven other *Drosophila* species have been sequenced providing a wealth of information, that is ripe for interrogation and studying *cis*-regulation (Clark et al., 2007). Using established techniques, regulatory elements can be fused to a reporter gene such as LacZ, GFP or other fluorescent protein, and the regulatory nature of these pieces of DNA can be observed. The use of P-elements has made it possible to

generate transgenic flies to investigate the regulatory elements in the context of the animal's development (Rubin and Spradling, 1982). This method has been fruitful.

Advances to the P-element method to generate transgenic flies have been made by site-directed integration of regulatory element-reporter fusions. This relatively new technique greatly reduces the time needed to screen lines. It also reduces the risk of generating a reporter line that is under the influence of unwanted positional effects, or "trapped" (Levis et al., 1985). Fly lines with landing sites were generated by two research groups. These lines facilitate the direct comparison of *cis*-regulatory elements in the same regulatory environment with no positional effects (Bischof et al., 2007; Groth et al., 2004). The number of sequenced genomes along with the ability to quickly generate transgenic flies makes *Drosophila* an excellent system for learning about developmental patterning and *cis*-regulation.

Using conservation among species to find new regulatory elements should provide a rapid means to identify putative regulatory elements. Several Dorsal transcriptional target genes have been studied in other *Drosophilid* species and are indeed conserved in their spatial expression. This degree of conservation provides reassurance that their regulatory regions may also be conserved. Yet the search for conserved regulatory elements has yielded mixed results. Blocks of DNA have been found which are capable of driving similar reporter gene expression that mimics endogenous gene expression, but these regions of DNA do not always contain the expected DNA binding sites (personal communication D. Papatsenko and Liberman and Stathopoulos, 2008). Is our knowledge limited or is conservation between binding sited not anticipated over these time-scales? It is a daunting task to experimentally validate and mutagenize every piece of DNA to determine the sites that are necessary and sufficient for generating a given transcriptional output. Recent advances in sequencing may hold the key to reducing the laborious task of identifying regulatory elements individually (see Discussion).

Gene regulatory network

Studying the gene regulatory network that describes the developmental patterning of a fruit fly serves as a tractable model for learning about regulatory control in other organisms. This is such an important area of study because we do not have a firm grasp on how the regulatory DNA functions. We do not know very many of the regulatory sequences and we know even less regarding which transcription factors bind to regions we have identified. The notable exception to this is the regulatory element controlling the *Endo16* gene from the purple sea urchin, which has been studied extensively for over a decade (Davidson, 2006; Yuh et al., 2001).

There is a wiring diagram of the gene regulatory network for dorsal-ventral patterning in *Drosophila* embryos that describes the functional regulatory relationships between transcription factors and their targets (Stathopoulos and Levine, 2005a). This network details which genes are regulated by Dorsal and other downstream transcription factors such as Twist. Genetic studies revealed that Dorsal controls the expression of at least 10 genes along the dorsal-ventral axis (sog: Francois et al., 1994; dpp: Huang et al., 1993; zen: Ip et al., 1991; rhomboid: Ip et al., 1992a; sna: Ip et al., 1992b; brinker: Jazwinska et al., 1999; twist: Jiang et al., 1991; single-minded: Kasai et al., 1992; Kasai et al., 1998; tolloid: Kirov et al., 1994; Pan and Courey, 1992; Thisse et al., 1991). Of these 10 genes, 7 regulatory elements were found by a variety of methods. Microarray screens provided evidence for approximately 40 additional targets (Stathopoulos and Levine, 2002c). These studies focused primarily on footprinting data and searching for clusters of Dorsal binding sites, which had been helpful for identifying regulatory elements for the anterior-posterior axis (Berman et al., 2002). Using bioinformatics, 11 more regulatory elements were predicted and validated (reviewed in Stathopoulos and Levine, 2004). It has been suggested more recently that this method, although successful under some circumstances, may not have the best predictive power. One reason for this is that there is no

correlation between the size of a particular regulatory element and the number of motifs found inside that region (Papatsenko and Levine, 2005b). The identification and validation of regulatory elements has been labor intensive, but the resulting 30 sequences give insight into the regulation of each expression domain along the dorsal-ventral axis. Despite the intricate map of predicted and known interactions, the gene regulatory network in its current format cannot account for the broad lateral expression of the presumptive neurogenic ectoderm.

In Chapter 2 we investigate how the neurogenic ectoderm is patterned. Since the ventral neurogenic ectoderm has been very well characterized, we decided to perform an analysis of the *cis*-regulatory elements that are capable of generating expression in broad lateral domain of the presumptive neurogenic ectoderm. Two regulatory elements had been identified for this expression domain when we started our investigation: one for *short gastrulation (sog)* and one for *thisbe (ths)*. We chose to analyze the *sog* regulatory element because it is a small, well-characterized regulatory element and thus more tractable for *cis*-regulatory manipulation. We learned that regulation of the *sog* gene requires Dorsal, functioning with a ubiquitous activator to generate a broad lateral expression domain.

Synthetic regulatory elements were useful for dissecting the *sog* regulatory element because they allow the characterization of binding sites in isolation from their native environment and provide a way for learning about the regulatory power of putative and established transcription factor binding sites. In the following chapter we discuss a few synthetic elements and determine the combinations of sites that are necessary and sufficient for controlling expression in the broad lateral domain of the presumptive neurogenic ectoderm. These synthetic experiments are easy to design and execute with simple PCR reactions. Coupling synthetic regulatory element generation to the site directed integration techniques discussed earlier allows for fast generation and maintenance of transgenic lines that report the expression of a regulatory element of interest.

The transcription factors highlighted in gene regulatory networks in general, and in the dorsal-ventral gene regulatory network in particular, are spatially distributed. The dorsal-ventral gene regulatory network does not account for the contributions of ubiquitously distributed transcription factors and does not consider timing of the onset of transcription, except in the case of transcriptional cascades that rely on the transcription of one factor for the activation of a downstream target. Ubiquitous activators may have been overlooked in the network model of transcriptional activation and spatial patterning. They could act to expand expression domains by acting in conjunction with graded transcription factors. An example of this type of interaction is seen with the maternal activator, Zelda, which we show is required along with Dorsal to control *sog* expression (see Chapter 2 and Liberman and Stathopoulos, 2008). Future work will determine whether other ubiquitous activators also function with Dorsal to pattern other genes in the presumptive neurogenic ectoderm (see Chapter 2 and Discussion).

We investigate Dorsal further in Chapter 3 asking the basic question: Can Dorsal concentration alone account for the expression domains we see along the dorsal-ventral axis? We designed experiments to test whether the amount of Dorsal in the nucleus could predict the eventual cell fate. Although Dorsal has been extensively well studied, these fundamental questions regarding Dorsal regulatory behavior remain unknown. We start our analysis by observing where the Dorsal gradient is located with regard to the gene expression of target genes which require Dorsal for proper positioning. We find that Dorsal is present in the nuclei of the presumptive ventral neurogenic ectoderm. This is the domain where *vnd* is expressed and the ventral half of the *sog* expression domain. We found it notable that Dorsal is only present at very low levels in nuclei where *ind*, a presumed Dorsal target, is expressed.

Considering this observation, we decided to test whether Dorsal is ever present in more dorsal nuclei. We computationally staged embryos during nuclear cycles 10 through 14. We found that Dorsal is present in these nuclei, but the levels are uniform through the dorsal regions

of the embryo and thus could not provide additional positional information for regulating the expression of *ind* or *sog*. We also find that gene expression is not directly correlated with Dorsal nuclear concentration when the nuclear concentration is genetically manipulated. In mutant embryos with supposedly uniformly low levels of Dorsal, both *vnd* and *ind* expression is seen. These two genes were previously thought to represent different threshold outputs (Stathopoulos and Levine, 2004). However, we show that there aren't two different levels of Dorsal present in these embryos. Our results indicate that Dorsal cannot be acting as a concentration dependent morphogen if multiple thresholds are seen in these embryos. As even Wolpert has noted: "morphogens may represent a crude positional information system" (Kerszberg and Wolpert, 2007). If we accept that morphogens are not precise positional determinants, as once believed, we are left with the task of determining what information these spatially distributed transcription factors relay to their targets and what other inputs are simultaneously regulating gene expression. The following chapters serve as my attempt at this pursuit.

Chapter 2

Design Flexibility in *cis*-Regulatory Control of Gene Expression: Synthetic and Comparative Evidence

Abstract

In early *Drosophila* embryos, the transcription factor Dorsal regulates patterns of gene expression and cell fate specification along the dorsal-ventral axis. How gene expression is produced within the broad lateral domain of the presumptive neurogenic ectoderm is not understood. To investigate transcriptional control during neurogenic ectoderm specification, we examined divergence and function of an embryonic *cis*-regulatory element controlling the gene *short gastrulation (sog)*. While transcription factor binding sites are not completely conserved, we demonstrate that these sequences are *bona fide* regulatory elements, despite variable regulatory architecture. Mutational analysis of conserved putative transcription factor binding sites revealed that sites for Dorsal and Zelda, a ubiquitous maternal transcription factor, are required for proper *sog* expression. When Zelda and Dorsal sites are paired in a synthetic regulatory element, broad lateral expression results. However, synthetic regulatory elements that contain Dorsal and an additional activator also drive expression throughout the neurogenic ectoderm. Our results suggest that interaction between Dorsal and Zelda drives expression within the presumptive neurogenic ectoderm, but they also demonstrate that regulatory architecture directing expression in this domain is flexible. We propose a model for neurogenic ectoderm specification in which gene regulation occurs at the intersection of temporal and spatial transcription factor inputs.

Introduction

Patterned specification of cell fate results from differential gene expression. Differential control of gene expression is accomplished by site-specific transcription factors, which bind DNA to regulate expression over developmental space and time and are themselves regulated at the level of expression or activity. *Cis*-regulatory regions determine how individual genes respond to varying levels and combinations of transcription factors found in different cells during development. However, cell type is discrete and developmental pattern is precise. Therefore, patterning depends on the function of *cis*-regulatory regions to integrate information from transcription factors to produce differential gene expression states in the developing embryo. The architecture of these regulatory regions is complex and the logic behind the organization needs to be determined empirically (e.g. Brown et al., 2007; Davidson, 2001; Deplancke et al., 2006; Ochoa-Espinosa and Small, 2006; Zinzen et al., 2006).

Dorsal-ventral axis patterning during *Drosophila* embryogenesis is a well-studied system that is poised for understanding *cis*-regulatory mechanisms driving development. Over 25 *cis*-regulatory sequences have been identified for over 60 genes known to control different aspects of dorsal-ventral patterning (Stathopoulos and Levine, 2005a). Three presumptive germ layers form along the dorsal-ventral axis in the developing *Drosophila* embryo: mesoderm in ventral regions, neurogenic ectoderm in lateral regions and ectoderm and amnioserosa in dorsal regions. The specification of these germ layers is dependent on the NFκB-like transcription factor, Dorsal, which localizes to the nucleus in a gradient with highest amounts in ventral regions and lowest amounts in dorsal regions (reviewed in Moussian and Roth, 2005). Although Dorsal has been studied extensively, questions remain about how this analog gradient of nuclear Dorsal can direct discrete target gene expression outputs.

Combinatorial interactions between Dorsal and other transcription factors surely contribute to the distinct outputs of gene expression. In ventral and ventral-lateral regions, a synergistic relationship between the bHLH transcription factor, Twist, and Dorsal, has been demonstrated to establish the mesodermal and ventral-neurogenic cell fates (Ip et al., 1992b; Jiang and Levine, 1993; Markstein et al., 2004). Furthermore, in dorsal regions of the embryo, the ectoderm and amnioserosa form as a result of repression by Dorsal and activation by ubiquitous transcription factors to regulate the expression of genes such as *decapentaplegic* (*dpp*) (Liang et al., 2008; Rusch and Levine, 1997). However, the regulatory architecture required to support expression in a broad lateral domain, encompassing the entire presumptive neurogenic ectoderm region of the early embryo, has not been clearly defined (reviewed in Stathopoulos and Levine, 2004).

Only two regulatory elements that direct expression in a broad lateral domain within the early embryo have been identified, those controlling expression of the genes *short-gastrulation (sog)* and *thisbe* (*ths*). These regulatory elements were found by searching for clusters of high-affinity Dorsal binding sites in the genome and have been validated (Markstein et al., 2002; Stathopoulos et al., 2002). These regulatory elements have similar binding site composition: both contain multiple Dorsal binding sites, sites for the ventral repressor, Snail, and the presence of an overrepresented sequence, TTCCAGC, also called GCTGGAA, which we will refer to as the T motif (Stathopoulos and Levine, 2002c). These *cis*-regulatory elements also contain the CAGGTAG motif and other similar heptamers, collectively referred to as TAGteam sites (De Renzis et al., 2007; ten Bosch et al., 2006). The maternal transcription factor, Zelda, also known as vielfaltig (Staudt et al., 2006), binds specifically to these heptamers and is a critical player in zygotic genome activation (Liang et al., 2008). However, the requirement of all these putative binding sites (i.e. Dorsal, Snail, T motif, Zelda) to direct *sog* and *ths* early embryonic expression has not been rigorously tested. One reason that more neurogenic ectoderm regulatory elements

have not been found could be that variable combinations of *cis* and *trans* factors are capable of directing expression in the presumptive neurogenic ectoderm.

It has been demonstrated that flexibility can occur in regulatory element structure with little to no effect on transcriptional output. Regulatory regions with variable binding site composition are capable of generating expression in the same tissue in *Caenorhabditis elegans* (Guhathakurta et al., 2002; Hunt-Newbury et al., 2007). Studies in sea urchin have found that flexibility in both *cis* and *trans* regulators can exist while still producing conserved expression of the *Endo16* gene (Romano and Wray, 2003). More recently, a study comparing *even-skipped* gene regulatory elements in *Drosophilids* and *Sepsids* showed that although there is minimal sequence conservation, functional conservation of regulatory elements remains (Hare et al., 2008). Additionally, an extensive study of co-expressed genes in *Ciona* demonstrates that different motif architectures are tolerated to generate co-regulation of genes (Brown et al., 2007). Such flexibility in the organization and composition of binding sites within *cis*-regulatory sequences might provide a method for "buffering" during development, allowing organisms to develop reproducibly even when the regulatory regions of DNA are altered throughout the course of evolution.

In this analysis, we explore the transcriptional architecture required to pattern the neurogenic ectoderm in *D. melanogaster* embryos. Specifically, our goal was to define the transcription factor binding sites necessary and sufficient to direct expression within the broad lateral domain of early embryos. We define the underlying logic within the minimal *cis*-regulatory element, which supports expression of *sog* in *Drosophila* early embryos, using both evolutionary comparisons and synthetic reporter constructs. Collectively, our results support the view that flexible regulatory element structures are capable of producing similar transcriptional outputs.

Methods

Regulatory element alignments and annotations

Cartwheel (http://cartwheel.caltech.edu/) and JASPAR (http://jaspar.genereg.net/) were used to generate Position Weight Matrices (PWMs) from *in vitro* binding data (Brown et al., 2005; Sandelin et al., 2004). These matrices were used to scan putative regulatory regions for motifs of interest. For a complete list of motifs, Cartwheel-generated consensus sequences, threshold values and probabilities of these matrices occurring randomly in a one kilobase (Kb) sequence, (see Table 1).

Table 1. Thresholds and False positive probabilities for each of the motifs

Name	Consensus	Probability	Threshold	Reference
TAGTeam	YAGGYAD	3.7E-04	14.7	(ten Bosch et al., 2006)
Snail	CAGGTG	9.8E-04	27.5	(Mauhin et al., 1993)
Snail	DCADRDNN	9.2E-04	21	Papatsenko personal comm.
Snail	CACCT	9.8E-04	match 4/4	(Markstein et al., 2002)
Snail	MMRCAWGT	2.4E-04	match 8/8	(Stathopoulos and Levine, 2005b)
Hb	GCATAAAAAA	<1 hit / kb	29.21	(Stanojevic et al., 1989)
Schnurri	GRCGMCWVWBHG TCTG	<1 hit / kb	25	(Pyrowolakis et al., 2004)
D-STAT	TTTCCCGGAAA	<1 hit / kb	42.94	(Yan et al., 1996)
Twist	ACATATG	8.5E-04	40.16	(Lee et al., 1997)
Daughterless	CACCTGC	6.1E-04	40.73	Senger personal comm.
bHLH	CANNTG	3.9E-03	match 6/6	(Murre et al., 1994)
Dorsal	GGGAATTCC	8.4E-04	49	Senger personal comm.
NFKappaB	GGGAATTTCC	<1 hit / kb	39	(Vlieghe et al., 2006)
Dorsal	GGGWDWWWCCM	<1 hit / kb	match 11/11	(Markstein and Levine, 2002)
TTCCAGC	TTCCAGC	6.1E-05	match 7/7	(Stathopoulos and Levine, 2002c)
Pointed	SNGGAWRY	9.0E-04	14.3	(Xu et al., 2000)
Su(H)	BTGTGGGAAMCGA GAT	<1 hit / kb	30	(Bailey and Posakony, 1995)

p*2*1000 = per site probability of finding a motif with these parameters at random

Homologous sequences were obtained for seven of the twelve *sog Drosophilid* sequences (Papatsenko and Levine, 2005a). A complete list of all the predicted *sog Drosophilid* homologous sequences is available (http://flydev.berkeley.edu/cgi-bin/Annotation/enhancers/sog.htm; D. Papatsenko, in preparation). Sequences were loaded onto the Cartwheel site and scanned for binding sites using the previously generated PWMs. Cartwheel generates false positive statistics

for each of the matrices (listed in Table 1). We used these statistics to set thresholds which correspond to one or fewer false positive match per kilobase of sequence for all of the putative binding sites. Snail does not have particularly good binding site predictions. To adjust for this, four motifs were used to find putative Snail binding sites. The same methods were used to find binding sites in the *thisbe (ths*) regulatory elements.

The *Neu3 D. melanogaster* regulatory element sequence we tested was used to find homologous regulatory elements in each of the twelve sequenced *Drosophilids.* Briefly, UCSC BLAT search was used to find sequences of high similarity in the other *Drosophilid* genomes (http://genome.ucsc.edu/cgi-bin/hgBlat). In the case of the identification of *D. virilis* homologous sequences, the Drosophila genome version "April 2004" must be selected.

Vector construction

All of the *even-skipped* (eve) promoter LacZ (eve.p-lacZ) fusion elements used a modified pLacZattB vector, with the eve minimal promoter inserted in place of the hsp70 minimal promoter (Bischof et al., 2007; Jiang et al., 1991). *cis*-regulatory modules were amplified from genomic DNA, cloned into the NotI site of the eve.p-LacZ.attB vector and verified by sequencing. Synthetic *cis*-regulatory elements were constructed from oligonucleotides and cloned into the pGEMT-easy vector (Promega) or directly into the BglII and NotI sites of the eve.p-LacZ.attB. All constructs were verified by sequencing.

Site-directed mutagenesis

Primers were designed to mutate sites within the *sog cis*-regulatory element using the QuickChange SiteDirected Mutagenesis Kit from Stratagene (for primer sequences see Supplemental Materials & Methods). Genomic DNA was used as a template for PCR reaction to amplify the *sog* regulatory element. It was cloned into the pGEMT-easy vector, which was

subsequently used as the template for mutagenesis reactions.

Generation of transgenic fly lines

Phi-C31 mediated site-specific integration of *cis*-regulatory element-reporter fusions was done as described into either ZH-attp51D or attp16 (Bischof et al., 2007; Groth et al., 2004; Markstein et al., 2008). Embryo injections were performed in house and with help from Rainbow Transgenic Flies (Newbury Park, CA) and Genetic Services Inc. (Sudbury, MA).

In situ hybridization

Digoxigenin-UTP-labeled *LacZ* antisense RNA probes were used to detect *LacZ* reporter gene expression as described previously with a few modifications (Jiang and Levine, 1993; Tautz and Pfeifle, 1989). Briefly, embryos were collected, aged to be 2-4 hours old, dechorinated in 100% Sodium hypochlorite (Sigma #239305) for 3 minutes, washed and transferred to a scintillation vial with 3 mL buffer (1.3XPBS, 67 mM EGTA pH 8.0), 4 mL heptane, 1 mL 37% formaldehyde solution. Embryos were fixed for 20 minutes and then MeOH was used to remove the vitelline membrane. *D. mojavensis* and *D. pseudobscura* embryos were fixed with 1.6 mL buffer, 8 mL heptane, 0.4 mL paraformaldehyde solution (Electron Microscopy Sciences #15713-S).

Fly lines

Drosophila species were obtained from the Drosophila Species Stock Center (https://stockcenter.ucsd.edu/info/welcome.php). Dorsal mutant analysis was performed with dl^1 cn^1 sca^1/CyO, l(2)DTS100[1] , and Twist mutant analysis was carried out using twi cn bw/ CyO ; both stocks are available from the Bloomington Stock Center.

Results

Broad lateral expression of *sog* is conserved among *Drosophilids*

The genomes of 12 *Drosophila* species have been sequenced, facilitating the analysis of coding and regulatory regions spanning approximately 40 million years of evolution (Clark et al., 2007). In *Anopheles, sog* expression is different from *D. melanogaster* in that it is found in ventral regions of the embryo (Goltsev et al., 2007). We decided to determine whether *sog* expression is conserved or divergent within early embryos from a phylogenetically representative set of seven of the twelve sequenced *Drosophilids*: *D. melanogaster, D. yakuba, D. simulans, D. ananassae, D. pseudobscura, D. mojavensis,* and *D. virilis*. The broad lateral expression pattern of *sog* in *D. melanogaster* was conserved when compared with endogenous *sog* expression in the *Drosophilids* we examined (Fig. 1D, G, J, and Supplemental Figure 1A, C, compare with Fig. 1A). We found that expression was maintained in a broad lateral stripe even when the size of the embryos varied. *D. yakuba*, *D. simulans,* and *D. mojavensis* embryos are all slightly smaller than *D. melanogaster* on average; *D. virilis* and *D. ananassae* embryos are longer along the anterior-posterior axis and thinner along the dorsal-ventral axis than *D. melanogaster*. Nevertheless, *sog* expression is absent from both ventral and dorsal-most regions of the embryos in these divergent *Drosophilids*, as observed in *D. melanogaster*. The sharp ventral border of *sog* expression due to Snail repression in ventral regions, seen in *D. melanogaster*, is also apparent in the other *Drosophilids* we tested.

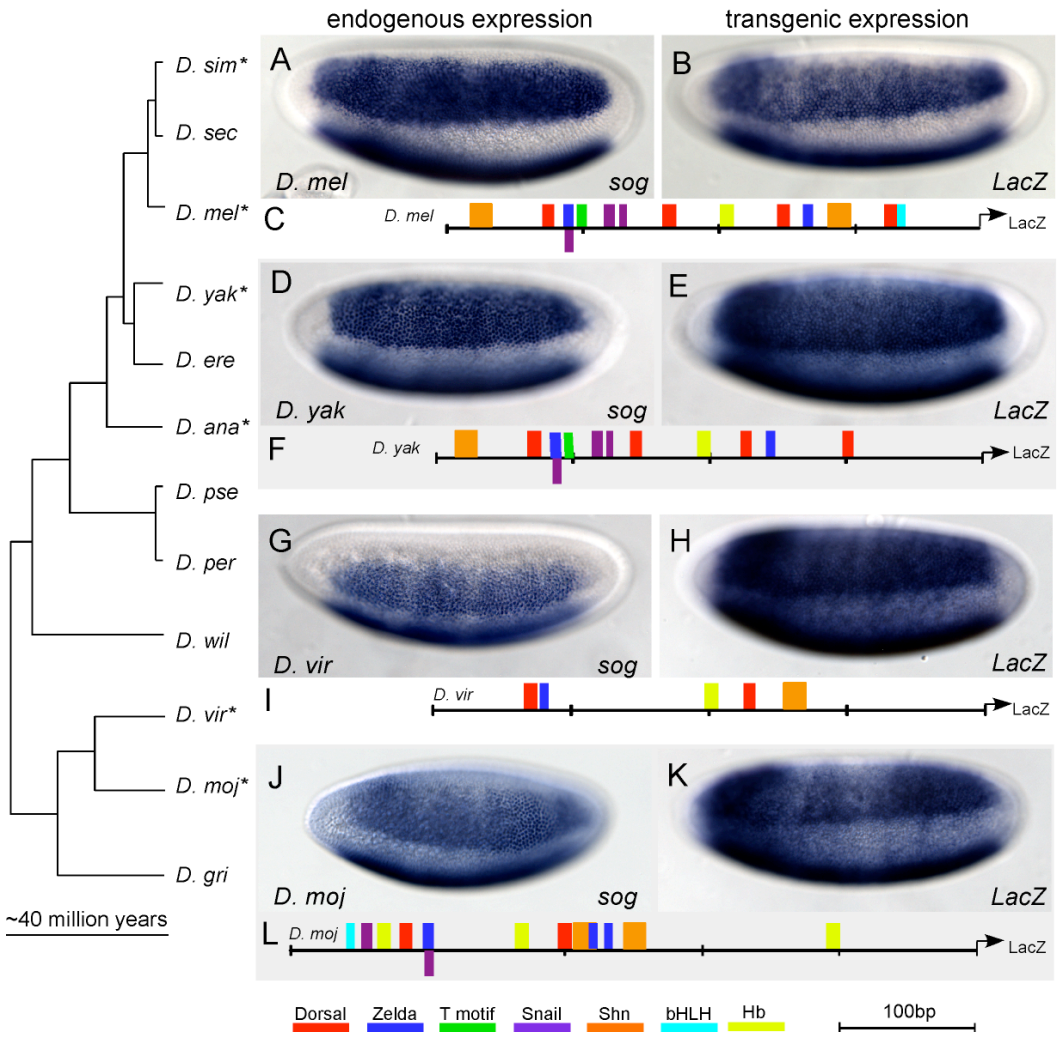


Fig. 1. Reporter fusions in *D. melanogaster* reveal conservation of expression and regulatory logic. Whole mount *in situ* hybridization using anti-sense probes to mRNA transcripts in various *Drosophilid* species. In (A, D, G, and J), the expression of *sog* is depicted. In (B, E, H, K), reporter gene expression is detected by *in situ* hybridization using a riboprobe recognizing the *lacZ* gene. *sog* is expressed broadly in lateral regions of the embryo in (A) *D. melanogaster (D. mel),* (D) *D. yakuba (D. yak)*, (G) *D. virilis (D. vir)*, and (J) *D. mojavensis (D. moj)*. Expression is absent from ventral and dorsal regions of the embryo and the anterior and posterior poles. Cartoons of putative minimal *sog* regulatory elements of (C) *D. mel,* (F) *D. yak*, (I) *D. vir*, and (L) *D. moj* which were fused to a *LacZ* reporter and integrated into *D. melanogaster* embryos (B, E, H, and K, respectively). Stronger expression (denoted by the asterisk) present in all transgenic embryos in a band at the anterior is associated with vector sequence and likely due to the *lacZ* gene, (Jiang et al., 1991). In this figure and all subsequent ones, embryos are oriented with anterior to the left and dorsal side up. The asterisk denotes stronger expression in a band at the anterior is associated with vector sequence, present in all transgenic embryos and most prominent at cellularization (Jiang et al., 1991). The embryos are tilted ventral-laterally in order to show repression in ventral regions; thus both lateral stripes of *sog* expression are in view, though the domain of expression located at the bottom of the images is only a partial view of the broad lateral expression domain. Zelda refers to all of the TAGteam motifs.

Validation of homologous *sog* regulatory regions

A *sog cis*-regulatory module that drives expression in the broad lateral domain of early *D. melanogaster* embryos was previously identified and verified in a genome-wide search for clusters of Dorsal binding sites (Markstein et al., 2002). This minimal *cis*-regulatory module from *D. melanogaster* was used to find homologous DNA sequences in the six other *Drosophilid* species (see Methods). We tested whether these putative *cis*-regulatory elements were able to support expression of a reporter in the presumptive neurogenic ectoderm of *D. melanogaster*.

Constructs containing DNA of the presumptive *sog cis*-regulatory modules isolated from six species were fused to a reporter gene (i.e. *LacZ* or *Cherry*) and integrated into the *D. melanogaster* genome by PhiC31 integration (see Methods). By using site-specific integration methods, we are confident that our comparative analysis of regulatory sequences is not confounded by positional effects, which can result when P-elements are used to generate transgenic lines as a result of random integration of the reporter gene construct into the genome (Levis et al., 1985). All the transgenic constructs direct expression similar to that supported by the minimal *sog cis*-regulatory module previously identified from *D. melanogaster* (Fig. 1E, H, K, and Supplemental Fig. 1B, D, compare with Fig. 1B), which itself is comparable in expression to the endogenous *sog* gene at this same stage (Figs. 1A and 2B; and Markstein et al., 2002). These results demonstrate that the homologous sequences are functionally conserved regulatory elements.

There are minor differences in the borders of expression for various regulatory element reporters. In particular, there is a slight difference in the ventral borders of both the *D. virilis* and *D. mojavensis sog cis*-regulatory element reporters as compared with the other reporter constructs (Fig. 1H and K, compare with 1B). The border appears less sharp than in the *D. melanogaster sog* regulatory element reporter and when compared to endogenous *sog* expression. Since the

endogenous expression boundary is discrete, we reasoned that there must be changes in *cis*- or *trans*-factors that influence the transgenic reporter expression. In fact, there are more putative binding sites for the Snail transcriptional repressor in the *D. melanogaster cis*-regulatory module (i.e. three sites), than are found in the *cis*-regulatory modules of *D. viliris* and *D. mojavensis* (i.e. one site and two sites, respectively). *D. virilis* and *D. mojavensis* may use other transcription factors, which are not functional in the context of the *D. melanogaster* embryo, to support repression in ventral regions (i.e. *cis* effects). Alternatively, changes in the Snail protein within these other species (i.e. *trans* effects) may contribute to changes in binding site preference such that we no longer can predict binding sites using the PWM defined by *D. melanogaster* data.

Despite subtle differences in the expression patterns supported by these divergent sequences, all the predicted *cis*-regulatory modules do indeed direct expression of a reporter in a broad lateral domain within early *D. melanogaster* embryos. We hypothesized that a core set of conserved binding sites and transcription factors bind to all the regulatory elements tested to drive reporter expression in this broad lateral expression domain.

Identification of conserved binding sites within homologous *sog* regulatory regions

In order to determine the requirements for patterned broad lateral expression, we set out to identify the functional set of transcription factor binding sites within the minimal *D. melanogaster sog* regulatory element. To date, several predicted transcription factors binding sites and overrepresented motifs, which act to pattern expression during *D. melanogaster* embryogenesis, have been identified (Markstein and Levine, 2002; Markstein et al., 2002; Mauhin et al., 1993; Muller et al., 2003; Ochoa-Espinosa and Small, 2006; Papatsenko and Levine, 2005a; Pyrowolakis et al., 2004; Stanojevic et al., 1989; Stathopoulos and Levine, 2002c; Stathopoulos and Levine, 2004; ten Bosch et al., 2006; Vlieghe et al., 2006; Xu et al., 2000; Yan

et al., 1996). We used the results of these *in vitro* binding studies as well as degenerate binding site predictions (Murre et al., 1994) to construct position weight matrices that describe the binding sites preferences exhibited by Dorsal, Zelda, Smad/Schnurri, D-STAT, Snail, bHLH proteins (including Daughterless and Twist), and Hunchback DNA-binding proteins (see Methods). We also analyzed whether the overrepresented T motif (TTCCGCA) was present, as this motif was found previously to be associated with the broad lateral expression in the early embryo (Stathopoulos and Levine, 2002c).

Using these position weight matrices (PWMs), we scanned for putative transcription factor binding sites in the *sog* regulatory element from *D. melanogaster* using the Cartwheel program (http://woodward.caltech.edu/canal/; Brown et al., 2005). We identified the four Dorsal binding sites, two sites for the Snail repressor, one T motif site, and two TAGteam sequences all of which had been previously identified within this *cis*-regulatory module. We identified several novel sites as well, including two binding sites for Schnurri, the transcriptional co-repressor, one degenerate bHLH site, an additional Snail site, and one Hunchback site, an activator which functions along the anterior-posterior axis (Fig. 2D and Supplemental Fig. 2).

We examined the conservation of these binding sites within the homologous regulatory elements from other *Drosophilids* in an effort to define the essential features of the minimal regulatory element. We searched for the same putative binding sites identified within the *D. melanogaster sog* regulatory element in the functionally conserved regulatory sequences from *D. simulans*, *D. yakuba*, *D. annanassae*, *D. psuedobscura*, *D. virilis* and *D. mojavensis* (see Fig. 1C, F, I, and L).

Our results reveal conserved clusters of binding sites among otherwise non-conserved sequence (see boxes, Fig. 2D and Supplemental Fig. 3). Using the Cartwheel program, we defined threshold cutoffs for matches to PWMs such that conserved binding sites were found and sites that were likely to appear in the sequence randomly were rejected (see Table 1 and Methods,

Brown et al., 2005). These conditions were used for all of the putative binding site sequences except Snail, as the binding data for this factor is not as well defined as for the others (see Methods). Using this program we found only one Dorsal binding site is conserved, in sequence and position, throughout the homologous *sog cis*-regulatory modules examined. One Zelda site is also conserved, in sequence and position, until the divergence of *D. virilis* and *D. grimshawi*; moreover this particular site retains close proximity to the conserved Dorsal site. A previous study of *cis*-regulatory modules regulated by the Dorsal transcription factor also identified conservation of one Dorsal and one linked TAGteam site within the *sog cis*-regulatory module (Papatsenko, 2007). In addition, we identified conserved binding sites for Snail and Schnurri. We found bHLH sites throughout the diverged sequences, however these sites were not conserved in position or exact sequence. Considering approximately 40 million years of evolution between *D. melanogaster* and *D. virilis* or *D. mojavensis*, the fact that specific DNA sequences are conserved suggests they were maintained by selection.

Neurogenic ectoderm specification involves dynamic expression

We examined the expression pattern of *sog* in embryos and document the dynamic nature of the transcript (Fig. 2A-C). At early stages, approximately nuclear cycle 9/10, *sog* is expressed ubiquitously throughout the embryo with strongest expression in ventral regions of embryos (Fig. 2A). At cellularization, nuclear cycle 14, *sog* is expressed in a broad lateral stripe, and later expression refines to encompass the mesectoderm (Fig. 2B and C); these expression patterns have been previously documented (Francois et al., 1994).

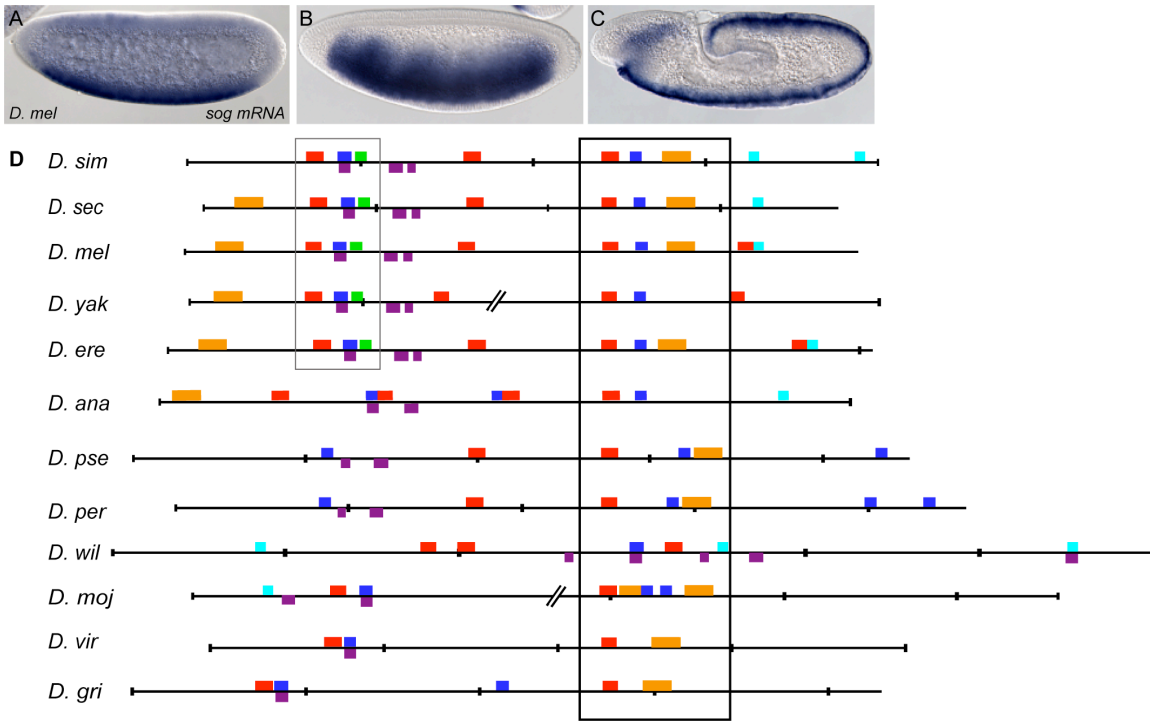


Fig. 2. Alignment of putative *sog cis*-regulatory elements from other *Drosophilids* reveals conservation and turnover of binding sites.
(A-C) *sog* expression is dynamic. Endogenous expression of the *sog* gene detected by *in situ* hybridization using a riboprobe within embryos of nuclear cycle ~11 (A), cycle 14/stage 5 (B), and during germ-band elongation after gastrulation (C). (D) Shown are the predicted binding sites for transcription factors and over-represented motifs that are associated with neurogenic ectoderm patterning within the *sog cis*-regulatory modules identified from 12 *Drosophilid* species. Position weight matrices (PWMs) were used to find putative transcription factor binding sites using the program Cartwheel (Brown et al., 2005). Alignments were generated on the UCSC genome browser webpage (http://genome.ucsc.edu). Cartoons were generated by Cartwheel and then colored according to the key. Gaps in the alignments, shown as broken lines, were introduced to help visualize conservation. Box domains represents well-conserved region of *sog cis*-regulatory element. The sites boxed in black are located in the most well-conserved region, and the sites boxed in grey are the second most well-conserved region. Note that closely associated Dorsal and Zelda binding sites are present in all of the alignments, though the location of these sites within the sequence can vary. Full alignment can be viewed in Supplemental Fig. 3.

We find that expression within all three of these domains (Fig. 2A-C) is controlled by the same *sog cis*-regulatory module; one regulatory element controls three distinct patterns of gene expression (Fig. 3A, 3B, and data not shown). The *sog cis*-regulatory module drives reporter gene expression in a ubiquitous domain within embryos at early stages (Fig. 3A), and this early ubiquitous expression as well as the other patterns of expression controlled by the reporter gene

are zygotic, as expression is present when the transgene is introduced paternally (data not shown). This observation suggested the hypothesis that an early ubiquitous activator may function together with Dorsal to regulate expression of *sog*.

Mutagenesis of sites reveals conservation of function and spatial organization

In order to dissect the core regulatory logic of the *sog* regulatory element, we took a candidate approach and mutated conserved binding sites within the *sog cis*-regulatory sequence in order to determine which sites are important for regulation. We mutated sites we thought most likely to promote expression in the neurogenic ectoderm taking into account two criteria: (1) whether the site was conserved in our comparative analysis of orthologous *sog cis*-regulatory sequences and (2) whether there was evidence to suggest the proteins that recognize these sites function to regulate expression along the dorsal-ventral axis. Predicted sites for Dorsal, Zelda, Schnurri/Smad, and Snail were all conserved, in sequence and relative position, in the comparisons of divergent *Drosophilid* sequences (Fig. 2D).

However, since our goal was to identify how activation of *sog* is produced in a broad lateral domain even as the levels of nuclear Dorsal diminish, we limited our analysis to putative activators of *sog* expression that might function during cellularization. The bHLH protein, Twist, functions with Dorsal to control expression of genes within the presumptive ventral neurogenic ectoderm, in a lateral stripe encompassing 5-7 cells (Jiang and Levine, 1993). In *twist* mutants embryos, *sog* expression remains broad in a lateral stripe ~15 cells wide. The only change in expression is that the ventral border extends slightly into ventral regions, presumably due to the fact that lower levels of Snail repressor are present (data not shown). Similarly, when the *sog* regulatory element is crossed into the *twist* mutant background, reporter expression remains broad but slightly expanded into ventral regions (data not shown). Considering this information

and given that the bHLH site was not conserved in the other *Drosophila* species, we chose not to investigate whether Twist contributes to *sog* expression. Schnurri has been documented to function as a transcriptional repressor only after embryos have completed germ-band elongation (Pyrowolakis et al., 2004); therefore we did not expect Schnurri to effect the early embryonic *sog* expression pattern. Snail protein likely represses *sog* in ventral regions, because *sog* expression is expanded in *snail* mutant embryos (Kosman et al., 1991). For these reasons, we chose to focus our efforts on the requirement of documented transcriptional activators Dorsal and Zelda, as well as on the T motif, since its function was undefined.

Consistent with Dorsal playing a key role in controlling dorsal-ventral patterning, we find that Dorsal sites are required to generate a broad lateral expression pattern. When all four Dorsal binding sites in the *sog* regulatory element are mutated, early ubiquitous expression of *sog* is unperturbed, but expression of the reporter at nuclear cycle 14 is restricted to the ventral neurogenic-ectoderm forming a narrow band of expression in 3-5 cells (Fig. 3D, compare with 3B). Furthermore, within the set of *sog cis*-regulatory modules sequences, we found that only one of the four Dorsal binding sites was conserved, both in sequence and position, in 11 of the 12 species examined (see Fig. 2D; black box). In order to test the significance of this highly conserved Dorsal binding site, we mutated this site to examine the effect on reporter expression. We found that mutating this site alone produced a severe reduction in expression at nuclear cycle 14, which was almost as acute as mutagenesis of all four Dorsal binding sites (Fig. 3J and 3D, respectively, compare with Fig. 3B). These results suggest that Dorsal transcription factor binding to these sites is crucial for broad expanded expression into lateral regions at cellularization.

We also analyzed the requirement for Zelda to direct *sog* expression. TAGteam sites are recognized by Zelda, a recently described transcription factor that is maternally deposited and thus ubiquitously expressed in the early embryo (Liang et al., 2008). The presence of TAGteam sites in *cis*-regulatory elements has been associated with ubiquitous expression in the early embryo

(De Renzis et al., 2007; ten Bosch et al., 2006). We mutagenized the two TAGteam sites (i.e. Zelda sites) present in the *sog cis*-regulatory element, and observed that ubiquitous early activation of the reporter is almost completely eliminated (Fig. 3E). At nuclear cycle 14, reporter gene expression is restricted to the ventral neurogenic ectoderm, as observed when Dorsal sites are mutated (Fig. 3F, compare with Fig. 3D). Our mutagenesis results indicate a role for Zelda in directing early ubiquitous expression (Fig. 3E), as well a secondary role for Zelda in controlling expression of *sog* in a broad lateral domain later (Fig. 3F).

Additionally, we find evidence that an unknown factor, which presumably binds to the T motif, is necessary for proper expression of the reporter in the presumptive neurogenic ectoderm. When the T motif is mutated, reporter expression is still broad, but the expression pattern exhibits modulation along the anterior-posterior axis (see Fig. 3H). This result suggests that there is regulatory input for *sog* from pair-rule transcription factors. It has been shown that mutations that effect dorsal-ventral patterning also influence anterior-posterior patterning both by altering nuclear movements and through transcriptional changes (Carroll et al., 1987; Keranen et al., 2006). Furthermore, transcription factors that pattern the anterior-posterior axis also bind regions in and around many genes that are dorsal-ventral axis determinants, and the reverse is also true (Li et al., 2008; Zeitlinger et al., 2007b). Considering the importance of early patterning events on the ultimate specification of cells, regulatory cross talk between anterior-posterior and dorsal-ventral factors could enable synchronous expression where necessary. Nevertheless, we conclude that the transcription factor that binds to the T motif likely does not contribute to the *sog* expression domain along the dorsal-ventral axis (i.e. the height of the broad lateral stripe), instead this site facilitates modulation of the *sog* expression domain along the anterior-posterior axis.

In addition to testing the necessity of these binding sites, we wanted to test whether the presence of these sites alone was sufficient to direct expression or if spacing of sites was

important. To address this question, we used the construct with mutated Dorsal sites and replaced four Dorsal binding sites proximal to the reporter. The resulting reporter expression is similar to the transgenic with mutated Dorsal sites (Fig. 3L). This result implies that distance between the Dorsal and Zelda sites (i.e. relative spacing) is indeed important for creating a broad lateral expression domain.

Analysis of *sog* expression and reporters in mutant backgrounds

Similar to what we observed in the *cis*-regulatory construct with mutagenized Zelda cites (Fig. 3F), a refined domain of expression for *sog* was recently identified in Zelda mutant embryos (Liang et al., 2008) suggesting that we have indeed disrupted Zelda binding to the *sog cis*-regulatory module sequence. Ubiquitous expression remains in all mutagenized *sog cis*- regulatory module transgenic embryos (Fig. 3A, C, G, I, and K) except those with mutagenized Zelda binding sites, suggesting that Zelda plays a major role in directing early expression of *sog*.

In *dorsal* mutant embryos, *sog* expression is completely eliminated at both early and later stages (data not shown; Francois et al., 1994). Expression of the *sog cis*-regulatory element reporter gene is also absent at all stages we tested in a *dorsal* mutant background (i.e. up to stage 6; data not shown). This complete loss of expression is much more severe than when the Dorsal binding sites are mutagenized within the *cis*-regulatory element (see Fig. 3C,D).

Collectively, our results suggest that Dorsal and Zelda function together to control *sog* expression within a broad lateral stripe. However, at early stages, the mechanism to generate ubiquitous expression remains unclear. It is conceivable that our mutagenesis experiments, which targeted high-affinity Dorsal binding sites, did not completely eliminate Dorsal binding or, alternatively, Dorsal might fulfill an additional role to indirectly influences the ability of Zelda or another factor to support *sog* expression (see Discussion). Nevertheless, a role for Dorsal and

Zelda proteins in supporting expression is clear.

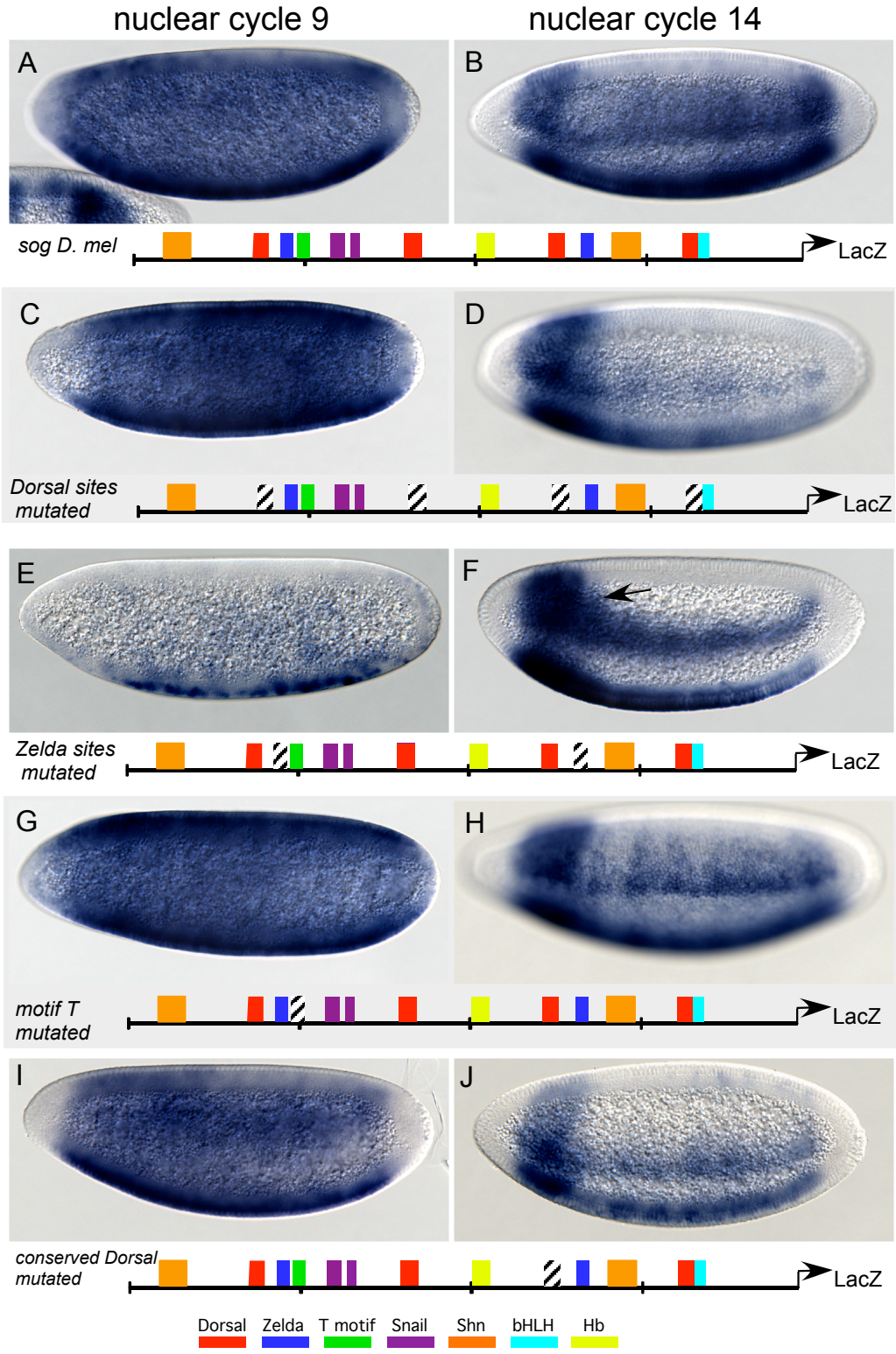


Fig. 3. Mutational analysis of conserved binding sites within the *sog cis*-regulatory module. The minimal *sog cis*-regulatory element, and versions containing mutations introduced by site-directed mutagenesis, were fused to a LacZ reporter and integrated into the *D. melanogaster* genome at positions 51D and 53C4 (ZH-attp51D and attp16 respectively) by site-directed transgenesis. Embryos at nuclear cycle 9 are depicted in (A, C, E, G, I, and K); embryos at nuclear cycle 14/stage 5 are depicted in (B, D, F, H, J, and L). The asterisks denotes stronger expression in a band at the anterior is associated with vector sequence, present in all transgenic embryos and most prominent at cellularization (Jiang et al., 1991). Cartoons below the embryo images represent predicted transcription factor sites within the *sog* cis-regulatory module; the particular sites mutated in each experiment are depicted by diagonal lines.
(A, B) *In situ* hybridizations reveal expression of the reporter at nuclear cycle 9 (A), which is ubiquitous except for repression in the anterior and the pole cells. Expression at nuclear cycle 14 (B) is supported in lateral regions of the embryo, within a broad domain. (C, D) The four Dorsal binding sites were mutated in the *cis*-regulatory element. Expression supported by this mutagenized sequenced was unaffected at nuclear cycle 9 (C), but at nuclear cycle 14 (D) expression is restricted to ventral regions of the neurogenic ectoderm.
(E, F) Zelda binding sites were mutated in the *cis*-regulatory element. Early activation of reporter expression is absent from all but the ventral-most regions of the embryo (E) and at nuclear cycle 14 (H) expression is also diminished compared to wild-type(B). The arrow marks the region of the lateral stripe, closer to the anterior, which shows expanded expression in more dorsal regions, compared with the width of the stripe closer to the posterior.
(G, H) The predicted T motif was mutated in the *cis*-regulatory element. No effect was identified at nuclear cycle 9 (G), but expression of the reporter by this sequence appeared modulated along the anterior-posterior axis (i.e. "stripy") at nuclear cycle 14 (H).
(I, J) The well- conserved Dorsal binding site (within the black box in Fig. 2D) was mutated in the *cis*-regulatory element. Reporter expression was examined at nuclear cycle 9 (I) and nuclear cycle 14 (J); at the later stage (J), expression is restricted to ventral regions of the neurogenic ectoderm.
(K, L) The construct with all the mutagenized Dorsal binding sites (C, D) was amended to contain four Dorsal binding sites proximal to the LacZ reporter. The expression at both stages is similar to the original construct (C, D).

Constructing synthetic regulatory elements from putative binding sites

The results of our analyses [i.e. the identification of conserved sites (Fig. 1 and 2) as well as sites required for broad lateral expression (Fig. 3)], together, suggested that Dorsal, Zelda, and possibly T motif sites are important for *sog* expression. Using this newly acquired information, we designed synthetic *cis*-regulatory elements to attempt to reconstruct the broad pattern found in the presumptive neurogenic ectoderm (Fig. 4).

Neither Dorsal nor Zelda and T motif alone are able to support expression in a broad lateral domain. When the four native Dorsal sites from the minimal *sog cis*-regulatory element

are used to drive reporter expression, early expression is broad, encompassing the ventral and ventral-lateral but not the dorsal-most region of the embryo (data not shown). This result suggests that Dorsal, possibly functioning with a bHLH transcription factor, is capable of generating broader expression, at least transiently, in the early embryo. Expression of this synthetic regulatory element at nuclear cycle 14 is detected in a ventral-lateral stripe, consisting of ~5 cells (Fig. 4A). This is similar to what was observed previously when the proximal element for the Twist *cis*-regulatory element, which includes Dorsal binding sites and Snail binding sites, was used to drive reporter expression (Jiang et al., 1991). Our result confirms that Dorsal binding sites alone are not sufficient to generate the broad lateral expression. When Zelda sites and T motif are used to direct a reporter, early expression is ubiquitous (data not shown), and expression is essentially ubiquitous at nuclear cycle 14, with slight repression in ventral regions of the embryo and some obvious anterior-posterior modulation (Fig. 4B). Although we do not know what factor binds T motif, the protein Zelda is ubiquitously expressed (Liang et al., 2008), suggesting that the expression we see is largely due to Zelda activation. Some of the predicted Snail sites, overlap with the predicted binding domain for Zelda; this likely accounts for the subtle ventral repression observed.

We multimerized the conserved sequence block containing Dorsal, Zelda, and T motif sites as well as one Snail site (delineated by gray box in Fig. 2D) to generate a synthetic *cis*-regulatory construct which was used to drive reporter expression. This synthetic reporter drives early ubiquitous expression (data not shown), which at nuclear cycle 14 refines to a broad lateral stripe of expression (Fig. 4C). Since mutagenizing the T motif did not appear to affect the width of the broad lateral stripe (see Fig. 3H), we tested whether Zelda and Dorsal could function without the T motif. When these two sites, Zelda and Dorsal, are multimerized, they also direct early ubiquitous reporter expression, and furthermore a broad lateral stripe is generated at nuclear cycle 14 (Fig. 4D).

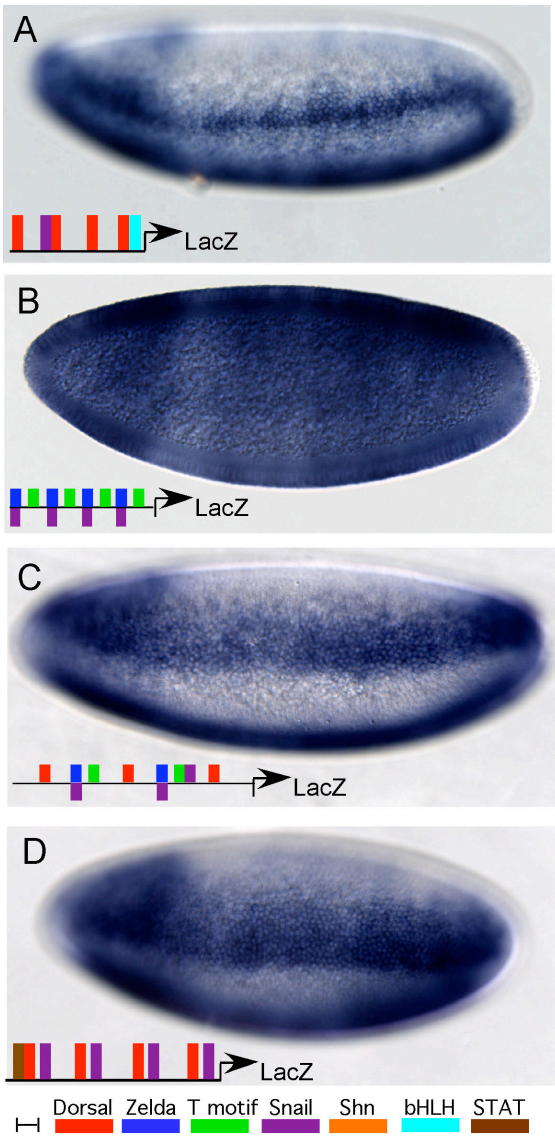


Fig. 4. Synthetic *cis*-regulatory elements constructed from conserved motifs and binding sites. Whole mount *in situ* hybridizations using a riboprobe to *LacZ* to analyze expression supported by various synthetic reporter constructs.
(A) Native Dorsal sites taken from the minimal *sog cis*-regulatory element direct reporter expression in the ventral regions of the neurogenic ectoderm. The exact sequences of the 4 Dorsal binding sites from the *sog cis*-regulatory module were used including 5 bp of linker sequence upstream and downstream from the predicted binding site. Thus, the sites were separated by 10 bp, or one helical turn of DNA. At least one Dorsal binding site was associated with a Snail binding site, which explains the repression observed ventrally. A bHLH site is associated with another Dorsal site.
(B) Multimerized Zelda and T motif sites direct ubiquitous reporter expression. A ~20 bp fragment of the endogenous *sog cis*-regulatory module in which Zelda and T motif sites are linked (see grey box, Fig. 2D) was multimerized so that four copies were assayed.
(C) Dorsal, Zelda, T motif, and Snail sites direct expression in a broad lateral stripe. A ~30 bp fragment of the endogenous *sog cis*-regulatory module in which Dorsal, Zelda, and T motif sites are closely associated (see grey box, Fig. 2D) was assayed. Two copies of this element together with one additional Dorsal site was used to construct a synthetic reporter. (D) Multimerized Dorsal and Zelda sites direct broad lateral reporter expression. A ~31 bp fragment of the endogenous *sog cis*-regulatory module in which Zelda and T motif sites are linked (see grey box, Fig. 2D) was multimerized so that three copies were assayed. Scale bar represents 50 bp.

If Zelda functions as an early temporal activator, this activating role might be replaceable by other activators. To test this idea, we designed other synthetic regulatory elements to direct expression in a broad lateral domain. We used the binding sites from a segment of the *brinker (brk)* regulatory element (Markstein et al., 2004) which contained Dorsal and Snail sites (Fig. 5A). Interestingly, the expression of this synthetic reporter encompasses a broad lateral domain,

where as the entire *brk cis*-regulatory sequence only generated expression in a ventral-lateral domain. We identified that a site for a ubiquitous maternal activator, D-STAT (Yan et al., 1996), was introduced in the process of generating the synthetic element. Therefore, we hypothesized that this STAT site, in combination with Dorsal and Snail sites, may be responsible for directing broad lateral reporter expression. To test this hypothesis directly, STAT sites were used in place of Zelda sites in a similar synthetic background (as Fig. 4D); expression was found to be broad, but occasionally exhibits anterior-posterior modulation (Fig. 5B and data not shown). This is not surprising considering the suggestion that STAT activity is modulated along the anterior-posterior axis by phoshorylation (Shi et al., 2008). This result suggests that a more general mechanism for creating expanded expression domains that are Dorsal-dependent may rely on interactions between Dorsal and other coactivators. For instance, multiple ubiquitous or broadly expressed activators may be competent to interact with Dorsal in order to support expression within the broad lateral domain in question here (see Discussion).

Also of note is the fact that we observed that all of the synthetic *cis*-regulatory elements we generated have expanded expression domains at the anterior and posterior poles. Such expression is not seen when the *sog cis*-regulatory module drives reporter expression, nor when *sog* mRNA expression is observed (Fig. 4, compare with Fig. 2B and 3). This result supports the idea that other transcription factor(s), functioning along the anterior-posterior axis, work to refine *sog* expression. In this particular case, the factor may function downstream of the terminal signaling cascade.

Flexibility of regulatory inputs provides insight for finding additional neurogenic ectoderm-specific regulatory elements.

Previous attempts to identify *cis*-regulatory elements that function in the neurogenic ectoderm focused on the identification of clusters of high-affinity Dorsal binding sites; the hypothesis was that multiple high-affinity Dorsal binding sites could support expression even

where levels of nuclear Dorsal are quite low as is the case in dorsal regions of the neurogenic ectoderm (Markstein et al., 2002; Stathopoulos and Levine, 2002a). This approach was successful in finding both *sog* and *ths cis*-regulatory elements. Yet the *cis*-regulatory elements that drive expression of other genes expressed in a broad lateral expression domain (Neu3: Figure 5, SoxN: Supplemental Figure 4, and pyramus: Stathopoulos and Levine, 2002a) could not be found in this manner. We hypothesized that one reason that the respective *cis*-regulatory elements have remained elusive is that multiple mechanisms may exist to support activation within a broad lateral domain of the early embryo.

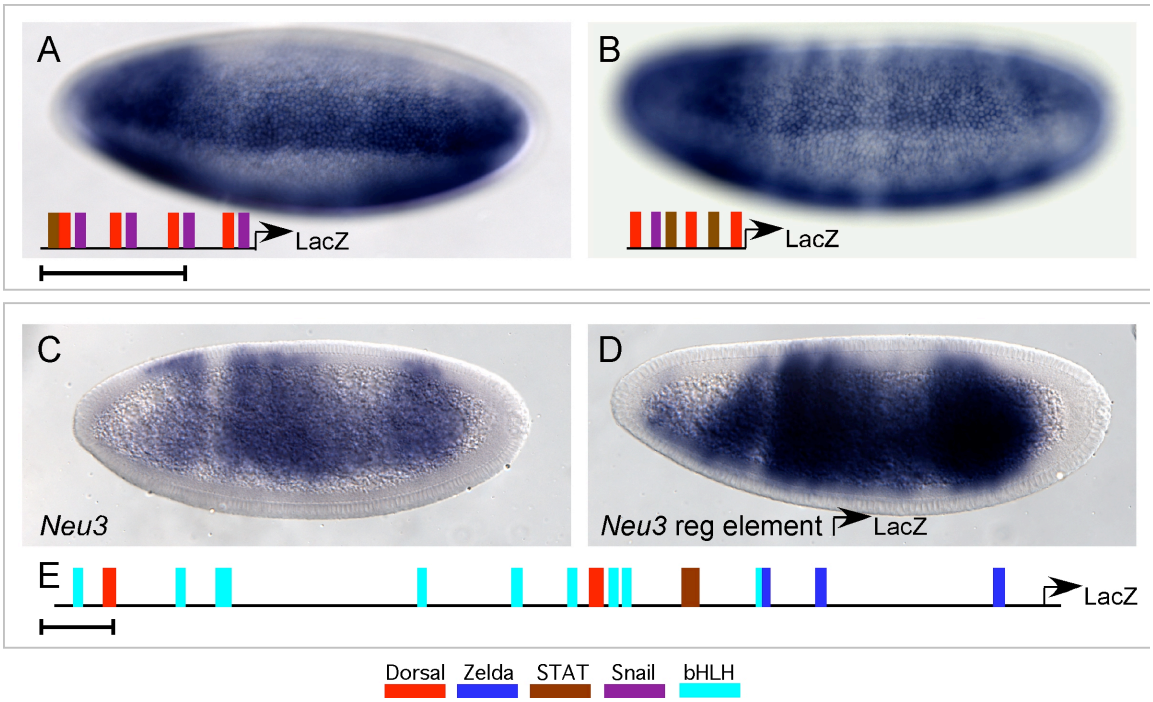


Fig. 5. Identification of a novel regulatory element that functions in the neurogenic ectoderm. (A) Dorsal sites, Snail sites and a STAT site drive expression in a broad lateral stripe. A ~20 bp element derived from the *cis*-regulatory module controlling expression of the *brinker (brk)* gene was multimerized; one STAT site was introduced in the course of cloning. (B) The conserved Dorsal site (grey box Fig. 2D) was used in a synthetic with a Snail site and two STAT sites. Expression is broad, but occasionally stripy. *In situ* hybridization patterns using riboprobes specific for the genes *Neu3* (C) and *LacZ* (D) are depicted. (C) *Neu3* is expressed in a broad lateral domain at cellularization. (D) Expression of LacZ supported by the identified *Neu3 cis*-regulatory element is depicted. A schematic of the binding sites found in the putative *cis*-regulatory element is shown at the bottom (E). Scale bars represent 100 bp.

In vivo binding data for Dorsal, Twist, and Snail transcription factors has facilitated the prediction of hundreds of *cis*-regulatory regions based on genome-wide occupancy of these factors (Zeitlinger et al., 2007b; A. Ozdemir and A. Stathopoulos, unpublished observation). Clusters of high-affinity Dorsal binding sites were not identified within the genomic sequences defined by the ChIP-chip analyses near any of the genes in question. However, we scanned the DNA sequences which were bound by the transcription factors in the ChIP studies and found several candidate regions that contained Dorsal as well as Zelda binding sites in proximity to the genes *SoxNeuro (SoxN), pyramus*, and *Neu3*. Although we tested four putative *cis*-regulatory elements, we validated only one regulatory region.

A ~2 kB fragment of genomic sequence from within an intron of the *Neu3* gene supports expression of a reporter gene in a domain similar to that of *Neu3* mRNA expression (Fig. 5D, compare with 5C). This regulatory region contains two weak Dorsal binding sites, a STAT site, three Zelda sites, as well as several bHLH binding sites (Fig. 5E). A comparison of putative homologous regulatory regions revealed little conservation of sequence (see Supplemental Fig. 5 for sequences and alignment). Dorsal sites are present in many of the homologous sequences from other *Drosophilids,* though the relative positions of these sites have changed, and their PWM scores were poor. One Zelda site appears to be conserved, both in sequence and position, but the distance between this site and the nearest Dorsal site is 198 bp, which is further than the distance in the replacement experiment (Fig. 3 K, L) suggestion that they are not able to function to generate broad lateral expression in this regulatory element. The STAT sites are also present in some of the other *Drosophilid* species, but their location is variable. This analysis demonstrates that multiple, high-affinity Dorsal binding sites are not required to support expression in a broad lateral domain within the early embryo, but instead provides further evidence suggesting that additional activators are functioning to drive expression in this domain.

Discussion

Even limited sequence conservation within *cis*-regulatory modules can provide insights into the underlying regulatory logic

Through a comparative analysis of orthologous *sog cis*-regulatory modules from twelve *Drosophilid* species, we identify core regulatory elements conserved in these sequences. Considerable binding site turnover has occurred during the approximately 40 million years of evolution, yet some sequences are conserved (see Fig. 2). This observation supported the hypotheses we investigated in this work which are, 1) that conserved sequences are functionally required and, 2) that variable architectures might generate the same or similar patterns of expression. Surprisingly, despite the opportunity for binding site turnover during the course of evolution, the *sog* regulatory regions from *D. virilis* can still be interpreted faithfully when used to drive reporter expression in *D. melanogaster.* We conclude from these experiments, despite flexibility in the *cis*-regulatory element structure, regulatory logic has been conserved during evolution of the *cis*-regulatory module sequences to support *sog* expression.

Though this comparative analysis identified limited sequence homology, we allowed what sequence conservation was present to guide our efforts to examine the core regulatory elements required for patterning the neurogenic ectoderm. Using site-directed mutagenesis to eliminate sites within the *sog cis*-regulatory sequence, we obtain results which suggest that Dorsal functions together with the ubiquitous activator Zelda to control *sog* expression within the neurogenic ectoderm (Fig. 3). Furthermore, we constructed synthetic *cis*-regulatory elements, consisting of Dorsal and Zelda or Dorsal and D-STAT sites, which are both able to support expression in the broad lateral domain of *Drosophila* early embryo (Fig. 4 and Fig. 5). From these results we conclude that broad lateral expression is achieved by a combination of Dorsal sites and sites for the ubiquitous activator Zelda, which suggests that a more general mechanism to create broad

expression may involve interactions between Dorsal and other broadly expressed transcription factors.

Dorsal functions with distinct transcriptional activators to support expression along the dorsal-ventral axis

Our mutagenesis and mutant analysis results demonstrate that Dorsal and Zelda support expression of *sog* along the dorsal-ventral axis (Figure 3 and data not shown). In the absence of Dorsal protein, expression of *sog* is gone; however when Dorsal binding sites were mutagenized, weak ventral-lateral reporter expression remains that could be due to unknown Dorsal binding sites that were not detected by our PWM searches or due to input from another transcription factor (Fig. 3D). In the absence of Zelda binding sites or in Zelda mutants, expression is slightly broader than when Dorsal sites are eliminated (Fig. 3F, compare with 3D; and Liang et al., 2008). This residual expression could be due to Dorsal and/or other transcription factor (e.g. bHLH) functioning to direct expression, in a Zelda-independent manner, within the ventral-neurogenic ectoderm; however, our data suggests that Twist is not likely involved, as the domain of *sog* expression along the dorsal-ventral axis is not severely affected in *twist* mutants (data not shown).

Previous genetic studies have demonstrated that Dorsal is required for specification of the presumptive neurogenic ectoderm, but binding sites for Dorsal alone are not sufficient to generate expression within the broad lateral domain of embryos. Dorsal has been shown to function synergistically with Twist to pattern the presumptive mesoderm and ventral neurogenic ectoderm (Jiang and Levine, 1993). Here, we present evidence that Dorsal and Zelda function synergistically to regulate expression that is able to encompass the entire presumptive neurogenic ectoderm domain. Some method of cooperativity likely exists between Dorsal and Zelda, at the

level of DNA binding or downstream, and is responsible for extending the expression domain into dorsal-lateral regions of the embryos, where the levels of nuclear Dorsal are low.

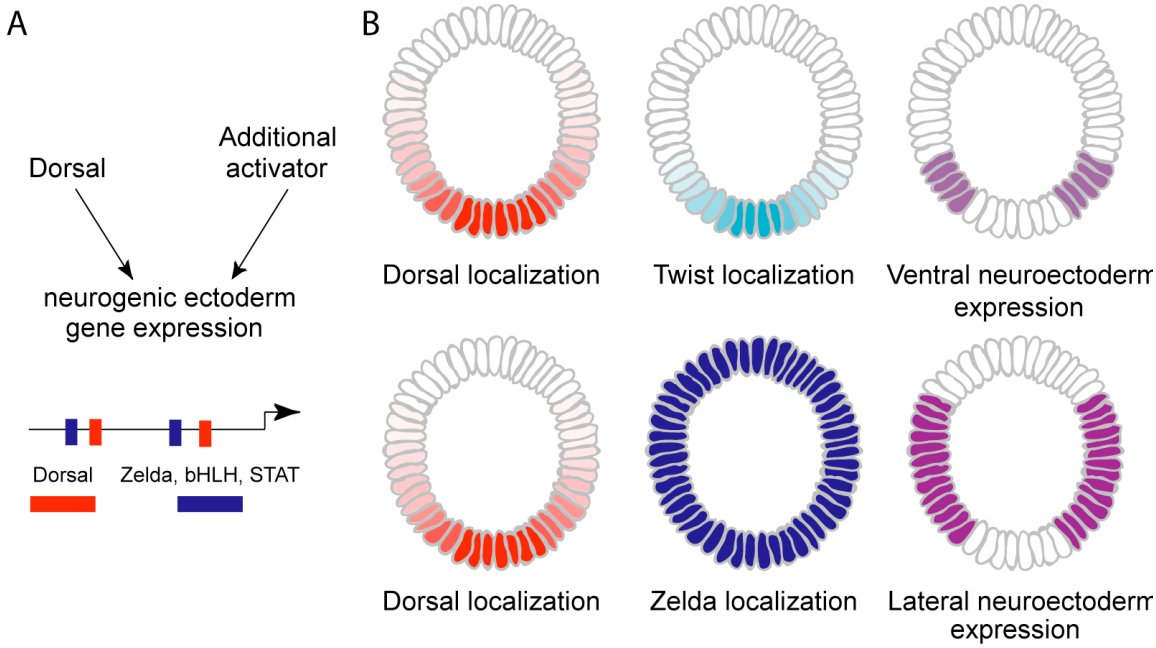


Fig. 6. Model of transcription factor participation in patterning the presumptive neurogenic ectoderm. (A) We propose that Dorsal and Zelda both activate expression in the presumptive neurogenic ectoderm. Zelda functions to initiate ubiquitous expression while Dorsal functions primarily as a regulator of spatial expression. Neither alone is sufficient to support expression of genes like *sog* or *ths* during cellularization. We suggest that other ubiquitous or broadly expressed activators may function with Dorsal in a general manner to regulate expression within different domains of the presumptive neurogenic ectoderm of *Drosophila* early embryo. (B) Schematic of Dorsal and Twist functioning together to generate expression in the ventral neurogenic ectoderm. We propose that Dorsal and Zelda function in an analogous manner to generate broad lateral expression.

We propose that Dorsal functions as a spatial regulator in the neurogenic ectoderm and that additional transcription factors like Zelda, act as co-activators to regulate the precise onset of expression (see Fig. 6A). Furthermore, we suggest that multiple ubiquitous or broadly expressed activators may function with Dorsal to support expression in a broad lateral domain (e.g. Zelda, STAT, and bHLH transcription factors such as Daughterless (Da), see Fig. 6A). We have demonstrated that STAT binding sites can also function together with Dorsal to drive expression in a broad lateral domain. Further support for this idea includes the observation that *sog* as well

as *ths* exhibit broad expression early (see Fig. 2A and Supplemental Fig. 4A). Sites for Zelda are also present in the *ths cis*-regulatory module, and these sites likely direct the almost-ubiquitous early expression of *ths* observed. Interaction of Dorsal with distinct co-activators may not only regulate the spatial domain of expression supported, but also the temporal output. Zelda along with Dorsal or a Dorsal target initiates the earliest zygotic expression detected; perhaps interactions between Dorsal and other activators facilitate expression within a broad lateral domain (or other defined pattern) at later time-points. We assert that gene expression is achieved at the intersection of the Dorsal nuclear gradient and the additional activator which could either be ubiquitous in the case of Zelda or localized in the case of Twist (Fig. 6B)

Flexibility in organization and composition of binding sites can complicate identification of co-expressed genes

Even equipped with this new knowledge, other *cis*-regulatory modules that support co-expression of genes *SoxN, pyramus* and *Neu3* have proven difficult to identify. To date, *SoxN* and *pyramus* regulatory elements remain unidentified. Flexible regulatory structures could account for some of the obscurity that has been encountered in the identification of *cis*-regulatory modules that support expression of genes within *Drosophila* early embryos. Flexibility in binding site composition, orientation and number of sites has also been demonstrated in the regulation of co-expressed genes in *Ciona* by extensive co-expression analyses (Brown et al., 2007). Possibly the observed flux in binding site composition and arrangement provides a mechanism that facilitates the introduction of mutations, which may be selected when a fitness advantage is provided to the developing embryo.

Recently, a second regulatory element for *sog* located upstream of the gene was identified which also drives expression in a broad lateral stripe in the presumptive neurogenic ectoderm of cellularized embryos (Hong et al., 2008; A.Ozdemir and A. Stathopoulos, unpubublished

observations). This novel regulatory element as well as the known regulatory element, the intronic enhancer examined in this study, probably function together to control the full expression pattern of *sog* in the developing embryo. While both *cis*-regulatory sequences contain Dorsal and Zelda binding sites, the novel enhancer contains many more bHLH sites (L. Liberman, unpublished observations), which is in stark contrast to the intronic *sog* regulatory element, which contains only one bHLH site and exhibits very little change of expression in *twist* mutant embryos. This new regulatory element presents further evidence that there exist multiple solutions for the developmental problem of producing spatially and temporally regulated expression. Future experiments will address whether these early embryonic enhancers controlling the expression of the *sog* gene within similar domains use the same mechanism (i.e. Dorsal + Zelda cooperativity) to support expression in a broad lateral stripe or whether different mechanisms are used.

Conclusion and implications for vertebrate biology

Evolutionary comparisons of sequences from diverged species can be very useful for the dissection of underlying *cis*-regulatory logic, as we have shown here; yet the important variable is that the proper comparisons of sequences must be made (i.e. species of appropriate evolutionary distance) and this is not always easy to define. In vertebrate systems, analyses of *cis*-regulatory modules usually focus on modules identified by methods that select for high degrees of conservation, which inherently have a low amount of flexibility. Recently, the identification of ultra-conserved regions, defined as greater than 200 base-pairs of conservation within non-coding DNA sequence, was used as a criterion to identify *cis*-regulatory modules in the mouse (e.g. Visel et al., 2008). Arguments have been made that deciphering the underlying regulatory logic from evolutionary comparisons of sequences, when conservation is too high, is hard to interpret.

However, we contend that the relevant comparisons are context-dependent. In our analysis of the *sog* and *Neu3* *cis*-regulatory modules, we found only limited sequence conservation was identified in comparisons of homologous sequences isolated from *D. melanogaster* and other *Drosophilids*. In the case of the *sog* early embryonic regulatory element, we analyzed in this study, 71 (of 395) base-pairs of non-contiguous sequence exhibits conservation. The degree of conservation that was retained however was useful for dissecting the underlying regulatory logic.

Identifying regulatory regions with flexible structure is more challenging than scanning for a stringent set of binding sites, but it may also reveal alternative mechanisms for specification that were not previously considered. Our prediction is that studies that dissect the flexibility of *cis*-regulatory modules may one day provide insights to facilitate dissection of vertebrate regulatory elements in general, including ones that exhibit flexibility of sequence. It seems plausible that stringently conserved regulatory elements control gene expression of certain classes of genes, like those required for certain essential processes. Flexible regulatory architectures may provide a mechanism for generating variability throughout evolution. Ultimately it will prove useful to make evolutionary comparisons with both highly conserved sites and flexible architectures to determine how each contributes to establishment or maintenance of gene regulation.

Chapter 3

Dorsal-ventral Positional Information Does Not Simply Reflect NFκB/Dorsal Nuclear Concentration

Abstract

The NF-κB related transcription factor, Dorsal, forms a nuclear concentration gradient in the early *Drosophila* embryo patterning the dorsal-ventral axis to specify mesoderm, neurogenic ectoderm and dorsal ectoderm cell fates. These patterning events are thought to be determined in a Dorsal concentration-dependent manner; however, the actual levels of nuclear Dorsal have not been quantified. Furthermore, existing models for Dorsal-dependent germ layer specification and patterning consider steady-state levels of Dorsal relative to target gene expression patterns, yet Dorsal gradient formation is dynamic as is gene expression. We devised a quantitative imaging method to characterize the dynamics of Dorsal nuclear gradient formation while simultaneously examining Dorsal target gene expression in nuclei along the dorsal-ventral axis. Unlike what has been observed in other insects such as *Tribolium*, we find that the Dorsal gradient maintains a constant bell-shaped distribution during embryogenesis. We also determined the relationship between levels of nuclear Dorsal and spatial domains of target gene expression. We find that some genes that require Dorsal for activation fall outside the graded localization of Dorsal, raising the question whether these genes are direct Dorsal targets. Additionally, we show that Dorsal levels change in time during embryogenesis such that steady state is not reached even at cellularization. These results suggest that the multiple gene expression outputs observed along the dorsal-ventral axis do not simply reflect the steady-state Dorsal nuclear gradient. Instead we propose that the Dorsal gradient supplies positional information throughout nuclear cycles 10-14 that compensatory combinatorial interactions between Dorsal and other factors effect differential gene expression along the dorsal-ventral axis.

Introduction

Over three centuries of scientists have discussed the mechanisms by which cell fate specification and patterning occur (rev. in ref Sander, 1996). The morphogen gradient model persists as a concept for describing how positional information is conferred to a field of cells enabling the specification of different cell types. In this model, a diffusible molecule forms a concentration gradient. Cells interpret their position by reading the morphogen concentration they experience and respond with differential gene expression outputs along the gradient. This concept is appealing in its simplicity, and has been used to explain cell fate specification and the patterning of these cells in animals (Ashe and Briscoe, 2006; Tickle, 1999; Wolpert, 1968).

The nuclear concentration of the NF-κB homolog, Dorsal, is present in a concentration gradient within the *Drosophila melanogaster* embryo. It is the only known maternal factor to provide positional information along the dorsal-ventral axis. Both the maternally-deposited *dorsal* mRNA and the translated protein are initially uniform within the early embryo. However, nuclear import of Dorsal selectively occurs in ventral regions of the embryo, resulting in a nuclear concentration gradient that is first visible around nuclear cycle 10, when the nuclei migrate to the periphery of the embryo (Moussian and Roth, 2005). Using transgenic flies with a Dorsal-GFP fusion protein, it has been observed that Dorsal shuttles constantly between the nucleus and the cytoplasm of pre-cellularized embryos. This shuttling is dynamic and occurs in all the nuclei, including those located in dorsal regions (DeLotto et al., 2007).

The asymmetries that result in the Dorsal gradient are initialized in the egg before

fertilization. During stage 10 of oogenesis, EGFR signaling through Gurken establishes a dorsal-ventral asymmetry in the follicular epithelium, a layer of somatic cells that surrounds the developing oocyte (Schupbach, 1987) (Sen et al., 1998). After fertilization, this dorsal-ventral information is relayed to the embryo in the ventrally-localized maturation of the Toll ligand, Spätzle (Nilson and Schupbach, 1998). Consequently, Toll activation directs the degradation of the IκB homolog, Cactus, allowing Dorsal to enter the nucleus (Whalen and Steward, 1993). Dorsal nuclear localization recurs five times, during interphase of each nuclear cycle from nuclear cycle 10 - 14 (DeLotto et al., 2007).

Dorsal is required for patterning the germ layers along the dorsal-ventral axis and controls gene expression in what is thought to be a concentration-dependent manner (Jiang et al., 1992; Pan and Courey, 1992). Dorsal has been shown to function as both an activator and a repressor of transcription (Huang et al., 1993; Jiang et al., 1992). In ventral regions of the embryo, where Dorsal concentration is high, Dorsal positively regulates the expression of the genes *twist* and *snail*, which encode a bHLH transcription factor and a Zinc-finger transcriptional repressor, respectively. Together, Twist and Snail specify the presumptive mesoderm. Lower levels of Dorsal in lateral regions of the embryo activate the expression of genes in the presumptive neurogenic ectoderm including *rhomboid* (*rho*) and *short gastrulation* (*sog*). In contrast, Dorsal functions as a repressor of presumptive dorsal ectoderm genes, *zerknüllt* (*zen*) and *decapentaplegic* (*dpp*), restricting their expression to dorsal regions where Dorsal protein levels are at their lowest.

An affinity model provides a simple, mechanistic explanation for the concentration-dependent readouts of the Dorsal nuclear gradient. This model dictates that Dorsal binds to regulatory regions of target genes with differential affinity. Target genes expressed in ventral regions, where Dorsal is in high abundance, are activated by a suite of low-affinity Dorsal binding sites. As the nuclear concentration decreases in ventral-lateral, lateral and dorsal regions of the embryo, the expression of target genes within these regions requires progressively higher affinity Dorsal binding sites in their regulatory regions (Jiang and Levine, 1993). Thus, the affinity model generally predicts that the regulatory regions of genes will contain Dorsal binding sites with affinities appropriate to their expression along dorsal-ventral axis. While further studies have been consistent with an affinity model for Dorsal, the actual binding affinity for its targets has not been tested *in vivo* (Ip et al., 1991; Jiang and Levine, 1993; Papatsenko and Levine, 2005b; Zeitlinger et al., 2007a).

The requirement of the Dorsal gradient for patterning the dorsal-ventral axis has received much attention (Roth, 2003), though relatively few groups have attempted to quantify the levels of nuclear Dorsal (DeLotto et al., 2007; Zinzen et al., 2006). Recently, a thermodynamic model was proposed based on fractional occupancy of Dorsal, Twist, and Snail on binding sites of regulatory element to control the expression of target genes in the ventral neurogenic ectoderm (Zinzen et al., 2006). However, this model uses data from cellularized embryos and does not consider the dynamics of gradient formation in the pre-cellularized embryo. Furthermore, the whole mount quantification profiles used for the model parameters do not distinguish functional Dorsal located in the nuclei from cytoplasmically-retained Dorsal, which is unable to support transcription.

Although this thermodynamic model, which requires synergistic interactions between Dorsal and Twist, provides a mechanism for patterning the ventral neurogenic ectoderm [i.e. genes *rho* and *ventral neuroblasts defective (vnd)*], Twist is expressed in ventral regions of the embryo and is not likely involved in patterning genes expressed in dorsal-lateral regions of the neurogenic ectoderm [i.e. genes *sog* and *intermediate neuroblasts defective (ind)*]. Instead, we propose that quantification of nuclear Dorsal levels throughout the entire embryo must be measured to determine whether different Dorsal levels direct distinct gene expression outputs.

Our aim was to determine whether the levels of Dorsal within a nucleus correlate with the gene expression. Here we develop methodology to measure the nuclear Dorsal levels in the embryo during nuclear cycles 10 - 14, prior to stable gradient formation at nuclear cycle 14. Previous studies using a Dorsal-GFP fusion revealed dynamic, regulated, nucleo-cytoplasmic shuttling (DeLotto et al., 2007), which suggests a functional role for Dorsal during this process. We chose to examine fixed embryos with wild type nuclear Dorsal. This approach has two advantages over live-imaging. First, we can observe both Dorsal protein level and gene expression in the same embryo. Second, we can image many more individuals to observe any variability that may exist at a given developmental stage. We used wild type and mutant embryos with genetically manipulated levels of nuclear Dorsal and asked if the amount of Dorsal protein could be used as a predictor of gene expression output. The results of our analysis show that Dorsal levels vary during interphase of each nuclear cycle, no steady state gradient is observed even at cellularization. These results suggest that a static reading of Dorsal

concentration alone cannot predict domains of target gene expression. We conclude from our data that steady state Dorsal concentration levels do not directly determine gene expression boundaries, as predicted by morphogen gradient models, and instead our data support the view that combinatorial interactions with other factors are necessary to account for patterning the dorsal-ventral axis.

Results

The Dorsal nuclear gradient supplies positional information for the dorsal-ventral axis in developing *Drosophila* embryos, yet the levels of nuclear Dorsal have not been defined. We measured the Dorsal nuclear concentration in order to determine how much Dorsal was in the nucleus during these cell-fate specification events. To this end, we performed antibody staining to view Dorsal and Histone proteins, while gene expression was observed by *in situ* hybridization. This approach allowed us to quantify nuclear Dorsal concentrations across the embryo and to compare these levels with the expression patterns of select target genes in the neurogenic ectoderm (see Supporting Information). This technique is useful for viewing protein and mRNA in the same embryo as can be seen in manual cross sections (Fig. 1A-B).

We collected three-dimensional (3D) stacks of confocal microscope images of embryos at nuclear cycles 10 - 14, (Fig. 1C). We computationally unrolled images to produce a two-dimensional (2D) picture of the a 3D embryo (Fig. 1D, see Methods and Luengo Hendriks et al., 2006). At the nuclear cycles we examined, all the nuclei have migrated to the periphery of the embryo. Thus all relevant information (i.e. histone

levels, Dorsal concentration, and gene expression data) is retained within these 2D representations (Fig. 1, compare D' with C', D" with C", and D"' with C"', respectively).

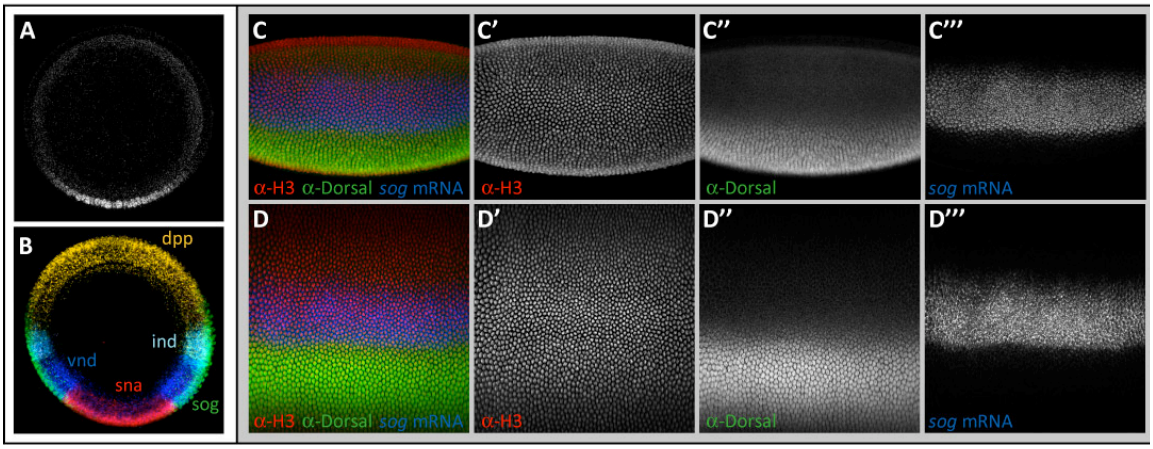


Figure 1. Cross-sections and whole mount *in situ* hybridizations and antibody staining. (A) Dorsal antibody staining visualized by manual cross-section. (B) mRNA *in situ* hybridization of gene expressed along the dorsal-ventral axis. (C) 3D whole mount *in situ* hybridization of *sog* gene expression, shown in blue, co-labeled with antibodies for Dorsal protein in green and Histone (H3) in red. (D) Computational unrolling of 3D images of whole mount embryos allows for protein and mRNA expression to be analyzed in 2D.

We find that the distribution of nuclear Dorsal at all stages is approximately bell-shaped, and consequently can be empirically fit to a Gaussian-like curve (see Fig. 2A-C, Methods and Supporting Information). In ventral-lateral regions of the embryo, where *vnd* expression and the ventral portion of *sog* expression are observed (Fig. 2A,B), the nuclear localization of Dorsal decreases sharply, consistent with initial Dorsal nuclear localization studies (Roth et al., 1989; Rushlow et al., 1989). However, in lateral regions of the embryo, where *ind* and the dorsal portion of *sog* are expressed, nuclear Dorsal protein levels decrease to the same basal levels observed in dorsal regions of the embryo (Figure 2B,C). In particular, it appears the entirety of *ind* expression is almost always seen in the regions where Dorsal is at basal levels, outside of the graded distribution of

Dorsal (Figure 2C). We find that nuclear Dorsal reaches basal levels reliably at ~100 μm from the ventral midline (Figure 2C,D).

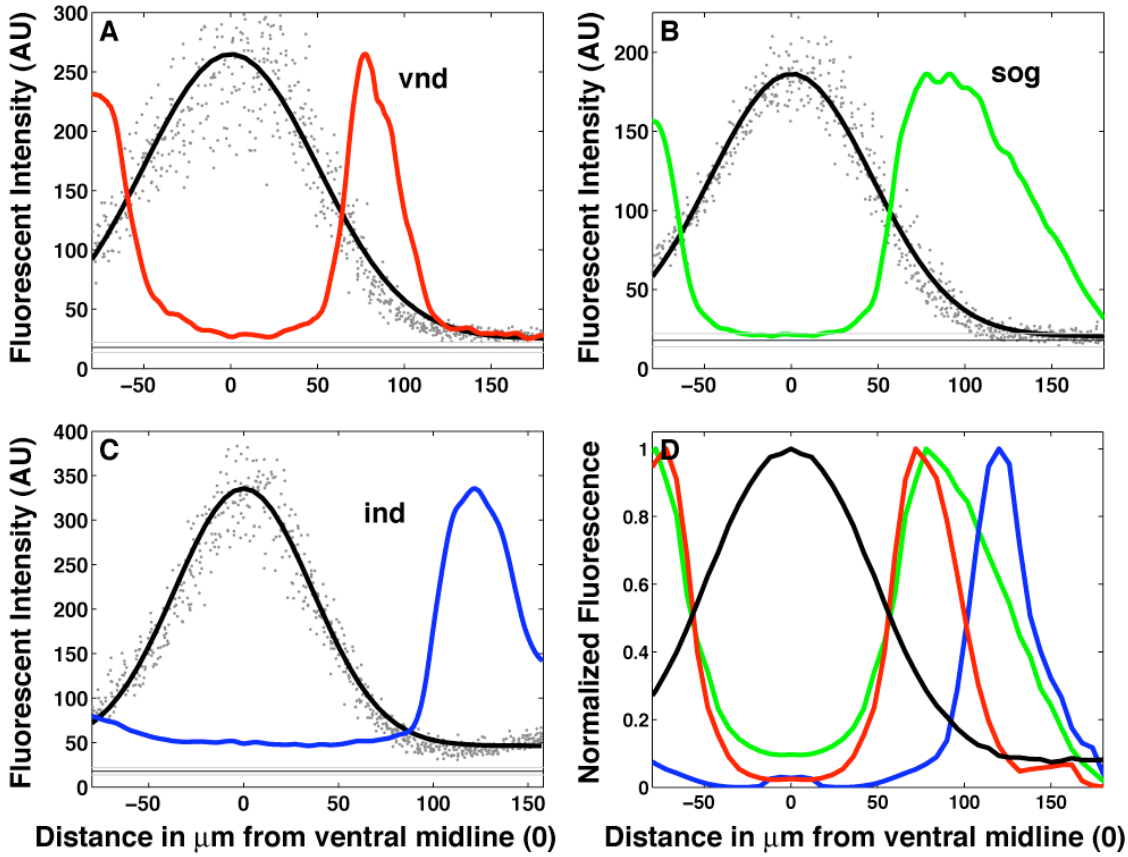


Figure 2: Dorsal quantification and target mRNA expression in wild type embryos shows the location of germ layer-specific target gene expression at nuclear cycle 14. Fluorescent intensity values of Dorsal within the nuclei are shown as grey dots. Colored lines represent the location where gene expression is detected. Numbers on the x-axis represent distance from the ventral midline (located at 0). *dl1/dl1* mutants were used to determine the background fluorescence in the absence of Dorsal protein (black line with standard deviation shown as grey lines). (A) *vnd* expression (red trace) starts within the steepest part of the Dorsal gradient and ends at the Dorsal border of Dorsal nuclear localization. (B) *sog* expression (green trace) spans from the ventral region of *vnd* expression to lateral regions of the embryo where Dorsal levels are uniform. (C) *ind* expression (blue trace) lies largely outside the Dorsal gradient. (D) Overlay of all three gene expression outputs onto a single plot with Dorsal localization in black. Note that the amplitudes of Dorsal concentration vary among the embryos shown. This variability is seen even when the embryos are all at the same nuclear cycle (see Figure 3).

It is important to note that these basal levels correspond to a non-zero concentration of nuclear Dorsal. The Dorsal antibody has some low level of non-specific

background staining, assayed by imaging embryos derived from homozygous dl^1/dl^1 mothers, which produce no Dorsal protein (Roth et al., 1989). However, nuclear Dorsal levels detected in all embryos exceed this background staining, even in the dorsal-most regions of the embryo (Figure 2). For the remainder of the paper, "basal levels" of Dorsal refer to the non-zero levels of nuclear Dorsal achieved in the dorsal portion of the embryo, and all subsequent gradients are plotted with this background subtracted.

Considering these observations, we asked whether the Dorsal gradient is initially broad at earlier nuclear cycles, only to refine later into the steep gradient observed at nuclear cycle 14. This could explain how Dorsal-dependent genes such as *ind* and *sog* could be expressed in lateral nuclei, where Dorsal signaling during nuclear cycle 14 is at basal levels, and thus cannot supply additional positional information to regulate the expression of target genes. However, when the Dorsal gradients from each of the embryos are normalized and plotted on the same graph, the widths of these gradients are constant throughout all nuclear cycles (Figure 3C). To quantify this observation, we used the empirically fit Gaussian parameters, finding the variation in gradient widths to be 14% (standard deviation divided by the mean), which can easily be attributed to natural variation. As a comparison, the natural variation in embryo sizes used in this study was also 14%. Furthermore, when grouped by nuclear cycle, the gradient widths are not significantly different from one another by ANOVA (Figure 3D).

In contrast to the constancy of gradient widths, we find the levels of nuclear Dorsal are surprisingly variable during all of the nuclear cycles, but most dynamic during nuclear cycle 14 (red traces in Figure 3A). We attribute this variance to the dynamic process of Dorsal gradient formation during each nuclear division cycle because technical

noise would not produce the same types of trends we observe. During mitosis, the nuclei break down, forcing Dorsal and other nuclear factors such as Bicoid into the cytoplasm (Roth et al., 1989). We propose that, following each nuclear division, Dorsal begins to accumulate in the nuclei, and as interphase proceeds, the levels of nuclear Dorsal monotonically increase. Therefore, the changes in levels of nuclear Dorsal that we measure can be explained by a random sampling of slightly different time points during each interphase. Our data reflects the fact that we are observing differing stages of a dynamic process. This is consistent with previous work showing that Dorsal protein localization during gradient formation is dynamic, but tends to increase during a single nuclear cycle (DeLotto et al., 2007).

We identified two novel trends in these data. Dorsal levels in the ventral-most nuclei increase during nuclear cycles 10 - 14 (Figure 3B and inset in Figure 3A). However, basal levels of Dorsal in lateral and dorsal nuclei, outside of the graded localization of Dorsal, decrease over this same period. As can be seen from the box-and-whisker plot (Figure 3B), there is a wide range in peak nuclear Dorsal concentration at each nuclear cycle, but the largest biological variability is seen in nuclear cycle 14. Nuclear cycle 14 is the longest in duration, as each progressive nuclear cycle takes more time than the previous one, ranging from approximately 8-50 minutes (Foe and Alberts, 1983). Considering that peak levels of nuclear Dorsal found within nuclei increase in time, we wondered how Dorsal could supply the positional information necessary to pattern the dorsal-ventral axis.

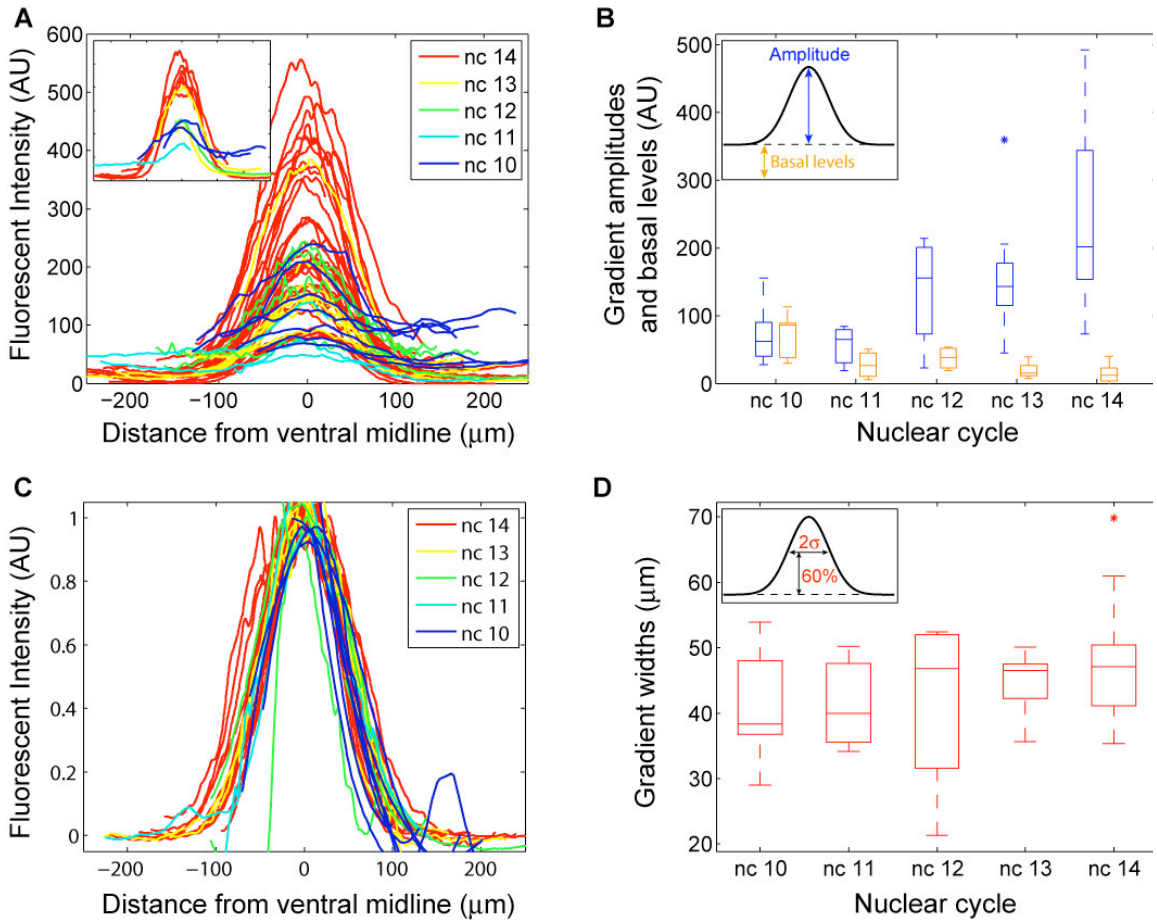


Figure 3. Developmental time course of Dorsal gradient shows no change in the width of the gradient across the ventral midline. (A) Whole mount quantification of Dorsal levels in computationally staged embryos from nuclear cycles 10-14 color coded by stage. The top 15% of Dorsal nuclear levels at each nuclear cycle is shown in the inset. (B) Box-and-whisker plot of Dorsal levels in ventral-most nuclei corresponding to the peak amplitude (blue). Basal levels represent Dorsal levels in lateral and dorsal regions of the embryo outside of the graded distribution of Dorsal at each nuclear cycle (yellow). Median intensity is shown as a horizontal bar in the box. Whiskers show the distribution of intensity values at each nuclear stage. Inset shows cartoon of Amplitude and Basal portion of signal. (C) When the peaks of each of the curves in (A) are normalized to 1, all curves fall along the same Gaussian curve with minor variation in width of the curve. (D) Box-and-whisker plots show Dorsal nuclear gradient widths remain constant throughout embryogenesis.

We quantified Dorsal in embryos with mutations in the Toll receptor and Dorsal itself to determine whether positional information could be seen in the absence of a Dorsal nuclear gradient. In dl^{1}/dl^{1} homozygous *null* mutant embryos, no Dorsal-dependent gene expression is observed. In *Tollrm9/Tollrm10* mutants, Toll signaling is

disrupted, resulting in what is thought to be uniformly low levels of Dorsal throughout the embryo based on the target gene outputs observed (Cowden and Levine, 2003; Jiang et al., 1992). In these mutant embryos, *ind* and *vnd* gene expression is observed in stripes along the anterior posterior axis in variable domains (Figure 4C-D), as has been previously noted (Cowden and Levine, 2003). These expression domains were explained by assuming that *vnd* is seen in a broad domain in embryos which have higher amounts of Dorsal than embryos which express *ind* broadly (Cowden and Levine, 2003). However, we find that *vnd* expression is only broadly expressed at early time points (data not shown) and that the majority of cellularized embryos express both genes in non-overlapping expression domains (Figure 4C-D). The expression of *ind* and *vnd* in the same *Tollrm9/Tollrm10* mutants is unexpected because these genes represent what was previously considered to be two Dorsal threshold outputs. These results do not omit a temporal dependence to Dorsal concentration. Still, these results suggest that there is not a simple concentration-dependent readout of Dorsal concentration levels to produce a particular threshold expression pattern.

In *Toll10b* mutants, the Toll receptor is constitutively active throughout the embryo, thus, Dorsal is not sequestered in the cytoplasm by Cactus. Although the levels of Dorsal had never been measured previously, only mesoderm cell fates are expressed which was thought to be the result of uniformly high Dorsal levels. From embryo-to-embryo, Dorsal levels vary over a two-fold range, a range similar to the biological variability observed in *Tollrm9/Tollrm10* mutant embryos (Figure 4). Despite this variance, all of the embryos express *snail* and only mesoderm cell fates are produced.

These embryos serve as another example of robust gene expression in the absence of precise Dorsal concentration.

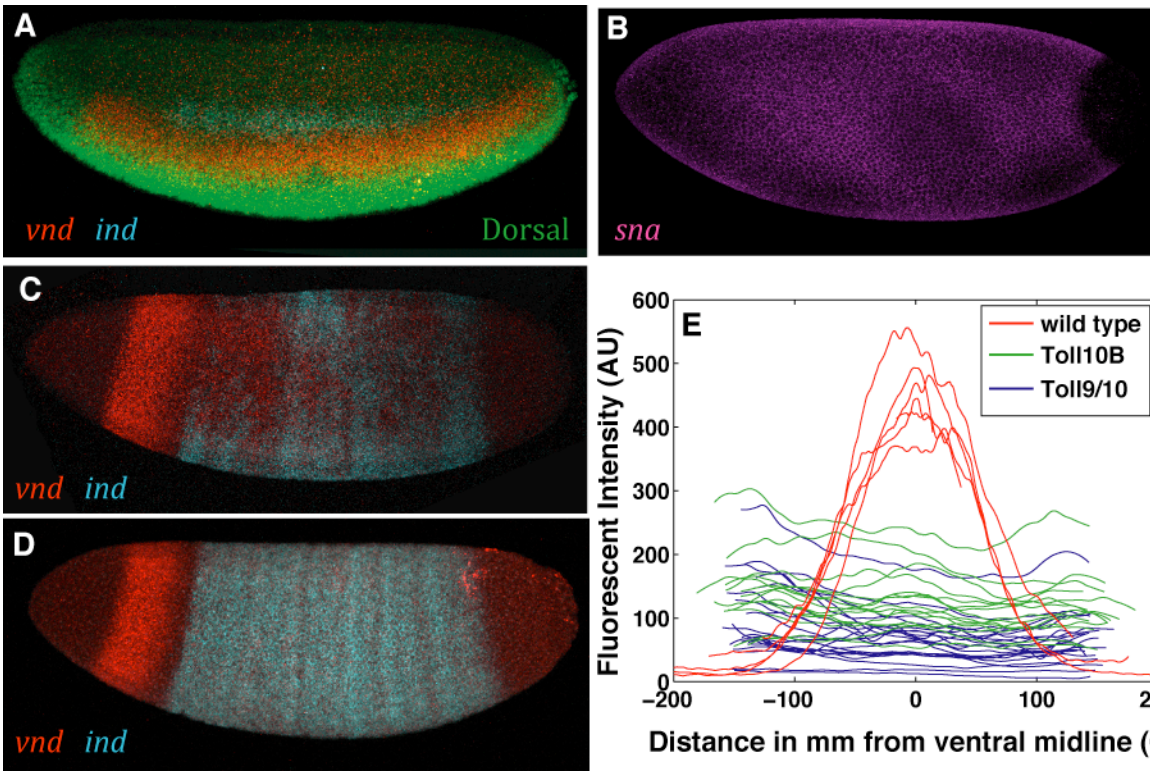


Figure 4. Dorsal nuclear localization in wild type and mutant embryos reveals a wide range of nuclear concentrations. (A) Dorsal nuclear localization in wild type (green) with wild type expression domains of *vnd* (red) and *ind* (blue). (B) *sna* expression in *Toll10b* mutant embryos is ubiquitous except for repression in posterior of the embryo. (C-D) *Tollrm9/Tollrm10* mutant embryos with variable expression of *ind* and *vnd*. (E) Nuclear localization of Dorsal in wild type (red) *Toll10b* (green) and *Tollrm9/Tollrm10* (blue) embryos at nuclear cycle 14.

In order to test the relationship between Dorsal concentration levels and gene expression outputs, we examined embryos with different copies of maternal *dorsal*. In heterozygous Dorsal embryos (dl^1/CyO), a simple model would predict half the amount of Dorsal in the embryo. However, in these embryos, the Dorsal gradient is flat in ventral regions, rather than bell-shaped and steep in ventral-lateral regions (Figure 5). We measured the expression domain of *sog* in these embryos and found its location and width to be indistinguishable from wild type, with *p*-values of 0.6 and 0.3, respectively (Figure

5B and Supporting Information). This invariance of gene expression corresponds to the observation that, in ventral-lateral regions, where gene expression boundaries (such as those between *sna* and *sog*) are delineated, the Dorsal gradient retains a steepness similar to that found in wild type (*p*-value 0.2, see Supporting Information).

We also investigated embryos carrying a copy of the transgenic Dorsal-GFP fusion construct (DeLotto et al., 2007). We found that the Dorsal gradients, retained their Gaussian shape, yet were significantly wider in these embryos than in wild type (*p*-value 0.05), with higher amplitudes (*p*-value 0.003, Figure 5). Additionally, we found that the expression of *sog* was widened dorsally in these embryos, (*p*-value 0.0006, Figure 5B) which is likely the result of a widened gradient.

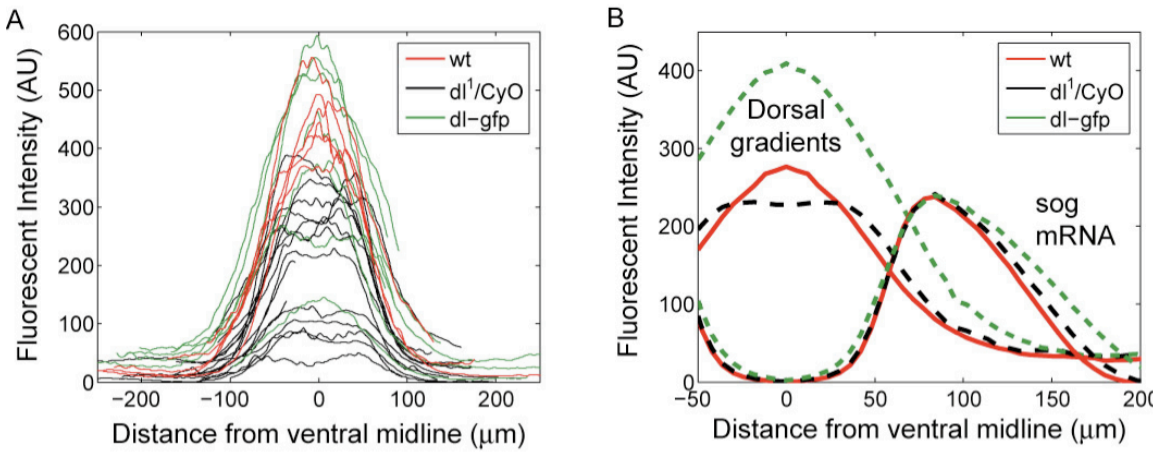


Figure 5. Mutant embryos with genetically manipulated Dorsal still produce gene expression outputs. (A) *dl*^{1}*/CyO* heterozygous embryos have a flattened plateau of Dorsal gradient instead of a peak in ventral regions of the embryo. Lower maximal fluorescent intensity is detected when compared to wild type localization (blue and red gradients). (B) *dl-gfp* embryos contain an additional copy of Dorsal and have significantly wider Dorsal gradients. In these embryos the dorsal border of *sog* expression is dorsally shifted.

Discussion

Dorsal is essential for supplying positional information to pattern the dorsal-ventral axis in what is thought to be a concentration dependent manner, which is why it is widely accepted as a morphogen, although it may be more accurate to refer to it simply as a graded transcription factor. Previous studies used cellularized profiles of whole mount embryos to quantify Dorsal, which does not account for the dynamics of Dorsal nuclear localization. We used whole mount staining and quantitative imaging to analyze the relationship between the amount of nuclear Dorsal and the gene expression outputs that Dorsal regulates. We computationally unrolled embryos to produce 2D images from 3D stacks. This facilitated staging the embryos by nuclear density and quantification of Dorsal fluorescence. We use standard Matlab techniques to segment the nuclei and extract the fluorescent intensities. In this way, we were able to examine Dorsal levels and gene expression outputs in cellularized embryos. Surprisingly, we found that the intermediate presumptive neurogenic ectoderm, where *ind* and the dorsal portion of *sog* are expressed, is consistently beyond the range of graded nuclear Dorsal (Figure 2). While small amounts of Dorsal are present in these nuclei, these levels are also present in the dorsal-most nuclei, and thus cannot supply additional positional information.

This led us to ask how the borders of *ind* and *sog* (and other presumptive neurogenic ectoderm genes) are positioned. Why do these nuclei in the intermediate neurogenic ectoderm express these genes at all? And if Dorsal signaling is responsible for their activation, what spatially localized factor prevents their expression from extending more dorsally? One possibility is that the Dorsal gradient is initially broader and then

narrows during embryogenesis as is seen in the short germ beetle, *Tribolium castaneum* (Chen et al., 2000; Nunes da Fonseca et al., 2008). However, our results dismiss this possibility by showing that there is little to no change in either the Gaussian shape or the extent of the Dorsal gradient during nuclear cycles 10-14, ranging to ~100 μm on either side of the ventral midline (Figure 3).

Alternatively, *ind* and the dorsal portion of *sog* could be indirect Dorsal targets, or directly activated by Dorsal, yet augmented or refined by the combinatorial action of other factors. For example, Dorsal functions synergistically with its target Twist to regulate mesodermal patterning (Ip et al., 1992b; Jiang and Levine, 1993), and in patterning the lateral regions of the neurogenic ectoderm, Dorsal is thought to act in conjunction with ubiquitous activators such as the maternal transcription factor Zelda (Liang et al., 2008) (Liberman and Stathopoulos, 2008). EGFR is an attractive candidate participating in *ind* regulation, as embryos deficient in EGFR signaling lack *ind* expression (Von Ohlen and Doe, 2000). Thus, *ind* and the dorsal portion of *sog* expression could result from EGFR signaling in *lieu* of or in addition to direct Dorsal activation.

It is possible that the basal levels of Dorsal in dorsal-lateral and dorsal regions are sufficient to activate *ind* expression. In this case, repression by a secondary factor, such as *Decapentaplegic* (*Dpp*), would be required to restrict *ind* to the intermediate neurogenic ectoderm. It is known that in *brk, sog* double mutant embryos that would have an hyperactive Dpp signaling, *ind* expression is lost (Carneiro et al., 2006; Mizutani et al., 2006; Von Ohlen and Doe, 2000). Furthermore, embryos derived from mothers carrying

maternally-provided dominantly active Thickveins receptor lack *ind* expression, as well as *vnd* and *sog* expression (Greg Reeves personal communications). However, Dpp cannot be the only dorsally-localized repressor of *ind*, as *dpp* mutant embryos maintain the correct dorsal border of *ind* (Von Ohlen and Doe, 2000). Moreover, it is known that *ind* is dorsally repressed by an unknown A-box factor (Stathopoulos and Levine, 2005b).

Our results show that, although the shape and width of the wild type Dorsal gradient is constant in time at any given dorsal-ventral axis location, the overall levels of nuclear Dorsal vary widely from embryo-to-embryo. We propose that this variability is the result of observing snapshots of a rapid, time-dependent process in which the net nuclear import of Dorsal during interphase causes a monotonic increase in Dorsal levels, followed by a rapid export during mitosis when the nuclear envelope breaks down. This phenomenon was also observed previously in single-nucleus time-lapses using a Dorsal-GFP fusion protein (DeLotto et al., 2007). Despite the rapid dynamics of measured nuclear Dorsal levels, the gene expression boundaries of Dorsal target genes along the dorsal-ventral axis remain surprisingly robust, suggesting that patterning the axis is not strictly dependent upon a steady dose of Dorsal concentration. This opposes the classical morphogen gradient model, in which a series of steady, absolute concentration thresholds determine gene expression boundaries. Alternatively, gene expression may result from "pre-steady state decoding" of the Dorsal gradient, as has been suggested for the Bicoid gradient (Bergmann, 2007). It is also possible that nuclei read or interpret the Dorsal gradient by integrating exposure to Dorsal over multiple nuclear cycles.

The variation in amplitude of Dorsal nuclear levels is not restricted to wild type embryos, but was observed in all embryos studied. This was true in particular for

embryos with supposedly uniform Dorsal levels (from *Toll*^{rm9}/*Toll*^{rm10}, and *Toll*^{10B} mothers). Interestingly enough, in the *Toll*^{rm9}/*Toll*^{rm10} background, both *ind* and *vnd* were frequently seen within the same embryo, yet in spatially distinct locations. Furthermore, *vnd* expression occurs earlier in development, and is later replaced in the trunk by *ind*. We do not observe *ind* in the anterior portion of the embryo even at the latest stages of development. This begs the question, why doesn't *ind* invade the region occupied by *vnd* in wild type embryos? Previous "ventral dominance" models held that *vnd* expression somehow took precedence over *ind* expression (Cowden and Levine, 2003), but our data show that this is not true, at least in this genetic background. One possibility is that, if EGFR were indeed the direct activator of *ind*, then in *Toll*^{rm9}/*Toll*^{rm10} embryos, where *rho* and *vnd* are ubiquitously expressed and EGFR signaling is likely much higher than that seen in wild type embryos, *ind* has enough EGFR signaling to overcome repression by *vnd*.

Another interesting facet of the *Toll*^{rm9}/*Toll*^{rm10} embryos is the anterior/posterior asymmetry. The expression patterns of *ind* and *vnd* observed in *Tollrm9/Tollrm10* embryos suggest that, at least in this genetic background, anterior-posterior determinants modulate expression of *ind, vnd* (Figure 6) and *sog* (data not shown). Previous groups have shown that anterior-posterior factors influence expression along the dorsal-ventral axis and bind to their regulatory regions and these factors could also be functioning here (Li et al., 2008; Mizutani et al., 2006; Zeitlinger et al., 2007b).

We also examined embryos with either one copy (*dl*^{1}/*CyO*) or three copies (*dl-gfp*) of maternally supplied Dorsal. We noted that, in the heterozygous embryos, the overall shape of the Dorsal gradient was not retained; the gradient in the ventral-most regions

was flat rather than smoothly-peaked. Despite this change in shape of the Dorsal nuclear gradient, or perhaps because of it, the gene expression outputs remain virtually unchanged from wild type. It appears that a compensatory mechanism exists to alter the shape of the Dorsal gradient when gene dosage is low: in the region of the embryo where graded Dorsal is presumably important (i.e., in the presumptive neurogenic ectoderm), the distribution of nuclear Dorsal is very similar to wild type (Figure 5). While it isn't immediately clear what form of regulation could be responsible for the redistrubution of nuclear Dorsal, it could be dependent on positive feedback loops through zygotic *cactus* or *Toll* expression. Compensatatory control of gene expression has been shown in previous work which demonstrated that when Dorsal levels are low, expression of Twist along the anterior posterior axis can control the expression of some Dorsal target genes (Stathopoulos and Levine, 2002b).

In contrast, embryos carrying one copy of *dl-gfp* have significantly wider and higher-amplitude gradients, and gene expression in these embryos is shifted dorsally (Figure 5). This is easily explained by the extra, transgenic copy of *dl*. However, the observed phenotype is also likely to be exacerbated by the nature of the Dorsal-GFP fusion protein, which lacks putative export sequences and failed to complement *dl* null mutant in our hands (see Supporting Information).

Our results call into question Dorsal's role as a classical morphogen. Although the expression of some genes is clearly affected by altering Dorsal nuclear concentration, concentration threshold dependence, as described in the classical morphogen model, cannot fully explain our results. Indeed, it is not evident that nuclei determine which genes to express by reading the concentration of any particular factor. Instead, it is likely

that Dorsal signaling is integrated over time as well as augmented by interactions with other transcription factors that function to regulate gene expression along the dorsal-ventral axis. The proposal that Dorsal functions with co-factors throughout embryogenesis and immunity is supported by multiple other studies (Ip et al., 1992b, Liberman, 2008 #43, immunity reference?). Additionally, embryonic patterning has been described in other systems without invoking morphogens to direct the specification of cell fates (Baugh et al., 2005; Bolouri, 2008). Our data support the view that combinatorial interactions between transcription factors at regulatory sites serves as the best explanation of differential gene expression and patterning in developing embryos (Ochoa-Espinosa et al., 2009; Zinzen et al., 2006). Our data suggest that the morphogen hypothesis may not be sufficient for explaining dorsal-ventral axis patterning of *Drosophila* embryos.

Methods

Antibody staining and fluorescent in situ hybridization (FISH)

Dual fluorescent in situ and antibody staining were performed using established methods omitting the Proteinase K procedure (Kosman et al, 2004; http://superfly.ucsd.edu/~davek/). Antisense RNA probes were used to detect *short gastrulation* (*sog*), *intermediate neuroblasts defective* (*ind*), *ventral nervous system defective* (*vnd*), *snail* and *zerknullt* (*zen*) gene expression. anti-Dorsal 7A4 monoclonal antibody generated by Ruth Steward was used to detect Dorsal protein localization (Developmental Studies Hybridoma Bank). Alexa Fluor 488 dye conjugated anti-mouse secondary (Invitrogen A21202) was used to detect primary antibody localization. Anti-

Histone H3 polyclonal rabbit antibody was used to detect histones and served as a nuclear marker (Abcam #ab1791-100). Alexa Fluor 555 dye conjugated anti-rabbit secondary (Invitrogen A31572) was used to detect histone localization. Alexa Fluor 647 dye conjugated anti Sheep secondary (Invitrogen 21448) was used to visualize RNA localization of target gene expression.

Image Acquisition and processing

The LSM 5 Pascal (Zeiss) microscope was used to acquire confocal z-stacks of fixed and labeled embryos. Briefly, confocal stacks were acquired to image through at least 50% of the embryo. For groups of yz-sections, the location of the periphery of the embryo was found computationally. We then used a keystone transformation to computationally "unroll" the embryo's peripheral shell slice by slice. This unrolled shell was then averaged in the proximo-distal direction. (For more information, see Supporting Information). This exchanges a 3D data set for a smaller, more easily manipulated 2D sheet.

Dorsal protein quantification

Dorsal was quantified in embryos in nuclear cycles 10-14. Starting from the 2D sheet representation of the 3D data set, the nuclei were segmented using standard protocols in Matlab (see Supporting information and Figure 6). Up to an additive constant, the Dorsal concentration in each nucleus was calculated to be proportional to the intensity of the Dorsal image in the location of the nucleus normalized by the intensity of the same nucleus in the Histone H3 image (for depth correction):

$$c_{dl,i} \propto I_{dl,i} / I_{hist,i} + k,$$

where $I_{dl,i}$ and $I_{hist,i}$ are the intensities of the i^{th} nucleus in the Dorsal and Histone images, respectively, and k is a constant describing non-specific antibody binding. We estimate the value of k by imaging embryos derived from dl^1 mothers.

The Dorsal nuclear gradients were fit to Gaussian-shaped curves to determine the following global properties of the gradient: amplitude, basal levels, presumptive location of ventral midline, and length scale of decay (width):

$$c_{dl}(x) \approx A \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) + B,$$

where A and B denote the amplitude and basal levels of the fitted Dorsal gradient, respectively, m denotes the location of the presumptive ventral midline, and s is the length scale, or width, of the gradient. For each imaged Dorsal gradient, the values of these parameters were optimized in the least squares sense. Because signal decay was problematic at the edges of the image, only the central 60% of the image (along the anterior-posterior axis) was used in the optimization.

Correction for variations in laser power

During each imaging session, a calibration was performed to measure the current laser power. The microscope mechanical shutters were checked periodically that the laser power (in mW) varies linearly with the percentage laser power slider in the Zeiss LSM 5 Pascal software. In addition, the dependence of Alexa fluor dye fluoresence emission was determined for our system as a function of incident laser power. Using these calibration

baselines, fluoresence intensity of embryos from different imaging sessions could be directly compared.

Fly lines

yw flies were used to quantify the wild type Dorsal gradient. Dorsal mutant heterozygous and homozygous mothers were generated using the dominant temperature sensitive lethal mutation from *dl^{1} cn sca /CyO DTS100* line (Bloomington Stock Center). *Sp/CyO; Tollrm9/TM3* (Gerttula et al., 1988) and *Sp/CyO; Tollrm10/TM3* (Schneider et al., 1991) flies were used to generate *Tollrm9/Tollrm10* females that produced maternally deficient embryos. *Toll10b/TM3* (Erdelyi, 1989) and *Toll10b/OR60* (Stathopoulos and Levine, 2004) flies were used to generate females with a constitutively active Toll receptor.

Acknowledgements

We would like to thank Scott Fraser for helpful discussions regarding imaging strategies.

Chapter 4

Discussion

My goal for this thesis was gain a better understanding as to how the neurogenic ectoderm is patterned in *Drosophila melanogaster*. I wanted to learn about the regulation of specification of this germ layer at the *cis*-regulatory level and at the protein level. In Chapter 2, we examined the regulatory element which controls the expression of the gene *sog*. We found that Dorsal functions in conjunction with a ubiquitous maternal activator, Zelda, to control gene expression in a broad lateral stripe encompassing the presumptive neurogenic ectoderm. We were also able to create synthetic regulatory elements that directed expression of reporter genes in the same location. We achieved this by using Dorsal and Zelda sites as well as sites for another ubiquitous activator. These results led us to conclude that there is a flexible nature to gene regulation. It remains to be seen whether the flexibility that we observed in synthetic elements also functions to regulate genes in the fly.

In Chapter 3 we measured the amount of Dorsal in the nuclei of the early embryo during nuclear cycles 10-14. This is a critical period of development, when the zygote begins to take control of its own gene expression. We learned that the simple goal of determining the amount of Dorsal in the nucleus was actually a complex problem. We found that Dorsal levels within the nuclei vary widely during the same nuclear cycle. Yet, even with this high amount of variability within the nucleus, we noticed an overall increase in the amount of nuclear Dorsal during this

time period. We also established that the shape of the gradient was constant in time, reliably fitting a Gaussian curve. Surprisingly, reducing the amount of Dorsal in the embryo did not effect gene expression despite dramatic change in the shape and amplitude of the gradient. In other mutant embryos, which have only partially functional Toll receptor and thus theoretically low levels of Dorsal, we observe gene expression of both *vnd* and *ind,* which were previously considered two different threshold readouts of nuclear concentration. In sum our data indicate that a simple concentration readout cannot explain the complex process of patterned gene expression along the dorsal-ventral axis.

Append ubiquitous activators to the network

There are numerous regulatory interactions responsible for patterning the dorsal-ventral axis. Understanding the regulatory nature of transcription factors and their targets is a challenging but exciting pursuit. The gene regulatory network wiring-diagram for dorsal-ventral patterning depicts the transcription factors and the regulatory regions they control. This representation is static, representing the hardwired nature of the interactions. Yet, the network itself is constantly responding to the current state of the cell. The ability to control expression of a given gene in a given location at a given time is crucial for normal development. The physical interactions shown in the diagram represent a subset of all the interactions that occur through developmental time.

Dorsal is a clearly an important regulatory protein, often enlisting the aide of additional proteins to control expression. Twist and Dorsal synergy serves as an excellent example of how Dorsal can function with other factors to direct the specification of the mesodermal cell fate. This work implicates the ubiquitous maternal activator, Zelda, in patterning the neurogenic ectoderm with Dorsal. Localized transcription factors serve an essential role in the regulation of spatial patterns in the embryo, but the role of ubiquitous transcription factors have been largely

overlooked in gene regulatory networks.

The maternal contribution of essential proteins and mRNA prepares the oocyte for the complex developmental processes it will undergo. In fact it has long been appreciated that early development is largely dependent upon maternal control:

> …Earlier stages [of sea urchin development], for which according to our results, specific chromosomes are not necessary, demonstrate a purely maternal character. . . . I would like to ascribe to the cytoplasm of the sea urchin egg only the initial and simplest properties responsible for differentiation. . . it provides the most general basic form, the framework within which all specific details are filled in by the nucleus.
>
> -Theodor Boveri 1902

In addition to the spatially restricted maternal products, which have received much attention, there are many ubiquitous maternal factors. Now we must determine what roles these maternal inputs play in gene regulation. One possibility is that these factors provide essential temporal regulation, enabling the global activation of certain genes that require precisely timed activation. These ubiquitous activators may function in sequence throughout early embryogenesis providing the necessary gene products to support the subsequent stage.

Finding regulatory elements amidst flexibility

This is a very exciting time for developmental biology. Although the genome sequence is almost ten years old, there is still much information within the genome that we do not understand. We are not able to faithfully identify regulatory regions using our current bioinformatics methods. Since these regulatory regions are needed to determine the interconnected nature of the gene regulatory network, they must be found empirically.

The most common method for binding site isolation to date is Systematic Evolution of Ligands by Exponential Enrichment (SELEX). SELEX determines binding sites by multiple iterations of binding, selection and isolation of an aptamer that has the highest affinity for the

protein of interest *in vitro*. This method may omit other *bona fide* binding sites for these transcription factors that do not bind as well under these conditions. Unless multiple proteins are added to the binding reaction, SELEX generally fails to isolate protein complexes. If proteins function in a synergistic fashion to regulate gene expression then SELEX may not be optimal. With the decrease in sequencing costs and the simultaneous advent of genome-wide sequencing technologies, we will soon be able to isolate many more regulatory elements. Chromatin IP paired with deep sequencing (ChIP-Seq) has the advantage of identifying the *in vivo* regulatory regions and provides binding site resolution of transcription factor binding throughout the entire genome. This technique should prove extremely useful to find binding sites that are bound at a given time in development. It holds the promise of identifying unknown binding sites that would not have been found *in vitro* or by bioinformatics methods, because it is unbiased to affinity, conservation or clusters of binding sites. ChIP-seq should help us find regulatory regions in a high-throughout manner. Once these elements have been identified we will be able to isolate their functional motifs by methods like those described in Chapter 2.

In an attempt to determine direct, versus indirect Dorsal target genes, I proposed a microarray experiment for my candidacy exam with time-course of gene expression with 10 minute time resolution. This experiment, although time and resource intensive, is worth the investment. Our current data set for Dorsal target genes encompasses a two hour time window of development does not adequately address the rapid development of *Drosophila* embryogenesis (Stathopoulos and Levine, 2002c; Zeitlinger et al., 2007b). The transcriptional profile of the early embryo changes every 10-15 minutes (Nasiadka and Krause, 1999), which means that an accurate understanding of gene expression would necessitate precise staging during each of these time points. This experiment coupled with Dorsal ChIP-seq experiments, which adhere to the same stringent staging, would provide the information to state with reasonable certainty which genes are bound by Dorsal and when they are regulated. This data set would permit the inference

of direct or indirect target genes based on binding and timing of expression. This would serve as a tremendous resource for the lab and the *Drosophila* community as a whole.

As has been shown, conservation of genomic sequences between species can be useful (see Chapter 2 for additional discussion). Even when there is not very much conservation, comparative genomic analysis can serve to highlight the evolutionarily significant components for regulation. Yet, for quickly evolving genes or regulatory regions, conservation may not always yield obvious conserved regions to test. When searching for additional regulatory elements, we were fortunate to find an element for *Neu3* that directed expression in a similar domain to that of endogenous *Neu3* expression. *Neu3* is conserved throughout the 12 *Drosophilid* species, and putative regulatory regions can be identified for the divergent species, yet these regions to do not appear to retain many binding sites (see Chapter 2; Appendix A). We do find sites for Zelda and Dorsal, but these sites are not in close proximity to each other, the closest being almost 200 bp apart, and are not well-conserved. There are several possible explanations why our comparative analysis did not suggest regions to perturb. The first is that the transcription factors themselves have evolved in these other species and no longer bind the sites that are predicted based on *D. melanogaster* analysis. We have reason to believe this is not the case, since Dorsal and Zelda sites within the *sog* regulatory region are faithfully conserved throughout the 12 sequenced *Drosophilids*. However, we cannot dismiss the possibility that, although *sog* and *Neu3* are expressed in the same general domain, different transcription factors regulate their expression. In the case of *Neu3*, further dissection of the regulatory element will determine which sites are necessary and sufficient for expression.

Considering the variable structure of regulatory elements that have been identified: *ths, sog* and *Neu3,* there does not appear to be one simple "regulatory code" which dictates transcription factors and their binding sites that direct the expression of genes in this broad lateral region. It is likely that many genes are controlled by the same transcription factors, but it is also equally likely

that certain transcription factors are replaceable and that the gene regulatory network can compensate if some factors are not present at a particular time or place. This is evident in the fact that less than 2% of zygotically transcribed genes are required for embryonic development (Wieschaus, 1996). This developmental robustness could be explained by a gene regulatory network that allows for multiple solutions to the same fundamental problem of controlling gene expression spatially and temporally. We have synthetic regulatory element evidence ubiquitous activators are essentially interchangeable and this may be true in native regulatory element as well. In other words, the system is flexible and adaptable to perturbations, permitting transcription factor substitutions when necessary.

Synthetic regulatory elements

Considering the variable structure of regulatory elements both native and synthetic which direct similar expression, it leaves me to wonder what inherent organization is necessary for gene regulation? Perhaps many genes are controlled by the same overlapping sets of transcription factors, but the number and spacing of regulatory sites is not as evolutionarily constrained as the presence or absence of sites. Synthetic regulatory elements serve as a powerful tool for learning the limits of flexibility in *cis*-regulation. Once we understand the boundaries of each component we can then use them to direct the expression of genes in the way that we choose. In order to establish the relationship between transcription factors and binding sites, I suggest the simplest of experiments utilizing regulatory sites that have already been identified.

The experimental design is simply to make transgenic flies that express fluorescent reporter genes under the control of synthetic regulatory elements. Synthetic regulatory elements will be made, with 1, 2, 4 and 8 Dorsal binding sites (D) respectively. The same will be done for Twist (T) and Zelda (Z). Starting with Dorsal and Twist, which are known to function synergistically,

synthetic elements will be made with alternating sites for each protein (DTDTDTDT). The expected result would be an expression output that was greater than the sum of the expression of Dorsal and Twist. Considering the synergistic nature of this relationship it would also provide insight to vary the distance between the Dorsal and Twist sites to learn if the proximity of sites is important for them to function cooperatively. Is there a specific distance, in base pairs, at which their cooperativity is maximized, or how far apart must sites be to loose cooperativity? Using this information we could then search the genome for to find how often these sites are found at this maximally cooperative distance.

In Chapter 2 we suggest that Dorsal and Zelda function together, but we could not say whether they function in a cooperative manner. Using synthetic regulatory elements we could test this hypothesis in a similar manner to how I propose the demonstration of cooperativity between Dorsal and Twist above. We already know that Dorsal and Zelda sites create broad lateral expression of reporter genes, but a fluorescent reporter would allow for a quantitative analysis of gene expression in live embryos.

The ultimate fly

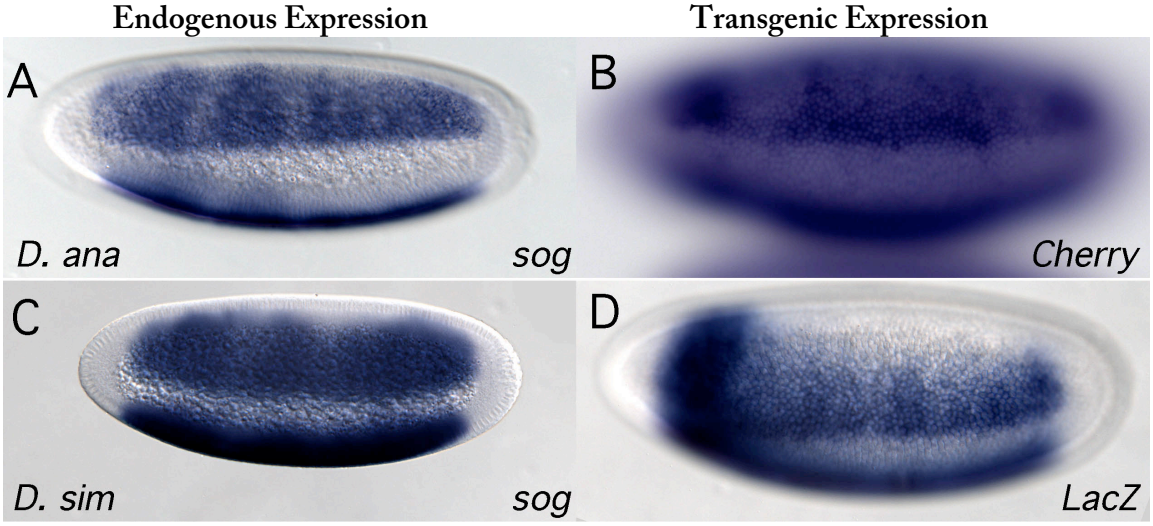
In the not too distant future, we could know the location, initiation and duration of control of transcription factors and the binding site it recognizes for every tissue in the fly. With this information we could construct regulatory DNA fused to a reporter, or gene of interest, to direct expression in any location at any time in the embryo. In addition to this being a really fun way to make a pretty time-lapse flashing fly, we could customize enhancers for spatio-temporal expression at any given time throughout development. This would be useful for studying the effects of mis-expressed genes at different times and locations in development. It could also support progressive medical treatments where tissue specific enhancers are coupled to inducible promoters to regulate the expression of genes for gene therapy purposes.

Appendix A: Supplemental Materials for Chapter 2

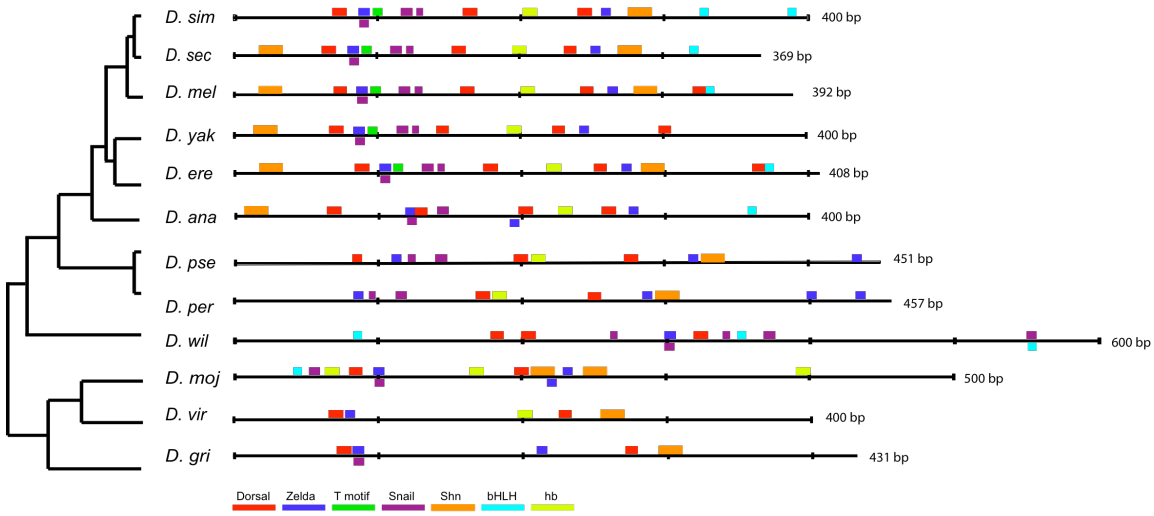
Introduction

The following figures and methods serve as supplementary materials for Chapter 2 of this thesis: Design Flexibility in *cis*-Regulatory Control of Gene Expression: Synthetic and Comparative Evidence. This information should be sufficient to replicate the results we observed. Negative results were included to reduce further efforts to find regulatory elements for Dorsal transcriptional targets in the future.

Supplementary Materials



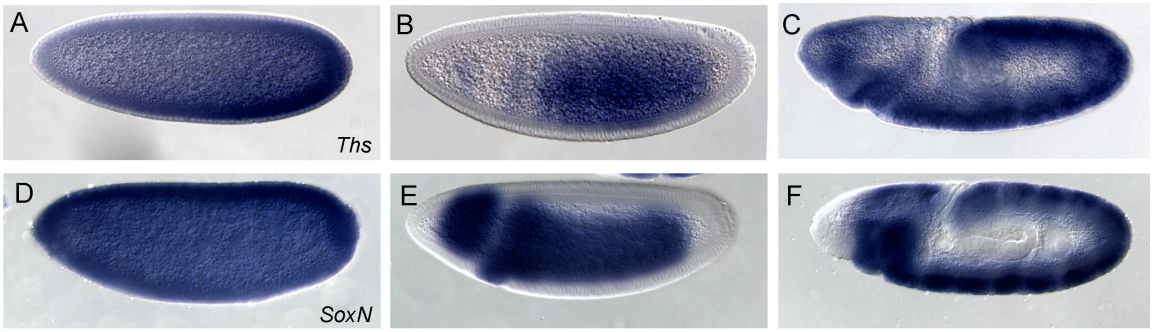
Supplemental Figure 1. Reporter fusions reveal conservation of expression and regulatory logic. Endogenous expression viewed by in situ hybridization of RNA probes in *D. annanassae* and *D. simulans* shows that the broad lateral expression seen in *D. melanogaster* is conserved. Regulatory element reporters show a similar expression pattern as well.



Supplemental Figure 2. Putative *sog* regulatory regions from the twelve sequenced *Drosophilids*. Tree depicts evolutionary relationships of the twelve sequenced genomes. Length of each regulatory region is shown in bp to the right of each cartoon (compare with Figure 2).

```
sog_wil         AACAACAACAATCACAGCCGGTAGATGCTTTTGCCAACCTCGTCACCTGCTCATTTTGGCCATATTTCTTGTTGTTGCATGCTTCAGCCAACCAGTGTTGGGTCGCCGTCATGCTCCACT 120

sog_sim         ----------------------------------------GTTGCCAATGCCATTGCGCATACGCCGTT-CGTCTAT--ATGG-CTAT-----ATGGCTATATGGCCGTGTA---------- 63
sog_sec         ----------------------------------------GTTGCCAATGCCATTGCGCATACGCCGTGTCGTCTAT-----ATGGCTATATGGCCGTGTA---------- 56
sog_mel         ----------------------------------------GTTGCCAATGCCATTGCGCATACGCCGTGTCGTCTAT--ATGG-CTAT-----ATGGCTATATGGC--TGTA---------- 62
sog_yak         ----------------------------------------CCAATGCCATTGCGCATACGCCGTGTCGTCTAT--ATGG-CTAT-----ATGGCTATATGGCGATATA---------- 60
sog_ere         ----------------------------------------GTTTCCAATGCCATTGCGCATACGCCGTGTCGTCTAT--ATGG-CTAA-----ATGGCTATATGGATATATGGATATATG--- 72
sog_pse         ----------------CACATTCCCACAACAGTCGTTGCATAAGTCTCTGCCTCCGCGTGCAGCCGCATTGCGCAT--ACG--CCGT-----GTCGCAAGCGGAAAGCGGG---------- 86
sog_per         ----------------------------------------GTCTCTGCCTCCGCGTGCA-GCCGCATTGCGCAT--ACG--CCGT-----GTCGCAAGCGGAAAGCGGG---------- 59
sog_ana         ----------------------------------------CATTGCGCATACGCCGTGTCGGCCAATGGCGG-CTGCTAGG-GTGGCTGCTAGAGGGTGGGTAGCGGGGAAA 67
sog_moj         ------------------------------AACTCGTAGGCAATCTAACATTGCGCATACGCCCCGTCCGTCAGC--TGG-CCATAAAC-ATGTATGTTTTTTGGTCGG---------- 75
sog_vir         ----------------------------------------TACGCCACGCCCGCCGGC-TGGG-CCATAAAA-GTGGGCCCTGTTTTTC---------------- 43
sog_gri         ------------------------------------TGCAACAGCTAAATTGCACATACGCCCCGATGCTCTCT--CGC-TCGTTTGCCATAAAAGTTGGTTCGCC---------- 67
sog_wil         CATGTTCGAGGAGTCGGATACGGGACGACGTTCCAATAGACCAGCCGGTAAGTGATATTC AATATGTTTTTGTCAAGTATTTATTTCTATTTTACGCTTATATTACACAAGTTAAGCAT- 240

sog_sim         --------------------------------TGG'GCGGGGAAATCCCCGTGATCGCAGGTAGAATTCCAGCC-GG'GCCGAGGCGGGACCTGCTCGCACCTCTAATCCCGCC--------- 144
sog_sec         --------------------------------TGG'GCGGGGAAATCCCCGTAATCGCAGGTAGAATTCCAGCC-GG'GCCGAGGCGGGACCTGCTCGCACCTCTAATCCCGCC--------- 137
sog_mel         --------------------------------TGG'GCGGGGAAATCCCCGTAATCGCAGGTAGAATTCCAGCC-GG'GCCGAGGCGGGACCTGCTCGCACCTCTAATCCCGCC--------- 143
sog_yak         --------------------------------TGG'GCTGGGAAATCCCCGTAATCACAGGTAGAATTCCAGCC-GG'GTCGAGGCGGGACCTGCTCGCACCTCTAATCCCGCC--------- 141
sog_ere         GCTATC----------------------------TGG'GCGGGGAAATCCCCGTAAG-GCAGGTAGAATTCCAGCC-GG'GTCGAGGCGGGACCTGCTCGCACCTCTAATCCCGCC--------- 158
sog_pse         --------------------------------AA'GCGGGGCAGAGCCCGCGGTGTCAGGTAGAATTAGGTGTCTA'GCCTAGGCGGGACCTGCTCGCACCCTAATCCCACCCT-------- 168
sog_per         --------------------------------AAAGCGGGGCAGAGCCCGCGGTGTCAGGTAGAATTAGGTGTCTACGCCTAGGCGGGACCTGCTCGCACCCTAATCCCACCCT-------- 141
sog_ana         -----ATCCCTATGCACCACCACCACTGCTAGTGCTAGTGCCACCAATGTGGTGGCAGGTAG---------AGAATCCCGAGGCGGGACCTGCTCGCACCCCTAATCCC----------- 165
sog_moj         ------------------------------CTTTCGGGTTTTCCGCTGGACAGCAGGTAGTGATGGCCGGCT---------------------------------------------- 117
sog_vir         ----------------------CTATTTCTTCTTACGGGTTTGGGTTTTTCCGGCTACCTGGCCG--------GACAGCCGCTGCACTGGCTCGCACCCTTAAGCCGGGC---------- 126
sog_gri         ----------------------------TTTCGGTATTTCCCCCGCAGGTAGCGGTTCGGTTTGTCGGTGCCTCGAGGCGGGTACTGGCTCGCACCCTTAATCCCCCAGCCCAACC----- 154
sog_wil         ACGCCCAGTTGTCTCTAGGTGCAATCTATTATGGGTTTTTCCCCTTCCCCCCGAGCAGGTAG 300

sog_sim         --AGGGTTT----------------TCGGGACATGGGATATTCCCGACGGCACAGCATAGCCCACACTCCCTTTTCTTTTTT-----ATTGTTGTGCCCAGTT-TTAATCCGGAAAGC 243
sog_sec         --AGGGTTT----------------TCGGGACATGGGATATTCCCGACGGCACAGCATAGCCCACACTCCCTTTTCTTTTTT-----ATTATTGTACCCAGTT-TTAATCCGGAAAGC 236
sog_mel         --AGGGTTT----------------TCGGGACATGGGATATTCCCGACGGCACAGCATAGCACTCCGTTTTCTTTTTTTTTTTTTATTATTATTGTGTCCAGTTTTAATCCGGAAAGC 248
sog_yak         --AGGGTTT----------------TCGGGACCTGGGCTATTCCCGACGGCCCAGCACTGC----ACTCCCTTTTTTT------------ATTGT-CCAGTTT-TTAATCCGGAAAGC 228
sog_ere         --GGGGTTT----------------TCGGGACATGGGATATTCCCGACGGCACAGCACACCACACATCAGCACTCCCTCTTT------TTATTGTGCCCAGTC-TTAATCCGGAAAGC 256
sog_pse         --ATGACCTGCCTCATCCCTGCCATAGAGAGAGATTCCCTTTTTTTTTATTATTATT--TTATTTTATTGTTTTGTGTGT----------TTTTCTGCTTGTCTGCTAATCCGGAAAAC 277
sog_per         --ATGACCTGCCTCTTCCCTGCCATAGAGAGAGATTCCCTTTTTTTTTATTATTATT--TTATTTTATTGTTTTGTGTGTG---------TTTTCTGCTTGTCTGCTAATCCGGAAAAC 251
sog_ana         TGGCCCCTGGACGAAGGGATGCCGCCTCTGCCTGGGATATTCCC-AGCCTATTATAAAGATTTT-------------------------TTTTAT-TGTGTGGACTAATCCGGAAAAC 261
sog_moj         --------------------CGC-----ACCCTAAGCCGGCAGCGTGCTCG----------CTCCTTTGGGTTTTTCAC------AATTTTTTATGCCCGCCTGCTAATCCGGAAAAT 218
sog_vir         ----GGTGCG------ATGCGATGCGGTGTGTGTGTGCGTGTGTGTTGTTGCA-TT-----GCCTTGCAGCTTTGGGTTTTTCTTCTTTTTATGCTTGTGT--TAATCCGGAAAAC 231
sog_gri         GGTAGAGATACACACACACACACACTGCACAACACAACACAACAGATTGTTTAGCAGGCA GATCTGCTGTCCTGGGTTTTTCATCGCTTTTTATGTTAGTGCTAATCCGGAAAAT 276
sog_wil         CAAAACACACATTTATTGTTGTGCACAACTAAGAGGTGCTGCTCGCACCCAGCTTAATCCCTCTTGCGTGTTGGTGGTG------GTGGTGCGGGGGGTTTTCCCTAATCCGGAAAAC 418

sog_sim         GGGAATTCCCTTCCGCTC----------------------------GCTGCCTGCACTGCGCTGCGCAGACGCGTCGGCGTCCGTAAGCCGCTTACCAAAAAGATATGGGTATAACCAAATG- 333
sog_sec         GGGAATTCCCTTCCGCTC----------------------------GCTGCCTGCACTGCGCTGCGCAGACGCGTCGGCGTCCGTAAGCCGCTTACCAAAAAGATATGGGTATAACCAAATG- 325
sog_mel         GGGAATTCCCTTCCGCTC----------------------------GCTGCCTGCACTGCGCTGCGCAGACGCATCGGCGTCCGTAAGCCGCTTACCAAAAAGATACGGGTATACCCAAATG- 337
sog_yak         GGGAATTCCCTTCCGCTC----------------------------GCTGCCTGCGCTGCG-CAGACGCGTCGGCGTTCGTAAGCCGCTTACCAAAATGATATGGGTATACCCAAAACGG--- 314
sog_ere         GGGAATTCCCTTCCGCTC----------------------------GCTGCCTGCGCTGCGCAGACGCGTCGGCGTCCGTGTTCGTGTTCCGTATCAGTGTCCGTGTCCGTAAGCCGC 346
sog_pse         GGGAATTCCCTTCCGCTCGATGCTCGATGCTATTTGCTTGTTGCGCTGCCTGCGCAGACGCATAGGCGTCGGCGTCGGCGTCGCTATCGTGGTCGTTGTTGTTGTTGTTGTCGAACCGT-- 391
sog_per         GGGAATTCCCTTCCGCTC-GATGCTCGATGC------TTGTTGCGCTGCCTGCGCAGACGCATAGGCGTCGGCGTTGGCGTCGCTATCGTGGTCGTTGTTTTTGCCGAACCGT-- 358
sog_ana         GGGAATTCCCTTCCGTTT----------------------------GCTGCCTGCCGGTCCGGGCCGGTCCTCAGACTCTAGGGAGACTGCGTAGAGAATAGAATCTGAAAGATACTGAGATA 351
sog_moj         GGGAATTCCCTTCCGACTGTTTG----------CTGCCTGCCCGCTGCCTGCTAGGCCCAGACACCGAGGCGTCCTCAATGTTTGGGCCTTTGTCATGGGCTAAGCTGCTTAGC------ 332
sog_vir         GGGAATTCCCTTCCGCTGG-CCTGCACAGCCAGACATCGAGGCGACGCCAA-CAAAAAGATATGTGGCAAGCAGCGCGGCTTTGTCTATGCCTCTGCCTCTGCCTCTGCTAAGC- 344
sog_gri         GGGAATTCCCTTCCGCTCG-CTCGCTCGCCTGTCTGTCTGCCTACCTTTTTCGTTCCTAGACACATCGGCAACCCCATCCCAAATAACACAAAAATTACAGTGTAGAAAAGAACTCG---- 387
sog_wil         GGGAATTCCCTTCCGCTCGACCTGATCAGAGGTTCGATTGGCCTGGCCTGGCCTGGCCTGCCTAGACGCGGCGGCGGCGGCAGCAGCAACGTCGGTGGCAGTTGCAACCATTTTACT---- 537

sog_sim         GATGCCTATATAGACCATTGGACCACTTATGGACCAGGGACCATAAAGCGGCACCCAATTGCAATTTG---------------------------------------------------- 400
sog_sec         GATGCCTATATAGACCATTGGGTATGGACCAGGGACCATAAAGC---------------------------------------------------------------------- 369
sog_mel         AAAGATACTGAGATACAGATAC-------------------------------CAATTGCAATTTGTTGCAACTCACACGCTTGATTAGTTGCAC---------------------- 395
sog_yak         ATTCCTATATAGACCATATACCATATACCATAGACCATACACCATATACCATAGACCATACACCATAGACTATAAAGCGGCACCCA------------------------------ 400
sog_ere         TTACCGAAAAGATATGGGTATACCCAAATGGGTTCCTATGTAGACCATGTAGCCTAT-AAAGC---------------------------------------------------- 408
sog_pse         AAGCTGCTTACAACACCGGCAACAAACCGTAG-CAAGAGTCCAGGCAGCATAGGCACAGGC---------------------------------------------------- 451
sog_per         AAGCTGCTTACAACACCGGCAACAAACCGTAG-CAAGAGTCCAGGCAGCATAGGCACAGGCACAGATATCTGTGGCAGGCAGGCACCCAGGCCGTCAGGC---------------- 457
sog_ana         CAGATAC---------------------------------------------CAATTGCAATTTGTTGCAACTCACACGCTTGATTAGTTGCAC---------------------- 400
sog_moj         AGCTCACACACAATATACAATATACCATATATAGATATATAGATCTCTCTCTCTATATATATTTATTTATTTGTATATATTTTATGTATATTTTTATGGATCAGACCCAG 425
sog_vir         TGCTTAGCAGCTCGACCCAGGCCCAGGCCCAAGCCCAGTTAAGAGCTCCTTGCGTT---------------------------------------------------------- 400
sog_gri         CTTTTGTCTATAAGCGGCTTAGTATGAATATCAACTCTCAATAC---------------------------------------------------------------------- 431
sog_wil         AGTTTTCCCACTCTAGCCACACAACAACAACAACAACAAACAATTTGTTCTAAACCCATACGCTTGATTA---------------------------------------------- 600

sog_moj         TTTTGTTGGCCCTGGCTGCTGTTACGATTCTTACGCTTGATTAGCCGCATTTCCTGTTTTTATATAAAGATTCTCTCATTTTCCTTTT 500
```

Supplemental Figure 3. Alignment of all twelve *sog* Drosophila regulatory elements. Alignment was generated by ClustalW (http://www.ebi.ac.uk/Tools/clustalw2) and then adjusted by hand to show conservation. Boxes are drawn around well-conserved sequences. Grey box represents the sequence used for the synthetic enhancer in Figure 4C in the main text. Color code is the same as other figures: **Dorsal**, **Zelda**, **bHLH**, **Tmotif**, **Snail**, **Schnurri**



Supplemental Figure 4. Endogenous expression of *Thisbe* and *SoxNeuro* expression in *D. melanogaster*. (A-F) *ths* and *SoxN* expression are dynamic. Endogenous expression of the *ths* and *SoxN* transcripts detected by *in situ* hybridization using a riboprobe within embryos of nuclear cycle ~10 (A,D), cycle 14/stage 5 (B,E), and during germ-band elongation after gastrulation (C,F). *SoxN* is maternally expressed. It is likely that the *Ths* expression observed (A) is due to zygotic expression, since repression is observed at the anterior (A).

```
neu3-sim_RC      ------------------------------------------------ATGCTGACTGCG 12
neu3-sec_RC      ------------------------------------------------------G 1
neu3-mel         ------------------------------------------------ATGCTGAATGTG 12
neu3_yak_RC      ------------------------------------------------ATTCCGATTCTG 12
neu3_ere         ------------------------------------------------------A 1
neu3_ana_RC      ------------------------------------------------------A 1
neu3-pse_RC      ------------------------------------------------------
neu3-per         ------------------------------------------------------
neu3_vir         --------------------CGTTTCTATATAATTCCAAATTTGTTATATGAAATAGGCC 40
neu3_moj         GCATCCCTTAGTGATCCCTTTTACATTTCTCACATATGAAATACTTTTCGATTTTCTGTC 60
neu3-gri         ------------------------------------------------------

neu3-sim_RC      ATTCTCTATTCAGATACAATAGA--TATTGTGTAGCTATCTTTTCATTCAGTCTTTTCAT 70
neu3-sec_RC      ATTCTCGATGCATATACAATAGA--TATTGTGTAGCTATCTTTTCATTGAGTCTTTTCAT 59
neu3-mel         ATTCTCAATTCTGATACAATAGA--TATTGTGTAGCTAGCTTTTCATTGAGTCTTTTCAT 70
neu3_yak_RC      ATTCTCGGTTCTGATGATGTGTGGCTACCTTTTCATTGAGTTTCGTTTTCATTTTCCCAT 72
neu3_ere         ATGCTCGGTTCTGATGATGTGA--------------TGAGTCT-----------TCCCAT 36
neu3_ana_RC      AATTTCAAAGAAAATGTATTGTGTTAAACAGTTTTTAGTCTCTATACCTCCTCTATGGAG 61
neu3-pse_RC      ----------------------------------------------------GGCTGTCCAT 10
neu3-per         ----------------------------------------------------GGCTGTCCAT 10
neu3_vir         GGCTTCAGCTTTGCTTCGATGCGTTTCAATGGCAACTTCCTGTTGCGTGTAATTGTGTAC 100
neu3_moj         TATTTTAGCTGTGCTTCACTGCATCACAAAGCCAACTTCCTGTTGCCTGTAATTGCATAC 120
neu3-gri         ------------------------------------------------------ATAC 4

neu3-sim_RC      C---AACGTTGCTGGCCTGTCAATCAATGTAAGGCCAATTATTTCTGCACATC---ACTG 124
neu3-sec_RC      C---AACGTTGCTGGCCTGTCAATCAATGTAAGGCGAATTATTTCTGCACATC---ACTG 113
neu3-mel         C---AACGTTTCTGGCCTGTCAATCAATGTA--GCCAATTATTTCTGCACATC---ACTG 122
neu3_yak_RC      C---AACGTTGA-GGCCTGTCAATCTATATAAAGCCAATTATTTCTGCACATC---ACTG 125
neu3_ere         C---AACGTTGCTGGCCTGTCGATCAATA-AAAACTAATTATTTCTGCACATC---ACTG 89
neu3_ana_RC      C---CATCCCTCTGTCCCGTCAATCAAAATTAAGCCAATTATTTCC-CACATC---GCTG 114
neu3-pse_RC      C---AAAAT------CATGTCCATCAA------GTCAATTATCTCCACACATC---ACTG 52
neu3-per         C---AAAAT------CATGTCCATCAA------GTCAATTATCTCCACACATC---ACTG 52
neu3_vir         CTGTAAAATCTCATTAATCTCAACCCCTAAACAATAAACCGTATGTTCACACA--GATTG 158
neu3_moj         CTGCAAAGCCTCATTATTCCCGGCTCGTAAACAATTGCCAAT--GCCCAGGTATTCAGTT 178
neu3-gri         CTGAAA--------TAATCTCATCCCGCAAACAATACACAATTTGTTCACATACGAACTA 56
```

```
neu3-sim_RC      GACT---AAATCAGTTGCACCCTCGGCTCTTTAAGAATTTCCTT-GTCAAGCGGAAAACTCCC 183
neu3-sec_RC      GACT---AAATCAGTTGTACCCTCGGCTCTTTAAGAATTTCCTTTGCCAAGCGGAAAACTTCC 173
neu3-mel         GACT---AAATCAGTTGCACCCTCGGCTCTTTAA--ATATGTTT-GCCAAGCGGAAAACTCCC 179
neu3_yak_RC      GCCT---AAATCAGTTGCACTCTCGGCTTTTCAC--ATTTTCTT-GCCAAGCGGAAAACTGCC 182
neu3_ere         GCCT---AAATCAGTTGCAGCCTCGGCTGTTCAA--ATTTTCTT-GCCAAGCGGAAAACTCCC 146
neu3_ana_RC      AGCC---AAATCAGCCGAAGCCT----------------------GCTTCGGGGATATCCCT 151
neu3-pse_RC      AGCCGGGAGATCCCTCGTACTGCCTGCCCGAGAGGTGTGCCGTT---CTGCCTATTTGCA 109
neu3-per         AGCCGGGAGATCCCTCGTACTCCCTGCCCGAGAGGTGTGCCGTT---CTGCCTATTTGCA 109
neu3_vir         GGCCTAAACAAATTTCACAGCTTGCA-TAGAGAAAACTACAAAT---CATGCTGTTGGT- 213
neu3_moj         GGCCTAAACAAAT-TGGCAGCATTCA-ATGGAAAAACTACAAAT---CA--TCAGAGAT- 230
neu3-gri         AACATTTTCTACTGCCACAGACAAAA-GACAGACACAGACAAAT---CAACCCATAGATA 112

neu3-sim_RC      TAGGTGAAAATTCAACAACAAAAGACAAATAAACTCTGTTCGCTGC-AATAATTCG- 238
neu3-sec_RC      TAGGTGAAAATTCAACAACAAAAGACAAATAAACTCTGTTCGCTGC-AATAATTCG- 228
neu3-mel         TAGGTGGAAATTCAACAACAAAAGACAAATAAACTCTGCTCGCTGC-AATAATTCG- 234
neu3_yak_RC      TAGGTGAAAATTCCACAACAATAGACAAATAAACGCAGCTCACTGC-ATTAATCTGC 238
neu3_ere         TAGGCGAAAATCCATCAACAAAAGGCAAATAAACACAGCTCACTGC-ATTAAACTG- 201
neu3_ana_RC      CCTTCGG-------------ACAAAGGACGAACAATCGCAACTCATTGCGAATAGTTTC- 197
neu3-pse_RC      ACCCCGG--AAAACTCGCCCA-AAGATAAACACAGACTCTGATCAATGC-CATAATGTG- 164
neu3-per         ACCCCGG--AAAACTCGCCCA-AAGATAAACACAGACTCTGATCAATGC-CATAATGTG- 164
neu3_vir         TAGTTATGAA---TCGCTCCATCTGTGTTATCTCAGGCAGAATGAATGCTGAAAATTGCA 270
neu3_moj         TATTTATGAA---TCGCTGCATCTGCATTATCTCTG----AATGAATGCTAGATATTG-- 281
neu3-gri         TTGTTTGGAAAAGCAGCAGCATCTGTGCTATCTCAGGCAGAATAAATGC-GAATGCTGCA 171

neu3-sim_RC      ------------------------CATTGTGCAAATTCATGGAGGCATCTGTATGTCCGT 274
neu3-sec_RC      ------------------------CATTGTGCAAATTCACGGAGGCATATGTATGTCCGT 264
neu3-mel         ------------------------CATTGTGCAAATTCATGGAGGCATCTGTATGTCCGT 270
neu3_yak_RC      ATTCTGTATATTCTGTACATTCTGTATTGTGCAAATTCATGGAGGCATCTGTGCGTCCGT 298
neu3_ere         ------------------------TATTGTGCAAATTCATGGAGGCATCTGTACGTTCGT 237
neu3_ana_RC      ------------------------CATTGTGGAAAT-CATGAAAGCATCTCTCCGATC-- 230
neu3-pse_RC      ------------------------CATTGTGCAAAATCATGGAGGCATCTCTCTCTCCCG 200
neu3-per         ------------------------CATTGTGCAAAATCAGGGAGGCATCTCTCTCTCCCG 200
neu3_vir         A----------TATAAATGGCATTGTGCAAAATCAGAAAGGCATGTAGGTAG-------- 312
neu3_moj         -------------------GCATTGTGCAAAATCAAAAAGGCATATGGGTAG-------- 314
neu3-gri         AAT------------CGCAAATCGCATTGTACAAAATCATAGAGCTAGGCAGGCAGCCAG 219

neu3-sim_RC      -----TGTCGTTTCAGATAGA----GATCATAAGTCTTTGCTGCACCTGCCATCTGCCTT 325
neu3-sec_RC      -----TGTCGTTTCAGATAGA----GATCATAAGTCTTTGCTGCACCTGCCATCTGCCTT 315
neu3-mel         -----TGTCGATTCAGATAGA----GATCATAAGTCCTTGCTGCACCTGCCATCTGCATT 321
neu3_yak_RC      -----TGGCGATTTAGATAAA----TATCATAAGTCTTTGCTGCACCTGCCATCTGCCTT 349
neu3_ere         -----TGCCAATTCAGATAAA----GATCATAAGTCTTTGCTGCACCTGCCATCTGCCTT 288
neu3_ana_RC      -----TAGGGATTCATA---------ATC-----TCTCAGCTGCACCTGCCTGCCGGCCC 271
neu3-pse_RC      -----CTCGGATAGAGATCAGA--TGGTCTCCTGTGGATTGAGCACCTGCCTGCTGCAGT 253
neu3-per         -----CTCGGATAGAGATCAGA--TGGTCTCCTGTGGATTGAGCACCTGCCTGCTGCAGT 253
neu3_vir         ---------GCTAGA------TAAAGATCATGACTGTTGGCTGCACCTGCCAG------- 350
neu3_moj         ---------GTCGGC-----------ATGCCAGGTAAACTGT---ATTGCACCTGCCTC- 350
neu3-gri         GCAGGCAAGGCTAGAAATAGAGATAAAGATCATGACTCTTATAGAATTGCACCTGCCTC- 278

neu3-sim_RC      CTCTCCATTACTCGGGGTCAAACAATGGCCTGACGGGTGGTCAGATCCATTTC-GGCCAG 384
neu3-sec_RC      CTCTCCATTACTCGGGGTCAAACAATGGCCTGACGGGTGGTCAGATCCATTTCTGACCAG 375
neu3-mel         TTCTTCATCACTCGGGGTCAAACAATTGCCTGACGGGTGGTCAGATCCATTTC-GGCCAG 380
neu3_yak_RC      -TCTCCATGGCTCGGGGTCAAACAATAGCCTGACGGGTGGTCAGATCCATTTC-GGCCAG 407
neu3_ere         -TCTCCATTACTCGGGGTCAAACAATAGCCTGACGGGTGGTCAGTTCCACTTC-GGCCAG 346
neu3_ana_RC      CTAGACAATAGCCCGG---------CGGGTGGCCAGATGGTCGAGTG-ATTGCCGGCCGG 321
neu3-pse_RC      TCATGGAATGCACAGA------CAAAGGCCCAACGGGTGGGC--TCCGGGGGTCAACAAT 305
neu3-per         TCACGGAATGCACAAA------CAAAGGCCCAACGGGTGGCC--TCCGGGGGTCAACAAT 305
neu3_vir         GGGCCAGGGGCAAACAA-----CAAAGCCACAGCGGGTGGC----------AAAAACTAA 395
neu3_moj         GG-CCAGGGGCAGACAA-----CAAAGCCTCAGCGGGTGGT------------------- 385
neu3-gri         GG-CCAGGGGCATCTAAA----CAAAGGCGCAGCGGGTGGCCAAGCTAAGCAAAAACTA- 332

neu3-sim_RC      GTCAG--CAC--GCTATCTGTGCAAA-TCTAATAGAA-CAAAGCAAATG---ATTTCGAG 435
neu3-sec_RC      GTCAG--CAC--GCTATCTGTGCAAA-TCTAATAGAA-CAAAGCAAATG---ATTTCGAG 426
neu3-mel         GTCTGTGCAC--GCTATCTGTGCAAA-TCTAATAGAA-CAAAGCAAATA---ATTTCGAG 433
neu3_yak_RC      GTCAG--CAC--GGAATCTGTGCAAAATCTAATAGAA-CAAAGCAAATG---ATTTCGAG 459
neu3_ere         GTCAG--CTC--GGAATCTGTGGAAAATCCAATAGAA-CAAGGCAAATG---ATTTCGAG 398
neu3_ana_RC      GTCCTCACCCTGGCAATCTGCAGAATCGACAACAAAG-CGCTCCGCGTTTACATTTCGAG 380
neu3-pse_RC      GCCGT--------GAATCTGGGCAAAAACTAGCAAAAACAAAGCAGACA---TTTTTGCA 354
neu3-per         GCCAT--------GAATCTGGGCAAAAACTAGCAAAAACAAAGCAGACA---TTTTTGCA 354
neu3_vir         --TAG----------AATAAAACACA--GCAACAATT-TACAGCCCGCT---GCTTCGAG 437
neu3_moj         --CAG----------AGCAAA--------TA--AATT-TACAGTTCTTT---GTCTCGAG 419
neu3-gri         --TAG----------AATAGAATAGA--GTA-CAATT-TACAGCTCCTC---GCTTCAAG 373
```

```
neu3-sim_RC      TGCGGAACAGCGGGCGGTGGTGGAG----------------GCACTTCAATGACAAG--- 476
neu3-sec_RC      TGCGGAACAGCGGGCGGTGGTGGAG----------------GCACTTCAATGACAAG--- 467
neu3-mel         TGCAGAACAGCGGGCGGTGGTGGAACCTTATTCCCGAGGAGGCACTTCAATGACAAG--- 490
neu3_yak_RC      TGCGGAACAACGGGCGGTGGTGGAACCTTCGCTTCGAC---ACACTTCAATGACAAG--- 513
neu3_ere         TGCGGAACAGCGGGCGGTGGTGGAACCCTA--TGCGAG---GCACT-CACTGACAAG--- 449
neu3_ana_RC      TGCGAAACAGGGGCCGGAGGAGGGCG-------GCCACTGCGGACCCCCGAGACGAGCCC 433
neu3-pse_RC      TTCTGCGTTTACATTTCGAGTGCG-----------------GCACGGCACAGCGGGGC-- 395
neu3-per         TTCTGCGTTTACATTTCGAGTGCG-----------------GCACGGCACAGCGGGGC-- 395
neu3_vir         TGGAGCGCAACGGGCGGGCG-TTCATCCATAGAGT---------ATATAAGACTAGG--- 484
neu3_moj         TGGAGCGCAACGGGCGGATGGTTTATCTACAGAGCCTGCGCCTAGCCCGAGCTCGGGAGC 489
neu3-gri         TGGAGCGCAGCGGGT------TTCGGGCA---GAT---------ACTCGAGGTTGAG--- 412

neu3-sim_RC      GACTTCTATTCTTCTGTCCTTCCCTCCATTTCTCCCGAGGACTTCTTAGTTGAACTGCAG 536
neu3-sec_RC      GACTTTCATTCTTCTGTCCTTCCCTCCATTTCTCCCGAGGACTTCTTAGTTGAACTGCAG 527
neu3-mel         GACTTCAATTCTTCTGTCCTTCCCTCCATTTCTCCCGAGGACTTCTTGGTCGAACTGCAG 550
neu3_yak_RC      GACTTCCATTCTTCCATCCTGCCCTCCATTTCATCCGTGGACTTCTAAGTTGCGCTGCAG 573
neu3_ere         GACTTCCATTCTTCTGTCCTGCCCTCCATTTCTCCCGAGGACTTCTCGGTTGAGCTGCAG 509
neu3_ana_RC      GAAGTCTAATCTAAGACACTTCAGTG-ACAAGGACCTTGCAGTTCTGTGTTACTTTCCGC 492
neu3-pse_RC      GACAATTGTCC--GGGCACTTGAATG-ACAAGGACTTT-CATTCCTGCA----GCTGCAG 447
neu3-per         GACAATTGTCC--GGGCACTTGAATG-ACAAGGACTTT-CATTCTTGCA----GCTGCAG 447
neu3_vir         CTGCAGGCGGACACTTCACTCACAAGGACTTGCATTAAATTGTTGCACGGTTTTGTTTTT 544
neu3_moj         TCACGCTGAGGCACTTCACTCACAAGGACTTGCGTTAAATAGTTGGCAAGGCCGCTCTTG 539
neu3-gri         CCA--------------------ATGATTTTGCTACAGCCTGTCA---AGTTTTGTGTGT 449

neu3-sim_RC      CT-------GGGACCAAGGT---CAAGGACTTGGGCAATTAACACCCTGCGCGTTGGCGG 586
neu3-sec_RC      CT-------GGGACCAAGGT---CAAGGACTTGGGCAATTAACACCCTGCGCGTTGGCGG 577
neu3-mel         CT-------GGGACCAAGGT---CAAGGACTTGGGCAATTAACACCCTGAGCGTTGGCGG 600
neu3_yak_RC      CT-------GGGACCAAGGT---CAAGGACTTGGGCAATTAACACCCTGAGCGATGGCGG 623
neu3_ere         CT-------GGGACCAAGGT---CGAGGACTTGGGCAATTAACACCCTGAGCGATGGCGG 559
neu3_ana_RC      TTTTCCGACAGTGCCAAGGT---CAAGTGCCTGTGCAATTAACACCCCGGGTCCT----- 544
neu3-pse_RC      ---------GGGCACAGGTT---CAAGTGTCCCGGCAATTAACACCCTAAGCG------- 488
neu3-per         ---------GGGCACAGGTT---CAAGTGTCCCGGCAATTAACACCCCAAGCG------- 488
neu3_vir         TTGTTTTTGTCAGCAAGGTCAACTAA--GTCTGGTCTGTCCGCATTACAGGTGCTTGG-- 600
neu3_moj         CACCT------ACTCAAGTTCTCTGACGGCAAGGTCAGCTCGTCGTGTCGGCCTTTTTAA 593
neu3-gri         CAG--------AGCAAGGTCAGCCAA--GTGTCGGCATTTTTAGGTGCCGCTTTGCCGCT 499

neu3-sim_RC      GTGGCATTTAGCACACAAACACAAGCACCCAGGCGGCATTTTTCACAATT--CCAGAGAT 644
neu3-sec_RC      GTGGCATTTAGCACACAAACACAAGCACCCAGGCGGCATTTTTCACAATT--CCAGAGAT 635
neu3-mel         GTGGCATTTAGCACACAAACACAAGCACCCAGGCGGCATTTTTCACAATT--CCAGAGAT 658
neu3_yak_RC      GTGGCATTTAGCACACAAACACAAGCACCCAGGTGGCATTTTTCACAATTTTCCAGAGAT 683
neu3_ere         GTGGCATTTAGCACACGAGCACAAGCACCCAGGCGGCATTTTTCACAATT--CCAGAGAT 617
neu3_ana_RC      -----------CCCACAGGCCTCGGCTCGCA-----CATTTTTCACAGCC--CGGAAGAC 586
neu3-pse_RC      ----------------------------------GGCACTTCTTACGGTTTTTGGCAACG 514
neu3-per         ----------------------------------GGTACTTCTTACGGTTTTTGGCAACG 514
neu3_vir         -TGCCTGTCGCATTTAGCACATTTTGAACAATGAGATTTT------------TGGCAAGT 647
neu3_moj         ATGCGCCCGCCATTTAGCACATTCTTAACAATGAGAATTT------------TGGCAAGT 641
neu3-gri         TTGCCGCTTTGGTTTAGCACAATCTTAACAATGAGGTTTTACTTATATTTTTTGGCAAGT 559

neu3-sim_RC      --TTGGCAAGTGCCTGC-GAAGGTGACGAGACCACCCACGGCTCGA--------ATACCT 693
neu3-sec_RC      --TTGGCAAGTGCCTGC-GAAGGTGAGTAGACCACCCACGGCTCGA--------ATACCT 684
neu3-mel         --TTGGCAAGTGTCTGCCGAAGGTGAGGAGACCACCCACGGCTCGG--------ATGCCT 708
neu3_yak_RC      --TTGGCAAGTGTCTGC-GAAGGTGAG--GACCTCCCACGGCTCGAG-------AATACCT 732
neu3_ere         --TTGGCAAGTGCCTGC-GAAGGTGGC--GACCTCCCACGGCTCGA--------ATACCT 664
neu3_ana_RC      --CTAGCAAGTGTCTGC-CAAGGTGAG--GACCACCCCTGGGCGGAAT-----GCTTCTT 636
neu3-pse_RC      --ACAGCAGGAGGTAGCCACAGGTGGCTGGGCCACCCATGAGCCTGGT---------TCC 563
neu3-per         --ACGGCAGGAGGTCTCTACAGGTGGCTGGGCCACCCATGAGCCTGGTCCCCTTGTGTCC 572
neu3_vir         GCCTGGCAACTACCGCTTTGCACTTGGCCCGTTTGGGATGGCCCATTGCGTG-TGGCTAT 706
neu3_moj         GTCTGTCAACTGCCGCC---CACTGCGTGCACACACCTTCATCTGT-GTGTG-----TGT 692
neu3-gri         GTTTGTTAACTACCGCTT-GCACTGGGCTT-CATGGGACCGCCCATAGGATAACAGTTGC 617

neu3-sim_RC      TCTTTGACTTGC-CCAGCACGTCACCTGCTCCTGCTTAAGATGATCTACGGAAAGGCCCA 752
neu3-sec_RC      TCTTTGACTTGC-CCAGCACGTCACCTGCTCCTGCTTAAGATGATCTATGGAAAGGCCCA 743
neu3-mel         TCCTTGACTTGC-CCAGCACGTCACCTGCTCCTGCTTAAGATGATCTACGGAAAGGCCCA 767
neu3_yak_RC      TCCTTGACTTGC-CCAGCACGCCACCTGCTGCTGCTTAGGAGGATCTACGGAAAGGCCCA 791
neu3_ere         TCCTTGACT-GC-CCAGCAGGCCACCTGCTCCTGCTGAAGATGATCTGCGGAAAGGCCCA 722
neu3_ana_RC      TCTCTGACGTGC-CAGGCATGCCACCGCCGCC-----AGGTGTCCCTGCGCGAGGTGATC 690
neu3-pse_RC      TTTGTGACTCGAGCCAGGACTTCTTATGCTGTTCCCTGATCATCTATGATCCAGGG---- 619
neu3-per         TTTGTGACTCGAGCCAGGACTTCGTATGCTGTTCCCTGATCATCTATGATCTAGGG---- 628
neu3_vir         CAGGGGACAGCTTGATGTGCG-----CTGTGTCCTGTGACACCGGCTGTGT-TATTGTAG 760
neu3_moj         GTGTGTACGTGTGCGTGTGCGTGTGTGTCCTGTGTGATACCAGCTGCGCGT-TTGTAG 751
neu3-gri         ATAGTGACACCTTGATATATGTGTGTCCTTGTCCTGCGACACCAGCTGCGTATTTTAAAG 677

neu3-sim_RC      CAACGAAAGTGACAAGTGACTAATATACACTTGGCCAAA------------TTAACTTG 799
```

```
neu3-sec_RC     CAAGGAAAGTGACAAGTGACTAACATACACTTGGCCAAAATGACTATACAAATTAACTTG 803
neu3-mel        CAAGGAAAGTGACAAGTGACTTATATACACTTGGCCAAAAATACTATACAAATTAACTTG 827
neu3_yak_RC     CAGCGAAAGTGACAAGTGAACAATATGCACTTGGCCAAAA-GACGGTGCCAATTAACTTG 850
neu3_ere        CAA-GAAAGTGACAAGTGGCCAGTATACACTTGGCCAAAA-AACTGTGCAAATTAACTTG 780
neu3_ana_RC     CGGGGAAATGCCCAAAGGACAAGTGCAGA---GGCCACAA----------ATATAGCCCT 737
neu3-pse_RC     ------------------------------------------------------------
neu3-per        ------------------------------------------------------------
neu3_vir        CTG-----------------------------GCACAGGTAGTG-CCGTGAGTTAAGCCG 790
neu3_moj        ------------------------------------------------------------
neu3-gri        ACACACACACGCATACACACACACGCACA-ACACACAGGTAGAT-CGTAGAGTTAAGCCG 735

neu3-sim_RC     AAATGAACCTTCGTTCATC---------------CTTTTAATAATAATAA-TA---GTTT 840
neu3-sec_RC     AAATGAACATTCGTTCATC---------------TTTTTAATAATAATAA-TAATAGTTT 847
neu3-mel        AAATGAACCTTCGCTCATC---------------TTTATAATAATAATAA---TCCTTTT 869
neu3_yak_RC     AAAGGAGCATTTGTTTATCTGCATGGCAATGCATCTTGCACTTAAAATAAATAATCGTTT 910
neu3_ere        AAGTTAACCTTCGCTCACTCGCACGGGAATGCTTCTTACACTTAAAGTAAATAATCGTTT 840
neu3_ana_RC     A-------CTGCGCTCA----------------------------AGTAC---------- 752
neu3-pse_RC     ------------------------------------------------------------
neu3-per        ------------------------------------------------------------
neu3_vir        TGA-AATTTCACACCCGCT-----------GTGGTTGGTTTTTGTATGAAGTG---TTTG 835
neu3_moj        ------------------------------------------------------------
neu3-gri        TGA-AATTTCACACCCGCT-----------GTGGTTGGGTTTT-TATGAAGTGGTGTTTG 782
                                                           *
neu3-sim_RC     TTCTAGGAAAAACCAGGAAGTTGCGAAGCAGTAAAGAAGATAT--ATTAAACCATTT-TC 897
neu3-sec_RC     TTCTAGGAAAAACCAGGAAGTTGTGAAGCAGTAAGGAAGATAT--ATTAACCATTTT-CC 904
neu3-mel        TTCTAGGAAAAGCCAGGAAGTTGTAAAGCAGTTAAGAACATATCTATTAAACCATTTTCC 929
neu3_yak_RC     TTCCAGGAAAATACTGGAAGTTGTGCAGCAGTAAAAAAGGTAT--ATTCAACTATTTGCG 968
neu3_ere        TTCCGGGAAACGGTT-------------------AAAAGGTAT--TTTAAACCATTTGCC 879
neu3_ana_RC     ----------------------------------------------TGCCATTTCCC 763
neu3-pse_RC     ------------------------------------------------------------
neu3-per        ------------------------------------------------------------
neu3_vir        ------------------------------------------------------------
neu3_moj        ------------------------------------------------------------
neu3-gri        ------------------------------------------------------------

neu3-sim_RC     ACAGTGTTTGTCCAAGAGCCCATTGAAAGCAACTACAGGTAGACACCTTTT-GGCTA 953
neu3-sec_RC     ACAGTGTATGTCCAAGGGCCCTTTGA---CTACTACAGGTAGACACCTTTT-GGCTA 957
neu3-mel        ACAGTGTATATCCACGAGCCCATACAAAGCAGATGCAGGTAGACACCTTTT-GGCTA 985
neu3_yak_RC     ACAGTGTATGTCCAAGGGCCCATTCAAAGCAATTGCAGGTAGACAACACCTTT-GGCTA 1026
neu3_ere        ACAGTGTATGTCCAAGGGCCCATTCAAAGCAATTGCAGGTAGA---CACCTTT-GGCTA 934
neu3_ana_RC     CAAGTGTAGTTCCAGGCAGCGCAACAAATCAATTCTGGGTAGACA---CCTAC-CGATA 818
neu3-pse_RC     -----AAAGGGACTCCCAATCTATTTACAGGTAGAGGTCTTTAGATCTTCAGTTAATCCG 674
neu3-per        -----AAAGGGACTCCCAATCGATTGACAGGTAGAGGTCTCTGGATATTCAGTTAATCCG 683
neu3_vir        TCAAGCTGAGCC------CAGGCTCAGGTAGTACCCAATATGGCCACTTAAGTCTCCCAA 889
neu3_moj        TTG-----------------------------GCACAGGTAAAGGTAGCGAGCTAAGCCG 782
neu3-gri        TCAAGTTGAGCCATCGGCTCAGTTCAGGTAGAACTCAATCTGGCCACTTAAATCTAGCAA 842

neu3-sim_RC     ATCCGTGAAATTTCACACCCGCCATGGTTGATTCTCATCAGGATCAGGTACA---AC 1007
neu3-sec_RC     ATCCGTGAAATTTCACACCCGCCATGGTTGATTCTCATCAGGATCAGGTACA---AC 1011
neu3-mel        ATCCGTGAAATTTCACACCCGCCATGGTTGATTCTCATCAGGATCAGGTACA---AC 1039
neu3_yak_RC     ATCCGTGAGTTTTCACACCCGCTGTGGTTGATTCTCATCAGGATCAGGCATC---CT 1080
neu3_ere        ATCCGTGAAATTTCACACCCGCCGTGGTTGATGCTCATCAGGATCAGGTACA---AC 988
neu3_ana_RC     ATCCGTGAAATTTCACACCCGCTGCGGTCGATTTTCATCAGGAGTTGTAGGTATAGGTAT 878
neu3-pse_RC     TGA-AATTTCACACCCGCT-----GTGGTTTCTCTTTCAGTTTCTATCGGGCACAGGTCT 728
neu3-per        TGA-AATTTCACACCCGCT-----GTGGTTTCTGTTTCCATTTCTATCGGGCACAGGTCT 737
neu3_vir        ACTCATTTCTGTTTATCAT-CGTTCGGGTTAGGGAACTGCGAAGTATTTTTTTCCAACTA 948
neu3_moj        ------------------------------------------------------------
neu3-gri        ------------------------------------------------------------

neu3-sim_RC     A------------AATCGACATTCCTGCCTACT-CAC---TTTTATTTATTTGGCAAAGG 1051
neu3-sec_RC     A------------AATCGACATTCCTGCCTACT-CAC---TTTTATTTATTTGGCAAAGG 1055
neu3-mel        A------------AATCGAGATTCCTGCCTACT-CAC---TTTTATTTATTTGGCAAAGG 1083
neu3_yak_RC     GT--------CAGGATCGACATTCCTGCCTACT-CAC---TTT-GGTTATTTGGCAAAGG 1127
neu3_ere        A------------AATCGACATTCCTGCCTACT-CAC---TTT-GGTTATTTGGCAAAGG 1031
neu3_ana_RC     AGGTTGTAGGTAAGGTTGAGGGTCCTACCTGCC-CAGCAGTTCTCATTATTTGCAAAAGG 937
neu3-pse_RC     CAGGATGGGATGAC--------------------AATCCTGTTTGCCGAAGGGATTTCCT 768
neu3-per        CAGGATGGGATGAC--------------------AATCCTGTTTGCCGAAGGGATTTCCT 777
neu3_vir        CTTTGGCCCATTGCCAAAAAGTAAACATTGCCAGCGGCTTTAAATGTTTGCCATTCTATC 1008
neu3_moj        ------------------------------------------------------------
neu3-gri        TCGTTGCCTATTGCCAAAA-GTAAACATTGCCAGAGGCTTTAACTGTTTGCCATTGACTC 956

neu3-sim_RC     ATTTCAATAAATCCCCTCCTTT---ACCC-------------ACTGCCTATTCT------ 1089
neu3-sec_RC     ATTTCGATAAATCTC-TCCTTT---ACCC-------------GCTGCCTAATCT------ 1092
```

```
neu3-mel        ATTTCGATATTTGCCGTATCCTCAAATCCCCTTCTCTATCC-GCTGCCTTTTTCC----- 1137
neu3_yak_RC     ATTCCAATATTTGCCATACCCTCAAATCCACTCCCTAGCCCAACAGCTTATTCTTATTCT 1187
neu3_ere        ATTTCGATATTTGCCTTACGCTCAAATCC---------CCTTACCGCTTATTCTTATTCT 1082
neu3_ana_RC     ATTTCGGTGTCTGTGACTCCTCT--------------------CTGTTGTTTCTC----- 972
neu3-pse_RC     TCTCTGTAATCCCCTGGGGCCCACTGTTTGCTCTGTCAACTTTATTATCTGACC-GTTTG 827
neu3-per        TCTCTGTAATCCCCTGGGGCCCACTGTTTGTTCTGTCAACTTTATTATCTGTCC-GTTTG 836
neu3_vir        TGA-AAATTCACACCCGCT-----------GTGGTTGGTTTTT--ATGAAG--------- 879
neu3_moj        ------------------------------------------------------------
neu3-gri        GTTAGCTCCAGATGATTGCCTGGGCAATCAGGCAGTATATCCATATAGATATACATATGT 1016

neu3-sim_RC     CCGGGGTTTCTATATCCCTGCTGCCAGGTTGTTTGCGCATTAAAATTGTCAACTTTACTA 1149
neu3-sec_RC     CCGGCAGGGATATA-CCCTCCTGCCAGGTTGTTTGCGCATTAAAATTGTCAACTTTACTA 1151
neu3-mel        CCTCGGTTTCTATATCCCTGCTCACAGGTTGTTTGCGCATTAAAAATGTCAACTTTACTA 1197
neu3_yak_RC     CTGGAGTTTCCTTATCTCTGCGCACAGTTTGTTTAAGCATTAAAATTGTCAACTGTGTTG 1247
neu3_ere        CTGGCATTTCCTTATCGCTGCCTGCAGTTTGTTTAAACATTAAAGTTGTCAACTGTGTTG 1142
neu3_ana_RC     CTCCCAGTCCCAAATCC------ACTGTTTACCCCGGCATTAAAATTGTCAACTCTGCTA 1026
neu3-pse_RC     GCTCATAAAAAAGCCAACAACCTGCCGCAGGTA--CATTTAGCT-----GATAAA----- 875
neu3-per        GCTCATAAAAAAGCCAACAACTTGCCACAGGTA--CATTTAGCT-----GATAAA----- 884
neu3_vir        TGGTACTCTTGAC-ATTACCTGGGCAATCAGGCA-------TATATAGCTACACTAAC-- 1058
neu3_moj        ------------------------------------------------------------
neu3-gri        ATATATATACTGACTACATACACACTATTCATACAGCTCCTGCACTCTTCAGCTCGGTTG 1076

neu3-sim_RC     T-CTGTGTTGTTTCA---CCAGCTTGGCTC-----------GTAAAGCAGCCCACAAACC 1194
neu3-sec_RC     T-CTGTGTTGTTTCA---CCAGCTTGGCTC-----------GTAAAGCAGCCAACAAACC 1196
neu3-mel        T-CTGTGTTGTTTCA---CCAGCTCGGCTC-----------GTAAAGCAGCCAACAAACC 1242
neu3_yak_RC     TTCTATACCAGACTATACCCGACTTTAATGCGCGACTCGTAGTAAAGCAGCCAACAAACC 1307
neu3_ere        TTTTATACCAG------CTCGACTCC----------------CAAAGCAGCCAACAAACC 1180
neu3_ana_RC     T-CTGGGCAGTGTTGTTTCTGTGCCAGGCC-----------GGCTCGTAAAAAACCAACC 1074
neu3-pse_RC     -------------AAAACACACACCTACCAA--------------GATATCAGGCTGTGG 908
neu3-per        -------------GAA-CACACACCTACCAACAACATCCAGCTATGATATCAGGCTATGG 930
neu3_vir        -CATATATGTTAGCTCC-TGCGCTGC-TTCATGAAAG----GTGCTGTTTG--TCG--TG 1107
neu3_moj        ------------------------------------------------------------
neu3-gri        GCTATTTGATGTTGTCAACTTTATTATCTCATAAAGGAAAAGCAACAATCTGGGGCATTT 1136

neu3-sim_RC     -ATCGCAGGTATGGGTCATTCATGTTGATAAAGGTGTTGTTGCCTTATCATAACCAAAGG 1253
neu3-sec_RC     -ATCGCAGGTATGGGTCATTCATGTTGATAAAGGTGTTGTTGCCTTATCATAACCAAAGG 1255
neu3-mel        -ATCGCAGGTATGGGTCATTCATGCTGATAAAGGTGTTGTTGCCTTATCATAACCAAAGG 1301
neu3_yak_RC     CATCGCAGGTGTGGGTCATTCCCGCTGATAAGGCCA-----ACACTATCACAACCAACGG 1362
neu3_ere        CATCGCAGGTGTGGGCCATGCCTGCTGATAAAGCCA-----ACACTATCGCAACCAAAGG 1235
neu3_ana_RC     TGAGTCAGCCATTGACGGTTGTGGCTGATAAGGCAA------TCGCACTACAACTTGTTG 1128
neu3-pse_RC     GCTATAGCATTGG------------------------------------------------ 921
neu3-per        GCTATAGCTCTGG------------------------------------------------ 943
neu3_vir        TCAACTT--TATTATCTGCCCTGG-AGCTCATAAAAGCAAAGCAAGAAACCGACATGCTG 1164
neu3_moj        ------------------------------------------------------------
neu3-gri        TTATTTTCAGCTGATAAGGCTCAACAGCACTTACTACAACAAGTGGCATTAAAAGTTGTT 1196

neu3-sim_RC     CAAAAGGCACAAAAAGGGCCGAGGGGGGAAATGCAACTTGGTTTATTGCGCAGATATCA 1312
neu3-sec_RC     CAAAAGGCACAA---CGGCCGAGGGGGGAAATGCAACTTGGTTTATTGCGCATATATCA 1311
neu3-mel        CACAAGGCACAACAAGGTCCGAGTGGGGAAATGCAACTTGGTTTATTGCACAGATATCA 1360
neu3_yak_RC     CCAAAGGAGCAACAAAGAGCGAGGGGGGAAATGCAACTTGGTTTATGGCGCAGATATCA 1421
neu3_ere        CAAAAGGAACAACAAAGGGCGAGGGGG-AAATGCAAC----TTTATTGCGCAGATATCA 1289
neu3_ana_RC     TAGTTGGTTAATTGCGCAGATATCG------------------------------------ 1153
neu3-pse_RC     ------------------------------------------------------------
neu3-per        ------------------------------------------------------------
neu3_vir        C-ATATTAGG--GGTGTTTCTAATTGATAAGCCCAACAACTAATCGCATTGAAACTTGTT 1221
neu3_moj        ------------------------------------------------------------
neu3-gri        GGTTTTAGGTTTGGTTTTTTAATCATGGCAGATTAGG----------------------- 1233
```

Supplemental Figure 5. Alignment of *Neu3* homologous regulatory regions from eleven of the twelve sequenced *Drosophilids*. Sequences were aligned using ClustalW then manually adjusted to align conserved binding sites. Dorsal, Zelda, bHLH, Snail, Pointed

Supplemental Methods

1. Mutagenesis Sequences

All sequences are NotI fragments mutagenized in pGEMT easy and then sub-cloned into peveLacZattB unless otherwise noted. Sequences below include the entire sequence between the cloning sites. The following is the list of transcription factors and overrepresented motifs that we examined in the manuscript: Dorsal (GGG WWW(3-5)CCC), Zelda (YAGGYAD), T-motif (TTCCAGC), Snail (CACCT, DCADRDNN, MMRCAWGT), Schnurri (GRCGNCNNNNNGTCTG), STAT (TTCNN(2-3)GAA), Hunchback (GCATAAAAAA), bHLH (CANNTG), Pointed (SNGGAWRY). Mutagenic primer sequences are underlined in the sequences below. Lowercase letters indicate mutagenized bases.

Melanogaster *sog* minimal enhancer (LL16)

GAATTCACTAGTGATTGTTGCCAATGCCATTGCGCATACGCCGTGTCGTCTATATGGCTATATGGCTATAT
GGCTGTATGGTGCGGGGAAATCCCCGTAATCGCAGGTAGAATTCCAGCCGGTGCCGAGGCGGGACCTGCTC
GCACCTCTAATCCCGCCAGGGTTTTCGGGACATGGGATATTCCCGACGGCACAGCATAGCACTCCGTTTTC
TTTTTTTTTTTTTATTATTATTGTGTCCAGTTTTAATCCGGAAAGCGGGAATTCCCTTCCGCTCGCTGCCT
GCACTGCGCTGCGCAGACGCATCGGCGTCCGTAAGCCGCTTACCAAAAAGATACGGGTATACCCAAATGGA
TGCCTGCCCATGTATATAGACCATTGGGTGGTATGGACCATGGACCATAAAGCATCGAATTCCC

Dorsal sites mutagenized (LL20)

GAATTCACTAGTGATTGTTGCCAATGCCATTGCGCATACGCCGTGTCGTCTATATGGCTATATGGCTATAT
GGCTGTATGGTGCGcccAAATCCCCGTAATCGCAGGTAGAATTCCAGCCGGTGCCGAGGCGGGACCTGCTC
GCACCTCTAATCCCGCCAGGGTTTTCGGGACATcccATATTCCCGACGGCACAGCATAGCACTCCGTTTTC
TTTTTTTTTTTTTATTATTATTGTGTCCAGTTTTAATCCGGAAAGCcccAATTCCCTTCCGCTCGCTGCCT
GCACTGCGCTGCGCAGACGCATCGGCGTCCGTAAGCCGCTTACCAAAAAGATACcccTATACCCAAATGGA
TGCCTGCCCATGTATATAGACCATTGGGTGGTATGGACCATGGACCATAAAGCATCGAATTCCC

Zelda sites mutagenized (LL127)

GAATTCACTAGTGATTGTTGCCAATGCCATTGCGCATACGCCGTGTCGTCTATATGGCTATATGGCTATAT
GGCTGTATGGTGCGGGGAAATCCCCGTAATCGCtGcTtGAATTCCAGCCGGTGCCGAGGCGGGACCTGCTC
GCACCTCTAATCCCGCCAGGGTTTTCGGGACATGGGATATTCCCGACGGCACAGCATAGCACTCCGTTTTC
TTTTTTTTTTTTTATTATTATTGTGTCCAGTTTTAATCCGGAAAGCGGGAATTCCCTTCCGCTCGCgGaCg
GCACTGCGCTGCGCAGACGCATCGGCGTCCGTAAGCCGCTTACCAAAAAGATACGGGTATACCCAAATGGA
TGCCTGCCCATGTATATAGACCATTGGGTGGTATGGACCATGGACCATAAAGCATCGAATTCCC

Minimal *sog* enhancer with 3rd Dorsal site mutated (LL137)

GAATTCACTAGTGATTGTTGCCAATGCCATTGCGCATACGCCGTGTCGTCTATATGGCTATATGGCTATAT
GGCTGTATGGTGCGGGGAAATCCCCGTAATCGCAGGTAGAATTCCAGCCGGTGCCGAGGCGGGACCTGCTC
GCACCTCTAATCCCGCCAGGGTTTTCGGGACATGGGATATTCCCGACGGCACAGCATAGCACTCCGTTTTC
TTTTTTTTTTTTTATTATTATTGTGTCCAGTTTTAATCCGGAAAGCcccAATTCCCTTCCGCTCGCTGCCT
GCACTGCGCTGCGCAGACGCATCGGCGTCCGTAAGCCGCTTACCAAAAAGATACGGGTATACCCAAATGGA
TGCCTGCCCATGTATATAGACCATTGGGTGGTATGGACCATGGACCATAAAGCATCGAATTCCC

TTCCAGC mutated in minimal *sog* enhancer (LL171)

GAATTCACTAGTGATTGTTGCCAATGCCATTGCGCATACGCCGTGTCGTCTATATGGCTATATGGCTATAT
GGCTGTATGGTGCGGGGAAATCCCCGTAATCGCAGGTAGAATcCtAaCCGGTGCCGAGGCGGGACCTGCTC
GCACCTCTAATCCCGCCAGGGTTTTCGGGACATGGGATATTCCCGACGGCACAGCATAGCACTCCGTTTTC
TTTTTTTTTTTTTATTATTATTGTGTCCAGTTTTAATCCGGAAAGCGGGAATTCCCTTCCGCTCGCTGCCT
GCACTGCGCTGCGCAGACGCATCGGCGTCCGTAAGCCGCTTACCAAAAAGATACGGGTATACCCAAATGGA
TGCCTGCCCATGTATATAGACCATTGGGTGGTATGGACCATGGACCATAAAGCATCGAATTCCC

Dorsal replacement construct (LL191)
GGGAATTCGATGCTTTATGGTCCATGGTCCATACCACCCAATGGTCTATATACATGGGCAGGCATCCATTT
GGGTATAGGGGTATCTTTTTGGTAAGCGGCTTACGGACGCCGATGCGTCTGCGCAGCGCAGTGCAGGCAGC
GAGCGGAAGGGAATTGGGGCTTTCCGGATTAAAACTGGACACAATAATAATAAAAAAAAAAAAAGAAAACG
GAGTGCTATGCTGTGCCGTCGGGAATATGGGATGTCCCGAAAACCCTGGCGGGATTAGAGGTGCGAGCAGG
TCCCGCCTCGGCACCGGCTGGAATTCTACCTGCGATTACGGGGATTTGGGCGCACCATACAGCCATATAGC
CATATAGCCATATAGACGACACGGCGTATGCGCAATGGCATTGGCAACAATCACTAGTCGGAAGGGAATTC
CCGCTTTCGGAAGGGAATTCCCGTATCCCGTCGGGAATTCCCGCTTTCGGAAGGGAATTCCCGCTTTACTA
GTGAATTC

2. Synthetic Enhancer Sequences

Synthetic Figure 4A: *sog*'s native Dorsal sites (LL131)
GGGAATTCGATTTTACTAGTGGTGCGGGGAAATCCCCGTAAGACATGGGATATTCCCGACGGAAAGCGGGA
ATTCCCTTCCGGATACGGGTATACCCAAATGACTAGTCCAATCACTAGTGAATTC
Primer 79: GGTGCGGGGAAATCCCCGTAAGACATGGGATATTCCCGACGGAAAG
Primer 80r: CATTTGGGTATACCCGTATCCGGAAGGGAATTCCCGCTTTCCGTCG

Synthetic Figure 4B: 4x(CAGGTAG/TTCCAGC) (LL160)
AGATCTCAGGTAGAATTCCAGCAACAGGTAGAATTCCAGCATCAGGTAGAATTCCAGCAACAGGTAGAATT
CCAGC
Primer211:
cgAGATCTCAGGTAGAATTCCAGCaaCAGGTAGAATTCCAGCatCAGGTAGAATTCCAGCaa
Primer212:
atGCGGCCGCGCTGGAATTCTACCTGttGCTGGAATTCTACCTGatGCTGGAATTCTACCTGtt

Synthetic Figure 4C: 2x(Dorsal Zelda T-motif) + Dorsal (LL179)
GAATTCACTAGTGATTGGGGAAATCCCCGTAATCGCAGGTAGAATTCCAGCCGGTCCGAGGGGGAAATCCC
CGTAATCGCAGGTAGATTCCAGCCGGTGCCGAGGGGGAAATCCCAATCGAATTCCC

primer215: GGGGAAATCCCCGTAATCGCAGGTAGAATTCCAGCCGGTGCCGAGGGGGAAATCCC
primer216r: GGGATTTCCCCCTCGGCACCGGCTGGAATTCTACCTGCGATTACGGGGATTTCCCC

Synthetic Figure 4D: 3x(Dorsal and Zelda)
GAATTCACTAGTGATTCGAGATCTTATGCGGGGAAATCCCCGTAATCGCAGGTAGAATGCGGGGAAATCCC
CGTAATCGCAGGTAGAATGCGGGGAAATCCCCGTAATCGCAGGTAGAACA

Primer258:
cgAGATCTtaTGCGGGGAAATCCCCGTAATCGCAGGTAGAATGCGGGGAAATCCCCGTAATCGCAGGTAGA
ATGCGGGGAAATCCCCGTAATCGCAGGTAGAAcaGCGGCCGCat
Primer259: cgAGATCTtaTGCGGGGA
Primer260r: atGCGGCCGCtgTTCTAC

STAT synthetic Figure 5A: dgm140/89 STAT + (dorsal snail)x4
Construct was not site directed (3 independent lines were tested)
GGGAATTCGATTGAATTCTGGGAATTCCCAGAGGTGCATGCCGGTTGTGGGAATTCCCAGAGGTGCATAGA
TCTTGGGAATTCCCAGAGGTGCATGCCGGTTGTGGGAATTCCCAGAGGTGCATACTAGTGAATTC

STAT synthetic Figure 5A':
TATGCGGGAAATCCCCGAGGTGTCGTTCCAAGAAAAGGGAAATCCCCGGTTCCAAGAAAAGGGAAATCCCG
CGCA

Primer318: cgAGATCTtaTGCGG
Primer319r: atGCGGCCGCtgCGC
Primer320:cgAGATCTtaTGCGGGAAATCCCCGAGGTGTCGTTCCAAGAAAAGGGAAATCCCCGGTTCC
AAGAAAAGGGAAATCCCGCGcaGCGGCCGCat

3. Divergent Sequences used for transgenic embryo construction

Ananassae *sog*

GGGAATTCGATTCATTGCGCATACGCCGTGTCGGCCAATGGCGGCTGCTAGGGTGGCGGCTAGAGGGTGGG
TAGCGGGGGAAATCCCTATGCACCACCACCACTGCTAGTGCTAGTGCCACCAATGTGGTGGCAGGTAGAGA
ATCCCGAGGCGGGACCTGCTCGCACCCCTAATCCCTGGCCCCTGGACGAAGGGATGCCGCCTCTGCCTGGG
ATATTCCCAGCCTATTATAAAGATTTTTTTTATTGTGTGGACTAATCCGGAAAACGGGAATTCCCTTCCGT
TTGCTGCCTGCCGGTCCGGGCCGGTCCTCAGACTCTAGGGAGACTGCGTAGAGAATAGAATCTGAAAGATA
CTGAGATACAGATACCAATTGCAATTTGTTGCAACTCACACGCTTGATTAGTTGCACAATCACTAGTGAAT
TC

Mojavensis *sog*

GAATTCACTAGTGATTAACTCGTAGGCAATCTAACATTGCGCATACGCCCCGTCCGTCAGCTGGCCATAAA
CATGTATGTTTTTTGGTCGGCTTTCGGGTTTTCCGCTGGACAGCAGGTAGTGATAGCCGGCTCGCACCCTT
AAGCCGGCAGCGTGCTCGCTCCTTTGGGTTTTTCACAATTTTTTATGCCCGCCTGCTAATCCGGAAAATGG
GAATTCCCTTCCGACTGTTTGCTGCCTGCCCGCTGCCTGCTAGGCCCAGACACCGAGGCGTCCTCAATGTT
TGGGCCTTTGTCATGGGCTAAGCTGCTTAGCAGCTCACACACAATATACAATATACCATATATAGATATAT
ATAGATCTCTCTCTCTATATATATATTTATTTATTTGTATATATTTATGTATATTTTTATGGATCAGACCC
AGTTTTGTTGGCCCTGGCTGCTGTTACGATTCTTACGCTTGATTAGCCGCATTTCCTGTTTTTATATAAAG
ATTCTCTCATTTTCCTTTTAATCGAATTCCC

Pseudobscura *sog* enhancer

GAATTCACTAGTGATTCACATTCCCACAACAGTCGTTGCATAAGTCTCTGCCTCCGCGTGCAGCCGCATTG
CGCATACGCCGTGTCGCAAGCGGAAAGCGGGAAAGCGGGGCAGAGCCCGCGGTGTCAGGTAGAATTAGGTG
TCTACGCCTAGGCGGGACCTGCTCGCACCCTAATCCCACCCTATGACCTGCCTCATCCCTGCCATAGAGAG
AGATTCCCTTTTTTTTTATTATTATTTTATTTTATTGTTTTGTGTGTTTTTCTGCTTGTCTGCTAATCCGG
AAAACGGGAATTCCCTTCCGCTCGATGCTCGATGCTAGATGCTTGTTGCGCTGCCTGCGCAGACGCATAGG
CGTCGGCGTCGCTATCGTGGTCGTTGTTGTTGTTGTTGTCGAACCGTAAGCTGCTTACAACACCGGCAACA
AACCGTAGCAAGAGTCCAGGCAGCATAGGCACAGGCAATCGAATTCCC

Simulans *sog* enhancer

GGGAATTCGATTGTTGCCAATGCCATTGCGCATACGCCGTGTCGTCTATATGGCTATATGGCTATATGGCC
GTGTATGGTGCGGGGAAATCCCCGTGATCGCAGGTAGAATTCCAGCCGGTGCCGAGGCGGGACCTGCTCGC
ACCTCTAATCCCGCCAGGGTTTTCGGGACATGGGATATTCCCGACGGCACAGCATAGCCCACACTCCCTTT
TCTTTTTTATTGTTGTGCCCAGTTTTAATCCGGAAAGCGGGAATTCCCTTCCGCTCGCTGCCTGCACTGCG
CTGCGCAGACGCGTCGGCGTCCGTAAGCCGCTTACCAAAAAGATATGGGTATAACCAAATGGATGCCTATA
TAGACCATTGGACCACTTATGGACCAGGGACCATAAAGCGGCACCCAATTGCAATTTGAATCACTAGTGAA
TTC

Virilis *sog*

GAATTCACTAGTGATTTACGCCACGCCCGCCGGCTGGGCCATAAAAGTGGGCCCTGTTTTTCCTATTTCTT
CTTACGGGTTTGGGTTTTTCCGGCTACCTGGCCGGACAGCCGCTGCACTGGCTCGCACCCTTAAGCCGGGC
GGTGCGATGCGATGCGGTGTGTGTGTGCGTGTGTGTTTTGTTGTTTGCCTTGCAGCTTTGGGTTTTTCTTC
TTTTTATGCTTGTGCTAATCCGGAAAACGGGAATTCCCTTCCGCTGGCCTGCACAGCCAGACATCGAGGCG
ACGCCAACAAAAAGATATGTGGCAAGCAGCGCGGCTTTGTCTATGCCATTGCCTCTGCCTCTGCCTCTGCT
AAGCTGCTTAGCAGCTCGACACAGGCCCAGGCCCAGTTAAGAGCTCCTTGCGTTAATCGAATTCCC

Yakuba *sog*

GGGAATTCGATTCCAATGCCATTGCGCATACGCCGTGTCGTCTATATGGCTATATGGCTATATGGCGATAT
ATGGTGCTGGGAAATCCCCGTAATCACAGGTAGAATTCCAGCCGGTGTCGAGGCGGGACCTGCTCGCACCT
CTAATCCCGCCAGGGTTTTCGGGACCTGGGCTATTCCCGACGGCCCAGCACTGCACTCCCTTTTTTTATTG
TCCAGTTTTTAATCCGGAAAGCGGGAATTCCCTTCCGCTCGCTGCCTGCGCTGCGCAGACGCGTCGGCGTT
CGTAAGCCGCTTACCAAAATGATATGGGTATACCCAAAACGGATTCCTATATAGACCATATACCATATACC
ATAGACCATACACCATATACCATAGACCATACACCATAGACTATAAAGCGGCACCCAAATCACTAGTGAAT
TC

4. Other enhancer sequences:

Ths enhancer ("high-affinity" Dorsal sites; confirmed previously Stathopoulos et al., 2002)

GGACCAGCACGAGCTACGCAGCCTCACACAGCAGGATAATAGGGAAAGAAAGGACAGGACAATGGAGGTTC
AGAAGAAGCGAGCAAATGCTGGAAAATGCAGTGACAACAGGTGCAAAATTATTTTTTGTTTGTTGCGAGTG
CGCGTGAAAATTTCCAGCTGGCCAGGGACAGGAATATGACCACTTAAGGCCTAATGTGCGAAAAGTTCCTT
TGTCAATTTACACGCCCTCTCCTCACCAAGCGACCGTGAAAACTTCATTCATTTGCATGGCTAAGCTCAGG
TAGCCGGGGATTATCCCTCGTTCTAACCAAAACCTCCTGCTACATTCGGGTTTATCCCACTTGTTTGGCTG
GGAATTTCCCCCGCAGATTTACGGTGGTCAGCCAAATCCCGGTTAGCTGGAAAAATATCGCAGAAATAAGG
GAAGGTACGGCTGCTAATGAATCCTGACATCTCAATCAATTTTGGGGGAATCGAAACGCAAGGGAGGTGGA
ACTTTCACCAC

Neu3_3R_10504881_10506240 (dorsal sites are "weak" with regard to PWM scores)

ATGCTGAATGTGATTCTCAATTCTGATACAATAGATATTGTGTAGCTAGCTTTTCATTGAGTCTTTTCATC
AACGTTTCTGGCCTGTCAATCAATGTAGCCAATTATTTCTGCACATCACTGGACTAAATCAGTTGCACCCT
CGGCTCTTTAAATATGTTTGCCAAGCGGAAAACTCCCTAGGTGGAAATTCAACAACAAAAGACAAATAAAC
TCTGCTCGCTGCAATAATTCGCATTGTGCAAATTCATGGAGGCATCTGTATGTCCGTTGTCGATTCAGATA
GAGATCATAAGTCCTTGCTGCACCTGCCATCTGCATTTTCTTCATCACTCGGGGTCAAACAATTGCCTGAC
GGGTGGTCAGATCCATTTCGGCCAGGTCTGTGCACGCTATCTGTGCAAATCTAATAGAACAAAGCAAATAA
TTTCGAGTGCAGAACAGCGGGCGGTGGTGGAACCTTATTCCCGAGGAGGCACTTCAATGACAAGGACTTCA
ATTCTTCTGTCCTTCCCTCCATTTCTCCCGAGGACTTCTTGGTCGAACTGCAGCTGGGACCAAGGTCAAGG
ACTTGGGCAATTAACACCCTGAGCGTTGGCGGGTGGCATTTAGCACACAAACACAAGCACCCAGGCGGCAT
TTTTCACAATTCCAGAGATTTGGCAAGTGTCTGCCGAAGGTGAGGAGACCACCCACGGCTCGGATGCCTTC
CTTGACTTGCCCAGCACGTCACCTGCTCCTGCTTAAGATGATCTACGGAAAGGCCCACAAGGAAAGTGACA
AGTGACTTATATACACTTGGCCAAAAATACTATACAAATTAACTTGAAATGAACCTTCGCTCATCTTTATA
ATAATAATAATCCTTTTTTCTAGGAAAAGCCAGGAAGTTGTAAAGCAGTTAAGAACATATCTATTAAACCA
TTTTCCACAGTGTATATCCACGAGCCCATACAAAGCAGATGCAGGTAGACACCTTTGGCTAATCCGTGAA
ATTTCACACCCGCCATGGTTGATTCTCATCAGGATCAGGTACAACAAATCGAGATTCCTGCCTACTCACTT
TTATTTATTTGGCAAAGGATTTCGATATTTGCCGTATCCTCAAATCCCCTTCTCTATCCGCTGCCTTTTTC
CCCTCGGTTTCTATATCCCTGCTCACAGGTTGTTTGCGCATTAAAAATGTCAACTTTACTATCTGTGTTGT
TTCACCAGCTCGGCTCGTAAAGCAGCCAACAAACCATCGCAGGTATGGGTCATTCATGCTGATAAAGGTGT
TGTTGCCTTATCATAACCAAAGGCACAAGGCACAACAAGGTCCGAGTGGGGAAATGCAACTTGGTTTATTG
CACAGATATCA

Putative SoxN2L8841307_8842609 (did not work)

TATTTAATTTATAACTTGCGTGCTAGCCAATCAGGATCCCTGATGTCATATATAAGCCTGACTTTCGCTGC
GATTAGGCTTATCGATCTCGGTGCGCCCTCATGTCTGAACGCCTCTGGCCCAAAACAGACCTCCCCGGAAA
GGGATTGGAAATCCCTGGCCGCATCCAAGCCGCAGAATTCAATGGAAAGCGACAGGCAGTCTGGCTACAGT
CCGCCCAGCTAATCTTATCGCCGCTGACGATGCAACTGTCGATAAGAGATCGCCAAATCAAACGAACCGCA
AACGGAGGATCCTTCGATCCGGAGCAGGATCGAGTGCACAGTAGTCCGCAAGCAATGATTACCCATGCACA
GTGCGTCCTTTGTCTACTAAATCGAATACCGCAAAAAAACACACAGGGATCAGAGGAGAAGTGGAGATAAG
AATCGAAATTGTAAAATGGTACGTAAAAATGCGAGGATTCATAGTTCCCAGAAAGGCCCTGGACCAACCCC
CTTCCACCAGGATTTCAGGACTTCAGGACATCGGGACAGGACCTACCCTCCCAGCTCTGAAATACTCCTAC
GCATCAGCCTCTGCCTGTTACATCCTGTGTGCAAAGGTGGCAAAAACTCGGAGCTACCTGCGGAGCATCGG
CGTAAAATGTGTCCCTCGTCCTGGGTCTCGGTAAAGATTTATTATCCTATACATTTCATGAATGCGATATG
CGAGTGAGCCGGCAACTGTATCTGCATCTGTGTTCGTATCACACTATCTGTGCCAGTGTGTGTGTGAGGGT
TTCCCTGGAAAACTACTACTACCACAACTACTGCCAACTAAGGCAGCAGCGAAAAGCAACAACAACACTGG
GAGCACAAAAGGTGTGCACCATAATGGCTGACAAAAAGTCGGCCTTAGCCTTTAATAGCACTGAGTTTAAC
AGCCAACACAAACATGGCCCAACTGCCATGCACTGCCCGAAAAAAAACCCATCTGCGGTTTGCATCAAGTG
CAAGTTCAACTTGCAGGGAAAGTTTCTTTTATGACTCTGTCTTAGGGGTGTCAATGGCTTCCAGTCCGAAA
CATCAAACGAATCATTTTATTAACTTAACTTTACATTACCTATATTGAACATACATTTGAAGTTTCATAAT
TTCAGAGTTGTCTAATAAGTTTTCTTTCAGTGCACAGCAGCCAACTCCATAGAACCCATTCGCTGCAATCC
AACAAAATGGCAACATTCACTTCTCACTTACAATTGAAGCTGGGGCTTCCTTCGTTTTTTTCGCATGGAAA
AAGTGAAAAGTGAAACTATTTGCTG

Putative Pyramus enhancer (did not work)

GATGATTTGTAAGCTGTGATTCAGTGACCAGCGATCTAATGAGACCTAATGGGGCATTCTGGAAGGAAAGGCGGATGCT
GAAGAGGTATTTTTAATGTTATACTAATAAGCTTTGTTGTTATAGTATCATGGAAGGAATTCAATCAACGCAACAAAGT
TTTAGCACTGCCATTTCAAACGTTTGATAATCGATGGCTCTATCTTTCGAACAAGGATTCTTTTAATATTTTTGGCATA
TACATAGAATGCCAAAAGGTTTGAAATGCAACGCTTTGGTCACATGCCACCCAAGATTTAACCGATCCCAATCCCTTAT
CAGCGTGGCTAATTACAAGTAAAAAAGTAAAGTGCCAGATAAAACACTTCTGCCCTCAGATAGCACACAAATATCGGTG
CTCAGTAGTTAGTCCAGGGAACTGTTGCAAAAAAAAAAAACTGTCAAGAAAAAATAGCTGGACAAAGGGAGCGACAGGG
GACGACGTGGGCTGAGGCGATAAGAGGAACCAGCCAGGTAGAGTCCGGATTCCAGCCTGGCCCAAATGTCCAACGAATT
GACAATTTATGTGACAGGCCGTGCATAATTGAGGCCTTTGGCAATTGGCAACTGCCGGTTGTTGGCCAGGTAAACCCGC
CGGGTAATAGGTATGTAATTTTTCCTGGTAAAACTTAATTTCATTGATGTCCTTGCACAGGATCGTCGTGTGCCTTCGG
CCGATGCGTCGTCGGTCGAATATCAAAGAATCGAATCCCATACAATTGAAGTGATTTATATGACCATTGATAGACTGAT
TTTTATAGACTACAGCCAATTAACAGCTGCCTCCGACTTCCGTTTTTGGTTGACGGACTCAATCTATTCTAGGCTCTAA
TTGTGGATAATTGCGGCTAGCGCCAAGGTCAAATGATTGATTTGTCCGGGCCGTAAACTTCTTGAAAGAGGGATGTGCC
AGCAATAAGACATAGTCCCCGGAAAACGGAACAGACCTCTCTACGTGAACGCAACTCAAGTGTTATTGCCTCCGGCCAC
TAAACCGACTTAGTGGGCAATTCTCTATTGTTAAGCTCTTTTAGTTTAGCGATAAGACATAATTTAGCCAAAGTTACCA
GTTCTAACCGGTTGTCCACTTATCTACAATGTATACTTCACGTTTCAGTTACACCCACTGCCGTATAAACTTTAGTATA
TATAGTAATTTAAACGGACGCACCTCTGATCTAAACGGTGCTTTCTGAATCACCTCAAGTGATAACTCTAATGTGTGCG
AATTCTTGGAAATTGAATTGCTGTGGTCATGGCTAAATGGAATTCGTGATTAGAAACCAATTTAGCACGCAATTGCGGA
AGCCCTCCAATGGAACTTGCACCTTTTCTAGTGCATATATATAAATAGGACTGTGGAAAGGGATTCCCCGCCTTATGGG
TAAATATTTGATAGTTCTTTGATTTTCGCATTTTAACGTTGCTTTTGTTACATTGTTACTCAAAAGGCATGCTGCCTGA
GAAAGTGTAAATTCTTTTCCGAAACCATTAAATTTATAAAAGTAACATAATTTATATTTATCGCATATTTTGGGCACCT
GACTTAAGAATTGGCTATTAAAAGTGCCGAGTCAGGTTGACAAACGCATTGGCACATATCTGGATATGCGATAACTTGA
CCTTGACCTATCACCACAAATTCGAATCTACAATCTAGGGTGTTGACCACGGCAACAGGGGTGCTTTTTGATGCGATAG
GTCTGCTCCATAAACAAGAAACAGCTAATGGCACAGGTATTTTTACTATTACCATTTACCGACGATAAACCACAACCAA
AAAGAAAAGAACAATAGAAAAGCTGGTCAAGTGGAGGCCTGTGGGCCAATAAACTGAGCAATTGCCGTTTGCGTAACGG
GCCCCCCTTAGTCGCATTTCCAATCGATAAGTGATAGTCAGACGAAGGTATTTATTTTGTCCAAATTGCGACTTAATGA
CCGCCCAAGAGGTTGTAGAAAGTGATGTCATCCGAGACCGCTGATAGCACCAATTAGCCACTTTAATTTCAGCTGGCGC
CCCAAAATGGAAAAATTACCAAATCCTTGTACAAATCTTGTTGCCAATGGGTTTATG

Supplemental Table 1. Fly lines and constructs generated in this work.

Fly line	pos	AttB DNA	pGEM DNA	seq	figure ref	ref	description
LL16	16	93	16	8/12/07	2B	this work	wildtype sog reg ele -evepLacZ
LL16	attp2	93	16	8/12/07	not shown	this work	wildtype sog reg ele -evepLacZ
LL93	51D	93	16	8/12/07	1B	this work	wildtype sog reg ele -evepLacZ
LL20	16	96	20	10/3/07	2C, 2D	this work	sog reg ele w/ dorsal binding sites mut
LL108	51D	108	PCR	yes	3A	this work	synthetic with native dorsal binding sites
LL115	51D	115	53	yes	SF2	this work	ana1-103f ana evepCherry
LL126	16	126	123	1/12/08	not shown	this work	123/89 (chopped sog)
LL127	16	127	125	1/25/08 r	2E, 2F	this work	125/89 (caggtag mutated)
LL127	51D	127	125	1/25/08 r	2E, 2F	this work	125/89 (caggtag mutated)
LL128	51D	128	122	11/29/07	4D	this work	122/89 (neu3-2)
LL129	51D	129	55	1/25/08	not shown	this work	pse/89
LL130	51D	130	58	1/25/08	1E	this work	yak/89
LL131	16	131	37	1/25/08	3A	this work	sog's Dl sites native (LL37)
LL137	51D	137	12	2/5/08	2I, 2J	this work	12(sog enhDl3*)/89
LL138	51D	138	54	2/5/08	1K	this work	54a(moj)/89 for
LL140	51D	140	56	2/5/08	SF2	this work	56b(sim)/89 for
LL141	51D	141	57	2/5/08	1H	this work	57c(vir1)/89 rev
LL160	51D	160	PCR	4/8/08	3B	this work	caggtag/tcc x4 bgl2/notI
LL171	51D	171	77	4/16/08	2G, 2H	this work	77b/89 (TTCCAGC mut sogwt (CW1))
LL179	51D	179	PCR	4/16/08	3C	this work	162a/89 (dct5 (2 copy)
LL189	51D	189	167	7/14/08	3D	this work	167/89 "L" caggtag/dorsal only
LL191	51D	191	41	8/4/08	2K, 2L	this work	all Dl* + Dl3x4 (20/17-5)
LL195	51D	195	PCR	8/8/08	not shown	this work	dorsal twist syn
LL196	51D	196	PCR	8/5/08	4B	this work	Dorsal stat syn
dgm140	p ele			des's notes	4A	this work	dl, STAT and sna syn from brinker reg ele

otherDNA and lines not in paper but related:

LL153		153	3/29/08		dgm140/89 STAT/Dl/snail
LL185	51D	185	5/14/08	not shown	182/89 dl3wt in sog mut enh (clone attp1)
LL186	51D	186	5/9/08	not shown	183/89 mini sog HI "mini3" forward
LL190	---	190		not in paper	ana/89 evep LacZ

Appendix B: Supplemental Information for Chapter 3

Table of Contents

1. Unrolling the embryos

As described in the methods section of the main paper, confocal z-stacks were obtained of the embryos. Roughly 110-120 slices were taken with approximately 0.9 microns between each slice. The first few slices were above the plane of the top of the embryo, while the last slice taken was just beyond the slice with the largest of embryo saggital section. The zoom level was fixed such h that the xy pixel size was roughly 0.3 microns, such that the z-scale was three times the xy scale. The resulting image stack was approximately 1024 x1024 x 115 pixels.

The outer edge of the embryo was found in the following manner. yz sections were taken in groups of about 16 x-slices and averaged together. In the z direction, each line of the image was repeated three times, ensuring that the image was isotropic (i.e., the yz pixel size of the new image became 0.3 microns on a side; see Figure S1A).

Using the lowest z pixel and center y pixel as a reference point, the image was divided into 30 radial slices (Figure S1A), and the intensity within each slice as a function of distance from this point was normalized to fall between zero and one (Figure S1B).

To determine where within this theta-slice the periphery of the embryo was located, two thresholds were set. The low threshold, at 0.05, denoted the "background" intensity. The distal-most location where the intensity crossed this threshold was counted as the distal-most point where the embryo edge could be. The next-most proximal point where this low threshold is also crossed defined the proximal-most location where the embryo edge could be, as long as somewhere in between, the intensity became higher than the high threshold, which was defined at 0.25. The intensity between these two bounds was then renormalized to be between zero and one, and the embryo periphery for this "theta-slice" was then defined to be the distal-most point where the new intensity crossed the value 0.25 (circle in Figure S1B). This procedure was repeated for each slice of theta.

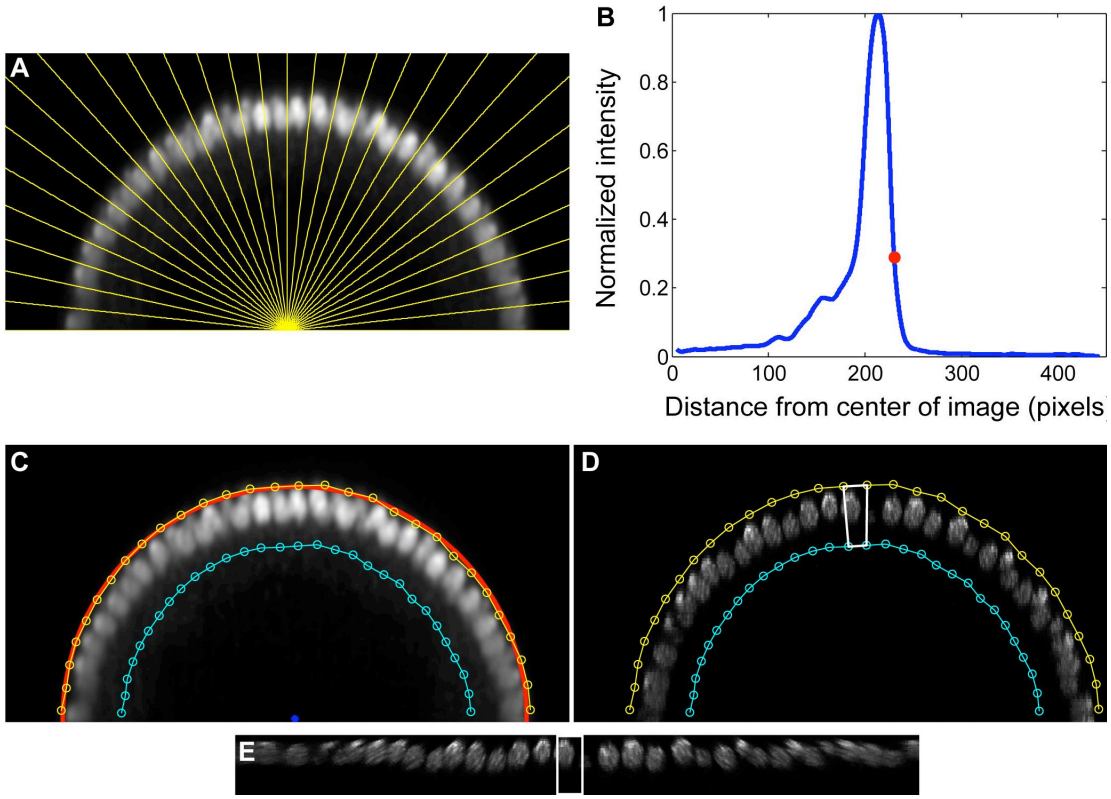


Figure S1: Unrolling the embryo. (A) Dividing a yz cross section into 30 domains. We define the "center" of the image as the lowest point in *z* and the midpoint in *y*. (B) The intensity of a single domain from (A) as a function of distance from the center of the image. Using the algorithm described in the text, we define the edge of the embryo as located at the point in red. (C) The periphery of the embryo. After the algorithm described is repeated for each domain, we obtain a series of points which define the periphery of the embryo (yellow circles). These points can be fit to a circle (red curve) which helps us determine the presumptive center of the embryo (blue dot) and the likely "inner" border of the nuclear layer (cyan). (D) Unrolling one yz-slice. Adjacent pairs of outer periphery points and the corresponding pair of inner points defines a quadrilateral (white box). Each of these quadrilaterals is slightly distorted to become a rectangle (see white rectangle in (E)). (E) Unrolled yz slice. Using the keystone-like distortion, the yz slice is converted into a strip of nuclei. The white rectangle here corresponds to the white quadrilateral shown in (D).

After these presumptive periphery points were found (yellow in Figure S1C,D), they were used to determine the best-fit circle to the periphery of the embryo (red in Figure S1C). Any point with a residual greater than 2.5 times the standard deviation of all residuals was thrown out, and a new best-fit circle was found. After no points fell outside this 2.5 standard deviation limit, the missing points were replaced with points that lie perfectly on the best-fit circle (example not shown). In total, these points were chosen as the periphery of the embryo in this grouping of yz-slices.

Along with fitting the periphery of the embryo to a circle, the coordinates for the presumptive "center" of the embryo were found (blue point in Figure S1C). An "inner periphery" was then determined by moving each outer peripheral point an average of 60 pixels closer to the presumptive center of the embryo (cyan in Figure S1C,D). These points of the inner and outer periphery delimited a series of quadrilaterals that contained the outer surface of the embryo (white box in Figure S1D). Using an affine transformation (a keystone-like transformation), each

quadrilateral was morphed into a rectangle with a height of 60 pixels and width equal to the distance between the two points along the periphery of the embryo that defined the outer edge of the quadrilateral (white rectangle in Figure S1E). Adjoining each of these rectangles yielded a strip of embryo periphery containing all the necessary information (i.e., all of the nuclei) that originated in the yz-slice (Figure S1E). It is important to note that, while the periphery of the embryo was found using groupings of roughly 16 yz-slices, the yz-slices were each unrolled individually using the peripheral points that were found using the group that the yz-slice originated from.

After each strip was found, the proximal-distal axis of the embryo has essentially become the z-axis of the strip, while the y-axis of the strip corresponds to the dorsal-ventral axis of the embryo. The information contained in this strip was averaged in the proximal-distal direction. At dorsal-ventral coordinates where nuclei were present, this averaging was weighted such that only proximal-distal intensities corresponding to the location of the nucleus were used. Thus, the strip was compressed into a one-dimensional (1D) "image" of average intensities in each color channel. The length of this 1D image roughly corresponds to the length of the arc characterized by the periphery of the embryo at the given yz-slice.

Note that, in general, this arc length will be different depending on the x-location (i.e., anterior-posterior location) because the width of the embryo varies with respect to anterior-posterior (AP) location. Therefore, to concatenate these 1D images into a full 2D image, the 1D images were stretched to correspond to the length of the longest 1D image. After this procedure was performed on the original volume of the embryo, the data were compressed into a two-dimensional sheet of intensities for each color channel (see Figure 1D of main paper).

2. Nuclear segmentation

The nuclei were segmented according to the estimated nuclear cycle. First, the "center" of the 2D sheet of nuclear staining was taken (the rectangle from 25% to 75% width and from 25% to 75% height). This part of the image corresponds to the pat of the embryo closest to the objective on the microscope, and thus with the least amount of signal loss due to (1) light scattering through the tissue and (2) poor z-resolution as compared to xy-resolution, both of which affect the periphery of the image.

This center of the image of nuclear stain was thresholded at a level predicted by Matlab's graythresh function. The resulting binary image (Figure S2A) approximated the segmented nuclei for the center of the image. After removing outliers of small area, the remaining number of objects in the binary image was used to estimate the nuclear density, and hence, the nuclear cycle. We have empirically found the following formula for an approximation of the nuclear cycle:

$$\log_2(nuclear\ density) + 19.9$$

where "*nuclear density*" is in objects per micron squared.

As the nuclei appear smaller and more densely packed at later nuclear cycles, we used this estimate of the nuclear cycle to determine the radius of the disk used to segment the full image. On images corresponding to nuclear cycles (n.c.) 10-11, we used a radius of 9 pixels, for n.c. 12, 8 pixels, for n.c. 13, 7 pixels, and for n.c. 14 (and images where the nuclear cycle was not determined), a radius of 5 pixels.

The full image was locally background subtracted using a tophat operation with a disk of radius 12 pixels, and then morphologically opened (Figure S2B) with a disk that had a different radius depending on the estimated nuclear cycle (see previous paragraph). The resulting image was then segmented using a regional maxima algorithm (imgregionalmax in the Matlab image processing toolbox). This resulted in a binary image in which the objects correspond to the individual nuclei (Figure S2C).

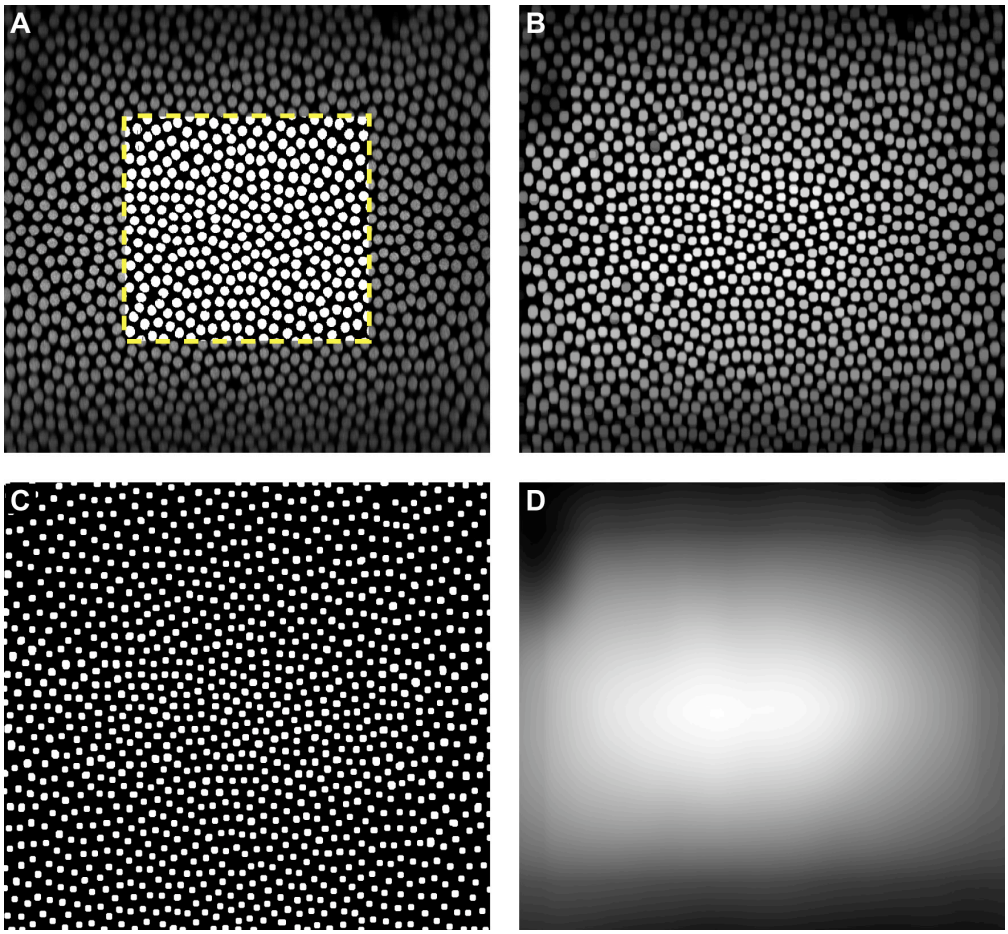


Figure S2: Segmenting the nuclei. (A) Approximate the staging of the embryo. The center of the image (bounded by yellow box) is extracted, and a threshold level is calculated on it. After the threshold is applied, this image becomes binary, allowing for the counting of the objects. (B) Morphologically opened image. This operation removes bright spots smaller than the nuclei. (C) Segmented nuclei. After morphological opening, local regional maxima are counted as nuclei. (D) Calibrating image. This image is effectively an interpolation of the intensity using the nuclei as reference points.

Using standard protocols, the location (centroid) and pixel list of each object were extracted. We measured the nuclear intensity and the Dorsal intensity in each nucleus as the average intensity of all the pixels included in the pixel list of that nucleus. The nuclei with intensities of less than 5% of the most intense nucleus were considered spurious and thrown out. (This 5% number was determined to be high enough such that a set of pixels not corresponding to a real nucleus would be less than this value, but the dimmest nuclei would still be brighter than this value.)

Using the coordinates and intensities of the nuclei, a calibrating image was constructed to interpolate the depth-dependent signal loss across the whole image (including portions of the image that do not contain nuclei; Figure S2D). This calibrating image was smoothened using a sliding window of 100 pixels in the x-direction (AP axis of embryo) and 50 pixels in the y-direction (dorsal-ventral (DV) axis of embryo). This image was then used to normalize the intensity of nuclear Dorsal as well as mRNA distributions.

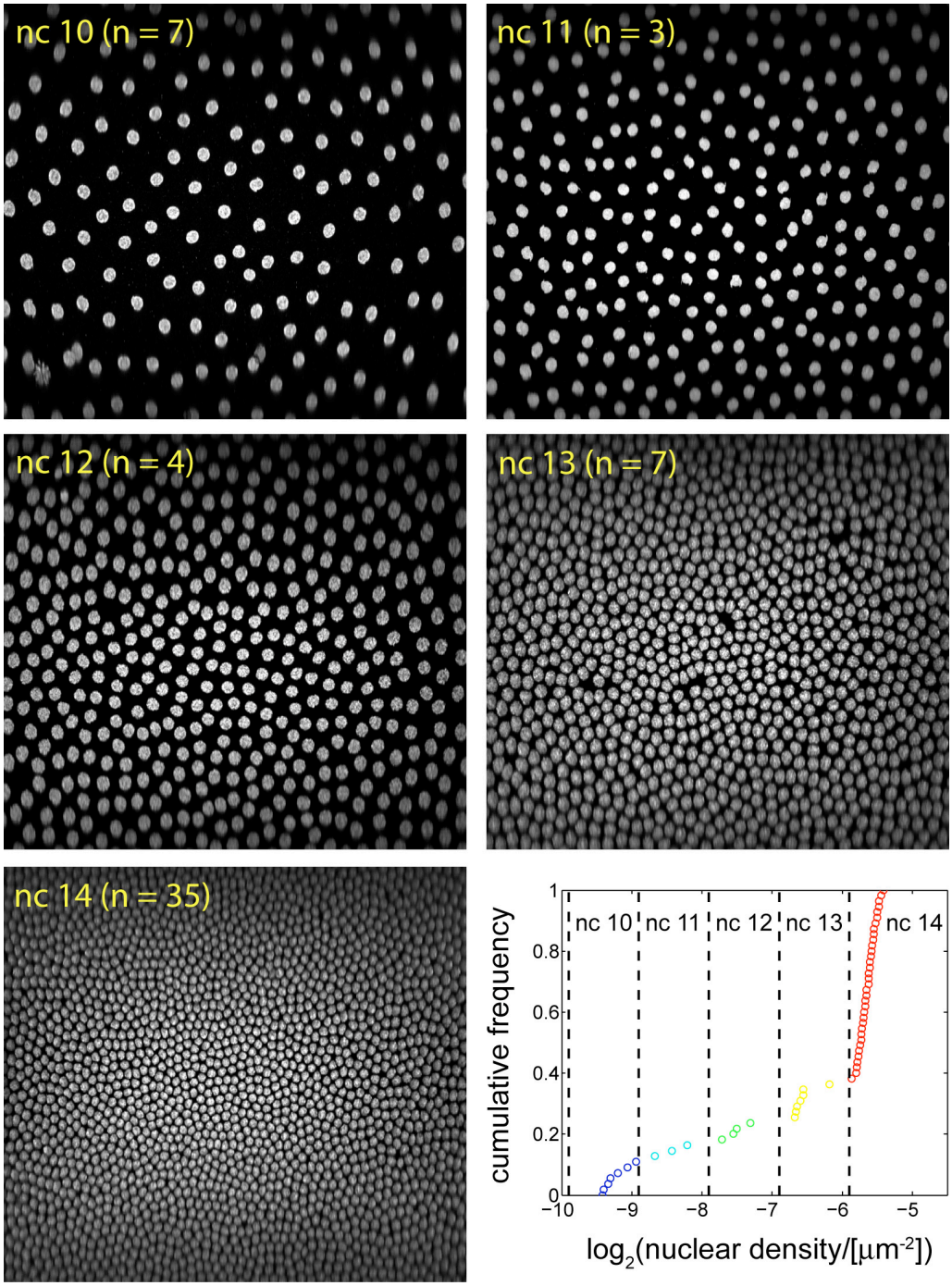


Figure S3: Staging the embryos by nuclear density. Examples of embryos from each of the five nuclear cycles are shown here. The graph depicts the cumulative frequency of nuclear density, showing that the embryos naturally group into the different nuclear cycles, each separated by a power of 2.

3. Using the nuclear density to stage the embryos

After the nuclei have been segmented and the spurious objects discarded, we obtain (among other things) a count of the number of nuclei in the image. Dividing by the area of the image, we can measure the nuclear density in number of objects per square micron. If we rank the nuclear densities of all wild type embryos and plot them on the same graph, we see the

nuclear densities fall into five distinct groups, roughly separated by a factor of two each (Figure S3, bottom right). From examining this graph, we obtain the empirical formula for the nuclear cycle shown in Section 2. This formula allows us to unequivocally determine the nuclear cycle of each embryo. As an example, we have shown an image of a representative embryo from each nuclear cycle (Figure S3).

4. Using the histone H3 antibody intensity for depth correction

In this study, we use the intensity of histone antibody staining to correct for depth-dependent signal loss of Dorsal antibody staining. This approach makes the following assumptions.

(1) The intensity of histone antibody staining is uniform across the embryo, so that changes in signal can be attributed only to light scattering through tissue and other depth-dependent signal attenuation.
(2) The signal loss is similar in all wavelengths. For example, we have chosen to visualize the nuclei with Alexa Fluor 555, Dorsal antibody with Alexa Fluor 488, and mRNA with Alexa Fluor 647. In particular, the Dorsal antibody staining must have signal loss similar to that of the histone staining.
(3) Bleaching is minimal.

If these assumptions hold, then the following equations are valid. The intensity of the histone image, I_{hist}, can be related to the concentration of histone H3, H, as follows:

$$I_{hist} = k(z)H + B_{L,555},$$

where $k(z)$ is an unknown function that describes the depth-dependent signal loss and $B_{L,\lambda}$ denotes laser background, which can be quantified and subtracted. We quantify this laser background (in each chnnel) as the most frequent intensity of any z-slice as it does not change with respect to depth.

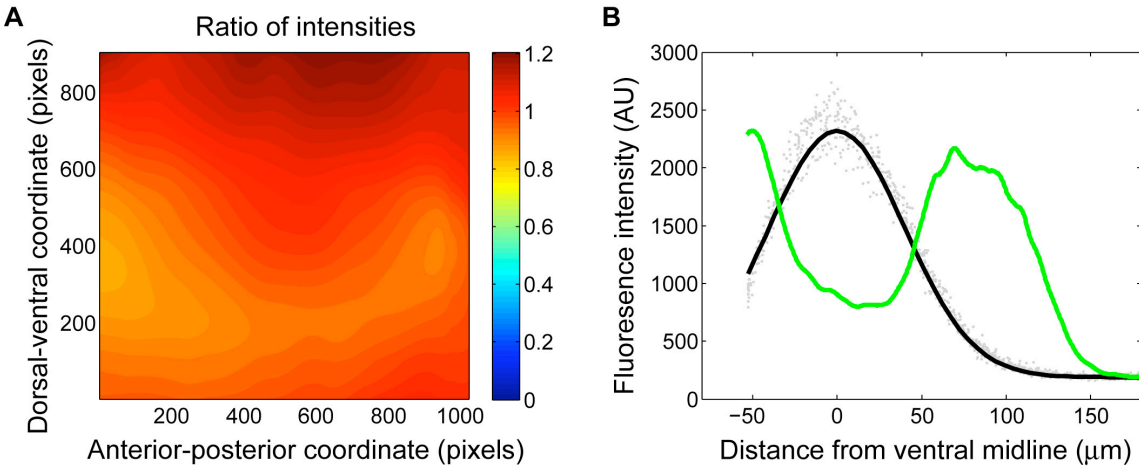


Figure S4: Testing the method of using histone staining to depth-correct. (A) Heatmap of the ratio of the green color channel to the red color channel in an embryo in which both channels visualize histones. In this plot, the normalized "calibrating image" from the green channel is divided by the normalized calibrating image from the red. As they are both normalized before the ratio is taken, we expect the ratio to be close to unity, which it is. (B) Trace of Dorsal nuclear gradient and *sog* mRNA expression from an embryo with the colors used to visualize Dorsal and histones switched. In this case, the calibrating image was generated from the green channel, and Dorsal staining in the red. The gradient appears normal.

The intensity of the Dorsal image, I_{dorsal}, as related to the concentration of Dorsal along the DV coordinate (denoted by $c(x)$ in these equations) is as follows:

$$I_{dorsal} = k(z)(Ac(x) + B) + B_{L,488},$$

where $k(z)$ is the same function that appears in the equation for the histone image (see assumption #2), B is some background level, and A is a proportionality constant.
Subtracting the two laser backgrounds and dividing the second equation by the first, we arrive at:

$$\frac{I_{dorsal}}{I_{hist}} \equiv r = ac(x) + b,$$

where we have defined r to be the ratio between the two intensities. Note that we have replaced the two unknown constants, A and B, with two others, a and b. However, since they were unknown to begin with, the form of the equation is the same in either case. Now we have a quantity, r , that is proportional to the Dorsal concentration, up to an unknown additive constant. If we assume that this background constant is simply due to non-specific antibody staining, then quantifying the value of r in embryos derived from dl^1/dl^1 mothers should, in principle, give us this constant. Even if it is due to factors other than non-specific antibody staining, as long as these factors are equal in both dl^1 embryos and all other embryos, then subtracting the value of r obtained from dl^1 embryos is the correct approach.

Addressing the validity of our assumptions, we take assumption #1 for granted. Assumption #3 remains valid if the laser power used to image the embryos remains relatively low. Assumption #2 was investigated by imaging embryos treated with histone H3 antibody and visualized with Alexa Fluor 488 and 555 to recognize the histone antibody. In these cases, we see that the effect is variable, but more importantly, we find that the normalized ratio of intensities in the two color channels does not greatly deviate from unity in these embryos (Figure S4A). We also imaged embryos in which the Dorsal antibody was visualized with Alexa Fluor 555, and histones with Alexa Fluor 488 (thereby swapping the colors of these two visualizations).

In these embryos, the Dorsal gradients (the values of *r*) also appeared normal, as well as the normalized mRNA profiles (Figure S4B).

5. Calibration of measurements on different days

In order to acquire a data set as large as the one used in the study, it was necessary to image on several different days. To control for day-to-day variations in laser power, we measured the laser output during each imaging session by sending the laser unimpeded into the transmitted light detector. We held all other imaging conditions (detector gains, amplifier offsets, and amplifier gains) constant across all images.

Therefore, for each imaging session and for each laser, we obtained a percent laser power reading, LP_0, and a detection intensity reading, I_0 (Figure S5A). We then expressed the laser output for that imaging session, Ω, in comparison to the "ideal laser" that would give an intensity of 255 for a 1% laser transmittance (with the detector and amplifier settings we were using):

$$\Omega = \frac{1\%}{LP_0}\frac{I_0}{255}.$$

For example, if, during a given imaging session, the 488 laser output measurement gives an intensity I_0 =255 at a percent laser transmittance of LP_0=5%, then the current laser output is 5 times weaker than the "ideal" laser, which would achieve an intensity of 255 at only 1% laser transmittance.

For each embryo imaged, the percent laser transmittance used to image that embryo, *LP*, was tuned to reveal the greatest imaging dynamic range possible, without raising the laser power so much that bleaching becomes a problem. We then defined a "reduced laser power" for each embryo in units of the "ideal laser":

$$LP_r \equiv \Omega\, LP = LP\frac{1\%}{LP_0}\frac{I_0}{255}.$$

These definitions allowed us to use a consistent measure for the incident laser power used to image each embryo.

In order for this approach to be valid, however, two other functions must be measured. First, we must be sure that the actual laser power transmitted is linear with respect to the percent laser power parameter tunable from the LSM 5 Pascal software. Our microscope system uses an MOTF (mechano-optical tunable filter) to change the percent laser power transmitted to the specimen. Therefore, the relationship between the tunable parameter on the software and the position of the MOTF (a mechanical shutter that can selectively block a fraction of the laser light) can be calibrated to be linear.

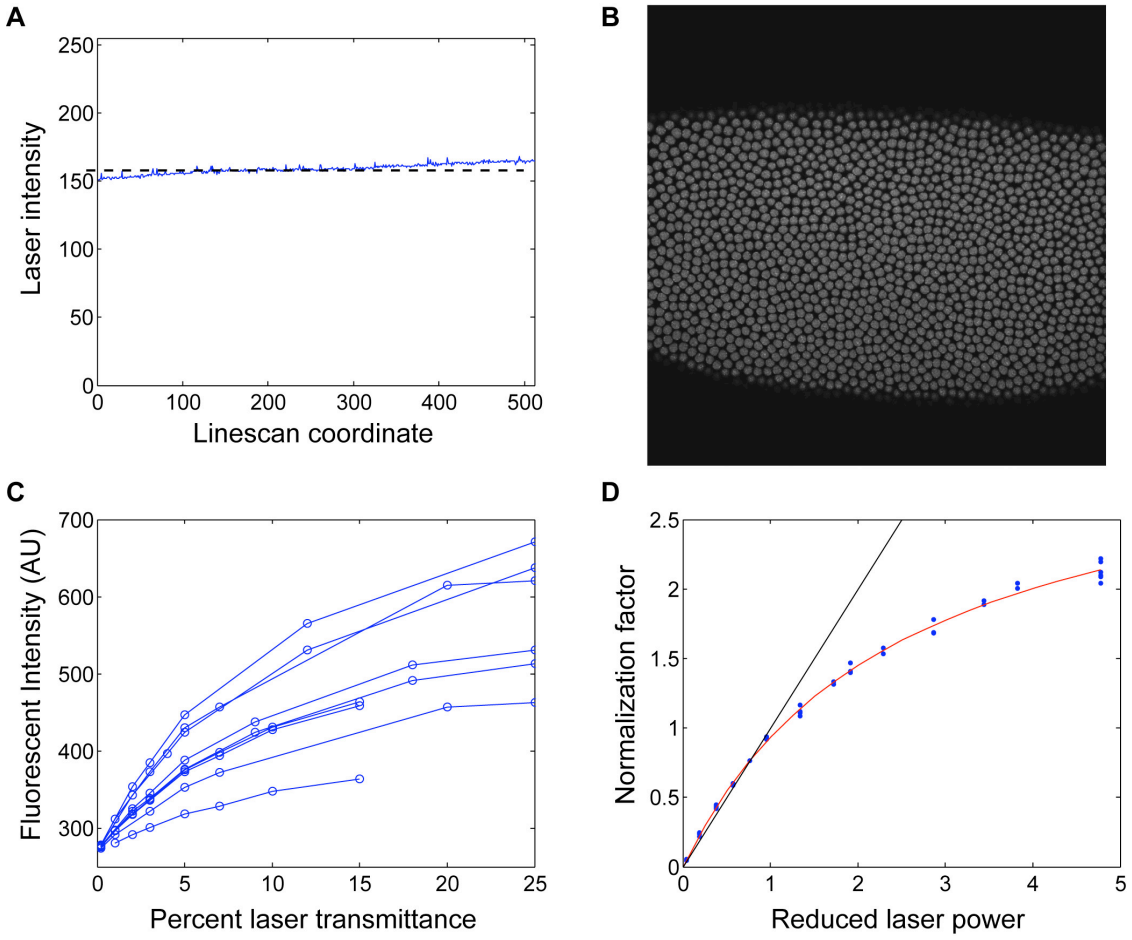


Figure S5: Calibration of measurements on different days. (A) Typical linescan performed for each laser during each imaging session. The average laser intensity across the entire linescanned "image" (dotted line) is taken to be I_0 for that day. (B) An example of a surface image of an embryo used to determine the relationship between incident laser strength and emission intensity. (C) Dependence of fluorescent intensity on percent laser tranmittance of the incident laser for the nine embryos used in this calibration. The differences can be explained by different average concentrations of the Alexa Fluor dye within each embryo. (D) When normalized properly, the nine traces in (C) collapse onto one curve (blue dots, data points). This curve is empirically fit to a saturating hyperbolic (red curve). The black line corresponds to $y = x$, showing that the dependence of the normalization factor on reduced laser power is linear in regimes near zero.

The second function that must be measured is the relationship between the incident laser power on the embryo and the fluorescence emission of the Alexa Fluor dyes. Others have reported that this is a linear relationship (1), however this was only tested at very low laser powers. In some regimes of laser power used for our study, this linear relationship breaks down (Figure S5D).

To determine the shape of this function, we imaged one slice on the surface of nine different embryos several times using different values of percent laser transmittance for the 488 laser (Figure S5B). For each embryo, this revealed a relationship between mean intensity of the slice and the tunable parameter (percent laser transmittance) in the LSM 5 Pascal software (Figure S5C). However, note that this relationship depends on the actual density of Alexa Fluor

dye in the embryo, and thus will be different from embryo-to-embryo. Also note that these functions are linear at low values of the laser power, and have a non-zero background level. That is, fitting a line to the points with laser power 5% or less, the y-intercept is not zero. However, this is the same laser background that was discussed earlier, and can easily be measured and subtracted from the signal values. Therefore, performing linear regression on data points with low laser values, while forcing the intercept to be the measured laser background, we can estimate the low-laser power behavior of these incident laser/emission functions for each embryo as a slope, *m*:

$$I \approx m\,LP + B_L.$$

Therefore, if we normalize the emission intensity of each embryo by the estimated slope for that embryo, then the incident laser/emission curves for all embryos collapse onto one curve (Figure S5C). We find that the relationship at these laser power levels can be approximated by a saturating hyperbolic:

$$\bar{I} \approx V\frac{LP_r}{K + LP_r},$$

where the reduced laser power, LP_r, was introduced above, and $\bar{I}$ is the background subtracted, slope-normalized emission intensity in units of the "ideal laser", or:

$$\bar{I} \equiv \frac{I - B_L}{m}\Omega.$$

We found that, for our microscope settings, the best-fit values of the parameters are *V=3.25* and *K=2.5* (red curve in Figure S5D).

In practical terms, this "$\bar{I}$" is a normalization factor. We can this calculate the normalization factor for each embryo imaged in this paper, given (1) the reduced laser power used to image that embryo, and (2) the saturating hyperbolic equation above. Then, the data (*r*) for that embryo is normalized by $\bar{I}$. This allows us to account for variations in laser output during different imaging sessions.

6. Fitting wild type Dorsal gradients to Gaussian-shaped curves

After measuring many wild type Dorsal gradients, it was noted that each appeared to be roughly bell-shaped. Therefore, we attempted to fit each to a Gaussian-like curve in order to fit global parameters to the curve. This was motivated by the fact that a gradient may have several different length scales associated with it. For example, how does one precisely and consistently measure the "width" of a Dorsal gradient? Is it the width of the gradient at half-maximal? Even small amounts of noise can give a drastic error in such a measurement. However, if the gradient always retains the rough shape of a known function of *x*, then the entire curve can be used to estimate some length scale (such as a width), rather than some arbitrary point half-way to the maximum of the curve. The same can be said of the amplitude and basal levels of a gradient, as well as the presumptive ventral midline: is the amplitude simply the highest data point? Are the basal levels simply the minimum data point? Is the ventral midline the location where the highest data point occurs? All of these would be easily distorted by only a very small amount of noise. While there are ways to minimize the sensitivity to noisy data, one way to solve these problems is by fitting the entire gradient to a known function of *x*, in this case a Gaussian shape:

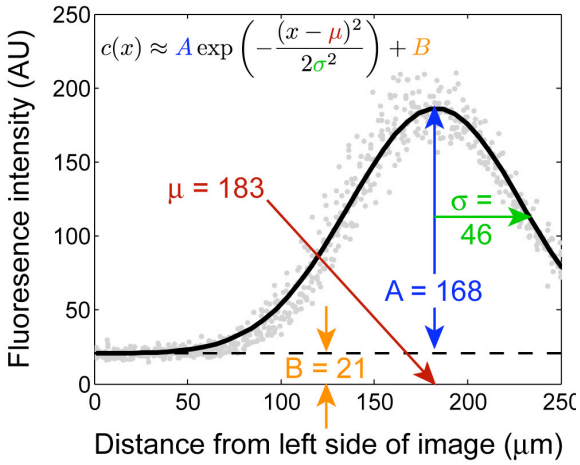


Figure S6: Illustration of fitting the Dorsal gradients to Gaussian curves. This fitting process is used to globally extract four quantities that characterize the Dorsal gradient: amplitude (A), basal levels (B), location of the peak (μ), and signal decay length (σ). Note that, for a Gaussian-shaped curve, the signal decay length is related to the width of the curve. After 60% decay, the width of the Gaussian curve is equal to 2σ.

$$c(x) \approx A\, exp\left(-\tfrac{1}{2}(x-\mu)^2/\sigma^2\right) + B.$$

Here we see that after fitting the Dorsal gradient to this equation, we extract four parameters that describe each gradient: an amplitude, *A*, basal levels, *B*, the location of the presumptive ventral midline, μ, and a length scale of the gradient, σ, which we have often called the gradient "width," but it is more accurately the length scale of signal decay (Figure S6). Note that, at this point, the similarity between the wild type Dorsal gradient and a Gaussian-shaped curve is strictly empirical. We are not proposing a physical mechanism that would dictate this shape to be Gaussian. We used the Matlab function, fit, and used the 'NonlinearLeastSquares' option to perform this fit as well as all others in this study.

7. Measuring domains of gene expression

The motivation behind fitting the Dorsal gradient to a Gaussian-like function was also present in our attempts to characterize the domains of gene expression of sog, *vnd*, and *ind*. For each of these genes, we found the "canonical" expression profile, or shape, by aligning and averaging several wild type expression profiles (Fig S7A). After these canonical shapes were found, we fit a given gene expression profile to the appropriate shape in a manner similar to what was done for the wild type Dorsal gradient.

For example, if the canonical *sog* profile (green curve in Figure S7A) were defined as $sog_0(x)$, then any *sog* expression domain could be fit to this canonical profile by the following equation:

$$sog(x) \approx A\, sog_0\left(\frac{x-\mu}{\delta}\right) + B,$$

where *A* and *B* are the amplitude and background levels, μ is the location of the presumptive "center" – in our case, we have chosen this to be the maximum – of the peak, and δ is a "stretching factor" that defines how wide or narrow the individual gene expression domain is with respect to the canonical form (Figure S7B). If $\delta > 1$, then the individual profile is wider than the canonical form, and if $\delta < 1$, then it is narrower. The data points that were used to fit each of these expression domains were the original gene expression profiles locally background subtracted. The width of the structuring element used to subtract the local background was chosen to be large enough to not disturb the overall shape of these gene expression domains. The examples for *vnd* and *ind* are similar, and can be found in Figure S7C,D, respectively.

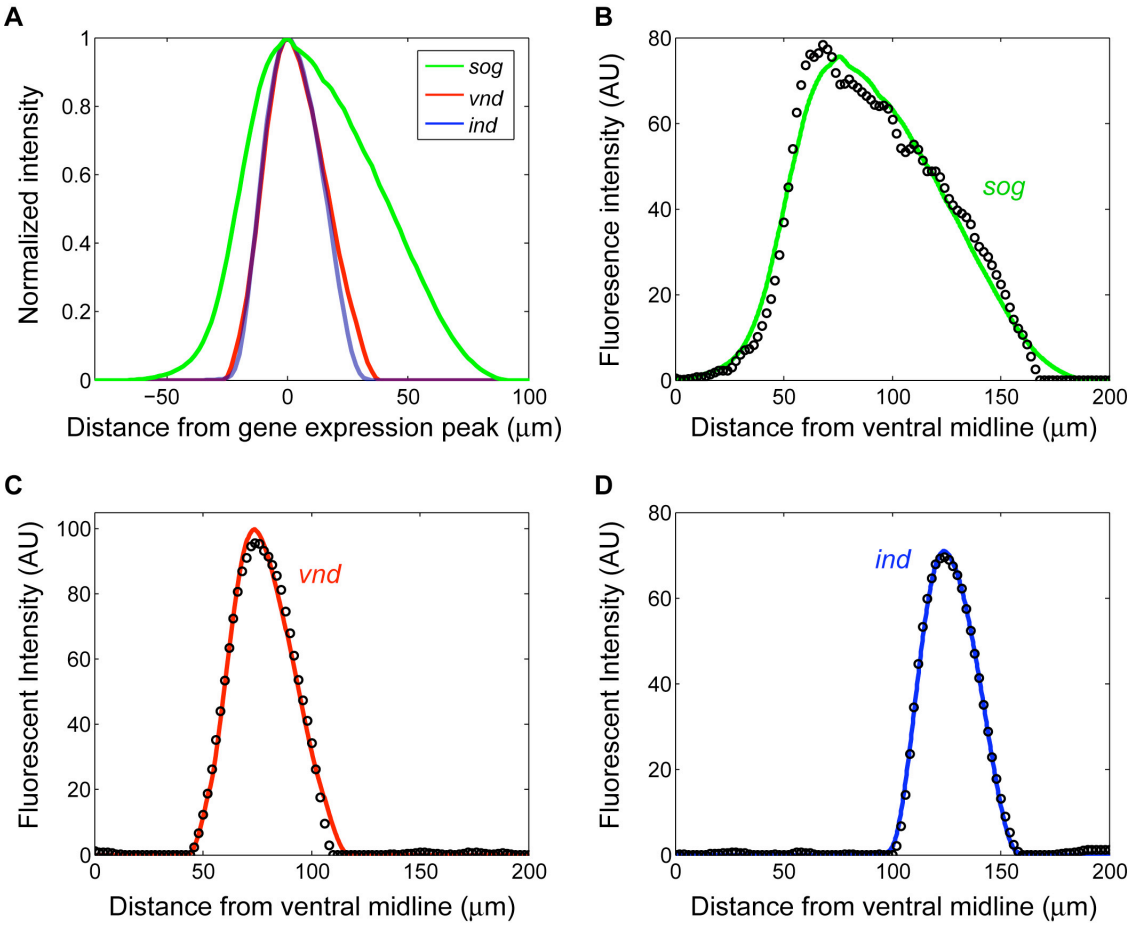


Figure S7: Fitting individual gene expression profiles to average, canonical data. (A) The canonical gene expression profiles for *sog* (green), *vnd* (red), and *ind* (blue). The x-axis denotes the distance, in microns, from the "center" of the peak. We have defined the center of the peak to refer to the location of the maximum. (B-D) Examples of fitting *sog* (B), *vnd* (C), and *ind* (D) expression patterns (black dots) to the canonical shapes (solid curves).

In a similar fashion, we fit the dl^1/CyO gradient to a canonical shape found by averaging (see Figure S8A), and for each of those embryos, we also found the parameters A, B, μ and δ. Here, the parameters $A, B,$ and μ are directly analogous to those found for the wild type and Dorsal-GFP gradients. On the other hand, δ is related to the parameter σ that characterizes the width of the wild type gradient, but is not directly analogous (see next section).

8. Statistical analysis of Dorsal gradients and mRNA expression patterns

We performed several statistical analyses on the parameters extracted from the fitting procedures described in the previous two sections. The analyses quoted in the main portion of the paper are as follows:

(1) ANOVA on the widths of the wild type gradients, grouped by nuclear cycle.
(2) *t*-test on the location of *sog* expression in dl^1/CyO versus wild type.
(3) *t*-test on the width of *sog* expression in dl^1/CyO versus wild type.
(4) *t*-test on the steepness of the Dorsal gradient in dl^1/CyO versus wild type.
(5) *t*-test on the width of the Dorsal gradient in *dl-gfp* versus wild type.
(6) *t*-test on the amplitude of the Dorsal gradient in *dl-gfp* versus wild type.
(7) *t*-test on the width of *sog* expression in *dl-gfp* versus wild type.

In test (1), we tested whether the value of σ from any nuclear cycle would be different from the others. Using ANOVA, we concluded that the value of σ remained constant throughout development (*p*-value: 0.3). However, we found that embryos from *dl-gfp* mothers had significantly wider gradients (test (5)). We performed a modified *t*-test (2), allowing for distinct samples sizes and distribution variances (all t-tests were performed with these relaxed assumptions), with the null hypothesis that the *dl-gfp* σ was not larger than that of wild type (one-tailed test), and found the *p*-value to be 0.05. Furthermore, we tested whether the amplitudes of Dorsal gradients in *dl-gfp* embryos would be larger than that of wild type (test (6), one-tailed test), and found that to be the case (*p*-value: 0.003).

In tests (2), (3), and (7), we compared the properties of *sog* mRNA expression patterns from either dl^1/CyO embryos or *dl-gfp* embryos to wild type. We asked whether these properties differed significantly from wild type (two-tailed tests), and only found the width of *sog* expression in *dl-gfp* to be distinct from the corresponding wild type value (*p*-value: 0.0006). We also conclude that the *location* of *sog* expression in *dl-gfp* is indistinguishable from wild type (not quoted in main paper; *p*-value: 0.5). When we speak of "location" of the mRNA profile, we are describing the parameter μ, as defined in the previous section. Note that, for the *sog* mRNA profile, the "location" is skewed to the ventral side of the profile. Thus, we conclude that, in *dl-gfp* embryos, the *sog* profile is mostly widened in the dorsal direction, with a similar ventral border to wild type (also see Figure 5 from main paper).

In test (4), we asked whether the length scale of signal decay (or "steepness") of the Dorsal gradient found in dl^1/CyO embryos would significantly differ from that found in wild type embryos. Because the shape of the gradient is non-Gaussian, we are only interested in whether the steepness is maintained through the "important" part of the gradient, that is, in the presumptive neuroectoderm (50-90 μm from the ventral midline). As mentioned in the previous section, the value of σ, which we used to characterize the width of the wild type gradients, is not directly comparable to δ – the parameter used to characterize the widths of gradients from heterozygous animals. Therefore, we used δ to approximate a value corresponding to σ for the heterozygous embryos in the following manner.

First, we note that the wild type Dorsal gradient is approximately Gaussian in shape. From the equation above, if we define *y* as:

$$y \equiv \frac{c(x) - B}{A},$$

and we define *z* as:

$$z \equiv x - \mu,$$

then the equation for the Dorsal gradient can be transformed into:

$$-\log y^2 = \left(\frac{z}{\sigma}\right)^2.$$

Taking the derivative with respect to z^2 and rearranging, we obtain:

$$\sigma = -\left(\frac{d(\log y^2)}{d(z^2)}\right)^{-1}.$$

This is an identity for Gaussian shaped curves, and thus holds true at any point *x* (meaning, this derivative is constant and equal to the parameter σ , no matter where you are on the curve). On the other hand, there is no reason why the shape of the Dorsal gradient from dl^1/CyO embryos should maintain this sort of property. However, we can calculate this derivative for every point within the presumptive neuroectoderm for dl^1/CyO embryos and determine what the value of σ "should be" if that curve were indeed Gaussian.

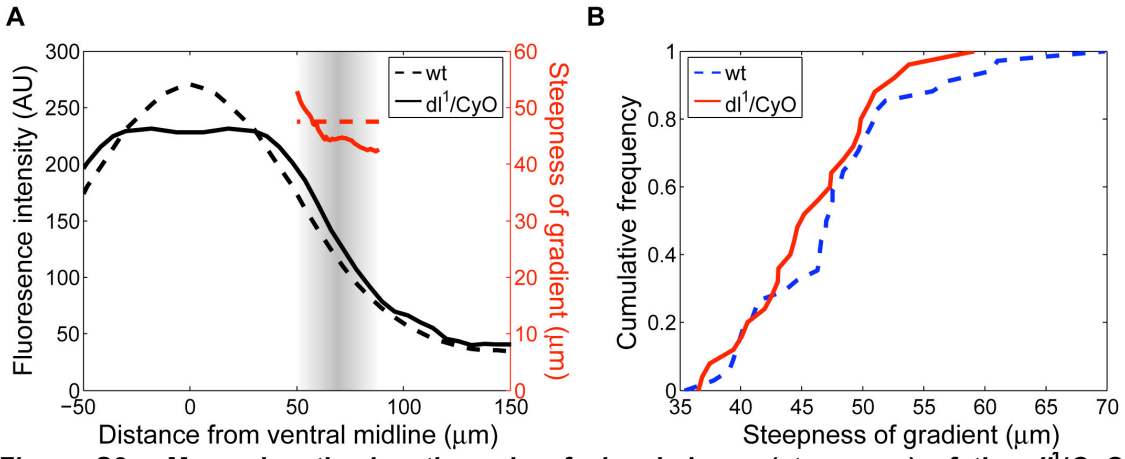


Figure S8: Measuring the length scale of signal decay (steepness) of the *dl*^{1}/*CyO* gradient within the neuroectoderm. (A) Definition of the steepness. In black, the average wild type (dashed) and *dl*^{1}/*CyO* (solid) n.c. 14 gradients are simultaneously plotted (left axis) against the DV coordinate. In red, the measure of the steepness for wild type (dashed) and *dl*^{1}/*CyO* (solid) Dorsal nuclear gradients are plotted (right axis) as functions of the DV coordinate. Note that the steepness of the wild type gradient is constant, while our measure of the steepness of the *dl*^{1}/*CyO* gradient varies with position. (B) Distributions of the steepness of the Dorsal nuclear gradient in the two genotypes. Note that the two distributions are nearly the same.

Therefore, we took this derivative for the average, "canonical" heterozygous Dorsal gradient (solid black curve in Figure S8A) within the presumptive neurogenic ectoderm (gray region in Figure S8A) to obtain a putative value of σ at each location, x (solid red curve in Figure S8A). Compare this to the steepness of average wild type Dorsal gradient (dashed black curve in Figure S8A), which is constant with respect to x (red dashed horizontal line in Figure S8A). While the steepness of the heterozygous gradient varies slightly in this region of the embryo, it is quite close to what we would expect it to be were it a wild type gradient. This makes sense, as the two gradients in this region appear very similar. We conclude that the median value of the gradient steepness in this region, 44.4 μm, is sufficient to characterize the changing value of σ for this average heterozygous gradient.

After assigning this value to the "steepness" of the neurogenic ectoderm region of the canonical heterozygous gradient, we can apply this calculation to each of the individual gradients through the value of the fitted parameter δ. As each embryo i has a different value of this stretching factor, δ_i, we can simply "stretch" the value of σ accordingly:

$$\sigma_i = \sigma_{avg}\delta_i,$$

where σ_i is the value of the steepness for embryo i, and σ_{avg} is the steepness for the average heterozygous gradient (equal to 44.4 μm). Plotting the distribution of σ_i's for both wild type and heterozygous embryos, we see that they are very similar (Figure S8B). Indeed, performing the t-test on these two populations shows that they cannot be distinguished (p-value 0.2).

Finally, although not quoted in the main paper, we also performed ANOVA on the amplitudes and basal levels of the wild type gradients, grouped by nuclear cycle. We found that both of these variables have significant differences among the nuclear cycles (p-values 0.0009 and 7×10^{-9}, respectively). However, this is plain to see from Figure 3D in the main paper.

9. Sequencing Dorsal and Dorsal-GFP fusion.

Below are the sequences for Dorsal-gfp protein fusion with the deleted N terminus in **BOLD** (used by DeLotto et al., 2007)and the full length Dorsal protein translation respectively. The Nuclear Export Sequences (NES 1-4) (defined by Xylourgidis et al., 2006) are highlighted in BLUE in the Dorsal protein translation below. RED letters denote the nuclear localization sequence (NLS).

MFPNQNNGAAPGQGPAVDGQQSLNYNGLPAQQQQQLAQSTKNVRKKPYVK
ITEQPAGKALRFRYECEGRSAGSIPGVNSTPENKTYPTIEIVGYKGRAVV
VVSCVTKDTPYRPHPHNLVGKEGCKKGVCTLEINSETMRAVFSNLGIQCV
KKKDIEAALKAREEIRVDPFKTGFSHRFQPSSIDLNSVRLCFQVFMESEQ
KGRFTSPLPPVVSEPIFDKKAMSDLVICRLCSCSATVFGNTQIILLCEKV
AKEDISVRFFEEKNGQSVWEAFGDFQHTDVHKQTAITFKTPRYHTLDITE
PAKVFIQLRRPSDGVTSEALPFEYVPMDSDPAHLRRKRQKTGGDPMHLLL
QQQQKQQLQNDHQDGRQTNMNCWNTQNIPPIKTEPRDTSPQPFGLSYRAP
PELTPSPQPLSPSSNYNHNSTPSPYNMASAVTPTNGQQQLMSPNHPQQQQ
QQQQYGATDLGSNYNPFAQQVLAQQQQHQQQQQQHQHQHQQQHQQQQQQQ
QQQQQQQSLQFHANPFGNPGGNSWESKFSAAAVAAAAATATGAAPANGNS
NNLSNLNNPFTMHNLLTSGGGPGNANNLQWNLTTNHLHNQHTLHQQQQLQ
QQQQQQYDNTAPTNNNANLNNNNNNNNTAGNQADNNGPTLSNLLSFDSGQ
LVHINSEDQQILRLNSEDL**H-**GFP PROTEIN

Dorsal protein sequence:

MFPNQNNGAAPGQGPAVDGQQSLNYNGLPAQQQQQLAQSTKNVRKKPYVK
ITEQPAGKALRFRYECEGRSAGSIPGVNSTPENKTYPTIEIVGYKGRAVV
VVSCVTKDTPYRPHPHNLVGKEGCKKGVCTLEINSETMRAVFSNLGIQCV
KKKDIEAALKAREEIRVDPFKTGFSHRFQPSSIDLNSVRLCFQVFMESEQ
KGRFTSPLPPVVSEPIFDKKAMSDLVICRLCSCSATVFGNTQIILLCEKV
AKEDISVRFFEEKNGQSVWEAFGDFQHTDVHKQTAITFKTPRYHTLDITE
PAKVFIQLRRPSDGVTSEALPFEYVPMDSDPAHLRRKRQKTGGDPMHLLL
QQQQKQQLQNDHQDGRQTNMNCWNTQNIPPIKTEPRDTSPQPFGLSYRAP
PELTPSPQPLSPSSNYNHNSTPSPYNMASAVTPTNGQQQLMSPNHPQQQQ
QQQQYGATDLGSNYNPFAQQVLAQQQQHQQQQQQHQHQHQQQHQQQQQQQ
QQQQQQSLQFHANPFGNPGGNSWESKFSAAAVAAAAATATGAAPANGNSN
NLSNLNNPFTMHNLLTSGGGPGNANNLQWNLTTNHLHNQHTLHQQQQLQQ
QQQQQYDNTAPTNNNANLNNNNNNNNTAGNQADNNGPTLSNLLSFDSGQL
VHINSEDQQILRLNSEDL**QISNLSIST**

Bibliography

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., et al., 2000. The genome sequence of Drosophila melanogaster. Science. 287, 2185-95.

Ashe, H. L., Briscoe, J., 2006. The interpretation of morphogen gradients. Development. 133, 385-94.

Baeuerle, P. A., Baltimore, D., 1996. NF-kappa B: ten years after. Cell. 87, 13-20.

Bailey, A. M., Posakony, J. W., 1995. Suppressor of hairless directly activates transcription of enhancer of split complex genes in response to Notch receptor activity. Genes Dev. 9, 2609-22.

Baugh, L. R., Hill, A. A., Claggett, J. M., Hill-Harfe, K., Wen, J. C., Slonim, D. K., Brown, E. L., Hunter, C. P., 2005. The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the C. elegans embryo. Development. 132, 1843-54.

Belvin, M. P., Anderson, K. V., 1996. A conserved signaling pathway: the Drosophila toll-dorsal pathway. Annu Rev Cell Dev Biol. 12, 393-416.

Bergmann, A., Stein, D., Geisler, R., Hagenmaier, S., Schmid, B., Fernandez, N., Schnell, B., Nüsslein-Volhard, C., 1996. A gradient of cytoplasmic Cactus degradation establishes the nuclear localization gradient of the dorsal morphogen in Drosophila. Mech Dev. 60, 109-23.

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., Eisen, M. B., 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci USA. 99, 757-62.

Bischof, J., Maeda, R. K., Hediger, M., Karch, F., Basler, K., 2007. An optimized transgenesis system for Drosophila using germ-line-specific {varphi}C31 integrases. Proceedings of the National Academy of Sciences. 104, 3312.

Bolouri, H., 2008. Embryonic pattern formation without morphogens. Bioessays. 30, 412-7.

Brown, C. D., Johnson, D. S., Sidow, A., 2007. Functional architecture and evolution of transcriptional elements that drive gene coexpression. Science. 317, 1557-60.

Brown, C. T., Xie, Y., Davidson, E. H., Cameron, R. A., 2005. Paircomp, FamilyRelationsII and Cartwheel: tools for interspecific sequence comparison. BMC Bioinformatics. 6, 70.

Carneiro, K., Fontenele, M., Negreiros, E., Lopes, E., Bier, E., Araujo, H., 2006. Graded maternal short gastrulation protein contributes to embryonic dorsal-ventral patterning by delayed induction. Dev Biol. 296, 203-18.

Carroll, S. B., Winslow, G. M., Twombly, V. J., Scott, M. P., 1987. Genes that control dorsoventral polarity affect gene expression along the anteroposterior axis of the Drosophila embryo. Development. 99, 327-32.

Chen, G., Handel, K., Roth, S., 2000. The maternal NF-kappaB/dorsal gradient of Tribolium castaneum: dynamics of early dorsoventral patterning in a short-germ beetle. Development. 127, 5145-56.

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., et al., 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 450, 203-18.

Cowden, J., Levine, M., 2003. Ventral dominance governs sequential patterns of gene expression across the dorsal-ventral axis of the neuroectoderm in the Drosophila embryo. Dev Biol. 262, 335-49.

Davidson, E., 2006. The regulatory genome : gene regulatory networks in development and evolution.

Davidson, E. H., 2001. Genomic regulatory systems : development and evolution. Academic Press, San Diego, Calif.

De Renzis, S., Elemento, O., Tavazoie, S., Wieschaus, E. F., 2007. Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. PLoS Biol. 5, e117.

DeLotto, R., DeLotto, Y., Steward, R., Lippincott-Schwartz, J., 2007. Nucleocytoplasmic shuttling mediates the dynamic maintenance of nuclear Dorsal levels during Drosophila embryogenesis. Development. 134, 4233-41.

Deplancke, B., Mukhopadhyay, A., Ao, W., Elewa, A. M., Grove, C. A., Martinez, N. J., Sequerra, R., Doucette-Stamm, L., Reece-Hoyes, J. S., Hope, I. A., 2006. A Gene-Centered C. elegans Protein-DNA Interaction Network. Cell. 125, 1193.

Driever, W., Thoma, G., Nüsslein-Volhard, C., 1989. Determination of spatial domains of zygotic gene expression in the Drosophila embryo by the affinity of binding sites for the bicoid morphogen. Nature. 340, 363-7.

Erdelyi, M. S., J., 1989. Isolation and characterization of dominant female sterile mutations of Drosophila melanogaster. I. Mutations on the third chromosome. Genetics. 122, 111-27.

Foe, V. E., Alberts, B. M., 1983. Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in Drosophila embryogenesis. Journal of cell science. 61, 31-70.

Francois, V., Solloway, M., O'Neill, J. W., Emery, J., Bier, E., 1994. Dorsal-ventral patterning of the Drosophila embryo depends on a putative negative growth factor encoded by the short gastrulation gene. Genes Dev. 8, 2602-16.

Gerttula, S., Jin, Y. S., Anderson, K. V., 1988. Zygotic expression and activity of the Drosophila Toll gene, a gene required maternally for embryonic dorsal-ventral pattern formation. Genetics. 119, 123-33.

Goltsev, Y., Fuse, N., Frasch, M., Zinzen, R. P., Lanzaro, G., Levine, M., 2007. Evolution of the dorsal-ventral patterning network in the mosquito, Anopheles gambiae. Development. 134, 2415-24.

González-Crespo, S., Levine, M., 1993. Interactions between dorsal and helix-loop-helix proteins initiate the differentiation of the embryonic mesoderm and neuroectoderm in Drosophila. Genes Dev. 7, 1703-13.

Groth, A. C., Fish, M., Nusse, R., Calos, M. P., 2004. Construction of transgenic Drosophila by using the site-specific integrase from phage phiC31. Genetics. 166, 1775-82.

Guhathakurta, D., Schriefer, L. A., Hresko, M. C., Waterston, R. H., Stormo, G. D., 2002. Identifying muscle regulatory elements and genes in the nematode Caenorhabditis elegans. Pac Symp Biocomput. 425-36.

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

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

Huang, J. D., Schwyter, D. H., Shirokawa, J. M., Courey, A. J., 1993. The interplay between multiple enhancer and silencer elements defines the pattern of decapentaplegic expression. Genes Dev. 7, 694-704.

Hunt-Newbury, R., Viveiros, R., Johnsen, R., Mah, A., Anastas, D., Fang, L., Halfnight, E., Lee, D., Lin, J., Lorch, A., et al., 2007. High-throughput in vivo analysis of gene expression in Caenorhabditis elegans. PLoS Biol. 5, e237.

Inoue, J., Kerr, L. D., Rashid, D., Davis, N., Bose, H. R., Verma, I. M., 1992. Direct association of pp40/I kappa B beta with rel/NF-kappa B transcription factors: role of ankyrin repeats in the inhibition of DNA binding activity. Proc Natl Acad Sci USA. 89, 4333-7.

Ip, Y. T., Kraut, R., Levine, M., Rushlow, C. A., 1991. The dorsal morphogen is a sequence-specific DNA-binding protein that interacts with a long-range repression element in Drosophila. Cell. 64, 439-46.

Ip, Y. T., Park, R. E., Kosman, D., Bier, E., Levine, M., 1992a. The dorsal gradient morphogen regulates stripes of rhomboid expression in the presumptive …. Genes Dev.

Ip, Y. T., Park, R. E., Kosman, D., Yazdanbakhsh, K., Levine, M., 1992b. dorsal-twist interactions establish snail expression in the presumptive mesoderm of the Drosophila embryo. Genes Dev. 6, 1518-30.

Jazwinska, A., Kirov, N., Wieschaus, E., Roth, S., Rushlow, C., 1999. The Drosophila gene brinker reveals a novel mechanism of Dpp target gene regulation. Cell. 96, 563-73.

Jiang, J., Kosman, D., Ip, Y. T., Levine, M., 1991. The dorsal morphogen gradient regulates the mesoderm determinant twist in early Drosophila embryos. Genes and Development. 5, 1881.

Jiang, J., Levine, M., 1993. Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. Cell. 72, 741-52.

Jiang, J., Rushlow, C. A., Zhou, Q., Small, S., Levine, M., 1992. Individual dorsal morphogen binding sites mediate activation and repression in the Drosophila embryo. EMBO J. 11, 3147-54.

Kasai, Y., Nambu, J. R., Lieberman, P. M., Crews, S. T., 1992. Dorsal-ventral patterning in Drosophila: DNA binding of snail protein to the single-minded gene. Proceedings of the National Academy of Sciences of the United States of America. 89, 3414-8.

Kasai, Y., Stahl, S., Crews, S., 1998. Specification of the Drosophila CNS midline cell lineage: direct control of single-minded transcription by dorsal/ventral patterning genes. Gene expression. 7, 171-89.

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

Kerszberg, M., Wolpert, L., 2007. Specifying positional information in the embryo: looking beyond morphogens. Cell. 130, 205-9.

Kirov, N., Childs, S., O'Connor, M., Rushlow, C., 1994. The Drosophila dorsal morphogen represses the tolloid gene by interacting with a silencer element. Molecular and cellular biology. 14, 713-22.

Klingler, M., Erdélyi, M., Szabad, J., Nüsslein-Volhard, C., 1988. Function of torso in determining the terminal anlagen of the Drosophila embryo. Nature. 335, 275-7.

Kosman, D., Ip, Y. T., Levine, M., Arora, K., 1991. Establishment of the mesoderm-neuroectoderm boundary in the Drosophila embryo. Science. 254, 118-22.

Lee, Y. M., Park, T., Schulz, R. A., Kim, Y., 1997. Twist-mediated activation of the NK-4 homeobox gene in the visceral mesoderm of Drosophila requires two distinct clusters of E-box regulatory elements. J Biol Chem. 272, 17531-41.

Lehmann, R., Nüsslein-Volhard, C., 1991. The maternal gene nanos has a central role in posterior pattern formation of the Drosophila embryo. Development. 112, 679-91.

Levis, R., Hazelrigg, T., Rubin, G. M., 1985. Effects of genomic position on the expression of transduced copies of the white gene of Drosophila. Science. 229, 558-61.

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

Liang, H. L., Nien, C. Y., Liu, H. Y., Metzstein, M. M., Kirov, N., Rushlow, C., 2008. The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. Nature. 456, 400-3.

Liberman, L. M., Stathopoulos, A., 2008. Design flexibility in cis-regulatory control of gene expression: Synthetic and comparative evidence. Dev Biol.

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

Markstein, M., Levine, M., 2002. Decoding cis-regulatory DNAs in the Drosophila genome. Curr Opin Genet Dev. 12, 601-6.

Markstein, M., Markstein, P., Markstein, V., Levine, M. S., 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proceedings of the National Academy of Sciences. 99, 763.

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

Markstein, M., Zinzen, R., Markstein, P., Yee, K. P., Erives, A., Stathopoulos, A., Levine, M., 2004. A regulatory code for neurogenic gene expression in the Drosophila embryo. Development. 131, 2387-94.

Mauhin, V., Lutz, Y., Dennefeld, C., Alberga, A., 1993. Definition of the DNA-binding site repertoire for the Drosophila transcription factor SNAIL. Nucleic Acids Res. 21, 3951-7.

McDonald, J. A., Holbrook, S., Isshiki, T., Weiss, J., Doe, C. Q., Mellerick, D. M., 1998. Dorsoventral patterning in the Drosophila central nervous system: the vnd homeobox gene specifies ventral column identity. Genes Dev. 12, 3603-12.

Mizutani, C. M., Meyer, N., Roelink, H., Bier, E., 2006. Threshold-dependent BMP-mediated repression: a model for a conserved mechanism that patterns the neuroectoderm. PLoS Biol. 4, e313.

Morisato, D., Anderson, K. V., 1995. Signaling pathways that establish the dorsal-ventral pattern of the Drosophila embryo. Annu Rev Genet. 29, 371-99.

Moussian, B., Roth, S., 2005. Dorsoventral axis formation in the Drosophila embryo--shaping and transducing a morphogen gradient. Curr Biol. 15, R887-99.

Muller, B., Hartmann, B., Pyrowolakis, G., Affolter, M., Basler, K., 2003. Conversion of an extracellular Dpp/BMP morphogen gradient into an inverse transcriptional gradient. Cell. 113, 221-33.

Murre, C., Bain, G., van Dijk, M. A., Engel, I., Furnari, B. A., Massari, M. E., Matthews, J. R., Quong, M. W., Rivera, R. R., Stuiver, M. H., 1994. Structure and function of helix-loop-helix proteins. Biochim Biophys Acta. 1218, 129-35.

Nasiadka, A., Krause, H. M., 1999. Kinetic analysis of segmentation gene interactions in Drosophila embryos. Development (Cambridge, England). 126, 1515-26.

Nilson, L. A., Schupbach, T., 1998. Localized requirements for windbeutel and pipe reveal a dorsoventral prepattern within the follicular epithelium of the Drosophila ovary. Cell. 93, 253-62.

Nunes da Fonseca, R., von Levetzow, C., Kalscheuer, P., Basal, A., van der Zee, M., Roth, S., 2008. Self-regulatory circuits in dorsoventral axis formation of the short-germ beetle Tribolium castaneum. Dev Cell. 14, 605-15.

Nüsslein-Volhard, C., 1979. Maternal Effect Mutations that Alter the Spatial Coordinates of the Embryo of Drosophila melanogaster. Determinants of Spatial Organization. 28.

Nüsslein-Volhard, C., 1991. Determination of the embryonic axes of Drosophila. Dev Suppl. 1, 1-10.

Nüsslein-Volhard, C., Frohnhöfer, H. G., Lehmann, R., 1987. Determination of anteroposterior polarity in Drosophila. Science. 238, 1675-81.

Ochoa-Espinosa, A., Small, S., 2006. Developmental mechanisms and cis-regulatory codes. Current Opinion in Genetics & Development. 16, 165.

Ochoa-Espinosa, A., Yu, D., Tsirigos, A., Struffi, P., Small, S., 2009. Sackler Special Feature: Anterior-posterior positional information in the absence of a strong Bicoid gradient. Proc Natl Acad Sci USA.

Ochoa-Espinosa, A., Yucel, G., Kaplan, L., Pare, A., Pura, N., Oberstein, A., Papatsenko, D., Small, S., 2005. The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. Proc Natl Acad Sci USA. 102, 4960-5.

Pan, D., Courey, A. J., 1992. The same dorsal binding site mediates both activation and repression in a context-dependent manner. EMBO J. 11, 1837-42.

Papatsenko, D., 2007. ClusterDraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors. Bioinformatics. 23, 1032-4.

Papatsenko, D., Levine, M., 2005a. Computational identification of regulatory DNAs underlying animal development. Nat Meth. 2, 529.

Papatsenko, D., Levine, M., 2005b. Quantitative analysis of binding motifs mediating diverse spatial readouts of the Dorsal gradient in the Drosophila embryo. Proc Natl Acad Sci USA. 102, 4966-71.

Pyrowolakis, G., Hartmann, B., Muller, B., Basler, K., Affolter, M., 2004. A simple molecular complex mediates widespread BMP-induced repression during Drosophila development. Dev Cell. 7, 229-40.

Romano, L. A., Wray, G. A., 2003. Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. Development. 130, 4187-99.

Roth, S., 2003. The origin of dorsoventral polarity in Drosophila. Philos Trans R Soc Lond, B, Biol Sci. 358, 1317-29; discussion 1329.

Roth, S., Stein, D., Nüsslein-Volhard, C., 1989. A gradient of nuclear localization of the dorsal protein determines dorsoventral pattern in the Drosophila embryo. Cell. 59, 1189-202.

Rubin, G. M., Spradling, A. C., 1982. Genetic transformation of Drosophila with transposable element vectors. Science. 218, 348-53.

Rusch, J., Levine, M., 1997. Regulation of a dpp target gene in the Drosophila embryo. Development. 124, 303-11.

Rushlow, C. A., Han, K., Manley, J. L., Levine, M., 1989. The graded distribution of the dorsal morphogen is initiated by selective nuclear transport in Drosophila. Cell. 59, 1165-77.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., Lenhard, B., 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 32, D91-4.

Sander, K., 1996. Pattern formation in insect embryogenesis: The evolution of concepts and mechanisms. International Journal of Insect Morphology and Embryology.

Schneider, D. S., Hudson, K. L., Lin, T. Y., Anderson, K. V., 1991. Dominant and recessive mutations define functional domains of Toll, a transmembrane protein required for dorsal-ventral polarity in the Drosophila embryo. Genes Dev. 5, 797-807.

Schupbach, T., 1987. Germ line and soma cooperate during oogenesis to establish the dorsoventral pattern of egg shell and embryo in Drosophila melanogaster. Cell. 49, 699-707.

Sen, J., Goltz, J. S., Stevens, L., Stein, D., 1998. Spatially restricted expression of pipe in the Drosophila egg chamber defines embryonic dorsal-ventral polarity. Cell. 95, 471-81.

Shi, S., Larson, K., Guo, D., Lim, S. J., Dutta, P., Yan, S. J., Li, W. X., 2008. Drosophila STAT is required for directly maintaining HP1 localization and heterochromatin stability. Nat Cell Biol. 10, 489-96.

Shine, I., Wrobel, S., 1976. Thomas Hunt Morgan : pioneer of genetics. University Press of Kentucky, Lexington.

Stanojevic, D., Hoey, T., Levine, M., 1989. Sequence-specific DNA-binding activities of the gap proteins encoded by hunchback and Kruppel in Drosophila. Nature. 341, 331-5.

Stathopoulos, A., Levine, M., 2002a. Dorsal Gradient Networks in the Drosophila Embryo. Dev Biol. 246, 57.

Stathopoulos, A., Levine, M., 2002b. Linear signaling in the Toll-Dorsal pathway of Drosophila: activated Pelle kinase specifies all threshold outputs of gene expression while the bHLH protein Twist specifies a subset. Development. 129, 3411-9.

Stathopoulos, A., Levine, M., 2002c. Whole-genome expression profiles identify gene batteries in Drosophila. Dev Cell. 3, 464-5.

Stathopoulos, A., Levine, M., 2004. Whole-genome analysis of Drosophila gastrulation. Curr Opin Genet Dev. 14, 477-84.

Stathopoulos, A., Levine, M., 2005a. Genomic Regulatory Networks and Animal Development. Developmental Cell. 9, 449.

Stathopoulos, A., Levine, M., 2005b. Localized repressors delineate the neurogenic ectoderm in the early Drosophila embryo. Developmental Biology. 280, 482.

Stathopoulos, A., Van Drenth, M., Erives, A., Markstein, M., Levine, M., 2002. Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo. Cell. 111, 687-701.

Staudt, N., Fellert, S., Chung, H. R., Jackle, H., Vorbruggen, G., 2006. Mutations of the Drosophila zinc finger-encoding gene vielfaltig impair mitotic cell divisions and cause improper chromosome segregation. Mol Biol Cell. 17, 2356-65.

Tautz, D., Pfeifle, C., 1989. A non-radioactive in situ hybridization method for the localization of specific RNAs in Drosophila embryos reveals translational control of the segmentation gene hunchback. Chromosoma. 98, 81.

ten Bosch, J. R., Benavides, J. A., Cline, T. W., 2006. The TAGteam DNA motif controls the timing of Drosophila pre-blastoderm transcription. Development. 133, 1967.

Thisse, C., Perrin-Schmitt, F., Stoetzel, C., Thisse, B., 1991. Sequence-specific transactivation of the Drosophila twist gene by the dorsal gene product. Cell. 65, 1191-201.

Tickle, C., 1999. Morphogen gradients in vertebrate limb development. Semin Cell Dev Biol. 10, 345-51.

Verma, I. M., Stevenson, J. K., Schwarz, E. M., Van Antwerp, D., Miyamoto, S., 1995. Rel/NF-kappa B/I kappa B family: intimate tales of association and dissociation. Genes Dev. 9, 2723-35.

Visel, A., Prabhakar, S., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E. M., Pennacchio, L. A., 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet. 40, 158-60.

Vlieghe, D., Sandelin, A., De Bleser, P. J., Vleminckx, K., Wasserman, W. W., van Roy, F., Lenhard, B., 2006. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. Nucleic Acids Res. 34, D95-7.

Von Ohlen, T., Doe, C. Q., 2000. Convergence of dorsal, dpp, and egfr signaling pathways subdivides the drosophila neuroectoderm into three dorsal-ventral columns. Dev Biol. 224, 362-72.

Weiss, J. B., Von Ohlen, T., Mellerick, D. M., Dressler, G., Doe, C. Q., Scott, M. P., 1998. Dorsoventral patterning in the Drosophila central nervous system: the intermediate neuroblasts defective homeobox gene specifies intermediate column identity. Genes Dev. 12, 3591-602.

Whalen, A. M., Steward, R., 1993. Dissociation of the dorsal-cactus complex and phosphorylation of the dorsal protein correlate with the nuclear localization of dorsal. J Cell Biol. 123, 523-34.

Wieschaus, E., 1996. Embryonic transcription and the control of developmental pathways. Genetics. 142, 5-10.

Wolpert, L., 1968. The French Flag problem: a contribution to the discussion on pattern development and regulation. Towards a theoretical biology: C. H. Waddington. 1, 125-133.

Xu, C., Kauffmann, R. C., Zhang, J., Kladny, S., Carthew, R. W., 2000. Overlapping activators and repressors delimit transcriptional response to receptor tyrosine kinase signals in the Drosophila eye. Cell. 103, 87-97.

Xylourgidis, N., Roth, P., Sabri, N., Tsarouhas, V., Samakovlis, C., 2006. The nucleoporin Nup214 sequesters CRM1 at the nuclear rim and modulates NFkappaB activation in Drosophila. Journal of cell science. 119, 4409-19.

Yan, R., Small, S., Desplan, C., Dearolf, C. R., Darnell, J. E., Jr., 1996. Identification of a Stat gene that functions in Drosophila development. Cell. 84, 421-30.

Yuh, C. H., Bolouri, H., Davidson, E. H., 2001. Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. Development. 128, 617-29.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J. W., Nechaev, S., Adelman, K., Levine, M., Young, R. A., 2007a. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat Genet. 39, 1512-6.

Zeitlinger, J., Zinzen, R. P., Stark, A., Kellis, M., Zhang, H., Young, R. A., Levine, M., 2007b. Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. Genes Dev. 21, 385-90.

Zinzen, R., Senger, K., Levine, M., Papatsenko, D., 2006. Computational Models for Neurogenic Gene Expression in the Drosophila Embryo. Current Biology. 16, 1358-1365.