May 15, 2004

A regulatory code for neurogenic gene expression in the Drosophila embryo

Michele Markstein, University of Massachusetts - Amherst
Robert Zinzen
Peter Markstein
Ka Ping Yee
Albert Erives, et al.

Available at: https://works.bepress.com/michele_markstein/4/
A regulatory code for neurogenic gene expression in the *Drosophila* embryo

Michele Markstein1,*,†, Robert Zinzen1,*, Peter Markstein2, Ka-Ping Yee3, Albert Erives1, Angela Stathopoulos1 and Michael Levine1,†

1Department of Molecular and Cellular Biology, Division of Genetics and Development, 401 Barker Hall, University of California, Berkeley, CA 94720, USA
2Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA
3Computer 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)

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 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 response to the same levels of the Dorsal gradient share a distinctive combination of cis-regulatory elements (Stathopoulos et al., 2002).
The research article discusses the identification and analysis of enhancers in the Drosophila melanogaster genome. 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). 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: CATACGT, YGGTDGAA, 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; Schnepf 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* 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* strain and the derivation of *stripe2-Notch*-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*, *Cl* and *vn*) or digested with restriction enzymes corresponding to restriction sites added to the 5’ ends of each primer pair. PCR products were cloned into pGEM® T-Easy (*brk*, *Ady*, and *Cl*) and digested with *NotI* and cloned into the gypsy-insulated pCaspeR vector E2G (a gift from Hilary Asher), or partially digested with *EcoRI* (*vn*) and cloned into the [-42veulacZ]-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 *BglII*, *SpeI* and *EcoRI* cloning sites in place of *NotI*. Enhancers were manipulated 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 (aagTAAATACGA-CTCAGTTAATAGGAGA) 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
Neurogenic gene expression in *Drosophila*

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/).
The CACATGT and YGTGDGAA motifs bind the known transcription factors, Twist and Suppressor of Hairless [Su(H)], respectively (Thiese 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 1, 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, MMMMCWGTG; 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 (NotchIC). rho expression is augmented and slightly expanded in the vicinity of the

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

<table>
<thead>
<tr>
<th>Motif</th>
<th>zen</th>
<th>sog</th>
<th>CG12443</th>
<th>brk</th>
<th>vnd</th>
<th>rho</th>
<th>vn</th>
<th>sim</th>
<th>Phm</th>
<th>Ady</th>
<th>False positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-box motifs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CANNTG</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>CAAATG</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>CACATGT</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>CACCGT</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>CAGTTG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>CAATGT</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CAGGTG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>CATCTG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>CAGCTG</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>CACATGT</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Su(H) motifs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CGTGGGAA</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TGTGGGAA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CGTGGAGA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YGTGDGAA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Clustered motifs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTGNCCY</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>CTGWCCY</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

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 vnd 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.
Neurogenic gene expression in *Drosophila* stripe2-Notch IC transgene (Fig. 3D). A similar expansion is observed for the *sim* expression pattern (Cowden and Levine, 2002).

**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
to be expressed in ventral regions of the neurogenic ectoderm: vein and sim (Fig. 4A-D) (Kasai et al., 1992; Schnepf 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) (Schnepf 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 gambiæ for clustering of Dorsal, Twist, Su(H), CG12443 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 CG12443 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 CG12443 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; Halton 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 presen 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 \textit{vnd} and \textit{brk} enhancers. Mutations in the CTGWCCY motifs nearly abolish the activity of the \textit{brk} enhancer, but have virtually no effect on the \textit{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 \textit{so} and \textit{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 \textit{rho}, \textit{vnd} and \textit{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 CACATG motif quite well (K. Senger, unpublished). The presence of this motif in the \textit{vnd}, \textit{brk} and \textit{sim} enhancers, and the fact that it functions as an essential element in the \textit{vnd} and \textit{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 CACATG 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 \textit{sim} expression within the mesectoderm depends on Notch signaling (Morel and Schweisguth, 2000; Cowden and Levine, 2002); however, the \textit{rho}, \textit{vnd} and \textit{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 \textit{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 \textit{so} and \textit{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 \textit{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 \textit{sim} on Dorsal might be atypical for genes expressed in the mesectoderm. For example, the \textit{m8} gene within the Enhancer of split complex may be regulated solely by Su(H) (e.g. Cowden and Levine, 2002). The \textit{Anopheles sim} enhancer might represent an intermediate between the \textit{Drosophila sim} and \textit{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 \textit{Anopheles gambiae} genomic DNA; Hilary Ashe at the University of Manchester for the 2EG vector; and Khoa Tran, Austin Luke and Rachel Bernstein for technical assistance. This work was funded by a grant from the NIH (GM46638).

References


