A regulatory code for neurogenic gene expression in the *Drosophila* embryo

Michele Markstein[1,*,†], Robert Zinzen[1,*], Peter Markstein[2], Ka-Ping Yee[3], Albert Erives[1], Angela Stathopoulos[1] and Michael Levine[1,†]

[1]Department of Molecular and Cellular Biology, Division of Genetics and Development, 401 Barker Hall, University of California, Berkeley, CA 94720, USA
[2]Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA
[3]Computer Science Division Office, University of California, Berkeley, 387 Soda Hall #1776, Berkeley, CA 94720-1776, USA
*These authors contributed equally to this study
[†]Authors for correspondence (e-mail: mlevine@uclink4.berkeley.edu and michele@opengenomics.org)

Accepted 12 February 2004

Development 131, 2387-2394
Published by The Company of Biologists 2004
doi:10.1242/dev.01124

Summary

Bioinformatics methods have identified enhancers that mediate restricted expression in the *Drosophila* embryo. However, only a small fraction of the predicted enhancers actually work when tested in vivo. In the present study, co-regulated neurogenic enhancers that are activated by intermediate levels of the Dorsal regulatory gradient are shown to contain several shared sequence motifs. These motifs permitted the identification of new neurogenic enhancers with high precision: five out of seven predicted enhancers direct restricted expression within ventral regions of the neurogenic ectoderm. Mutations in some of the shared motifs disrupt enhancer function, and evidence is presented that the Twist and Su(H) regulatory proteins are essential for the specification of the ventral neurogenic ectoderm prior to gastrulation. The regulatory model of neurogenic gene expression defined in this study permitted the identification of a neurogenic enhancer in the distant *Anopheles* genome. We discuss the prospects for deciphering regulatory codes that link primary DNA sequence information with predicted patterns of gene expression.

Supplemental data available online

Key words: *Drosophila*, *Anopheles*, cis-regulation, enhancers, neurogenic ectoderm, mesectoderm, Twist, Su(H), Dorsal

Introduction

Comparative genome analyses have revealed remarkable constancy in the genetic composition of different animals. Vertebrates contain an average of 25,000 to 30,000 protein-coding genes, and most of these genes can be aligned with one another even among distantly related groups (e.g. Mural et al., 2002; Aparicio et al., 2002). This constancy extends to invertebrates. Although vertebrates contain about twice the number of genes as invertebrates, this increase in number is primarily due to the duplication of 'old' genes rather than the invention of new ones (e.g. Dehal et al., 2002). Thus, it would appear that animal diversity depends on the differential expression of a common set of genes during evolution.

Differential gene activity is primarily controlled by enhancers, which are typically 500 bp in length and contain roughly ten binding sites for two or more sequence-specific transcription factors (reviewed by Levine and Tjian, 2003). The total number of enhancers might be a critical determinant of organismal complexity. Based on well-characterized genes such as *even skipped* and *fushi tarazu*, which are regulated by multiple enhancers, one might estimate the *Drosophila* genome to contain 30,000-50,000 enhancers (e.g. Davidson, 2001). The use of comparative genome methods to understand animal diversity would be greatly facilitated by the existence of 'cis-regulatory codes' that link DNA sequence data with inferred patterns of gene activity. The dorsoventral patterning of the early *Drosophila* embryo provides a well-defined system for applying computational methods to the problem of predicting gene activity from DNA sequence information (Markstein et al., 2002; Markstein and Levine, 2002).

Dorsoventral patterning is controlled by the sequence-specific transcription factor Dorsal (reviewed by Stathopoulos and Levine, 2002). The Dorsal protein is distributed in a broad nuclear gradient in the early embryo, with peak levels in ventral regions, and progressively lower levels in more lateral and dorsal regions. This regulatory gradient initiates the differentiation of several embryonic tissues by regulating the expression of over 30 target genes in a concentration-dependent fashion (e.g. Casal and Leptin, 1996; Stathopoulos et al., 2002). Some of these target genes are activated by high levels of the Dorsal gradient within the presumptive mesoderm, whereas others are activated by intermediate or low levels of the gradient in ventral and dorsal regions of the neurogenic ectoderm, respectively. Previous studies identified seven of the estimated 30 Dorsal target enhancers in the *Drosophila* genome (reviewed by Rusch and Levine, 1996; Stathopoulos and Levine, 2002). Their analysis raised the possibility that co-regulated enhancers responding to the same levels of the Dorsal gradient share a distinctive combination of cis-regulatory elements (Stathopoulos et al., 2002).

Two of the previously identified enhancers are associated with the *rhomboid* (*rho*) and *ventral nervous system defective* (*vnd*) genes (White et al., 1983; Bier et al., 1990). Both enhancers are activated by intermediate levels of the Dorsal gradient in ventral regions of the neurogenic ectoderm (Ip et al., 1992; Stathopoulos et al., 2002). The present study identified a third enhancer, from the *brinker* (*brk*) gene (Jazwinska et al., 1999), which directs a similar pattern of expression. The three co-regulated enhancers share three sequence motifs, in addition to Dorsal binding sites: CACATGT, YGTGDGAA and CTGWCCY (Stathopoulos et al., 2002). The first two motifs bind the known transcription factors, Twist and Suppressor of Hairless [Su(H)], respectively (Thisse et al., 1987; Bailey and Posakony, 1995). All three motifs are shown to function as critical regulatory elements, thereby providing direct evidence that Twist and Su(H) are essential for the specification of the neurogenic ectoderm. A whole-genome survey for tightly linked Dorsal, Twist, Su(H) and CTGWCCY motifs identified only seven clusters in the entire *Drosophila* genome. Three correspond to the 'input' enhancers: *rho*, *vnd* and *brk*. Another two clusters are shown to correspond to new neurogenic enhancers associated with the *vein* (*vn*) and *single-minded* (*sim*) genes (Kasai et al., 1992; Schnepp et al., 1996). Additionally, the defined computational model for neurogenic gene expression permitted the identification of an orthologous *sim* enhancer in the distantly related *Anopheles* genome.

Materials and methods

Fly stocks

Strain yw^{67} was used for P-element transformations and in situ hybridization in *Drosophila melanogaster*, as described previously (e.g. Stathopoulos et al., 2002). Construction of the *stripe2-Notch*^{IC} strain and the derivation of *stripe2-Notch*^{IC}-expressing embryos was described (Cowden and Levine, 2002).

Cloning and injection of DNA fragments

Genomic *D. melanogaster* DNA was prepared from a single anesthetized *yw* male as described (Gloor et al., 1993). Mosquito DNA was derived from the *Anopheles gambiae* PEST strain (a gift from Anthony James). DNA fragments encompassing identified clusters were amplified from genomic DNA with the primer pairs listed (see supplemental data at http://dev.biologists.org/supplemental/). PCR products were purified with the Qiagen™ QiaQuick® PCR purification kit, and either cloned into the Promega™ pGEM® T-Easy vector (*brk*, *Ady*, *C1* and *vn*) or digested with restriction enzymes corresponding to restriction sites added to the 5′ ends of each primer pair. PCR products cloned into pGEM® T-Easy (*brk*, *Ady* and *C1*) were digested with *Not*I and cloned into the gypsy-insulated pCaSpeR vector E2G (a gift from Hilary Ashe), or partially digested with *Eco*RI (*vn*) and cloned into the [–42evelacZ]-pCaSpeR vector (Small et al., 1992). The remaining PCR products were directly digested and cloned into a modified version of the E2G vector called newE2G, which contains *Bgl*II, *Spe*I and *Eco*RI cloning sites in place of *Not*I. Enhancers were mutagenized in pGem® T-Easy using the Stratagene™ QuickChange® Multi Site-directed Mutagenesis Kit and the primers indicated (see supplemental data at http://dev.biologists.org/supplemental/). Constructs were introduced into the *D. melanogaster* germline by microinjection as described previously (e.g. Ip et al., 1992; Jiang and Levine, 1993; Rubin and Spradling, 1982). Between three and nine independent transgenic lines were obtained for each construct.

Whole-mount in situ hybridization

Embryos were hybridized with digoxigenin-labeled antisense RNA probes as described (Jiang et al., 1991). An antisense *lacZ* RNA probe was used to examine the staining patterns of transgenic embryos. To examine the patterns of endogenous gene expression, probes were generated by PCR amplification from genomic DNA. A 26 bp tail encoding the T7 RNA polymerase promoter (aagTAATACGA-CTCACTATAGGGAGA) was included on the reverse primer. PCR products were purified with the Qiagen™ PCR purification kit and used directly as templates in transcription reactions. Between 500 bp to 3 kb of coding sequence was used as a template for each probe.

Computational identification of shared motifs and enhancers

To identify shared motifs, we developed a program called MERmaid (available at www.opengenomics.org) which finds all n-mers of any length that are present or absent in specified groups of sequences. In this study, we considered two classes of motifs: 'exact match' motifs, in which every position in the motif is filled by one specific nucleotide; and 'fuzzy' motifs, in which up to two positions in the motif can be occupied by any of the four nucleotides. The *vn* and *sim* enhancers could be identified in genome-wide searches for clusters of sequence motifs using the parameters indicated in the text and supplement, and online search tools freely available at www.flyenhancer.org (Markstein et al., 2002). A similar tool is available for the mosquito genome at www.mosquitoenhancer.org.

Results

Previous studies identified two enhancers, from the *rho* and *vnd* genes, that are activated by intermediate levels of the Dorsal gradient in ventral regions of the neurogenic ectoderm (Ip et al., 1992; Stathopoulos et al., 2002). The present study identified a third such enhancer from the *brk* gene. This newly identified *brk* enhancer corresponds to one of the 15 optimal Dorsal-binding clusters described in a previous survey of the *Drosophila* genome (Markstein et al., 2002) (Fig. 1C). Although one of these 15 clusters was shown to define an intronic enhancer in the *short gastrulation* (*sog*) gene, the activities of the remaining 14 clusters were not tested. Genomic DNA fragments corresponding to these 14 clusters were placed 5′ of a minimal *eve-lacZ* reporter gene, and separately expressed in transgenic embryos using P-element germline transformation. Four of the 14 genomic DNA fragments were found to direct restricted patterns of *lacZ* expression across the dorsoventral axis, which are similar to the expression patterns seen for the associated endogenous genes (Fig. 1).

The four enhancers respond to different levels of the Dorsal nuclear gradient. Two direct expression within the presumptive mesoderm where there are high levels of the gradient. These are associated with the *Phm* and *Ady43A* genes (Fig. 1D,E). The third enhancer maps ~10 kb 5′ of *brk*, and is activated by intermediate levels of the Dorsal gradient (Fig. 1C, Fig. 2A), similar to the *vnd* and *rho* enhancers (Fig. 2C,E). Finally, the fourth enhancer maps over 15 kb 5′ of the predicted start site of the *CG12443* gene (Stathopoulos et al., 2002), and directs broad lateral stripes throughout the neurogenic ectoderm in response to low levels of the Dorsal gradient (Fig. 1B). In terms of the dorsoventral limits, this staining pattern is similar to that produced by the *sog* intronic enhancer (Fig. 1A).

The remaining ten clusters failed to direct robust patterns of expression and are thus referred to as 'false-positives' (data not

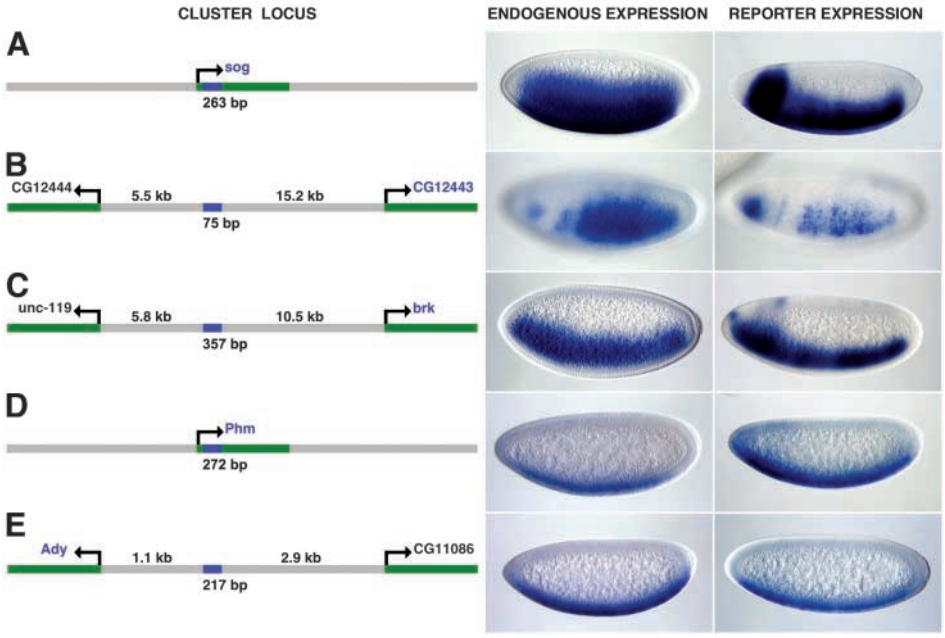


Fig. 1. Dorsal binding clusters identify regulatory DNAs. Diagrams on the left show the locations and sizes of five Dorsal binding clusters (depicted as blue boxes with sizes indicated below) identified in an earlier study (Markstein et al., 2002). In situ hybridization assays were performed to identify the expression profiles of the protein-coding genes (indicated as green boxes) located near the different clusters. Those genes found to be differentially expressed along the dorsal-ventral axis are shown in the middle column ('endogenous expression'). Genomic DNA fragments that encompass each of the five Dorsal-binding clusters were fused with a *eve-lacZ* reporter gene and expressed in transgenic embryos. Reporter gene expression (right column) was visualized by in situ hybridization using a digoxigenin-labeled *lacZ* antisense RNA probe. There is a close correspondence between the expression patterns of the endogenous genes and the staining patterns obtained with the fusion genes: *sog* (A) and *CG12443* (B) are expressed throughout the neurogenic ectoderm; *brk* (C) is expressed in the ventral neurogenic ectoderm; and *Phm* (D) and *Ady43A* (E) are expressed in the mesoderm. Lateral views of cellularizing embryos oriented with anterior to the left and dorsal up are shown.

shown). As analysis of spacing and orientation of the Dorsal sites alone did not reveal features that could discriminate between the false positives and the enhancers, we examined whether additional sequence motifs could aid in this distinction. We developed a program called MERmaid, which identifies motifs over-represented in specified sets of sequences. MERmaid analysis identified a group of motifs, which was largely specific to the *brk*, *vnd* and *rho* enhancers, suggesting that the regulation of these coordinately expressed genes is distinct from the regulation of genes that respond to different levels of nuclear Dorsal.

The *rho*, *vnd* and *brk* enhancers share common cis-regulatory elements

The *rho*, *vnd* and *brk* enhancers direct similar patterns of gene expression (Fig. 2). The *rho* and *vnd* enhancers were previously shown to contain multiple copies of two different sequence motifs: CTGNCCY and CACATGT (Stathopoulos et al., 2002). A three-way comparison of minimal *rho*, *vnd and brk* enhancers permitted a more refined definition of the CTGNCCY motif (CTGWCCY), and also allowed for the identification of a third motif, YGTGDGAA (Table 1, and supplemental data at http://dev.biologists.org/supplemental/).

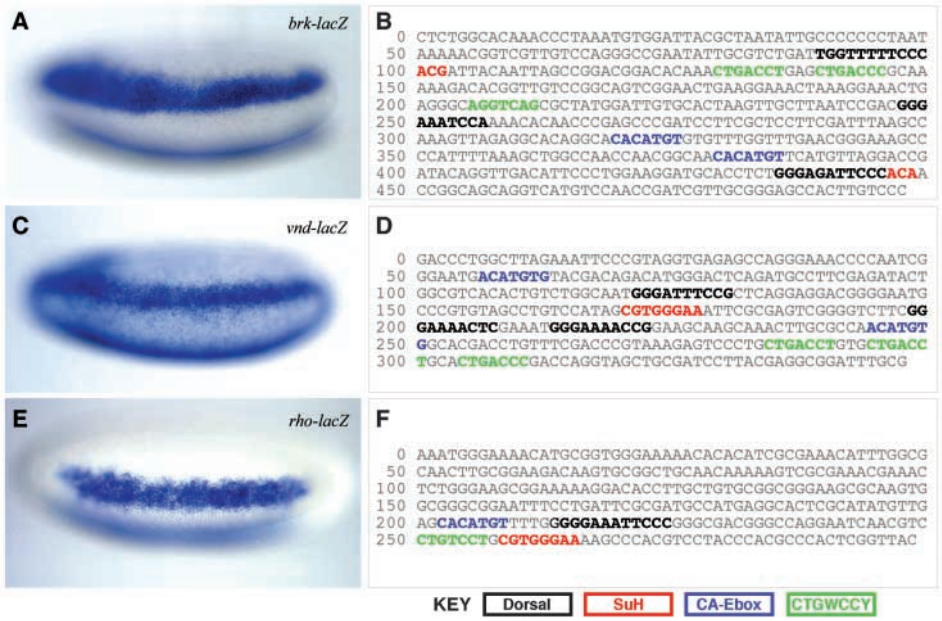


Fig. 2. The coordinately expressed *brk*, *vnd* and *rho* enhancers share sequence motifs. Embryos in A, C and E express *lacZ* fusion genes containing the enhancer sequences indicated in B (*brk*, 498 bp), D (*vnd*, 348 bp) and F (*rho NEE*, 299 bp), respectively. Reporter gene expression was visualized by in situ hybridization, as described in Fig. 1. The three enhancers direct similar lateral stripes of *lacZ* expression. Each enhancer contains at least one copy of each of: CTGWCCY (indicated in green) and binding sites for Dorsal (black), Su(H) (red) and Twist (CA-core E-box, blue). Ventrolateral views of cellularizing embryos oriented with anterior to the left and dorsal up are shown.

Table 1. Occurrence of shared motifs in Dorsal target enhancers and false-positive clusters

	zen 624 bp	*sog* 392 bp	*CG 12443* 343 bp	*brk* 498 bp	*vnd* 348 bp	*rho* 299 bp	*vn* 497 bp	*sim* 631 bp	*Phm* 443 bp	*Ady* 217 bp	False positives 6612 bp
E-box motifs											
CANNTG	3	1	5	4	3	5	3	5	6	1	43
CAAATG	1	1	2	0	0	1	0	1	0	1	14
CACATG	1	0	0	2	2	1	1	1	1	0	4
CACTTG	0	0	1	2	0	2	1	0	0	0	8
CAGCTG	1	0	1	0	0	0	0	0	2	0	4
CAGTTG	0	0	0	0	0	0	0	0	1	0	5
CAATTG	0	0	0	0	0	0	0	1	0	0	4
CATATG	0	0	0	0	0	1	0	1	2	0	0
CAGGTG	0	0	1	0	0	0	1	1	0	0	1
CATCTG	0	0	0	0	1	0	0	0	0	0	1
CACGTG	0	0	0	0	0	0	0	0	0	2	0
CACATGT	0	0	0	2	2	1	1	1	1	0	3
Su(H) motifs											
CGTGGGAA	0	0	0	1	1	1	1	2	0	0	0
TGTGGGAA	0	0	0	1	0	0	0	1	0	0	0
CGTGAGAA	0	0	0	0	0	0	0	1	0	0	0
TGTGAGAA	0	0	0	0	0	0	0	1	0	0	0
YGTGDGAA	0	0	0	2	1	1	1	5	0	0	0
Clustered motifs											
CTGNCCY	0	0	1	4	3	2	3	1	0	0	9
CTGWCCY	0	0	1	3	3	1	2	1	0	0	2

The frequency of specific sequence motifs belonging to the E-box, Su(H) and CTGWCCY motif families are shown for Dorsal target enhancers, as well as the group of 10 false-positive Dorsal clusters (FP-clusters C1-C10, see supplemental data at http://dev.biologists.org/supplemental/). The combined presence of sequences matching the E-box motif CACATGT, the Su(H) motif YGTGDGAA, and CTGWCCY (each highlighted in yellow) distinguishes the ventral neurogenic ectoderm enhancers (*brk*, *vnd* and *rho*; shaded in gray) from: mesodermal enhancers (*Phm* and *Ady43A*), enhancers responsive to the lowest levels of nuclear Dorsal (*zen*, *sog* and *CG12443*), and from the false-positive clusters of Dorsal-binding sites. As described in the text, a genome-wide search for clusters containing each of these motifs identified enhancers for *vn* and *sim*, which, like *brk*, *vnd* and *rho*, are responsive to intermediate levels of the Dorsal gradient and are expressed in the ventral neurogenic ectoderm.

The CACATGT and YGTGDGAA motifs bind the known transcription factors, Twist and Suppressor of Hairless [Su(H)], respectively (Thisse et al., 1991; Bailey and Posakony, 1995). All three motifs are over-represented in authentic Dorsal target enhancers directing expression in the ventral neurogenic ectoderm, as compared with the 10 false-positive Dorsal-binding clusters (Table 1). As indicated in Table I, some of the false-positive clusters contain motifs matching either Twist or CTGWCCY; however, none of the false-positive clusters contain representatives of both of these motifs. The *rho* enhancer is repressed in the ventral mesoderm by the zinc-finger Snail protein (Ip et al., 1992). The four Snail-binding sites contained in the *rho* enhancer share the consensus sequence, MMMCWTGY; the *vnd* and *brk* enhancers contain multiple copies of this motif and are probably repressed by Snail as well.

The functional significance of the shared sequence motifs was assessed by mutagenizing the sites in the context of otherwise normal *lacZ* transgenes (Fig. 3). Previous studies suggested that bHLH activators are important for the activation of *rho* expression, as *rho-lacZ* fusion genes containing point mutations in several different E-box motifs (CANNTG) exhibited severely impaired expression in transgenic embryos (Ip et al., 1992; Gonzalez-Crespo and Levine, 1993; Jiang and Levine, 1993). However, it was not obvious that the CACATGT motif was particularly significant as it represents only one of five E-boxes contained in the *rho* enhancer. Yet, only this particular E-box motif is significantly over-represented in the *rho*, *vnd* and *brk* enhancers (Table 1). *vnd-lacZ* and *brk-lacZ* fusion genes were mutagenized to eliminate each CACATGT motif, and analyzed in transgenic embryos (Fig. 3B,F). The loss of these sites causes a narrowing in the expression pattern of an otherwise normal *vnd-lacZ* fusion gene (Fig. 3B; compare with A). By contrast, the *brk* pattern is narrower in central and posterior regions, but relatively unaffected in anterior regions (Fig. 3F; compare with E). The *brk* enhancer contains two copies of an optimal Bicoid-binding site, and it is possible that the Bicoid activator can compensate for the loss of the CACATGT motifs in anterior regions (M.M., unpublished).

Similar experiments were performed to assess the activities of the Su(H)-binding sites (YGTGDGAA) and the CTGWCCY motif. Mutations in the latter sequence cause only a slight reduction and irregularity in the activity of the *vnd* enhancer (Fig. 3C), whereas similar mutations nearly abolish expression from the *brk* enhancer (Fig. 3G). Thus, CTGWCCY appears to be an essential regulatory element in the *brk* enhancer, but not in the *vnd* enhancer (see Discussion). Mutations in both Su(H) sites in the *brk* enhancer caused reduced staining of the *lacZ* reporter gene (Fig. 3H), suggesting that Su(H) normally activates expression. Further evidence that Su(H) mediates transcriptional activation was obtained by analyzing the endogenous *rho* expression pattern in transgenic embryos carrying an *eve* stripe 2 transgene with a constitutively activated form of the Notch receptor ($Notch^{IC}$). *rho* expression is augmented and slightly expanded in the vicinity of the

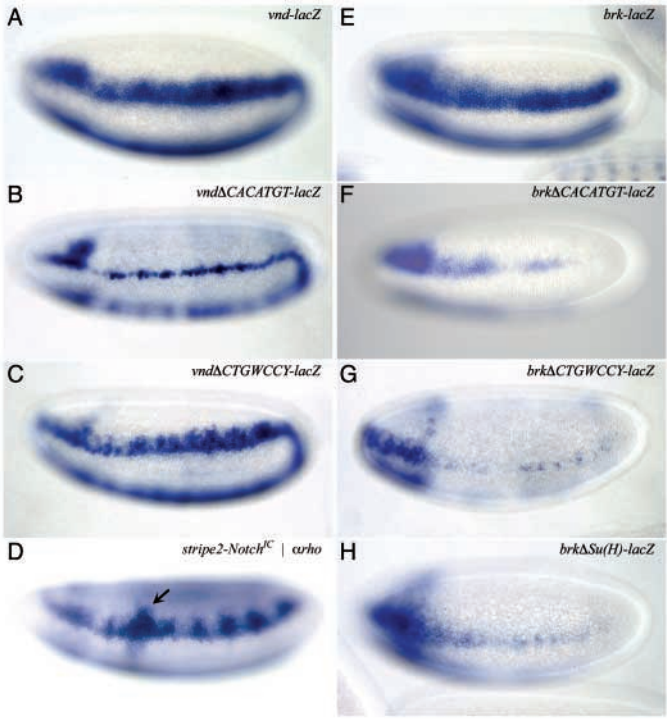


Fig. 3. The shared sequence motifs correspond to essential cis-regulatory elements. The shared sequence motifs in the *vnd* (A-C, 743 bp) and *brk* (E-H, 498 bp) enhancers were mutated as indicated, and the effects on enhancer activity were assayed by in situ hybridization as described in Fig. 1. Ventrolateral views of embryos oriented with anterior to the left and dorsal up are shown. All of the embryos (except D) are undergoing cellularization. (A-C). A larger, more robust *vnd* enhancer than shown in Fig. 2 was used. The wild-type *vnd* enhancer directs lateral stripes of *lacZ* reporter gene expression (A). By contrast, point mutations that eliminate each of the two CACATGT motifs disrupt the activities of an otherwise normal *vnd-lacZ* fusion gene (B). Staining is restricted to the ventral-most regions of the neurogenic ectoderm, similar to the normal *sim* expression pattern (see Fig. 4). Mutations in the three CTGWCCY motifs in the *vnd* enhancer cause subtle changes in the *lacZ* staining pattern, including a slight narrowing and some irregularity in expression (C). (E-H). The embryos express different *brk-lacZ* fusion genes. The wild-type *brk* enhancer directs a staining pattern that is similar to the one produced by the *vnd* enhancer (E, compare with A). Mutations in the two CACATGT motifs disrupt the activities of the *brk* enhancer and cause a loss of *lacZ* staining, especially in the posterior half of the embryo (F, compare with E). Point mutations in the CTGWCCY motifs nearly abolish expression from an otherwise normal *brk-lacZ* fusion gene (G). Finally, mutations in the two Su(H)-binding sites cause a loss of expression in the posterior half of the embryo (H), similar to the altered pattern obtained with mutations in the Twist (CACATGT) binding sites (F). The transgenic embryo in D expresses a *stripe2-Notch*^{IC} fusion gene that causes constitutive activation of Notch signaling in the stripe 2 region. The embryo was hybridized with a digoxigenin-labeled *rho* antisense RNA probe. Expression is slightly expanded in the region where the *stripe2-Notch*^{IC} transgene is active (arrow).

stripe2-Notch^{IC} transgene (Fig. 3D). A similar expansion is observed for the *sim* expression pattern (Cowden and Levine, 2002).

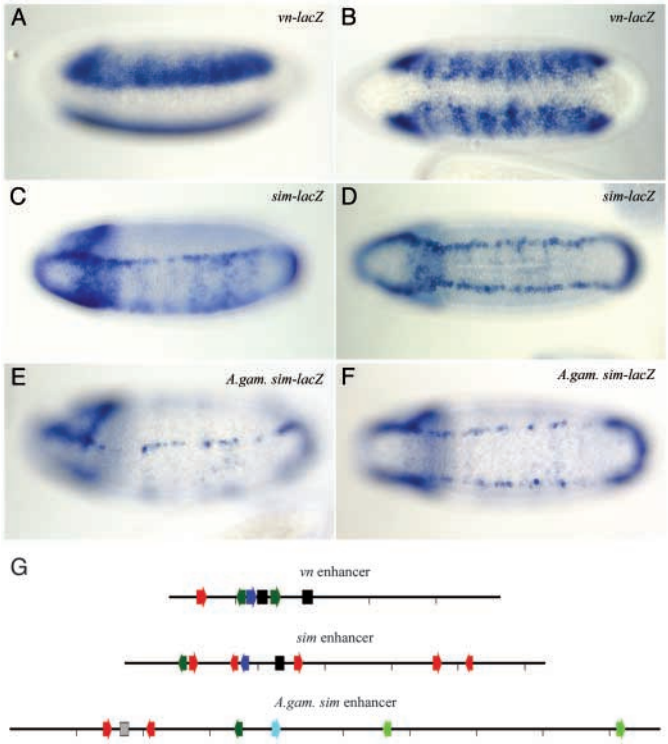


Fig. 4. Expression directed by newly identified fly and mosquito enhancers. The newly identified enhancers for *vn* (497 bp), *sim* (631 bp) and *A. gambiae sim* (976 bp) were fused to *lacZ* reporter genes. Embryos transgenic for these reporter constructs were analyzed by in situ hybridization, as described in Fig. 1. All embryos are depicted with anterior to the left. (A,C,E) Ventrolateral views of cellularizing embryos; (B,D,F) ventral views of gastrulating embryos. The *vn* enhancer drives expression in the ventral neurogenic ectoderm (A,B), similar to *brk*, *vnd* and *rho* (compare with Fig. 2A,C,E). The enhancer is located in the first intron of *vn*. The *sim* enhancer (C,D) drives expression in the mesectoderm, the ventral-most line of cells of the neurogenic ectoderm. The enhancer is located 5′ of the *sim* gene. Weak and variable staining is also detected in more ventral regions of early embryos (C), possibly due to the loss of crucial Snail repressor sites. The *Anopheles sim* enhancer (E,F) drives irregular expression in the mesectoderm, similar to the pattern obtained with the *Drosophila sim* enhancer. The enhancer is located 5′ of a putative *sim* ortholog. The relative arrangement and orientations of sequence motifs in the *vn*, *sim* and *Anopheles sim* enhancers are depicted in G: Dorsal motifs (black boxes), Su(H) motifs (red arrows), CA-Eboxes (CACATGT, dark-blue arrows) and CTGWCCY sites (green arrows). Additionally, the location of a sub-optimal Dorsal site (light gray box), a close relative to the CA-Ebox (CACATGG, light blue arrow), and two close matches to the CTGWCCY motif (CTGNCCY, light green arrows), are shown for the *A. gambiae sim* enhancer.

Identification of the *vein* and *sim* enhancers

To determine whether the shared motifs would help identify additional ventral neurogenic enhancers, the genome was surveyed for 250 bp regions containing an average density of one site per 50 bp and at least one occurrence of each of the four motifs for Dorsal, Twist, Su(H) and CTGWCCY. In total, only seven clusters were identified (see supplemental data at http://dev.biologists.org/supplemental/). Three of the seven clusters correspond to the *rho*, *vnd* and *brk* enhancers. Two of the remaining clusters are associated with genes that are known

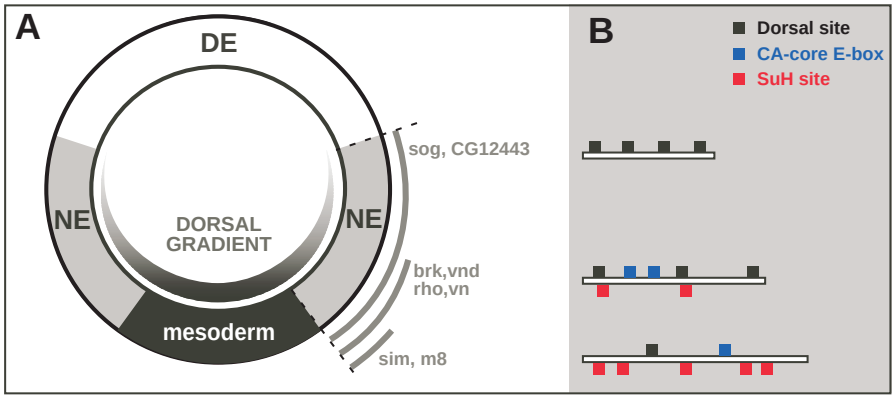


Fig. 5. Model for differential gene expression in the neurogenic ectoderm. (A) Cross-section through a cellularizing embryo. The nuclear Dorsal gradient is shown with peak levels in ventral regions and lower levels in more lateral regions. The presumptive neurogenic ectoderm (NE) exhibits at least three distinct patterns of gene expression: *sim* and *m8* are expressed only in the ventral-most line of cells in the NE, the mesectoderm; *brk*, *vnd*, *rho* and *vn* are expressed in the 5-6 cell wide ventral domain of the NE; and *sog* and *CG12443* are expressed in broad lateral stripes throughout the NE. DE, dorsal ectoderm. (B) A stylized representation of the enhancers active in the NE. Enhancers active in the mesectoderm (e.g. *sim*) contain a large number of Su(H)-binding sites (red boxes), but few optimal dorsal sites (black boxes). By contrast, enhancers that direct broad expression throughout the NE (*sog* and *CG12443*) contain several optimal Dorsal sites, but no Su(H) sites. Enhancers that direct expression in an intermediate pattern, i.e. in ventral regions of the NE (*rho*, *vnd*, *brk* and *vn*), contain a mixture of high-affinity and low-affinity Dorsal sites, as well as a few Su(H) sites. Additionally, CA-Eboxes (CACATGT, blue boxes) and the CTGWCCY motif (not shown) are only found in the mesectodermal and ventral neurogenic ectodermal enhancers, and not in the enhancers driving broad expression in the NE. This implies that genes exhibiting overlapping expression patterns (e.g. *sog* and *brk*) are not activated solely by a gradient of nuclear Dorsal, but also by a variety of transcription factors, and also that they are activated in the same regions by different means.

to be expressed in ventral regions of the neurogenic ectoderm: *vein* and *sim* (Fig. 4A-D) (Kasai et al., 1992; Schnepp et al., 1996). Both clusters were tested for enhancer activity by attaching appropriate genomic DNA fragments to a *lacZ* reporter gene and then analyzing *lacZ* expression in transgenic embryos. The cluster associated with *vein* is located in the first intron, about 7 kb downstream of the transcription start site. The *vein* cluster (497 bp) directs robust expression in the neurogenic ectoderm, similar to the pattern of the endogenous gene (Fig. 4A,B) (Schnepp et al., 1996). The cluster located in the 5′ flanking region of the *sim* gene (631 bp) directs expression in single lines of cells in the mesectoderm (the ventral-most region of the neurogenic ectoderm), just like the endogenous expression pattern (Fig. 4C,D) (Kasai et al., 1992). These results indicate that the computational methods defined an accurate regulatory model for gene expression in ventral regions of the neurogenic ectoderm of *D. melanogaster* (see Discussion).

To assay the generality of our findings, we scanned genomic regions encompassing putative *sim* orthologs from the distantly related dipteran *Anopheles gambiae* for clustering of Dorsal, Twist, Su(H), CTGWCCY and Snail motifs. One cluster located 865 bp 5′ of a putative *sim* ortholog contains one putative Dorsal binding site, two Su(H) sites, three CTGWCCY motifs (or close matches to this motif), a CACATG E-box (Fig. 4G) and several copies of the Snail repressor sequence MMMCWTGY. A genomic DNA fragment encompassing these sites (976 bp) was attached to a minimal *eve-lacZ* reporter gene and expressed in transgenic *Drosophila* embryos (Fig. 4E,F). The *Anopheles* enhancer directs weak lateral lines of *lacZ* expression that are similar to those obtained with the *Drosophila sim* enhancer (Fig. 4E,F; compare with C,D). These results suggest that the clustering of Dorsal, Twist, Su(H) and CTGWCCY motifs constitute an ancient and conserved code for neurogenic gene expression.

Discussion

This study defines a specific and predictive model for the activation of gene expression by intermediate levels of the Dorsal gradient in ventral regions of the neurogenic ectoderm. The model identified new enhancers for *sim* and *vein* in the *Drosophila* genome, as well as a *sim* enhancer in the distant *Anopheles* genome. Five of the seven composite Dorsal-Twist-Su(H)-CTGWCCY clusters in the *Drosophila* genome correspond to authentic enhancers that direct similar patterns of gene expression. This hit rate represents the highest precision so far obtained for the computational identification of *Drosophila* enhancers based on the clustering of regulatory elements (e.g. Berman et al., 2002; Halfon et al., 2002). Nevertheless, it is still not a perfect code.

Two of the seven composite clusters are likely to be false-positives, as they are associated with genes that are not known to exhibit localized expression across the dorsoventral axis. It is possible that the order, spacing and/or orientation of the identified binding sites accounts for the distinction between authentic enhancers and false-positive clusters. For example, there is tight linkage of Dorsal and Twist sites in each of the five neurogenic enhancers. This linkage might reflect Dorsal-Twist protein-protein interactions that promote their cooperative binding and synergistic activities. Previous studies identified particularly strong interactions between Dorsal and Twist-Daughterless (Da) heterodimers (Jiang and Levine, 1993; Castanon et al., 2001). Da is ubiquitously expressed in the early embryo and is related to the E12/E47 bHLH proteins in mammals (Murre et al., 1989). Dorsal-Twist linkage is not seen in one of the two false-positive binding clusters.

The regulatory model defined by this study probably failed to identify all enhancers responsive to intermediate levels of the Dorsal gradient. There are at least 30 Dorsal target enhancers in the *Drosophila* genome, and it is possible that 10 respond to intermediate levels of the Dorsal gradient (e.g. Stathopoulos et al., 2002). Thus, we might have missed half of all such target enhancers. Perhaps the present study defined just one of several ‘codes’ for neurogenic gene expression.

The possibility of multiple codes is suggested by the

different contributions of the same regulatory elements to the activities of the *vnd* and *brk* enhancers. Mutations in the CTGWCCY motifs nearly abolish the activity of the *brk* enhancer, but have virtually no effect on the *vnd* enhancer (see Fig. 3). Future studies will determine whether there are distinct codes for Dorsal target enhancers that respond to either high or low levels of the Dorsal gradient. Indeed, it is somewhat surprising that the *sog* and *CG12443* enhancers essentially lack Twist, Su(H) and CTGWCCY motifs, even though they direct lateral stripes of gene expression that are quite similar (albeit broader) to those seen for the *rho*, *vnd* and *brk* enhancers (see below and Fig. 5).

This study provides direct evidence that Twist and Su(H) are essential for the specification of the neurogenic ectoderm in early embryos. The Twist protein is transiently expressed at low levels in ventral regions of the neurogenic ectoderm (Kosman et al., 1991). SELEX assays indicate that Twist binds the CACATGT motif quite well (K. Senger, unpublished). The presence of this motif in the *vnd*, *brk* and *sim* enhancers, and the fact that it functions as an essential element in the *vnd* and *brk* enhancers, strongly suggests that Twist is not a dedicated mesoderm determinant, but that it is also required for the differentiation of the neurogenic ectoderm. However, it is currently unclear whether the CACATGT motif binds Twist-Twist homodimers, Twist-Da heterodimers or additional bHLH complexes in vivo. Su(H) is the sequence-specific transcriptional effector of Notch signaling (Schweisguth and Posakony, 1992). The restricted activation of *sim* expression within the mesectoderm depends on Notch signaling (Morel and Schweisguth, 2000; Cowden and Levine, 2002); however, the *rho*, *vnd* and *brk* enhancers direct expression in more lateral regions where Notch signaling has not been demonstrated. Nonetheless, mutations in the two Su(H) sites contained in the *brk* enhancer cause a severe impairment in its activity. This observation raises the possibility that Su(H) can function as an activator, at least in certain contexts, in the absence of an obvious Notch signal.

The Dorsal gradient produces three distinct patterns of gene expression within the presumptive neurogenic ectoderm (summarized in Fig. 5A). We propose that these patterns arise from the differential usage of the Su(H) and Dorsal activators. Enhancers that direct progressively broader patterns of expression become increasingly more dependent on Dorsal and less dependent on Su(H) (indicated in Fig. 5B). The *sog* and *CG12443* enhancers mediate expression in both ventral and dorsal regions of the neurogenic ectoderm, and contain several optimal Dorsal sites but no Su(H) sites. By contrast, the *sim* enhancer is active only in the ventral-most regions of the neurogenic ectoderm, and contains just one high-affinity Dorsal site but five optimal Su(H) sites. The reliance of *sim* on Dorsal might be atypical for genes expressed in the mesectoderm. For example, the *m8* gene within the Enhancer of split complex may be regulated solely by Su(H) (e.g. Cowden and Levine, 2002). The *Anopheles sim* enhancer might represent an intermediate between the *Drosophila sim* and *m8* enhancers, as it contains optimal Su(H) sites but only one weak Dorsal site. This trend may reflect an evolutionary conversion of Su(H) sites to Dorsal sites, and the concomitant use of the Dorsal gradient to specify different neurogenic cell types. A testable prediction of this model is that basal arthropods use Dorsal solely for the specification of the mesoderm and Su(H) for the patterning of the ventral neurogenic ectoderm.

We thank Kate Senger and John Cowden for sharing unpublished results; Fred Biemar for advice; Anthony James at UC Irvine for the gift of *Anopheles gambiae* genomic DNA; Hilary Ashe at the University of Manchester for the E2G vector; and Khoa Tran, Austin Luke and Rachel Bernstein for technical assistance. This work was funded by a grant from the NIH (GM46638).

References

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A. et al. (2002). Whole-genome assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301-1310.

Bailey, A. M. and Posakony, J. W. (1995). Suppressor of hairless directly activates transcription of enhancer of split complex genes in response to Notch receptor activity. *Genes Dev.* **9**, 2609-2622.

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 clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* **99**, 757-762.

Bier, E., Jan, L. Y. and Jan, Y. N. (1990). *rhomboid*, a gene required for dorsoventral axis establishment and peripheral nervous system development in *Drosophila melanogaster*. *Genes Dev.* **4**, 190-203.

Casal, J. and Leptin, M. (1996). Identification of novel genes in *Drosophila* reveals the complex regulation of early gene activity in the mesoderm. *Proc. Natl. Acad. Sci. USA* **93**, 10327-10332.

Castanon, I., Von Stetina, S., Kass, J. and Baylies, M. K. (2001). Dimerization partners determine the activity of the Twist bHLH protein during *Drosophila* mesoderm development. *Development* **28**, 3145-3159.

Cowden, J. C. and Levine, M. (2002). The Snail repressor positions Notch signaling in the *Drosophila* embryo. *Development* **129**, 1785-1793.

Davidson, E. H. (2001). *Genome Regulatory Systems: Development and Evolution*. San Diego: Academic Press.

Dehal, P., Satou, Y., Campbell, R. K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D. M. et al. (2002). The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157-2167.

Gloor, G. B., Preston, C. R., Johnson-Schlitz, D. M., Nassif, N. A., Phillis, R. W., Benz, W. K., Robertson, H. M. and Engels, W. R. (1993). Type I repressors of P element mobility. *Genetics* **135**, 81-95.

Gonzalez-Crespo, S. and Levine, M. (1993). Interactions between Dorsal and helix-loop-helix proteins initiate the differentiation of the embryonic mesoderm and neuroectoderm in *Drosophila*. *Genes Dev.* **7**, 1703-1713.

Halfon, M. S., Grad, Y., Church, G. M. and Michelson, A. M. (2002). Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.* **12**, 1019-1028.

Ip, Y. T., Park, R., Kosman, D., Bier, E. and Levine, M. (1992). The Dorsal gradient morphogen regulates stripes of *rhomboid* expression in the presumptive neuroectoderm of the *Drosophila* embryo. *Genes Dev.* **6**, 1728-1739.

Jazwinska, A., Rushlow, C. and Roth, S. (1999). The role of *brinker* in mediating the graded response to Dpp in early *Drosophila* embryos. *Development* **126**, 3323-3334.

Jiang, J. and Levine, M. (1993). Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the Dorsal gradient morphogen. *Cell* **72**, 741-752.

Jiang, J., Kosman, D., Ip, Y. T. and Levine, M. (1991). The Dorsal morphogen gradient regulates the mesoderm determinant twist in early *Drosophila* embryos. *Genes Dev.* **5**, 1881-1891.

Kasai, Y., Nambu, J. R., Lieberman, P. M. and Crews, S. T. (1992). Dorsal-ventral patterning in *Drosophila*: DNA binding of Snail protein to the *single-minded* gene. *Proc. Natl. Acad. Sci. USA* **89**, 3414-3418.

Kosman, D., Ip, Y. T., Levine, M. and Arora, K. (1991). Establishment of the mesoderm-neuroectoderm boundary in the *Drosophila* embryo. *Science* **254**, 118-122.

Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature* **424**, 147-151.

Markstein, M. and Levine, M. (2002). Decoding cis-regulatory DNAs in the *Drosophila* genome. *Curr. Opin. Genet. Dev.* **12**, 601-605.

Markstein, M., Markstein, P., Markstein, V. and Levine, M. (2002). Dorsal binding clusters identify potential target genes in the *Drosophila* embryo. *Proc. Natl. Acad. Sci. USA* **99**, 763-768.

Morel, V. and Schweisguth, F. (2000). Repression by suppressor of hairless and activation by Notch are required to define a row of *single-minded* expressing cells in the *Drosophila* embryo. *Genes Dev*. **14**, 377-388.

Mural, R. J., Adams, M. D., Myers, E. W., Smith, H. O., Miklos, G. L., Wides, R., Halpern, A., Li, P. W., Sutton, G. G., Nadeau, J. et al. (2002). A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**, 1661-1671.

Murre, C., McCaw, P. S., Vaessin, H., Caudy, M., Jan, L. Y., Jan, Y. N., Cabrera, C. V., Buskin, J. N., Hauschka, S. D., Lassar, A. B. et al. (1989). Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. *Cell* **58**, 537-544.

Rubin, G. M. and Spradling, A. C. (1982). Genetic transformation of *Drosophila* with transposable element vectors. *Science* **218**, 348-353.

Rusch, J. and Levine, M. (1996). Threshold responses to the Dorsal regulatory gradient and the subdivision of the primary tissue territories in the *Drosophila* embryo. *Curr. Opin. Genet. Dev*. **6**, 416-423.

Schnepp, B., Grumbling, G., Donaldson, T. and Simcox, A. (1996). Vein is a novel component in the *Drosophila* epidermal growth factor receptor pathway with similarity to the neuregulins. *Genes Dev*. **10**, 2302-2313.

Schweisguth, F. and Posakony, J. W. (1992). Suppressor of Hairless, the *Drosophila* homolog of the mouse recombination signal-binding protein gene, controls sensory organ cell fates. *Cell* **69**, 1199-1212.

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

Stathopoulos, A. and Levine, M. (2002). Dorsal gradient networks in the *Drosophila* embryo. *Dev. Biol.* **246**, 57-67.

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

Thisse, B., el Messal, M. and Perrin-Schmitt, F. (1987). The twist gene: isolation of a *Drosophila* zygotic gene necessary for the establishment of dorsoventral pattern. *Nucl. Acids Res*. **15**, 3439-3453.

Thisse, C., Perrin-Schmitt, F., Stoetzel, C. and Thisse, B. (1991). Sequence-specific transactivation of the Drosophila twist gene by the dorsal gene product. *Cell* **65**, 1191-1201.

White, K., DeCelles, N. L. and Enlow, T. C. (1983). Genetic and developmental analysis of the locus *vnd* in *Drosophila melanogaster*. *Genetics* **104**, 433-448.

SUPPLEMENT

Supplement section overview:

Section I. Primers used in this study

Primers for enhancers defined in this study:

CG12443left(SpeI):	aatactagtGGACAGGACAATGGAGGTTCAGAAGAAGCGAGC
CG12443right(EcoRI):	aatgaattcGGCTGACCACCGTAAATCTGCGGGGGAAATTCCCAGCC
BRK-left(182):	CTCTGGCACAAACCCTAAATGTGGATTACGCT
BRK-right(183):	GGGACAAGTGGCTCCCGCAACGATCGGTTGGA
VND(348)-Lft-(SpeI):	aatactagtGACCCTGGCTTAGAAATTCCCGTAGGTGAG
VND(348)-Rt-(EcoRI):	aatgaattCGCAAATCCGCCTCGTAAGGATCGCAGCTACC
VND(743)-Lft-(SpeI):	ttcactagtGTCCCGGCAAATTGTCATATAATCGAGTGATTAATG
VND(743)-Rt-(EcoRI):	gaagaattcGTAGGTGAGAGCCAGGGAAACCCCAATCGGG
PHM-L(SpeI):	aatactagtGCTGCTCCTGCTTATCGGAGTGATCAGTGTGG
PHM-R(EcoRI):	aatgaattcACGGCGCAGGTTCGAGAAGTGCAATCTACAGG
ADY-Left(186):	CCAAGGTGCAATTTTGTATGCAGTGCGGCT
ADY-Right(187):	CCGAAACTCCCTTGCCAAAAGCCCCACGAT
SIM-RZ F06(SpeI):	taaactagtCCCCGGCATATGTTACGCACATTTACAGCGTATG
SIM-RZ R07(EcoRI):	TAAGAATTCGGTTACAGGCAAACAGCAAACTGAAAACAAATGG
VEIN-topAE:	TTATTGAAAGTGCCGAAGTTAGCGG
VEIN-botAE:	CTGAAACGTGGTTAAAGTGGCC
A.g.SIM-RZ F45(SpeI):	aagactagtACTGTTTACCGAGTGGAAAGGTCCTAACG
A.g.SIM-RZ R45(EcoRI):	tatgaattcATCGCGCCCGCACGATCGACTGCGCAGGAAGC

Primers for false-positive/control clusters C1-C10:

C1-TAK1-LEFT:	CGCGCTACCAAGCCCCCATAGCCACCACCA
C1-TAK1-RIGHT:	CCACACACTCCCCTAGTCTGTGAATTGCA
C2-island1L(SpeI):	aatactagtACTGGAAATCATGTCACGTGCCTGGCGTGCC
C2-island2(RI):	aatgaattcGATTCCCATACCGATTCCAAATGCGAATGCGG
C3-RuntLEFT1(SpeI):	aatactagtGGTTTTCCCGATACTTTTAGGAAATCCCTTTCAGC
C3-RuntRIGHT1(RI):	aatgaattcCGTGATAAGCTATTCTAAATCGAATCGTTGGGG
C4-KrH2-L(SpeI):	attactagtCAGCTTAGAACAAGGGAATTTCCCTCTAGC
C4-KrH2-R(RI):	atagaattcGCCACTTGAGGGAAGATATATGTGGGTATGGG
C5-PpD5-LFT(SpeI):	aatactagtCTTTGAAGCCCACCGACCCATTTGTCCCGCCC
C5-PpD5-RT(RI):	aatgaattcAGGGGGAGCTGCTTTGCGGAGGGGTTCGGA
C6-Fas3-L(SpeI):	aatactagtGACCAAGCTAATCACAAAAGGCTGAGAGAATGG
C6-Fas3-R(RI):	aatgaattcGGTCGAGGCTATGGAAACGCTTCCGATTCGGG
C7-CG5549-L(SpeI):	aatactagtCCTGACCGACGGGAAGGAAAAGTTTGCCGGCAC
C7-CG5549-R(RI):	aatgaattcGATCAAATGAGATGAGTTGAGATGAGATGGCGTTG
C8-Cli-L(Bgl2):	aatagatctCCACGGAGACAAAGACAAAGACAGAGAGTCC
C8-Cli-R(SpeI):	aatactagtGGCTAAGCCGCGGAGGTGTGGAGGGTCGGCCGG
C9-CG1924-L(SpeI):	aatactagtCCCCATGCATTTTACCATCCTTTACCATCGGC
C9-CG1924-R(RI):	aatgaattcTTGTGGCAATCTTGCCAACCATGTGTAGCAC
C10-Ben-L(SpeI):	aatactagtCCCAATGGATAAGGCGTCGGACTTCGGATCC
C10-Ben-R(RI):	aatgaattcAGGCAGCAAGAGCCAACGAACAAAGTTGGG

Primers for mutagenesis:

VNDΔE(CA)-RZ F04mut:	GAAACCCCAATCGGGAATGgatccaaTACGACAGACATGGGACTCAG
VNDΔE(CA)-RZ F05mut:	GCAAGCAAACTTGCGCCAtggaaaaGCACGACCTGTTTCGACCCG
BRKΔE(CA)-(227):	GCCAAAGTTAGAGGCACAGGCtatctaaCGTGTTTGGTTTGAACGGG
BRKΔE(CA)-(229):	GGCCAACCAACGGCtgctaaaGTTCATGTTAGGACCG
VNDΔCTG-RZ F21mut:	cctGTGctgacctGCAgtgtgtgGACCAGGTAGCTGCG
VNDΔCTG-RZ F22mut:	CCCGTAAAGAGTCCCTGgggggggGTGccccccGCAgtgtgtgGACCAGG
BRKΔCTG-RZ F46mut:	CAATTAGCCGGACGGACACAAAaaaaaaaGAGCaaaaaaaGCAAAAAGACACGGTTGTCC
BRKΔCTG-RZ F47mut:	GAAGGAAACTAAAGGAAACTGAGAGAGAGTTCTCGCTATGGATTGTGCAC
BRKΔSu(H)-RZ F44mut:	GCGactagtTGGTTTTTCCCAtttTTACAATTAGCCGGACGG
BRKΔSu(H)-RZ F45mut:	GATGCACCTCTGGGAGATTCCCtttACCGGCAGCAGGTCATG

Section II. Dorsal target enhancer sequences

>ZEN dorsal ectoderm
TCTAGAATGAACGAAAACAGTATCTGGTTTTCCCGAAAATCTTATGAATTTAAAAATGCACTTTATTGCACATACTCACACATGCCTGCCATAAAATATG
ATTCGCGATTTTTCCGCGAACACCCGCGGATCATAAAACATTTGCACCAGCTGCCTGTGTTTATTCACCTACCTGAAACCCATACTCTTATCGCCTGATC
CTCGCGCGGTCGCACTATTTAGGTAGACACTGTACAGGCAGCACTMGCGGGGCGCCGCCAGGCAACTTGTTGCCCCTATTTTCTTTGATATAAGTTTTGG
GAAATCCAGAAGTCCAATAACGGGGCCTATATGAACGAATATTGATTGGGTTTCTCCCAGTTATAGAGTTTTTATTGATCTTGGGCGCGTTTTAGTAGGT
GGATTATGGATGTCCTGGGAAAACCAAGCTCTGAATCCATTCTTACTTATCACAAGACTTGTTTTCCTGTGCCCTTCATCACACCTAGTTTAACTACTAT
AATTAATAAACTACTTTTTAATCCGCCATTTATGAAGATAATTTATAGCATCTATGCATCTTTCAAATATCGCGAGATATCGCAAAGTATTCATAAATTA
GTTTTTTATTGATATTGTAAAGCT

>SOG broad lateral neurogenic ectoderm
GTTGCCAATGCCATTGCGCATACGCCGTGTCGTCTATATGGCTATATGGCTATATGGCTGTATGGTGCGGGGAAATCCCCGTAATCGCAGGTAGAATTCC
AGCCGGTGCCGAGGCGGGACCTGCTCGCACCTCTAATCCCGCCAGGGTTTTCGGGACATGGGATATTCCCGACGGCACAGCATAGCACTCCGTTTTCTTT
TTTTTTTTTTATTATTATTGTGTCCAGTTTTAATCCGGAAAGCGGGAATTCCCTTCCGCTCGCTGCCTGCACTGCGCTGCGCAGACGCATCGGCGTCCGT
AAGCCGCTTACCAAAAAGATACGGGTATACCCAAATGGATGCCTGCCCATGTATATAGACCATTGGGTGGTATGGACCATGGACCATAAAGC

>CG12443 broad lateral neurogenic ectoderm
GGACAGGACAATGGAGGTTCAGAAGAAGCGAGCAAATGCTGGAAAATGCAGTGACAACAGGTGCAAAATTATTTTTTGTTTGTTGCGAGTGCGCGTGAAA
ATTTCCAGCTGGCCAGGGACAGGAATATGACCACTTAAGGCCTAATGTGCGAAAAGTTCCTTTGTCAATTTACACGCCCTCTCCTCACCAAGCGACCGTG
AAAACTTCATTCATTTGCATGGCTAAGCTCAGGTAGCCGGGGATTATCCCTCGTTCTAACCAAAACCTCCTGCTACATTCGGGTTTATCCCACTTGTTTG
GCTGGGAATTTCCCCCGCAGATTTACGGTGGTCAGCCAAATCC

>BRK neurogenic ectoderm
CTCTGGCACAAACCCTAAATGTGGATTACGCTAATATTGCCCCCCCCTAATAAAAACGGTCGTTGTCCAGGGCCGAATATTGCGTCTGATTGGTTTTTCCC
ACGATTACAATTAGCCGGACGGACACAAACTGACCTGAGCTGACCCGCAAAAAGACACGGTTGTCCGGCAGTCGGAACTGAAGGAAACTAAAGGAAACTG
AGGGCAGGTCAGCGCTATGGATTGTGCACTAAGTTGCTTAATCCGACGGGAAATCCAAAACACAACCCGAGCCCGATCCTTCGCTCCTTCGATTTAAGCC
AAAGTTAGAGGCACAGGCACACATGTGTGTTTGGTTTGAACGGGAAAGCCCCATTTTAAAGCTGGCCAACCAACGGCAACACATGTTCATGTTAGGACCG
ATACAGGTTGACATTCCCTGGAAGGATGCACCTCTGGGAGATTCCCACAACCGGCAGCAGGTCATGTCCAACCGATCGTTGCGGGAGCCACTTGTCCC

>RHO NEE neurogenic ectoderm
AAATGGGAAAACATGCGGTGGGAAAAACACACATCGCGAAACATTTGGCGCAACTTGCGGAAGACAAGTGCGGCTGCAACAAAAAGTCGCGAAACGAAAC
TCTGGGAAGCGGAAAAAGGACACCTTGCTGTGCGGCGGGAAGCGCAAGTGGCGGGCGGAATTTCCTGATTCGCGATGCCATGAGGCACTCGCATATGTTG
AGCACATGTTTTGGGGGAAATTCCCGGGCGACGGGCCAGGAATCAACGTCCTGTCCTGCGTGGGAAAAGCCCACGTCCTACCCACGCCCACTCGGTTAC

>VND(348) neurogenic ectoderm
GACCCTGGCTTAGAAATTCCCGTAGGTGAGAGCCAGGGAAACCCCAATCGGGAATGACATGTGTACGACAGACATGGGACTCAGATGCCTTCGAGATACT
GGCGTCACACTGTCTGGCAATGGGATTTCCGCTCAGGAGGACGGGGAATGCCCGTGTAGCCTGTCCATAGCGTGGGAAATTCGCGAGTCGGGGTCTTCGG
GAAAACTCGAAATGGGAAAACCGGAAGCAAGCAAACTTGCGCCAACATGTGGCACGACCTGTTTCGACCCGTAAAGAGTCCCTGCTGACCTGTGCTGACC
TGCACTGACCCGACCAGGTAGCTGCGATCCTTACGAGGCGGATTTGCG

>VND(743) neurogenic ectoderm
GTAGGTGAGAGCCAGGGAAACCCCAATCGGGAATGACATGTGTACGACAGACATGGGACTCAGATGCCTTCGAGATACTGGCGTCACACTGTCTGGCAAT
GGGATTTCCGCTCAGGAGGACGGGGAATGCCCGTGTAGCCTGTCCATAGCGTGGGAAATTCGCGAGTCGGGGTCTTCGGGAAAACTCGAAATGGGAAAAC
CGGAAGCAAGCAAACTTGCGCCAACATGTGGCACGACCTGTTTCGACCCGTAAAGAGTCCCTGCTGACCTGTGCTGACCTGCACTGACCCGACCAGGTAG
CTGCGATCCTTACGAGGCGGATTTGCGTTTAATTGTTGATGGTATTAGGCAAATCAAAACTCGGGGTCTGACCGGGACTAGGTGTCAATAATCCAGCGAT
TTGGGTGCACTTATTCAAAGTTAATTCCGGGGGAAATGTGCGCGTTTTCGGTTCCGAAGCATGCCTGCAGGATGCACACCCCCCACCTCCTTATCTTCTT
AACAACGGCAAGTGCAAAAATCTGTGAAAGTCAGAGCGCTACAGGTAGTGCAGGTAGTTTCCTTTGCATATCCCGACCAACAGGGACCTCCTTTTGTTAA
ACCTTCCGGCCATTCACACGATTGACACAGGATGTCGCTGCAATAAGCATGAAACAGGGAAAAATCGTTCCACGTCCTCTAAGGAGCCATCTTTATACTC
GGGGAGTCATTAATCACTCGATTATATGACAATTTGCCGGGAC

>VN neurogenic ectoderm
TTATTGAAAGTGCCGAAGTTAGCGGGCATTTCACTTACCTGCGTGGGAAAATCGACTAATCTGCGACCGCCCCGAGGAGTCAGTTTTTGTTTTTAGAGCG
GTAAAGGACAGGTAACGGGCCACATGTCTGGCCGGAAATTCCCCGTTGACCCCTGACCCCGTGTCCTTATGACGAATTCGTCACTTGGCGTGAGCACACC
TGGATTTCCCACCGCTTAGCCAGCGGAAATTCCAAAACACCTCCGGCCCACTGGCCCTCAAAATTGTTATATGCTCTGCTACGATGAAGCAGAAGCAGAA
GCAGCAGTGTTTTATTGGCGGAAGCATCCGCCAAATTGCACCCAATCTGCAGTTTGAAGTGCTCAAAACCCCCACCGCTCCCCTGTGAATTTCCGCCGGC
CGGCAAGGTGACCGTGTGCTAAAACAAAATTTTTATATCGAAATTGCCGCCGGTCAACGCGCGCGCCTGCCCAATGGCCACTTTAACCACGTTTCAG

>SIM mesectoderm
CCCCGGCATATGTTACGCACATTTACAGCGTATGGCGATTTTCCGCTTTCCACGGCCACGGCCACAGCTTCCCACCTGATAGGACAGCTCGGCAATGTGT
GGGAATCGCAGTGAGGTGCCGGTAGGAGTGGCAGGTAAGCCTGGCCGCCTCGCAAGTTTCTCACACTTCCAGGACATGTGCTGCTTTTTTGGCCGTTTTT
CCCCGACTGGTTATCAATTGGCCGATTGGAAATTCCCCGATGGCGATGCGCTAGCGTGAGAACATGAGCTGCGAGCATCGGGTTTTTAGCATATCCATAC
CTGTGGCTCGTCCTGATGGGAAGCGAGAAGCAGCAGGATCGGATGTAGGATGCAGGATATAGGGTATAGGCGCTGTTGCGCCTCACCCGCAACACCCACA
TTAGCATCGGACCAGCGTCCAGTGTCCTGTTAATTGCTTTATGGACTCTCCACTTTCCGCTGCGTGGGAATCTTTGCTCATCCTACCTGTTTCCATGCCA
CACCAACCCATTCCCACAGCATTGTCCTCCTTATGTGAAACTCTCTAGTTCAAGTTCAGTGTGAATATTTGTGTTGACTTTATTTTTAAACTTTTGGCCA
TTTGTTTTCAGTTTGCTGTTTGCCTGTAACC

>A.GAM. SIM mesectoderm
ACTGTTTACCGAGTGGAAAGGTCCTAACGAGTGTTTTAGTGCGGTGCAATAAAATCAGGTACGGCACCGATACCGAGCACGTTCGAAGGGTGGGAAATTT
GGAACCGGCCCGGTTCATGGCGGTTGATGATAGGCATAAGCGTGGGAATGAATGAACTTCGGTAAGGATTTTCCCGAAACGGCTCGATGCTGTGGAGCAG
CAAATTCTCACGTTTCCTCAGCCTCGGTCACACACATTCCCCGATGGGTTGCAGTGTAGAGACAAACATTTTTATGCAAAGTAGCATCCTGGCGGAAGCA
ACCTGCGAGGCATAAGATGGTGCCACGCCGCCACTACCTGTCCCGGCGGTTGTTTGGTATGGTTCGAGCGTTAATACCTCTTCAAATATGGTGAACACAT
GGTGACAAGCTGCTTACGCTTTTGGACACAACTAGCCAAGTACCGTTAGTTTTGTGTTTTATTGTACTGCCATGAGATTAGTCGTTGATTTTTATCAAAC
AGCCTATCCTAAGGAAGGCTTATCTGGGAAGTTGTTGTACGTATTTTATAATCCGATCACTGCCCTTTCTCAACACCGATCAATTTGATTGCTGATGCAT
TTGTACACTCAACTATTGCATGTTTGCATAAATACGCCCATAAATGCACTGCCAAAATGTGCACAGTGTGCATGATGGGAAGGCAAACAGTTTCCATTTC
CTTTTCGTCCATCGCTATTTTTCTGCTTTCCTTTCCTCCTGTACCGTACGCACCGCCCATATTTTACAACCGACACCGATCGTGTTGCATTTTCACCTGC
AATACCCATCCCAGGCCACGCCGAGACCTGTGACCTGTTGTCGCCGTCAAACAATCGCAAAATGCAATTACAGTTGAGCTATTGGACGAGCAAGGTGCAA
CGATACGGTTGGGGCAGTGATTGCAGTATTTTCCCAACACACACGCTTCCTGCGCAGTCGATCGTGCGGGCGCGAT

>PHM mesoderm
GCTGCTCCTGCTTATCGGAGTGATCAGTGTGGATGGCCTTGTGAAAGAGGGGGATTACCAAAACTCCCTTTATCAACAGAATCTCGAGTCGAACTCCGCA
ACAGGCGCAACGGCTTCGTTTCCATTCCTGATGCCCAACGTTTCGCCCCAGACCGTAAGTCCAGGTTCCTGAAAAACGCCATACATATACATCATGGAGT
TCGCTCCATATGGATGAACAGATTCGCCCAGCTGCAGCTGTGCGCATACAATTCCAATCAGATTAGAAAAGTCCTCGATATAGGCTCTGGTTTTTCCCGT
ATCTTCCGCGCGGGGAAAACCCATAAGAGTACTGTGCAAATTTTTCCATATGTTCAGTTGTACCCGAATACACACAATTCGCCACCCACATGTAAGGCCC
TGAAACCGAAACCTGTAGATTGCACTTCTCGAACCTGCGCCGT

>ADY mesoderm
CCAAGGTGCAATTTTGTATGCAGTGCGGCTATGCTGGGGGATTTCCCATTTGGCTGCGATTGGCTGCAGGGAAAACCCAGTTTCCAAAAGAGTTCTCTGC
GAAAAACCCCCAATGCAGTGCAGTTTATTTTTTCGATGCGTGGTTTTTCCCACTCCAAACCAGATACCATAATGTATTCTCTTATTTATCGTGGGGCTTT
TGGCAAGGGAGTTTCGG

Section III. False positive cluster/control sequences

False-positive Dorsal clusters of 3 or more occurrences of GGGW(D)WWWCC in 400 bp, which failed to display enhancer activity in the early *Drosophila* embryo

>FP-C1
CGCGCTACCAAGCCCCCATAGCCACCACCACACCACCCTCTGGAATCACCCACTTTCGGGCCCAGATTGCATCCACTTTAAAATCCATTTATGACTGAAA
AGCAACAAAGGGTGGCGCAAAAAGTTCGGATTGGACTTTCAGCTTTGGCTTCTGGGGGTTTTTCCAGCTCTTTTGGGATCGAGGGTTGGTTTCACCCCCC
GGGGGGATGGGGGGTGGGTGGTTCGGGGATTTCCCCGTCCTGCTCATTTCACGCATGCAAAGGGAATGGGAGTCCCCCGGTGCTGAAACCCTTCGTAATG
AGTATTCAAAATGCAATTCACAGACTAGGGGAGTGTGTGG

>FP-C2
ACTGGAAATCATGTCACGTGCCTGGCGTGCCGAAAAACAAAAAAAATAAATAAAAAGGGGGAAAAAAAAAAACAAAAATTGCGTCAACAGTGGCTTAATT
ATTTTTTATTCAGTTTATGTTGCCAGTGCGACTCCTTTTGGCAGTGTCCTGCGTTGGCATTCCTTGGCTTCTTTGTTGGTTTTTCCCGTTGTGCTTCCTG
TTTATGGGATTGCTCGTGTCTACCAGGATCTCTTCTAAGTATATGTTTTGGGATTACCAGGGGGGATTACCATCCCTTTTATGAGTGGAGGGACGAGCGT
AGTGCATACCAATGTCCTGCTCCTTGCTGGACATGTCACTCATTCGGGCTGAAGGACTGAAGTTGCAACTGCCACACTGTGCTACAACTCTGCTGAATTG
TTCCGAAACCAGGAGACCCGAGATGAAGTCGCACTACCATCCGCATTCGCATTTGGAATCGGTATGGGAATC

>FP-C3
GGTTTTCCCGATACTTTTAGGAAATCCCTTTCAGCTTAACTACGAAACATTTGAAAACTATCCAATCCCTGAATCAAAATGAAAAGCCATTTTTTCTTTT
AATTAAAGTTTCCATAATATTTCTTAGATTCTAACTTTTTAGGATTCCACCTTAATGAACTAAAACGCCTATTTTAATTTGGATTTACCCTTTCAAAAAA
TGAACTTCCTATCACAGTTGGCCAGCAGGCCAGCAGAAATCGAAAATTAAATGATTATAAAATGCGACCGTGAATTGCCTCACATACTGAGAATCTAAAT
AAAATTATCAGACCAAAACAATTGGAAAACCCAACCGCGAATTAATATTTTCATTTTAACTTATCTCCGAGCATTAGCGATAACAACTATTTGTGGTTCT
TAGCTGGGGGGGAAATCCAAGTGTCTGGCACTTTCATCAAATAATCTCTTTTCTAATCAATCAAAACAATCGTTGAGTAACCCCAACGATTCGATTTAGA
ATAGCTTATCACGA

>FP-C4
CAGCTTAGAACAAGGGAATTTCCCTCTAGCAAATATGGTTTCCACTCAAATGCCTTTAAATCGGTTGAAACGCGCTCAAAGGGAATTTCCCTCCAATTGG
ATTCAATGGGCATGACATAACCGGCTGTGAGTCCGCATGTGCGTGTGTGTGCGTGTGTTGGAGTACAAGTGCTTCCATTTATTTATTTATTTATTTTTTT
TTTTGGCTGTGGCCAGTGACCCCTTGGTTAACGGTATCCATAGTATTATGTAAGTTGCCTGGGGGATTTCCCCTTACTTCACTGATACTTTTCTATACTC
CATGATGATCTGGTTTTCAATTCTACCCGCAGTCTTTGGCTATTTCCATTATTTTGATGTTAGAATCATTCAACTTATGTATTGTATATATGATGATCTC
TGATGGAGCTTTATTTAAATAGTTTTTTTGTTGTATCGTCAAGTTTGGGGGAGGAACTACTTAACTATGTGGCGCCCATACCCACATATATCTTCCCTCA
AGTGGC

>FP-C5
CTTTGAAGCCCACCGACCCATTTGTCCCGCCCAGAATGTATTTAATTGCAAGCGGCGCTCAAATAGGATTTGTGCATTGTTTTCTGTGGTGGATTTAATG
CATGGCCCGGGGAAAAACCCATTTCCTCGCCCACTCCCGGGTTTGTTTGCATTTTTTCCTCAGCCACCATTCGGAATCTATTTTAATGTCAGTTTAATTC
AATGCCCCAGCGGTTCTCTAATGCGCTGCGGGAGCTTTGAAGGGTTCCCTTGTGCGGCAAATGTAATTAAATTGTGGCCAAAGTGGGATAATCCCCATCG
TTCTTTTCGATATGGGTACTTAGGGCATATATGGGACTATGTGATATGGTTTTAGAAGCAGCATTGCGGTTTAAAATCACACAATGAAATGCGAAGAATG
CAATCAACGCAATTTCACGTGTTACTTGATTTGATGCGACCTAGTTTATAGAAAGTAAGCTATTATTTAAGTGGAAAACCCTCTTTTCGTTTCCATCTAA
ATACCAATTTTAAGGAATTTCAACTCAATCGCCTCACTGGGGACTCAAAGTCGAACAGCTTTCAACCGTCTCTCGAAATCTGAAGCAGTCAAAGGCGCTT
GGCTTATTTCCGTTTCCGCTTGCTGCCATATTTTCCGAACCCCTCCGCAAAGCAGCTCCCCCT

>FP-C6
GACCAAGCTAATCACAAAAGGCTGAGAGAATGGATTTCCCATCCTTTTTTGTTTTGTTCGTTTTCCGCTTTTCCCTGATGACATAATTCGGAGCGCTCGA
ATTTCTGCCGCCTGACTTTACACACTCTGACATAATTAGGGTGAACTACCCCCTGCCACTGGCCATTTAAGTCCCCTATCCGTGAACCCTCTTAAAATGG
CATCGACACACCCCTGAACACTTGTGTGCGCAGCTGGCTGTCTTCATTTAAAAATCAAATGCCATGCGCTGCAATTGTTTAATGAAAACTCCAGTTTTTT
GTGTAGAGTGTAATATAAAATTATGCAAATTCGCCACGCTTTTCAATGAAGTAACAAAGGTGGGAAAGCTGAGTTTTGTTATTAATAAGAAAGGGATGTC
ATACTTTAATAAATGAATGTATTCCAATGCCATTGGAATAGAAAAAAGAAACGAAGTTGCACAACGACTTTTTACACAAAAACTATTTTATTAGCATTGC
AAGTGTCGCCCTAACATTTAAAATTCAAAATCTGCATACACTTTAACTGTTCAAAATATTATAACATTTTTCTCCTTGGTTATTGCGGAGTCTTTTGCAT
TCGCTCTTGAATAAGGTATCGCATTATCTGACAAGACTTACCCTCACCTCAACCGTAGTAAAGTTGGGGTATTCCCCAAGAAGAACTGACATTTATTACG
GCTCCATTTCCCTTTTGAGAAGAGTCCTGTCCTGCATAGAAATGCAAGTCAGTAAAGTCCTGGAAAATCCACAAGGATTTGCTCGGTGATAGCTGAAAGC
TTTCTCCCATTCCCCACATCGAGGTTTGGATTTCCCTTTCGGGGATTTTCCACTCGTAGTTGATTGTTGTCAAAGGTCATGAAATATGCAAACAATATAC
TAGCACATTGTTTGTTTCCGATCCGATACGATTAGCTAATCGAAGCGATAAACTCAGTAGAAAAATCACGGCTGACCTGCTTTCCATTTTTAATGGCATC
GTAGGTCCCTCGCTATTAGTAGATTAATAGTTCGCAGGCTTTTAGTCGAATTGCAAGTGCTGACTTTTCAATTGTCTGATGTTTTGGCAAATGGAAGTAG
CTACGGTTAAATGCTAATCACCAAATTGCCATTTGATTGATAAACTGGCAGACACCCGCTCTCAATCGAAAGGACAAGCCACAAAAACCCGAATCGGAAG
CGTTTCCATAGCCTCGACC

>FP-C7
CCTGACCGACGGGAAGGAAAAGTTTGCCGGCACTTTTCACGCTTTCCTATGGAAACTTTGGCGCTTCCGCCTGCCACCTCTCGATTTTCATTGTCAGGCG
ATTGTGTTTCGACCATTTCGAGCATTTCCACAAAAGTTCGCACAAAGCTAAATGTTTATTTGTACTGCAGTTGGACTGCCCCTCCGCCTGTTTGTTTGTT
TGATCGTTTGTTTCGCCTTAGTTTTTTGTAGTTTTTTTTTTGCGTTTTGTTCAACTTTTCCGGGTTTTTCCCTGTGCGTCGCCGCTGAATCGCATTCTTC
TTGGGCCAGGAATCATTAGGAACTATTTGCTAAGCAAATGCCACTGGCATTGCGGAGTGGGCTGAAATTGTTGCAGCCGGAATGCTGGACCGCAAAAAAA
AACTGAATGGCAATGGGGAAAACCAGGCCAGGAGCAAAGCAACCTTGCTGGAAAATGGCGCTGCGAATTGGTATTGTGGTTTTCCCTCATTTTCCCCATA
TTCCCCCAGCGGCGCTGCCCCCGTGGTGCTGCATTTTCCCTGCTCCCCTGCGAACTGAGATCATCATTGCCGCATTAACAGCCCCATTGGCACACTCACT
TGGCCAAATGCACATTTGACTGGCTAATTACGAGGCTGTCAACAAATTGGCATAAAAAACAACGCCATCTCATCTCAACTCATCTCATTTGATC

>FP-C8
CCACGGAGACAAAGACAAAGACAGAGAGTCCAGCGGACGAAGAAGCGTGCCATGACAATCTTGGGTCAACGGCTGCTTTAACCCTTTCATGGGCAAGAGG
GTTCCCATCACACCACTCACTCACTCAAGAGGCTATGCAGACAAAAATAAATCCTAAGATCATGACAATCATTTAAAGCTTTAATGATGACATGACATAC
AAATTTATGATAAGTTATTTTACTTAAAACTGCATAATCCCAGTACCTATAGCTCGTTAAAATTACGACCTGAAATGAAGCCAAAAAGTGATGCTCCCCA
AGGGTTAACTGGAATCTCCGATCTGAACCTATCTGGCCTGTTAATTGATTGTTACATTCCCCTTGTGCGAGCGAATGGCTGTTTATGTTGGTGGCTGTGT
CCGGACAGCTTGCCCCACACTCCGTGGGATTAACCCCCACCACGGATTAACTCAGAACTCCACCTTCGCCAGCTGCGATGCAGACGTGTTGTTTTTTATG
ATTGCCCATTTTTGACAATGAATAAAAATTATTATAGTGTCGCACGTTTTGGTGGCTGTAATTTAATTAAAACGATGTCGCCACGACCCGCATCACGATC
GCGGCGTGCTCCATGAGGGTTTTCCCGGAATTCAGCACTCCCCCCCCCACCCCCCTGGCGAGCTTCATACAGGTGAGAAATGGTGGACAGTTCGGAAAAGT
GACTAATATATAAAGTACGATTTGTGCGTGCGGGAAAATCCCAGCTGGGAAAAAAGGGATGCGAATGCGGATGGGAAAACAAACGCAGCACTTGAGAAGT
CAGGACTTTTTCGAATATTTGCGCTTGCTTTTGGGAGAACGGAAGTGCGGCAAATACAGCTGGATTCGTTAGGCAAACAAAATCGGAATAAAATCATTTT
TGCCCTTCGTTTAATGCAACTGCCCTCGGTGGCCAGATGGCTGAGGAAAAGTGGAAGGTGGACGAGGACAAGGATGAGGATGAGCAGCCGGGAAAAACTT
GTGCGGCTTCTGTGCCGTGAGCCTGTGAACCTGTTTACACCAAACTCGGAAAGAGAAAGAAAAAACGAACCGACTGAAACCATGAATGAACTGCAACTTG
GACTTGGGCGGGCATTGTCACTGTCGATACAGATATATAGAAATGCACTCAAGGTAACTTCCTACTTGGCCCGTATCCGATTTCATCCCAACCTCTTTTC
CCCTCACAGATTATTGCACCGCCGGCCGACCCTCCACACCTCCGCGGCTTAGCC

>FP-C9
CCCCATGCATTTTACCATCCTTTACCATCGGCTTTAACTTTAGTTTTGCGGCTTTTGGTCATATTTTCAGCTTCAACTTCTTCTTCTTCTTTGGGGT
TTTCCCCCTTTGTTTTTCCGGTTTCTCGGCATTTCTTCTTCTTGCTTTTTTTTCCTGTTTTCGGTTCCGTTTTTCGGCAAAGTTAATTAAAATGCTTAGTT
GATTTCCTAAAGGGATAACCAGTTGGCCGGTGTCTGGGCAAACTCCTTGTTGCTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGCTCTTGTGTGGG
TGTGTTTCTGTACATGTGTGTATGTGTGTTTTTGTATGTGTGTGATTTACATGTGGGGACTCATAAATGTTTTGCGATACTTACATTACTATTGCTTGAA
ATTCGTTCGCACTTTCGTATTCAGTGGGGAAAATCCCGTACGCATTTTCACTTTTCACATAAGAAACTTTTCCAATATGGGGAAATTGAAATGCCATTAT
TTGTGCTACACATGGTTGGCAAGATTGCCACAA

>FP-C10
CCCAATGGATAAGGCGTCGGACTTCGGATCCGAAGATTGCAGGTTCGAGTCCTGTCACGGTCGAAGCTCAGGCTACATTTTTTTTAAATTATATTTTGTT
CGTCCTAGAATATATTAATATGGGAGATTCCCTAGCCCAACCCATTTGTGTAACCTGAGAAATTGGGAATTTGGGACGGCGGTTGCGTAACTGACCGTGT
GGCCCAATGGATAAGGCGTCGGACTTCGGATCCGAAGATTGCAGGTTCGAGTCCTGTCACGGTCGAAGCTCAGGCTATATTTTTTTTAAATTATATTTTG
TTCGTCCTAGAATATATTTATATGGGAGATTCCCTAGCCCAACCCATTTGTGTAACCTGAGAAATTGGGAATTTGGGACGGGGGTTACGTAACCGACCGT
GTGGCCCAATGGATAAGGCGTCGGACTTCGGATCCGAAGATTGCAGGTTCGAGTCCTGTCACGGTCGTACCTCAGTATTTAATTTTTTTTGAACTTATTT
TTCGTTCGTCCTATAATATATTAATATGGGAGATTCCCTAGCCCCACTCATTTGTGTAACCTGAGTGCGGTAAGCAGCAATCGTAACCAATTGGCATACC
GAATTGAAAGATTTATTGGACTTTTACATGGGTCGTCCATGGACGAATCAACATGTGGCTGCCACCGCAAGAAGCCCAACTTTGTTCGTTGGCTCTTGCT
GCCT

Section IV. Bioinformatics

OpenGenomics software, Fly Enhancer (www.flyenhancer.org) and Mosquito Enhancer (www.mosquitoenhancer.org), were used to conduct queries for clustering of specified DNA sequence motifs of the *Drosophila melanogaster* and *Anopheles gambiae* genomes, respectively. Diagrams shown in Figure 1 are adapted from Fly Enhancer output. The OpenGenomics MERmaid program (freely available for download at www.opengenomics.org) identifies n-mers of any length that are present or absent in specified groups of sequences. In this study, we considered two classes of motifs: "exact match" motifs in which every position in the motif is filled by one specific nucleotide (A, C, T, or G), and "fuzzy" motifs in which up to two positions in the motif can be occupied by any of the four nucleotides (denoted by the IUPAC symbol "N").

Section IVa. Results of MERmaid analysis

Table S1. Frequency of shared motifs found in comparisons of *brk*, *vnd*, and *rho* enhancers

	Condition 1			Condition 2			Condition 3		
nMER	A	B	C	A	B	C	A	B	C
6	157	731	0	16	536	0	17	578	0
7	75	2711	38	8	778	2	8	588	3
8	40	3725	1434	4	435	102	1	218	48
9	18	1979	1411	1	141	68	0	61	48
10	10	809	672	0	35	22	–	8	5
11	3	311	281	–	8	7	–	2	2
12	1	118	114	–	1	1	–	1	1
13	0	47	47	–	0	0	–	0	0
14	–	26	26						
15	–	15	15						
16	–	9	9						
17	–	7	7						
18	–	4	4						
19	–	2	2						
20	–	1	1						
21	–	0	0						

Beginning with 6-mers, this table shows the frequencies of all n-mer motifs that meet the following conditions: Condition 1 – motifs occur at least once in any pair of enhancers for *brk*, *vnd* and *rho*; Condition 2 – motifs occur at least once in each of the enhancers for *brk*, *vnd*, and *rho*; and Condition 3 – motifs occur at least twice in each enhancer or any pair of enhancers for *brk*, *vnd* and *rho*.

For each condition, three types of motifs are listed: (A) Exact motifs (non-degenerate); (B) Fuzzy motifs which can have up to two wildcard or "N" positions; and (C) Fuzzy motifs from column B that do not occur three or more times in the 10 false-positive sequences. The numbers in this Table represent all motifs that meet the specified conditions. Several of these motifs are related by shifts and sequence. This table represents raw numbers of motifs, without attempts at alignment.

Table S2. Motif-classes shared between *brk*, *vnd*, and *rho*

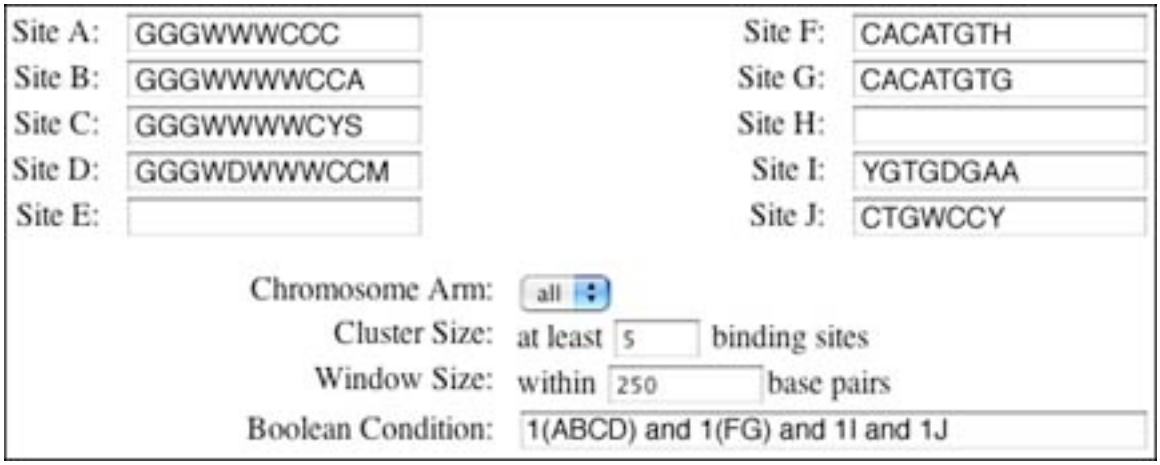
class	feature	aligned motifs	Tot	brk	vnd	rho	FP	genome
1	Extended E-box(CA)	CNCATGTNT	6	2	1	2	0	12583
		CACATGT	10	3	2	1	3	16384
2	Match to Su(H)	CGTGGGAAAWNC	3	1	1	1	0	149
		CGTGGGAAA	3	1	1	1	0	910
3	Match to Dorsal	TGGGAAAWWC	5	1	1	1	0	2622
		CGGGAAWTYC	4	1	1	1	0	1004
		CNGGNAAT	15	2	2	2	0	60582
		CGGGAA	18	2	3	2	6	40668
		GGGCWTTYCC	3	1	1	1	0	1118
		GGGAAAT	19	1	1	1	9	28614
		GGGAAT	28	2	3	1	1	258514
		GGAAWGNC	5	2	2	0	1	26120
		GAAATTC	11	0	2	2	5	29259
4	Dorsal Half-sites	CKNCGGGAA	7	1	2	1	1	4352
		AAATGGG	9	1	1	1	5	30229
		AAYGGGAAARC	3	1	1	1	0	708
		GACGGG	8	1	1	1	1	21962
		GACGGGVMA	5	1	1	1	1	3470
		CCAGGG	8	2	2	0	2	33813
5	Clustered Motifs	GCTGACCYGY	5	2	2	0	1	465
		CTGNCCY	16	4	3	2	9	82829
		CTGWCCY	9	3	3	1	2	33022
6	Palindromic	TTRCGCYAA	3	1	1	1	0	2987
		RANTCGCGA	4	0	2	2	0	4623
7	Other	CCTGWWTCG	3	1	1	1	0	1076
		GNCWTTCCC	4	2	2	0	0	6733
		AAANGAMAC	5	2	0	2	1	21169
		CATTTNGNG	4	1	1	1	1	24568
		ANTTSCGC	9	0	2	3	1	30178
		CGGNNGGA	7	2	0	3	1	35770
		CTGTGC	10	1	1	1	2	55970
		CNGNAGG	11	2	3	0	2	161358

The motifs shown in this table are representative of computer-assisted and manually adjusted alignments of *all* the motifs meeting Conditions 2 or 3 in Table S1. The motifs were manually grouped into seven classes. Motifs in classes 1–3 match consensus sequences for the known transcription factor binding sites Twist, (CACATGT E-box), Su(H), and Dorsal, respectively. Motifs in class 4 include portions of Dorsal half-sites; motifs in class 5 are related to the previously identified CTGNCCY (Stathopoulos et al., 2002) motif. This motif is tightly clustered in the *brk* and *vnd* enhancers. Motifs in class 6 contain partial symmetry; and the remaining motifs meet conditions 2 or 3 but do not have obvious distinguishing features. Within each class, representative motifs are shown. The frequency of each motif in the enhancers for *brk, vnd, rho,* and the false-positive clusters are shown as a sum total and for each enhancer separately. In addition the frequency of each motif in the *D. melanogaster* genome is shown.

Section IVb. Parameters used for *Drosophila melanogaster* genome query

Using Fly Enhancer, the *D. melanogaster* genome was scanned for composite clusters of Dorsal, Twist, Su(H) and CTGWCCY motifs using the parameters below

Input:



Output Summary (for graphics, enter input at www.flyenhancer.org):

```
Chromosome arm:   2L  2R  3L  3R  4    X  all
-------------------------------------------------------------------------
site A (GGGWWWCCC):   447   401   466   566   15  441   2336
site B (GGGWWWWCCA):    693   645   763   872   29  663   3665
site C (GGGWWWWCYS):   1919  1664  1993  2217   47 1769   9609
site D (GGGWDWWWCCM):   734   618   732   795   20  665   3564
site E (WWWWWWWWWCCC):  4633  3696  4888  5514  337 4452  23520
site F (CACATGTH): 2097  1876  2174  2761  100 2020  11028
site G (CACATGTG):  464   444   537   664   19  550   2678
site H (YGTGDGAA): 4039  3618  4264  4989  173 3859  20942
site J (CTGWCCY):    6217  5988  6466  7901  253 6197  33022
-------------------------------------------------------------------------
Clusters of at least
5 sites in 250 bases:               58
-------------------------------------------------------------------------
Clusters satisfying
1(ABCDE) and 1(FG) and 1(H) and 1(J):          12
-------------------------------------------------------------------------
After merging overlapping clusters:             7
```