2009

Hierarchy and CIS-Regulation in Drosophila Segmentation: Rules for Pattern Formation and Clues to Evolution

Mark David Schroeder

HIERARCHY AND CIS-REGULATION IN *DROSOPHILA* SEGMENTATION: RULES FOR PATTERN FORMATION AND CLUES TO EVOLUTION

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Mark David Schroeder

June 2009

Hierarchy and Cis-regulation in *Drosophila* Segmentation:
Rules For Pattern Formation and Clues to Evolution

Mark David Schroeder, Ph.D.
The Rockefeller University 2009

In few systems is it possible to analyze the global *cis*-regulatory structure of developmental transcription networks. One system where this is in principle possible is segmentation in *Drosophila melanogaster*, although to date such an undertaking has not been attempted. Here using computational algorithms to analyze the transcriptional regulatory regions of genes of the gap and pair rule classes such an analysis is carried out. Computational analysis, transgenic reporter element assays, site directed mutagenesis, genetics, and time courses of *in situ* hybridizations of central genes in carefully staged embryos are combined to understand how the *cis*-elements function together to achieve patterning of the anterior posterior axis. The transition from the non-periodic gap patterns to the seven striped periodic patterns of the pair rule genes is analyzed in detail. This step in the genetic hierarchy is of particular interest as it generates the segmental pattern that underlies the *Drosophila* body plan. The analysis clarifies the primary and secondary pair rule classification system and suggests certain organizational principles in pair rule *cis*-regulation.

This thesis is dedicated to my family.

Acknowledgements

I'd like to acknowledge a number of people for their help in carrying out the work presented in this thesis. Jak Fak, Micheal Pearce, and HongQing Fan did the experimental work described in Chapter 2. Christina Greer carried out many of the *in situ* hybridizations done in Chapter 3 and was a great help in with the injections and balancing required to generate the transgenic reporter lines. I'm very grateful to Ulrike Gaul, for her guidance and patience during the work. It was always enjoyable to have Ulrich Unnerstall around for in depth discussions. I'd also like to acknowledge Nikolaus Rajewsky, Saurabh Sinha, Massimo Vergassola, and Eric Siggia for developing Ahab and Stubb, which made much of this work possible.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

This thesis focuses on the role of hierarchy in generating the anterior-posterior (a-p) axis in the embryo of the fruit fly, *Drosophila melanogaster*. Establishment of the a-p axis is referred to as segmentation due to the segmental, or repeated, organization of insect body plans. Despite much work on the topic, the details of how this reiterated pattern is established are not well understood.

The earliest steps in segmentation occur during the syncytial blastoderm when the embryo is one large cell filled with dividing nuclei. By the 10^{th} nuclear division cycle roughly one thousand nuclei are positioned at the periphery of the embryo generating a two dimensional array. At this point zygotic transcription begins and a relatively small set of transcription factors are expressed in specific patterns. These transcription factors are able to diffuse between adjacent nuclei and thereby refine the initial patterns. Unlike most developmental contexts transcriptional cross regulation can generate pattern directly without intervening signal transduction pathways. As the a-p and dorsal ventral (d-v) axes are largely independent at this time, establishment of the a-p axis can be analyzed primarily in one dimension. These simplifications make *Drosophila* segmentation a good model system for understanding pattern formation within a purely transcriptional paradigm.

Transcriptional regulation in segmentation has been studied extensively, with a wealth of binding site data and promoter dissections. These data are sufficient to enable the use of computational algorithms to predict transcriptional

regulatory regions in genomic DNA from binding site data. Using existing algorithms and binding site preferences of transcription factors from the literature, a complete dissection of the transcriptional control regions of the core segmentation genes leading up to the establishment of the initial periodic patterns was a goal of this work. The binding site content of this comprehensive set of transcriptional *cis*-regulatory elements is then analyzed with the same computational methods to better understand how the patterns are encoded. Such a detailed network wide dissection and analysis has not been carried out previously in any developmental system.

1.1 History of Segmentation

The study of segmentation has a rich history and an important place in the modern study of developmental biology. The history nicely frames some of the ideas presented in the thesis and provides a useful context for this work. This brief review draws heavily from the "History short stories" section at the end of "The Making of a Fly" by Peter Lawrence (Lawrence, 1992), although available primary sources were also reviewed.

In the early genetics research, the analysis of transmission of genes between generations was largely separate from analysis of the function of these genes in the organism. Thomas Hunt Morgan's 1926 book "The Theory of the Gene" (Morgan, 1928) has an apt quote in this respect: "Between the characters, that furnish the data for the theory, and the postulated genes, to which the characters are referred, lies the whole field of embryonic development. The

theory of the gene, as here formulated, states nothing with respect to the way in which the genes are connected with the end product or character. The absence of information relating to this interval does not mean that the process of embryonic development is not of interest for genetics. A knowledge of the way in which the genes produce their effects on the developing individual would, no doubt, greatly broaden our ideas relating to heredity and probably make clearer many phenomena that are obscure at present, but the fact remains that the sorting out of the characters in successive generations can be explained at present without reference to the way in which the gene affects the developmental process." Therefore genetic research focused on inheritance of genes, putting aside the important topic of the genetics of embryonic development.

Most developmental biology research was embryological in nature, with techniques focused on transplantation and various crude mechanical manipulations of the embryo (Sander, 1976). The main exception was in the context of fate mapping, where genetic methods for labeling specific subsets of cells generated an interface where the two groups came together. The use of mosaic animals with a combination of wildtype and mutant tissue in fate mapping was pioneered by Sturtevant in the late 1930s, but only entered into more widespread use in trying to understand development in the 1960s and 1970s. In the 70s it became apparent that in early *Drosophila* development there were no strict cell lineages, but rather a series of cell fate restrictions in which groups of cells were specified to increasingly restricted fates in a stepwise fashion (Gehring, 1975). It was shown through fate mapping with UV laser microbeam and gynandromorphs (male-female mosaics) that 3-4 cell wide

regions at the cellular blastoderm are already specified to individual segments, suggesting that this fate restriction occurs very early with high cellular precision (Lohs-Schardin et al., 1979; Szabad et al., 1979; Wieschaus and Gehring, 1976).

The theory of compartments, which came out of mosaic studies of the development of adult *Drosophila* structures, is particularly noteworthy. Antonio Garcia-Bellido and co-workers initially generated this theory while working on *Drosophila* wing imaginal discs (Crick and Lawrence, 1975; Garcia-Bellido et al., 1973). Imaginal discs are sacs of epithelial cells that form adult structures, which develop autonomously during larval stages and are combined together during pupation to form the adult body plan. In most imaginal discs, anterior and posterior cells will not mix and are restricted to fates generated by these regions. The gene *engrailed* (*en*) is required for posterior fates and the differential adhesion underlying the cell sorting properties. Clones of *en* mutant cells in the anterior portion of the disc show no phenotype, but clones in the posterior generate anterior fates and mix with anterior cells indicating a loss of proper cell fate restriction. These properties indicate that there are specific genes, called selector genes, which restrict the fate of cells to those generated in specific regions of the developing organism. Compartments and stepwise fate restriction together suggested a stepwise hierarchical system that generates increasingly detailed patterns selected by specific genes as an organism develops. However, at this time most genetics still focused on adult morphology and few genes with selector function were known.

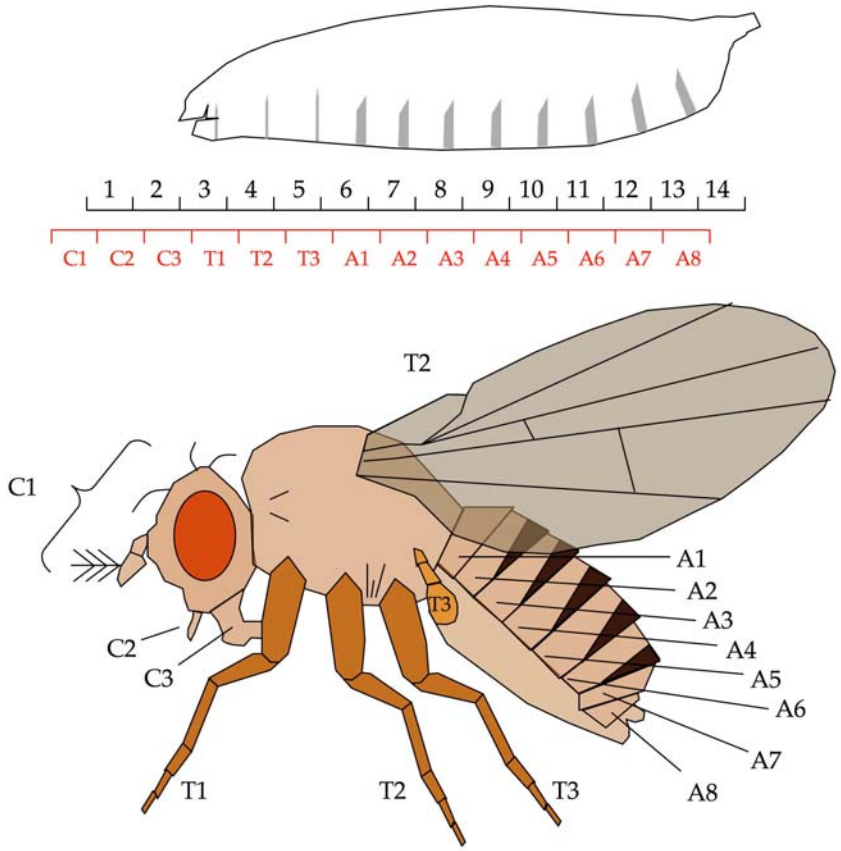


Figure 1

The body plans of larval and adult *Drosophila melanogaster*. Although originally defined as a series of segments corresponding to overt morphological units such as legs, the body plan is in fact specified as a series of parasegments. *Ubx* mutants affect parasegments 5 and 6, such that the posterior of the T2 leg is transformed to the posterior of the T1 leg and all of the T3 is transformed to the T1 leg. Therefore the molecular and anatomical morphological units are offset from each other.

The principle selector genes known at this time were the homeotic genes, most famously those of the bithorax complex, which was extensively studied by Ed Lewis starting in the 1940s (Lewis, 1978, 1998). The homeotic gene *Ultrabithorax* (*Ubx*), for example, is required to specify haltere vs. wing fate, and in its absence the haltere is transformed into wing. Strikingly, the order of the genes in the complex corresponds to the order of the structures they specify along the a-p axis of the organism. Originally the *Drosophila* body plan was assigned into a series of segments, where serially homologous structures such as the three sets of legs were each assigned into different units. Later more detailed analysis of phenotypes indicated that the morphological units recognized as segments did not correspond to the units defined molecularly (Kerridge and Morata, 1982; Morata and Kerridge, 1981), but rather offset units named parasegments (Figure 1). Although there was clearly interdependence between the genes in the homeotic complex, the exact nature of the interactions was difficult to decipher at this time and only became clear much later. Therefore, even in one case where a number of related selector genes were known, their interactions were not easily studied.

The availability of genes involved in setting up the segmental body plan then changed dramatically with the Nobel Prize winning genetic screens led by Christiane Nusslein-Volhard and Eric Wieschaus in the early 1980s. The segmentation screens attempted to define all zygotic genes that caused specific defects in the larval cuticle pattern when mutated (Jurgens et al., 1984; Nusslein-

Volhard and Wieschaus, 1980; Nusslein-Volhard et al., 1984; Wieschaus et al., 1984). The initial screen paper described 15 mutants, which surprisingly included the majority of the core set of genes that establishing the periodic pattern (Nusslein-Volhard and Wieschaus, 1980). This was many fewer than assumed previously, but many more genes assigned to a common developmental genetic program than known in the decades of prior research. Later more technically demanding screens for maternal genes required for the establishment of polarity within the embryo were also carried out (Nusslein-Volhard et al., 1987; Schupbach and Wieschaus, 1986). Beyond a simple catalog of genes, this work ushered in an era of great progress in understanding the molecular basis of developmental biology. Therefore with an appropriate set of screening criteria there was now a recipe for genetic dissection of the core genes involved in specific developmental processes.

The determination of this comprehensive set of genes provided a link between genetics and development. The role of genes not just in transmission of characteristics, but also in genetic programs that patterned organismal form was now an addressable question. Both *en* and a number of genes causing homeotic defects were found in this screen. As a result, the isolated cases that were previously known were now set in a genetic framework. Although the role of *en* was initially determined in imaginal discs, the fact that *en* was now seen to have a role in patterning the larval body plan indicated that some genes were used at multiple stages of development. This was the beginning of the realization that many core developmental genes were reused repeatedly at different

developmental stages. The role of genetics in development could now be addressed in a systematic fashion.

1.2 Review of Segmentation

The genetic hierarchy that unfolded from the segmentation screen consists of four classes of mutations including the maternal, gap, pair rule, segment polarity, and homeotic classes (Figure 2). Each class within the hierarchy effects a more restricted portion of the body plan as a series of steps, which elaborate increasingly detailed structures. The maternal genes are responsible for establishing embryonic polarity, the gap genes for specifying broad regions of the embryo, the pair rule genes for specifying alternating segments, and finally the segment polarity genes for specifying portions of every segment. The pair rule genes came as a particular surprise because, in their absence, every second segment of the body plan was effected. For instance in *even-skipped*, the second, fourth, sixth, and eighth abdominal segments are absent. That there was a developmental stage with a two segment organization had not been suggested by the extensive embryology done on insects prior to this time. Similarly, the irregular nature of the gap genes did not relate directly to the morphological features of the embryo. In contrast, both the segment polarity and homeotic gene classes cause defects that correspond to defined regions of segments and simply lead to mis-specification of one portion of the body plan into another. Therefore the gap and pair rule genes identified a set of unexpected positional cues.

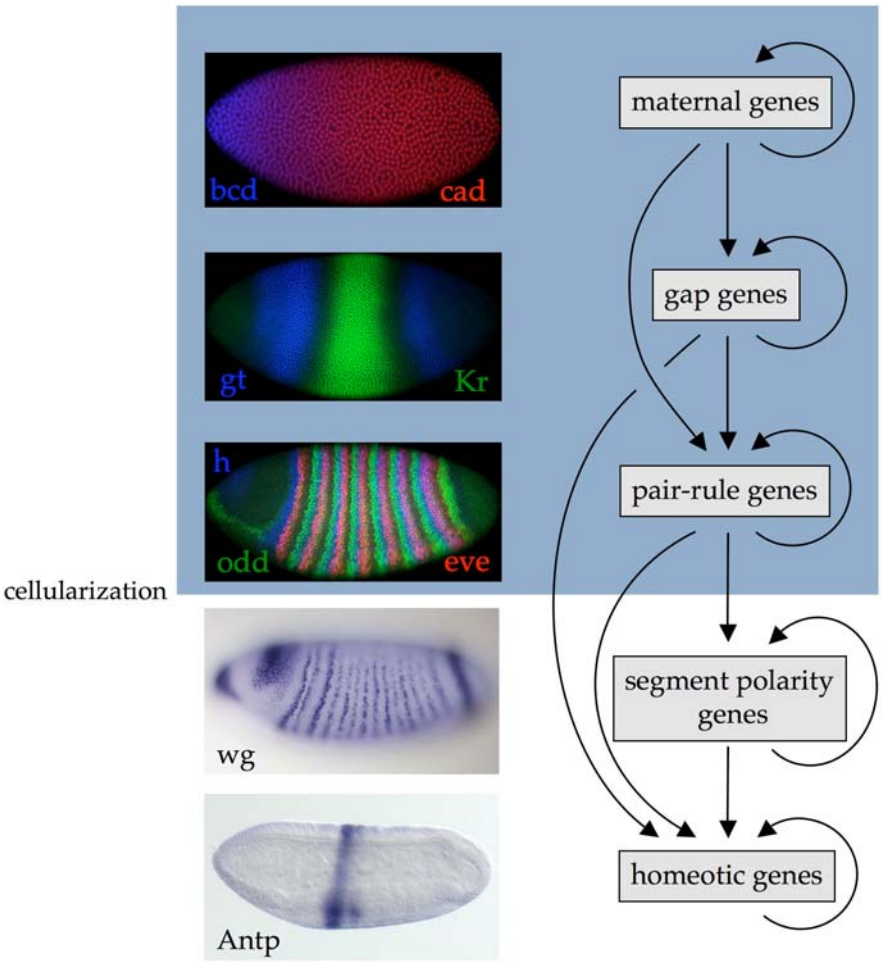


Figure 2

A schematic of the segmentation hierarchy. Embryos are oriented with the anterior to the left and dorsal side facing up as is the convention followed for all pictures of embryos. All tiers within the hierarchy are generated by a combination of cross-regulation within the tier and regulation by preceding tiers. At each step within the hierarchy the patterns are refined into more precise domains of expression. The maternal genes establish gradients of the transcriptional activators BCD and CAD through post transcriptional regulation, which then establish polarity within the embryo at the syncytial stage. The gap and pair rule classes also act during the syncytial stage, primarily as transcriptional repressors. Following cellularization the segment polarity genes, which include the hedgehog and wingless signaling pathways, become active. Fluorescent embryo stainings are from (Surkova et al., 2008)

The segmentation screen was built on work that sought to understand the fate map of the *Drosophila* embryo. Although genetic screens had been attempted earlier, most of the previous work in *Drosophila* focused on mutations affecting adults. The choice of a screen focused on early development was based on the idea that rapidly developing organisms provide their eggs with large quantities of the essential factors necessary for cellular function. Therefore genes expressed during early zygotic development are biased towards those required in spatially restricted patterns (Wieschaus, 1996). The segmentation screens focused on the larval cuticle and built upon prior fate mapping experiments that sought to understand how the pattern of the larval cuticle was specified. The work on fate mapping early events revealed the importance of the blastoderm stage of development as this the time when cell fate restrictions begin to occur (Gehring, 1975; Wieschaus and Gehring, 1976) and supported prior work in *Drosophila* indicating that there were no strict lineages. How is it that genes programmed groups of cells to specific fates and restricted their developmental capacities?

class	gene name	symbol	principle domain
maternal	*bicoid*	*bcd*	homeodomain
	caudal	*cad*	homeodomain
maternal/gap	*hunchback*	*hb*	C2H2 zinc finger
gap	*Kruppel*	*Kr*	C2H2 zinc finger
	knirps	*kni*	NHR
	giant	*gt*	basic leucine zipper
	tailless	*tll*	NHR
	huckebein	*hkb*	C2H2 zinc finger
head gap	*buttonhead*	*btd*	C2H2 zinc finger
	cap 'n' collar	*cnc*	basic leucine zipper
	collier	*col*	bHLH
	crocodile	*croc*	forkhead domain
	empty spiracles	*ems*	homeodomain
	forkhead	*fkh*	forkhead domain
	orthodenticle	*otd*	homeodomain
primary pair rule	*hairy*	*h*	bHLH
	even-skipped	*eve*	homeodomain
	runt	*run*	RUNX domain
secondary pair rule	*fushi-tarazu*	*ftz*	homeodomain
	odd-skipped	*odd*	C2H2 zinc finger
	sloppy-paired	*slp*	forkhead domain
	paired	*prd*	PRD homeodomain
	odd-paired	*opa*	C2H2 zinc finger
segment polarity	*engrailed*	*en*	homeodomain
	gooseberry	*gsb*	PRD homeodomain
	wingless	*wg*	secreted ligand
	armadillo	*arm*	cytoskeletal
	hedgehog	*hh*	secreted ligand
	patched	*ptc*	hh receptor
	fused	*fus*	kinase
	cubitus interruptus	*ci*	C2H2 zinc finger

Table 1

List of selected segmentation genes. Only transcription factors effecting the a-p axis are given for the maternal class. All gap and pair rule genes found in the initial screens are transcription factors. The segment polarity genes found in the original screens are listed, but additional members of the hedgehog and wingless signaling pathways that were found later are not. Although not discussed in the text, the set of head gap genes are given as well.

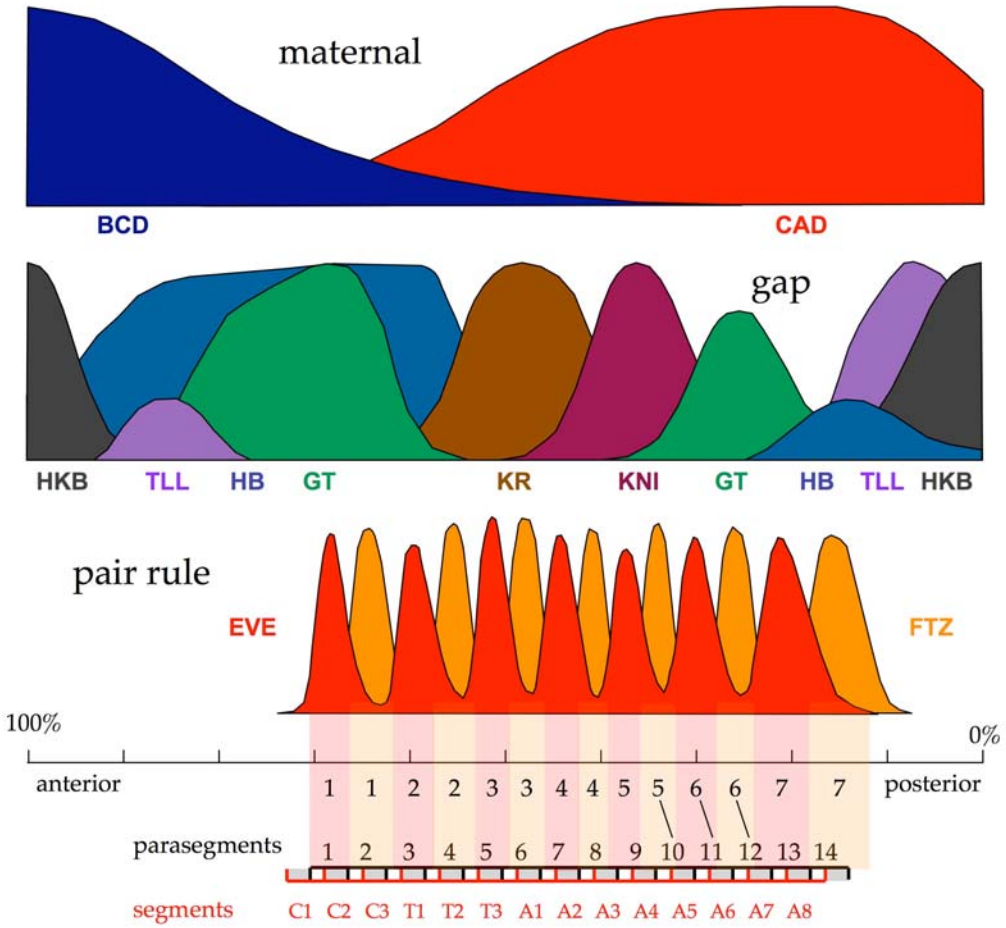


Figure 3

A schematic of the maternal, gap, and pair rule patterns based on data from the FlyEx database (Myasnikova et al., 2001). The anterior of the embryo is to the left, as will be the convention in all schematics of expression patterns shown in the thesis. The strength of expression indicated by height of the plotted domain. For the pair rule class, only EVE and FTZ are shown for clarity. As *eve* and *ftz* define individual parasegments, they mark the units that define the body plan of the fly. Below the patterns, the parasegments and segments are labeled in an idealized regular fashion. The gene *en* is expressed in the posterior compartment, which corresponds to the white region in the schematic of the segments and parasegments.

The genes involved in transcriptional control of zygotic patterning are schematized in Figure 3. Although the maternal class is a large one, the polarized maternal transcription factors that mediate the effect are limited to *bicoid* (*bcd*), *caudal* (*cad*), and *hunchback* (*hb*) (Table 1, Figure 3). All three genes form early gradients with BCD and HB concentrations maximal in the anterior and CAD concentration maximal in the posterior. Both *bcd* mRNA and protein are localized to the anterior pole of the embryo, generating a roughly exponential gradient towards the posterior. The BCD gradient acts primarily through transcriptional activation to activate anteriorly expressed genes like *hb,* but also represses translation of the ubiquitous maternal *cad* mRNA forming a reciprocal gradient that peaks at the posterior. There is also posterior class of genes that block the translation of *bcd* and *hb* mRNA in the posterior portion of the embryo. In all three genes, the translational repression leads to higher rates of mRNA degradation, thereby generating both protein and mRNA gradients. Therefore, the maternal system sets up the graded activity of three proteins, BCD and CAD, which act as activators, and HB, which can act as both an activator and a repressor.

The gap gene class consists entirely of transcription factors and includes *hb*, as well as *Kruppel* (*Kr*), *knirps* (*kni*), *giant* (*gt*), *tailless* (*tll*), and *huckebein* (*hkb*) (Table 1, Figure 3). The maternal regulation through the *torso* (*tor*) signal transduction pathway, which is specifically activated at the termini, generates the localized expression of *hkb* and *tll*. Expression of the remaining gap genes is thought to be patterned solely through the maternal gradients and gap gene cross regulation. Except for *hb,* most of these genes have been demonstrated to

act primarily as repressors, but, based largely on tissue culture experiments, it has been proposed that *Kr* can activate {La Rosee-Borggreve, 1999 #160; Sauer, 1991 #257}. There is also a group of head gap genes that have little if any role in regulating the segmented portion of the embryo and will therefore not be discussed. The maternal and zygotic classification of *hb* is due to the fact that it is contributed both maternally and zygotically and is therefore active at both stages. The only other transcription factor similarly provided at both stages is *cad* (Schulz and Tautz, 1995) and in both cases the transcriptional regulation reinforces the early, polarized pattern generated by the maternal translational control.

The pair rule class was originally limited to the transcription factors *hairy* (*h*), *even-skipped* (*eve*), *runt* (*run*), *fushi-tarazu* (*ftz*), *odd-skipped* (*odd*), *paired* (*prd*), *sloppy-paired* (*slp*), and *odd-paired* (*opa*) (Table 1). Since the original screen there have been additional maternally and zygotically expressed genes found that generate weaker more irregular pair rule phenotypes {Baumgartner, 1994 #11; Yan, 1996 #319}. This newer set of genes will not be discussed, as they are not patterned at the syncytial stage on which this thesis focuses. It was pointed out early on that *eve* and *ftz* determined the anterior boundaries of the parasegments through their regulation of *en* (Lawrence et al., 1987), which suggested they were particularly important members of the pair rule class (Figure 3). It has since been shown that the relative concentration of *eve* and *ftz* determines the size of parasegments (Hughes and Krause, 2001). Given the parasegmental organization of the fly body plan, establishment of these regions is central to this process.

The segment-polarity class of genes consists of *engrailed* (*en*) and *gooseberry* (*gsb*), as well as members of the *wingless* (*wg*) and *hedgehog* (*hh*) signaling pathways (Table 1). Only the gap and pair rule classes consist solely of transcription factors, whereas both the classes above and below them include signaling molecules. This matches up nicely with the importance of the gap and pair rule classes at the syncytial stage where transcription factors alone can "signal" between nuclei by diffusion. The segment polarity gene *en* is particularly important in establishing the body plan as it remains on throughout development and is important for organizing both larval and adult structures. As mentioned earlier *en* defines the posterior region of compartments. During embryonic development and within the wing disc *en* expression similarly establishes a *hh* gradient that organizes much of the pattern in both contexts (Blair, 1995; Sanson, 2001). The compartmental boundaries and the parasegmental boundaries are congruent and the embryonic *en* expression domain is maintained to form the expression domain in the posterior of the imaginal discs (Figure 3).

When the gap and pair rule genes were cloned and their patterns were determined, there was a strong correspondence between the expression patterns and phenotypes, particularly in the gap and pair rule classes (Figure 3). The gap genes are expressed in differentially positioned graded domains with substantial overlap. The next transition in the hierarchy is the remarkable jump to the much sharper periodic patterns of the pair rule genes, which are responsible for establishing the repeated patterns central to the segmentation process. Five of

the eight pair rule genes transition from their seven-stripe pattern to a segmental one and direct the 14 or 15 striped patterns of most segment polarity genes. One atypical case among the pair rule genes is that of *opa,* which is required to activate the *en* stripes in the odd parasegments. This highlights the interesting fact that the transcriptionally patterned segment polarity genes are differentially regulated in their even and odd stripes. Therefore segmental patterns initiate with an inherently pair rule organization suggesting a connection between these two classes. The correspondence between the patterns and the phenotypes also supports the interpretation that it is these factors themselves that somehow define the identity of the cells they are expressed in.

The pair rule class itself has been split into two classes, based on their contribution to establishing the early seven-striped pattern (Ingham, 1988). This categorization was initially based on molecular epistasis experiments, where it was determined that *h, eve,* and *run* were required to generate the initial patterns of all patterned pair rule genes, whereas the secondary pair rule genes were not required to generate the initial seven-striped pattern of *h, eve, and run* {Ingham, 1988 #535; Ingham, 1988 #120}. This lead to *h, eve,* and *run* being categorized as primary pair rule genes and *ftz*, *odd, prd, slp*, and *opa* being categorized as secondary pair rule genes. Therefore in contrasts to the important role of *ftz* in specifying parasegment size, *ftz* was suggested to have a lesser role in specifying the initial pattern.

At roughly this time a series of *h* "region specific alleles" were found in which *h* function and expression were only lost in particular portions of embryo

(Howard et al., 1988). This was used to argue that the upstream control region of *h* was modular and that it contained independent regions that interpreted the gap gene patterns in different portions of the embryo. Soon after it was shown that the upstream region of *eve* had distinct DNA *cis*-elements that could autonomously drive individual stripes of a LacZ reporter gene (Figure 4) (Goto et al., 1989; Harding et al., 1989). The *cis*-elements generating one or two stripes form a class of early acting "stripe specific elements" that demonstrated the existence of modular *cis*-regulatory elements sufficient to generate sub-portions of the complex patterns of these genes.

In order to understand how the stripe specific elements decode the gap gene patterns, the stripe 2 element was trimmed down to a minimal element of 480 bp sufficient to drive a correctly delimited pattern (Small et al., 1991). This element was small enough to allow comprehensive analysis of its binding site content both through DNAse footprinting and site directed mutagenesis experiments (Arnosti et al., 1996; Small et al., 1992). The mapping of sites showed a striking overlap of activator and repressor sites and lead to a model where gap repressors would block expression in part through competitive binding (Figure 4). The extensive site directed mutagenesis allowed study of the function of the maternal and gap inputs without the indirect effects unavoidable in mutant analysis. From this work it emerged that stripe 2 was generated by broad anterior activation mediated by BCD and HB binding sites, and then delimited by the flanking gap repressors GT and KR. This provided a fairly straightforward model, where repression is the primary generator of pattern of the *cis*-elements regulated by the maternal and gap genes.

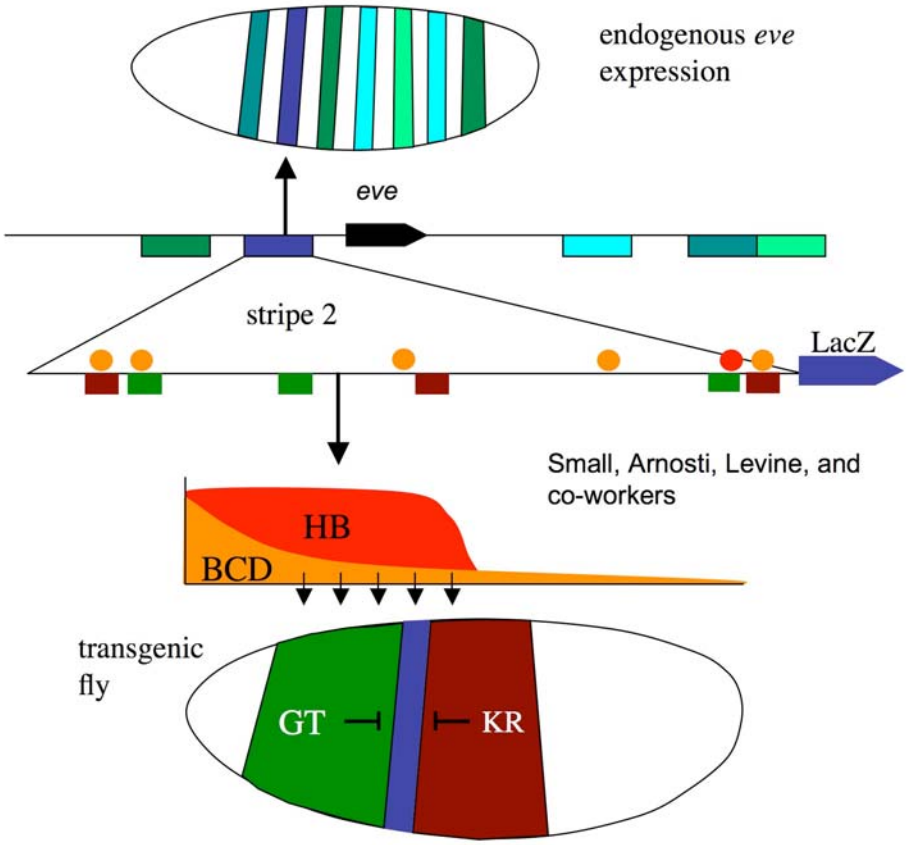


Figure 4

A schematic of the *eve* locus and its cis-regulatory elements. The early pattern initiated just prior to nuclear cycle 14 is generated by a series of stripe specific elements. The regulation of *eve* stripe 2 is shown schematically. It can autonomously recapitulate part of the *eve* pattern when driving a LacZ reporter gene. Broad activation mediated by binding sites for BCD (orange) and HB (red) is delimited by repression by overlapping GT (green) and KR (brown) sites.

Although there has been little systematic site directed mutagenesis of *cis*-elements, DNAse footprinting has been used to find binding sites for the maternal and gap genes in a large number of elements (Hartmann et al., 1994; Hoch et al., 1992; Hoch et al., 1990; La Rosee et al., 1997; La Rosee-Borggreve et al., 1999; Langeland et al., 1994; Pankratz et al., 1989; Pankratz et al., 1990; Rivera-Pomar et al., 1995; Stanojevic et al., 1989; Treisman and Desplan, 1989). When these elements are crossed into the corresponding mutants, the expression typically expands into the region where the mutant gap gene regulator is expressed supporting the *eve* stripe 2 model.

In the dissection of *eve*, *ftz*, *prd*, and *run*, another class of "seven-stripe" elements was found that could drive the complete seven-stripe pattern (Goto et al., 1989; Gutjahr et al., 1994; Hiromi et al., 1985; Klingler et al., 1996). Therefore the pair rule genes have both stripe specific elements that can generate their pattern in a modular fashion and seven-stripe elements that can generate their whole pattern in an inherently periodic fashion. The only pair rule gene thought to lack a seven-stripe element is *h* {Howard, 1990 #115; Pankratz, 1990 #217; Riddihough, 1991 #241}. Although most pair rule genes seem to act primarily as repressors in pair rule cross regulation, there is a limited role of activation as well {Vanderzwan-Butler, 2007 #669}.

The existence of stripe specific elements in the primary pair rule genes *h*, *eve*, and *run* (Fujioka et al., 1999; Goto et al., 1989; Harding et al., 1989; Howard and Struhl, 1990; Klingler et al., 1996; Pankratz et al., 1990; Riddihough and Ish-Horowicz, 1991) lead to the model that these genes interpret the maternal and

gap gradients whereas the secondary pair rule genes simply read off the primary patterns. This view was supported by the fact that only seven-stripe elements were found in *ftz, slp,* and *prd* when they were originally dissected (Gutjahr et al., 1994; Hiromi and Gehring, 1987; Hiromi et al., 1985; Lee and Frasch, 2000). Although it has been argued that the *ftz* pattern is not a simple consequence of reading off the primary pair rule patterns (Yu and Pick, 1995), no clear mechanism has been shown for how the early pattern arises. After the establishment of the initial seven-stripe pattern it is clear that this cross regulation occurs regardless of position in the hierarchy, indicating a high degree of temporal modulation of pair rule cross regulation. The interplay between stripe specific elements and seven-stripe elements in generating the pair rule patterns is not well understood, though in the case of *eve* there is a strict requirement for early EVE expression driven by the stripe specific elements in order for the stripes to be generated by the seven-stripe element (Fujioka et al., 1995).

Therefore there is a well supported model that the gap genes work as repressors to delimit broadly activated *cis*-elements to generate more refined domains of expression. This occurs first at the level of gap gene cross regulation and then later in patterning the stripe specific elements of the pair rule genes. Complex patterns are initially built up by modular control regions, however, the seven-stripe elements of the pair rule genes indicate that once more complex patterns have been generated simpler regulatory interactions can utilize these patterns to maintain and refine existing complex patterns.

Despite being a textbook example of development, there are still many unanswered questions in segmentation. In particular pair rule regulation has been quite difficult to decipher given the complexity of their regulation and the fact that there are two temporally distinct levels of regulation. Initial work on the primary and secondary pair rule classifications has been called into question based on various criteria (Nasiadka et al., 2002). For instance ectopic expression studies have indicated that the secondary pair rule gene *odd* can regulate all the primary pair rule genes at early time points (Saulier-Le Drean et al., 1998). Similarly it has been pointed out that the early *ftz* pattern is not a simple consequence of regulation by the primary pair rule genes (Yu and Pick, 1995). However given the difficulty of studying such complex regulation, genetic work to study this process has grown increasingly difficult and hard to interpret although some progress has been made (Jaynes and Fujioka, 2004). Here we seek to take an alternate approach based on completing the dissection of the *cis*-regulatory elements of all the pair rule genes to clarify how the pattern is encoded in each gene.

1.3 Algorithms used

The computational portion of this work attempting to predict *cis*-elements rests on two related programs, Ahab (Rajewsky et al., 2002) and Stubb (Sinha et al., 2003), developed in the Siggia lab. As used here they differ only in implementation although Stubb can also analyze pairwise alignments and utilize evolutionary information. Although Ahab was developed first and used initially, later work used Stubb when it became available as it is a superior

implementation. A thorough description of how these algorithms function is outside the scope of this introduction, but a brief description is given to familiarize the reader with the method and explain why they are a good choice of tool. For simplicity the name Ahab is used to describe both algorithms as Stubb is identical to it as far as the features outlined here.

Ahab is based on a statistical description of a binding site called a position weight matrix or PWM (Stormo, 2000). The matrices describe the expected frequency of each base at each position based on a set of aligned binding sites. This formulation assumes independence of binding preference for nucleotides within a binding site, which seems adequate in most cases (Benos et al., 2002). Given independence, the probability of a sequence is simply the product of the probability of the bases at each position in the PWM. The probabilities used in the PWM can be estimated by the frequency of each nucleotide at each position in a set of known aligned sites. As the sum of the score over all possible sequences is one and many sequences are possible, any given sequence has a relatively low probability. The typical approach to using PWMs is to use a log odds ratio, which is a ratio of the probability from the PWM to the probability of seeing the sequence at random. One then sets a threshold for each PWM to be analyzed independently and then integrates the predicted sites in a second post processing step. The models constructed in such a fashion are often *ad hoc* and based on a set of empirical rules.

Ahab in contrast uses an integrative model that analyzes larger sequences containing clusters of binding sites in a unified probabilistic framework. Markov

Chains are a class of statistical model for describing a series of observed states, in which the next state depends solely on the current state. The probability of the next state is parameterized by a transition probability. The algorithms used here are based on an extension of these models called Hidden Markov Models (HMMs). In these models a series of observations occur in which the state generating them is hidden. In this case each hidden state can "emit" the observations with a certain probability. Similar to the Markov Chain, the transitions between the hidden states are described by a set of transition probabilities. A nice explanation this class of models is presented in "Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids" (Durbin et al., 1998).

The fact that the same set of observations can be explained by multiple different states in an HMM, generates a situation where there is not a unique mapping of the states to the set of observations. The process of assigning the set of hidden states to a given sequence is called decoding. There are two basic approaches, decoding based on the most likely set of hidden states or an alternative called posterior decoding. In posterior decoding dynamic programming is used to estimate the probability of a given sequence over all possible mappings of the observations to the hidden states. Ahab uses posterior decoding allowing a robust estimate of site probabilities regardless of assignment of the hidden states that allows overlapping binding sites from different factors to all contribute to the final probability. Given that overlapping sites occur in this system, this is an important feature of the method.

One intuitive example of HMMs is the tossing of two indistinguishable six sided dice, one of which is loaded to give 6 more frequently at the cost of the other choices. If one knows the frequency that each die generates a six (the emission probability) and how often the tosser switches the dice (the transition probability), HMMs can be used to calculate which die was more likely to be in use for each toss over a series of tosses. In the case of *cis*-element prediction, the sequence is that of DNA, and the hidden state is whether a binding site is present at a particular location or not. Within a site the PWM determines the probability of a given base and the probability that a PWM occurs in a given location depends on the transition probability assigned to the PWM.

The model behind Ahab is based on the assumption that the presence of a binding site alters the probability of each base occurring at a given position in the site depending on the preferences of the binding factor. In the case of the dice, a normal fair die can be considered the "default" model for what a sequence of roles should look like. There is no such clear default in DNA sequences. As the entire sequence does not correspond to binding sites, an additional background model is included. The background model is somewhat simpler than a PWM in that it is simply the frequency of seeing a given base at the current position based on the previous k-1 nucleotides in the sequence. This model is called a Markov Model and the number k is the order of the model. Therefore single nucleotides correspond to order zero, pairs to order one, and so forth. By default both Ahab and Stubb generate the background model from the sequence under analysis and additional flanking sequences. Stubb can also generate "global" model from any set of provided sequences.

In order to predict whether or not an element exists, Ahab and Stubb compare the probability of the sequence being generated by the background model alone versus a model including both the background model and the set of PWMs for the transcription factors of interest (Figure 5). The improvement in the probability of observing the sequence when including the set of PWMs over the background alone is quantified by the free energy, which is the log of the ratio of the two probabilities. The term free energy was chosen based on the similarity of the algorithm to the calculation of free energy in statistical mechanics. There are theoretical arguments for a relationship between the binding of the transcription factors to the DNA and the score calculated by Ahab and Stubb. This gives the general intuition that higher the scores correspond to sequences that would be more extensively bound by the transcription factors parameterized by the PWMs.

One particularly nice aspect of this approach is that it integrates the strength and number of multiple sites for multiple factors into the free energy score, which allows meaningful ranking of predictions within a region unlike other approaches based on PWMs or matching of consensus patterns (Berman et al., 2002; Berman et al., 2004; Markstein et al., 2002). The algorithm also calculates the posterior probability of each putative site within the window using posterior decoding. This process calculates the probability a given PWM (or background) generated a particular sequence normalize by the probability of all possible ways the sequence could have been generated. Therefore, the sum of all posterior probabilities for all factors at this site sum to one and this probability is

a measure of the likelihood of a given binding site versus all other members of the dictionary under consideration. The sum of the posterior probabilities for each possible site of a factor over the entire sequence, called the dictionary score, is a robust estimation of the occurrence of sites for the entire region based on all possible assignments of sites in the sequence. Other approaches that utilize cutoffs and score sites independently have no natural way to combine the set of inputs into a common measure in this fashion. By having a single score for all sites predicted in the sequence, it is more straightforward to compare the input from a given factor into different *cis*-elements than when predicting sites independently.

Therefore the Ahab and Stubb algorithms are ideal for predicting *cis*-elements as the best predictions are clear from the free energy. Further given a set of elements, they are ideal for comparing the strength of input from different factors into the element. These features make this class of algorithm an ideal tool for the goal of dissecting and better understanding regulatory elements that have a complex set of known inputs as is the case in the segmentation network.

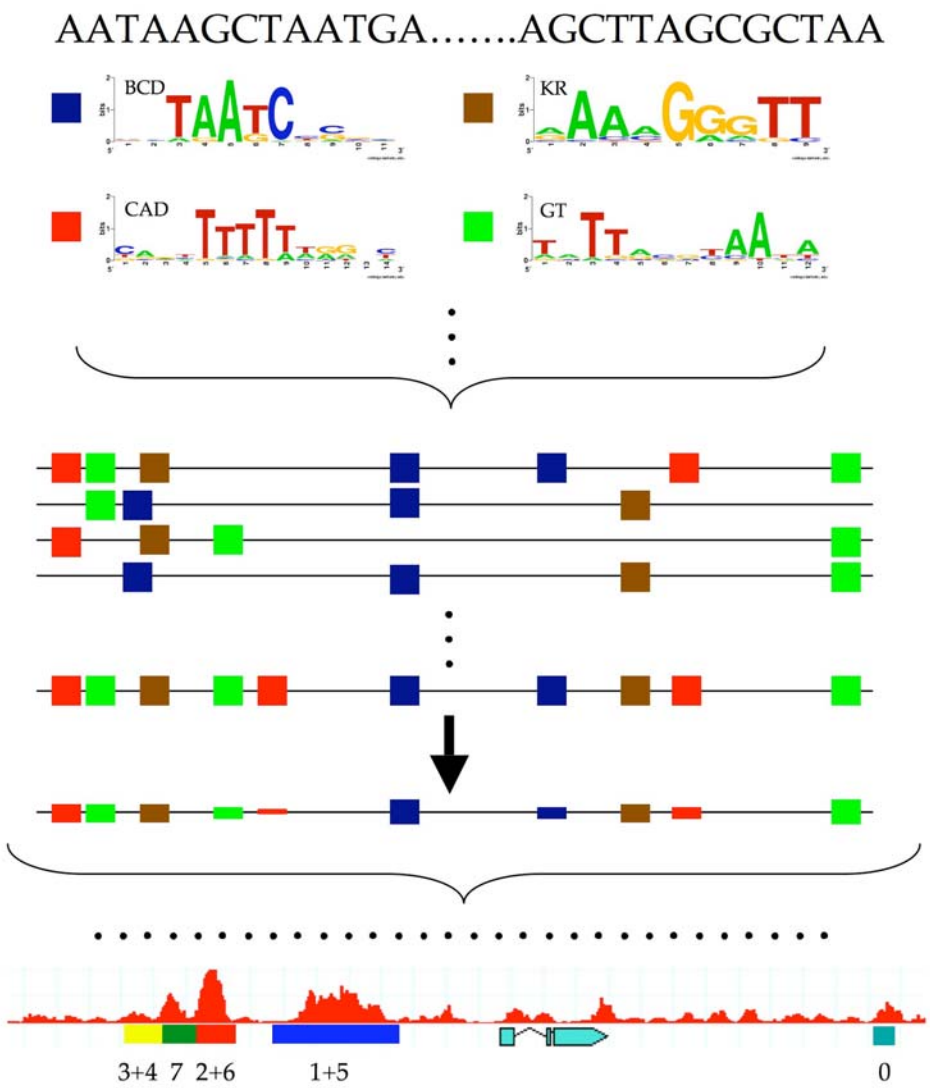


Figure 5

A schematic of the logic behind the Ahab and Stubb algorithms. The algorithm takes a window of sequence and a set of PWMs as input (top). Dynamical programming is used to calculate the probability of generating the sequence over all possible ways of assigning the sequence to the different PWMs and the background (middle). The free energy is a measure of how much more likely the sequence is when the PWMs are included in the calculation. The free energy as a function of position in the genome is called the free energy profile and is plotted in red for the *h* locus (at bottom). If the PWMs do not contribute at all the free energy is zero. The score of the 2+6 element depicted as a red rectangle below the free energy profile is above thirty and therefore roughly 10^{12} times more likely than the background model alone.

1.4 Organization

Although the segmentation system has been extensively studied, there are still many unknowns. The original dissections of the segmentation genes did not determine all the elements necessarily to establish the expression patterns of many factors including some at the core of the pathway. Determination of *cis*-elements composition and their binding site composition clarifies genetics, where indirect effects often complicate interpretation of pattern changes in mutants.

Chapter 2 describes the use of Ahab to drive experimental dissection of a large number of *cis*-elements throughout the segmentation hierarchy including the two important gap genes *gt* and *kni*. In the course of this work the discovery of stripe specific elements in *odd*, which are inconsistent with its original classification as a secondary pair rule gene, led to interest in revisiting the pair rule classification system. Therefore Chapter 3 revisits the pair rule hierarchy driven in large part by an attempt to complete the cloning of the stripe specific elements that generate its pattern. The complexity of pair rule regulation is well suited to a reductionist approach, as the function of individual components clarifies the inherent complexities of the process. Finally Chapter 4 looks at the composition of the maternal and gap regulated elements of both the gap and pair rule genes to better understand the organization of these two tiers in the hierarchy and how they relate.

Overall the goal of the work is to use computational methods to better understand the segmentation network. Of particular interest is how patterning

makes use of hierarchical relationships to build up the patterns that underlie segmentation. The transition from the nonperiodic gap gene patterns to the periodic patterns of the pair rule genes involves an impressive jump in complexity and refinement that is still poorly understood.

Chapter 2: Validation of Ahab for *cis*-element dissection

In the original paper describing Ahab (Rajewsky et al., 2002) a number of different approaches to finding *cis*-regulatory elements were described. The different approaches varied in the amount of prior information used to discover the elements. The publication focused on algorithmic issues and did not do a validation of the efficiency of the algorithm in predicting *novel cis*-elements although it clearly was effective in recovering *known* elements. Therefore there were a number of questions left open about the ability to use Ahab as a tool to dissect regulatory networks. As there were just starting to be publications of this type following up on the publication of the *Drosophila melanogaster* genome (Myers et al., 2000; Rubin et al., 2000), this was a relatively novel field of inquiry in general and few other studies to compare to.

The approach of predicting *cis*-elements using Ahab together with well defined PWMs utilized the most prior knowledge and was therefore the most specific and effective. Therefore it was the natural starting point for validating the algorithms that had been developed. Could Ahab predict new elements effectively within the known genes? Also beyond validation, there was a keen interest in understanding how to use Ahab to address some of the outstanding questions in the field. How is positional information encoded in the elements? How is cross-regulation across the tiers of the hierarchy organized?

2.1 Genome wide analysis

The original development of Ahab and its use genome wide was motivated by the availability of the *Drosophila melanogaster* genome. Genomics have been transformational in biology and genome wide approaches were an exciting area of study. The idea of finding new important segmentation genes by searching through the genome was a very exciting idea. Could Ahab find genes that were missed by the original genetics? Although the original screen papers (Jurgens et al., 1984; Nusslein-Volhard et al., 1984; Wieschaus et al., 1984) and later work on pair rule genes (Vavra and Carroll, 1989) presented good genetic evidence that the screen was saturating, genetics can miss genes where various forms of redundancy exist. Also the initial screen threw out a lot of genes with variable phenotypes that are still likely to play a role in segmentation.

To test the ability of Ahab to predict novel genes patterned by the segmentation network, *in situ* hybridizations were carried out for predictions from the predictions published in the initial Ahab publication. Five of the nineteen genes analyzed are patterned, for a rate of 26% in this set (Figure 6). At this time the Berkeley Drosophila Genome Project (BDGP) also released their first set of *in situ* data (Tomancak et al., 2002), which was a useful resource in that it contained a large data set of expression patterns that were generated in an unbiased fashion that could be compared to our list of genes. The BDGP database at the time contained 237 blastoderm patterned genes out of 2,993 assayed. This sampling covered roughly $1/5^{th}$ of the genome and had a rate of

8% of genes showing blastoderm patterns. In the BDGP set there were also 5 patterned genes out of a total of 28 from the list of predictions giving a rate of 18%. Combined the two sets have 10 patterned genes out of 47, for a rate of 21% (Table 2). By Poisson statistics the probability of predicting that fraction of patterned genes based on an 8% rate of patterned genes is 5.4×10^{-3}, indicating Ahab predicts patterned genes at a significant rate.

Examining the *in situ* hybridization data from the top hits, the novel predictions were typically weaker, a more complex mix of a-p and d-v patterns, and were not transcription factors. Although the role of more peripheral genes in establishing body pattern is interesting, the goal of the project was to use Ahab to better understand transcriptional regulation within the network and were therefore peripheral to the goals of the project and not studied further. Although Ahab was clearly predicting targets of the segmentation network genome wide, the rate and type of molecules predicted encouraged a shift towards regulation of the core segmentation components discovered in the original segmentation screens.

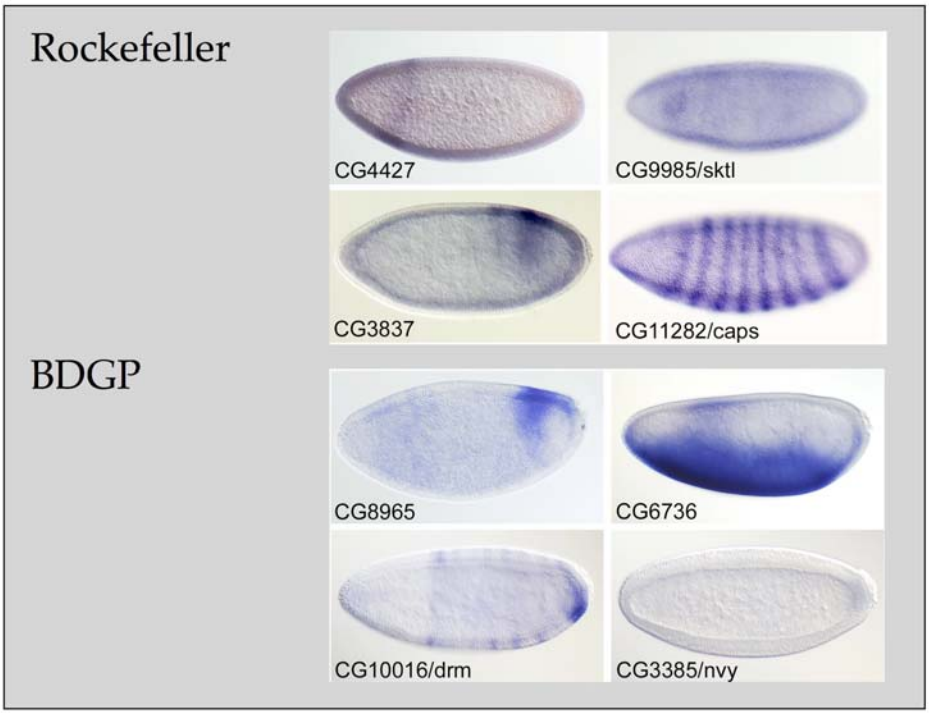


Figure 6

Patterned genes adjacent to genome wide Ahab predictions. The locations of the predictions for the genes are as follows: CG4427 - intragenic, *sktl* - 8kb upstream, CG3837 - 240 bp downstream, *caps* 21 kb upstream, CG8965 - intragenic, CG6736 - 5 bp upstream, *drm* - 7 kb upstream, *nvy* - 5 kb upstream. As elsewhere all embryos are displayed with anterior at left and the dorsal side at top.

Rank	Gene	CG no.	Position	Location	Multiple Hits	Source	Blastoderm Expression
6	Cyp6v1	CG1829	1630	downstream	rank 14	RU	NE
10	Sox21b	CG6419	41968	upstream		BDGP	NE
11	CG4871	CG4871	1253	upstream		RU	NE
12	CG7526	CG7526	3250	upstream		BDGP	NE
14	Cyp6v1	CG1829	5562	upstream	rank 6	RU	NE
16	Lar	CG10443	1749	upstream		RU	NE
19	ome	CG17705	17728	downstream		BDGP	NE
21	CG10191	CG10191	2455	upstream		BDGP	ubiquitous
22	mRpS26	CG7354	2531	upstream		BDGP	ubiquitous
26	CG4730	CG4730		exon 2		RU	NE
28	CG12604	CG12604	6103	downstream		BDGP	NE
29	drm	CG10016	6146	upstream		BDGP	pair rule
31	CG8965	CG8965		intron 1		BDGP	ap/dv patch
38	CG2083	CG2083	4080	upstream		BDGP	NE
40	svp	CG11502		intron 2		BDGP	NWR
42	caps	CG11282	20891	upstream		RU	pair rule
49	Dab	CG9695	635	upstream		RU	NE
55	ed	CG12676	57647	upstream	rank 93	RU	NE
56	Fur1	CG10772	21144	upstream		BDGP	NE
57	CG8586	CG8586	12	upstream		RU	NE
58	CG2118	CG2118		exon 5		RU	NE
59	CG9759	CG9759	356	downstream		BDGP	NWR
64	Sdc	CG10497		intron 2		RU	NE
68	CG1907	CG1907	10641	upstream		BDGP	NWR
69	faf	CG1945		intron 17		BDGP	ubiquitous
70	CG5151	CG5151	12484	upstream		RU	NE
71	NetA	CG18657	14574	upstream		BDGP	d/v stripe
80	CG9892	CG9892	249	upstream		RU	NE
82	RpS12	CG17672	2702	upstream		BDGP	ubiquitous
84	PFE	CG15151	34643	upstream		BDGP	NE
85	rk	CG8930	13815	upstream		BDGP	NWR
85	bgm	CG4501	5966	upstream		RU	d/v stripe
87	CG6736	CG6736	8	upstream		BDGP	ap/dv
92	CG3837	CG3837	241	downstream		RU	dv/ap patch
93	ed	CG12676	16132	downstream	rank 55	RU	NE
94	bab1	CG9097	53660	downstream		BDGP	NWR
95	CG11337	CG11337	18925	upstream		BDGP	NE
95	Gprk2	CG17998	9075	upstream		RU	NE
97	CG12870	CG12870	1121	downstream		BDGP	NWR
101	CG8782	CG8782	3558	downstream		BDGP	NE
108	nvy	CG3385	4601	upstream		BDGP	ap/dv
115	CG4427	CG4427		exon 2		RU	gap-like
128	CG9586	CG9586	27093	upstream		BDGP	NWR
139	CG3622	CG3622	2212	upstream		BDGP	NWR
144	sktl	CG9985	8123	upstream		RU	ap/dv patch
145	CG15160	CG15160	19676	upstream		BDGP	NWR
146	grp	CG17161	9898	downstream		BDGP	ubiquitous

Table 2

Table of genome wide Ahab predictions with *in situ* data. In the blastoderm expression column NE stands for not expressed. NWR stands for not worth revisiting, a description used for BDGP *in situ* data where the gene was either ubiquitously expressed or not expressed.

2.2 Overview of the Segmentation *cis*-element screen

Another obvious application of Ahab was in dissecting additional genes known to have a role in segmentation. The set of PWMs chosen for *cis*-element prediction consisted of BCD, CAD, TOR-RE, DSTAT, HB, KR, KNI, GT, and TLL. This eliminated the matrix for DORSAL, a primary transcriptional regulator of d-v fate in favor and added DSTAT and GT, which were also important a-p regulators not used in the original genome wide analysis. Of these matrices KNI and TLL were relatively non-specific and clearly over predicted sites, whereas DSTAT and GT were overly specific and seemed to be under predicted.

The primary targets of this set of transcription factors are the gap and pair rule genes, so a set of 29 segmentation genes with gap and pair rule patterns from the known segmentation genes was chosen for analysis. In addition, a further annotation of additional segmentation enhancers from the literature for this set of genes was carried out resulting in the addition of a number of core target elements such as *eve* stripe 1, *run* stripe 1, *h* stripe 2, *Kr* AD2, *ftz* ps4, and the *oc* early element. All known *cis*-regulatory elements in the gap and pair rule genes fall within the 20 kb upstream and 10 kb downstream of the target gene or the region defined by the two adjacent genes. Therefore the study focused on this region around the known segmentation genes. This set of 29 transcriptional control regions included roughly 750 kb of sequence as well as the set of *cis*-elements delineated for the analysis are given in tables within the Appendix.

A cutoff of 15 was chosen for the free energy score based on the distribution of free energy scores for the known *cis*-elements as it gave the best recovery of known elements as a fraction of total predictions. The cutoff was approximately 4 standard deviations above the genome wide mean and resulted in 52 predictions (Figure 7). These predictions recovered 22 of the 31 known *cis*-elements with 20 different predictions (some predictions overlapped multiple modules). An additional 3 elements had clear peaks in their free energy profiles that were just below the cutoff (e.g. white arrowhead Figure 7f), leaving only 6 modules with no clear sign of signal. In the cases where the elements are not predicted, the issue is likely to be in part that they are regulated largely by additional factors not included in the set of PWMs resulting in lower signal. The *hkb* ventral element, has important input from the d-v factor *dorsal*, which was not included in the set of PWMs used. In the *ems* head module, only two BCD and two TLL sites were determined experimentally suggesting a relatively small amount of input from the factors included in the set of PWMs. Given the poor quality of the TLL matrix, the fact that no TLL input is predicted is an additional reason the element is not recovered.

Despite missing some elements, more than two thirds were recovered indicating that Ahab is an effective predictor of elements within the network. One way to determine significance is to compare the correspondence between the actual predictions to a similar number of random predictions. To minimize penalizing predictions that might be functional and avoid selection bias, only control regions previously characterized in the literature were included in the analysis. As different extents of overlap between the predictions and the known

elements are important, the number of times where a similar level of overlap to the actual predictions was achieved in comparison to the randomized predictions was tracked.

When including all elements and previously dissected genes, equivalent or greater overlap was not seen in 10^8 randomizations, indicating a correspondingly small probability of getting such results at random. As some of the PWMs were generated by sequences within some of the known *cis*-elements the significance could be partly attributable to the sites used to construct the matrices. To control for this issue, the corresponding elements and predictions were left out. In genes where this included all known *cis*-elements, *kni*, *hb*, and *tll*, the entire control region was left out of the comparison. In this case the frequency of seeing equal or better overlap was 4.9×10^{-6} indicating the algorithm does not simply recover the sequences input through the PWMs.

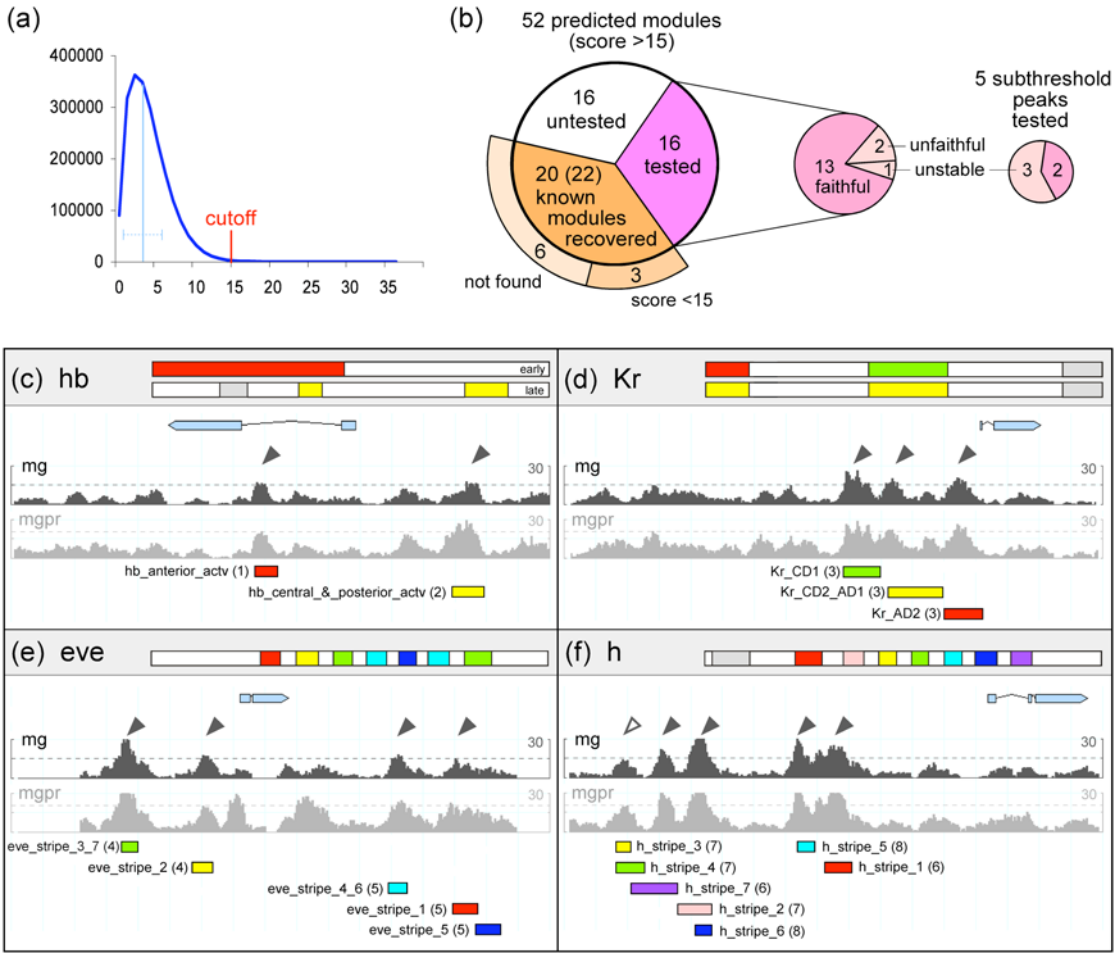


Figure 7

Overview of *cis*-element predictions. (a) Histogram of free energy scores genome wide with cutoff (red line), mean (blue line), and standard deviation (dashed blue line) marked. (b) Pie chart summarizing the predictions over the set of 29 genes chosen for analysis. (c-f) The control regions for selected gap and pair rule genes and the free energy profiles shown for both the mg and mgpr runs. A dashed line marks the free energy cutoff for each run with the predictions marked by black (above threshold) and white (below threshold) arrowheads. Colored bars depict the cis-elements. The color coding corresponds to the portions of the endogenous pattern they generated, which is schematized in the header (anterior on left, posterior on right). References for the known elements: (1) (Schroder et al., 1988), (2) (Margolis et al., 1995), (3) (Hoch et al., 1990), (4) (Goto et al., 1989), (5) (Fujioka et al., 1999), (6) (Riddihough and Ish-Horowicz, 1991), (7) (Howard and Struhl, 1990), (8) (Langeland et al., 1994).

In addition to recovering known elements, an additional 32 putative elements were predicted, half of which were tested. Five subthreshold predictions were also tested, to make a total of 21 tested elements. The predictions were cloned into the Casper hs43GAL transformation vector (Thummel and Pirrotta, 1991), which contains a LacZ reporter and the hs43 basal promoter. The Casper p-element transformation vectors allow random insertion of the element into the genome when co-injected with a transposase helper plasmid into flies prior to cellularization. To control for insertional effects, where flanking genomic sequence altered the expression pattern of the reporter gene, at least three lines for each construct were tested. Except in a few cases where noted the expression patterns were quite consistent between lines.

Of the 16 above threshold elements tested, 13 drove proper pattern for a success rate of over 80%, which suggests additional untested predictions are likely to be functional as well (Figure 7). Five subthreshold free energy peaks were also tested, two of which drove proper pattern, for a success rate of 40%. Although the marked drop in success rate indicates the cutoff was well selected, there is still clearly signal in the subthreshold peaks. It is notable that all elements drove blastoderm expression indicating that in all cases there was input within the element driving expression at this stage. The improper patterns could be classified into two types: unfaithful modules where the boundaries of pattern were improperly delineated and unstable elements that showed insertion dependent variation. Of the three above threshold predictions that were improperly expressed, 2 were unfaithful and one was unstable, whereas all of the

subthreshold elements were unstable. This suggests that as increasing signal is lost and the element includes less of the normal regulators there is a transition from the *cis*-element acting as a coherent unit to it interacting with inputs within the region of insertion.

Given the additional 15 elements found in the analysis, the total number of maternal and gap regulated *cis*-elements was taken from 31 to 46, an increase of almost 50%. As a central issues in understanding *cis*-element function is the relationship between input and output, having additional elements expressed in different locations with different complements of sites is of great utility. Following a description of the expression of the various constructs, the analysis of the enlarged set of *cis*-elements for compositions rules with Ahab will be presented.

2.3 *cis*-dissections of segmentation gene control regions

The gap gene *gt* has a fairly complex pattern for a gap gene, although it has never been dissected. Early in the syncytial blastoderm stage, *gt* is expressed in a strong, broad domain in the head and a narrower domain in the posterior. The anterior domain splits into two stripes later as cellularization begins. Then, towards the end of cellularization, a new domain of expression arises at the anterior pole. There are 3 predictions in the *gt* region, *gt* -1, *gt* -3, and gt -6. In addition there is a broad subthreshold peak, which was tested, *gt*-10 (Figure 8a). These four elements can account for all domains of *gt* expression. The *gt* -3 element corresponds to the posterior domain, the *gt* -6 element to the later far

anterior domain, and the *gt* -10 to the broad anterior domain. In addition the *gt* -1 element produces both the anterior and posterior domains from a single element.

Among our predictions were some in head gap genes. *cap 'n' collar* (*cnc*) is a head gap named for its expression pattern, which consists of an anterior cap and more posterior collar. The *cnc* +5 element generates both domains of expression (Figure 8b). The gene *ocelliless* (*oc*), which corresponds to the head gap mutant *otd*, has a single anterior domain of expression at the blastoderm stage. An element that generates this pattern that falls roughly 4kb upstream of the basal promoter was previously described (Gao and Finkelstein, 1998), but is not a significant prediction. Within the first intron we have a significant prediction, the *oc* +7 element (Figure 8c), which also generates the same pattern indicating multiple elements contribute to this pattern.

The gene *Dichaete* (*D*), is expressed in both a broad domain through the trunk region and a head patch. Our *D* +4 element is the only significant prediction in the region and recapitulates the broad trunk domain (Figure 8d). The gene *cad*, in addition to being an important maternal morphogen, has a later expression domain corresponding to its role as a homeotic gene in the most posterior segment (Moreno and Morata, 1999). The *cad* +14 element, which is the only significant prediction in the region, recapitulates this late stripe of expression (Figure 8e).

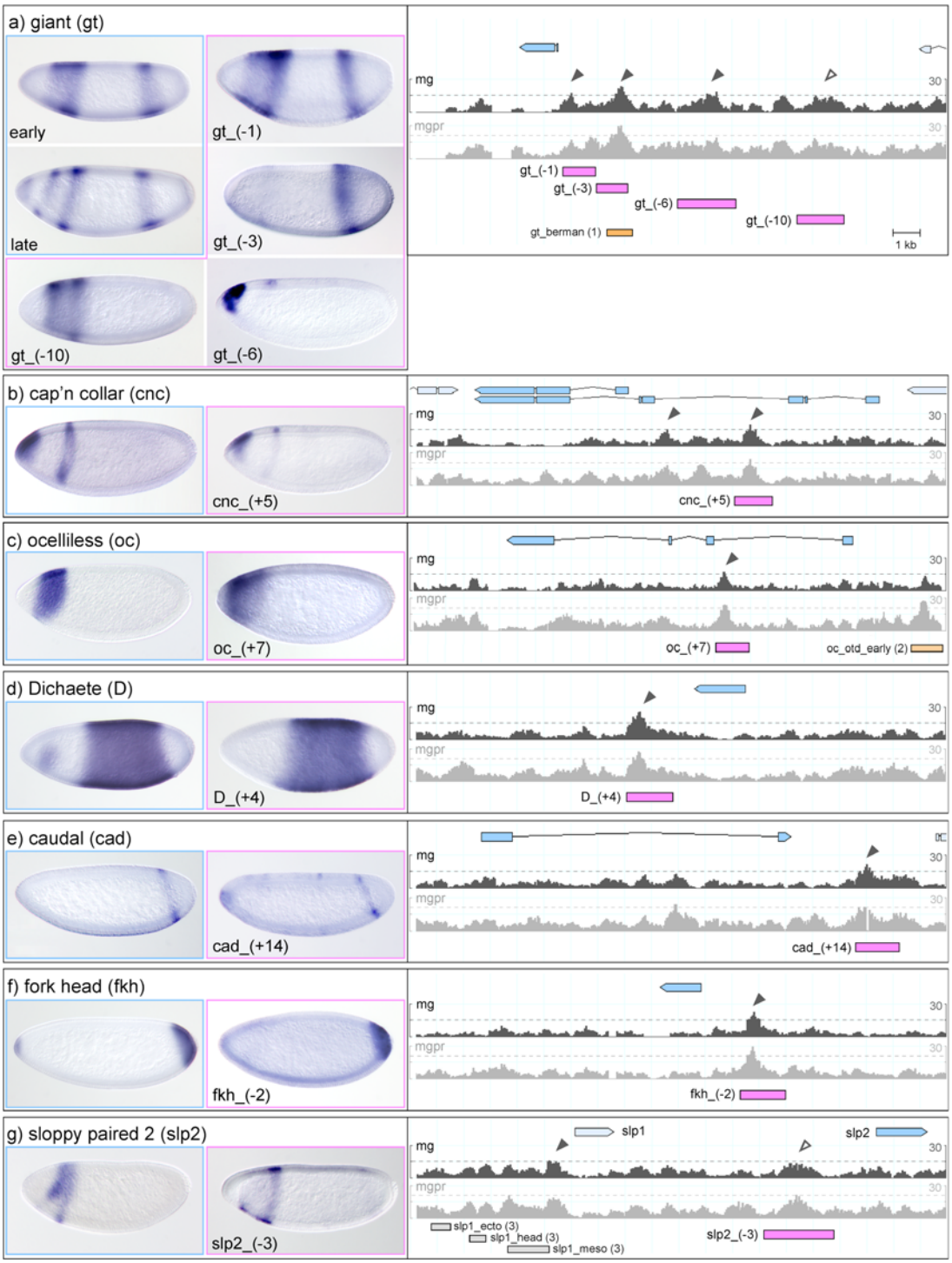


Figure 8. (figure legend on p. 43)

Figure 8

Ahab driven *cis*-element dissections of segmentation control regions. At left *in situ* hybridizations for both endogenous expression pattern of the gene (framed in blue) and the expression patterns of the LacZ reporters (framed in pink). Embryos are oriented with the anterior to the left and the dorsal side at top. At the right, the genomic region with the gene marked in blue and the tested *cis*-elements marked in pink to match the framing of the corresponding embryonic expression patterns. The free energy profiles for the maternal and gap, or mg, run is shown in black and the maternal, gap, and pair rule, or mgpr, run in grey. The free energy cutoff for each run is shown with a dashed line on the free energy profile. In the mg run above threshold predictions are marked by black arrowheads and below threshold peaks that were tested are marked with white arrowheads. Previously known elements are marked in orange, if they are driven by maternal and gap input, or grey, if they are driven by pair rule input. References: (1) (Berman et al., 2002), (2) (Gao and Finkelstein, 1998), (3) .(Lee and Frasch, 2000).

The secondary pair rule genes paralogs *slp1* and *slp2* are adjacent in the genome and expressed in similar patterns. Their patterns initiate as a broad head domain during the syncytial blastoderm stage, before generating a seven-stripe, and then segmental pattern as development progresses. Although *slp1* had previously been dissected, the gene *slp2*, had not. The, *slp2* -3 element, corresponds to a broad subthreshold peak in the free energy profile upstream of *slp2*, which generates a head domain similar to the early *slp2* expression pattern (Figure 8g).

Although the gap gene *kni* had previously been dissected (Pankratz et al., 1992; Rivera-Pomar et al., 1995), the sequence generating the anterior domain had not been mapped. The *kni* -5 element generates the anterior domain (Figure 9a), which is patterned along both the a-p and d-v axes. Despite the strong d-v bias in pattern, the element is a significant prediction using only a-p PWMs. The *kni* +1 element, which falls in the intron, drives expression that corresponds to both domains of *kni*, but is inappropriately restricted to the proper pattern. The anterior domain lacks input that would confine it along the d-v axis, whereas the posterior region is artificially broad along the a-p axis. The free energy profile also shows a clear peak corresponding to the previously determined *kni* kd element, which generates the posterior domain. The gene *knirps-like* (*knrl*) has a pattern very similar to that of *kni*, but with a weaker posterior domain. The *knrl* +8 element generates an improper pattern that is not delimited correctly in the head and includes pair rule like striping not seen in the endogenous *knrl* pattern. There is another above threshold prediction in the gene, also within the first intron, which was not tested.

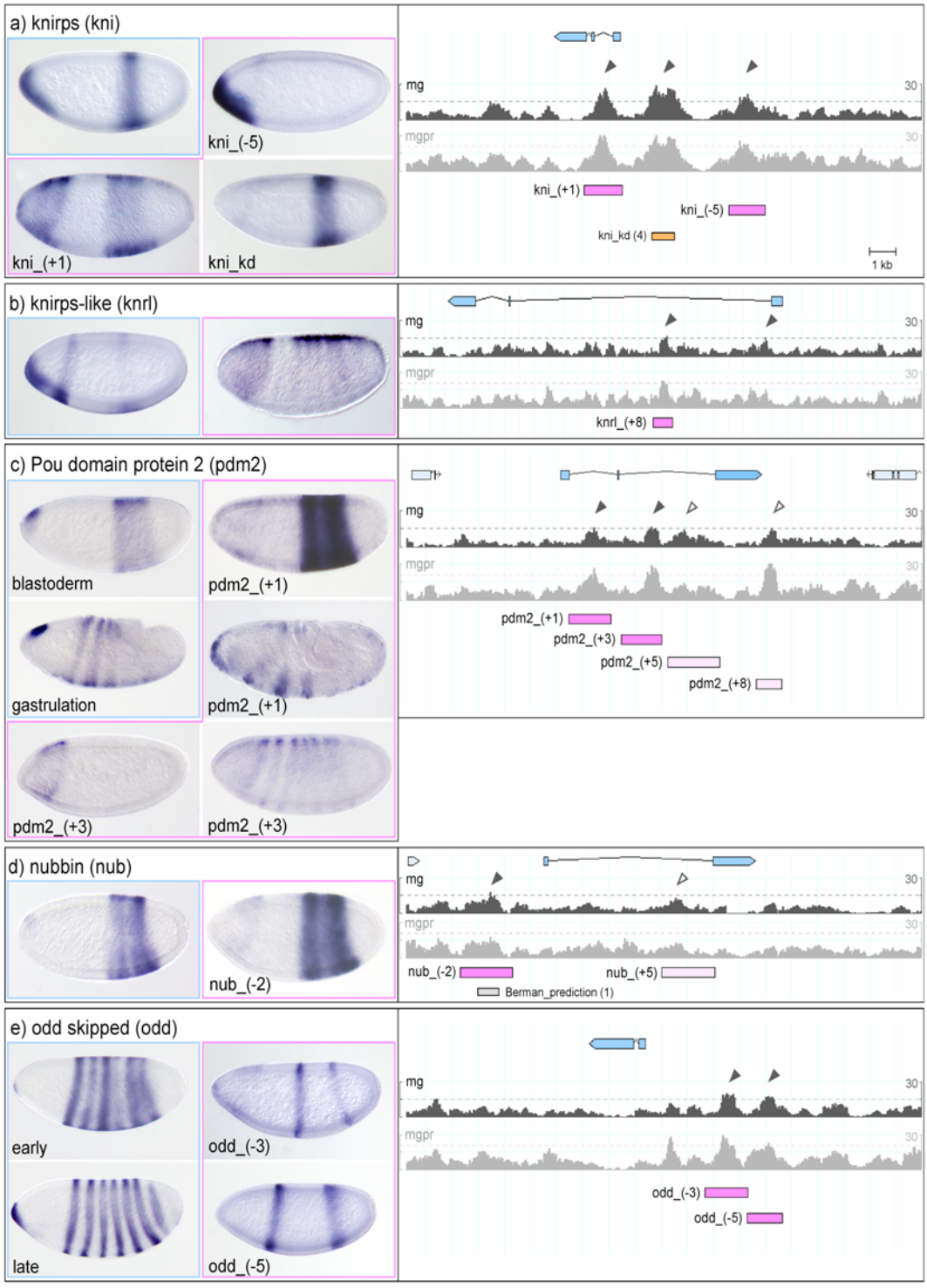


Figure 9

Additional *cis*-elements. See legend for Figure 8. (a) *kni*, (b) *knrl*, (c) *pdm2*, (d) *nub*, (e) *odd*. References: (4) (Pankratz et al., 1992) and (Rivera-Pomar et al., 1995)

The genes *Pou domain protein 2* (*pdm2*) and *nubbin* (*nub*), have been implicated in having a gap gene role in segmentation through ectopic expression, but do not have strong the typical gap gene phenotype (Cockerill et al., 1993). Both genes are initially expressed in a broad posterior gap like domain and an anterior head patch. Later the *pdm2* pattern develops into a series of stripes that are more strongly expressed in the anterior. In the *pdm2* gene four elements were tested (Figure 9c), two above threshold peaks, *pdm2* +1 and +3, and two below threshold peaks, *pdm2* +5 and +8. The *pdm2* +1 peak recapitulates the early and late *pdm2* pattern, whereas the *pdm2* +3 and the subthreshold elements generate variable line dependent patterns (Figure 9c). The *nub* -2 element recapitulates the endogenous *nub* pattern, whereas an additional subthreshold peak *nub* +5 generates variable line dependent patterns.

It is perhaps notable that there are two types of defect in pattern that are seen in the non-functional elements. The unfaithful *cis*-elements, which generate improper, but consistent spatial patterns, are clearly missing inputs, but still function autonomously. This maintains the important property seen in all the functional elements of being buffered against inputs into the region surrounding the transgene insertion site. In contrast the line dependent variable patterns are susceptible to differences in the genomic region surrounding the insertion site. In both cases there are likely to be missing inputs, but the difference in outcome suggests a subtle differences in what is missing with important functional consequences.

The secondary pair rule gene *odd* is initially expressed in a seven-stripe pattern that transitions to a segmental pattern as cellularization completes. There were two predictions, *odd* -3 and -5 (Figure 9e), that surprisingly functioned as stripe specific elements similar to those known to regulate primary pair rule genes. The *odd* -3 element generates stripes, 3 and 6. The *odd* -5 element drives stripe 1 and a broad stripe 5. It is generally thought that the secondary pair rule genes generate their seven-stripe pattern through regulation of a seven-stripe element by the primary pair rule genes. However, these two elements make clear that the majority of *odd* stripes are driven in a fashion similar to the primary pair rule genes.

In total the dissections demonstrate that Ahab is an effective algorithm for predicting *cis*-elements and allows efficient dissection of the regulatory regions of known segmentation genes. As the regions included in the reporters were defined by the free energy profile and generated proper pattern in most cases, this indicates Ahab also provides important information in delineating the extent of *cis*-regulatory elements. Some of the predictions like those in *odd* further highlight the role of certain regulatory connections that were less clearly understood from the initial genetic characterizations.

	posterior probability > 0.25		posterior probability > 0.5	
Factor	Recovery	PWM specificity	Recovery	PWM specificity
DSTAT	2/2	0.88	2/2	0.81
KR	13/23	0.78	12/23	0.60
TOR-RE	4/4	0.78	4/4	0.64
BCD	24/39	0.75	21/39	0.60
HB	30/43	0.70	22/43	0.45
GT	4/6	0.69	3/6	0.54
CAD	12/21	0.59	10/21	0.43
TLL	11/17	0.52	8/17	0.34
KNI	14/27	0.48	11/27	0.29

Table 3

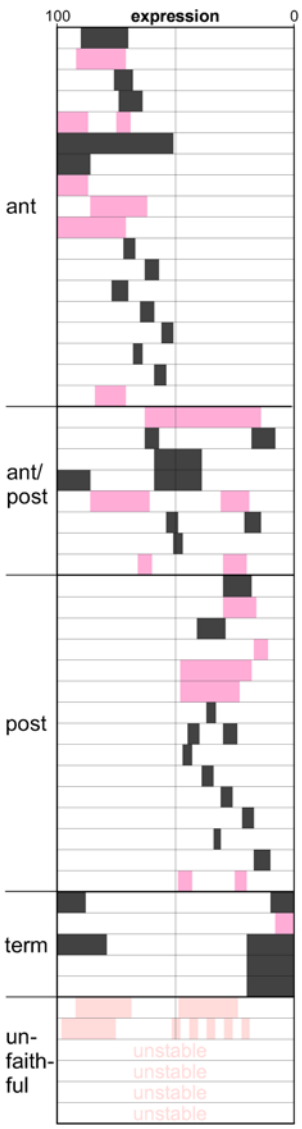
Recovery of known binding sites. The table shows the fraction of sites recovered by Ahab, with posterior probability cutoffs of 0.25 and 0.5. Given the much higher probability of labeling sequence as background, even a cutoff of 0.25 is relatively stringent. The PWM specificity is a measure of what fraction of the Ahab dictionary value is above the posterior probability cutoff. For instance the value 0.6 in column 5 for KR means that 60% of the sum of sites over all posterior probabilities is generated by sites with a probability greater than a half.

2.4 Binding site composition and pattern of *cis*-elements

That Ahab predicts modules so efficiently based on binding site clustering indicated a robust prediction of binding sites. Footprinting of the known *cis*-elements was never done in a comprehensive fashion, so the relationship between the patterns of the elements had not been systematically compared to the expression patterns they generated. Given the enlarged set of *cis*-elements, it was of great interest to use Ahab for such an analysis. To verify that Ahab predicted binding sites effectively, recovery of the footprinted sites by Ahab was

verified by looking at sites above a given posterior probability threshold (Table 3).

As discussed in the introduction, Ahab produces a set of dictionary values for each PWM based on the sum of the posterior probability of all sites within the given sequence. This number integrates the total strength and number of sites in a single value, which is robust to prediction of weak sites because they are given very low weights. A PWM describes the probability distribution over all sequences of a given length. More specific PWMs assign fewer sequences higher probabilities, while less specific PWMs assign more sequences lower probabilities. As can be seen in the table showing site recovery for Ahab, the matrices KNI and TLL predict more confidence sites of low confidence, whereas a matrix like DSTAT predicts more than 80% of the sites with greater than 50% confidence. This gives some indication of what proportion of the dictionary values are of high versus low confidence. As can be seen in the table for KR, which is relatively specific, a good fraction of the known sites fall below a 0.25 posterior probability, indicating that some of the lower confidence sites are functional. As Ahab already provides an estimate of input that is weighted by confidence, it does not make sense to threshold. In the case of specific PWMs it would not affect the estimates significantly, but would lose functional sites. In contrast in the case of unspecific matrices it would throw out a large proportion of the signal.



gene	module	=mel	=pse	size	Δ gene	Dstat	TorRE	Tll	Bcd	Hb	Gt	Kr	Kni	Cad
oc	oc_otd_early	(X)		1838	-3248			8.23	5.95	4.39	1.88	5.63	8.96	
oc	oc_(+7)	X	X	1833	7601		1.56	12.46	13.32		0.94		8.89	
ems	ems_head			304	-7112								3.36	
btd	btd_head	X	X	1800	-2801		0.91	14.80	14.33	2.88		5.46	4.09	3.75
cnc	cnc_(+5)	X	X	1421	-3966		2.24	13.39	5.68	4.00	2.12	2.97	1.23	
hb	hb_anterior_actv	X		721	3191			3.86	6.65	4.07	3.60		2.52	
Kr	Kr_AD2	X	X	1231	37			21.50		12.48	0.88		4.59	2.42
gt	gt_(-6)	X	X	2182	-4438			6.93	15.28	7.88	1.62	3.01	15.02	
gt	gt_(-10)	(X)	X	1745	-8904			13.29	16.20			0.59	8.71	
kni	kni_(-5)	X	(X)	1403	-4140	2.26	2.26	10.06	10.90	5.62		2.35	8.23	
eve	eve_stripe1	X	X	801	7495			2.87	5.23			5.97		
eve	eve_stripe2	X	(X)	663	-866				5.82	2.29	1.73	3.96		
h	h_stripe1	X	X	876	-4303			4.89				14.47		
h	h_stripe2	X	X	1081	-8712	1.28		1.19				10.49	12.55	10.67
h	h_stripe3	(X)	X	476	-11265					6.49	1.74			2.52
run	run_stripe1		X	1617	-12850			1.63	5.32	1.91		5.33	12.97	
ftz	ftz_ps4_actv		(X)	413	-5977	0.96			1.11	3.01	1.23			
slp2	slp2_(-3)	(X)	X	2640	-1556		3.68	16.48	8.67	9.48		12.65		2.25
D	D_(+4)	X	X	1725	4454			9.13	12.47	6.67	4.00	6.21	5.69	6.01
hb	hb_centr_&_post	X	X	1023	-3006			13.34		9.38		5.47	5.69	
Kr	Kr_CD1	X	X	1159	-3174			10.21	11.94	12.17	2.27		7.18	3.93
Kr	Kr_CD2_AD1	X	X	1707	-1193		1.88	4.22	8.03	13.16	3.29	5.54	6.09	
gt	gt_(-1)	X	(X)	1239	-163	2.59		16.32		2.27		8.44	3.32	
eve	eve_stripe3_7	X	X	511	-3229	1.76				5.59			9.60	
run	run_stripe3	X	X	2315	-8970	0.74	0.74	3.41	2.79	21.35		2.65	12.34	2.33
odd	odd_(-5)	X	X	1384	-3873			14.80	1.16	1.22		9.91		2.56
gt	gt_berman	X	X	945	-1815	2.61		8.00		4.79		7.02		7.19
gt	gt_(-3)	X	X	1186	-1410	1.74		6.51	2.20	7.18		5.45		6.11
kni	kni_kd	X	X	877	-1177			11.47	4.49	10.11		6.66	1.87	
cad	cad_(+14)	X		1637	15544			9.38			3.70		27.63	4.40
nub	nub_(-2)	X		1985	-1218		1.43	13.32		2.10		4.76	20.07	12.16
pdm2	pdm2_(+1)	X	(X)	1623	1917			13.80	1.56	12.42		1.16		3.34
eve	eve_stripe5		X	800	8201			4.07		3.12		5.90		
eve	eve_stripe4_6	X	X	602	5260	1.01		3.29		1.09			1.56	5.10
h	h_stripe4	(X)	X	912	-10829					13.57	2.13			3.80
h	h_stripe5	X	X	564	-5460			3.41	2.36	4.31		7.63	4.23	0.93
h	h_stripe6	X	X	538	-8712	1.29						2.69	8.55	10.25
h	h_stripe7	X		1474	-9792	0.95		4.21		15.88		7.03	3.96	
run	run_stripe5		X	1336	-11515			5.78		3.54		3.33		0.53
run	run_stripe7	X	X	5586	-8881	0.62		16.24	8.78	32.03		12.87	29.12	5.76
odd	odd_(-3)	X	X	1650	-2265	2.44		9.50		11.30	0.97	0.56	10.55	4.40
hkb	hkb_ventral_elem			590	-1546		2.52						2.07	1.53
fkh	fkh_(-2)	X	X	1708	-1444		4.78	17.45		4.48	1.87	2.86	6.65	
tll	tll_P2	X	(X)	2760	409		9.23	7.23	13.85				5.98	7.87
tll	tll_K2	X	(X)	2460	-2350		2.32	21.57	2.98	5.21		7.31	16.79	6.95
tll	tll_P3	X	(X)	1376	409		4.30			1.82				7.10
kni	kni_(+1)	X	X	1479	1407	0.73		7.82	4.66	9.91		2.31	7.06	6.40
knrl	knrl_(+8)	X	(X)	1298	8638		2.25	12.96	2.42	4.50			12.08	
nub	nub_(+5)	(X)		2045	6540			9.65		4.64		9.96		3.22
pdm2	pdm2_(+3)	X	(X)	1533	3846		0.77	2.43	2.43	8.77	0.94	8.79	9.28	1.67
pdm2	pdm2_(+5)	(X)	(X)	1987	6092			14.94	1.56	3.59		10.51	14.02	3.60
pdm2	pdm2_(+8)	(X)		973	8465							5.62	7.84	7.33

Figure 10

Relationship between Ahab binding site predictions and pattern. At left the expression pattern driven by the *cis*-element is depicted schematically. In the center is a table describing basic features of the element. The fields =mel and =pse describe whether or not the element was recovered in *Drosophila melanogaster* and *pseudoobscura* respectively. An X corresponds to elements above the threshold for prediction, whereas (X) corresponds to a subthreshold peak. At right is a table of dictionary values computed by running Ahab over the delineated module. The dictionary values are colored by strength of prediction with darker colors for stronger predictions.

As the *cis*-elements delineated in reporter constructs autonomously generate pattern, the positional information they read out is contained within the sequence tested. Therefore Ahab was run over the sequences tested in each case. In order to measure the patterns generated by the elements, the *in situ* stainings for each was measured at a comparable time (Materials and Methods). The *cis*-element expression patterns and binding site predictions used in the analysis are presented in Figure 10. The expression patterns are grouped by region of expression into anterior, anterior and posterior, posterior, and terminal classes. The most clear diagnostic pattern is the input from the maternal activators BCD, CAD, and TOR-RE. The anterior modules show a clear enrichment for BCD sites and depletion of CAD sites although there are exceptions. The anterior and posterior elements show no clear preference of input. The posterior elements show a clear enrichment for CAD and a lower level of BCD input. However, in all three cases there are exceptions that use inputs other than the most intuitive one. The only set of elements that have complete coherence with the natural activator are the terminal modules, which all contain TOR-RE sites. It is even more difficult to see relationships for the gap genes, which are expressed in more complicated patterns.

The point of interest is how the strength of input relates to the pattern of expression of the targets. To attempt to represent this complex relationship in a graphical fashion a simple method for depicting how the strength of input and the expression pattern of the elements was employed. As the expression data we generated is binary on/off patterns along the a-p axis, the data naturally defines a set of elements that is expressed at each position in the embryo. Therefore one

can calculate the average dictionary value predicted by Ahab for each factor over the set of elements expressed at each position in the embryo. A plot of the spatial average of input in comparison to the expression pattern of the regulating factor is shown in Figure 11a.

This analysis helps visualize the correspondence between the input factor distribution and the expression of target elements in a way that is less evident in the raw data. In the case of the maternal gradients, BCD, CAD, and TOR-RE, there is a clear correspondence between the strength of predicted input and the pattern of the input factor consistent with their role as activators (Figure 11a). This is most notably true for TOR-RE, which tracks very well with the inferred pattern of TOR-RE activity (Materials and Methods). The distribution of BCD and CAD input is also well correlated with the expression of target modules for the factors. In both cases the average Ahab dictionary values are inflated at the opposite pole due to the fact that both factors target a number of elements expressed at both termini.

Our analysis shows an anti-correlative correspondence between the HB, GT, and KR protein expression and the averaged Ahab dictionary values (Figure 11a). In the case of KNI and TLL, there is no clear relationship between the prediction of sites and their expression pattern. This is largely attributable to the unspecific nature of these matrices (Figure 11b, compare KR which is specific to KNI and TLL). The data is thereby consistent with these factors acting as repressors, when there is clear spatial signal as to where these factors act.

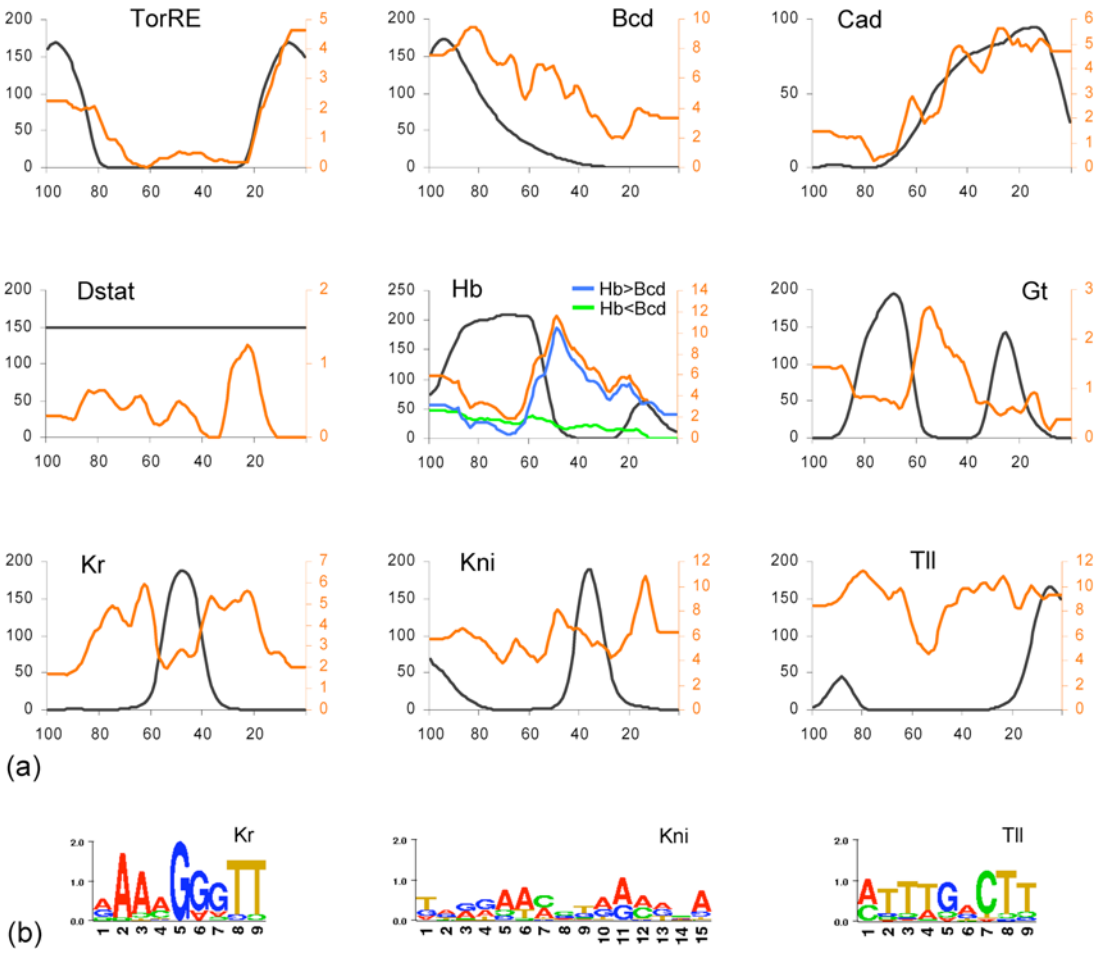


Figure 11

Graphical depiction of input predicted by Ahab. (a) The distribution of input factors (black) compared to the average Ahab dictionary value over all modules expressed at each position along the a-p axis (orange) in % EL (100% is the anterior tip). The protein data is from FlyEx (Myasnikova et al., 2001) except for TOR-RE, which is based on a quantification of the pattern of *capicua* (*cic*), protein expression. The factors BCD, CAD, and TOR-RE, the distributions are positively correlated. In HB, GT, and KR the distributions are negatively correlated. In the case of HB, which combinatorially regulates with BCD, the averages are also shown separately for elements with more HB input than BCD (blue) and elements with less HB input than BCD (green). (b) Sequence logos for the KR, KNI and TLL PWMs. The height of each column is scaled by the information content and the letters are sorted and scaled by frequency.

Initially many gap genes appearing to act as both activators and repressors based on the expression of other gap genes in mutants. In most cases, indirect effects more parsimoniously explain activation, but it is still not entirely clear whether some factors can act as both activators and repressors. The case with the most evidence for dual function is HB, which is thought to act in a combinatorial fashion with BCD (Simpson-Brose et al., 1994; Zuo et al., 1991). In HB, there are a relatively large number of *cis*-elements predicted to have relatively small amounts of input in the region HB is expressed. Given the implied role of BCD in altering the activity of BCD, their co-occurrence was plotted differentially (Figure 11a). This context dependent rule improves the negative correlation for HB acting alone and leaves a positive correlation with the modules where HB and BCD co-occur.

A technical question at the time this work was done was whether binding sites or elements could be better predicted using phylogenetic information. Given the availability at the time of the *Drosophila pseudoobscura* genome, there was an opportunity to test this question. A previous study done in the Siggia lab indicated that although *cis*-elements and binding sites were significantly better conserved than random sequences, there was extensive overlap between the two distributions (Emberly et al., 2003). Therefore any method that filtered results using conservation would lose elements that were functional. Indeed introducing an ascertainment bias towards conserved features of transcriptional regulatory networks would cloud our understanding of how evolution shapes

function. The prediction information shown in Figure 10 demonstrates the simple fact that Ahab predictions in the two species do not recover the same elements.

2.5 Discussion of Ahab *cis*-dissection screen

The efficient prediction of *cis*-elements elements allows a directed effort to systematically determine elements throughout the segmentation network allows, which would otherwise be too laborious a task to take on. The fact that the above threshold Ahab predictions drove appropriate patterns in over 80% of the elements tested shows that it is an effective tool for this task. Increasing the number of segmentation enhancers by 50% significantly moves forward the goal of delineating the *cis*-regulatory components of segmentation network. However, the most appealing aspect of the analysis is providing a unified framework for understanding the programming of the *cis*-elements. Segmentation is often presented as understood, but there is still no quantitative understanding of how the set of transcription factors generate the patterns despite the extensive work done on the system.

The analysis of binding site content presented is a significant step forward in understanding how position is encoded in segmentation *cis*-elements. One difficulty in genetics is separating out direct and indirect effects under mutant conditions. This has led to some confusion over the function of various gap gene factors. The analysis presented in this chapter indicates that HB, KR, and GT

target *cis*-elements with patterns that are anti-correlated to their own expression patterns. This suggests that they act primarily as repressors from a simple analysis that can pinpoint the spatial relationships between the direct connections in the transcriptional network. This interpretation is supported by the positive correlation between the maternal activators BCD, CAD, and TOR-RE, which are known to result in activation within their domain of activity. One limitation of the binding site analysis is that the KNI and TLL PWMs used are unspecific. It is clear that one of the most important factors in such an analysis is having high quality PWMs, as they are the basis of both the prediction of elements and binding sites.

How does Ahab compare with other methodologies? One difficulty in answering such questions is the heterogeneity of approach between different studies using different methodologies. The most direct studies for comparison in the case of Ahab and prediction of segmentation enhancers are those using *cis*-analyst (Berman et al., 2002; Berman et al., 2004). The first study is the more apt comparison as it also was focused on prediction of *cis*-elements using a similar set of PWMs (BCD, CAD, KR, HB, and KNI) without introducing conservation as a filter on predictions. The method employed was based on prediction of binding sites using a threshold and then predicting regions containing more than a certain number of binding sites in a given window.

At their lowest stringency, which required 12 binding sites, they predicted 12/20 enhancers that were analyzed by both studies, whereas Ahab recovered 16/20. It is also clear from looking at the predictions that Ahab was more precise

in delineating the boundaries of the known elements than *cis*-analyst. When validating their results with transgenic constructs, their success rate was 4/27 (15%) compared to 13/16 (80%) for Ahab at the defined threshold and 15/21 (71%), when including the subthreshold tests. However, the selection of elements was done differently and therefore the results are not completely comparable. Furthermore, the work with *cis*-analyst simply counted numbers of sites above a threshold and did not attempt to systematically study the relationship between composition and expression pattern. Although a true straightforward comparison of different prediction methodologies has not been carried out, it is clear that Ahab is a superior method. The probabilistic framework for integrating site number and strength into a single measure is much more sensitive than approaches involving thresholding.

Although there is more to say on many of the topics brought up in this section, they will be held off until the discussion of the third and fourth chapters. In the third chapter, the interesting result of *odd* containing stripe specific elements will be followed up with a revisitation of the pair rule hierarchy. In principle this effort will be focused on determining the complete set of stripe specific elements within the pair rule genes. In addition to understanding the details of how pattern is encoded within a given *cis*-element, there is the equally important question of how pattern is established within a larger transcriptional network. The pair rule genes establish the repeated patterns central to later development from the non-periodic maternal and gap inputs and are therefore a well defined genetic network encoding pattern at a network level. Given the success of Ahab at determining elements and the fact that this open issue could

be addressed with a more thorough *cis*-dissection of these genes, this seemed like a perfect biological question to address using the Ahab methodology.

Chapter 3: Revisiting the pair rule hierarchy

There are three basic criteria traditionally used to classify the pair rule genes: cuticle phenotypes, molecular epistasis, and *cis*-element composition. Based on the original cuticle phenotypes seen in the segmentation screen, the pair rule class was established based on loss of portions of the cuticle pattern with a two segment repeat. The main criterion used in formulating the current primary pair rule classification was phenotype in molecular epistasis experiments, which examine whether mutants in one gene effect the establishment of the pattern of another gene. Mutants of the primary pair rule genes were shown to cause defects in the initial patterns of all pair rule genes. In contrast the secondary pair rule genes have been described as not causing defects in the early patterns of the primary pair rule genes. After the initial classification, the presence of stripe specific elements was added as an additional criterion, which indicated the primary pair rule genes established the periodic patterns directly from nonperiodic patterns. In contrast, the prototypic secondary pair rule gene *ftz*, was found to only have a seven-stripe element, which supported a role limited to transmitting, but not establishing, periodicity.

The classifications as they stand were not immediate. The first indication that pair rule patterns were generated in a piecemeal fashion were the region specific alleles of *h*, in which discrete loss of stripes occurred as rearrangements removed increasing portions of the *h* upstream region (Howard et al., 1988). This suggested that individual stripes were generated by an inherently nonperiodic mechanism through a modular transcriptional control region. Soon after *h* and

run were proposed to be the primary pair rule genes that initiated the periodic patterns (Ingham, 1988). The criteria were that their patterns resolved "slightly earlier" than the other pair rule genes and they "appear to have a major function in establishing the striped patterns of other members of the class" (Ingham, 1988). Why *eve* was not included at that time is not clear as *eve* mutants were known to cause defects in the early patterns of *h* and *run* (Ingham and Gergen, 1988).

Later, *eve* was included when it was shown to have stripe specific elements generating stripes 2, 3, and 7, which were directly bound by gap genes and depended on gap function for proper expression (Goto et al., 1989; Harding et al., 1989; Stanojevic et al., 1989). *h* was then shown to have a full set of stripe specific elements, which were bound and regulated by gap genes (Howard and Struhl, 1990; Pankratz et al., 1990; Riddihough and Ish-Horowicz, 1991). The extensive work on the *eve* stripe 2 element demonstrated that the gap genes acted primarily as repressors and there was a quantitative interplay between activation through binding sites for the maternal factors and repression through gap binding sites to properly position the stripe borders (Arnosti et al., 1996; Small et al., 1992; Small et al., 1991; Stanojevic et al., 1991). Later *eve* was shown to have a complete set of stripe specific elements (Fujioka et al., 1999). These results strongly supported the inclusion of *eve* and established a model where the primary pair rule genes generate the periodic patterns in a piecemeal fashion from the nonperiodic maternal and gap patterns.

However, there are a number of issues with this model that have not been resolved. Stripe specific elements in *run* were only found for stripes 1, 3, and 5, while the region containing all three elements also generates a weak stripe 7 (Klingler et al., 1996). Therefore it was unclear if primary pair rule genes all had a complete repertoire of stripe specific elements. Additionally, early patterns had been seen in *ftz*, *odd*, *slp*, and *prd*, which had been shown to be dependent on gap gene patterns in some cases (Carroll and Scott, 1986; Coulter et al., 1990; Grossniklaus et al., 1992; Gutjahr et al., 1993). Therefore, the secondary pair rule genes did directly read off the nonperiodic pre-pattern indicating that their patterns are not solely established in an inherently periodic fashion. As the establishment of the periodic patterns from the nonperiodic patterns of the maternal and gap genes is the central function of the primary pair rule genes, it is critical to understand what types of input pattern the different pair rule genes as the periodic pattern is established.

The most complex story in the literature is *ftz*, which is often put forward as the prototypic example of a secondary pair rule gene (Klingler et al., 1996; Pankratz and Jackle, 1990). However, it has been shown that the early *ftz* pattern is not a simple consequence of regulation by primary pair rule genes (Yu and Pick, 1995). Further, there is evidence for *ftz* regulation by stripe specific elements. In the original *ftz* dissections an element was found that drove expression in the stripe 2 region, but only in the reverse orientation (Hiromi and Gehring, 1987; Pick et al., 1990). Therefore this element was thought to relate to the adjacent *Antp* gene that is expressed in parasegment 4, which corresponds to the *ftz* stripe 2 expression domain. A subsequent study found an element just

downstream of *ftz*, which generates expression in the region of stripes 1 and 5, but was attributed to being a relic of *ftz* function as a homeotic gene (Calhoun and Levine, 2003). As *ftz* regulation solely by a seven-stripe element was taken as support for the existing classifications, the manner of *ftz* is an important factor in clarifying the consistency of the current hierarchy.

What should the criteria be for primary versus secondary pair rule status? The inclusion of *eve* primarily based on *cis*-elements indicates that *cis*-regulation is a central criterion, but the fact that *eve* mutants effect the early *h* and *run* patterns leaves open how to balance conflicting criteria. The consistent inclusion of *run* indicates that having a complete repertoire of stripe specific elements is not explicitly required. The central role of the primary pair rule class is establishment of the periodic patterns from non-periodic patterns. Therefore, the timing and nature of *cis*-regulation are both central criteria that provide a concrete basis for classification. In contrast, genetics can miss interactions due to redundancy of function or lack of true null alleles, which make absence defects in loss of function circumstances somewhat ambiguous. In practice a systematic comparison of the timing, *cis*-regulation, and molecular epistasis is necessary before making decisions on *a priori* arguments.

3.1 *Cis*-dissection of the pair rule hierarchy

Given the *odd cis*-dissection, it was of interest to systematically determine the stripe specific input into the secondary pair rule genes. The first obvious question is whether *odd* had a complete compliment of stripe specific elements like *h* and *eve* or whether the maternal and gap input was of a more limited

nature. Additional predictions in the *odd* locus were not as strong based on free energy score and composition compared to the two elements found in the initial segmentation *cis*-element screen (Schroeder et al., 2004). Three additional constructs were tested and none drove stripes (Figure 12). However, the *odd* stripe specific elements generate stripes 1, 3, 5, and 6, which initiate first (Figure 12). This temporal difference suggests a causal relationship between the stripe specific elements consistent with the idea of the maternal and gap input drive the early striped patterns of the pair rule genes. Therefore the time when the *odd* pattern is limited to the four stripes generated by the stripe specific elements could help define when stripes are generated only by the stripe specific elements. Assuming that the different classes of *cis*-element are activated with similar timing between genes, the patterns in the different pair rule genes at this time point could be used as a heuristic to define the set of stripes presumably generated by stripe specific elements. The elements corresponding to these stripes could then be searched for in a directed fashion using Stubb.

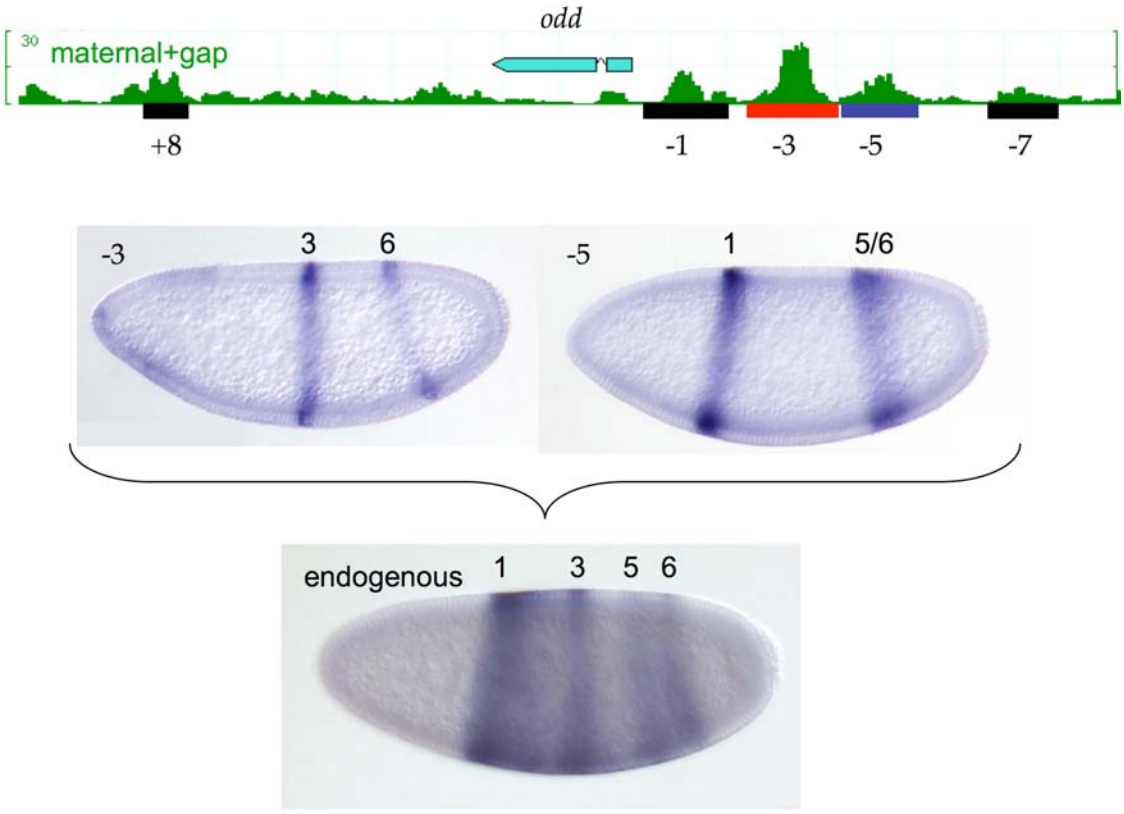


Figure 12

Additional constructs tested in the *odd* locus. Free energy profile from Stubb for the maternal and gap PWMs shown in green for the odd locus. Additional tested constructs are shown in black, but none drove stripes. However, the -3 element (red) drives stripes 3 and 6, and the -5 element (blue) dives stripes 1 and 5, which are the four earliest stripes that arise in the endogenous pattern. This suggests that the full compliment of *odd* stripe specific elements consists of two elements driving four of the seven stripes. As in other figures, the embryo stainings are *in situ* hybridizations with a LacZ probe for the reporter constructs and an *odd* probe for endogenous. Embryos are oriented with the anterior to left and dorsal up.

3.2 Temporal analysis of pair rule patterns

Given that early expression of stripes in *odd* corresponded well with the presence of stripe specific element expression, the timing of stripe formation was analyzed for all pair rule genes by *in situ* hybridization in carefully staged embryos (Figure 13). Cellularization, Stage 5 of *Drosophila melanogaster* development, begins at roughly two hours and ten minutes After Egg Laying (AEL) and lasts forty minutes (Campos-Ortega and Hartenstein, 1997). It is during this time when the seven-stripe patterns of the pair rule genes are established (Nasiadka et al., 2002). Cellularization can be divided into four morphologically distinct phases (Lecuit and Wieschaus, 2000), which provides a useful method for precise staging of embryos. Phase 1 is roughly 5 minutes long, whereas the other phases are all just over ten minutes at 25 C. Phase 1 is recognizable by the spherical shape of the nuclei, phase 2 by elongation of the nuclei, phase 3 by progression of the plasma membrane along the nuclei, and phase 4 by extension of the plasma membrane past the nuclei to roughly 35 um (Lecuit and Wieschaus, 2000). The end of each phase was chosen for documentation as they have better defined nuclear and membrane morphologies than the intermediate time points. It is difficult to precisely stage embryos in early phase 1 and before, so these time points are not systematically addressed here. Broader patterns arise earlier, but they are weaker than the striped patterns arising at this time.

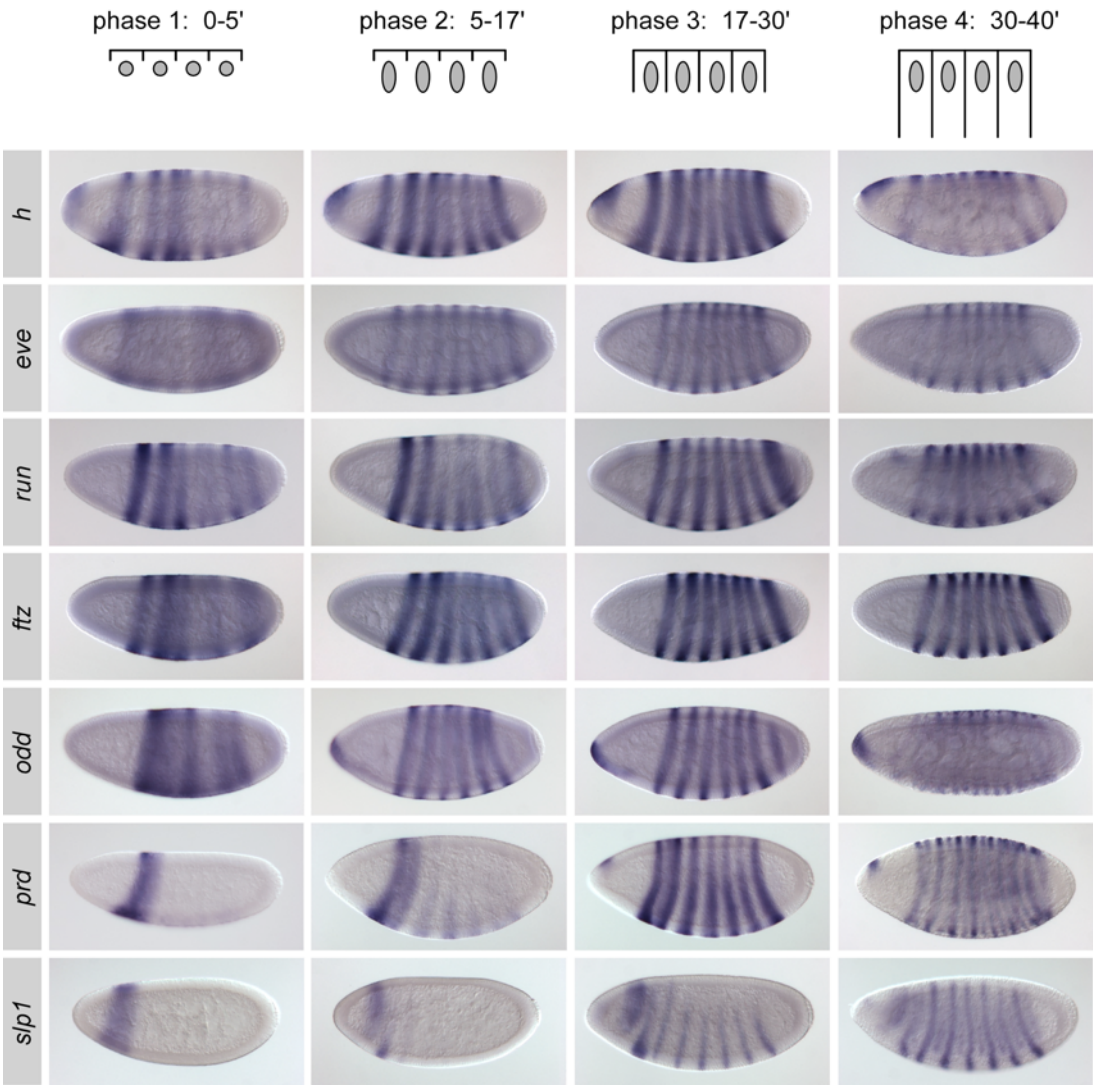


Figure 13

Time course of pair rule expression patterns. At top a schematic of the four phases of cellularization (Lecuit and Wieschaus, 2000). In phase 1 the nuclei are round, in phase 2 the nuclei extend, in phase 3 the plasma membrane invaginates along the nucleus, and in phase 4 the membrane extends to roughly 35 μm. Below, *in situ* hybridizations for the patterned pair rule genes. Embryos are oriented with the anterior to the left and the dorsal side up. Each row contains the patterns for a given gene and each column contains the patterns for a given time point. *ftz* and *odd* have similar kinetics to the primary pair rule genes *h, eve,* and *run* whereas, *prd* and *slp*, have similar kinetics.

In *h, eve, run, ftz,* and *odd* the spatio-temporal dynamics of stripe formation are quite similar, with irregular striped patterns apparent by the end of phase 1 and the regular mature pattern arising by the end of phase 3 (Figure 13). The seventh stripe of *odd*, which is the most posterior stripe within the set, shows a delay only arising at the end of phase 3. The genes *slp1* and *prd* on the other hand are expressed only in a single head domain during phase 1 with the seven-stripes arising relatively synchronously during phase 3 (Figure 13). Therefore, based on the timing of pattern formation, the grouping of *h, eve, run, ftz,* and *odd* in an early class and *prd* and *slp* into a late class seems more natural than the original primary and secondary pair rule gene classes. As the *ftz* and *odd* patterns arise with those of the primaries, they play a role in establishing the periodic patterns regardless of differences in their role in pair rule cross regulation.

In *h* and *eve*, which both contain a complete repertoire of stripe specific elements, expression corresponding to all stripes is present in phase 1 (Figure 13). In phase 2 their patterns sharpen considerably including the splitting of *h* stripes 3 and 4, which is known to require pair rule input (Hartmann et al., 1994). In phase 1 the *odd* expression corresponds solely to the 4 stripes generated by the stripe specific elements, while in phase 2 additional stripes are seen. This suggests that phase 1 best represents the time when stripe specific elements act alone in this locus. In *ftz* and *run* domains of expression corresponding to known stripe specific elements is present by the end of phase 1, but additional expression is present as well. In phase 2 all seven stripes are present for both genes, but the full pattern is not completely resolved. The patterns of *prd* and *slp*

at phase 1 also reveal the regions known to be generated by maternal and gap input. In both cases the expression corresponds to broad domains overlapping the stripe 1 region that split during phase 2 when pair rule based refinement is seen in other pair rule genes as well. Therefore the phase 1 pattern seems to define that generated primarily by the maternal and gap input rather than the pair rule input.

The fact that additional domains of expression exist for *run* and *ftz* suggest they define patterns, which are driven by maternal and gap input. In *run* early expression exists for all seven-stripes, although it is weaker and less well resolved in the region corresponding to stripes 4 and 5. This suggests stripe specific elements exist for stripes 2, 4, 6, and 7 as well as those already known for stripes 1, 3, and 5. In *ftz* expression is present at phase 1 in regions corresponding to all stripes except stripe 4 indicating stripes 2, 3 and a broad 6/7 domain are driven by stripe specific elements. Together this suggests a number of stripe specific elements are missing from the pre-existing *cis*-dissections.

There is something of a continuum of patterns at phase 1 with *h, eve,* and *run* being fully striped, *ftz* having all but one, *odd* having more than half, and *prd* and *slp* only having broad anterior domains in the stripe 1 region. It is notable that in phase 2, the patterns of *ftz* and *odd* continue to be more fully striped than *prd* and *slp* further supporting a greater role in the establishment of the early periodic pattern. However, the full regular seven striped pattern of all pair rule genes occurs in phase 3. Although *slp* and *prd* generate most of their seven stripes at phase 3 in a synchronous manner, they are extensively patterned at the time

when the regular seven striped patterns are also evident in the primary pair rule genes. Therefore, there is no clear separation between when the different pair rule genes establish their regular striped patterns. However, during phase 2 there is specification of almost all stripes in *h, eve, run, ftz,* and *odd*, indicating they have an early role in coarsely defining stripe positions.

This analysis indicates that the grouping of the pair rule genes based on their expression characteristics is not consonant with the original grouping based on molecular epistasis. Further it suggests that there are missing stripe specific elements for *ftz* and *run*. The main function ascribed to the primary pair rule genes is the establishment of periodic patterns, which is initiated by the stripe specific elements. However, the similar timing of pattern formation in *odd*, which lacks a full complement of stripe specific elements, emphasizes that there is a role to be played by seven-stripe elements during the initiation of the regular seven-striped pattern. Therefore a better characterization of what elements exist and their timing of expression are necessary to clarify what role these elements play in establishing the periodic pair rule patterns.

3.3 *Cis*-regulation of the *ftz* locus

The phase 1 *ftz* pattern shows expression in stripes 1, 2, 3, 5, and a broad 6-7 domain. The previously determined stripe 1+5 element explains stripes 1 and 5, while the ps4 element suggests that there is also a stripe 2 element in the associated region. However this still leaves elements undiscovered for stripe 3 and the broad 6-7 domain. The free energy profile of the *ftz* locus has 3 main peaks, two of which overlap the previous elements (Figure 14). The *ftz -6*

element overlaps the ps4 element and drives expression of not only stripe 2, but stripe 7 as well. As the construct functions in the correct genomic orientation, the unusual orientation dependence of the ps4 element was an artifact of the delineation. The *ftz -7* element lies entirely outside the limits of the original dissection and generates strong expression of stripes 3 and 6/7. Finally, the ftz +3 element drives expression in stripes 1+5 consistent with the previously described element. Therefore as the timing analysis indicated, *ftz* has extensive stripe specific input driving 6 of the 7 stripes.

One somewhat unusual feature of the -6 and -7 elements is the shared generation of stripe 7. The *ftz* -7 element initially generates a broad domain that is expressed throughout the stripe 6 and 7 region similar to that seen in the early endogenous pattern. Later during cellularization the broad domain splits into stripes 6 and 7 before stripe 7 fades. The *ftz* -6 element in contrast initially only generates stripe 2 with stripe 7 arising later. This indicates that the spatial and temporal dynamics of the endogenous stripe 7 are generated in combination between the two elements. The splitting of the broad 6-7 domain into two stripes is strongly suggestive of pair rule input. It is notable that although the -6 and -7 elements are separable, an earlier construct containing both elements drove much stronger expression of these stripes than either element alone (not shown). Together these results suggest that although the elements can act autonomously to generate the individual stripes they may act in concert in the endogenous locus.

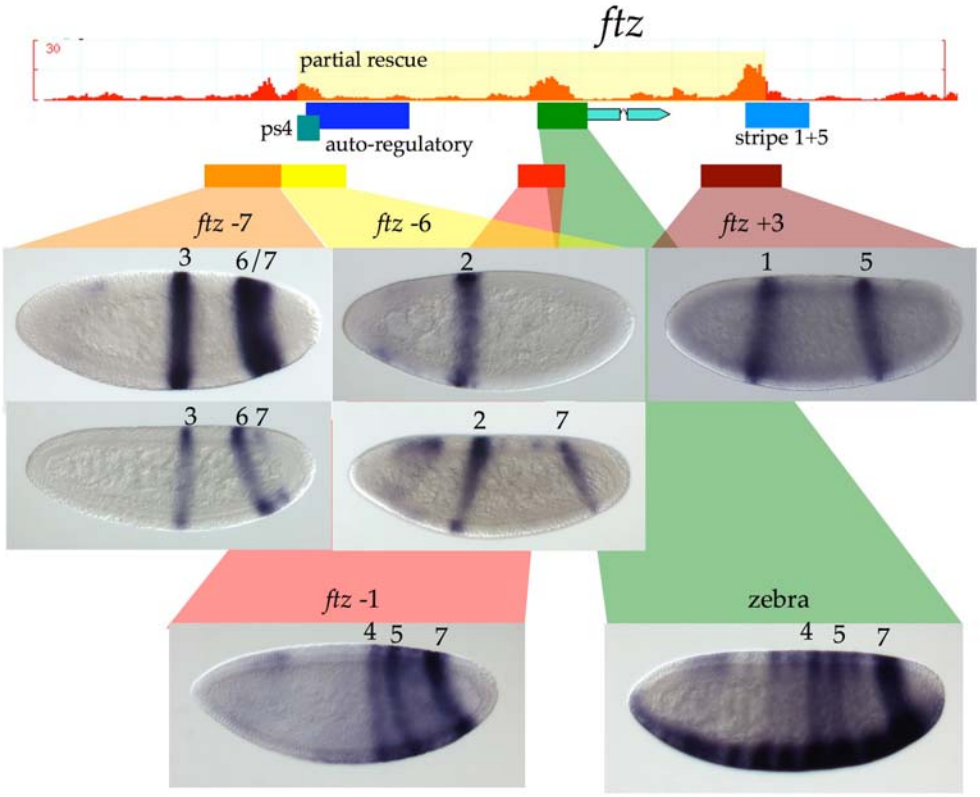


Figure 14

ftz dissection. At top schematic of the *ftz* locus with the free energy profile of a maternal and gap run. Just below the free energy profile are colored rectangles representing the known elements (Calhoun and Levine, 2003; Hiromi and Gehring, 1987; Hiromi et al., 1985). The next row of rectangles represents the Stubb based *cis*-dissection. *in situ* hybridizations to LacZ for the constructs are matched to their constructs by color coded callouts. Embryos are oriented with anterior to the left and dorsal side up. The *ftz* -7 construct generates stripe 3 plus a broad 6/7 early, which then splits before stripe 7 fades. The -6 construct drives stripes 2 early and then 7 later when it is fading from the -7 construct. The +3 construct generates stripes 1 and 5. The *ftz -1* element drives expression in a broad modulated domain with peak expression in the regions corresponding to *ftz* stripes 4, 5, and 7. The zebra element generates a pattern with very strong ventral expression and weaker expression along the rest of the d-v axis. The most strongly expressed stripes are 4, 5, and 7 similar to the -1 element.

The 6 stripes driven by the -7, -6, and +3 elements account for the early endogenous *ftz* expression, but stripe 4, which arises later in phase 2, is still unaccounted for. Although *ftz* stripe 4 arises later than expression driven by most stripe specific elements, there is no *a priori* reason that the stripe is not generated by a later expressed stripe specific element. The only other free energy peak within the region overlaps zebra element, which was originally shown to drive a late 7-stripe pattern at germband extension (Hiromi et al., 1985). This element is known to contain CAD binding sites (Dearolf et al., 1989) and is predicted to have both CAD and weak GT input. Consistent with the striped pattern and previous analysis of this region HAIRY input is predicted within the region as well.

The expression of the zebra element during cellularization is dominated by very strong ventral staining in a modulated striped pattern (Figure 14). When tracked over time the striped pattern corresponds to the later mesodermal striped pattern that has been emphasized in previous work (Pick et al., 1990). There is also a modulated striped pattern that is biased towards the posterior, consistent with activation by CAD from the posterior. Stripes 4, 5, and 7 are the most strongly expressed, consistent with this element having a role in generating the endogenous stripe 4. As the element has a mixed nature suggestive of both seven-stripe and stripe specific qualities an attempt was made to separate out the CAD and GT input from the HAIRY input.

The *ftz* -1 element, similar to the zebra element, generates a broad domain with a modulated pattern strongest in the position of stripes 4, 5, and 7 (Figure

14). That the element is not limited to stripe 4 and still shows pair rule modulation despite an attempt to separate the two types of input suggests that the element requires both inputs for proper early expression. Unlike the zebra element, the ventral staining as well as staining outside the stripe 4-7 region is substantially reduced. The unnatural expression of both the -1 and zebra elements suggest they lack important inputs and are portions of a larger *cis*-element. The different insertions of the *ftz* -1 element all preserve the main features described, but there is greater variation between different transgenic lines than seen in most constructs. The insertional dependence also suggests that important inputs may be missing as suggested in the unfaithful elements seen in the original segmentation *cis*-element screen (Chapter 2).

Based on the revisited *cis*-dissection of *ftz,* stripe specific elements clearly exist that generate 6 of the 7 stripes. Therefore *ftz* extensively interprets the nonperiodic patterns of the maternal and gap genes to generate all but one of its stripes through stripe specific elements in marked contrast to that suggested in the original *cis*-dissections. That the zebra element drives the early stripe 4 and receives maternal input does indicate an important early function for this element. As stripe 4 arises while the early pair rule patterns are still resolving, the seven-stripe element of *ftz* is involved in the establishment of the striped pattern. Therefore maternal and gap input into both stripe specific and seven-stripe elements drives the early seven-striped pattern of *ftz*. Although this input differs slightly from that of *h* and *eve* in extent, the *ftz* pattern clearly generates the vast majority of its periodic pattern from nonperiodic inputs in an inherently primary pair rule fashion.

3.4 *Cis*-regulation of the *run* locus

run stripes 1, 2, 3, 6, and 7 are strong and defined the phase 1, whereas stripes 4 and 5 are weaker and not completely resolved. In the original dissection stripe specific elements were found for stripes 1, 3, and 5, with the combined region generating a weak stripe 7 (Klingler et al., 1996). In addition a large 5 kb seven-stripe region was found that contains multiple portions that can each generate seven-stripes (Klingler et al., 1996). Based on the time course of pair rule patterns (Figure 13) additional stripe specific elements are likely to exist for 2, 4, 6, and possibly 7.

In the *run* upstream region there are 4 strong free energy peaks that fall within the original dissection (Figure 15). The *run* -17 element drives stripe 4 indicating that there is an element driving this stripe even though the early stripe 4 expression is weak and poorly resolved. The *run* -9 element corresponds to a tighter delineation of stripe three, which is well resolved and extremely strongly expressed. The third peak falls unexpectedly within the large seven-stripe region and contains strong prediction of both maternal and gap sites. Consistent with the location of the prediction, the *run* -3 element is able to generate seven-stripes (Figure 15). However, the various stripes at phase 2 are expressed at different intensities consistent with the maternal and gap input modulating the early expression. Finally the *run* +3 element at the 3' end of the gene did not drive expression in stripes, but rather weak head staining that does not correspond to the endogenous *run* pattern. Therefore the four strongest predictions in the

region do not include stripe specific elements for stripes 2 and 6 or a separable element for stripe 7.

The -3 element generates stripes in phase 2 including relatively strong stripes 2 and 6. However, the timing and strength is not consistent with the very strong expression of stripes 2 and 6 in the endogenous *run* pattern in phase 1. Therefore additional regions were tested for stripe specific expression. The large *run* downstream region had a number of free energy peaks with simple predicted composition, typically dominated by the poor quality KNI and TLL PWMs. As all known stripe 2 elements had strong BCD and KR input, which are good specific PWMs, the tested predictions were biased towards those containing this input. However the *run* -16, -6, and +6 elements do not drive additional stripes. The *run* -16 element, which overlaps the *run* stripe 1 element by 764 bp, also drives stripe 1. When improved KNI, TLL, and GT matrices were generated (Chapter 4, Materials and Methods) the downstream peaks were re-examined. The *run* +30 element corresponded to a broad peak with diverse inputs appropriate for generating stripe 2 and was tested. This element generates stripes 2 and 7, indicating that these stripes are indeed generated by stripe specific elements consistent with the timing analysis. Although no element was found for stripe 6, the strong early expression of this stripe makes it quite likely such an element exists. As KNI is a likely input in patterning stripe 6, the downstream elements containing this input contain some reasonable candidates.

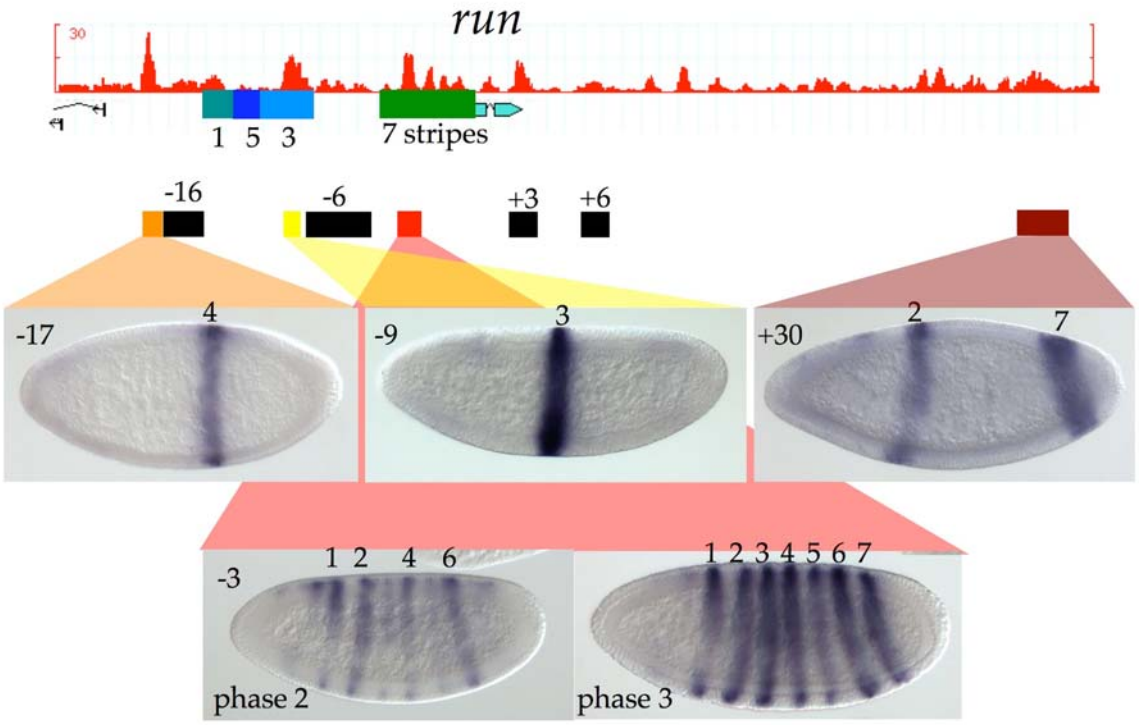
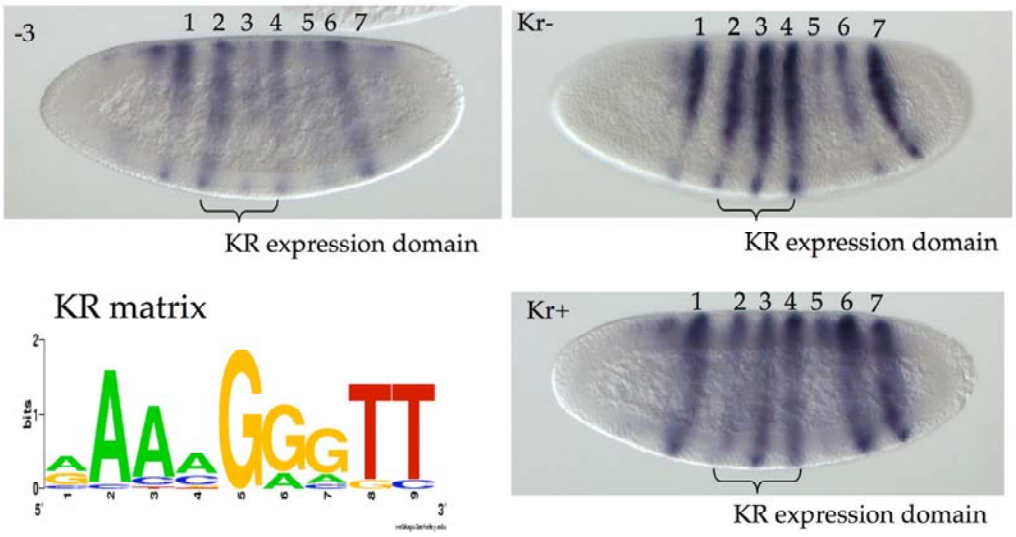


Figure 15

run dissection. At top schematic of the *run* locus with the free energy profile of a maternal and gap run shown in red. The transcription start of the next locus over is present on the left. Just below the free energy profile are colored rectangles representing the known elements (Klingler et al., 1996). The next row of rectangles represents the Stubb based *cis*-dissection. *in situ* hybridizations to LacZ for the constructs are matched to their constructs by color coded callouts. Black elements did not drive new domains of expression. Embryos are oriented with anterior to the left and dorsal side up. The *run* -17 construct generates stripe 4, the -9 construct is a tighter delineation of stripe 3, and the +30 construct generates stripes 2 and 7. The *run*-3 element drives a modulated 7 stripe pattern in phase 2 with stripes 3, 5, and 7 weak, which strengthens into a full seven-stripe pattern during phase 3 and 4.

Together with the original dissection (Klingler et al., 1996) the elements discovered here support the notion that *run* has a complete stripe specific element repertoire. From the phase 1 expression, which is weak and unresolved in the stripe 4 and 5 region, these are the two stripes least likely to have elements. However, the *run* -17 element demonstrates that stripe 4 has a stripe specific and a stripe 5 element was delineated in the original dissection. Although the combination of the stripe 1, 3, and 5 elements generate a weak stripe 7, the *run* +30 element demonstrates that stripe 7 is generated by a more traditional stripe specific element as well. The strong stripe 2 seen early is similarly supported by the stripe 2 generated by the +30 element, lending further support that strong early expression is indicative of the existence of stripe specific elements. Therefore stripe 6 is likely to also be generated by a stripe specific element. The seven-stripe element is quite large and expressed throughout the time when the stripe specific elements are active indicating the two types of input must somehow be integrated within the locus. That the early pattern driven by the seven-stripe element appears to be modulated by gap input suggests that the relative timing of stripes from the two types of elements may be coordinated by this input.



	Kr 1	Kr 2	Kr 3	Kr 4
normal	GAACGGAGT	CAAAGGATT	AAAAGGGTT	AAAAGGATC
Kr-	GAACtGAGT	CAAAtGATT	AAAAtGGTT	AAAAtGATC
Kr+	GAAaGGgtT	aAAAGGgTT	AAAAGGGTT	AAAAGGATt

Figure 16

run-3 Kr mutagenesis. *in situ* hybridizations are shown for the -3, Kr-, and Kr+ constructs. A sequence logo representing the PWM for KR binding site preferences and the specific sequences generated in the element are also shown. The Kr- differs from the -3 construct by only 4 single base substitutions in the four strongest predicted KR sites. A central G that exists in all footprinted KR sites was changed to a T in each of the four sites. In the Kr+ construct, the KR 1, 2 and 4 sites were changed to consensus, while the KR 3 site was already a consensus site. The KR 1 site was left with a G in position one as an A is only marginally preferred in that position.

KR is thought to act primarily as a repressor, which is consistent with the relative strengthening of stripes 2, 3, and 4, in the Kr- construct. In the Kr+ construct stripes 2, 3, and 4 are expressed at strong levels and only stripe 5 is weak. This suggests that the role of KR in this element may more complex than simple repression.

The large and complex nature of the *run* seven stripe region is a unique and defining feature of this locus. The *run* -3 portion of the seven-stripe element contains very strong maternal and gap input making it distinct from that of other pair rule genes. The full seven-stripe element is expressed very early in a broad domain (Klingler et al., 1996). In phase 2 the expression driven by the *run* -3 element is a modulated pattern with weaker expression in stripes 3, 5, and 7 (Figure 16). As stripe 3 falls in the center of the KR domain and KR sites are predicted in the element, it was possible that KR repression within this region was playing a causative role in the modulation of the early pattern. Although the simplest experiment to do would be to cross the reporter construct into a *Kr* mutant, the indirect effects through the pair rule genes would make any interpretation of the results difficult. KR is a convenient choice for site directed mutagenesis as there is a single central G nucleotide conserved in all known KR sites.

There are four KR sites predicted in the element with reasonable confidence and all were mutated in the context of a single construct (Figure 16). As expected, the construct where the sites were abolished drives stronger expression in stripe three at phase 2 as well as stronger expression in stripes 2 and 4, which also fall within the KR domain. To examine in more detail how the KR input modulates the expression driven from this construct, the weaker sites in the element were modified to match the consensus. In the KR+ construct with the strengthened sites, there is a strengthening of stripe 3 relative to stripe 2,

rather than a weakening of stripe 3 (Figure 16). This result is not consistent with a strict relationship between KR binding site strength and repression of stripe 3.

Regardless, the modulated seven stripe pattern driven by the -3 element together with the strong predicted maternal and gap input indicates unexpected inputs and regulation of a seven-stripe element. That the modulated pattern can be varied by a few point mutations in the KR sites supports a role for gap gene input. Gap gene input could help coordinate the onset of the seven stripe regulation with that of the stripe specific elements. It is notable that *run* stripe 3 plays an important role in splitting the expression of the *h* stripe 3+4 element into two stripes. Therefore the timing of stripes from expressed by the *run* seven-stripe element may balance a need for early integration and initiation of pattern by the stripe specific elements.

3.5 Seven-stripe element dissections

Although stripe specific elements have been studied more extensively, seven-stripe elements have been identified in the regulatory regions of *eve, run, ftz, prd,* and *slp* (Goto et al., 1989; Gutjahr et al., 1994; Hiromi et al., 1985; Klingler et al., 1996; Lee and Frasch, 2000), which span all tiers within the hierarchy. *h* and *odd* are the only patterned pair rule genes in which a seven-stripe element has not been described. In the case of *h,* neither the original *h* region specific alleles (Howard et al., 1988) or *cis*-dissections (Howard and Struhl, 1990; Pankratz et al., 1990; Riddihough and Ish-Horowicz, 1991) suggest a role for an

autonomous seven-stripe element. In contrast, in the case of *odd,* the lack of a complete set of stripe specific elements suggests the existence of such an element.

There is predicted pair rule input into *h* from upstream of the stripe specific elements all the way through the basal promoter (Appendix). Since we computationally detect pair rule input proximal to the *h* basal promoter and 15 kb upstream (Appendix), which had not been tested, we examined the basal element alone and the two in combination. However, neither generates a seven-stripe pattern (not shown). Consistent with this lack of a dedicated 7-stripe element, *h* has the most transient expression of all the pair rule genes, fading as cellularization completes (Figure 13). This supports the previous work, which suggested that *h* lacks a seven-stripe element (Howard et al., 1988; Howard and Struhl, 1990; Pankratz et al., 1990; Riddihough and Ish-Horowicz, 1991).

The lack of a complete compliment of stripe specific elements implies that *odd* is likely to have a seven-stripe element capable of generating the remaining stripes. As much less work has been done on pair rule input into seven-stripe elements there is less prior information guiding the search for such an element at the level of binding site composition. The two main regulators described for the *odd* seven-stripe pattern are *ftz* and *eve* (Manoukian and Krause, 1992; Nasiadka and Krause, 1999). Both factors are homeodomain containing transcription factors that recognized similar sequences (Liang and Biggin, 1998; Walter and Biggin, 1996; Walter et al., 1994). *ftz* has been described as an activator of *odd* and has a very similar expression pattern. In contrast, *eve* has been described as a repressing *odd* and is they are expressed in near reciprocal patterns. Therefore a

combination of *ftz* and *eve* input seems to be a straightforward recipe for generating the *odd* seven-stripe pattern.

As there is strong *eve* and *ftz* input in the 1 kb proximal to the basal promoter, this region was a strong candidate for such an element. However, this region had sufficient predicted gap input that it was tested in the search for stripe specific elements, but did not drive stripes (Figure 12). As this construct contained a heterologous hs43 basal promoter, an in frame fusion of *odd* including the basal promoter, intron, and first kb was tested, but the construct did not drive stripes (Figure 17). Previous work on *ftz* targets indicated that adjacent FTZ and FTZ-F1 sites were important for activation. The stripe specific elements contain predicted FTZ-F1 sites, which although not clustered with the FTZ input, suggested there might be an interaction between the input proximal to the basal promoter and the stripe specific elements. An extension of the in frame fusion including the first 5 kb of *odd* did not generate seven-stripes. The reporter construct did generate a weak stripe two, which was not seen in either of the individual stripe specific elements. In addition this construct drives ventral expression in the head similar to that seen in the endogenous *odd* pattern. Much weaker ventral head staining was also seen in the 3+6 element (not shown). Despite the fact that the elements drive certain expression domains not seen when included independently, the complete *odd* pattern was not generated.

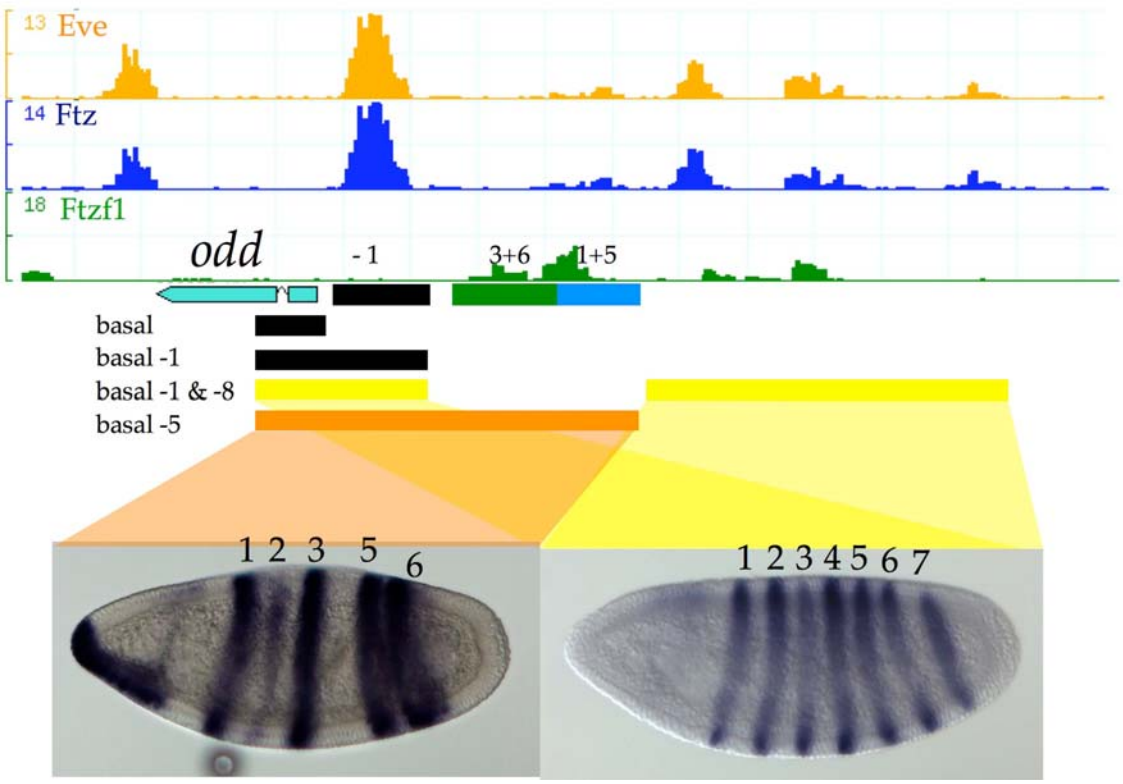


Figure 17

odd seven-stripe element delineation. The Stubb free energy profiles shown are for runs of individual PWMs. The EVE and FTZ PWMs are from bacterial one hybrid screens (Noyes et al., 2008a). Previously used EVE and FTZ PWMs based on footprinted sites predict similarly strong input into the -1 element, but do not predict substantial input into the -8 region. Constructs are schematized by rectangles, with the elements tested as stripe specific elements shown in line with the schematization of the *odd* gene. The black rectangles depict constructs that do not generate staining that generates aspects of the *odd* pattern at the blastoderm stage. The odd basal -5 element drives a weak stripe 2 and a strong ventral head domain, which are not seen in any of the sub elements. The odd basal -1 & -8 element drives all seven-stripes. The *odd* -1, basal, and basal -1 elements do not drive the seven-stripe pattern. One line of the *odd* -8 element was generated and it drove a much weaker seven stripe pattern compared to the basal -1 & 8 construct.

At this point a bacterial one hybrid paper was published which contained new FTZ and EVE matrices (Noyes et al., 2008a). Using these matrices two new clusters of FTZ sites were predicted upstream of the stripe specific elements that overlapped clusters of FTZ-F1 sites, the arrangement previously shown to generate activation by FTZ (Florence et al., 1997; Yu et al., 1997). The -8 kb element alone drives a weak and uneven seven-stripe pattern (not shown), but in combination with the *odd* basal -1 element it drives a strong seven-stripe pattern (Figure 17). As this construct does not contain the stripe specific elements, it demonstrates that *odd* has a separable seven-stripe region.

3.6 Timing of *cis*-element expression

Given a dissection of most of the *cis*-regulatory components that drive the initial seven stripe pair rule patterns, it is possible to take a reductionist approach to how the pattern is established. The mature seven-stripe pattern of the pair rule genes is generated during phase 3. Therefore the *cis*-elements that generate pattern prior to this time are all involved in establishing the periodic pattern. There are two important and distinct roles in the establishment of the periodic pattern. The first is initiation of stripe expression, whereas the second is the refinement and positioning of the stripes.

The stripe specific element dissections make clear that the maternal and gap system establish a much larger fraction of stripes than previously appreciated. The correspondence between phase 1 expression and the existence

of stripe specific elements indicates that all stripes initiated during phase 1 are driven by stripe specific elements. During phase 2 some additional stripes are initiated *de novo* by the seven-stripe elements of *ftz* and *odd*. The analysis of the *run* -3 element indicates the *run* seven-stripe element also drive expression of stripes in phase 2. The full *run* seven-stripe element is expressed in a broad domain throughout the segmented region of the embryo prior to the onset of the striped expression indicating the element is active during the whole course of *run* stripe establishment (Klingler et al., 1996). Therefore the endogenous *run* pattern is likely to integrate the activity of both stripe specific and seven-stripe elements. In contrast the lack of a seven-stripe element in *h* suggest that all pattern refinement occurs through modulation of the stripe specific elements. These results indicate that both stripe specific and seven-stripe elements are important in the initiation and refinement of stripes in the establishment of the periodic patterns.

The timing of the *eve* seven-stripe element indicates that it does not become active in all seven-stripes until phase 4 (Figure 18). PRD has been shown to activate the *eve* seven-stripe element (Fujioka et al., 1996) and the activity of the element tracks that of *prd* with a delay. Phase 3 expression is limited primarily to stripe 1 while the posterior stripes arise synchronously in phase 4 with a ventral bias. As the endogenous *eve* pattern is refined significantly in phase 2 and regular by phase 3, its clear that most of the refinement modulates the activity of the stripe specific elements. Therefore in both *h* and *eve*, the generation of the refined pattern is generated primarily through the activity of the stripe specific elements.

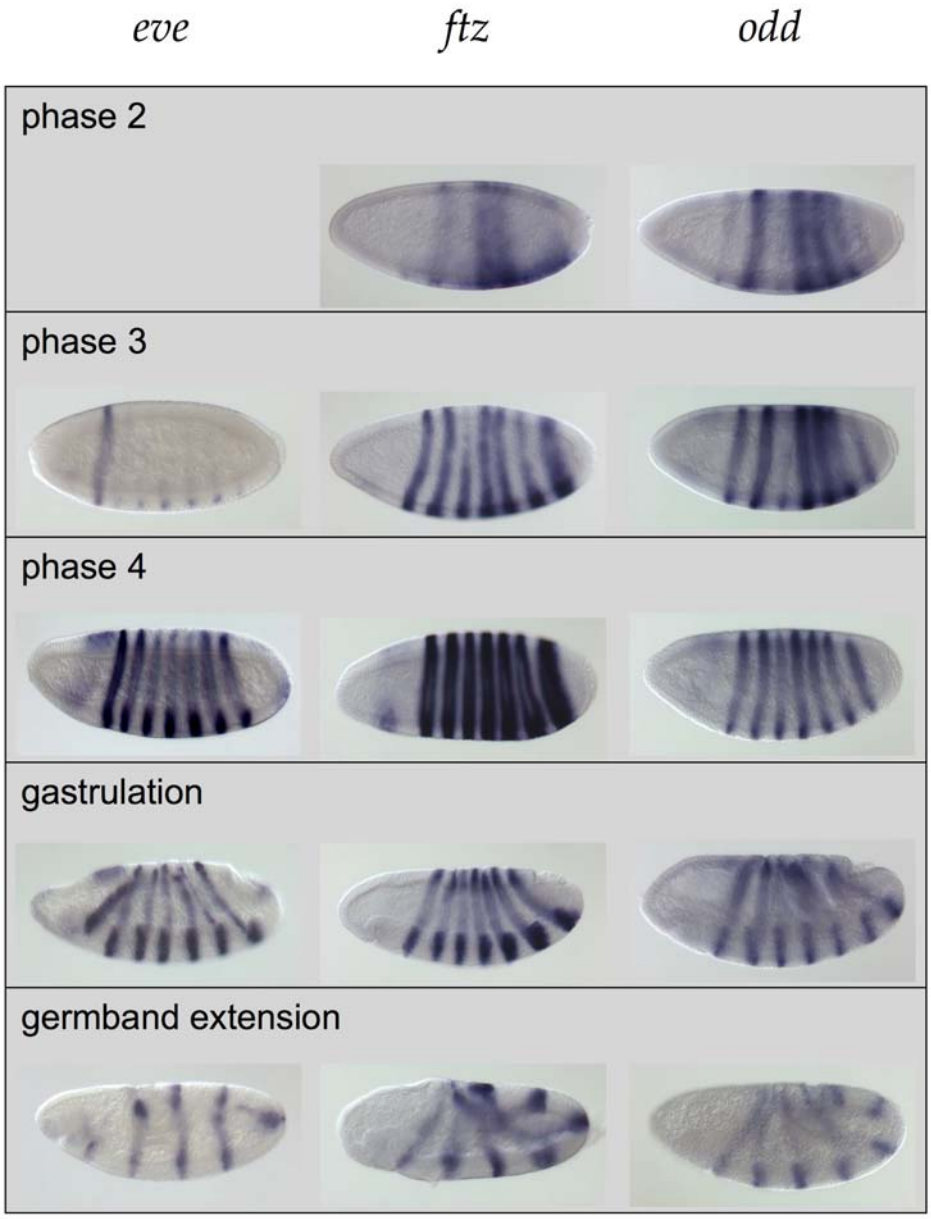


Figure 18

Seven-stripe element time course. Time course of the seven-stripe elements of *eve*, *ftz*, and *odd*. The *ftz* and *odd* elements become active during phase 2 with an extensive modulated seven-stripe pattern at phase 3. In contrast, the *eve* seven-stripe element is only expressed in stripe 1 at phase 3, with the other stripes arising synchronously during phase 4. In call cases the seven-stripes a regular at phase 4 and remain expressed through germband extension. The *ftz* element is the *ftz* lacA construct (Hiromi et al., 1985) contains the ps4 element, which may enhance expression of stripe 2 and 7.

The patterns of the *ftz* and *odd* elements initiate with a modulated pattern early in phase 2 and phase 3, but a full seven-stripe pattern is not apparent until phase 4 (Figure 18). Due to the incomplete nature of the stripe specific input into *odd* it is clear that there cannot be a simple switch from regulation by stripe specific elements to the seven-stripe element. In phase 2 the endogenous stripes 1 and 3 are quite strongly expressed, but the seven-stripe element does not drive a strong stripe 1 or 3 at this time. If the seven-stripe element did not contribute to patterning at all at this time, there would be no *cis*-element to drive stripes 2 and 4 strongly. This suggests that the early pattern is driven by both the stripe specific and seven-stripe elements at the same time. The case is similar in *ftz*, where stripes 3 and 6 are both weak at times when the endogenous *ftz* pattern has strong expression in those regions. How multiple *cis*-elements interact in the same locus to generate patterns has received little attention in the segmentation field, but the timing of the *run*, *ftz*, and *odd* seven-stripe elements suggest such interactions are likely to be important in establishment of the periodic patterns.

It is also notable that the expression driven by the seven-stripe elements of *ftz* and *odd*, which are both *ftz* targets (Nasiadka and Krause, 1999), do not track the expression of *ftz* itself. Stripe 3 of the elements arises much later than other stripes despite strong phase 1 expression of stripe 3 in the endogenous *ftz* pattern. This is particularly relevant to *ftz*, which has been emphasized as autoregulatory (Schier and Gehring, 1992, 1993). Between the existence of the stripe specific elements and the timing of different stripes in the *ftz* seven-stripe element, it is clear that the early *ftz* expression pattern at early time points is driven by very little auto-activation.

The complex early patterns generated by both the *ftz* and *odd* even-stripe elements indicate that they integrate a combination of inputs that are not inherently periodic. In *run*, *ftz,* and *odd*, maternal and gap input is predicted within the region that drives the seven-stripe pattern. Together with the experiments on the *run* seven-stripe element, this suggests that early acting seven-stripe elements receive maternal and gap input that modulates the intensity of the different stripes. Such combined input into the two types of elements could help coordinate the patterning driven by the two types of elements. However, the early modulation could also simply reflect the early non-periodic patterns of the pair rule genes themselves. The lack of concordance between the *ftz* pattern and the *ftz* and *odd* seven-stripe element driven patterns early on suggests that if this is the case, it is at least due to combinatorial interactions. In either case, the seven-stripe elements are clearly important integrators of early patterning information that are involved in the transition to periodic patterns.

The analysis of seven-stripe elements helps clarify the role of different *cis*-elements in forming the periodic pattern. The generation and refinement of the *h* and *eve* patterns involves pair rule inputs modulating stripe specific elements. In contrast *run*, *ftz,* and *odd* are patterned by a combination of stripe specific and seven-stripe elements, which are likely to both contribute to the pattern at the same time. Finally, the *slp* and *prd* seven-stripe patterns arise during phase 3, presumably by the activity of their seven-stripe elements. The mature periodic pair rule patterns only occur when all pair rule genes are expressed periodically,

which is the result of the activity of a diverse mix of stripe specific and seven-stripe elements. Generation of the periodic seven stripe patterns involves an integration of stripe specific and seven-stripe inputs, and there is no clear temporal separation between times when the two classes of *cis*-element are active in driving pattern.

The original molecular epistasis experiments focused on the mature pattern during phase 3 and phase 4, when the pattern was clearly periodic. The analysis of *cis*-element timing makes clear that this time point includes the activity of a different set of *cis*-elements in each pair rule gene. It does not make sense to focus on the role of pair rule cross regulation in phase 1 as the patterns seem to be primarily maternal and gap driven. By phase 2, it is clear that many of the pair rule patterns expressed at this time are driven by seven-stripe elements. The early activity of the *run* seven-stripe element indicates even true of a classic primary pair rule gene and that seven-stripe elements have an important role in generating early patterns.

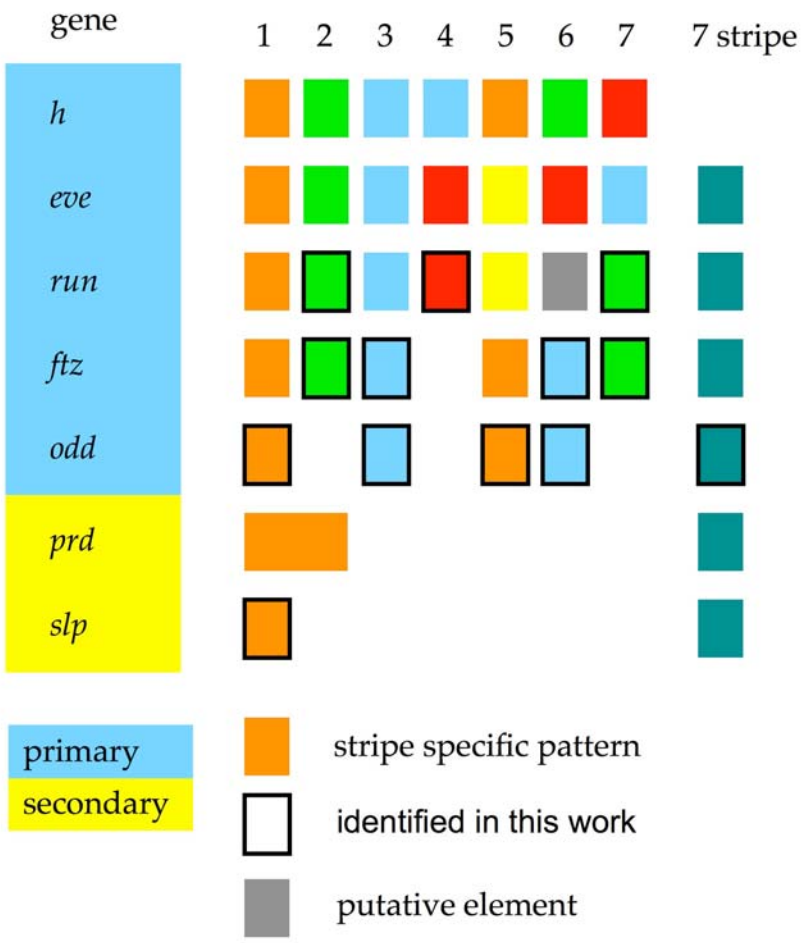


Figure 19

Summary of pair rule *cis*-dissections. Boxes represent stripes generated by stripe specific elements, color coded by stripes generated by the same element. As is clear by the black rimmed boxes, the full extent of maternal and gap input into the pair rule genes is much greater than previously recognized and includes more pair rule genes than in the original classification. The dissections presented in this work drive more than a third of the stripes generated by stripe specific elements and make clear that the majority of the *odd* and *ftz* stripes are initiated by maternal and gap input.

3.7 Summary of pair rule *cis*-regulation

The dissections and analysis of timing indicate that the maternal and gap input into the pair rules drives a much larger fraction of stripes than previously appreciated. While the initial pair rule dissections indicated that roughly half of the *h, eve, run, ftz,* and *odd* stripes were generated by stripe specific elements, the directed search for all stripe specific elements shows that the vast majority of stripes in these genes are patterned by maternal and gap input (Figure 19). Both *prd* and *slp* also have their most anterior domain specified by maternal and gap input such that all pair rule genes read the pre-existing nonperiodic pre-pattern. The extensive use of the maternal and gap patterns by the pair rule genes allows much more positional information to be transmitted to the periodic patterns than previously appreciated. The inclusion of *ftz* and *odd* also fills in a "fourth" position in the regular tiled array that these genes generate during cellularization (Figure 20).

Previous analyses of pair rule interactions suggest that there are two offset patterns generated in part by cross repression: *h* & *run* and *eve* and *odd* (Figure 20). Although the original molecular epistasis experiments suggested most of the cross regulatory interactions, the repression of *eve* by *odd*, was suggested primarily by ectopic expression experiments. Together these analyses suggest the pair rule patterns are set up by two cross repressive pairs of pair rule genes, that interact through cross regulation between *eve* and *run*. It is notable that in both of the cross repressive pairs there is one gene, *h* or *eve*, that generates the

early seven stripe pattern solely by stripe specific input and cross regulating with another gene that utilizes both stripe specific and seven-stripe elements, *run* or *ftz*/*odd*. Therefore the pairs seem to transition to periodicity in part by the temporal overlap of activities of stripe specific and seven-stripe elements. Further, *eve* and *run*, which connect the two pairs, also cross repress with *eve* utilizing only stripe specific elements and *run* seeming to utilize both a complete repertoire of stripe specific elements and an early acting seven-stripe element. Together this set of interactions generates a comprehensible framework for understanding how the seven stripe patterns are established and transition to periodicity.

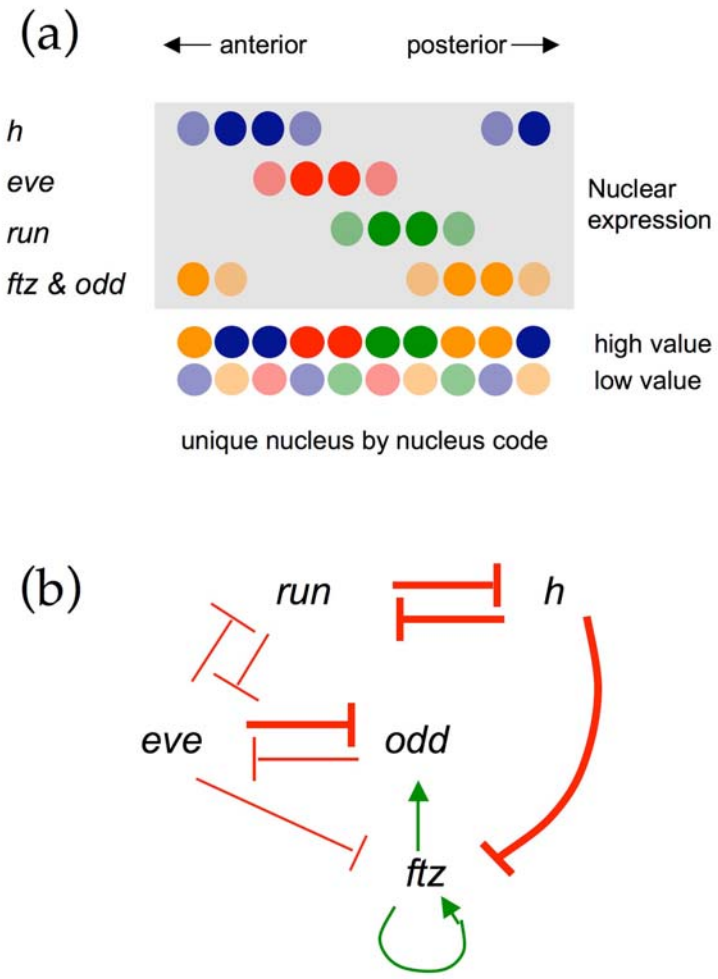


Figure 20

Pair rule cross-regulatory schema. (a) The *h*, *eve*, *run*, and *ftz*/*odd* patterns generate a tiled array of patterns with a four on, four off repeat shifted by two nuclei from gene to gene. The patterns are maximal in the central two nuclei with weaker expression in the two flanking nuclei. The overlaps give polarity to each stripe, and, in principle, the low and high expression values could give a unique expression code to each nucleus within the pattern. (b) Compiled schema of interactions, the offset patterns of *run* and *h* are generated by strong cross repression. The offset patterns of *eve* and *odd* are similarly generated by cross repression. Interactions compiled from the literature (Jaynes and Fujioka, 2004; Nasiadka et al., 2002).

3.8 Global analysis of molecular epistasis

Despite a large number of genetic studies on pair rule cross regulation, there has never been a complete comparison of the patterns and defects caused by all patterned pair rule genes. The previous studies are heterogeneous in staging and analysis of protein versus mRNA patterns, with no study looking at all patterned pair rule genes together. Furthermore, much of the work was done in the late 1980s with less sensitive RNA *in situ* protocols. As the regular strong periodic pattern is generated during phase 3, this is the time that was assayed for pattern defects. Phase 3 is preferable to phase 4 because it is prior to the onset of *eve* seven-stripe element. As the *eve* seven-stripe element is regulated in large part by PRD, this early time point is prior to the full activity of *prd* and *slp* in their later striped domains, although it is clear that they are already partially active by this time.

In order to clearly assess the complete nature of the defects, only complete loss of function alleles were used. This is particularly important for the classification of genes by presence of defects as remaining activity could reduce or eliminate the existence of such effects. In cases lacking molecularly characterized protein null alleles, deletions or transcript nulls were used. In *h*, *eve*, and *prd*, there were molecularly characterized null alleles allowing one to look at the pattern of the transcript that generated non-functional protein. In the case of *ftz*, which has been demonstrated to auto-regulate, the ability to look at the *ftz* pattern under mutant conditions is important. Previous work had

suggested that *ftz*^{9H34} was a protein null, but it had not been sequenced. Therefore the *ftz* locus was sequenced in strains containing this allele and was found to contain an amber (gln to stop) mutation after 53 amino acids eliminating most of the 410 amino acid protein including the whole homeodomain (Materials and Methods). Therefore the role of *ftz* auto-regulation in generating the early *ftz* pattern can be clearly addressed. For *run* a transcript null was used, whereas in *odd*, and *slp* deletions were used. Therefore these loci can not be analyzed in their own loss of function mutants. The *slp* deletion removes both *slp1* and *slp2*, two neighboring paralogs with very similar patterns and protein sequences.

Previous work with *odd* utilized point mutants and it was unclear if these were full loss of function alleles. The original *odd* phenotypes described involved deletion of less cuticle pattern than most other pair rule genes (Nusslein-Volhard et al., 1985). The cuticle phenotype of the p-element insertion line used to map *odd*, in contrast showed a complete loss of the odd denticle bands (Coulter et al., 1990). Unfortunately, this allele is no longer available, but a local deletion that removes *odd* and also results in the complete loss of odd denticle bands was used (Green et al., 2002). It is notable that this deletion also takes out two other paralogs, *drumstick* (*drm*) and *sister of odd and bowl* (*sob*), that are adjacent to *odd*. *sob* is expressed in an almost identical pattern to *odd*, except weaker early during blastoderm formation and stronger later during the segmental pattern post cellularization. There is an additional *odd* paralog *brother of odd with entrails limited* (*bowl*), that was not removed in these experiments and has been shown to have a function as a segment-polarity gene and also is expressed in a weak seven

striped pattern at the end of cellularization (Hatini et al., 2005). Therefore, although this deletion is null for *odd* and *sob* there may be some remaining function through *bowl*.

The central question for relevant to determining the hierarchy is whether mutants in secondary pair rule genes cause early defects in the primary pair rule gene patterns. The *h, eve,* and *run* mutants lead to rather dramatic alterations in the early patterns of all pair rule genes (Figure 21). The defects in *odd* and *ftz* mutants are much more subtle and limited. Somewhat surprisingly, both *slp* and to a lesser degree *prd* cause stronger defects in the patterns of the primary pair rule genes that clearly correspond to their anterior domains. In *slp* mutants there is a clear anterior shift of anterior pair rule stripes, particularly in *eve,* and a spacing defect in the first two *run* stripes. In *prd* mutants, *h* stripe one is clearly expressed more weakly than more posterior stripes, which is dramatic in that it is always one of the strongest stripes in wildtype embryos. The clear defects of primary pair rule gene patterns in *slp* and *prd* mutants compromises a clear classification based solely on molecular epistasis.

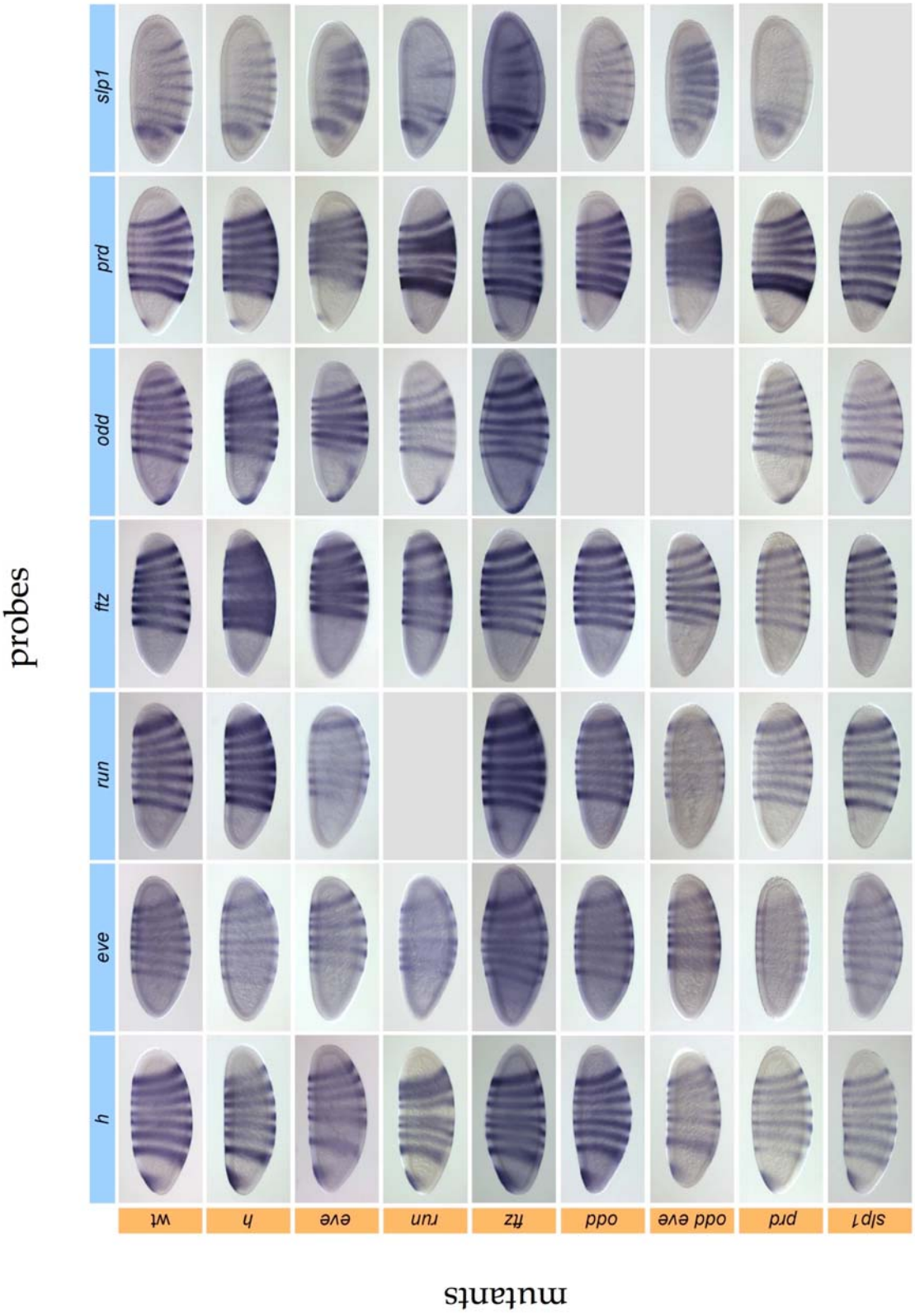


Figure 21. (figure legend on p. 98)

Figure 21

Molecular Epistasis of pair rule genes carried out at phase 3 of cellularization (to be viewed horizontally from the right). *in situ* hybridizations for all pair rule genes in all mutants. Embryos are oriented with the anterior to the left and dorsal to the top. Each row is a genotype and each column is the expression of a specific gene over the genotypes in each row. *odd, run,* and *slp* are not shown in their own mutant backgrounds as the genotypes do not produce transcript. All fixation and staining was carried out at the same time under identical conditions, except for *ftz*. The *ftz* stainings were carried out later and matched to the original conditions as best as possible. The results generally match the published literature, although there is evidence for all pair rule genes causing phenotypes at this stage. Of particular note are the notable defects in the anterior stripes in the primary pair rule genes in *slp* and *prd* mutants. In *slp*, the anterior stripes are shifted forward and clear spacing defects are present in *eve, run, ftz, odd,* and *prd*. In *prd, h* stripe 1 is weaker although it is normally one of the strongest stripes at all time points.

The types of defects seen in various mutants consist of both irregularities in intensity or width of stripes as well as improper positioning of stripes. In some extreme cases these effects can lead to loss or fusion of stripes. Mutations in *eve* and *run* lead to the strongest defects showing the greatest loss of periodicity as well as mis-regulation of stripe intensity. *h* shows strong defects in stripe intensity, although periodicity is essentially maintained. However, all seven stripes form for all pair rule genes in all mutant backgrounds (Figure 21). Therefore the extensive maternal and gap input drives all stripes directly or indirectly even in the absence of individual pair rule genes. In particular *h, eve, run,* and *odd* show clear separation of all stripes in all mutants. In *ftz, prd,* and *slp*, there is a more extensive fusion in some mutants, but a clear modulation of intensity is always present indicating individual stripes are not completely lost.

One consistent feature noted for *eve* mutants in the early literature was loss of stripe 1 in *eve, run,* and *ftz*, as well as stripe 2 of *h*. The loss of stripe 1 also occurs in *odd* and *prd,* while in contrast stripes 1 and 2 are fused in *slp*. The fused stripes in *slp* overlaps all of the missing stripes and expands dorsally over time as the loss of stripes starts ventrally and extends dorsally. This is consistent with a role of SLP in repressing these stripe specific elements and the anterior defects seen in *slp* mutants. As SLP has been previously shown to repress these stripes when expressed ectopically and SLP sites are predicted in these stripe specific elements, the role of *slp* in regulating these stripes is strongly supported. Therefore the clearest case of loss of stripes points to indirect effects generated by improper patterning.

The patterning defects seen in different mutants match up well with schema presented earlier (Figure 20) although regional differences are not captured. In *h* mutants *run* is expressed quite strongly and the *h* pattern is somewhat weakened as a result. In *run* mutants the *h* stripes are not as consistently separated and are more strongly expressed in some cases. However, the cross repression is not reciprocal with *run* expansion much more significant in *h* mutants, than *h* expansion in *run* mutants. Further the expansion of *run* in *h* mutants still leads to an essentially regular pattern, whereas in *run* mutants regularity of *h* is lost. It has previously been shown that in *run, h* double mutants the pattern of both *eve* and *ftz* are more periodic than in either single mutant (Ingham and Gergen, 1988). This suggests that imbalance due to the loss of *run, h* cross regulation leads to mis-regulation of the other gene that leads to mis-regulation of other genes.

A similar spatial arrangement is seen between *eve* and *odd*, which are expressed in a mutually exclusive pattern with cross repression indicated by previous work. *eve* mutants lead to an expansion of *odd* indicating *eve* represses *odd*, but *odd* mutants do not lead to a strong expansion of *eve*. However, the *odd, eve* double mutants are more regular than *eve* mutants indicating that the mis-patterning of *odd* in *eve* mutants effects the pattern of other pair rule genes including *h* and *run*. Further, the *eve* pattern in *odd, eve* double mutants has a fusion of the first two *eve* stripes not seen in any single mutants suggesting *odd* may play some role in setting this posterior border. Therefore some defects seen in a primary pair rule gene seem to be attributable to a mis-regulated secondary pair rule gene effecting the patterns of primary pair rule genes. This indicates

that the secondary pair rules do regulate the primaries, but may normally refine patterns in more subtle ways that are less conspicuous. When they are improperly expressed however, they help contribute to the irregularities in pattern seen in these mutants.

One significant difference compared to the commonly accepted interactions is the apparent effect of *eve* mutants on *run*. *run* expression never initiates properly and is extremely weak from phase 1 (Figure 22). No other pair rule pattern is altered significantly early enough to account for the effect. The simplest explanation is that *eve* directly activates *run*, which is consistent with previous ectopic expression studies (Manoukian and Krause, 1992). This notably contrasts with evidence that *eve* acts solely as a repressor (Fujioka et al., 2002).

There are only a few situations in which there is a clear correspondence between *cis*-elements and the pattern of defects. For instance in *run* mutants, *ftz* stripes 4, 5, and 7 are expressed very strongly. These same stripes correspond to those generated by the zebra element region and indicate that in *run* mutants these stripes are not properly restricted. In *h* mutants *run* and *odd* are both very expanded and *eve* is clearly repressed differentially in different stripes. The most strongly expressed stripes are generated by common stripe specific elements for stripes 4 & 6 and 3 & 7. However, in most cases there is no straightforward obvious correspondence between the defects and the organization of the stripe specific elements.

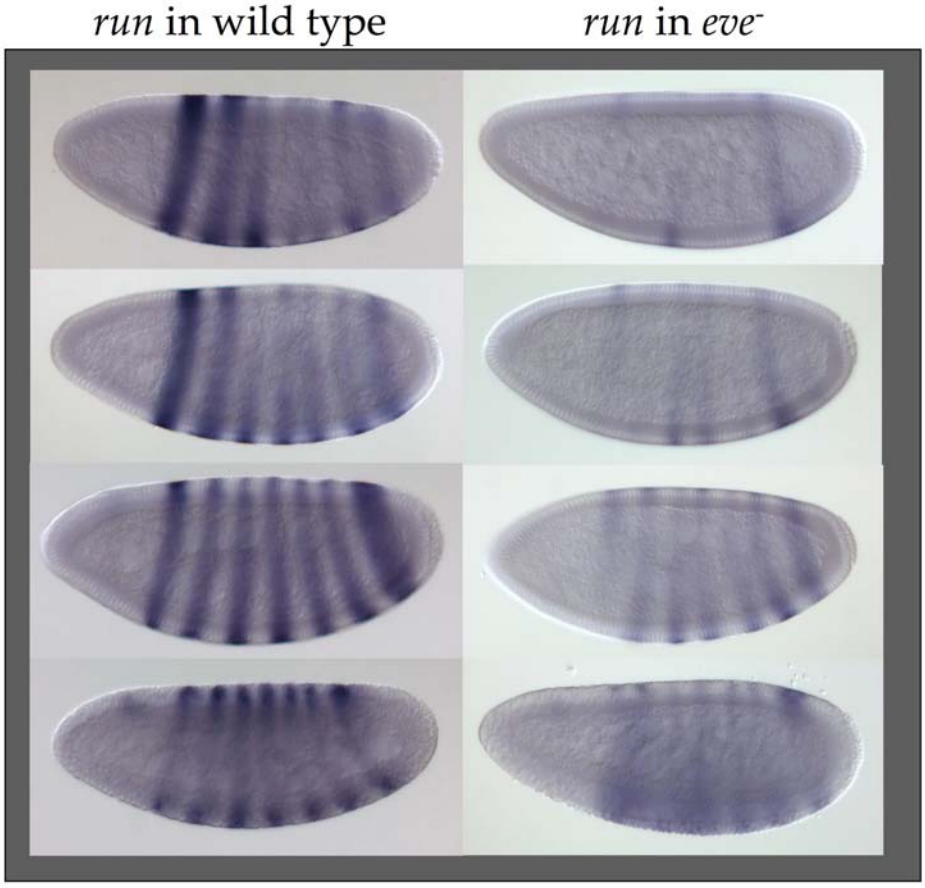


Figure 22

run expression in an *eve* mutant. The time course shows phase 1, 2, 3, and 4 *run* patterns from top to bottom. The left side shows the wild type pattern, while the right side shows the patterns in an *eve* mutant. Even phase 1 *run* expression is significantly weaker, even though at this time there is no significant alteration of other pair rule patterns that can explain the weakness of *run*. This suggests that *run* is activated by the early diffuse *eve* pattern.

In *ftz* mutants there is no clear defect in *ftz* pattern and in *odd* the defects are limited to stripes 1 and 2. The lack of defect in *ftz* pattern is notable in that earlier work on *ftz* had emphasized auto-regulation in driving the seven-stripe element (Pick et al., 1990; Schier and Gehring, 1993). The original work on establishment of pair rule patterns at this early stage supported a role for *ftz* autoregulation (Ingham and Gergen, 1988). However, these studies used the ftz^{W20} allele, which is a p-element insertion just upstream of the basal promoter that may have direct defects in *ftz* regulation. That no defects are obvious with an allele that only perturbs *ftz* function, but not regulation in other ways, makes clear that the early *ftz* pattern is not dependent on auto-regulation. This is consistent with the *cis*-dissection and time course of the seven-stripe element, which both suggest other more important inputs early.

Therefore the systematic analysis of pair rule patterns indicates a few clarifications to the original epistasis experiments. The *ftz* and *odd* patterns are *regulated* in a fashion similar to the primary pair rule genes even if they seem to have a more minor roles as *regulators*. Second the anterior expression domains of *slp* and *prd* correspond to regions where clear defects are seen in primary pair rule genes in loss of function conditions for these genes indicating molecular epistasis does not lead to a completely consistent categorization of primary and secondary pair rule genes. Finally the secondary pair rule genes are themselves responsible for some of the defects seen in primary pair rule gene mutants. This argues that they function early, but their removal leads to more subtle defects.

Therefore, it is again a matter of degree that separates the primary and secondary classes rather than a clean separation.

3.9 The pair rule hierarchy

The *cis*-dissection of the stripe specific elements alone makes very clear that *ftz* and *odd* are extensively regulated by the maternal and gap system. The similarity of regulation and the timing of expression of these genes to the role attributed to primary pair rule genes rather than that of a secondary pair rule genes (Ingham, 1988; Pankratz and Jackle, 1990; Small and Levine, 1991) argue strongly for their inclusion in the primary class. The existence of stripe specific elements alone demonstrates a primary character that is indisputable. As the role of primary pair rule genes was to interpret the maternal and gap patterns to establish the initial periodic pattern, their similar timing and regulation in establishing the early periodic patterns argues strongly for their inclusion in the primary class. In the discussion to follow, "primary pair rule genes" will include *ftz* and *odd* as they are indisputably regulated by stripe specific elements and establish their patterns in large part through interpretation of the nonperiodic maternal and gap patterns. In order to refer to *h, eve,* and *run* alone, the phrase "classic primary pair rule genes" will be used.

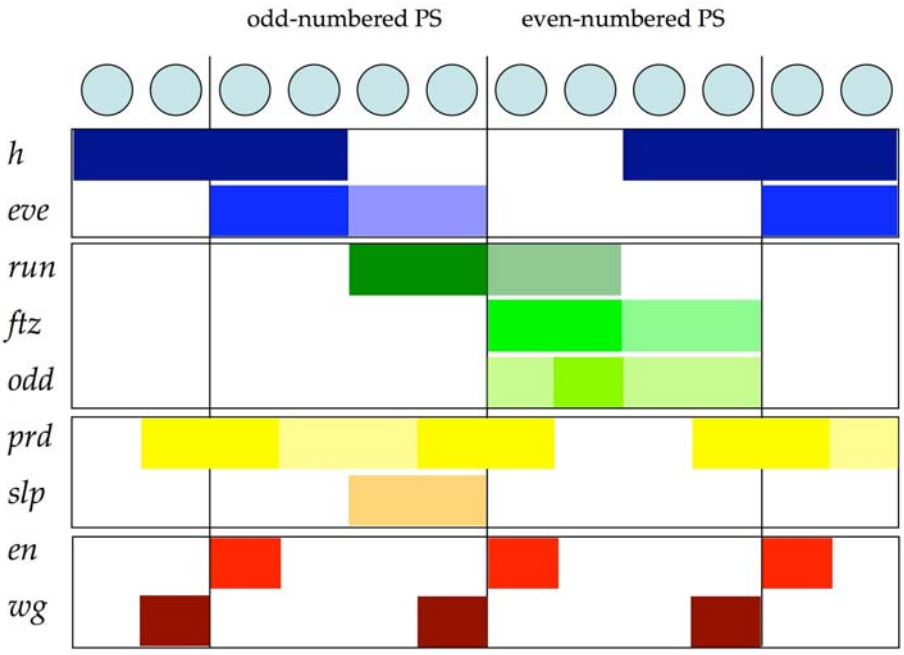


Figure 23

Spatial layout of pair rule patterns. Expression domains are shown in colored rectangles. The lighter shading indicates regions where expression is lost as cellularization completes. The initial patterns of *h, eve, run, ftz,* and *odd* form a series of offset 4 on/4 off nuclear patterns that tile the segmented region of the embryo. Stripe specific elements alone generate the initial periodic pattern of *h* and *eve*. In contrast, *run, ftz,* and *odd* have both stripe specific and seven-stripe elements active during the formation of their patterns. The pair rule genes *prd* and *slp* have their pattern generated almost entirely by stripe specific elements. The segment polarity genes *en* and *wg* are shown for reference.

The primary pair rule genes are expressed in a tiled series of seven-stripes with a 4 nuclei on, 4 nuclei off pattern (Figure 23). As one shifts from the most anteriorly expressed, *h*, to the most posteriorly expressed, *odd*, the emphasis in establishment of pattern shifts from dominated by stripe specific elements to requiring the seven-stripe element. In *h*, there is no seven-stripe element and the pattern is generated solely by stripe specific elements. In *eve*, the seven-stripe element is not active until phase 4, after the maximal regular seven-stripe pattern has formed, again indicating a reliance on stripe specific elements for establishing the periodic pattern. In contrast, the seven-stripe pattern of *run* is generated by a both a full repertoire of stripe specific elements and an early acting seven-stripe element. This indicates that unlike *h* and *eve*, which rely solely on stripe specific elements, *run* relies on both types of input to establish the 7 stripe pattern. As *run* has always been included as a primary pair rule gene, this indicates that seven-stripe elements have an important role to play in the *establishment* of the seven stripe patterns. This trend towards earlier and greater dependence on seven-stripe elements is continued with *ftz* and *odd* to a degree that now some stripes are formed solely by the seven-stripe element.

The trend in some ways continues with the shift to *slp* and *prd*, which are expressed in domains highly overlapping with *h* and *eve* respectively. Their most anterior domain, generated by maternal and gap input, regulates all anterior pair rule stripes. The later posterior stripes are completely dependent on the seven-stripe elements and are completely downstream of all pair rule genes. Therefore the trend is continued, but split completely into two separable phases of activity. The combination of their upstream role at the anterior

boundary and their downstream role in striped expression clearly relates to the complexity of hierarchy when analyzing periodic interactions. The fact that the periodic pattern along the a-p axis is set up in a relatively synchronous fashion in *Drosophila melanogaster* makes it somewhat unclear why this correspondence exists. It is notable that in ancestral modes of insect development anterior segments are set up first and posterior segments are added progressively from a growth zone established in the posterior of the embryo at the end of the blastoderm stage (Sander, 1976; Sommer and Tautz, 1993). The increasing role of the seven-stripe element along the a-p axis may relate to an ancestral role of the seven-stripe elements in the establishing new stripes within the growth zone.

The patterns of *h* and *run* are strongly anti-correlated and the two genes clearly cross repress (Figure 24). Similarly, the *eve* and *ftz*/*odd* patterns are strongly anti-correlated and *eve* represses both genes. There is also reasonable evidence that *odd* represses *eve*, although the role in positioning *eve* is relatively minor. In both cases the anterior pair rule gene is regulated strictly by stripe specific elements as the pattern forms, whereas in the posterior pair rule gene(s) are generated by a combination of stripe specific and seven-stripe inputs. The connection between the two anti-correlated pairs are through cross regulatory interactions between *eve* and *run* and repression of *ftz* and *odd* by *h*. From the epistasis experiments presented here, it is clear that *eve* and *run* have important roles in generating periodicity. Their cross-regulation is the only clear strong interaction between the classic primary pair rule genes and therefore a critical one in integrating the information from their combined set of stripe specific elements. It is notable that as shown with *eve*, *odd* double mutants and

previously reported for *run, h* double mutants, much of the periodicity is returned to the pattern. This suggests that there are balancing interactions between the anti-correlated pairs that are critical for maintaining proper regularity of pattern.

There is a more direct relationship between early and late pair rule function and the role of stripe specific versus seven-stripe elements. Neither *eve* nor *h*, regulate *en* directly, whereas *run*, *ftz*, and *odd* do (Figure 24). The stripes of *en* arise as cellularization is completing and the regulation of *en*, by *ftz* and *odd* is particularly important as neither interaction is redundant with other inputs. Although *run* also directly regulates *en*, the anterior border of the *en* domain is also set by *slp* (Jaynes and Fujioka, 2004). Therefore, the early *h, eve,* and *run* patterns can be more easily optimized for blastoderm stage patterning without interfering with their later role. In contrast, the *ftz* and *odd* patterns themselves need to be precisely positioned at single cell resolution as cellularization completes.

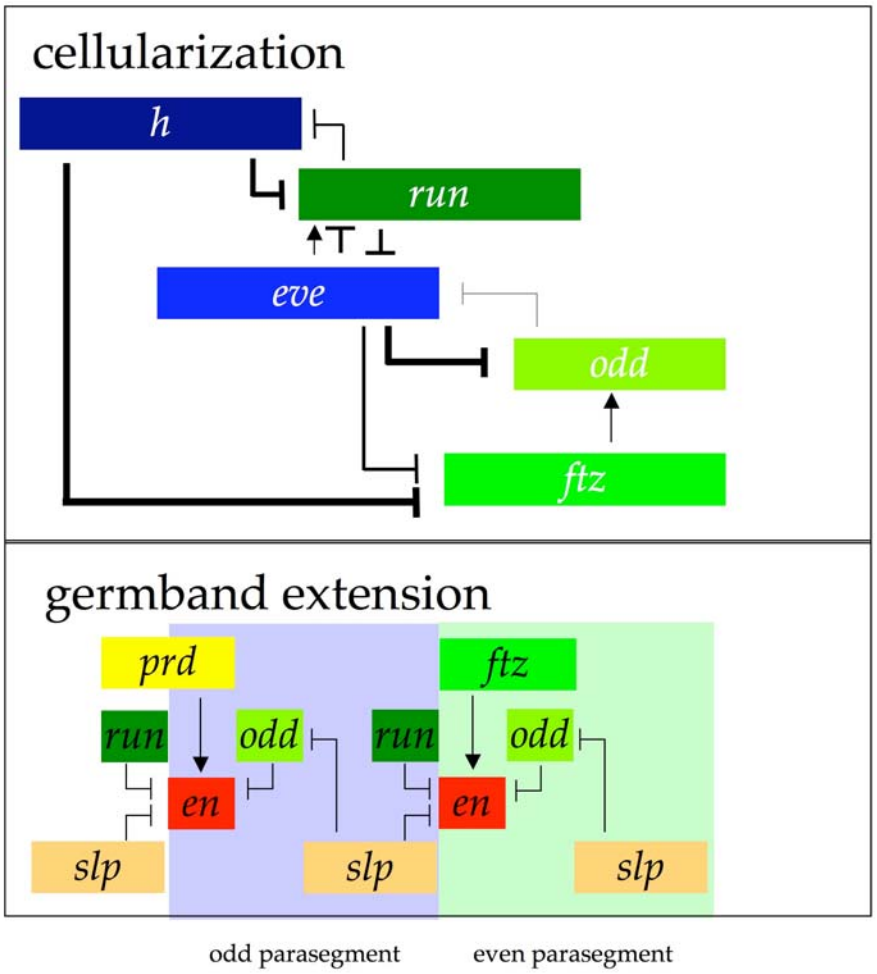


Figure 24

Top: Pair rule cross regulatory schema, the anti-correlated pairs *h* & *run* and *eve* and *odd* cross repress. The colored rectangles represent domains of expression. In both cases the more anteriorly expressed member of the pair is expressed strictly from stripe specific elements while the posterior member is expressed from both stripe specific and seven-stripe elements. The anterior member also more strongly represses the posterior component. The interactions between *eve* and *run* are critical to generating periodicity and link the positioning of the two sets of genes.

Bottom: The later function of pair rule genes shown to directly regulate the *en* pattern. It is notable that *run*, *ftz*, and *odd* all directly regulate *en*, whereas *eve* and *h* do not. This matches well with *h* and *eve* being more optimized for early function. *run*, which has both an early role in generating periodicity and an early acting seven-stripe element regulates *en* redundantly with *slp*.

The simple categories therefore may miss nuances that are important to understanding the role of each gene within the network. The "primary" and "secondary" pair rule system in some ways emphasized the importance of *h*, *eve*, and *run* above the other pair rule genes in a fashion that unfairly de-emphasizes the importance of the other pair rule genes, which each serve a unique and critical function. The fact that *ftz* and *eve* label alternate parasegments during cellularization indicates an analogous and important role for each gene. Previous studies have indicated that relationship between *eve* and *ftz* activity during cellularization sets the size of parasegments (Hughes and Krause, 2001). Therefore, whether acting directly or indirectly, the expression of these two genes establish the initial morphological and molecular unit of the repeated body plan in an analogous way.

The original *ftz* rescue experiments, in which the stripe 3+6 element was missing, show defects in segments A1 and A7, where these stripes are expressed (Hiromi et al., 1985). This indicates that the early *ftz* expression is required for proper patterning by *ftz*. The fact that *ftz* and *odd* are both FTZ targets and need to be precisely patterned to properly position *en* suggests that their role might be in very fine scale adjustments of pattern. Therefore, even if they do regulate the primary pair rule genes, they may be involved in making fine scale adjustments to the pattern. If this is the case, it would require precise measurements that have not been made so far on pair rule cross regulation. Such measurements are difficult to do given the dynamic nature of the pair rule patterns, which vary early and only gradually become precise.

The final point that requires mention is the general model of how pair rule genes act. In genetic analysis pair rule genes have always been ascribed either a role of activation or repression. However, in the extensive work done using ectopic expression, there have been consistent indications of changes in function that may involve combinatorial interactions (Manoukian and Krause, 1992; Nasiadka and Krause, 1999; Swantek and Gergen, 2004; Vanderzwan-Butler et al., 2007). Furthermore, target selection is clearly dependent in some cases on protein-protein interactions of pair rule proteins with each other and other transcription factors that regulate pair rule *cis*-elements (Copeland et al., 1996; Florence et al., 1997; Yu et al., 1997). Indeed, the transition of *ftz* from having a homeotic role in ancestral insects to a role as a pair rule gene has been shown to depend in part on evolution of a protein-protein interaction with Ftz-F1 (Lohr and Pick, 2005; Lohr et al., 2001). These studies paint a very different picture of pair rule regulation, but there still no well defined examples where other interpretations have been ruled out. In most cases, the genetic evidence for these factors acting as both activators and repressors are lacking. One exception is *run*, which clearly acts as a direct activator in sex determination (Kramer et al., 1999), but has a clear role in repressing multiple targets in segmentation (Jaynes and Fujioka, 2004). Therefore, it is of interest to supply genetic evidence that additional pair rule proteins can switch function. This would argue further for more combinatorial models where pair rule genes can act as either activators or repressors.

Our result that *eve* mutants result in weakening of *run* activation provides further evidence for such relationships. This interaction has also been seen previously in ectopic expression experiments in which *eve* is expressed early (Manoukian and Krause, 1992). Previous work with *eve,* in which it was rescued with the *eve* homeodomain fused to the repressor domain of *en*, suggested that *eve* was a dedicated repressor (Fujioka et al., 2002). However, this experiment leaves the *eve* homeodomain intact and homeodomains are known to mediate protein-protein interactions in some cases (Ohneda et al., 2000; Plaza et al., 2008; Zappavigna et al., 1994). Therefore, one interpretation of the data is that *eve* may interact with other proteins through the homeodomain to activate *run* transcription. Although this is a relatively simple piece of data relating to a complex topic, it encourages further work looking at combinatorial interactions in pair rule proteins. Whether pair rule genes act in direct combinatorial interactions in regulating their targets is a very important distinction. As there is consistent data pointing to this interpretation, it could be another complex feature of pair rule regulation that has obscured a clear model of how these proteins work together to establish the periodic patterns of development.

In sum, the data presented here argues strongly for a reclassification of the *ftz* and *odd* as primary pair rule genes. These genes clearly interpret the maternal and gap gradients to help establish the periodic patterns of the embryo, which is the main function attributed to the primary class. The detailed analysis of the timing, *cis*-regulation, and genetics of pair rule pattern formation paints a much more nuanced view of how pattern is formed. Each pair rule gene plays a distinct and unique role in establishing the initial patterns. That these roles are

organized along the a-p axis suggests one possible causative reason for the differences based on evolutionary considerations. Another, not necessarily incompatible, reason for the differences may relate to whether they regulate segment polarity genes directly or indirectly. Whatever the role the different genes play, this analysis clarifies the role and extent of different *cis*-elements in establishing the periodic patterns. The patterns apparent in this extended catalog of elements thereby suggests new directions for understanding the *cis*-architecture and cross regulation of the pair rule sub-hierarchy, which is so central to segmentation.

Chapter 4: Binding site analysis

Together with the previous literature, the set of *cis*-elements regulated by the maternal and gap genes that establish the gap and pair rule genes is essentially complete. In Chapter 2, an analysis of maternal and gap regulated *cis*-elements at a composite level was presented. However, it is of interest to understand how these elements encode position in a more detailed fashion. The original predictions were in part limited by unspecific matrices for KNI and TLL, as well as an overly specific matrix for GT. In addition, the gap gene HKB, which has generally been studied in less detail than the other gap genes, was not analyzed. The publication of bacterial one hybrid (B1H) PWMs for all gap genes (Noyes et al., 2008b) allowed improvement of the KNI, GT, and TLL matrices and provided a HKB matrix.

Although the B1H PWMs are very specific on their own, these matrices were very helpful in revisiting the known binding sites. In the case of KNI and TLL the B1H PWMs were used to align the known sites to generate improved matrices. The sites in the original PWM were misaligned in part due to a reasonable number of weak sites from the literature. Therefore, the B1H sites were also included in the alignments for matrix generation as the stronger sites in this data set maintained the core preferences despite the addition of weaker sites. This was considered preferable to making arbitrary choices of what footprinted sites to include and yielded a good correspondence with targets described in the literature. Both the realignment and the addition of the B1H sequences contributed to the improvement in the matrices.

In the case of GT, the small number of footprinted sites led to the combined matrix remaining too specific. The *eve* stripe 2 element, which has been shown to contain functional GT sites by site directed mutagenesis (Arnosti et al., 1996; Small et al., 1992; Small et al., 1991; Stanojevic et al., 1991), was not predicted as a GT target using the combined matrix. However, the B1H GT matrix was completely palindromic and there was a clear palindromic nature to the earlier PWM generated from the footprinted sites as well. Therefore a trivial way to double the sampling of sites was to include all reasonable matches to the B1H PWM from the footprinted sites in both orientations. This GT matrix matches up well with known GT targets, though the best described targets are the *eve* stripe 2 element, the *Kr* CD1, and *Kr* CD2 elements, which all contain footprinted sites.

For prediction of binding sites Stubb was run differently than for module prediction. The original design of Ahab and Stubb is to fit maximize the free energy for each window to generate a sensitive measure of whether the region is better explained by containing sites from the dictionary of PWMs versus a simple background model. However, this fitting distorts the estimation of the posterior probability of a PWM match to the same sequence in a number of ways. First, the set of transition probabilities, which correspond to an estimate of the fraction of nucleotides assigned to the PWM, is fit and therefore varies between windows. The difference in transition probabilities between elements was handled by setting them to the same low value (0.0005) for all elements. Second, the use of a number of PWMs can lead to partial site overlaps in some windows,

but not others so sites were predicted for each factor separately. Finally a local background can effect site prediction as background is the most likely assignment for most of the sequence. Therefore the same global background was used for all elements. Finally, instead of running the matrices over windows of a fixed length, the actual mapped elements were used. Together these adjustments in usage generate consistent comparable binding site predictions within each element.

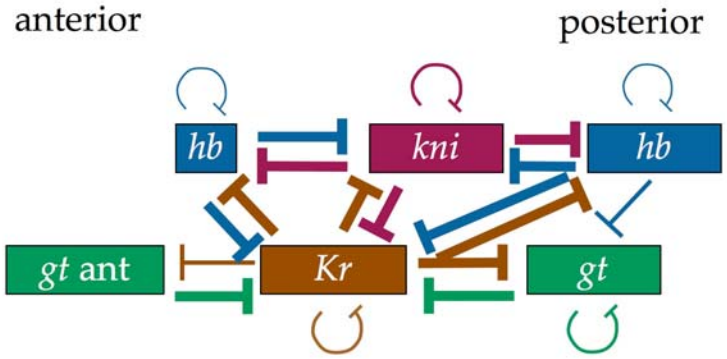
4.1 Gap Gene elements

Although much genetic work has been done on the gap genes and the general outline of their regulatory interactions are clear, there is no agreed upon network in the literature. The focus here will be on *hb*, *Kr*, *kni*, and *gt*, which are the main gap regulators that pattern the segmented region of the embryo. Despite many modules being footprinted, the majority of elements have not been. Even in extensively footprinted elements all possible regulators have not been tested and new inputs compared to those footprinted are predicted in every element. The footprinting of the modules was not done in a quantitative or consistent fashion so the strengths of different sites were not determined in a comparable fashion. Therefore, the use of Stubb to predict sites over this enlarged set of elements is a unique data set in both completeness and consistency.

This data is particularly useful in comparison to the genetic data as redundant interactions and indirect effects cloud what regulatory interactions

exist in mutant analyses. Therefore site predictions, which indicate direct interactions are a useful for understanding what the connections are between different genes. In particular it is unclear what role the maternal system plays in initially biasing the gap domains versus how cross repression resolves initial positioning as all gap genes are activated by the maternal input.

The predicted network is shown in Figure 25 with stronger predicted interactions shown with thicker lines. For simplicity in *gt* and *kni* only the interactions predicted from the single domain elements are depicted. In *gt* the two domain element has the same predicted gap repressors except for GT itself. The *kni+1* is improperly expressed and is likely to be missing key inputs and was therefore not included. Finally only the late acting *hb* element is depicted as it is the primary target of the gap regulation. The *hb* anterior element is predicted to contain only weak gap gene inputs from KR and GT. As the evidence supports a primarily repressive role for gap genes the interactions will be schematized as repressive in nature. Because the predicted sites do not directly address mode of action, this is a reasonable simplification that fits with the analysis presented in Chapter 1 (Figure 11a). Further, early examples such as the activation of *kni* by *Kr*, were later indicated to be indirect (Capovilla et al., 1992; Pankratz et al., 1992).



cis-element	Bcd	Cad	Dstat	Hb	Kr	Kni	Gt	Tll	Hkb
gt anterior	2.1				1.1	1.3		1.0	1.0
hb late		0.9	0.5	1.4	3.0	1.1			
Kr CD2/AD1	1.4	1.6		3.2	1.8	1.3	1.4	2.1	
Kr CD1	1.6	1.5		3.3			0.8	1.2	0.5
kni	1.8	2.1	1.0	3.4	3.3	1.0		3.6	
gt posterior	0.9	1.9	1.3	2.2	2.9		0.7	0.6	

Figure 25

Gap cross regulatory schema. At top a schematization of the rough location of domains for the gap genes with anterior to the left and posterior to the right. The predicted interactions are depicted with blocked arrows that are weighted by the predicted strength of the interaction. Although HB has been shown to activate in some cases and some interactions may not be strictly repressive, for simplicity of depiction all connections are shown as repressive. Below the schema a table of the predicted inputs based on Stubb dictionary values is provided. The single domain elements are listed for *gt* and *kni* as they match up best with effects seen in mutants. For *Kr*, both the CD1 and CD2/AD1 elements are included in the depiction. A complete table for all gap elements is available in the Appendix.

The network is quite extensively interconnected, with strong cross repressive interactions predicted in particular for *hb*, *Kr*, and *kni*. The most "connected" node is *Kr*, which is involved in cross repression with all other gap genes and sits in the center of the embryo. Notably one of the two predictions of the highly specific HKB PWM is *Kr*, although HKB is expressed distantly from *Kr* at the poles. A similar arrangement exists for *kni*, although it is not predicted to receive GT input, but instead very strong TLL input (not schematized) that could set the posterior border. As the maternal systems generate activities maximal the poles, the generation of central gradients is of great importance, which is dramatically demonstrated in the predicted network.

Unlike the central domains of *Kr* and *kni*, the *gt* domains, which are expressed closer to the poles seem to be positioned in larger part by activating inputs. It was originally hypothesized that *bcd* sets the posterior borders of many genes directly through limiting activation, but this does not seem to be the case (Driever et al., 1989; Ochoa-Espinosa et al., 2005). However, the predicted input into the *gt* anterior domain includes very strong BCD input and only relatively weak KR input, suggesting limiting activation is important in setting the posterior border. The strong KNI input into the *gt* anterior domain is likely to be related to the splitting of this domain late as a stripe of *kni* arises around the same time in the same location. The *Kr* central domain and the *gt* posterior domain are involved in strong cross repressive interactions, but the more posterior location of the *gt* domain again seems to be due to activating inputs. Whereas *Kr* is predicted to have intermediate to strong BCD and CAD inputs, *gt* is predicted to have very strong CAD input and intermediate to weak BCD

input. Therefore the central *Kr* and *kni* domains seem to be differentially positioned by repressive inputs from the termini, the positioning of the *gt* domains seems to be due more to limiting activation.

Therefore the relative importance of activating and repressing inputs seem to be different for the central versus terminal domains. The relative positioning of *Kr* and *kni*, which both have central positions and are involved in strong mutually repressive interactions may well due to biases in activation. While *Kr* has balanced intermediate activation by both BCD and CAD, *kni* has very strong CAD input, somewhat stronger BCD input, as well as DSTAT input. This suggests a posterior bias for *kni*, but it is unclear how exactly the two domains resolve. In all cases it is apparent that multiple inputs are involved in setting the stripe borders from the high density of predicted interactions. With this better idea of what interactions exist it is useful to compare the results to the published genetic data.

Previous work had highlighted the cross repression between two specific gap gene pairs, which are expressed in reciprocal domains (Clyde et al., 2003; Kraut and Levine, 1991a). The first pair is *Kr* and *gt* in which the central *Kr* domain is cradled between the two *gt* domains. A similar arrangement occurs between the central *kni* domain and the anterior and posterior *hb* domains. The cross-regulatory interactions in both cases have been shown to be important in positioning the domains. In both cases however, there are other strong stabilizing inputs indicating that the repressive interactions do not function alone.

In the case of *hb*, removal of *kni* leads to an anterior expansion of the posterior *hb* domain, removal of *Kr* leads to expansion of both domains towards the center, and removal of both *kni* and *Kr* leads to expression throughout the central region between the anterior and posterior domains (Clyde et al., 2003). In the case of *kni,* removal of zygotic *hb* leads to an anterior expansion, while the posterior border is still set properly (Hulskamp et al., 1990). The posterior border of *kni* is set by the terminal system (Rothe et al., 1994). Therefore, the cross-repression is always supported by additional interactions and is not solely limited to the cross-repressive pair. Interestingly, the setting of each border is dominated by one of the two genes. The border between the anterior *hb* domain and the posterior *kni* domain is set primarily by *hb*, but also positioned by *Kr* repression of *hb*. In contrast the anterior border of the posterior *hb* domain is set primarily by *kni*, while the abutting border of the *kni* is set by *tll* input. There is therefore a primary direction of flow from one regulator to the other. The manner in which *Kr* regulates *kni* is less clear as indirect effects occur and both activation and repression have been claimed (Capovilla et al., 1992; Pankratz et al., 1989).

A similar situation exists in the case of *gt* and *Kr*, where other inputs are important in positioning the interface between the two domains. *Kr* is initially patterned correctly in a *gt* mutant (Eldon and Pirrotta, 1991), but shows an anterior expansion later (Wu et al., 1998). The early *Kr* pattern is set at the anterior by *hb* and expands anteriorly in a *hb* mutant (Hulskamp et al., 1990). Similarly the posterior border of *Kr* expands posteriorly in a *kni* mutant (Gaul

and Jackle, 1987), arguing that both borders of *Kr* are set in part by inputs other than *gt*. There is an anterior expansion of the posterior *gt* domain in a *Kr* mutant, but no effect on the anterior domain (Eldon and Pirrotta, 1991). Therefore, the anterior border between the genes is set more by GT and the posterior by KR, again indicating a primary direction of flow from one gene to the other.

With respect to the role of activation, genetics support the idea of biasing inputs although indirect effects are surely involved as well. In a *cad* maternal and zygotic mutant embryos the posterior *gt* domain is almost completely lost, the *kni* posterior domain is significantly weakened, and the *Kr* domain is only marginally weakened (Olesnicky et al., 2006). Therefore the more posterior domains are increasingly more dependent on CAD activation. In embryos lacking *bcd* function the anterior *hb* and *gt* domains are completely lost, whereas *Kr* is still fairly strongly expressed (Hulskamp et al., 1990; Struhl et al., 1992). Therefore the anterior domains are more dependent on BCD activation than the centrally expressed KR. As in a *kni* mutant the posterior border of the anterior *gt* domain is still set correctly, KNI is unlikely to play a critical role in setting this border (Kraut and Levine, 1991b). Interestingly, the *kni* posterior domain is more sensitive to loss of BCD activation than *Kr*, indicating *kni* is more limited by activation in by both maternal gradients although why is not clear. While the maternal gradients do not seem to provide fine-grained positional input, they clearly bias the zygotic gap patterns sufficiently that gap cross regulation leads to a unique ordering of the expression domains.

The primary difference between the predictions presented here and our previous work (Schroeder et al., 2004; Segal et al., 2008) is that "self" predictions are consistently predicted. Such inputs are predicted for every domain except the *gt* anterior domain. As the new matrices seem to be of good quality, it is unlikely to be spurious prediction. Prediction of self-regulation is suggestive of auto-activation, but alternate explanations exist. Negative feedback at high expression levels can flatten gradients and thereby increase the spatial extent over which targets can effectively read the concentration differences (Eldar et al., 2004). Outside of *hb,* feedback of gap genes onto their own regulation has not been systematically studied. The prediction of this input into a large number of the gap *cis*-elements warrants further investigation.

This analysis therefore supports a simple model where the tiled patterns of the gap genes are generated by a chain of offset cross-repressive interactions. At the interface between any cross repressive pair one gene typically dominates the interactions. The positioning of this dominant gene is then generated primarily by other inputs that are positioned by other mechanisms. There is thereby a primarily a unidirectional flow of information through the genes despite many cross repressive interactions. This cross repression presumably aids in refinement of the position of each border. It is interesting that the pair rule interactions seen in the molecular epistasis experiments also followed this basic logic.

4.2 Stripe specific elements

The large number of stripe specific elements is a perfect opportunity to look at how patterns are encoded over a range of positions spanning the segmented portion of the embryo. Given the common functional role of the stripe specific elements in coordinating the establishment of the pair rule patterns, they are a perfect data set for examining how the inputs that establish a transcriptional sub network are organized. A general overview of the results supports the *eve* stripe 2 model, where the stripe-specific elements read off the long-range activating gradients of the maternal factors and the shorter-range repressive gradients of the gap factors with most of the positional information generated by repressive gap input (Arnosti et al., 1996; Small et al., 1992; Small et al., 1991; Small and Levine, 1991). In order to better visualize gap input into stripe specific elements, the expression patterns of the *cis*-elements were plotted in conjunction with the patterns of the predicted gap inputs (Figure 26).

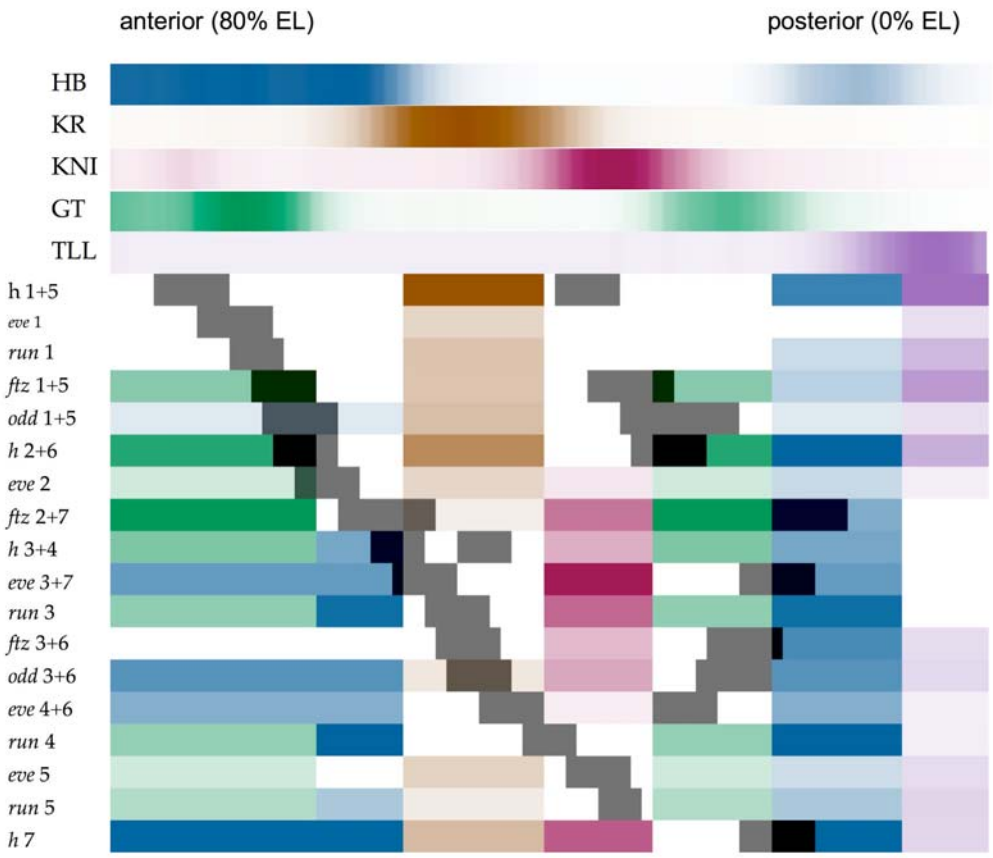


Figure 26

Stripe specific element regulation schema. Graphical depiction of stripe specific expression compared to repressive gap inputs plotted along the a-p axis (100% EL is the anterior tip of the embryo and 0% EL is the posterior pole of the embryo).

Top: The gap gradients are schematically plotted with different intensities corresponding to different expression levels. These domains are then represented in an on/off pattern below to allow representation of strength of predicted input.

Bottom: The patterns of cis-elements are plotted in black and sorted by anterior boundary of expression. Darker colored gap domains represent more predicted input. When BCD and HB are both predicted in the anterior HB is not plotted under the model it would activate. A clear anti-correlation between predicted gap input and *cis*-element expression is clear, with use of gap inputs shifting as one moves posteriorly through the embryo.

The most evident feature of the schematic is the almost complete anti-correlation between the predicted gap input and the stripe specific element expression shown in Figure 25. This is consistent with gap genes acting primarily as repressors to define the borders of stripe expression. The one clear exception to this rule is the role of HB in activation of certain *cis*-elements such as *eve* stripe 2 (Arnosti et al., 1996; Small et al., 1992). Since HB can switch providing activation in the presence of BCD sites (Simpson-Brose et al., 1994), HB was not plotted in the figure when BCD sites are predicted. When looking at the pattern of *cis*-element expression in relation to the gap gene inputs along the a-p axis, there is a steady shift of what gap gene is not predicted, leaving a region free of repression in which expression occurs. Given the tiled nature of the gap gene patterns there are adjacent gap domains well positioned throughout most of the trunk region of the embryo. However, there is an absence of clear repressive inputs to set the stripe 1 borders of *h, eve,* and *run*.

The most notable lack of repressive input is anterior to the GT domain. Previous work on *eve* stripe 2 has implicated SLP as an important input in this region (Andrioli et al., 2004; Andrioli et al., 2002). The results in Chapter 3 support the role of SLP in setting the anterior border of stripe 1 of *eve*, *run*, *ftz*, and *odd* as well as stripe 2 of *h*. However, these borders are still defined in *slp* mutants and *h* stripe 1 does not appear to be a target of SLP repression, indicating there are other anterior inputs. Although not schematized, HKB input is predicted in all elements generating stripe 1 based on the B1H PWM. As HKB is expressed at both the anterior and posterior terminus, this input would help

fill a gap role in setting the anterior stripe 1 boundary, which is not clearly supplied by the gap genes of the trunk. Despite testing the majority of gap and pair rule mutants, previous work on the *eve* and *run* stripe specific elements found no mutant conditions where the anterior or posterior, stripe 1 boundaries were not maintained (Fujioka et al., 1999; Klingler et al., 1996). *hkb* was notably not tested in either case, although the involvement of *slp* suggests that multiple inputs may set this border together. Our results suggest HKB and SLP play an important role in specifying this boundary, which the gap genes of the trunk are poorly positioned to do. In general the stripe 1 borders seem to have more redundant inputs than the other stripes.

Stripe-specific elements driving two stripes that straddle the KR and KNI gradients generate roughly half the stripes with the two interior borders specified by one repressive interaction. This arrangement has been noted in the *eve* locus (Clyde et al., 2003), where the 4+6 and 3+7 stripe elements of *eve* have been shown to read off two different concentrations of KNI to generate a pair of nested domains, with the outside borders specified by HB. The 3+6 stripe elements of both *ftz* and *odd* utilize this same arrangement to generate a pair of stripes from a common element. This is an efficient way to generate 8 borders of expression from symmetric use of two inputs. It is notable that KNI and HB make up one of the cross-repressive pairs seen in the gap gene network indicating that the stripe specific elements read off reciprocal gradients established by cross-repression. Additional stripe elements also use this combination of inputs including *h* 3+4, *run* 3, and *h* 7. Therefore there are a class

KNI regulated stripe specific elements, which generate stripes flanking the posterior *kni* domain.

The dual stripe elements reading off the KR gradient generate stripes 1 and 5 of *h*, *ftz,* and *odd*, as well as stripes 2 and 6 from *h*. In the case of both the *h* 2+6 element and the *ftz* 1+5 element, the exterior borders are specified by GT in a relatively symmetric fashion. Therefore some members of the second set of dual stripe elements make use of the KR and GT cross-repressive pair. However, as the stripes generated by this set of dual stripe elements are not as symmetrically positioned on the KR and GT gradients as those of the KNI, HB case, the stripes cannot all be generated in a completely symmetric fashion due to their offset positions in the embryo. It is therefore notable that the *eve* and *run* stripe 1 elements are predicted to have KR input, but no GT input. This is consistent with the primarily anti-correlative arrangement between gap gene patterns and the expression of their targets as *eve* and *run* stripe 1 fall within the anterior GT domain. In contrast the *eve* and *run* stripe 5 elements are predicted to have both KR and GT input and therefore do make use of the same interactions consistent with lack of co-expression with GT. Therefore there is also a large class of KR stripe specific elements, which generate stripes flanking the posterior *Kr* domain.

The famous *eve* stripe 2 element also uses the GT, KR combination, but seems to fit in better with the *ftz* and *run* 2+7 elements, which generate two borders by straddling the KR, KNI, and GT gradients. Although delineated as generating stripe 2 alone, the region containing the *eve* stripe 2 element can also contribute to stripe 7 expression when included in larger constructs (Goto et al.,

1989; Harding et al., 1989; Hare et al., 2008; Janssens et al., 2006). It is notable in this regard that the *eve* 3+7 element is expressed more weakly in stripe 7 than stripe 3 although both are expressed at similar levels in the endogenous pattern (Small et al., 1996). It is also interesting that the *ftz* 3+6/7 element and 2+7 elements clearly interact to generate stripe 7 together (Chapter 3) as seems to be the case in the *eve* locus. The 2+7 stripe elements thereby share some commonalities with both classes of element, with stripe 2 set by GT and KR, but with the central region also repressed by KNI.

Stripe 7 of all pair rule genes overlap the posterior *hb* domain, which has a lower expression level than the anterior HB domain. Although there is evidence that *hb* helps set the posterior border of stripe 7 in *eve* (Clyde et al., 2003; Small et al., 1996), the posterior border is still maintained in *hb* mutant embryos. One needs to remove *tor* for the posterior border of stripe 7 to expand to the posterior pole of the embryo (Small et al., 1996). The same is true for *h*, where the terminal group has to be removed for strong posterior expansion of the stripe (La Rosee et al., 1997). In *run* the posterior border also seems to be set in part by *hkb* (Klingler and Gergen, 1993). However, the prediction of TLL and HKB into the stripe 7 elements appears incomplete, with neither predicted into the *eve* 3+7 element. Unlike the stripe 1 elements, which are predicted to contain HKB sites, the stripe 7 generating elements are not predicted to have HKB except for the *run* 2+7 element. Stripe 7 is much more closely expressed to the posterior terminus than stripe 1 to the anterior terminus, such that the stripe 1 elements would require much stronger input to set their borders. As the HKB B1H PWM is highly specific, the weaker sites likely to occur in these elements might go undetected.

The same problem might be the case for TLL input into the *eve* 3+7 and *ftz* 2+7 elements. Therefore the genetics and predictions together suggest that as in stripe 1, the terminal group genes are important in defining the posterior boundary of the segmented region.

Therefore the gap inputs are organized in a consistent fashion throughout the set of elements. Most notably there are three classes of dual stripe elements with their central region carved out by either KR, KNI, or both together with GT. Since the *cis*-elements drive expression at different positions, it is of interest to compare the positioning of stripes on the gap gradient and the strength of the input they receive. For both KNI and KR, there are a large number of elements with their borders set primarily by the same domain making it straightforward to look at the correspondence between input and positioning. In both HB and GT, this is not as straightforward as the posterior domains are not as strongly expressed as the anterior domains. Additionally, the posterior border of a large fraction of HB and GT regulated stripes receive input from HKB and TLL, further complicating such an analysis. Finally for HB, the switch between activation and repression further complicates analysis of HB input and output.

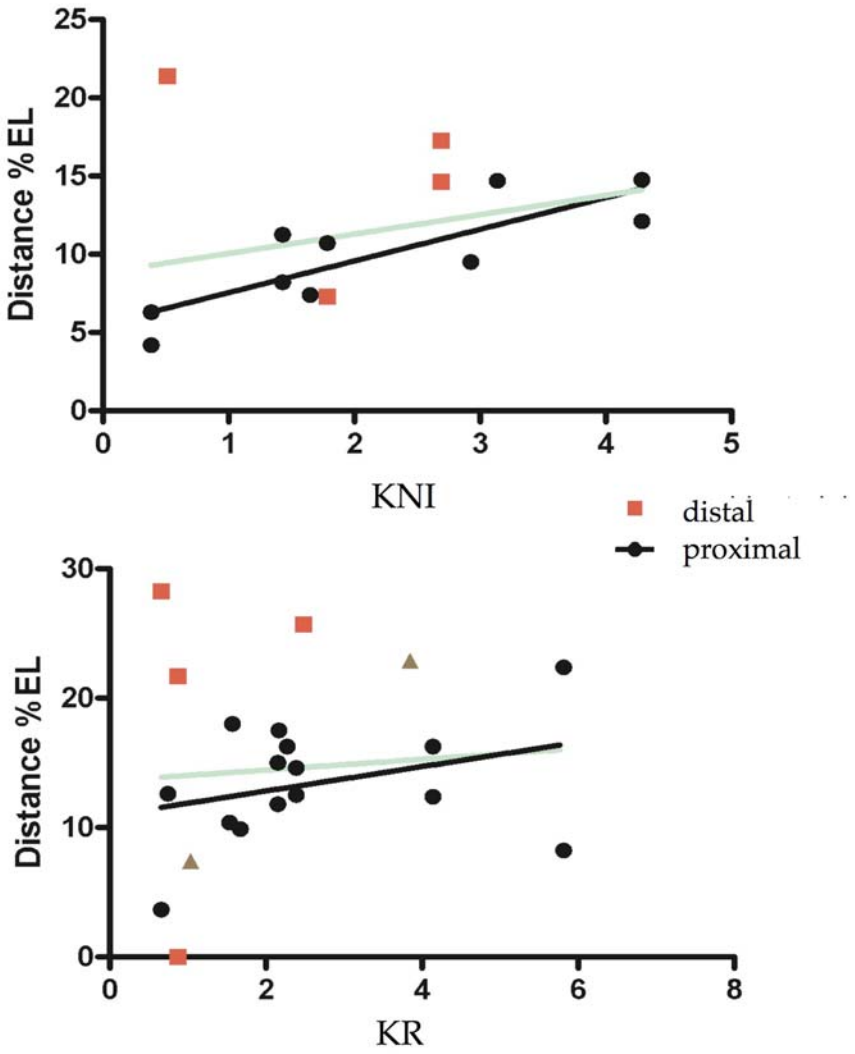


Figure 27

Correlation between positioning and Stubb dictionary values. The Stubb dictionary values for each element were plotted versus the distance between the measured boundary of *cis*-element expression and the center of the relevant gap domain. The black circles are where the gap input under analysis is most proximal to the border of expression, whereas the red squares are cases where another gap input is more proximal. Black lines and grey lines depict the linear fits for the proximal only and full data sets. In the KR graph, the grey triangles represent the *h* 1+5 element split into individual elements. The KR specified borders, are properly positioned when the elements are separated, but the posterior border of stripe 5 is not.

In the case of KNI, there is a strong correlation for all elements where KNI is the primary input (Figure 27, R^2=0.69, $p < 5x10^{-3}$). However, including elements that receive additional more proximal repressive input (KR, GT), results in a loss of significant correlation (R^2=0.11, not significant). The same is true for KR, namely a significant correlation for all elements with KR as the adjacent input (Figure 27, R^2=0.36, $p < 5x10^{-2}$) and no significant correlation for elements that receive additional intervening repressive input (R^2=0.06, not significant). Simple linear fits on binding site data predicted with no free parameters demonstrate there is a basic correspondence between element position and site prediction with minimal parameter fitting.

This quantitative relationship supports the model that gap domains act principally as repressive gradients to position stripes globally along the a-p axis. The KNI correlation is stronger than the KR correlation, which might be due in part to the more symmetrically positioning of the dual stripe elements around the KNI than the KR expression domains. The implication of lesser correlation is that KR-repressed stripes may require more additional input from other gap or pair rule factors to obtain their proper final position. It is notable in this regard that in *h* mutants in Chapter 3, where *run* and *odd* patterns were strongly expanded, *eve* stripes 1, 2, and 5 were much more repressed than stripes 3, 4, 6, and 7. This suggests repressive pair rule inputs into the stripes regulated by KR are stronger than those into the KNI regulated stripes.

The analysis suggests a fairly stereotypic use of gap inputs in organizing the different pair rule stripes. The stripes are generated by distinguishable classes of elements, which correspond in part to pairs of gap genes that are involved in cross repressive interactions. These graded domains are read off at distinct thresholds to generate a unique set of offset stripes across the embryo. The exact combinations of inputs transition as stripes cross the boundaries between adjacent gap domains, with almost all stripes clearly fitting into three classes. The only element that does not fit clearly into one of the three classes is the *run* stripe 4 element. This element is similar to the KR class elements that generate stripe 5 elements, with repression by GT and HB, except lacking KR input as it straddles the border of the KR domain.

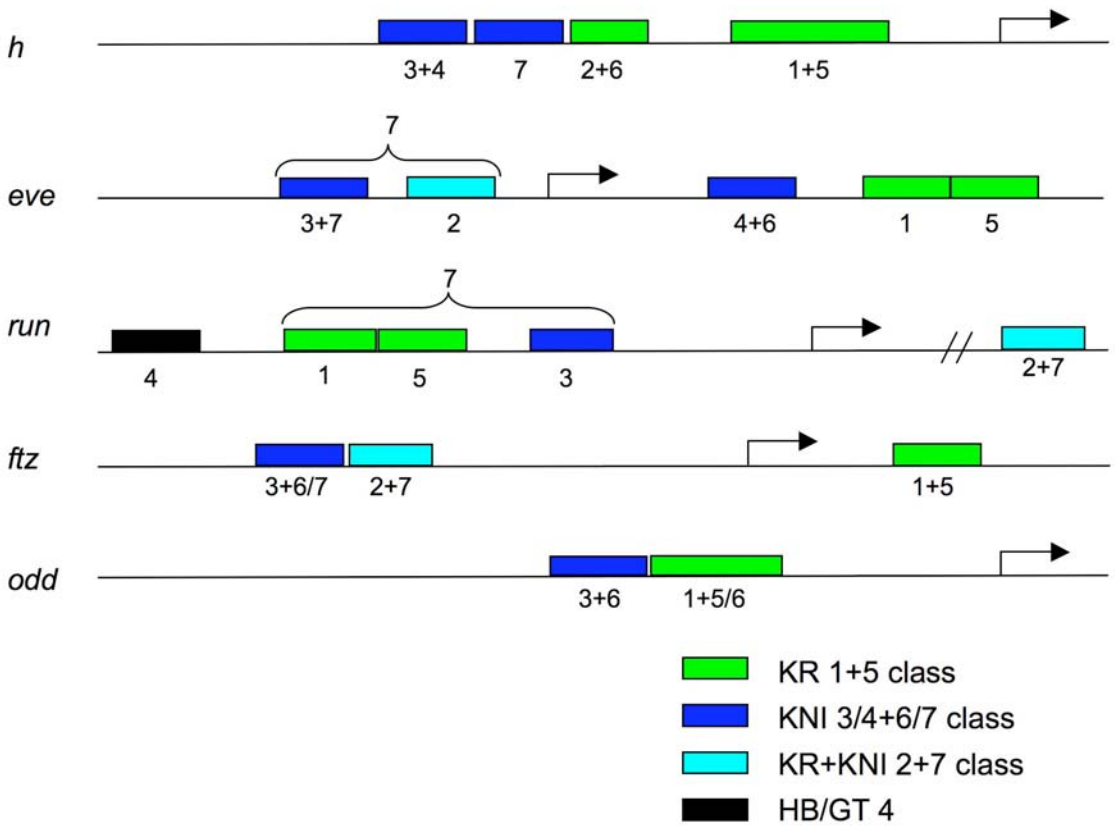


Figure 28

Genomic organization of stripe specific elements. A line depicts the genomic region for each gene. The rectangles represent the stripe specific elements and are colored by the classes seen in the binding site analysis. The stripe specific elements generating stripes 1 and 5 co-occur in all pair rule regulatory regions. The prediction of *Kr* input into all of these elements suggests common regulation may be the reason for co-occurrence. In *eve* and *run* an ancestral 1+5 element could have subfunctionalized into two adjacent elements generating independent stripes. That the elements regulated by KNI also constitute a class of two stripe elements further supports the idea that posterior stripes may have picked up secondary domains of expression during evolution. In *h* and *run*, where the elements generating stripe 3 do not also drive a posterior stripe, the adjacent sequences have also been shown to drive expression in stripe 7. Finally the stripe 2+7 elements constitute a class of elements. The *run* stripe 4 element is the only single stripe element driven at this position and has a unique input composition.

4.3 Genomic organization of *cis*-elements

The genomic organization of the different classes of *cis*-element suggests that elements with similar regulators are more likely to be adjacent (Figure 28). This association is already implicit in the dual stripe elements that established the three classes discussed in the previous section. Most suggestive is that the stripe 1 and 5 elements are adjacent in both *eve* and *run*, while the stripes 1 and 5 are generated by common elements in *h, ftz,* and *odd*. In *h*, the stripe 2+6 and 1+5 elements are also adjacent as are the stripe 3+4 and 7 elements. There is of course the question of whether the relationship is coincidental or meaningful. The probability of the stripe 1 and 5 elements being adjacent in single stripe elements or generated by the same element can be calculated by a simple permutation calculation. The fraction of arrangements where this association is maintained compared to all possible permutations of the stripe labels as currently delineated. This calculation gives a result of 1.8×10^{-4} with the delineations as given and is at least 7.6×10^{-3} under worst case assumptions (Materials and Methods).

Although this arrangement is unlikely, the meaning can be interpreted in multiple ways. The common regulation of both stripes in the dual stripe elements suggests that there is a functional basis for the grouping. The simplest model is a generative one where the co-occurrence of stripes is based on the posterior stripe expression arising from an element that was already expressed in the anterior. At one point in evolution stripe 1 elements existed for *h, eve, run, ftz,* and *odd*, which were delimited at the posterior by *Kr*. During evolution the blastoderm stage has patterned an increasing portion of the a-p axis. In *Tribolium*

castaneum, which is representative of a more ancestral developmental mode, the *Kr* expression domain is at the posterior of the embryo (Cerny et al., 2005; Sommer and Tautz, 1993). Therefore, the KR domain has shifted to increasingly anterior positions within the blastoderm stage on the lineage leading to *Drosophila*. Given the results of the binding site analysis, stripe specific element expression is mainly limited by repressive inputs. Therefore a KR regulated stripe 1 element would be expected to generate a posterior domain of expression as the *Kr* domain shifted away from the posterior of the embryo. In the *eve* and *run* stripes 1 and 5 elements, this suggests that new elements have arisen as a result of a 1+5 element splitting into two independent elements.

As seen in Chapter 3, stripe specific elements also receive significant pair rule input. Therefore the ectopic domains generated by the shift in repressive KR input could be resolved by pair rule cross regulation. It is notable in this regard that in various gap mutants, initial overlap among the pair rule genes is increasingly resolved over time (Klingler and Gergen, 1993). As evolutionary changes would accumulate as a series of smaller less abrupt changes in gap gene expression, the variation in gap expression would accrue more gradually. Therefore the variation that was likely to occur should remain within the bounds normally resolved by pair rule cross regulation. It is notable in this regard that the early maternal and gap directed pair rule patterns are more variable and become increasingly stable over time (Surkova et al., 2008).

Given that dual stripe elements seem more efficient, why would the elements subfunctionalize into two single stripe elements? As discussed in the

previous section, the *eve* and *run* stripe 1 expression fall within the GT domain, whereas GT input sets the posterior border of *eve* and *run* stripe 5. In contrast the 1+5 element of *ftz* can read off GT in a more symmetric fashion. Indeed, the split elements are less symmetric on the KR domain and could not set both borders accurately if the KR gradient was read off the same to position both stripes. Thus, conflicting needs for repressive input may drive the sub-functionalization of dual-stripe into single-stripe elements in order to optimize for gap gene input.

The "unsplit" *h* 1+5 element is an outlier in this respect as the most asymmetrically positioned of all. However, the KR input into the region critical for generating stripes 1 and 5 is indeed separable (Langeland et al., 1994; Pankratz et al., 1990), suggesting that the inputs may be in the process of separating. The element is annotated as a 1+5 stripe element because splitting the elements leads to expansion of the posterior border of stripe 5 (Langeland et al., 1994). Interestingly stripe 1 is strongly repressed by ODD ectopic expression and stripe 5 somewhat less so (Meng et al., 2005; Saulier-Le Drean et al., 1998), suggesting pair rule input may help set the posterior stripe 5 border.

The KNI class of elements is more symmetric around the KNI domain and there is only evidence for one case of such subfunctionalization. The *h* stripe 7 and 3+4 elements are adjacent and are both positioned by KNI. The *h 7* element has much stronger predicted KNI input and is positioned farther from the center of the KNI domain, while *h 3+4* has less predicted KNI input and is positioned much closer to the KNI domain. The maintained set of elements, are the 3+7 and 4+6 elements of *eve* and the 3+6 elements of *ftz* and *odd*. It is notable that they

form a series of nested domains that fall into the correct pair rule register despite being generated by reflection. Therefore, unlike the 1+5 elements, the arrangement generates correct positioning with less need for pair rule input to refine them.

A similar process is suggested in the gap genes themselves by the consistent existence of dual domain elements, which are partially redundant with domain specific elements. It is notable that dual domain elements, with a domain of expression anterior and posterior to *Kr*, exist in all core gap genes - *hb*, *kni*, *gt*, and *tll*. These elements all contain *Kr* sites, except the unfaithfully expressed *kni* +1 element, suggesting shifts in *Kr* could lead to an additional posterior domain of expression as in the stripe 1+5 case. The three adjacent *Kr* elements also suggest a splitting of two domain elements into single domain elements as the central domain only element (CD1) and the anterior domain element (AD2) are separated by a two domain element (CD2 + AD1) (Hoch et al., 1990). Given the relatively smaller number of gap elements compared to the stripe specific elements, their location in the genome is uninformative except in this case.

4.4 **Discussion of binding site analysis**

The analysis of the binding site composition of the stripe specific elements in respect to their expression supports the idea that gap genes act primarily as repressors. The gradual transition along the a-p axis in which sites are not predicted for the co-expressed gap gene suggests a relatively simple spatial correspondence between input and output. The correlation seen between the

position of the stripe specific element borders and the center of the most proximal predicted repressor supports the model of gap genes acting as repressive morphogen gradients. However, the precise quantitative correspondence between input and output has not been uncovered despite a large body of work on the subject (Arnosti and Kulkarni, 2005).

The expanded set of stripe specific elements is an ideal data set for studying the relationship between sequence and pattern. The analysis here involves essentially no fitting of parameters emphasizes the importance of good binding site data. The results support the model from previous work on the *eve* stripe specific elements (Arnosti et al., 1996; Clyde et al., 2003; Small et al., 1992, 1996) and captures important basic features of element function. This consistency both supports the simple analysis and provides a basic framework for understanding the regulation of these elements across the whole set. The results were in part a result of additional data that allowed generation of improved PWMs. The large scale determination of PWMs for cohesive sets of transcription factors is a boon to research of this kind (Berger et al., 2008; Meng et al., 2005; Noyes et al., 2008a; Noyes et al., 2008b).

The data here suggest a relatively straightforward model for the generation of new *cis*-elements through an intermediate where an element directs two "homologous" domains of expression under the same set of regulatory interactions. This reuse of regulatory interactions makes much clearer how new positions can be generated within an embryo. Rather than requiring the generation of completely new sets of interactions for a whole host of factors,

existing positions are in effect generated by partial reuse of pre-existing interactions. However, it is clear from the case of both the gap *cis*-elements and the stripe 1/5 elements, that the pre-existing interactions are not used in precisely the same way. The polarity that is established in the ordering of the stripe 1 and 5 elements is maintained indicating that other reinforcing inputs used differentially in the two regions allow truly new positions to be generated.

The independence of stripe specific elements is at the heart of how the pair rule genes establish periodic patterns from the gap inputs. The broad gap gene domains cover multiple pair rule stripes, such that repression of pair rule genes at the locus level would disallow the formation of some pair rule stripes. The fact that most gap genes have been shown to act as short range repressors, which are act over a few hundred base pairs, helps explain the modularity of stripe specific elements (Courey and Jia, 2001; Gray and Levine, 1996). Through compensatory gain and loss of gap sites within a *cis*-element, dual stripe elements can partition the inputs important for each stripe into distinct increasingly independent units. Although the assumption in most work analyzing binding site content is that the changes are neutral, it is possible that selection favors subdivision of elements. The alleviation of constraints for generating two stripes with the same interactions would clearly improve flexibility in relative shifts between gap domains. The more independent arrangement would thereby have an entropic advantage as there are presumably many more ways to specify the stripes independently by different sets of

interactions than in pairs where both stripes are constrained to be regulated similarly.

The fact that the gap domains are organized in offset pairs that direct sets of pair rule stripes indicates that coherent changes among the patterns would shift expression together. However, it is notable that the cross repression between the pairs is not symmetric and is stabilized by other inputs in each case. This would allow repositioning of the domains independently. The fact that gap domains in the anterior and posterior are generated by optimized interactions allow each domain to be shifted independently, while the dual domain elements are also suggestive of an evolutionary origin for these domains.

Chapter 5: Discussion

The work presented here started with the validation of Ahab for computationally dissecting the *cis*-regulation of *Drosophila melanogaster* segmentation genes. This validation significantly increased the number of known segmentation elements and helped provide initial rules concerning the relationship between binding site content and expression pattern across a large number of elements. The stripe specific elements found in *odd* through these dissections indicated a clear inconsistency with its classification as a secondary pair rule gene. In general, *cis*-dissections are part of a reductionist approach to break down the complex regulation of segmentation into individual components that can be more clearly understood. This approach is quite useful in the case of the pair rule genes due to the complexity of their regulation.

The original *cis*-dissections of *eve* and *h* demonstrated that the periodic patterns are established by modular transcriptional control regions (Goto et al., 1989; Harding et al., 1989; Howard and Struhl, 1990; Pankratz et al., 1990). However, the molecular epistasis experiments that were the basis of the pair rule classification did not break down the process into these more readily understandable components (Carroll and Scott, 1986; Carroll and Vavra, 1989; Howard and Ingham, 1986; Ingham and Gergen, 1988; Ingham, 1988). Therefore the complex relationships that involved in establishing pattern were less clear. The attempt described here to complete the dissection of the *cis*-regulatory inputs into the pair rule genes is thereby a natural continuation of the process started in the eighties, but left unfinished. By focusing on the clear discrete criteria of *cis*-

elements in relation to the process of pattern formation the hierarchy was now modified. As *ftz* and *odd* clearly decode the non-periodic patterns of the maternal and gap genes, they establish the seven striped patterns. Although their role in regulating other pair rule genes appears more minor than those of *h, eve,* and *run*, they are critical components of the periodic pattern that is generated and integrate unique positional information from the maternal and gap system as the patterns form. Further, the work here and previous work using ectopic expression (Andrioli et al., 2004; Saulier-Le Drean et al., 1998) indicate a clear role for *odd*, *prd,* and *slp* in regulating the early patterns of even the classic set of primary pair genes. Therefore *odd* and *ftz* should be included in the set of primary pair rule genes. *odd* and *ftz* thereby fill in the fourth position in a tiled set of overlapping patterns that generate a unique nucleus by nucleus code with a two segment repeat.

Based on the *cis*-dissections (Chapter 3; Klingler et al., 1996) and the timing of stripe expression, *run* is likely to have a complete repertoire of stripe specific elements like *h* and *eve*. Unlike *h* and *eve*, *run* also has a seven-stripe element that is active while the seven stripe patterns are generated. This element appears to integrate both maternal and gap inputs in addition to that from the pair rule genes. By characterization of the timing of seven-stripe elements in *run, ftz,* and *odd* it is evident that they contribute to the establishment of the periodic pattern. Further, in *ftz* and *odd* it is likely that stripe specific and seven-stripe elements interact during the establishment of periodic pattern to generate the full endogenous pattern. Therefore how the seven-stripe elements and the stripe specific elements interact is a question that extends past the reductionist

approach presented here towards understanding how the pieces work together to generate the endogenous patterns.

The interaction of different *cis*-elements has not been systematically studied. Some information is present in rescue experiments that were conducted for a number of the loci. In *eve*, rescue experiments including different portions of the upstream region that contains the seven-stripe element as well as the stripe specific elements generating stripes 2, 3, and 7 were carried out (Fujioka et al., 1995). Three rescue constructs were generated, one containing all the elements, one containing only the stripe specific elements, and one containing only the seven stripe element. The element only containing the seven-stripe element could note rescue at all whereas, the other two constructs generated fairly similar rescue. Interestingly, the stripe specific elements drove weak expression in the region normally generating stripes 5 and 6, which resolved into a relatively proper stripe 5 and a weak stripe 6 in the presence of the seven-stripe element. Therefore, even in *eve*, where the pattern resolves prior to the seven-stripe element becoming active, the presence of the seven-stripe element helps reinforce, resolve, and refine improper expression.

In *run*, the region containing the seven-stripe element alone can generate significant rescue with defects most notable in the first abdominal denticle belt (Butler et al., 1992). It is notable that stripe 3 is never expressed strongly in the rescued animals and corresponds to the region of the embryo that generates the defective denticle belt (Butler et al., 1992). A similar mixed role for specification is seen in *ftz*, which also has an early acting seven-stripe element. The original *ftz*

rescue experiments never included the stripe 3+6 element and show defects in segments A1 and A7, which correspond to this expression (Hiromi et al., 1985). Therefore, these genes show a more shared dependence on the stripe specific and seven-stripe elements than *eve*. Although in *run* and *ftz*, the correspondence between these defects and the *cis*-elements was not apparent at the time, in retrospect this points towards a combined role for the elements that matches up with their role in generating the early pattern. In *ftz*, the fact that defects exist point to a functional requirement for the stripe specific elements and therefore a clear role in establishing the periodic patterns. However, the important role of the seven-stripe element is indicated as well.

Finally the organizational relationship between the gap and pair rule tiers was analyzed by predicting the binding site composition of the maternal and gap regulated elements of these two classes of gene. Predicted gap input was anti-correlated to the pair rule patterns suggesting stripe borders were set primarily by repression. This was supported by the correlation between the position of the border of stripe expression and the strength of predicted input for KNI and KR. Elements were further classified into three groups based on input, KR regulated, KNI regulated, and KR+KNI regulated. The KNI and KR groups, which constitute the bulk of elements, integrate inputs from the two offset pairs KNI+HB and KR+GT. Therefore there is a correspondence between the organization of gap cross regulation and pair rule stripe formation.

Interestingly, the early expressed pair rule patterns themselves consist of two offset pairs. In each case one member of the pair appears to provide somewhat

stronger input into the other to establish the pattern, whereas the second appears to have a more minor role in refinement. In the gap genes, both *hb* and *Kr* seem to have important roles in regulating genes in the other pair, whereas *gt* and *kni* do not. In the pair rule genes, there are more interconnections with *h, eve,* and *run* all helping to tie the two pairs together to maintain correct offsets. By coupling two offset patterns a large number of positions of downstream target genes can be specified by reading off combinations of the set of genes. By having one gene dominate the cross repression, there is a primarily directional flow of information that gives hierarchy to the interactions.

The different classes of stripe specific element each have members that generated two stripes - one in the anterior of the embryo and one in the posterior. The organization of these elements in the genome revealed a clustering of inputs and suggests that some single stripe elements may have evolved from elements capable of generating two stripes. This result suggests a model of evolution in which the new positions encoded in the blastoderm fate map arose by reuse of existing regulatory interactions to generate posterior positions. Whether or not this model is correct, determining a larger set of stripe specific elements generates a near optimal dataset for understanding evolution of *cis*-elements across a similarly regulated set of genes. The modular aspect of the pair rule transcriptional regulatory regions and the large number of domains to their pattern give more data points on how *cis*-elements evolve in a sub-network than any other well studied example.

The work is based on a growing area of research into computational techniques for prediction of *cis*-regulatory elements. There have been few studies to systematically compare the variety of techniques that have been developed to determine which performs best. A recent study attempted to compare techniques over a large set of *Drosophila cis*-elements that function in a wide range of developmental contexts (Li et al., 2007). The analysis presented indicated that the *cis*-elements of the segmentation system had a much higher level of binding site clustering than typical of the whole set. Interestingly, despite the very strong performance of Ahab and Stubb in predicting elements, they showed it was less bias than CIS-ANALYST, PFR-Searcher, and Fly Enhancer towards the high level of clustering specific to segmentation (Berman et al., 2002; Grad et al., 2004; Li et al., 2007; Markstein et al., 2002). Therefore, the performance of algorithms like Ahab and Stubb seem more generalizeable than approaches strongly tied to clustering. New methods continue to be developed with different approaches, such as more explicit physical models of TF binding (Hallikas et al., 2006; Palin et al., 2006) and HMMs with the ability to learn complex grammars (Won et al., 2008), but Stubb is generally considered one of the more sophisticated algorithms (Aerts et al., 2007; Hallikas et al., 2006; Palin et al., 2006; Won et al., 2008).

One area that has shown more recent growth is Chromatin Immunoprecipitation followed by microarray hybridization (ChIP-chip) or ChIP sequencing (ChIP-Seq), both of which determine the location of transcription factor binding at a genome wide level (Barski et al., 2007; Lee et al., 2002; Ren et al., 2000). One particularly nice study in *Drosophila* analyzing *twist* targets

combined loss of function data and a temporal time course of Chip-chip data, with discovering 12 newly identified *cis*-elements and proposing a dynamic genome wide map of *twist* function. Whether other *cis*-elements were tested was not reported, but this method was clearly quite effective in finding elements, by combining multiple genome wide data types. It is very clear that such technologies will help push analysis of transcriptional networks forward.

An analysis of segmentation by ChIP-chip has also been carried out for a number of the same factors used in this study (Li et al., 2008). The study did not attempt to determine new *cis*-elements, so the success rate cannot be directly compared. One advantage of these techniques is that they provide a measure of the actual *in vivo* occupancy of regions of the genome by the transcription factors of interest. However, they are more expensive and require specific antibodies to avoid non-specific binding. Interestingly, there seems to be consistent detection of transcription factor binding to genomic regions in the absence of clear occurrence of their binding motifs, which was also seen in the segmentation study (Li et al., 2008). Therefore, these techniques do not necessarily provide a direct measure of binding site occurrence, but do indicate the possibility that there is some recruitment in the absence of DNA binding. This is an important issue for models that assume recruitment of transcription factors is based solely on protein-DNA interactions. Instead some protein-protein interactions may be important as well. However, this feature can be explicitly included if the protein-protein interactions are known.

More than a third of the segmentation *cis*-elements analyzed in the ChIP-chip study were detected to have binding of all the factors tested, which include BCD, CAD, HB, KR, KNI, and GT. Another third only lacked KNI, but were bound by all the other factors, and only 6 of 43 known elements were bound by fewer than 5 of the 6 factors. It is not clear whether the systemic difference in binding between KNI and the other factors examined in the study is meaningful. However, the ChIP-chip work neither ranked the strength of binding to different elements nor was able to suggest any composition rules. As ChIP-chip predicts regions of binding, rather than sites, it is hard to directly compare results, but both methods clearly have their own strengths. The *in vivo* nature of ChIP-chip gives an experimental measure of recruitment, whereas approaches based on binding site predictions are more closely tied in to how position is encoded in the sequence of the *cis*-elements.

Optimally a model that can predict the expression pattern of a gene based solely on genomic sequence, the set of transcription factors present, and the regulatory history of the locus is desirable. Much effort has gone into modeling in a number of systems, but the *Drosophila* work is the most relevant and comparable to that presented here, so it will be the focus of discussion. Work in this area has been promoted by the determination of increasing numbers of *cis*-regulatory elements in this work and that of others (Markstein et al., 2002; Markstein et al., 2004; Schroeder et al., 2004; Segal et al., 2008; Senger et al., 2004).

In the *Drosophila* d-v axis, grammar rules have been proposed where composition, spacing, and orientation of sites have been proposed to play a role

in function (Erives and Levine, 2004; Markstein et al., 2004). A relatively simple model based on occupancy explains the expression of three elements active in one domain of the neurogenic ectoderm based on the three regulators DORSAL, TWIST, and SNAIL (Zinzen et al., 2006). The study needed fit cooperativity parameters to accurately reproduce the patterns. As the expression of only three elements needed to be fit and no evidence for the lack of over-fitting was provided, it is unclear how accurately the model reflects the true details of regulation, but the results certainly point towards issues for further study. The study did offer some comparison to different *Drosophilids* to argue that the cooperativity matched generally with differences in spacing, but the spacing of sites was not explicitly modeled. Later work on the orthologous enhancers in different *Drosophilids* demonstrated that small insertions and deletions, which altered the spacing between sites, were causal in changes in expression by making similar mutations in the *melanogaster* elements (Crocker et al., 2008). However, this study did not model expression of the elements.

In the a-p axis, there have been two primary attempts to model *cis*-elements. The first only modeled the *eve* stripe 2 element and built a model that fit the expression of the element quite well and also reproduced previously published data on site directed mutagenesis and one mutant background that were not used in the fitting (Janssens et al., 2006). This model was a nice step forward in understanding the expression pattern of elements, however, it was based on only a single element and might not generalize. The second study attempted to model 44 elements with the same set of parameters for all factors acting in all elements (Segal et al., 2008). This study also demonstrated a clear ability to fit the data

well, but required fitting of the PWMs to achieve good success. As the fitting of PWMs might raise concern of overfitting, held out data and tenfold cross validation were used to demonstrate that the results were statistically significant. Therefore, this study indicated that the model used in that analysis generalized over the entire set of elements.

The work presented here on the binding site predictions within the gap and pair rule *cis*-elements does not aim at explicit modeling, but rather understanding the relationship between components in a network. Although modeling is an important end goal, features of network organization can be analyzed based solely correspondence between input and output without fitting PWMs or other parameters. A simple analysis that highlights the correspondence between binding site strength and position of expression avoids makes clear the important features without depending on fitting of parameters that are hard to demonstrate or verify. Like both the previous a-p studies, it is explicitly clear in the work presented here that modeling of the HB switch from repression to activation is critical for maintaining a correspondence between the predicted input and the output of the elements (Janssens et al., 2006; Segal et al., 2008). By focusing in on a cohesive set of elements important for a specific developmental transition aspects of network organization are apparent that were not seen in previous studies. Therefore, by focusing on organizational features rather than the mechanism of transcription, interesting features in *cis*-element organization became clear.

Many studies now focus on large scale genome wide analyses of transcriptional cross regulation. In contrast there has been a lot of past work in segmentation analyzing each gene independently, where detailed information on how single genes were regulated was explored. One major goal of this work has been to explore an interesting intermediate position to understand a distinct transcriptional sub-network. This approach has been fairly fruitful, clarifying past conflicts in the data regarding the pair rule hierarchy and demonstrated unrecognized aspects of how their regulation is organized.

Chapter 6: Outlook

Although many open questions were answered, many new questions have been opened up. How do stripe specific and seven stripe elements interact? What is the relationship between the gap and pair rule genes in later development? Is there a tendency for dual stripe elements to separate out into single stripe elements? How do segmentation genes switch from activation to repression?

There have been no true quantitative studies of expression driven from different constructs, in part because insertional effects could not be controlled easily. New targeted transgenic approaches based on Cre and the phage PhiC allow repeated targeting of the same site in the genome with good efficiency (Fish et al., 2007; Groth and Calos, 2004; Oberstein et al., 2005). By looking at both mutagenized *cis*-elements and constructs containing different sets of the delineated elements, the interaction between sites and elements could be studied both systematically and quantitatively. In addition, there has been relatively little work looking at how differences in *cis*-element composition play out functionally when used in rescue scenarios. By placing a targeting site on the same chromosome as mutant alleles one could also look at how variants of the *cis*-element repertoire effect function of the network in a fairly efficient fashion.

With the large number of *cis*-elements available a number of distinct questions can be addressed systematically using the targeted transgenesis. One straight forward feature to test is gap sites that target their own *cis*-elements by

site directed mutagenesis. How these sites modulate the activity of the enhancers is a nice direct test that could help clarify whether these genes auto-activate, repress, or have a more complex effect on their own pattern. Given the likelihood that the stripe specific elements and seven-stripe elements interact in *run*, *ftz,* and *odd*, the expression driven by combinations of these different types of elements together and separately could be compared. Additionally *cis*-element combination could be combined with mutagenesis of specific sites to better understand how the interactions depended on specific inputs. In addition to quantitatively comparing expression of such combinations in wild type flies, they could be used to drive rescue constructs to see what role they play in the dynamics of cross regulation. Finally, it would be of interest to look at the role of the gap and pair rule genes in other contexts. It would be possible to design rescue constructs that only rescue the initial blastoderm function, but lack other control elements that function later. As both the gap and pair rule genes function in the nervous system and other contexts, it is of interest to see what aspects of their regulation are similar and different in later contexts.

It is more difficult to test whether the model of stripe specific element evolution, but with the rapidly expanding number of sequenced genomes it is possible. Analysis of the 12 sequenced *Drosophila* genomes indicates that gross details of *cis*-element organization in the gap and pair-rule genes are conserved in these species. One way to demonstrate that the stripe 1 and 5 elements have subfunctionalized is to find species where the order of the elements is reversed, due to the binding sites separating out in opposite directions within the genomic DNA. It appears that genomic rearrangements in the regulatory regions of the

pair rule genes are uncommon as the order of the stripe specific elements has been maintained over the 100 million years of evolution between *Sepsidae* and *Drosophila melanogaster* (Peterson et al., 2009). Alignments within even the *cis*-regulatory regions are possible over this large span, so ruling out inversions would be possible. Alternatively the idea could be tested by finding direct examples of transitions between dual and separable stripe specific elements in at least three species, where the inferred ancestral state is a dual stripe element. However, if the separation of elements is driven for similar stripes, it is unlikely to necessarily find dual stripe elements that have not subfunctionalized.

One particularly interesting issue is the switching between activation and repression seen in genes like HB (Simpson-Brose et al., 1994; Zuo et al., 1991). It is interesting that BCD and HB together activate targets throughout a larger domain than BCD acting alone (Simpson-Brose et al., 1994). Since HB alone represses (Figure 29), activation by the pair must be extended in part by recruitment of BCD at concentrations where BCD alone cannot bind. This suggests that HB might help recruit BCD to the DNA through a protein-protein interaction. The extra binding energy of the protein-protein interaction might help allow recruitment when BCD concentrations are limiting. The physical interaction of BCD and HB would also generate new binding surfaces that could recruit different co-activators or co-repressors than recruited when HB binds alone. The nicest aspect of this hypothesis is that protein-protein interactions between BCD and HB alone and in the presence of DNA can be tested easily. If this does uncover important interactions similar experiments could be

systematically carried out for sets of segmentation genes to determine whether other instances of this phenomena occur.

Together these experiments would help to take the set of components in this complex process and begin to address how *cis*-elements and binding sites interact to more systematically reconstruct the complex interactions that occur in early segmentation. Although segmentation is not the premiere developmental system it was in the 1980s, it is still one of the best understood transcriptional paradigms and has many interesting questions left open.

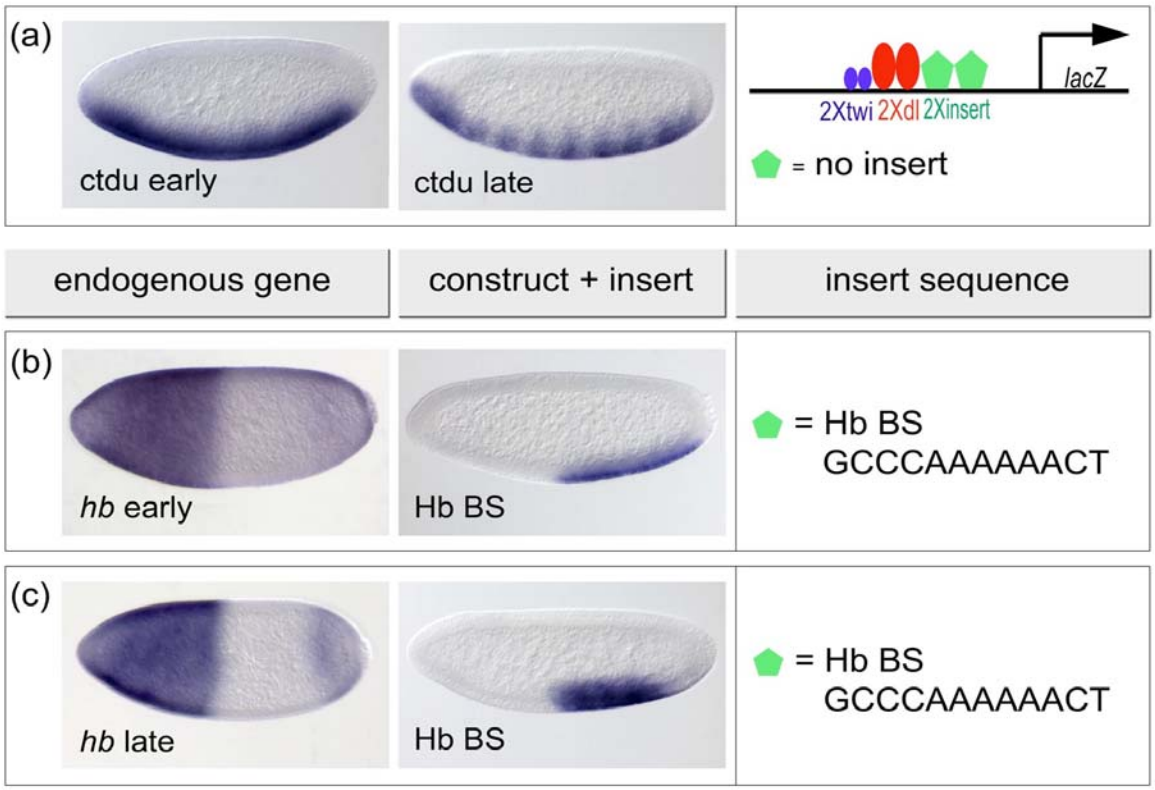


Figure 29

HB sites repress in the context of an artificial d-v enhancer. A pair of HB binding sites inserted into the artificial 2xTWI2xDL enhancer converts the ventral stripe of expression into a ventral stripe restricted to expression outside the *hb* domain of expression. (a) The ctdu construct alone drives a ventral stripe both early (left) and late (middle). Schematic of the ctdu construct (right). (b) The early *hb* expression (left) is confined to an anterior domain that matches the region where repression is seen at this stage in the ctdu enhancer containing a pair of HB sites (middle). The HB binding site inserted in the element (right). (c) same as (b), but shown at a later stage. Repression in the posterior domain is apparent as well.

Materials and Methods

Analysis with Stubb

For *cis*-element prediction runs were done as in (Schroeder et al., 2004), except the Stubb program (Sinha et al., 2003) was used instead of Ahab (Rajewsky et al., 2002). The Stubb program computes a free energy score for a given sequence that provides a measure of the density and strength of binding sites in the sequence. The *free energy* score is the log of the ratio of the probability of the sequence with a set of PWMs and a background model to the probability of the sequence derived solely from a Markov Model generated from genomic sequence. The probability of the sequence when including the PWMs requires fitting the probability of site occurrence for each PWM and the background. In Stubb these probabilities are fit by an iterative method where an estimate of the probability is generated from the current set of probabilities through an expectation maximization approach. The *free energy profile* is the result of moving a sliding window of 500 bp along the sequence in 50 bp increments and plotting the free energy score for each window. For finding *cis*-elements, Stubb works best when given binding preferences (PWMs) for multiple transcription factors and then determines clustering of sites as statistically significant by comparing against a background model derived from local genomic sequence (Sinha et al., 2004).

For predicting binding site composition, Stubb was run differently than for *cis*-element prediction. The use of multiple matrices, a local background

model, and fitting of site occurrence probabilities leads to distortion of the binding site score for the same sequence in different *cis*-elements. To alleviate this problem Stubb was run individually for each matrix with the same fixed priors of 0.0005 (0.9995 for the background) using a zero order Markov Model background model generated from all segmentation genes. Instead of running over a window of 500 bp, the element as delimited experimentally was run as a single piece.

To generate the KNI and KR correlation plots, the binding site predictions (dictionary scores) were plotted against the distance between the center of the KNI expression domain and the most adjacent border of the expression generated by the element. The value for the center of the KNI domain was taken from the FRDWT 10 % strip 14A 4 time class of (Myasnikova et al., 2001) and the cis-element predictions were taken from our measurements.

Generation of new matrices

Our original KNI, GT, and TLL PWMs were not of good quality. Matrices from bacterial one hybrid (B1H) screens (Noyes et al., 2008b) were used to revisit the DNAse footprinted sites for these factors. There were two types of problems with the matrices: it was always clear that we lacked an adequate number of sites for GT, whereas in the case of KNI and TLL it became clear during this effort that the sites were originally misaligned yielding a matrix that was "out of focus." The footprinted sequences from the literature were aligned with the B1H matrices and then a composite PWM consisting of the aligned sequences from

the footprints as well as the B1H data were generated. In KNI and TLL, these matrices performed well and were used.

In the case of GT the PWM from the one hybrid data was a perfect palindrome and therefore did not choose a strand when aligning the sequences. Second, given the small number of GT footprints, the combined PWM was still overly specific and did not predict any GT sites in the *eve* stripe 2 element, despite this being the best characterized GT target. For these reasons the GT PWM used here was generated only from the footprinted sites, but they were included in both orientations as suggested by the B1H data. Given that our original PWM was also clearly palindromic in nature, this is a reasonable approach to double the amount of data for estimating the binding preferences at each position. There are few known sites for HKB and the published B1H matrix was used without modification.

Likelihood of *cis*-elements co-occurrence

The calculation of stripes 1 and 5 co-occurring by chance was based on calculating the fraction permutations of the stripe labels over the *cis*-elements that result in co-occurrence of stripes 1 and 5. Co-occurrence was defined as any case where a two stripe element was assigned both 1 and 5 or two adjacent single stripe elements were assigned 1 and 5. In case the delineations are incorrect a "worst case" scenario was also calculated to get a lower bound on the probability. The *h* 3+4 element, which is split by pair rule input and does not fit cleanly into separable stripes established by the gap genes was treated as

generating only one stripe. The encoding of *h* stripe 2, although assigned primarily to the stripe 6 region (Howard and Struhl, 1990), was not unambiguously mapped in all dissections and appeared more delocalized (Lardelli and Ish-Horowicz, 1993; Pankratz et al., 1990; Riddihough and Ish-Horowicz, 1991). Therefore *h* stripe 2 was also left out of this minimal set. Finally, the rest of the stripes were all assumed to be separable, even though there is good evidence that some are not (Fujioka et al., 1999; Langeland et al., 1994). These worst case assumptions still result in a probability of 7.6×10^{-3}, which is relatively unlikely to happen by chance.

Analysis of expression patterns

RNA *in situ* hybridization using digoxigenin labeled RNA probes and alkaline phosphatase detection were carried out as before (Schroeder et al., 2004). Briefly, antisense RNA probes were generated by in vitro transcription and cleaned up using the Qiagen RNeasy kit. The yield from this procedure was calculated using UV spectrophotometry and *in situ* hybridizations were carried out with 2ng/μL probe in hybridization buffer A. Hybridization was carried out overnight and followed by a day of 8 hour long washes in hybridization buffer B. After a series of washes in PBT, the probe was then detected with anti-DIG Alkaline Phosphatase for an hour, and washed further with PBT. The colorimetric reaction was carried out with 4.5 μL of NBT solution and 3.5 BCIP solution in 1 mL of staining buffer. Typically the reaction took 5-15 minutes although in a few cases times of up to 1.5 hours were required. For the *in situ* analysis of the mutants, timed stainings were used to try to get uniform results for all genotypes. The development times used were 4:30 for *eve*, 10:00 for *h*, 4:15

for *ftz*, 5:30 for *run*, 6:00 for *odd*, 10:00 for *prd* and *slp*. The NBT, BCIP, and anti-DIG Alkaline Phosphatase antibody are all available from Roche Applied Science. Staining buffer consists of 0.1M Tris pH 9.5, 50 mM MgCl2, 1M NaCl, and 0.1% Tween20.

Embryo treatment prior to hybridization was carried out according to standard procedures. Timed embryo collections were dechorianated in 50% bleach for 2 minutes and then washed with double distilled water. They were then fixed in a 1:1 mixture of 5% paraformaldehyde (PFA) in poly buffered saline (PBS) and heptane for 25 minutes. The embryos were then devitillenized by shaking for one minute in a 1:1 mixture of heptane and methanol followed by washing and storage in methanol. Prior to hybridization they were further subjected to 20 minute post fixation in 5% PFA in PBS + 0.1% Tween20 (PBT), 5 minute treatment with 5 ng/μL proteinase K, followed by a second 20 minute post fixation. The embryos were then prehybridized for 1-2 hours at 65 C with hybridization solution A. Hybridization solution B is 50% formamide, 5X SSC, 0.1% Tween 20, in ddH_2O. Hybridization solution A is the same as solution B supplemented with 100 μg/mL salmon sperm DNA and 50 μg/mL heparin.

To measure the patterns for *cis*-elements a series of measurements were made using Zeiss Axiovision software. A straight line through the points at the embryo perimeter defined each stripe boundary where expression dropped to roughly 50%. A second line was drawn through the central longest axis of the embryo. The distance from the posterior of the embryo to each stripe boundary

was measured as well as the total length of the embryo. From this the %EL position of each boundary was calculated. Three embryos were measured at roughly phase 3 and averaged. For the literature mapped *cis*-elements, the most appropriately staged embryo shown for the construct in the initial publication was used.

The staging of blastoderm embryos followed (Lecuit and Wieschaus, 2000). In phase 1 the nuclei have not yet begun to elongate; in phase 2 nuclear elongation occurs; in phase 3 the plasma membrane begins to invaginate between the nuclei until it reaches the base of the nucleus; in phase 4 the plasma membrane extends to roughly 35 μm and cellularization completes. For the purpose of having a clearly defined point for each stage, the end points were used. Phase 1 was taken as the point where nuclei show the first sign of elongation. Phase 2 was taken as the point where the nuclei have completed elongation. Phase three was when the plasma membrane reached the base of the nucleus. Phase 4 was taken as when the plasma membrane was invaginated roughly 35 μm.

Cloning

Sequences of the primers are available in Appendix A. All elements were cloned into TOPO (Invitrogen, Carlsbad CA), sequence verified, and then subcloned into a P-element transformation vector. In all cases except for the seven-stripe elements of *h* and *odd*, the P-element vector was Casper hs43GAL (Thummel and Pirrotta 1991). When testing for 7-stripe elements in *h* and *odd*,

fragments including the endogenous basal promoter were fused into a LacZ reporter in Casper 4.

The megaprimer method (Barik, 1996) was used for the site directed mutagenesis. Roughly 100 bp megaprimers were generated by PCR with mutagenic primer pairs at one or either end. The megaprimers were used at a concentration of 15 pmol/μL to amplify larger regions off a plasmid at 5 ng/μL by otherwise standard PCR. The megaprimers are double stranded, but particularly when amplifying regions off of a plasmid, there were no problems with nonspecific bands. The larger regions generated by the megaprimer method were roughly 1kb and 500 bp overlapping PCR fragments containing the necessary mutations. These fragments were then fused using weave PCR with the original flanking primers. The final sequence was confirmed by sequencing prior to subcloning into the transformation vector.

Fly strains and molecular characterization

The following mutant fly strains were used *eve*^{3}, *h*^{25}, *run*^{3}, *Df(3R)SCB* which removes *ftz*, *ftz*^{11}, *ftz*^{9H34}, *odd*^{7L}, *drm*^{p2} which removes *odd* (Green et al., 2002), *prd*^{4}, and *Df(2L)ed1* for *slp1* and *slp2*. To determine the molecular lesion associated with *ftz*^{9H34} a region spanning the *ftz* locus was cloned by single fly PCR from heterozygotes into TOPO. Twelve clones were sequenced and two alleles were detected. One allele was also present in clones from a control *h* mutant strain containing the same balancer, whereas a second containing a glycine to stop mutation at position 53, was unique to the *ftz*^{9H34} strain.

Appendices

PWMs used in this study

The PWMs used in the original Ahab *cis*-element screen. The counts for each nucleotide at each position within the binding site are listed.

Bcd				
position	A	C	G	T
1	2	10	10	8
2	5	12	4	9
3	3	0	0	27
4	27	0	3	0
5	30	0	0	0
6	0	0	7	23
7	1	29	0	0
8	3	13	3	11
9	4	15	11	0
10	2	9	13	6
11	6	14	6	4

Cad				
position	A	C	G	T
1	0	12	4	5
2	12	3	5	1
3	6	1	8	6
4	4	6	1	10
5	0	0	1	20
6	1	1	1	18
7	2	1	0	18
8	0	0	1	20
9	5	0	0	16
10	8	3	0	10
11	8	1	11	1
12	6	0	12	3
13	4	9	3	5
14	4	9	0	8

TorRE				
position	A	C	G	T
1	0	0	0	6
2	0	1	4	1
3	0	5	1	0
4	0	0	0	6
5	0	5	0	1
6	6	0	0	0
7	6	0	0	0
8	0	0	1	5
9	0	0	6	0
10	6	0	0	0
11	6	0	0	0

Hb				
position	A	C	G	T
1	17	12	14	0
2	0	0	4	39
3	0	1	0	42
4	0	1	0	42
5	0	0	0	43
6	1	0	0	42
7	11	3	5	24
8	7	8	8	20
9	3	6	25	9
10	11	11	7	14
11	5	9	10	19
12	8	18	4	13

Kr				
position	A	C	G	T
1	10	3	7	0
2	19	1	0	0
3	17	2	0	1
4	12	6	1	1
5	0	0	20	0
6	4	0	16	0
7	3	2	15	0
8	0	0	2	18
9	0	2	0	18

Kni				
position	A	C	G	T
1	5	0	7	15
2	11	1	8	7
3	11	3	12	1
4	9	0	13	5
5	18	5	4	0
6	18	2	0	7
7	11	14	1	1
8	4	13	7	3
9	2	3	8	14
10	14	0	9	4
11	17	0	10	0
12	13	12	1	1
13	10	0	10	7
14	7	11	5	4
15	19	1	5	2

Gt				
position	A	C	G	T
1	1	0	0	5
2	3	0	0	3
3	0	0	0	6
4	0	0	3	3
5	6	0	0	0
6	0	6	0	0
7	0	0	4	2
8	1	1	0	4
9	5	1	0	0
10	4	1	0	1
11	2	1	1	2
12	6	0	0	0

Tll				
position	A	C	G	T
1	12	8	0	0
2	1	2	2	15
3	1	2	1	16
4	5	1	0	14
5	2	3	15	0
6	11	1	5	3
7	1	17	0	2
8	0	2	1	17
9	0	3	2	15

Additional PWMs from the bacterial one hybrid screens as well as the new KNI, TLL, and GT matrices generated using those PWMs. The three papers the

matrices were drawn from are (Meng et al., 2005; Noyes et al., 2008a; Noyes et al., 2008b).

Eve_Noyes_Cell_2008				
position	A	C	G	T
1	2	1	4	7
2	6	1	9	6
3	3	10	0	9
4	1	0	0	21
5	22	0	0	0
6	22	0	0	0
7	0	2	1	19
8	0	2	11	9
9	17	0	4	1
10	4	4	4	1

Ftz_Noyes_Cell_2008				
position	A	C	G	T
1	4	4	9	15
2	4	10	10	10
3	2	3	0	29
4	1	0	0	33
5	34	0	0	0
6	34	0	0	0
7	0	0	0	34
8	0	0	18	16
9	29	0	5	0

Run_Bgb_Meng_Nat_Biotechnol_2005				
position	A	C	G	T
1	11	3	0	9
2	22	0	1	0
3	21	0	2	0
4	0	23	0	0
5	0	23	0	0
6	7	0	16	0
7	0	23	0	0
8	23	0	0	0
9	16	0	7	0
10	7	3	8	5

Odd_Meng_Nat_Biotechnol_2005				
position	A	C	G	T
1	2	0	21	0
2	0	21	2	0
3	2	0	0	21
4	18	0	0	5
5	0	23	0	0
6	0	9	0	14
7	0	0	22	1
8	3	2	13	5
9	10	3	1	9

Prd_Meng_Nat_Biotechnol_2003				
position	A	C	G	T
1	15	4	13	5
2	26	1	4	6
3	4	1	1	31
4	9	3	0	25
5	4	29	1	3
6	0	1	25	11
7	7	3	8	19
8	1	16	3	17
9	20	15	0	2
10	0	30	7	0
11	6	6	21	4
12	3	13	19	2
13	0	8	3	26

Slp1_Noyes_NAR_2008				
position	A	C	G	T
1	8	3	14	16
2	0	0	0	41
3	3	0	38	0
4	0	0	0	41
5	0	0	1	40
6	0	0	2	39
7	27	0	4	10
8	1	22	7	11
9	17	8	15	1
10	4	13	3	21
11	15	3	4	19

Opa_Noyes_NAR_2008				
position	A	C	G	T
1	0	1	15	2
2	10	7	1	0
3	1	16	0	1
4	1	15	1	1
5	0	16	1	1
6	0	18	0	0
7	0	18	0	0
8	0	14	0	4
9	4	0	13	1
10	0	14	2	2
11	3	1	4	10
12	2	0	16	0

Kni_new				
position	A	C	G	T
1	31	7	4	11
2	43	1	3	6
3	32	2	3	16
4	9	16	11	17
5	3	9	10	31
6	37	5	11	0
7	0	1	52	0
8	33	1	10	9
9	8	4	25	16
10	0	52	0	1
11	45	2	5	1
12	12	21	17	3

Gt_new				
position	A	C	G	T
1	3	0	2	11
2	9	2	3	2
3	1	0	0	15
4	0	1	2	13
5	9	1	5	1
6	2	9	2	3
7	3	2	9	2
8	1	5	1	9
9	13	2	1	0
10	15	0	0	1
11	2	3	2	9
12	11	2	0	3

Tll_new				
position	A	C	G	T
1	34	10	9	12
2	47	1	10	7
3	55	5	5	0
4	54	4	7	0
5	5	0	60	0
6	2	2	2	59
7	1	53	2	9
8	60	0	3	2
9	53	3	5	4
10	35	14	6	10

Binding site predictions in *cis*-elements

module	Bcd	Cad	Dstat	Hb	Kni	Kr	Gt	Tll	Hkb
eve_late	0.56			0.86	0.73				
eve_stripe1	1.22		0.79			1.57		0.90	1.92
eve_stripe2	1.44			1.14	0.51	1.54	0.40	0.49	
eve_stripe3_7		0.94	1.77	3.11	4.29				
eve_stripe4_6		2.84	0.95	2.56	0.38			0.46	
eve_stripe5	0.45	0.72	0.41	1.04		1.67	0.40	0.99	0.46
ftz_+3	0.89	1.75	1.26	1.40		2.15	0.97	2.73	
ftz_-1		1.62	1.17	1.63			0.57		
ftz_-6	0.84	1.08	0.90	2.42	2.69	0.66	1.82		
ftz_-7	0.81	1.69		3.33	1.43			0.96	
ftz_distal-proximal		1.19		1.68			1.28		
ftz_ps4_activator	0.66	0.44	0.87	1.21	1.40	0.23	0.90		
ftz_zebra_element		1.34	1.17	1.53					0.53
h_stripe1+5	2.33	2.56	2.30	3.55		5.82		3.69	
h_stripe2+6	1.31	4.95	2.69	4.31		4.14	1.57	2.20	
h_stripe3+4		1.57		2.80	1.65		1.03		2.35
h_stripe7		1.88	1.06	4.57	3.14	2.48		1.17	
odd_-1		1.24		1.19		1.43	1.03	1.22	
odd_-3		2.15	1.49	3.37	1.79	0.87		1.09	
odd_-5		0.76		0.71		2.39		0.89	1.15
odd_basal-1		1.68		1.43		1.87	1.13	1.62	
prd_+4	1.03			0.62		2.28		0.87	
run_-17		2.69	0.90	4.60			0.87	0.49	
run_-3	2.56	1.94	1.09	1.45		1.23	0.75	0.75	
run_-3_Kr-	2.56	1.81	1.09	1.44			0.76	0.74	
run_-9		1.79	0.80	4.36	2.93		0.90		
run_7stripes	3.80	4.49		6.43		2.57	2.36		1.17
run_stripe1	1.90		0.92	1.09		2.17		1.97	1.17
run_stripe3		2.54	0.89	5.45	2.92		1.33		
run_stripe5		1.24		1.78		0.75	0.65	1.19	0.59

Table 4

Maternal and gap input into stripe specific elements as described in the binding site analysis section. The original matrices were used except for KNI, GT, and TLL which are the new versions given in the Appendix of PWMs. The HKB matrix was taken directly from (Noyes et al., 2008b).

module	Hairy	Eve	Run	Ftz	Ftzf1	Odd	Opa	Prd	Slp1
eve_late			1.37			0.98	0.92	1.54	
eve_stripe1	0.74		0.67		1.20	0.96	0.39		0.92
eve_stripe2		0.48		0.33	0.41				
eve_stripe3_7						0.25	0.32	0.95	0.42
eve_stripe4_6			0.59	0.29			0.36	0.31	
eve_stripe5					1.34	0.63	0.49		
ftz_+3		0.94			1.24				1.00
ftz_-1	0.62			0.65					1.83
ftz_-6				0.74					
ftz_-7		0.78		0.89					
ftz_distal-proximal	1.11			1.28	3.75		1.14		
ftz_ps4_activator									
ftz_zebra_element	1.38				1.38		1.14		1.26
h_basal	3.94		1.51		1.08		1.28	1.64	
h_stripe1+5	2.98				1.48	1.85			
h_stripe2+6	1.57		1.35		1.94				2.48
h_stripe3+4	0.74	0.62	1.73		0.53				0.55
h_stripe7		1.24	2.13	0.73				0.95	0.73
h_up	4.23	4.73		3.98	3.24	3.00		2.95	4.37
odd_-1		1.05		1.44		1.54			
odd_-3	0.75				1.30				1.04
odd_-5	1.59				1.66			0.95	1.12
odd_-7			0.89		1.46				1.42
odd_basal-1		1.29		1.57		2.04			
odd_basal-5	3.40			2.61	3.59	3.26			3.28
run_-17	0.51	0.77		0.74	0.93	0.64			0.79
run_-3	1.93					1.00	1.01		
run_-3_Kr-	1.93			0.56		1.00	1.00		
run_-6	3.03		2.14		1.82		2.50		3.48
run_-9			0.88				0.52		
run_7stripes	3.84		3.24			2.46	4.49		2.72
run_stripe1		0.91	0.91	1.25	0.89		0.74	1.03	1.25
run_stripe3			1.97		1.20				1.85
run_stripe5				0.67	0.81				

Table 5

Pair rule input into elements. The PWMs used are listed in the Appendix of PWMs except HAIRY, which was from a previous study and had always matched up well with described HAIRY targets (Van Doren et al., 1994).

region	Hairy	Eve	Run	Ftz	Ftzf1	Odd	Prd	Slp1	overlaps
h -20	0.91		0.66			0.49		0.91	
h -19								0.52	
h -18	0.48				0.84			0.62	up
h -17	1.56	0.58		0.66	0.50	0.49	0.97	0.63	up
h -16	1.03	0.57			0.99				up
h -15	0.49	0.56				0.54		0.69	up
h -14		1.24	0.89	1.06				1.04	up
h -13		1.25		0.99		0.89		0.94	up
h -12		0.67		0.79	0.57				up
h -11		1.07		1.19	0.59	0.51		0.56	s3+4, s7
h -10	0.48	1.04	1.95	0.53			0.68	0.67	s3+4, s7
h -9	1.02		0.92		0.65		0.59		s2+6, s7
h -8		0.61		0.97				0.77	s2+6
h -7					1.06	0.63		1.55	s1+5, s2+6
h -6	0.66			0.53	0.53				s1+5, s2+6
h -5	1.15								s1+5
h -4	1.17				0.64	1.08	0.49		s1+5
h -3		0.72		0.70				0.83	
h -2	1.18	0.50						0.75	
h -1		1.00		0.60				1.20	
h -0	1.53		0.95		0.59	0.60	0.58		basal
h +0		0.50	0.93			0.73	1.22		
h +1	0.72		1.21						
h +2	1.20	0.52		0.52		0.53		1.36	
h +3			0.47		0.53			1.31	
h +4	0.87		0.54		0.82	1.11		0.89	
h +5	1.47	0.78		0.96					
h +6	1.72	0.68		0.70					
h +7		0.79				0.73			
h +8	1.36	1.11	0.77	0.52			0.57		
h +9	0.78				0.70	0.55		1.67	
h +10		0.58				1.10		0.56	
h +11		0.51	1.21	0.72					
h +12	0.45						0.62		
h +13				0.52	0.68				
h +14	1.75	0.76	1.28		0.77		0.76		
h +15		0.50							
h +16	0.74	0.60	0.89	0.69			0.50	1.21	
h +17	2.50		0.51			0.61			
h +18		1.22	0.63	1.04				0.64	
h +19	0.53	0.49				0.51	0.76		
h +20	0.79								

Table 6

Binding site analysis over 1 kb fragments of the *hairy* locus using pair rule PWMs.

region	Hairy	Eve	Run	Ftz	Ftzf1	Odd	Prd	Slp1	overlaps
eve -14				0.59	0.51		0.92	0.61	
eve -13							0.47	1.05	
eve -12	0.66	0.53		0.84	0.91		0.53		
eve -11					0.79	1.12			
eve -10						0.70	0.48		
eve -9								1.28	
eve -8								1.24	
eve -7			1.79		1.47	1.03			
eve -6								0.77	late
eve -5			1.02			0.86	1.39		late
eve -4			0.63				0.57		late
eve -3			0.84				1.09	0.54	s3+7
eve -2				0.54					
eve -1		0.52			0.66				s2
eve -0	1.22		0.53		0.58				
eve +0	1.21	0.60		0.70	2.41		0.67		
eve +1	0.58				0.54			1.31	
eve +2			0.79					0.49	
eve +3			1.46	0.55					s4+6
eve +4	0.48		0.92	0.57	1.07	0.81			
eve +5	0.59		0.78		1.90	1.03	0.49	0.96	s1
eve +6			1.21		1.35	0.65			s1, s5
eve +7		0.61		0.74					
eve +8					0.87				
eve +9							0.72	0.48	
eve +10					1.10	0.62			
eve +11	0.72				1.86	0.80	0.64		
eve +12			1.01					0.95	
eve +13		0.90		1.43				1.00	
eve +14			0.56	0.57			1.00	1.36	

Table 7

Binding site analysis over 1 kb fragments of the *eve* locus using pair rule PWMs.

region	Hairy	Eve	Run	Ftz	Ftzf1	Odd	Prd	Slp1	overlaps
run -20								0.79	
run -17		0.96							-17
run -16	0.48			0.59	1.58	0.60		1.11	-17
run -15								0.59	
run -14				0.89	0.58			1.01	s1
run -13	0.47	0.78	0.69	1.07	0.54		0.56	1.05	s1
run -12				0.65	0.54				s1, s5
run -11					1.01				s3, s5
run -10			0.71					1.06	s3
run -9			1.14					0.76	-9, s3
run -8		0.62	0.85	0.79	0.69			0.48	
run -7			0.66					0.73	
run -6	1.12				0.60				
run -5	1.59							1.50	7s
run -4			1.12		0.70	0.64		1.56	7s
run -3	0.64		0.89	0.52		0.96	0.53		-3, 7s
run -2	1.79								-3, 7s
run -1		0.78		0.49					7s
run -0	1.06		0.83		0.74			0.55	7s
run +0		0.60	0.85	1.01				1.27	
run +1		0.47	1.11		1.00				
run +2	0.92		0.52	0.59			0.46	0.85	
run +3	0.58	1.13	0.48					0.65	
run +4			1.28		1.07	0.60		0.96	
run +5		0.55				0.64	0.62		
run +6								0.88	
run +7			0.48		0.66		0.64		
run +8	0.66		0.87					1.24	
run +9	0.83		1.00			0.82			
run +10		0.75	0.74			0.93			
run +11		0.79	1.44	0.55		0.70		0.83	
run +12			0.52			0.51	0.72	0.86	
run +13	0.51	0.96	1.52	0.73				0.60	
run +14		1.08		1.30					
run +15				0.53	1.48			0.60	
run +16		0.99	0.91	1.03	1.13	0.52	1.16	0.92	
run +17			0.75			1.88	0.47	0.53	
run +18			0.77						
run +19	0.49		0.58	0.73					
run +20	3.45		0.69		1.36		1.79	0.52	

Table 8

Binding site analysis over 1 kb fragments of the *run* locus using pair rule PWMs.

region	Hairy	Eve	Run	Ftz	Ftzf1	Odd	Prd	Slp1	overlaps
ftz -20								0.62	
ftz -19		0.55	1.06	0.81				0.87	
ftz -18	0.84		1.08	0.76	1.35			1.29	
ftz -17				0.80					
ftz -16		0.53						0.49	
ftz -15							0.58		
ftz -13		0.54	0.79	0.73				0.66	
ftz -12								0.59	
ftz -11		0.61	0.49	0.51				0.52	
ftz -10		0.60		0.62			1.16	0.49	
ftz -9		0.53		0.57	0.67			0.90	
ftz -8		0.57							-7
ftz -7				0.69					-7
ftz -6									-6, -7, auto
ftz -5	0.68			0.76				0.69	-6, auto
ftz -4		0.49		0.46	2.01	0.65			auto
ftz -3		0.92			0.85				auto
ftz -2	1.13			0.76				0.53	
ftz -1		0.59		0.85				0.87	-1, zebra
ftz -0	1.35				1.37			1.24	-1, zebra
ftz +0		0.67		0.48					+3
ftz +1					0.65				+3, s1+5
ftz +2		0.64		0.59			0.74	1.30	+3, s1+5
ftz +3							0.49		
ftz +4		0.82		0.67					
ftz +5		1.02	0.61	1.19			0.55	0.56	
ftz +6		0.88		1.31					
ftz +7		0.61		0.74	0.46		0.52		
ftz +8					0.51			0.81	
ftz +9		0.54					0.87		
ftz +10					0.83				
ftz +11	0.86							1.43	
ftz +12	0.54					0.61		1.41	
ftz +13		0.78		0.76	0.55			0.53	
ftz +14			0.58						
ftz +15			0.51		0.67	0.90	0.48	0.69	
ftz +16		0.49		0.54				1.68	
ftz +17								0.77	
ftz +18			0.68	0.60				0.71	
ftz +19		0.99		0.72	1.26			0.73	
ftz +20		0.99		1.37		0.96			

Table 9

Binding site analysis over 1 kb fragments of the *ftz* locus using pair rule PWMs.

region	Hairy	Eve	Run	Ftz	Ftzf1	Odd	Prd	Slp1	overlaps
odd -20	0.83		1.06			0.76			
odd -19	0.50	0.53		0.87			0.97		
odd -18						0.60			
odd -17	0.48				0.82				
odd -16									
odd -15		0.63							
odd -14		1.03		0.84		1.03			
odd -12			0.79			0.83		1.48	
odd -11			0.70		0.49				
odd -10	1.10			0.51				1.02	
odd -9	0.53		0.49	0.63		0.90	0.78		
odd -8	1.27							0.83	
odd -7		0.53	0.92		1.30				-7
odd -6	0.62				1.24		0.61	1.25	-7
odd -5	0.80	0.85		0.57				0.96	-5
odd -4					0.61	0.48	0.89	0.53	-5
odd -3	1.60				1.72	0.53		0.77	-3, -5
odd -2	0.51	0.49		0.52	0.61			0.85	-3
odd -1		0.76		1.35				0.58	-1
odd -0						1.36			-1, basal
odd +0	0.54	0.58	0.58						basal
odd +1		0.51		0.77	0.77			0.52	
odd +3			0.84						
odd +4						0.51			
odd +5			0.97		1.01		0.48	1.10	8
odd +6	0.64						0.53	0.67	
odd +7	0.52		0.57			0.65			
odd +8	1.78			0.67		0.46		0.89	
odd +9		0.60		0.64				0.47	
odd +10		0.56	0.77						
odd +11	0.57				0.96				
odd +12		0.60		0.56	1.06		0.52		
odd +13	1.25								
odd +14						0.68	0.81	0.47	
odd +15			0.53					1.21	
odd +16			1.08		0.57				
odd +17				0.74		1.13			
odd +18		0.46	0.57			0.53			
odd +20		0.67	1.04		1.09				

Table 10

Binding site analysis over 1 kb fragments of the *odd* locus using pair rule PWMs.

module	Bcd	Cad	Dstat	Hb	Kr	Kni	Gt	Tll	Hkb
gt_anterior	2.15				1.11	1.30		1.00	1.05
gt_anter_post			2.13	0.87	2.30			0.70	
gt_posterior	0.89	1.88	1.26	2.19	2.89		0.74	0.60	
hb_anterior	1.72			0.99	0.54		0.42		
hb_late		0.89	0.46	1.40	2.99	1.09			
kni_anterior	1.61	0.85	1.61	2.02	0.95	0.69			
kni_anter_post	1.07	2.84		3.51		1.42		0.87	
kni_posterior	1.82	2.07	1.02	3.39	3.28	0.98		3.62	
Kr_AD2		1.64		1.90		0.90	0.64	0.93	
Kr_CD1	1.56	1.52		3.27			0.83	1.21	0.53
Kr_CD2_AD1	1.36	1.56		3.23	1.78	1.28	1.43	2.15	
tll_CD1_anter	2.14	0.93		0.74				0.42	
tll_D3_anter_post		0.86		0.49	0.31				
tll_K11_post		1.28	0.88	0.91	1.67			0.32	

Table 11

Maternal and gap input into the gap gene *cis*-elements as described in the binding site analysis section. The original matrices were used except for KNI, GT, and TLL which are the new versions given in the Appendix of PWMs. The HKB matrix was taken directly from (Noyes et al., 2008b).

Regions from the segmentation *cis*-element screen

gene	chromosome	begin	end	length
btd	X	9579705	9601080	21376
cad	2L	20768523	20785689	17167
cnc	3R	19010300	19028372	18073
croc	3L	21457768	21480908	23141
D	3L	14158701	14178468	19768
ems	3R	9707901	9739675	31775
eve	2R	5861246	5875515	14270
fkh	3R	24398743	24422780	24038
ftz	3R	2681934	2701803	19870
Gsc	2L	580549	601112	20564
gt	X	2310755	2335340	24586
h	3L	8649204	8682345	33142
hb	3R	4507600	4536731	29132
hkb	3R	166226	181203	14978
kni	3L	20675936	20708895	32960
knrl	3L	20581242	20628330	47089
Kr	2R	21094475	21126356	31882
noc	2L	14470862	14504019	33158
nub	2L	12615220	12637085	21866
oc	X	8515373	8564950	49578
odd	2L	3594402	3617286	22885
opa	3R	658984	700812	41829
Optix	2R	3902233	3939898	37666
pdm2	2L	12674826	12689538	14713
prd	2L	12078635	12095932	17298
run	X	20546509	20578333	31825
slp1	2L	3815407	3828109	12703
slp2	2L	3828110	3848795	20686
tll	3R	26668059	26690095	22037

Table 12

Genomic regions used in the original *cis*-element screen in Release 5 coordinates.

Coordinates of *cis*-elements

gene	chromosome	begin	end	length
ftz_-7	3R	2681761	2683378	1618
ftz_-6	3R	2683373	2684612	1240
ftz_-1	3R	2688614	2689688	1075
ftz_+3	3R	2692616	2694360	1745
h_up	3L	8650357	8656519	6163
h_basal	3L	8668300	8670477	2178
odd_-7	2L	3613177	3614413	1237
odd_-1	2L	3606953	3608462	1510
odd_+8	2L	3598286	3599055	770
odd_basal-1	2L	3606032	3608462	2431
odd_basal-5	2L	3606032	3611858	5827
odd-8	2L	3612375	3617476	5102
prd_+4	2L	12080376	12081687	1312
run_-17	X	20548261	20549257	997
run-16	X	20549285	20551803	2519
run_-9	X	20555735	20556596	862
run_-6	X	20557443	20560872	3430
run_-3	X	20561710	20563080	1371
run_-3_Kr-	X	20561710	20563080	1371
run_+3	X	20567796	20569156	1361
run_+6	X	20571490	20572650	1161

Table 13

Genomic positions of new constructs in release 5 coordinates.

cis-element	chromosome	begin	end	length
eve_stripe1	2R	5873440	5874240	801
eve_stripe2	2R	5865217	5865879	663
eve_stripe3_7	2R	5863006	5863516	511
eve_stripe4_6	2R	5871404	5872005	602
eve_stripe5	2R	5874147	5874946	800
ftz_stripe1+5	3R	2692616	2694360	1745
ftz_stripe2+7	3R	2683373	2684612	1240
ftz_stripe3+67	3R	2681761	2683378	1618
h_stripe1+5	3L	8662058	8665028	2971
h_stripe2+6	3L	8659411	8662070	2660
h_stripe3+4	3L	8657463	8658374	912
h_stripe7	3L	8657938	8659411	1474
odd_stripe1+5	2L	3610420	3611803	1384
odd_stripe3+6	2L	3608812	3610461	1650
prd_anterior	2L	12080376	12081687	1312
run_stripe1	X	20551039	20552655	1617
run_stripe3	X	20555735	20556596	862
run_stripe4	X	20548261	20549257	997
run_stripe5	X	20552655	20553990	1336
run_stripes2+7	X	20594595	20597303	2709
Kr_CD1	2R	21110142	21111300	1159
Kr_CD2_AD1	2R	21111575	21113281	1707
Kr_AD2	2R	21113281	21114511	1231
kni_anterior_posterior	3L	20687055	20688533	1479
kni_posterior	3L	20689640	20690671	1032
kni_anterior	3L	20692603	20694005	1403
hb_anterior	3R	4520323	4521043	721
hb_late	3R	4526520	4527542	1023
tll_K11_post	3R	26675265	26675744	480
tll_CD1_anter	3R	26676777	26677272	496
tll_D3_anter_post	3R	26677663	26678030	368
gt_anterior_posterior	X	2323048	2324286	1239
gt_posterior	X	2324294	2325502	1209
gt_anterior	X	2331789	2333533	1745

Table 14

Release 5 coordinates of the *cis*-elements used in the binding site analysis

Primers for *cis*-elements

Lower case bases in primers correspond to restriction sites except for the sequences related to the run_-3 mutagenesis where they correspond to mutated bases.

```
>h_up-1.5
ACAAGGGAAAGGGGATGTGG
>h_up-1.3 joined to h_up-2 by a natural SacII site in the overlap between the two
fragments
AGTCACGCATGGAAGCGAAC
>h_up-2.5
CCAAGCCCCAATAACCCAAG
>h_up-2.3 joined to h_up-3 by a natural SpeI site in the overlap between the two
fragments
AACCCTTCCGATTGCTCCAC
>h_up-3.5
AAACACGACCTAATTGCGATCAAC
>h_up-3.3
ACCTGCGACTGCAAGCAAAG
>h_basal.5
ctcgagCCGCAGATACACAGTACACAGCACAA
>h_basal.3
CCAGAATGTCGGCCTTTTCCAA
>h_basal2.3 used with h_basal.5 to PCR off of h_basal.5 & h_basal.3 cloned into TOPO to
subclone a precise fragment for in frame fusion
cctgaggGTTCGACTGCAAGAGGCAAGAAA
>odd_-7.5
CGATCCGGCTATTAGGGCACTGTTT
>odd_-7.3
GGCGACTCAAGGTCACTGGCGTAT
>odd_-1.5
gcggccgcGCGGAAATGCTTCACCTGGAAA
>odd_-1.3
actagtCCGACCGAATGTGCCTGTGGAT
>odd_+8.5
GCAGCACCCACCCAAAACAACAA
>odd_+8.3
GGTGGGTTCACATGGCCAGAAGAT
>odd_basal_-1.5
tctagaGCGGAAATGCTTCACCTGGAAA
>odd_basal_-1.3
cctgaggCGTCAGCATGGGGGAGAGACTT
>odd_basal_-5.5
agatctGGGTTGGCTCCGACTCCGTTTATT
>odd_basal_-5.3 joined to -3 with an XbaI site in the overlapping region
TCGTAGTTTTTGGTGCATTCGAGGA
>odd_basal_-3.5
TGCACCCGTCTTCTTCCTTTTCGT
>odd_basal_-3.3 joined to -1_alt with a BamHI in the overlapping region (there are two
adjacent BamHI sites resulting in a 8 bp deletion)
CGACTTTCCAGGTGAAGCATTTCC
>odd_basal_-1_alt.5
CCGAGAGTTCCAGACACACGCACT
>odd_basal_-1_alt.3 inserted into odd_basal_-1 construct in Casper with a SpeI site in
both vectors
TCCAGCAGTAAGCAAGCGGAAACC
>prd_+4.5
gaattcCCCAAAAGGACCAAACCGAAATGTT
>prd_+4.3
actagtCGAAGTGTACCTGCTGATCCGATGAT
>run_-17.5
gaattcGCACCTCATTAGCAGCCGCACATT
>run_-17.3
```

actagtCCCCGGAAGCCAAGGTAAACAGAA
>run_-16.5
TCTGCTCATCCATTGACTTTTGTG
>run_-16.3
CGAATCAGCCGCGTTAATTG
>run_-9.5
CCACATCCTTCGTCGCTTCCTCTT
>run_-9.3
CCTGCTCGCCCTCTGTTCTGCTTTA
>run_-6.5
CGAGACGCGAGTTAATCAAGCATTTTT
>run_-6.3
ctcgagCGGCGTTGTGTAAGTGAATTGTGGTTT
>run_-3.5
CCTCGAGCGTAGGTGCCTCTCTTT
>run_-3.3
actagtCGCCGCTCTCAGCTGGACATTA
>run_+3.5
GGAGCAGCCTCATCAGTGGGATAGAA
>run_+3.3
GGACCACCAGCGGACAGATTGTTA
>run_+6.5
ATTGTAACTATTGGCTTGACTGC
>run_+6.3
AACCACAGCGAGGATTAAAGC
>run_-3_KR- construct mutated bases in lower case
CCTCGAGCGTAGGTGCCTCTCTTTCGTTCGCCTGGTGCCGTCTCTTTCTGCGAGGGGGAG
GGATCTTGCGCACGTATATGAATAATTCAACTGATCGCCGTGGTGATTGGGGGATGAACt
GAGTACTGGTTTGTCCGTTCCTCGGAGGTGCGGAATGCACAGATCGTAGTTCTGATACCC
ATTCATTTCGAGGAAATTTATTCAGGAACGTCACGGTATTTGCATAAGTGAAATCGTATC
TCACAGTTAGCAGCCTGTATCGTAGATTGATTACTAAAATACTTTTCTAAATAATCTGCA
CTAAGATATAGTTCAGATTGCGTAAGATCGGTAAGTACAGAAGCTTTTAATCGCACTGGA
AGTTTTATTTACCGCTCACGACATTTGCATAGATGAAACCGTATCTTACAGATTCAAGTA
GCTTGCGTACTAAACTACTGTACAAAAAATATCTGCACTAAGAAATAGTCGGGAGTGATT
TTTTATGGTGAAGTACAGAAGTAAATTCAGCTAATTAAGCCTCGCTCTTTTTTGTTTTAA
ATTAACTGCTCAACAGCAGTGAGAAAAACATTGCATATATTGAGTACATGAATACAGGTT
ACTGTCGCCTTAATCTCCCATCGGTTAATCCCGCTAAAAAGCGAAGTCCTTTGAGTTTTT
GGTGGCCAGGTAGGCACTTTCCGTATCAGATGCTCGTTGCTTATTTTTTGGGAACATATT
TTATGGCCCGTGGCGGCGGCTAATCGGCCAAAATATTTGCGGGGCGTGCTGCTTAATCCG
GGCGATTGACTTTCATAAGCAAAtGATTAAGATTGCGCGGTTGGACTACCTGTTTTGAGG
TGCGATATCAGTACATGCGATGGTACATCTGAGGGCCAGGTACGTCAAAGCCAGTAAACC
CATAGTTTTCCCACTTTTTTGGGGCCGCAAAAAAGCATCGGAGGGCCCATAAAAAAAtGG
TTGATCaTTTTGGCTGCTGTGGGCTCGTAGCGAGTTCGGTATGGAGATCAGGTACTGCCT
GGTGCTCGGTGATCCCTATGAGGCGGTCCTGCGGGTCCTGCGATGCGCGTGCTGCGGCAG
CTCCTGTCGCAAATTGCCAAGGAATCGCAGCAGGATCCAAGAAGCGACGACAGGAGCGCT
GATTTCCCGGGAAACCGGCAATCGTCAATCGGCCATCGGCAATCGCGTCCTTGTCGCACG
CCCGCTAAACCTGCGCTGTCCTGCCATATATCCCGGGGCTATATGGTGTGAAATCGGTGT
AGGGACACGAGGTCCTTCGCAGCGAGCGGCCGCGCACGTACAAAAGGCAGCGCTGCCGAT
ACACTGGATTTACTGGAAGTGTGCGTCTATAATGTCCAGCTGAGAGCGGCG
>run_-3.120.3 mutagenic base in lower case
CTCaGTTCATCCCCCAATCACCAC
>run_-3.804.new.5 mutagenic base in lower case
CATAAGCAAAtGATTAAGATTGCGCGG
>run_-3.958-967.3 mutagenic bases in lower case
AtGATCAACCaTTTTTTTATGGGCCCTCC
>run_-3.958-967.5 mutagenic bases in lower case
AtGGTTGATCaTTTTGGCTGCTGTGG
>run_-3.KR+
CCTCGAGCGTAGGTGCCTCTCTTTCGTTCGCCTGGTGCCGTCTCTTTCTGCGAGGGGGAG
GGATCTTGCGCACGTATATGAATAATTCAACTGATCGCCGTGGTGATTGGGGGATGAAaG
GgtTACTGGTTTGTCCGTTCCTCGGAGGTGCGGAATGCACAGATCGTAGTTCTGATACCC
ATTCATTTCGAGGAAATTTATTCAGGAACGTCACGGTATTTGCATAAGTGAAATCGTATC
TCACAGTTAGCAGCCTGTATCGTAGATTGATTACTAAAATACTTTTCTAAATAATCTGCA
CTAAGATATAGTTCAGATTGCGTAAGATCGGTAAGTACAGAAGCTTTTAATCGCACTGGA
AGTTTTATTTACCGCTCACGACATTTGCATAGATGAAACCGTATCTTACAGATTCAAGTA
GCTTGCGTACTAAACTACTGTACAAAAAATATCTGCACTAAGAAATAGTCGGGAGTGATT
TTTTATGGTGAAGTACAGAAGTAAATTCAGCTAATTAAGCCTCGCTCTTTTTTGTTTTAA
ATTAACTGCTCAACAGCAGTGAGAAAAACATTGCATATATTGAGTACATGAATACAGGTT
ACTGTCGCCTTAATCTCCCATCGGTTAATCCCGCTAAAAAGCGAAGTCCTTTGAGTTTTT
GGTGGCCAGGTAGGCACTTTCCGTATCAGATGCTCGTTGCTTATTTTTTGGGAACATATT
TTATGGCCCGTGGCGGCGGCTAATCGGCCAAAATATTTGCGGGGCGTGCTGCTTAATCCG

GGCGATTGACTTTCATAAGaAAAGGgTTAAGATTGCGCGGTTGGACTACCTGTTTTGAGG
TGCGATATCAGTACATGCGATGGTACATCTGAGGGCCAGGTACGTCAAAGCCAGTAAACC
CATAGTTTTCCCACTTTTTTGGGGCCGCAAAAAAGCATCGGAGGGCCCATAAAAAAAGGG
TTaATCCcTTTGGCTGCTGTGGGCTCGTAGCGAGTTCGGTATGGAGATCAGGTACTGCCT
GGTGCTCGGTGATCCCTATGAGGCGGTCCTGCGGGTCCTGCGATGCGCGTGCTGCGGCAG
CTCCTGTCGCAAATTGCCAAGGAATCGCAGCAGGATCCAAGAAGCGACGACAGGAGCGCT
GATTTCCCGGGAAACCGGCAATCGTCAATCGGCCATCGGCAATCGCGTCCTTGTCGCACG
CCCGCTAAACCTGCGCTGTCCTGCCATATATCCCGGGGCTATATGGTGTGAAATCGGTGT
AGGGACACGAGGTCCTTCGCAGCGAGCGGCCGCGCACGTACAAAAGGCAGCGCTGCCGAT
ACACTGGATTTACTGGAAGTGTGCGTCTATAATGTCCAGCTGAGAGCGGCG
>run-3.5.122.3
CAAACCAGTAacCCtTTCATCCCCC
>run-3.5.806.5
CTTTCATAAGaAAAGGgTTAAGATTGCGCG
>run-3.5.963.3
CACAGCAGCCAAAAGGgTtAACCC

Bibliography

Aerts, S., van Helden, J., Sand, O., and Hassan, B.A. (2007). Fine-tuning enhancer models to predict transcriptional targets across multiple genomes. PLoS One *2*, e1115.

Andrioli, L.P., Oberstein, A.L., Corado, M.S., Yu, D., and Small, S. (2004). Groucho-dependent repression by sloppy-paired 1 differentially positions anterior pair-rule stripes in the Drosophila embryo. Dev Biol *276*, 541-551.

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

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

Arnosti, D.N., and Kulkarni, M.M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J Cell Biochem *94*, 890-898.

Barik, S. (1996). Site-directed mutagenesis in vitro by megaprimer PCR. Methods Mol Biol *57*, 203-215.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Benos, P.V., Bulyk, M.L., and Stormo, G.D. (2002). Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res *30*, 4442-4451.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., *et al.* (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266-1276.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and 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 U S A *99*, 757-762.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., and Celniker, S.E. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol *5*, R61. Epub 2004 Aug 2020.

Blair, S.S. (1995). Compartments and appendage development in Drosophila. Bioessays *17*, 299-309.

Butler, B.A., Soong, J., and Gergen, J.P. (1992). The Drosophila segmentation gene runt has an extended cis-regulatory region that is required for vital expression at other stages of development. Mech Dev *39*, 17-28.

Calhoun, V.C., and Levine, M. (2003). Long-range enhancer-promoter interactions in the Scr-Antp interval of the Drosophila Antennapedia complex. Proc Natl Acad Sci U S A *100*, 9878-9883.

Campos-Ortega, J.A., and Hartenstein, V. (1997). The Embryonic Development of Drosophila melanogaster, 2nd edn (New York, Springer).

Capovilla, M., Eldon, E.D., and Pirrotta, V. (1992). The giant gene of Drosophila encodes a b-ZIP DNA-binding protein that regulates the expression of other segmentation gap genes. Development *114*, 99-112.

Carroll, S.B., and Scott, M.P. (1986). Zygotically active genes that affect the spatial expression of the fushi tarazu segmentation gene during early Drosophila embryogenesis. Cell *45*, 113-126.

Carroll, S.B., and Vavra, S.H. (1989). The zygotic control of Drosophila pair-rule gene expression. II. Spatial repression by gap and pair-rule gene products. Development *107*, 673-683.

Cerny, A.C., Bucher, G., Schroder, R., and Klingler, M. (2005). Breakdown of abdominal patterning in the Tribolium Kruppel mutant jaws. Development *132*, 5353-5363.

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

Cockerill, K.A., Billin, A.N., and Poole, S.J. (1993). Regulation of expression domains and effects of ectopic expression reveal gap gene-like properties of the linked pdm genes of Drosophila. Mech Dev *41*, 139-153.

Copeland, J.W., Nasiadka, A., Dietrich, B.H., and Krause, H.M. (1996). Patterning of the Drosophila embryo by a homeodomain-deleted Ftz polypeptide. Nature *379*, 162-165.

Coulter, D.E., Swaykus, E.A., Beran-Koehn, M.A., Goldberg, D., Wieschaus, E., and Schedl, P. (1990). Molecular analysis of odd-skipped, a zinc finger encoding segmentation gene with a novel pair-rule expression pattern. Embo J *9*, 3795-3804.

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

Crick, F.H., and Lawrence, P.A. (1975). Compartments and polyclones in insect development. Science *189*, 340-347.

Crocker, J., Tamori, Y., and Erives, A. (2008). Evolution acts on enhancer organization to fine-tune gradient threshold readouts. PLoS Biol *6*, e263.

Dearolf, C.R., Topol, J., and Parker, C.S. (1989). The caudal gene product is a direct activator of fushi tarazu transcription during Drosophila embryogenesis. Nature *341*, 340-343.

Driever, W., Thoma, G., and Nusslein-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*, 363-367.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). Biological Sequence Analysis (Cambridge University Press).

Eldar, A., Shilo, B.Z., and Barkai, N. (2004). Elucidating mechanisms underlying robustness of morphogen gradients. Curr Opin Genet Dev *14*, 435-439.

Eldon, E.D., and Pirrotta, V. (1991). Interactions of the Drosophila gap gene giant with maternal and zygotic pattern-forming genes. Development *111*, 367-378.

Emberly, E., Rajewsky, N., and Siggia, E.D. (2003). Conservation of regulatory elements between two species of Drosophila. BMC Bioinformatics *4*, 57.

Erives, A., and Levine, M. (2004). Coordinate enhancers share common organizational features in the Drosophila genome. Proc Natl Acad Sci U S A *101*, 3851-3856.

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

Florence, B., Guichet, A., Ephrussi, A., and Laughon, A. (1997). Ftz-F1 is a cofactor in Ftz activation of the Drosophila engrailed gene. Development *124*, 839-847.

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

Fujioka, M., Emi-Sarker, Y., Yusibova, G.L., Goto, T., and Jaynes, J.B. (1999). Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. Development *126*, 2527-2538.

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

Fujioka, M., Miskiewicz, P., Raj, L., Gulledge, A.A., Weir, M., and Goto, T. (1996). Drosophila Paired regulates late even-skipped expression through a composite binding site for the paired domain and the homeodomain. Development *122*, 2697-2707.

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

Gao, Q., and Finkelstein, R. (1998). Targeting gene expression to the head: the Drosophila orthodenticle gene is a direct target of the Bicoid morphogen. Development *125*, 4185-4193.

Garcia-Bellido, A., Ripoll, P., and Morata, G. (1973). Developmental compartmentalisation of the wing disk of Drosophila. Nat New Biol *245*, 251-253.

Gaul, U., and Jackle, H. (1987). Pole region-dependent repression of the Drosophila gap gene Kruppel by maternal gene products. Cell *51*, 549-555.

Gehring, W.J. (1975). Determination of primordial disc cells and the hypothesis of stepwise determination. In Insect Development, P.A. Lawrence, ed. (Blackwell Scientific Publications), pp. 99-108.

Goto, T., Macdonald, P., and Maniatis, T. (1989). Early and late periodic patterns of even skipped expression are controlled by distinct regulatory elements that respond to different spatial cues. Cell *57*, 413-422.

Grad, Y.H., Roth, F.P., Halfon, M.S., and Church, G.M. (2004). Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D.pseudoobscura. Bioinformatics *20*, 2738-2750.

Gray, S., and Levine, M. (1996). Transcriptional repression in development. Curr Opin Cell Biol *8*, 358-364.

Green, R.B., Hatini, V., Johansen, K.A., Liu, X.J., and Lengyel, J.A. (2002). Drumstick is a zinc finger protein that antagonizes Lines to control patterning and morphogenesis of the Drosophila hindgut. Development *129*, 3645-3656.

Grossniklaus, U., Pearson, R.K., and Gehring, W.J. (1992). The Drosophila sloppy paired locus encodes two proteins involved in segmentation that show homology to mammalian transcription factors. Genes Dev *6*, 1030-1051.

Groth, A.C., and Calos, M.P. (2004). Phage integrases: biology and applications. J Mol Biol *335*, 667-678.

Gutjahr, T., Frei, E., and Noll, M. (1993). Complex regulation of early paired expression: initial activation by gap genes and pattern modulation by pair-rule genes. Development *117*, 609-623.

Gutjahr, T., Vanario-Alonso, C.E., Pick, L., and Noll, M. (1994). Multiple regulatory elements direct the complex expression pattern of the Drosophila segmentation gene paired. Mech Dev *48*, 119-128.

Hallikas, O., Palin, K., Sinjushina, N., Rautiainen, R., Partanen, J., Ukkonen, E., and Taipale, J. (2006). Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell *124*, 47-59.

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

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

Hartmann, C., Taubert, H., Jackle, H., and Pankratz, M.J. (1994). A two-step mode of stripe formation in the Drosophila blastoderm requires interactions among primary pair rule genes. Mech Dev *45*, 3-13.

Hatini, V., Green, R.B., Lengyel, J.A., Bray, S.J., and Dinardo, S. (2005). The Drumstick/Lines/Bowl regulatory pathway links antagonistic Hedgehog and Wingless signaling inputs to epidermal cell differentiation. Genes Dev *19*, 709-718.

Hiromi, Y., and Gehring, W.J. (1987). Regulation and function of the Drosophila segmentation gene fushi tarazu. Cell *50*, 963-974.

Hiromi, Y., Kuroiwa, A., and Gehring, W.J. (1985). Control elements of the Drosophila segmentation gene fushi tarazu. Cell *43*, 603-613.

Hoch, M., Gerwin, N., Taubert, H., and Jackle, H. (1992). Competition for overlapping sites in the regulatory region of the Drosophila gene Kruppel. Science *256*, 94-97.

Hoch, M., Schroder, C., Seifert, E., and Jackle, H. (1990). cis-acting control elements for Kruppel expression in the Drosophila embryo. Embo J *9*, 2587-2595.

Howard, K., and Ingham, P. (1986). Regulatory interactions between the segmentation genes fushi tarazu, hairy, and engrailed in the Drosophila blastoderm. Cell *44*, 949-957.

Howard, K., Ingham, P., and Rushlow, C. (1988). Region-specific alleles of the Drosophila segmentation gene hairy. Genes Dev *2*, 1037-1046.

Howard, K.R., and Struhl, G. (1990). Decoding positional information: regulation of the pair-rule gene hairy. Development *110*, 1223-1231.

Hughes, S.C., and Krause, H.M. (2001). Establishment and maintenance of parasegmental compartments. Development *128*, 1109-1118.

Hulskamp, M., Pfeifle, C., and Tautz, D. (1990). A morphogenetic gradient of hunchback protein organizes the expression of the gap genes Kruppel and knirps in the early Drosophila embryo. Nature *346*, 577-580.

Ingham, P., and Gergen, P. (1988). Interactions between the pair-rule genes *runt, hairy, even-skipped* and *fushi tarazu* and the establishment of periodic pattern in the *Drosophila* embryo. Development *104*, 51-60.

Ingham, P.W. (1988). The molecular genetics of embryonic pattern formation in Drosophila. Nature *335*, 25-34.

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

Jaynes, J.B., and Fujioka, M. (2004). Drawing lines in the sand: even skipped et al. and parasegment boundaries. Dev Biol *269*, 609-622.

Jurgens, G., Wieschaus, E., Nusslein-Volhard, C., and Kluding, H. (1984). Mutations affecting the pattern of the larval cuticle in Drosophila melanogaster: II. Zygotic loci on the third chromosome. Roux's Archives of Developmental Biology *193*, 283-295.

Kerridge, S., and Morata, G. (1982). Developmental effects of some newly induced Ultrabithorax alleles of Drosophila. J Embryol Exp Morphol *68*, 211-234.

Klingler, M., and Gergen, J.P. (1993). Regulation of runt transcription by Drosophila segmentation genes. Mech Dev *43*, 3-19.

Klingler, M., Soong, J., Butler, B., and Gergen, J.P. (1996). Disperse versus compact elements for the regulation of runt stripes in Drosophila. Dev Biol *177*, 73-84.

Kramer, S.G., Jinks, T.M., Schedl, P., and Gergen, J.P. (1999). Direct activation of Sex-lethal transcription by the Drosophila runt protein. Development *126*, 191-200.

Kraut, R., and Levine, M. (1991a). Mutually repressive interactions between the gap genes giant and Kruppel define middle body regions of the Drosophila embryo. Development *111*, 611-621.

Kraut, R., and Levine, M. (1991b). Spatial regulation of the gap gene giant during Drosophila development. Development *111*, 601-609.

La Rosee, A., Hader, T., Taubert, H., Rivera-Pomar, R., and Jackle, H. (1997). Mechanism and Bicoid-dependent control of hairy stripe 7 expression in the posterior region of the Drosophila embryo. Embo J *16*, 4403-4411.

La Rosee-Borggreve, A., Hader, T., Wainwright, D., Sauer, F., and Jackle, H. (1999). hairy stripe 7 element mediates activation and repression in response to

different domains and levels of Kruppel in the Drosophila embryo. Mech Dev *89*, 133-140.

Langeland, J.A., Attai, S.F., Vorwerk, K., and Carroll, S.B. (1994). Positioning adjacent pair-rule stripes in the posterior Drosophila embryo. Development *120*, 2945-2955.

Lardelli, M., and Ish-Horowicz, D. (1993). Drosophila hairy pair-rule gene regulates embryonic patterning outside its apparent stripe domains. Development *118*, 255-266.

Lawrence, P.A. (1992). The Making of a Fly: the Genetics of Animal Design (Blackwell Scientific Publications).

Lawrence, P.A., Johnston, P., Macdonald, P., and Struhl, G. (1987). Borders of parasegments in Drosophila embryos are delimited by the fushi tarazu and even-skipped genes. Nature *328*, 440-442.

Lecuit, T., and Wieschaus, E. (2000). Polarized insertion of new membrane from a cytoplasmic reservoir during cleavage of the Drosophila embryo. J Cell Biol *150*, 849-860.

Lee, H.H., and Frasch, M. (2000). Wingless effects mesoderm patterning and ectoderm segmentation events via induction of its downstream target sloppy paired. Development *127*, 5497-5508.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I.*, et al.* (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799-804.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565-570.

Lewis, E.B. (1998). The bithorax complex: the first fifty years. Int J Dev Biol *42*, 403-415.

Li, L., Zhu, Q., He, X., Sinha, S., and Halfon, M.S. (2007). Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses. Genome Biol *8*, R101.

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, Z., and Biggin, M.D. (1998). Eve and ftz regulate a wide array of genes in blastoderm embryos: the selector homeoproteins directly or indirectly regulate most genes in Drosophila. Development *125*, 4471-4482.

Lohr, U., and Pick, L. (2005). Cofactor-interaction motifs and the cooption of a homeotic Hox protein into the segmentation pathway of Drosophila melanogaster. Curr Biol *15*, 643-649.

Lohr, U., Yussa, M., and Pick, L. (2001). Drosophila fushi tarazu. a gene on the border of homeotic function. Curr Biol *11*, 1403-1412.

Lohs-Schardin, M., Cremer, C., and Nusslein-Volhard, C. (1979). A fate map for the larval epidermis of Drosophila melanogaster: localized cuticle defects following irradiation of the blastoderm with an ultraviolet laser microbeam. Dev Biol *73*, 239-255.

Manoukian, A.S., and Krause, H.M. (1992). Concentration-dependent activities of the even-skipped protein in Drosophila embryos. Genes Dev *6*, 1740-1751.

Margolis, J.S., Borowsky, M.L., Steingrimsson, E., Shim, C.W., Lengyel, J.A., and Posakony, J.W. (1995). Posterior stripe expression of hunchback is driven from two promoters by a common enhancer element. Development *121*, 3067-3077.

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

Markstein, M., Zinzen, R., Markstein, P., Yee, K.P., Erives, A., Stathopoulos, A., and Levine, M. (2004). A regulatory code for neurogenic gene expression in the Drosophila embryo. Development *131*, 2387-2394.

Meng, X., Brodsky, M.H., and Wolfe, S.A. (2005). A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol *23*, 988-994.

Morata, G., and Kerridge, S. (1981). Sequential functions of the bithorax complex of Drosophila. Nature *290*, 778-781.

Moreno, E., and Morata, G. (1999). Caudal is the Hox gene that specifies the most posterior Drosophile segment. Nature *400*, 873-877.

Morgan, T.H. (1928). The Theory of the Gene (Yale University Press).

Myasnikova, E., Samsonova, A., Kozlov, K., Samsonova, M., and Reinitz, J. (2001). Registration of the expression patterns of Drosophila segmentation genes by two independent methods. Bioinformatics *17*, 3-12.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., *et al.* (2000). A whole-genome assembly of Drosophila. Science *287*, 2196-2204.

Nasiadka, A., Dietrich, B.H., and Krause, H.M. (2002). Anterior-posterior patterning in the Drosophila embryo. Advances in Developmental Biology and Biochemistry *12*, 155-204.

Nasiadka, A., and Krause, H.M. (1999). Kinetic analysis of segmentation gene interactions in Drosophila embryos. Development *126*, 1515-1526.

Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008a). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell *133*, 1277-1289.

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

Nusslein-Volhard, C., Frohnhofer, H.G., and Lehmann, R. (1987). Determination of anteroposterior polarity in Drosophila. Science *238*, 1675-1681.

Nusslein-Volhard, C., Kluding, H., and Jurgens, G. (1985). Genes affecting the segmental subdivision of the Drosophila embryo. Cold Spring Harb Symp Quant Biol *50*, 145-154.

Nusslein-Volhard, C., and Wieschaus, E. (1980). Mutations affecting segment number and polarity in Drosophila. Nature *287*, 795-801.

Nusslein-Volhard, C., Wieschaus, E., and Kluding, H. (1984). Mutations affecting the pattern of the larval cuticle in Drosophila melanogaster: I. Zygotic loci on the second chromosome Roux's Archives of Developmental Biology *193*, 267-282.

Oberstein, A., Pare, A., Kaplan, L., and Small, S. (2005). Site-specific transgenesis by Cre-mediated recombination in Drosophila. Nat Methods 2, 583-585.

Ochoa-Espinosa, A., Yucel, G., Kaplan, L., Pare, A., Pura, N., Oberstein, A., Papatsenko, D., and Small, S. (2005). The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. Proc Natl Acad Sci U S A *102*, 4960-4965.

Ohneda, K., Mirmira, R.G., Wang, J., Johnson, J.D., and German, M.S. (2000). The homeodomain of PDX-1 mediates multiple protein-protein interactions in the formation of a transcriptional activation complex on the insulin promoter. Mol Cell Biol *20*, 900-911.

Olesnicky, E.C., Brent, A.E., Tonnes, L., Walker, M., Pultz, M.A., Leaf, D., and Desplan, C. (2006). A caudal mRNA gradient controls posterior development in the wasp Nasonia. Development *133*, 3973-3982.

Palin, K., Taipale, J., and Ukkonen, E. (2006). Locating potential enhancer elements by comparative genomics using the EEL software. Nat Protoc *1*, 368-374.

Pankratz, M.J., Busch, M., Hoch, M., Seifert, E., and Jackle, H. (1992). Spatial control of the gap gene knirps in the Drosophila embryo by posterior morphogen system. Science *255*, 986-989.

Pankratz, M.J., Hoch, M., Seifert, E., and Jackle, H. (1989). Kruppel requirement for knirps enhancement reflects overlapping gap gene activities in the Drosophila embryo. Nature *341*, 337-340.

Pankratz, M.J., and Jackle, H. (1990). Making stripes in the Drosophila embryo. Trends Genet *6*, 287-292.

Pankratz, M.J., Seifert, E., Gerwin, N., Billi, B., Nauber, U., and Jackle, H. (1990). Gradients of Kruppel and knirps gene products direct pair-rule gene stripe patterning in the posterior region of the Drosophila embryo. Cell *61*, 309-317.

Peterson, B.K., Hare, E.E., Iyer, V.N., Storage, S., Conner, L., Papaj, D.R., Kurashima, R., Jang, E., and Eisen, M.B. (2009). Big genomes facilitate the comparative identification of regulatory elements. PLoS ONE *4*, e4688.

Pick, L., Schier, A., Affolter, M., Schmidt-Glenewinkel, T., and Gehring, W.J. (1990). Analysis of the ftz upstream element: germ layer-specific enhancers are independently autoregulated. Genes Dev *4*, 1224-1239.

Plaza, S., Prince, F., Adachi, Y., Punzo, C., Cribbs, D.L., and Gehring, W.J. (2008). Cross-regulatory protein-protein interactions between Hox and Pax transcription factors. Proc Natl Acad Sci U S A *105*, 13439-13444.

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

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al.* (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-2309.

Riddihough, G., and Ish-Horowicz, D. (1991). Individual stripe regulatory elements in the Drosophila hairy promoter respond to maternal, gap, and pair-rule genes. Genes Dev *5*, 840-854.

Rivera-Pomar, R., Lu, X., Perrimon, N., Taubert, H., and Jackle, H. (1995). Activation of posterior gap gene expression in the Drosophila blastoderm. Nature *376*, 253-256.

Rothe, M., Wimmer, E.A., Pankratz, M.J., Gonzalez-Gaitan, M., and Jackle, H. (1994). Identical transacting factor requirement for knirps and knirps-related Gene expression in the anterior but not in the posterior region of the Drosophila embryo. Mech Dev *46*, 169-181.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., *et al.* (2000). Comparative genomics of the eukaryotes. Science *287*, 2204-2215.

Sander, K. (1976). Specification of the Basic Body Pattern in Insect Embryogenesis. In Advances in Insect Physiology, pp. 125-238.

Sanson, B. (2001). Generating patterns from fields of cells. Examples from Drosophila segmentation. EMBO Rep *2*, 1083-1088.

Saulier-Le Drean, B., Nasiadka, A., Dong, J., and Krause, H.M. (1998). Dynamic changes in the functions of Odd-skipped during early Drosophila embryogenesis. Development *125*, 4851-4861.

Schier, A.F., and Gehring, W.J. (1992). Direct homeodomain-DNA interaction in the autoregulation of the fushi tarazu gene. Nature *356*, 804-807.

Schier, A.F., and Gehring, W.J. (1993). Analysis of a fushi tarazu autoregulatory element: multiple sequence elements contribute to enhancer activity. Embo J *12*, 1111-1119.

Schroder, C., Tautz, D., Seifert, E., and Jackle, H. (1988). Differential regulation of the two transcripts from the Drosophila gap segmentation gene hunchback. Embo J *7*, 2881-2887.

Schroeder, M.D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E.D., and Gaul, U. (2004). Transcriptional control in the segmentation gene network of Drosophila. PLoS Biol *2*, E271. Epub 2004 Aug 2031.

Schulz, C., and Tautz, D. (1995). Zygotic caudal regulation by hunchback and its role in abdominal segment formation of the Drosophila embryo. Development *121*, 1023-1028.

Schupbach, T., and Wieschaus, E. (1986). Germline autonomy of maternal-effect mutations altering the embryonic body pattern of Drosophila. Dev Biol *113*, 443-448.

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

Senger, K., Armstrong, G.W., Rowell, W.J., Kwan, J.M., Markstein, M., and Levine, M. (2004). Immunity regulatory DNAs share common organizational features in Drosophila. Mol Cell *13*, 19-32.

Simpson-Brose, M., Treisman, J., and Desplan, C. (1994). Synergy between the hunchback and bicoid morphogens is required for anterior patterning in Drosophila. Cell *78*, 855-865.

Sinha, S., Schroeder, M.D., Unnerstall, U., Gaul, U., and Siggia, E.D. (2004). Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. BMC Bioinformatics *5*, 129.

Sinha, S., van Nimwegen, E., and Siggia, E.D. (2003). A probabilistic method to detect regulatory modules. Bioinformatics *19*, i292-301.

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

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

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

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

Sommer, R.J., and Tautz, D. (1993). Involvement of an orthologue of the Drosophila pair-rule gene hairy in segment formation of the short germ-band embryo of Tribolium (Coleoptera). Nature *361*, 448-450.

Stanojevic, D., Hoey, T., and Levine, M. (1989). Sequence-specific DNA-binding activities of the gap proteins encoded by hunchback and Kruppel in Drosophila. Nature *341*, 331-335.

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

Stormo, G.D. (2000). DNA binding sites: representation and discovery. Bioinformatics *16*, 16-23.

Struhl, G., Johnston, P., and Lawrence, P.A. (1992). Control of Drosophila body pattern by the hunchback morphogen gradient. Cell *69*, 237-249.

Surkova, S., Kosman, D., Kozlov, K., Manu, Myasnikova, E., Samsonova, A.A., Spirov, A., Vanario-Alonso, C.E., Samsonova, M., and Reinitz, J. (2008). Characterization of the Drosophila segment determination morphome. Dev Biol *313*, 844-862.

Swantek, D., and Gergen, J.P. (2004). Ftz modulates Runt-dependent activation and repression of segment-polarity gene transcription. Development *131*, 2281-2290.

Szabad, J., Schupbach, T., and Wieschaus, E. (1979). Cell lineage and development in the larval epidermis of Drosophila melanogaster. Dev Biol *73*, 256-271.

Thummel, C.S., and Pirrotta, V. (1991). Technical Notes: New pCaSperR P-element vectors. Drosophila Information Newsletter *2*, available at http://flybase.org/data/associated_files/DIN_compilation.htm#dinvol2.

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S.E., Richards, S., Ashburner, M., Hartenstein, V., Celniker, S.E., *et al.* (2002).

Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol *3*, RESEARCH0088.

Treisman, J., and Desplan, C. (1989). The products of the Drosophila gap genes hunchback and Kruppel bind to the hunchback promoters. Nature *341*, 335-337.

Van Doren, M., Bailey, A.M., Esnayra, J., Ede, K., and Posakony, J.W. (1994). Negative regulation of proneural gene activity: hairy is a direct transcriptional repressor of achaete. Genes Dev *8*, 2729-2742.

Vanderzwan-Butler, C.J., Prazak, L.M., and Gergen, J.P. (2007). The HMG-box protein Lilliputian is required for Runt-dependent activation of the pair-rule gene fushi-tarazu. Dev Biol *301*, 350-360.

Vavra, S.H., and Carroll, S.B. (1989). The zygotic control of Drosophila pair-rule gene expression. I. A search for new pair-rule regulatory loci. Development *107*, 663-672.

Walter, J., and Biggin, M.D. (1996). DNA binding specificity of two homeodomain proteins in vitro and in Drosophila embryos. Proc Natl Acad Sci U S A *93*, 2680-2685.

Walter, J., Dever, C.A., and Biggin, M.D. (1994). Two homeo domain proteins bind with similar specificity to a wide range of DNA sites in Drosophila embryos. Genes Dev *8*, 1678-1692.

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

Wieschaus, E., and Gehring, W. (1976). Clonal analysis of primordial disc cells in the early embryo of Drosophila melanogaster. Dev Biol *50*, 249-263.

Wieschaus, E., Nusslein-Volhard, C., and Jurgens, G. (1984). Mutations affecting the pattern of the larval cuticle in Drosophila melanogaster: III. Zygotic loci on the X-chromosome and fourth chromosome. Roux's Archives of Developmental Biology *193*.

Won, K.J., Sandelin, A., Marstrand, T.T., and Krogh, A. (2008). Modeling promoter grammars with evolving hidden Markov models. Bioinformatics *24*, 1669-1675.

Wu, X., Vakani, R., and Small, S. (1998). Two distinct mechanisms for differential positioning of gene expression borders involving the Drosophila gap protein giant. Development *125*, 3765-3774.

Yu, Y., Li, W., Su, K., Yussa, M., Han, W., Perrimon, N., and Pick, L. (1997). The nuclear hormone receptor Ftz-F1 is a cofactor for the Drosophila homeodomain protein Ftz. Nature *385*, 552-555.

Yu, Y., and Pick, L. (1995). Non-periodic cues generate seven ftz stripes in the Drosophila embryo. Mech Dev *50*, 163-175.

Zappavigna, V., Sartori, D., and Mavilio, F. (1994). Specificity of HOX protein function depends on DNA-protein and protein-protein interactions, both mediated by the homeo domain. Genes Dev *8*, 732-744.

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

Zuo, P., Stanojevic, D., Colgan, J., Han, K., Levine, M., and Manley, J.L. (1991). Activation and repression of transcription by the gap proteins hunchback and Kruppel in cultured Drosophila cells. Genes Dev *5*, 254-264.