Comprehensive identification of *Drosophila* dorsal–ventral patterning genes using a whole-genome tiling array

Frédéric Biemar*, David A. Nix†, Jessica Piel*, Brant Peterson*, Matthew Ronshaugen*, Victor Sementchenko†, Ian Bell†, J. Robert Manak†, and Michael S. Levine*‡

*Division of Genetics and Development, Department of Molecular Cell Biology, Center for Integrative Genomics, University of California, Berkeley, CA 94720; and †Affymetrix, Inc., Santa Clara, CA 95951

Contributed by Michael S. Levine, June 5, 2006

Dorsal–ventral (DV) patterning of the *Drosophila* embryo is initiated by Dorsal, a sequence-specific transcription factor distributed in a broad nuclear gradient in the precellular embryo. Previous studies have identified as many as 70 protein-coding genes and one microRNA (miRNA) gene that are directly or indirectly regulated by this gradient. A gene regulation network, or circuit diagram, including the functional interconnections among 40 Dorsal target genes and 20 associated tissue-specific enhancers, has been determined for the initial stages of gastrulation. Here, we attempt to extend this analysis by identifying additional DV patterning genes using a recently developed whole-genome tiling array. This analysis led to the identification of another 30 protein-coding genes, including the *Drosophila* homolog of Idax, an inhibitor of Wnt signaling. In addition, remote 5′ exons were identified for at least 10 of the ≈100 protein-coding genes that were missed in earlier annotations. As many as nine intergenic uncharacterized transcription units were identified, including two that contain known microRNAs, miR-1 and -9a. We discuss the potential functions of these recently identified genes and suggest that intronic enhancers are a common feature of the DV gene network.

gene network | microRNA | noncoding RNA

Dorsal–ventral (DV) asymmetry is established by complex interactions of at least 17 maternal genes that produce a localized ligand, Spätzle (Spz), in ventral regions of the perivitelline matrix surrounding the early embryo. Spz induces Toll signaling and the subsequent formation of a broad nuclear gradient of the Dorsal (Dl) protein, the *Drosophila* homolog of NF-κB (1). The Dl nuclear gradient establishes the territories of the prospective mesoderm, neuroectoderm, and dorsal ectoderm by activating or repressing zygotic gene expression in a concentration-dependent manner. Previous genetic screens, subtractive hybridization assays, and microarray analyses identified as many as 70 protein-coding genes that are differentially expressed across the DV axis of early embryos undergoing cellularization and the initial phases of gastrulation. Most of those DV patterning genes encode transcription factors or components of cell signaling pathways, and many are likely to be direct targets of the Dorsal gradient (2).

The advent of whole-genome tiling arrays provides a unique opportunity to identify microRNAs (miRNAs) and other noncoding RNAs that are regulated by the Dl gradient. In addition, these arrays present several opportunities for gene discovery not provided by traditional microarray screens. First, significant genes can be identified by using lower signal-to-noise cutoff values, because neighboring transcription units (TUs) serve as internal controls for even subtle elevations in tissue-specific expression. Second, there is no bias introduced by gene prediction models for the identification of protein-coding sequences. Third, it is possible to identify tissue-specific splicing isoforms for genes that display ubiquitous transcription. Fourth, the detailed visualization of gene structure permits the identification of novel exons. And finally, tiling arrays contain nonprotein coding genes such as those that specify miRNAs. Indeed, miR-1, a mesoderm-specific miRNA, is directly activated by high levels of the gradient in the mesoderm where it influences the activities of genes required for the differentiation of the dorsal vessel, the *Drosophila* heart (3–5). miR-1 expression is regulated by at least two distinct tissue-specific enhancers located in distal and proximal regions of the 5′ flanking region, respectively. The distal enhancer contains a cluster of linked Dorsal and Twist activator sites (4).

The control of DV patterning by the Dl gradient represents one of the best-defined gene regulation networks in metazoan development (6). It therefore provides a good opportunity to assess the role of noncoding genes in embryogenesis. For example, what fraction of all genes engaged in a specific developmental process specify noncoding RNAs? To address this question, we have used a recently developed whole-genome tiling array containing the entire *Drosophila* genome in combination with the same experimental strategy used in a previous study (7). The array contains >3 million 25-mer oligonucleotides covering ≈106 Mb of the fly genome, excluding repetitive DNA, at an interrogation resolution of one oligo approximately every 35 bp. In contrast to previous subtractive hybridization assays and microarray screens, which were restricted to the identification of protein-coding genes, this array permits the unbiased mapping of transcription of both coding and noncoding genes that are selectively expressed in specific tissues across the DV axis of early embryos.

Using this approach, we identified at least 29 additional protein-coding genes that are differentially expressed across the DV axis, thereby bringing the total to ≈100 such genes. At least 10 of the genes contain remote 5′ exons that were missed in earlier annotations. These include *crossveinless-2* (*cv-2*) and *N-cadherin* (*cadN*), which are expressed in the dorsal ectoderm and mesoderm, respectively. Finally, the tiling array identified potential noncoding RNAs, including at least two miRNA genes, miR-1 and -9a, that display restricted expression in the mesoderm or ectoderm. We discuss potential functions for some of the identified protein-coding genes and miRNAs and suggest that the previously uncharacterized 5′ exons help maintain the linkage of TUs with dedicated intronic enhancers.

Conflict of interest statement: No conflicts declared.

Abbreviations: DV, dorsal–ventral; miRNA, microRNA; TU, transcription unit; transfrag, transcribed fragment; CR, computational RNA.

Data deposition: The microarray data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE5434). The .cel files can be accessed at http://transcriptome.affymetrix.com/download/publication/dros_dvpattern_genes.

‡To whom correspondence should be addressed. E-mail: mlevine@berkeley.edu.

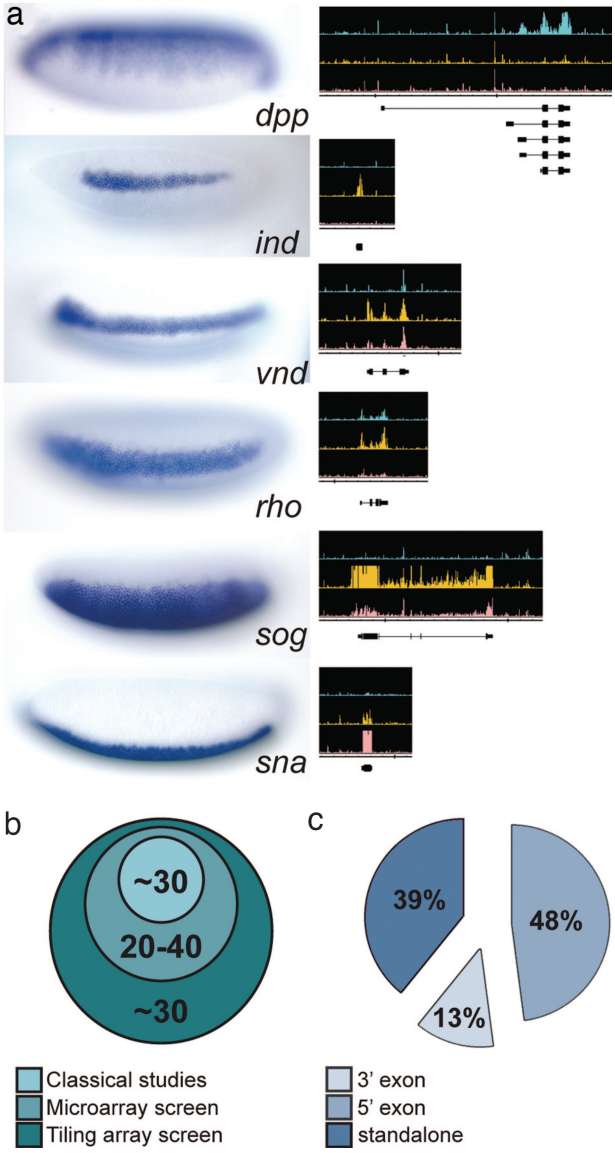

Fig. 1. Identification of Dl targets using a whole-genome tiling array. (*a*) (*Left*) The expression patterns of six previously characterized Dl target genes: *dpp* (dorsal ectoderm); *ind*, *vnd*, *rho*, *sog* (neuroectoderm); and *sna* (mesoderm). Embryos are all oriented with anterior to the left, and dorsal is up. (*Right*) Shown, for each of the six genes, are the RNA signal graphs from three mutant backgrounds as viewed in the Affymetrix Integrated Genome Browser (Affymetrix). The top (blue) graph represents total cellular transcripts from $pipe^{-}/pipe^{-}$ mutants; the middle (orange) graph represents transcripts from $Toll^{rm9}/Toll^{rm10}$ mutant embryos; and the bottom (pink) graph represents transcripts in $Toll^{10B}$ mutant embryos. (*b*) Classical genetic studies previously characterized ≈30 Dl target genes (small circle, pale green). Microarray analysis identified between 20 and 40 additional targets (intermediate circle, light green; ref. 7). In the present study, the unbiased survey of the entire genome using tiling arrays identified as many as ≈30 additional protein-coding genes (large circle, dark green). (*c*) In addition to the protein-coding Dl target genes, 23 uncharacterized transfrags were identified that represent previously unidentified 3′ exons (13%; 3 of 23) or 5′ exons (48%; 11 of 23) of known protein-coding genes. The remaining nine transfrags correspond to putative new genes (39%; 9 of 23).

Results and Discussion

The Dl nuclear gradient differentially regulates a variety of target genes in a concentration-dependent manner (summarized in Fig. 1*a*). The gradient generates as many as five different thresholds of gene activity, which define distinct cell types within the presumptive mesoderm, neuroectoderm, and dorsal ectoderm. As done previously (7), total RNA was extracted from embryos produced by three different maternal mutants: $pipe^{-}/pipe^{-}$, $Toll^{rm9}/Toll^{rm10}$, and $Toll^{10B}$. $pipe^{-}/pipe^{-}$ mutants completely lack Dl nuclear protein and, as a result, overexpress genes that are normally repressed by Dl and restricted to the dorsal ectoderm. For example, the *decapentaplegic* (*dpp*) TU is strongly "lit up" by total RNA extracted from $pipe^{-}/pipe^{-}$ mutant embryos (Fig. 1*a*; blue graph, *Top Right*). The intron-exon structure of the transcribed region is clearly delineated by the hybridization signal, most likely because the processed mRNA sequences are more stable than the intronic sequences present in the primary transcript. There is little or no signal detected with RNAs extracted from $Toll^{rm9}/Toll^{rm10}$ (neuroectoderm; orange graph) and $Toll^{10B}$ (mesoderm; pink graph) mutants. Instead, these other mutants overexpress different subsets of the Dl target genes. For example, $Toll^{rm9}/Toll^{rm10}$ mutants contain low levels of Dl protein in all nuclei in ventral, lateral, and dorsal regions. These low levels are sufficient to activate target genes such as *intermediate neuroblasts defective* (*ind*), *ventral neuroblasts defective* (*vnd*), *rhomboid* (*rho*), and *short gastrulation* (*sog*) but insufficient to activate *snail* (*sna*; Fig. 1*a*). In contrast, $Toll^{10B}$ mutants overexpress genes (e.g., *sna*) normally activated by peak levels of the Dl gradient in ventral regions constituting the presumptive mesoderm.

To identify potential Dl targets, ranking scores were assigned for the six possible comparisons of the various mutant backgrounds, *pipe* vs. $Toll^{rm9}/Toll^{rm10}$, *pipe* vs. $Toll^{10B}$, $Toll^{rm9}/Toll^{rm10}$ vs. $Toll^{10B}$, $Toll^{rm9}/Toll^{rm10}$ vs. *pipe*, $Toll^{10B}$ vs. $Toll^{rm9}/Toll^{rm10}$, and $Toll^{10B}$ vs. *pipe*, using the TiMAT software package (see *Materials and Methods*). As a first approximation, only hits with a median fold difference of 1.5 and above were considered. For further analysis, we selected the top 100 TUs for each of the comparisons, with the exception of $Toll^{rm9}/Toll^{rm10}$ vs. *pipe* for which the TiMAT analysis returned only 43 hits that meet the cutoff (see Tables 1–6, which are published as supporting information on the PNAS web site). To refine our search for TUs specifically expressed in the mesoderm, where levels of nuclear Dl are highest, we selected only those present in the $Toll^{10B}$ vs. $Toll^{rm9}/Toll^{rm10}$ and $Toll^{10B}$ vs. *pipe*, but not *pipe* vs. $Toll^{rm9}/Toll^{rm10}$ comparisons. For TUs induced by intermediate and low levels of nuclear Dl in the neuroectoderm, we selected those present in both the $Toll^{rm9}/Toll^{rm10}$ vs. $Toll^{10B}$ and $Toll^{rm9}/Toll^{rm10}$ vs. *pipe*, but not *pipe* vs. $Toll^{10B}$ comparisons. For TUs restricted to the dorsal ectoderm, only those present in the *pipe* vs. $Toll^{rm9}/Toll^{rm10}$ and *pipe* vs. $Toll^{10B}$, but not $Toll^{rm9}/Toll^{rm10}$ vs. $Toll^{10B}$, were selected. Finally, the TUs corresponding to annotated genes already identified in the previous screen were eliminated to focus on annotated genes not previously considered as potential Dorsal targets (Table 7, which is published as supporting information on the PNAS web site), as well as transcribed fragments (transfrags) not previously characterized (uncharacterized transfrags; Table 8, which is published as supporting information on the PNAS web site). Using these criteria, we identified 45 previously annotated protein-coding genes (Table 7), along with 23 uncharacterized transfrags (Fig. 1*c*). Of the 45 protein-coding genes, 29 exhibited localized patterns of gene expression across the DV axis (Fig. 1*b*), whereas the remaining 16 were not tested (Table 7).

The previous microarray screen relied on high cutoff values for the identification of authentic DV genes (7). For example, only genes exhibiting 6-fold up-regulation in $pipe^{-}/pipe^{-}$ mutant embryos were tested by *in situ* hybridization for localized expression in the dorsal ectoderm. Many other genes displayed >2-fold up-regulation but were not explicitly tested for localized expression. The whole-genome tiling array permitted the use of much lower cutoff values (Table 7*A*). For example, CG13800, which was identified by conventional microarray screens, falls

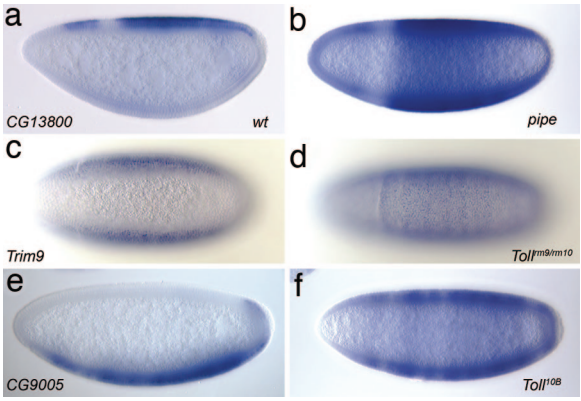

Fig. 2. Examples of protein-coding genes. Cellularizing embryos are all oriented with anterior to the left and represented in lateral (*a*, *b*, *e*, and *f*; dorsal is up) or ventral (*c* and *d*) views. (*a* and *b*) CG13800 is expressed in the dorsal ectoderm in WT embryos (*a*) and expands along the entire DV axis in $pipe^{-}/pipe^{-}$ mutants (*b*). (*c* and *d*) *Trim9* is restricted to the neuroectoderm in WT embryos (*c*) and shows expansion in adjacent territories in $Toll^{rm9}/Toll^{rm10}$ mutants (*d*). (*e* and *f*) CG9005 is present only in the mesoderm in WT embryos (*e*), but its expression spans the entire DV axis in $Toll^{10B}$ mutants (*f*).

just below the original cutoff value but displays 5-fold up-regulation in $pipe^{-}/pipe^{-}$ mutants in our analysis. *In situ* hybridization assays reveal localized expression in the dorsal ectoderm (Fig. 2*a*). This pattern is greatly expanded in embryos derived from $pipe^{-}/pipe^{-}$ mutant females (Fig. 2*b*), as expected for a gene that is either directly or indirectly repressed by the Dl gradient. Genes exhibiting even lower cutoff values were also found to display localized expression. Among these genes is a Wnt homologue, *Wnt2*, which is augmented only 2.25-fold in mutant embryos lacking the Dl nuclear gradient.

The 4-fold cutoff value used in the previous screen for candidate protein-coding genes expressed in the neuroectoderm also excluded genes expressed in this tissue (Table 7*B*). The *Trim9* gene exhibits just a 2-fold increase in mutant embryos derived from $Toll^{rm9}/Toll^{rm10}$ females. Nonetheless, *in situ* hybridization assays reveal localized expression in the neuroectoderm of WT embryos (Fig. 2*c*). As expected, expression is expanded in $Toll^{rm9}/Toll^{rm10}$ mutant embryos (Fig. 2*d*). Another gene, CG9973, displays just 1.8-fold up-regulation but is selectively expressed in the neuroectoderm (data not shown). CG9973 encodes a putative protein related to Idax, an inhibitor of the Wnt signaling pathway (Fig. 5, which is published as supporting information on the PNAS web site). Idax inhibits signaling by interacting with the PDZ domain of Dishevelled (Dsh), a critical mediator of the pathway (8, 9). As mentioned above, a *Wnt2* homologue is selectively expressed in the dorsal ectoderm. Recent studies identified a second *Wnt* gene, *WntD*, which is expressed in the mesoderm (10, 11). Thus, the CG9973/Idax inhibitor might be important for excluding Wnt signaling from the neuroectoderm. Such a function is suggested by the analysis of Idax activity in vertebrate embryos (12).

Additional genes were also identified that are specifically expressed in the mesoderm. Among these is CG9005, which encodes an unknown protein that is highly conserved in different animals, including frogs, chicks, mice, rats, and humans (data not shown). It displays <2-fold up-regulation in $Toll^{10B}$ embryos but is selectively expressed in the ventral mesoderm of WT embryos (Fig. 2*e*). Expression is expanded in embryos derived from $Toll^{10B}$ mutant females (Fig. 2*f*).

Other protein-coding genes were missed in the previous screen because they were not represented on the *Drosophila* Genome Array used at the time. These include, for instance, CG8147 in the dorsal ectoderm and CG32372 in the mesoderm (see Table 7).

An interesting example of the use of tiling arrays to identify tissue-specific isoforms is seen for the *bunched* (*bun*) TU. *bun* encodes a putative sequence-specific transcription factor related to mammalian TSC-22, which is activated by TGFβ signaling. It was shown to inhibit Notch signaling in the follicular epithelium of the *Drosophila* egg chamber (13, 14). Three transcripts are expressed from alternative promoters in *bun*, but it appears that only the short isoform (bun-RC) is specifically expressed in the dorsal ectoderm. A number of *bun* exons are ubiquitously transcribed at low levels in the mesoderm, neuroectoderm, and dorsal ectoderm. However, the 3′-most exons are selectively up-regulated in $pipe^{-}/pipe^{-}$ mutants (data not shown). It is conceivable that Dpp signaling augments the expression of this isoform, which in turn, participates in the patterning of the dorsal ectoderm.

In addition to protein-coding genes, the tiling array also identified uncharacterized TUs not previously annotated (Table 8). Some of them are associated with ESTs, providing independent evidence for transcriptional activity in these regions. For 14 of these transfrags (61%), visual inspection of neighboring loci using the Integrated Genome Browser (see *Materials and Methods*) suggested coordinate expression of a neighboring protein-coding region (i.e., overexpressed in the same mutant background). Two such examples are represented in Fig. 3. The *N-Cadherin* gene (*CadN*) has a complex intron-exon structure consisting of ≈20 different exons (Fig. 3*a*). The strongest hybridization signals are detected within the limits of exons, but an unexpected signal was detected ≈10 kb upstream of the 5′-most exon (red horizontal arrow, Fig. 3*a*). It is specifically expressed in the mesoderm, suggesting that it represents a previously unidentified 5′ exon of the *CadN* gene. Support for this contention stems from two lines of evidence. First, *in situ* hybridization using a probe against the 5′ exon detects transcription in the presumptive mesoderm, the initial site of *CadN* expression (Fig. 3*c*). Second, using primers anchored in the 5′ transfrag as well as the first exon of *CadN*, we obtained confirmation by RT-PCR that the recently identified TU is part of the *CadN* transcript (data not shown). This recently identified 5′ exon appears to contribute to the 5′ leader of the *CadN* mRNA. It is possible that this extended leader sequence influences translational efficiency as seen in yeast (15). Because there seems to be a considerable lag between the time when *CadN* is first transcribed and the first appearance of the protein, we suggest that this extended leader sequence might inhibit translation. An interesting possibility is that it does so through short upstream ORFs, as has been shown for several oncogenes in vertebrates (16–18).

A 5′ exon was also identified for *crossveinless-2* (*cv-2*), a component of the Dpp bone morphogenetic protein (BMP) signaling pathway. *cv-2* binds BMPs and functions as both an activator and inhibitor of BMP signaling. It is specifically required in the developing wing disk to generate peak Dpp signaling in the presumptive crossveins. *cv-2* is also expressed in the dorsal ectoderm of early embryos, but its role during embryonic development has not been investigated (19). The whole-genome tiling array identified a 5′ exon located ≈10 kb 5′ of the transcription start site of the *cv-2* TU (Fig. 3*b*). Using RT-PCR and *in situ* hybridization assays, we confirmed that the exon is part of the *cv-2* transcript (data not shown and Fig. 3*c*). It is possible that the exon resides near an embryonic promoter that is inactive in the developing wing discs. Future studies will determine whether this 5′ exon influences the timing or levels of Cv-2 protein synthesis.

In addition to the identification of 10 5′ exons associated with previously annotated genes such as *CadN* and *cv-2*, three other transfrags appear to correspond to 3′ exons, and nine of the

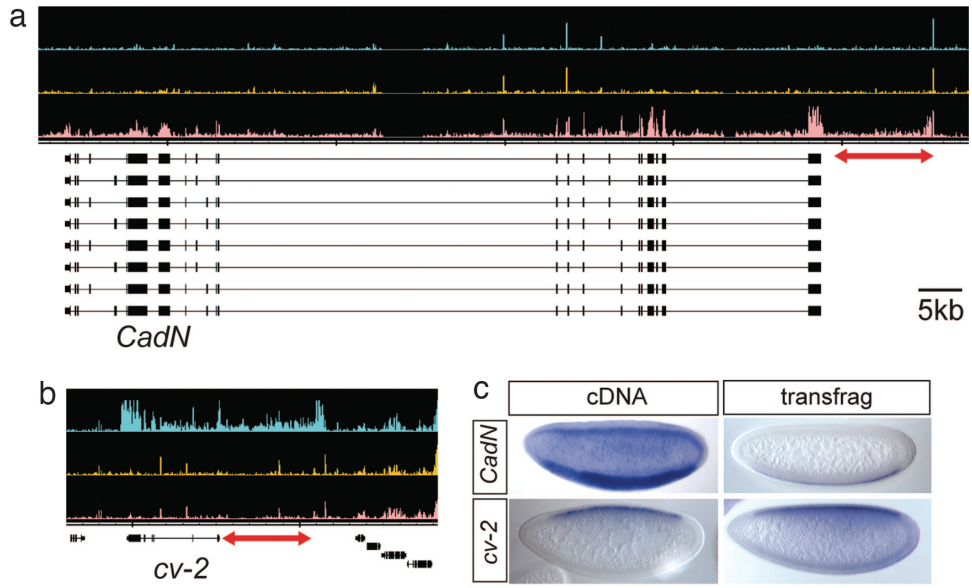

Fig. 3. Uncharacterized transfrags often correspond to novel 5′exon of known protein-coding genes. (*a* and *b*) RNA signal graphs from the three mutant backgrounds for the *CadN* (*a*) and *cv-2* (*b*) loci, suggesting extended transcription (red double arrows) 5′ of the known transcription start site; both genes are transcribed from the minus strand. (*c*) Cellularizing embryos hybridized with riboprobes directed against the cDNA (*Left*) or the recently identified transfrag (*Right*) of *CadN* (*Upper*) and *cv-2* (*Lower*). Expression is detected in the mesoderm and dorsal ectoderm, respectively. All of the embryos are oriented with anterior to the left, and dorsal is up.

RNAs seem to arise from autonomous TUs (Table 8). Three of these represent annotated computational RNA (CR) genes: CR32777, CR31972, and CR32957. CR32777 corresponds to *roX1*, which is ubiquitously expressed at the blastoderm stage, hence it represents a false positive (20, 21). The other two potential noncoding RNAs were recently identified independently in two other studies, and although the expression of CR32957 could not be detected by *in situ* hybridization (22), CR31972 transcripts are detected in the mesoderm (ref. 23; Table 8). There is no evidence that these transcripts are processed into miRNAs, but noncoding genes corresponding to known miRNA loci were also identified in the screen. Transfrag 22 corresponds to the miR-9a primary transcript (*pri-mir9a*) and is detected in both the dorsal- and neuroectoderm (Fig. 4*a*). Expression of *pri-mir9a* is ubiquitous in embryos derived from $pipe^{-}/pipe^{-}$ or $Toll^{rm9}/Toll^{rm10}$ females (data not shown and Fig. 4*b*). Transfrag 8 corresponds to *pri-mir1*, which is present in the mesoderm (Fig. 4 *c* and *d*).

A third noncoding transcript (Transfrag 12) maps next to a known miRNA, miR-184. It is selectively expressed in the mesoderm (Fig. 4*e*) and overexpressed in $Toll^{10B}$ mutants (Fig. 4*f*). The mesodermal expression of miR-184 was reported recently (24). It is possible that Transfrag 12 corresponds to *pri-mir-184*, and that secondary structures in the miRNA region preclude detection on the array. This is seen for several other miRNA precursors expressed at various stages during embryogenesis (J.R.M., unpublished results). Alternatively, Transfrag 12 might represent the fragment resulting from Drosha cleavage of the *pri-mir-184* to produce the miR-184 precursor hairpin (*pre-miR-184*). A similar situation has been observed for the *iab4* locus (25, 26). Like miR-1, miR-184 is selectively expressed in the ventral mesoderm. It will be interesting to determine whether the two miRNAs jointly regulate some of the same target mRNAs.

The identity of the last three transfrags is less clear. Visual inspection using the Integrated Genome Browser suggests expression of Transfrag 10 in the mesoderm, Transfrag 21 in the neuroectoderm, and Transfrag 11 in both the dorsal ectoderm and neuroectoderm. However, *in situ* hybridization assays confirm the predicted expression pattern only for Transfrag 11 (data not shown). Computational analyses designed to estimate the likelihood of translation (see *Materials and Methods*) suggest a protein-coding potential for Transfrag 10 [Likelihood Ratio Test (LRT) $P < 0.001$] and possibly Transfrag 11 (LRT $P < 0.01$), whereas Transfrag 21 could not be analyzed because of lack of conservation in other *Drosophila* species (Table 8 and Fig. 6, which is published as supporting information on the PNAS web site).

In this work, we describe an attempt to identify nonprotein coding genes involved in patterning the DV axis of the *Drosoph-*

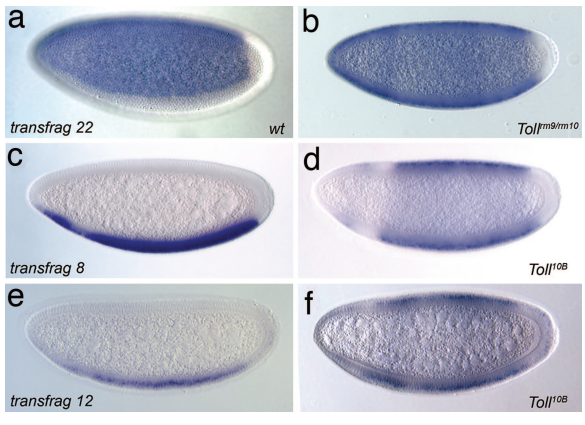

Fig. 4. Examples of noncoding transfrags. Cellularizing embryos are all oriented with anterior to the left and dorsal up. (*a* and *b*) *transfrag 22*/*pri-mir-9a* is expressed in both the dorsal and neuroectoderm in WT embryos (*a*) and expands along the entire DV axis in $pipe^{-}/pipe^{-}$ (not shown) and $Toll^{rm9}/Toll^{rm10}$ mutants (*b*). (*c* and *d*) *transfrag 8*/*pri-mir-1* is specifically expressed in the mesoderm in WT embryos (*c*) and shows expansion in adjacent territories in $Toll^{10B}$ mutant embryos (*d*). (*e* and *f*) Similarly, *transfrag 22* is present only in the mesoderm in WT embryos (*e*) but expands along the entire DV axis in $Toll^{10B}$ mutant embryos (*f*).

ila embryo using an unbiased approach to survey the entire genome. This study, along with earlier analyses, identified as many as 100 protein-coding genes and five to seven noncoding genes that are differentially expressed across the DV axis of the early *Drosophila* embryo. Roughly half of the noncoding RNAs correspond to miRNAs, although <1% of the annotated genes in the *Drosophila* genome encode miRNAs (27, 28). Future studies will determine how these RNAs impinge on the DV regulatory network.

Recent studies have identified large numbers of noncoding transcripts in the mouse and human genomes (29–38). If the present study is predictive, less than one-fourth of the transcripts correspond to novel noncoding RNAs of unknown function, akin to CR31972 and Transfrag 11 expressed in the mesoderm and ectoderm, respectively. Most of the noncoding transcripts are likely to derive from intronic sequences because of the occurrence of cryptic remote 5′ exons as seen for the *CadN* and *cv-2* genes. At least 10% of the DV protein-coding genes were found to contain such exons. As a result, these genes contain large tracts of intronic sequences that might encompass regulatory DNAs such as tissue-specific enhancers. The FGF8-related gene, *thisbe* (*ths*), represents such a case. A neurogenic-specific enhancer that was initially thought to reside 5′ of the TU actually maps within a large intron because of the occurrence of a remote 5′ exon (39). We suggest that such exons are responsible for the evolutionary "bundling" of genes and their associated regulatory DNAs. Gene duplication events are more likely to retain this linkage when regulatory DNAs map within the TU. In contrast, enhancers mapping in flanking regions can be uncoupled from their normal target gene by chromosomal rearrangements.

Materials and Methods

***Drosophila* Stocks.** The following mutant stocks were used: $Toll^{10B}$, $Toll^{rm9}/Toll^{rm10}$, and $pipe^{386}/pipe^{664}$. WT embryos were obtained from the yw^{67} strain.

Whole-Genome Tiling Array. Total RNA was extracted from $pipe^{386}/pipe^{664}$, $Toll^{rm9}/Toll^{rm10}$, and $Toll^{10B}$ mutant embryos, as described (7). First-strand cDNA synthesis and subsequent treatments were described previously (4).

Analysis of Tiling Microarray Data. Processing of the microarray data were performed in three basic steps using TiMAT (http://bdtnp.lbl.gov/TiMAT): data normalization, sliding window summary statistics, and enriched region identification. To normalize the data, all cel files were grouped together, and the perfect match intensities were quantile-normalized and median-scaled to 100. Mismatch intensities were discarded. To identify regions enriched relative to each other, all pairwise comparisons were made between *pipe*, $Toll^{rm9}/Toll^{rm10}$, and $Toll^{10B}$ data (i.e., pipe vs. rm9/rm10, pipe vs. 10B, rm9/rm10 vs. 10B, rm9/rm10 vs. pipe, 10B vs. rm9/rm10, and 10B vs. pipe). Cel files for a particular pairing were divided into treatment and control. Their intensities were mapped to the genome, and a ratio score was calculated for each oligo by dividing the average treatment by the average control. To minimize noise, a sliding window of 675 bp, containing ≈19 oligos, was advanced, one oligo at a time, across each chromosome (similar results were obtained by using a window of 250 bp containing seven oligos). A trimmed mean of the grouped oligo ratios was used to score each window. To collapse overlapping windows into enriched regions, windows that (*i*) intersect by >100 bp, (*ii*) exceed a low threshold of 1.25×, and (*iii*) contain more than five oligos were joined. An enrichment score (median fold difference) for each interval was calculated by identifying the best 225-bp subwindow within the interval based on the median of the associated oligo ratio scores. The intervals were ranked by using this enrichment score.

Computational Analysis of Likelihood of Translation. A strategy similar to the one described by Tupy *et al.* (22) was used to establish a likelihood of translation for previously unannotated transfrags. A 500-bp-long sequence from the second exon of the *even-skipped* (*eve*) gene was used as a positive control for protein-coding potential. First, we asked whether the longest ORF in each transfrag exceeds the median ORF length in 10,000 randomizations of that sequence. In addition, we used conservation in three other *Drosophila* species (*Drosophila ananassae*, *Drosophila pseudoobscura*, and *Drosophila virilis*) to ask whether evolution of transfrag sequences was best described by constraint associated with translation. Orthologous intergenic regions were assigned in each species by a synteny-based method anchored on orthologous gene models determined by a modified reciprocal blast approach (Venky Iyer, University of California, Berkeley; http://rana.lbl.gov/~venky/annotation). Orthologous region pairs [*Drosophila melanogaster* (*D. mel*/*D. ananassae* (*D. ana*), *D. mel*/*D. pseudoobscura* (*D. pse*), and *D. mel*/*D. virilis* (*D. vir*)] for each transfrag were exhaustively searched for most similar ORF pairs by three-frame translation and all-by-all Needleman–Wunsch pairwise alignment. Likelihood ratio tests were performed comparing likelihoods, computed using PAML 3.15 (40), for sequences evolving under fixed K_a/K_s of ($\omega = 1$; no constraint on putative amino acid changes) vs. likelihood of sequences evolving under variable K_a/K_s ($\omega < 1$; sequence under purifying selection) (41). Significance was assigned to sequences with two or more pairwise likelihood ratio tests with $P < 0.01$.

Whole-Mount *in Situ* Hybridization. All probe templates were obtained from PCR-amplified genomic fragments cloned into pGEM T-Easy vector (Promega). PCR primers were derived by using Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi); a list of primers used is available upon request. For each template, both sense and antisense RNA probes were *in vitro*-transcribed by using T7 or SP6 RNA polymerase and digoxigenin-UTP (Roche Molecular Biochemicals). Embryos were collected for 2 h and aged for an additional 2 h. Fixed embryos were hybridized with the riboprobes as described (42).

RT-PCR Analysis. Total RNA from 2- to 4-h WT embryo collections was isolated by using TRIzol reagent (Invitrogen). Extracted RNA was treated with RNase-free DNase I (Ambion, Austin, TX) for 30 min at 37°C and purified by using the RNeasy Mini kit (Qiagen, Valencia, CA). RT-PCR was performed by using the Supersript One Step RT-PCR kit (Invitrogen). Nested PCR was performed with internal primers on a diluted template from the first round (1:100) using Platinum Taq (Invitrogen). Individual PCR products were gel-extracted (Qiagen), cloned into the pGEM T-Easy vector (Promega), and sequenced. Sequences were analyzed by using vector NTI (Invitrogen) and GENEPALETTE (43); www.genepalette.org). A list of the primers used is available upon request.

Protein Alignment and Phylogenetic Inference. Idax and Idax-related protein sequences used in alignment and phylogenetic reconstruction were gathered from METAZOME, Ver. 1.1 (www.metazome.net). Alignments were performed by using CLUSTALX (43) on the two clusters most related to the CG9973 zinc finger. Phylogenetic relationships were inferred by using maximum likelihood (ML) from a 48-aa alignment containing the zinc-finger domains. Support for ML trees used quartet-puzzling reliability values from 10,000 puzzling steps. The quartet-puzzling ML analysis was performed with TREE-PUZZLE (44). Accession numbers for sequences may be obtained from METAZOME, Ver. 1.1. The putative CG9973 homo-

logues (labeled as Idax) constitute cluster ID 1910033, and the closely related CXXC5-labeled proteins are members of cluster ID 1907992.

We thank Robert Zinzen, Ben Haley, and Stephen Small for useful comments on the manuscript and Hari Tammana for help with data deposition. Maps detailing sites of transcription for the fly genome were constructed as part of an ongoing genomics project in the laboratory of Tom Gingeras (Affymetrix, Inc.) and were accomplished with the assistance of V.S. This work was funded by National Institutes of Health Grant GM46638 (to M.S.L.) and in part with Federal Funds from the National Cancer Institute, National Institutes of Health, under Contract N01-CO-12400, and by Affymetrix, Inc. (to Tom Gingeras).

1. Moussian, B. & Roth, S. (2005) *Curr. Biol.* **15,** R887–R899.
2. Stathopoulos, A. & Levine, M. (2004) *Curr. Opin. Genet. Dev.* **14,** 477–484.
3. Sokol, N. S. & Ambros, V. (2005) *Genes Dev.* **19,** 2343–2354.
4. Biemar, F., Zinzen, R., Ronshaugen, M., Sementchenko, V., Manak, J. R. & Levine, M. S. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 15907–15911.
5. Kwon, C., Han, Z., Olson, E. N. & Srivastava, D. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 18986–18991.
6. Stathopoulos, A. & Levine, M. (2005) *Dev. Cell* **9,** 449–462.
7. Stathopoulos, A., Van Drenth, M., Erives, A., Markstein, M. & Levine, M. (2002) *Cell* **111,** 687–701.
8. Hino, S., Kishida, S., Michiue, T., Fukui, A., Sakamoto, I., Takada, S., Asashima, M. & Kikuchi, A. (2001) *Mol. Cell. Biol.* **21,** 330–342.
9. Wallingford, J. B. & Habas, R. (2005) *Development (Cambridge, U.K.)* **132,** 4421–4436.
10. Gordon, M. D., Dionne, M. S., Schneider, D. S. & Nusse, R. (2005) *Nature* **437,** 746–749.
11. Ganguly, A., Jiang, J. & Ip, Y. T. (2005) *Development (Cambridge, U.K.)* **132,** 3419–3429.
12. Michiue, T., Fukui, A., Yukita, A., Sakurai, K., Danno, H., Kikuchi, A. & Asashima, M. (2004) *Dev. Dyn.* **230,** 79–90.
13. Treisman, J. E., Lai, Z. C. & Rubin, G. M. (1995) *Development (Cambridge, U.K.)* **121,** 2835–2845.
14. Dobens, L. L., Hsu, T., Twombly, V., Gelbart, W. M., Raftery, L. A. & Kafatos, F. C. (1997) *Mech. Dev.* **65,** 197–208.
15. Law, G. L., Bickel, K. S., MacKay, V. L. & Morris, D. R. (2005) *Genome Biol.* **6,** R111.
16. Brown, C. Y., Mize, G. J., Pineda, M., George, D. L. & Morris, D. R. (1999) *Oncogene* **18,** 5631–5637.
17. Child, S. J., Miller, M. K. & Geballe, A. P. (1999) *J. Biol. Chem.* **274,** 24335–24341.
18. Morris, D. R. & Geballe, A. P. (2000) *Mol. Cell. Biol.* **20,** 8635–8642.
19. O'Connor, M. B., Umulis, D., Othmer, H. G. & Blair, S. S. (2006) *Development (Cambridge, U.K.)* **133,** 183–193.
20. Meller, V. H., Wu, K. H., Roman, G., Kuroda, M. I. & Davis, R. L. (1997) *Cell* **88,** 445–457.
21. Amrein, H. & Axel, R. (1997) *Cell* **88,** 459–469.
22. Tupy, J. L., Bailey, A. M., Dailey, G., Evans-Holm, M., Siebel, C. W., Misra, S., Celniker, S. E. & Rubin, G. M. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 5495–5500.
23. Inagaki, S., Numata, K., Kondo, T., Tomita, M., Yasuda, K., Kanai, A. & Kageyama, Y. (2005) *Genes Cells* **10,** 1163–1173.
24. Aboobaker, A. A., Tomancak, P., Patel, N., Rubin, G. M. & Lai, E. C. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 18017–18022.
25. Cumberledge, S., Zaratzian, A. & Sakonju, S. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 3259–3263.
26. Ronshaugen, M., Biemar, F., Piel, J., Levine, M. & Lai, E. C. (2005) *Genes Dev.* **19,** 2947–2952.
27. Aravin, A. A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J. & Tuschl, T. (2003) *Dev. Cell* **5,** 337–350.
28. Lai, E. C., Tomancak, P., Williams, R. W. & Rubin, G. M. (2003) *Genome Biol.* **4,** R42.
29. Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. & Gingeras, T. R. (2002) *Science* **296,** 916–919.
30. Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., *et al.* (2002) *Nature* **420,** 563–573.
31. Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., *et al.* (2004) *Genome Res.* **14,** 331–342.
32. Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K., *et al.* (2004) *Nat. Genet.* **36,** 40–45.
33. Schadt, E. E., Edwards, S. W., GuhaThakurta, D., Holder, D., Ying, L., Svetnik, V., Leonardson, A., Hart, K. W., Russell, A., Li, G., *et al.* (2004) *Genome Biol.* **5,** R73.
34. Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., *et al.* (2004) *Science* **306,** 2242–2246.
35. Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., *et al.* (2005) *Science* **308,** 1149–1154.
36. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., *et al.* (2005) *Science* **309,** 1559–1563.
37. Washietl, S., Hofacker, I. L., Lukasser, M., Huttenhofer, A. & Stadler, P. F. (2005) *Nat. Biotechnol.* **23,** 1383–1390.
38. Ravasi, T., Suzuki, H., Pang, K. C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M. C., Gongora, M. M., *et al.* (2006) *Genome Res.* **16,** 11–19.
39. Stathopoulos, A., Tam, B., Ronshaugen, M., Frasch, M. & Levine, M. (2004) *Genes Dev.* **18,** 687–699.
40. Yang, Z. (1997) *Comput. Appl. Biosci.* **13,** 555–556.
41. Nekrutenko, A., Makova, K. D. & Li, W.-H. (2002) *Genome Res.* **12,** 198–202.
42. Jiang, J., Kosman, D., Ip, Y. T. & Levine, M. (1991) *Genes Dev.* **5,** 1881–1891.
43. Rebeiz, M. & Posakony, J. W. (2004) *Dev. Biol.* **271,** 431–438.
44. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997) *Nucleic Acids Res.* **25,** 4876–4882.
45. Strimmer, K & von Haseler, A. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 6815–6819.