Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome

Benjamin P. Berman*, Yutaka Nibu*, Barret D. Pfeiffer†, Pavel Tomancak*‡, Susan E. Celniker†§, Michael Levine*, Gerald M. Rubin*†‡, and Michael B. Eisen*§¶

*Department of Molecular and Cell Biology, †Berkeley *Drosophila* Genome Project, and ‡Howard Hughes Medical Institute, University of California, Berkeley, CA 94720; and §Department of Genome Sciences, Life Sciences Division, Lawrence Orlando Berkeley National Laboratory, Berkeley, CA 94720

Contributed by Gerald M. Rubin, November 14, 2001

A major challenge in interpreting genome sequences is understanding how the genome encodes the information that specifies when and where a gene will be expressed. The first step in this process is the identification of regions of the genome that contain regulatory information. In higher eukaryotes, this cis-regulatory information is organized into modular units [cis-regulatory modules (CRMs)] of a few hundred base pairs. A common feature of these cis-regulatory modules is the presence of multiple binding sites for multiple transcription factors. Here, we evaluate the extent to which the tendency for transcription factor binding sites to be clustered can be used as the basis for the computational identification of cis-regulatory modules. By using published DNA binding specificity data for five transcription factors active in the early *Drosophila* embryo, we identified genomic regions containing unusually high concentrations of predicted binding sites for these factors. A significant fraction of these binding site clusters overlap known CRMs that are regulated by these factors. In addition, many of the remaining clusters are adjacent to genes expressed in a pattern characteristic of genes regulated by these factors. We tested one of the newly identified clusters, mapping upstream of the gap gene *giant (gt)*, and show that it acts as an enhancer that recapitulates the posterior expression pattern of *gt*.

The development of multicellular organisms is, to a large extent, dictated by a carefully choreographed progression of domain- and tissue-specific gene expression. To understand development, it is therefore necessary to understand the logic and mechanisms of this transcriptional network. Much of the information that determines when and where genes will be expressed is encoded in an organism's genome sequence. Although we now have genome sequences for many important metazoans, our understanding of how this information is encoded is extremely limited. Cracking this "cis-regulatory code" is a major problem in biology.

A paradigmatic model for studying transcriptional control of development is the early *Drosophila* embryo. Most of the important players have been identified by exhaustive genetic analysis, and there are sophisticated tools for characterizing the sequence features controlling the transcriptional network organized by these key developmental regulators. Although the early *Drosophila* embryo is relatively simple, many of the genes involved in early development of the fly are known to control development in other animals (1). Thus, it is likely that an understanding of the developmental cis-regulatory code in *Drosophila* will be applicable to other higher eukaryotes, including humans.

Careful genetic and biochemical dissection of numerous genes involved in *Drosophila* development suggests some general principles for how cis-regulatory regions are organized. For example, the cis-regulatory region of the pair-rule gene *even-skipped* (*eve*)—expressed in seven stripes in the blastoderm embryo—is organized into a series of discrete sequence regions of roughly 500 bp in length, each of which controls a distinct component of *eve*'s expression pattern (2–7). This modular organization of cis-regulatory regions is observed in many developmental genes in *Drosophila*, and in other organisms (8). In general, several transcription factors bind to each of these cis-regulatory modules (CRMs), and there are often multiple binding sites for each of these factors (8). Presumably, multiple bound transcription factors act combinatorially to confer specific transcriptional activity. For example, the enhancer controlling expression of *eve*'s second stripe contains at least three binding sites for Hunchback (Hb) and Giant (Gt), five for Bicoid (Bcd), and six for Krüppel (Kr) (4, 9).

It has been proposed that the high local density of transcription factor binding sites required for the proper function of these CRMs could be used as the basis for identifying novel CRMs (10–13). Here, we demonstrate the utility of this approach by examining the genome-wide distribution of binding sites for five transcription factors known to act together in the early *Drosophila* embryo.

Materials and Methods

Collection and Alignment of Transcription Factor Binding Sites. Bcd, Cad, Hb, Kr, and Kni binding sequences determined by *in vitro* DNase protection assays were compiled from a previous study (14) and additional sources. These sequences and their sources are listed in Fig. 5, which is published as supporting information on the PNAS web site, www.pnas.org.

Binding site sequences for Bcd, Hb, and Kr were aligned by using the pattern discovery tool MEME (v 3.0; ref. 15), with the following command line settings "-mod zoops -revcomp -dna." The "-minsites" parameter was set to 80% of the total number of sites collected for each transcription factor. This setting allowed for up to 20% of binding site sequences that aligned poorly to be omitted as potential sources of experimental error. For Bcd, 51/51 sites were aligned; for Hb, 93/93 sites were aligned; for Kr, 29/37 sites were aligned. A -bfile or background model file was used, which included mono-nucleotide, di-nucleotide, and tri-nucleotide frequencies determined from the intergenic *Drosophila melanogaster* genomic sequence, as annotated in Berkeley *Drosophila* Genome Project (BDGP)/Celera Release 1 (Rel 1.; ref. 16). Individual binding site sequences for Cad and Kni were aligned manually.

Abbreviations: CRM, cis-regulatory module; PWM, position weight matrix.

See commentary on page 546.

¶To whom reprint requests should be addressed at: Lawrence Berkeley National Laboratory, Mailstop 84-171, One Cyclotron Road, Berkeley, CA 94720. E-mail: mbeisen@lbl.gov.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Construction of Position Weight Matrices (PWMs) and Searching. PATSER (v. 3b; ref. 17) was used to construct PWMs from sequences as aligned as described above, and to search genomic sequence for matches to the PWM. PATSER was run with the following command line options: "-c -d2 -l 4." An "alphabet" file (specified with the command line parameter "-a") was used to provide the following background frequencies: A/T = 0.297, G/C = 0.203. These frequencies were determined from the intergenic *D. melanogaster* genomic sequence as annotated in Rel. 1.

PATSER was run on Rel. 1 genomic sequence, and CIS-ANALYST was used to identify all potential binding sites with *P* value < site_p. CIS-ANALYST examines sequence windows of length wind_size, retaining only those containing at least min_sites binding sites. CIS-ANALYST then collapses all overlapping windows into a single "cluster."

Collection of CRMs. Test CRM boundaries were determined as described in the studies listed in Table 1, which is published as supporting information on the PNAS web site. If the CRM had been sequenced as part of a prior study, we aligned this sequence with Rel. 1 genomic sequence and used the aligned segment from BDGP/Celera sequence (all sequences matched perfectly or with greater than 99% identity). If the CRM element had not been previously sequenced, we identified the restriction sites bordering the element, and extracted the genomic sequence occurring between these sites.

Test CRM Independent Matrices. In analyzing the overlap between binding site clusters and our test CRMs, we sought to avoid evaluating a particular CRM with PWMs built by using binding sites from that CRM. For each CRM, we constructed a separate set of PWMs that excluded binding sites derived from that CRM and used these PWMs to determine whether the CRM overlapped a binding site cluster. The sole exception was the Kni PWM for the *eve* stripe 3/7 CRM, because all Kni example binding sequences were derived from the *eve* stripe 3/7 CRM.

Genome-Wide Searches. CIS-ANALYST was used to search 93 Mb of noncoding DNA from Rel. 1 for clusters of Cad, Bcd, Hb, Kr, and Kni by using the parameters site_p = 0.0003 and wind_size = 700, and values of min_sites from 12 to 18. CIS-ANALYST was also used to search for clusters of Bcd, Hb, Kr, and Kni by using the parameters site_p = 0.0003, wind_size = 700, and min_sites = 13. The set of 28 clusters in Table 2 (which is published as supporting information on the PNAS web site) represents the union of the results for searches for Cad, Bcd, Hb, Kr, and Kni with min_sites = 15 and Bcd, Hb, Kr and Kni with min_sites = 13.

Whole-Mount *in Situ* Hybridizations and DNA Microarray Hybridizations. Embryonic whole-mount *in situ* RNA hybridizations and DNA microarray hybridizations were performed as described on the Berkeley *Drosophila* Genome Project web site (http://www.fruitfly.org/).

Giant Transgenics and Mutant Embryos. A 1.1-kb DNA fragment located upstream of the transcription start site of *gt* (from −2.7 to −1.6 kb) was amplified from *y w* fly genomic DNA by PCR by using two primers containing synthetic *Asc*I and *Not*I restriction sites: ttaggcgcgccagaaacttaccatcacttcg, attgcggccgccccattcaggggggattggg. The PCR product was digested with *Asc*I and *Not*I, and inserted in their native orientation into the *Asc*I-*Not*I site of a modified CaSpeR-AUG-bgal transformation vector (18) containing the eve basal promoter, starting at −42 bp and continuing through codon 22 fused in-frame with lacZ (19). The P-element transformation vectors were injected into yw embryos, as described previously (19, 20).

Results

The transcription factors Bicoid (Bcd), Caudal (Cad), Hunchback (Hb), Krüppel (Kr), and Knirps (Kni) act at very early stages of *Drosophila* development to define the anterior–posterior axis of the embryo (reviewed in ref. 21). Bcd (22) and Cad (23–25) are maternal activators broadly distributed in the anterior and posterior portions of the embryo, respectively. Hb, Kr, and Kni are zinc-finger gap proteins that act primarily as repressors in specific embryonic domains (reviewed in ref. 26). Aided greatly by a prior study (14), we collected sequences of previously described binding sites for these five factors present in the cis-regulatory regions of known target genes. We aligned the binding sequences for each factor by using the motif-assembly program MEME (15), and modeled the binding specificities of each factor with a PWM. PWMs are a useful way to represent binding specificities and provide a statistical framework for searching for novel instances of the motif in genome sequences (27, 28). The sequences used and PWMs produced are shown in Fig. 5.

We used the freely available program PATSER (17) to search the genome for sequences that matched these PWMs, and developed a web-based visualization tool, CIS-ANALYST (http://www.fruitfly.org/cis-analyst/) to display the location of predicted binding sites along with genome annotations in selected genomic regions. PATSER assigns a score to each potential site that reflects the agreement between the site and the corresponding PWM. These scores approximates the free energy of binding between the factor and site (27, 29), and CIS-ANALYST uses a user-defined cutoff parameter (site_p) to eliminate predicted low-affinity sites.

Using CIS-ANALYST, we examined the distribution of Bcd, Cad, Hb, Kr, and Kni binding sites in a 1-Mb genomic region surrounding the well-characterized *eve* locus at a site_p value of 0.0003 (Fig. 1). At this relatively high-stringency value, most experimentally verified binding sites are retained; at more restrictive values, many of these sites would be lost. Fig. 1*A* shows all predicted binding sites for all five factors and reveals that binding sites for these factors are densely and widely distributed across this region of the genome.

To investigate whether binding site clustering could help to explain the specificity of these factors for *eve*, we incorporated a simple notion of binding site clustering into CIS-ANALYST, allowing searches for segments of a specified length containing a minimum number of predicted binding sites. When we searched the 1-Mb region surrounding *eve* for dense clusters of predicted high-affinity sites (at least 13 Bcd, Cad, Hb, Kr, or Kni sites in a 700-bp window), three discrete regions were identified (Fig. 1 *B* and *C*). Strikingly, these three clusters were all adjacent to *eve*, and overlapped the previously characterized stripe 2, stripe 3 + 7, and stripe 4 + 6 enhancers.

To generalize and quantify these promising results, we compiled a broader collection of 19 well-defined CRMs from 9 *Drosophila* genes known to be required for proper embryonic development (see Table 1). Each of these CRMs is sufficient to direct the expression of a distinct anterior–posterior pattern in early embryos, and genetic evidence suggests that each is regulated by at least one of Bcd, Cad, Hb, Kr, and Kni. Mutation and *in vitro* DNA binding studies completed on a subset of the CRMs provide evidence for a direct regulatory relationship. The same clustering criteria that were successful for identifying CRMs in *eve* (700-bp regions with at least 13 predicted binding sites) identified clusters overlapping 14 of these 19 known CRMs (binding site plots for each of these CRMs are shown in Fig. 6, which is published as supporting information on the PNAS web site).

A search of the entire genome for 700-bp windows containing at least 13 predicted binding sites identified 133 clusters in

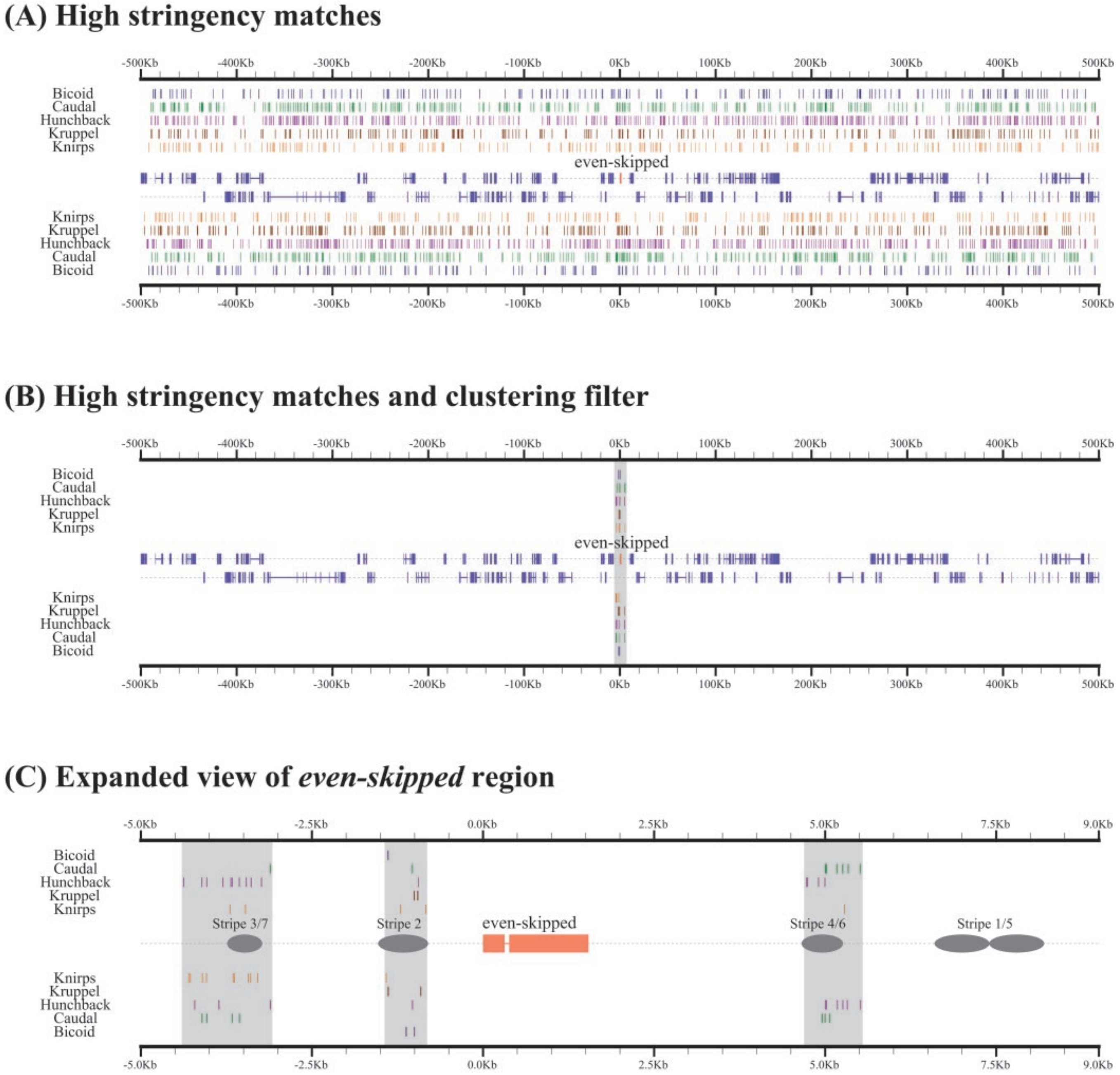


Fig. 1. Distribution of predicted transcription factor binding sites and binding site clusters in the vicinity of *eve*. (*A*) Predicted high-affinity ($P < 0.0003$) binding sites for the transcription factors Bcd, Cad, Hb, Kr, and Kni in 1 Mb of genomic sequence surrounding the gene *even-skipped (eve)* are displayed as colored boxes. Blue boxes in the center of the panel represent positions of annotated exons, with *eve* highlighted in red. Binding sites and genes shown above the midline map to the forward DNA strand; those below the midline map to the reverse strand. (*B*) Sites from *A* that occur in 700-bp windows containing at least 13 predicted binding sites. (*C*) Expanded view of region containing all clusters in *B*, with positions of known *eve* enhancers marked with gray ellipses.

addition to the 19 described above, or ≈1 per 700 kb of noncoding sequence. As expected, when more stringent clustering criteria are used, both the number of known CRMs recovered and the number of novel clusters identified decrease (see Fig. 2). We chose to examine further the novel clusters identified with a density of at least 15 binding sites per 700 bp, a level at which half of the known CRMs are still recovered. Binding site plots for the 22 novel clusters identified at this high stringency condition, and 6 additional novel clusters identified with an equally stringent search by using only Bcd, Hb, Kr, and Kni (see *Materials and Methods*) are shown in Fig. 7 (which is published as supporting information on the PNAS web site).

Twenty-three of these 28 clusters fall in regions between genes, whereas the remaining 5 fall in introns. There are therefore 49 genes that either contain a novel cluster of binding sites or flank an intergenic region that does. We examined the expression patterns of these 49 genes in early embryos by whole-mount RNA *in situ* hybridization and DNA microarray hybridization. The locations of these clusters and details and expression patterns of these adjacent and flanking genes are presented in Table 2. At least 10 of the 28 clusters were adjacent

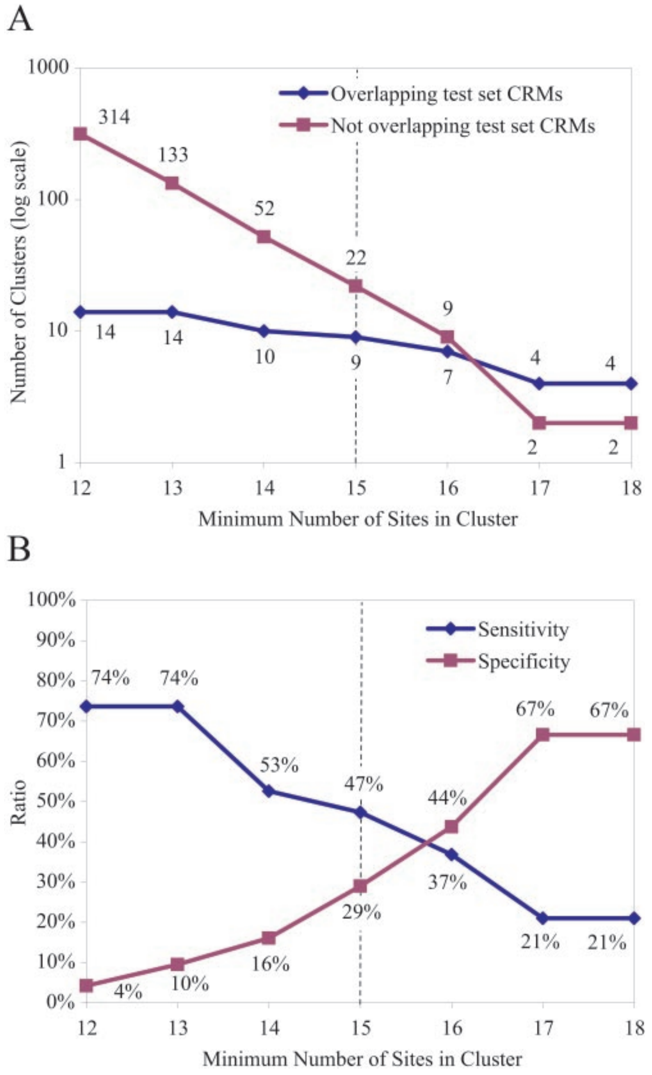


Fig. 2. Binding site clusters identified as a function of binding site density. (*A*) Number of binding site clusters in 93 Mb of noncoding genomic DNA at varying densities. Number of clusters overlapping test CRMs is shown in blue. Number of additional clusters is shown in pink. (*B*) Sensitivity (test set CRMs recovered divided by total number of test set CRMs) is shown in blue. Specificity (test set CRMs recovered divided by total number of clusters identified) is shown in pink. It is important to note that these sensitivity and specificity measures are computed assuming that only previously known CRMs are true positives. Because there are almost certainly additional bona fide CRMs in this set, the actual specificities and sensitivities of the method are expected to be better. Dotted line indicates density level chosen for the exploration of novel clusters described in the text.

to a gene that showed localized anterior–posterior expression in the syncitial or cellular blastoderm stages (see Fig. 3), consistent with early regulation by maternal effect or gap transcription factors. Although the numbers are small, this is significantly more than the 1 or 2 expected if the positions of clusters had been chosen at random.‖

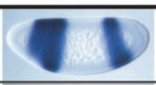
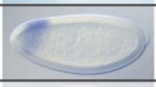
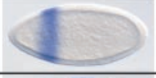
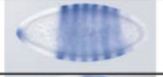
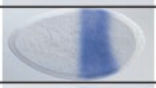
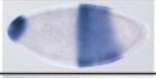
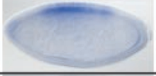
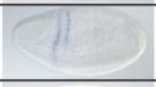
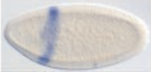
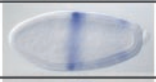
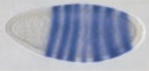
Cyto-logical position	Flanking gene(s)	Cluster start position	Cluster end position	Whole mount in-situ hybridization hour 0-4*	Known Regulators §
3A	gt	-2,759	-1,815		*bcd, cad, hb, Kr*
7F	otd	+37,924	+38,583		*bcd, hb*
8F	btd	-34,117	-33,376		*bcd, hb*
24A	odd	-3,394	-2,634		
33E	pdm1	-2,519	-1,825		*hb, kni*
33F	pdm2	-3,596	-2,873		*hb*
47A	psq	+45,838	+46,493		
75C	rpr	+6,616	+7,368		
84A	Dfd	-22,070	-21,370		*bcd, hb*
84B	Antp	+131,348	+132,041		*hb, kr, kni*
	ftz	+3,696	+4,389		*cad*

Fig. 3. Expression patterns of selected genes flanking novel binding site clusters. We examined the expression patterns of 49 genes adjacent to one of the 28 novel binding site clusters described in Table 2 in syncytial and cellular blastoderm embryos (whole mount RNA *in situ* images are available in Table 2 (which is published as supporting information on the PNAS web site) and on the Berkeley *Drosophila* Genome Project website (http://www.fruitfly.org/). Eleven of these genes representing 10 clusters had early embryonic expression patterns characteristic of genes regulated by maternal and gap transcription factors and are shown here. §, References for flanking genes are as follows: *gt* (25, 30, 37–40), *otd* (41–43), *btd* (44, 45), *pdm1* (46), *pdm2* (46), *Dfd* (47–49), *Antp* (49, 50), *ftz* (51–53), *odd* (54), and *psq* (55)

One of these clusters is located ≈2 kb upstream of the gap gene *giant* (*gt*; Fig. 4*A*). During cellularization, *gt* is expressed in two broad domains, one in the anterior and one in the posterior portion of the embryo (Fig. 4*B*). The pattern of expression of the posterior expression domain is known from a genetic analysis to be determined by the activities of Cad, Hb, and Kr (30). However, the cis-regulatory sequence controlling this posterior expression pattern has not been precisely identified. We sought to evaluate whether this cluster of binding sites might be the *gt* posterior enhancer. A 1.1-kb fragment containing this cluster was placed in a reporter construct containing the *eve* minimal promoter fused to a lacZ reporter gene. As shown in Fig. 4*D*, the expression pattern of this construct largely recapitulates the early expression pattern of the *gt* posterior expression domain. In the absence of Kr function, the anterior border of the *gt* posterior domain shifts anteriorly, indicating repression by Kr (Fig. 4*C* and ref. 30). The construct containing our *gt* posterior enhancer exhibits a similar shift in the absence of Kr (Fig. 4*E*).

Discussion

A central conundrum in understanding transcriptional regulation is that exquisite specificity in where and when genes are

‖Only 15 of 828 whole mount *in situ* hybridizations performed on randomly selected genes by the Berkeley *Drosophila* Genome Project (http://fruitfly.berkeley.edu) show localized anterior/posterior expression during the blastoderm stage. Based on this result, and accounting for the proportion of noncoding DNA in intergenic regions and introns, we estimate that a randomly selected 700-bp fragment has a 3.2% chance of being adjacent to a gene expressed in such a pattern.

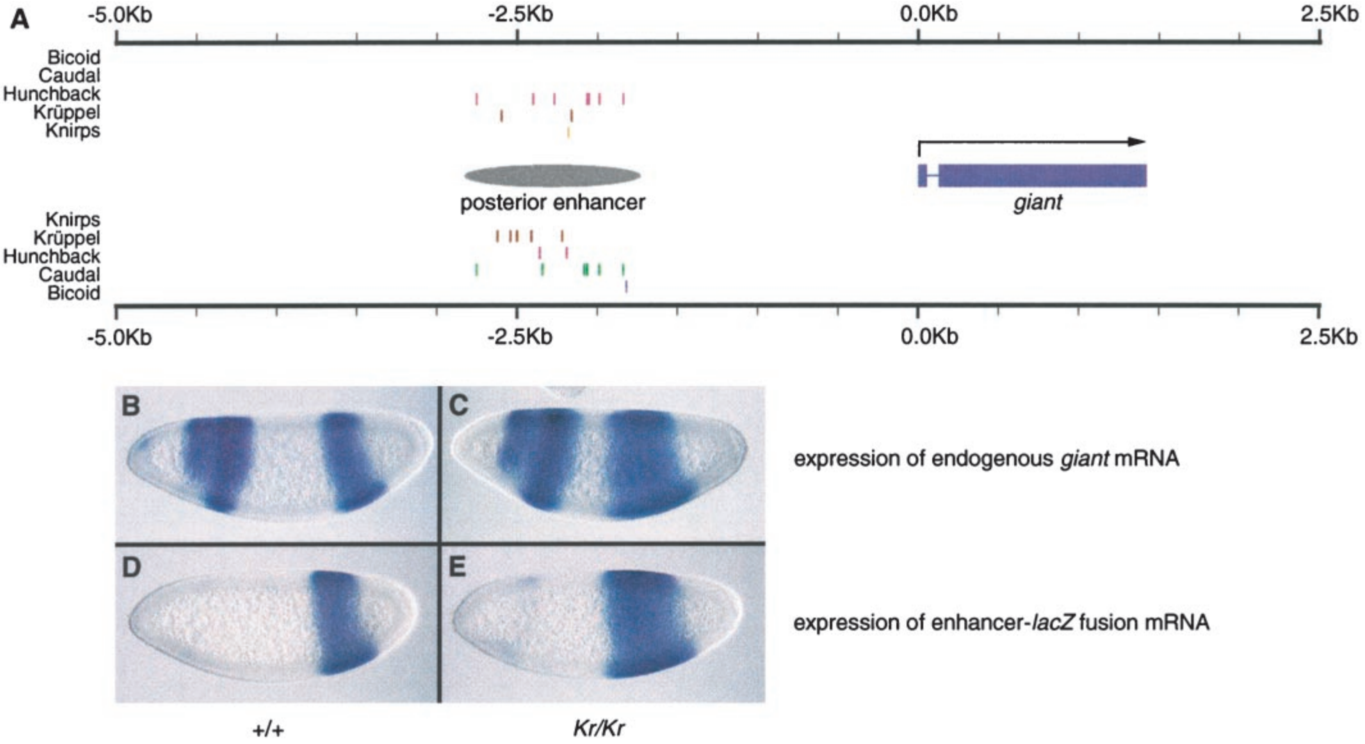


Fig. 4. Identification of a novel enhancer controlling posterior expression of *giant.* (*A*) Cluster of binding sites found between 2.9 Kb and 1.8 Kb upstream of *giant*. The DNA segment surrounding the cluster (labeled ''posterior enhancer'') was cloned into a lacZ fusion construct and introduced into the genome via germline transformation as described in *Materials and Methods*. (*B* and *C*) Expression of *giant* in syncitial blastoderm stage embryos as determined by RNA *in situ* hybridization. *B* shows a wild-type embryo, and *C* shows a Kr^1/Kr^1 embryo lacking Krüppel (Kr) function. Without repression by *Kr*, the anterior border of the posterior expression domain shifts anteriorly. (*D* and *E*) Expression of lac Z in embryos containing construct from *A*. *D* shows a wild-type embryo, and *E* shows a Kr^1/Kr^1 embryo. Expression of the lac*Z* construct in the mutant embryo shows similar expansion to that seen in *gt*.

GENETICS

DEVELOPMENTAL BIOLOGY

expressed is achieved through the action of sequence-specific DNA binding proteins whose sequence specificities are often not highly specific. Many transcription factors, like those examined here, bind DNA as monomers and recognize relatively short and degenerate sequences. Multiple predicted binding sites for each of the five transcription factors we examined are found adjacent to virtually every gene in the genome. Yet, only a handful of genes are regulated by these factors. The most striking result of this study is that variation in the local density of these binding sites appears sufficient to account for much of the specificity in their activities. At least a third, and likely more, of high-density clusters of maternal effect and gap transcription factor binding sites examined here correspond to bona fide cis-regulatory modules active in the early embryo. Although we did not observe patterned transcriptional activity in the early embryo for genes associated with a number of these clusters, Bcd, Cad, Hb, Kr, and Kni are also active later in development, and some clusters may represent regulatory elements active in these later processes. For instance, Hb and Kr are expressed in developing neuroblast cells much later in embryogenesis, and both are thought to contribute to fate determination of these cell populations (31).

Although it has been previously proposed that binding site clustering could be used to identify cis-regulatory modules (12), it is striking just how effective we have found this approach to be. The identification of functionally significant noncoding regions, especially those that control transcription, is one of the major challenges in understanding genome sequences, and we are optimistic that approaches like the one presented here will generalize to later stages of *Drosophila* development and to other organisms. Many of the transcription factors active in early *Drosophila* embryogenesis are also active in other organisms. In addition, binding site clustering has been observed in CRMs involved in processes later in *Drosophila* development and in the development of other organisms.

It is likely that improved methods for identifying binding site clusters will yield even better results than those presented here. The 700-bp window that we used is not appropriate for all CRMs. We are currently implementing a statistical model that will identify significant clusters of binding sites in windows of arbitrary size. This model will also consider the degree of agreement between prospective binding sites and the motif model, with increased significance assigned to clusters containing predicted high-affinity sites. Our binding site models will also be improved to account for any positional dependence occurring within transcription factor binding sites themselves; the current PWM model assumes that each base is independent. Models that incorporate some of these features have been successfully applied to other systems (10, 11, 13, 32). Additional functional exploration of computationally identified clusters will provide new examples that will permit further refinement of these methods.

Although binding data exist for some additional *Drosophila* transcription factors (33), a major roadblock to further evaluating and developing this approach is the paucity of available binding data for most transcription factors. We have initiated a project to gather accurate sequence specificity information for all of the ≈120 transcription factors believed to act in early *Drosophila* embryogenesis. With these data, it will be possible to search the genome for significant clusters of binding sites for additional combinations of factors. However, we cannot simply look for clusters of any combination of factors. Although high-density clusters of binding sites for *particular* combinations of factors are relatively rare, we expect every region of the genome to contain significant clusters of binding sites for *some* combinations of these 120 transcription factors. The analysis presented here worked in large part because we already knew that Bcd, Cad, Hb, Kr, and Kni act together. We plan to systematically identify additional sets of coacting factors by analyzing the expression patterns of transcription factors and through further genetic studies. The imminent availability of the *Drosophila pseudoobscura* genome sequence will provide additional means for distinguishing biologically relevant clusters from those that occur by chance, because only functionally significant clusters should be found in both genomes. In addition, the identification of blocks of noncoding DNA conserved between *D. melanogaster* and *D. pseudoobscura* will be useful in subsequent studies because recent analyses in many species suggest that such sequences are significantly enriched for functional transcription factor binding sites and CRMs (34, 35).

The grammar of the cis-regulatory code is clearly more complex than simply the density of transcription factor binding sites. The relative positioning of sites within cis-regulatory modules has been demonstrated to be significant in many cases. This result is to be expected, because we know that there are often important protein–protein interactions between bound factors that can influence CRM function. However, some plasticity in the positioning of binding sites is tolerated in some situations (14, 36). The analysis of orthologous CRMs in multiple species should help to further elucidate the rules governing CRM structure (as in ref. 14). Ultimately, we would like to incorporate these rules into our methods for identifying CRMs. However, to achieve a sufficient understanding of the architecture of cis-regulatory modules, we need to expand the number of identified and characterized CRMs. We believe that CRM detectors based on binding site clustering are a useful first step along this path.

We thank Amy Beaton, Elaine Kwan, Steven Richards, Richard Weiszman, Joseph Nunoo, and Todd Laverty for assistance with RNA *in situ* hybridizations; Chris Mungall for assistance with genome annotations; and Dmitri Papatsenko for sharing unpublished data and ideas. The binding site plots included in this work were created by using the GFF2PS (http://www1.imim.es/software/gfftools/GFF2PS.html), and we thank the author, Josep Francesco Abril Ferrando, for his personal assistance. B.P.B. wishes to thank Ed Lewis, Michael Brodsky, Erwin Frise, and Derek Chiang for helpful discussions. We also thank Casey Bergman for his critical reading of the manuscript. The CIS-ANALYST tool, and all sequences and other data used in this work are available at http://www.fruitfly.org/cis-analyst. This work was supported by National Institutes of Health Grants GM34431 (to M.S.L.), P50 HG00750 (to G.M.R.), and HL667201 (to M.B.E.); Department of Energy contract DE-AC03-76SF00098 (to M.B.E.); and by the Howard Hughes Medical Institute. M.B.E. is a Pew Scholar in the Biomedical Sciences.

1. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. (2001) *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design* (Blackwell Scientific, Oxford).
2. Harding, K., Hoey, T., Warrior, R. & Levine, M. (1989) *EMBO J.* **8,** 1205–1212.
3. Goto, T., Macdonald, P. & Maniatis, T. (1989) *Cell* **57,** 413–422.
4. Stanojevic, D., Small, S. & Levine, M. (1991) *Science* **254,** 1385–1387.
5. Small, S., Blair, A. & Levine, M. (1996) *Dev. Biol.* **175,** 314–324.
6. Sackerson, C., Fujioka, M. & Goto, T. (1999) *Dev. Biol.* **211,** 39–52.
7. Fujioka, M., Emi-Sarker, Y., Yusibova, G. L., Goto, T. & Jaynes, J. B. (1999) *Development (Cambridge, U.K.)* **126,** 2527–2538.
8. Davidson, E. H. (2001) *Genomic Regulatory Systems: Development and Evolution* (Academic, San Diego).
9. Ludwig, M. Z., Patel, N. H. & Kreitman, M. (1998) *Development (Cambridge, U.K.)* **125,** 949–958.
10. Crowley, E. M., Roeder, K. & Bina, M. (1997) *J. Mol. Biol.* **268,** 8–14.
11. Wasserman, W. W. & Fickett, J. W. (1998) *J. Mol. Biol.* **278,** 167–181.
12. Wagner, A. (1999) *Bioinformatics* **15,** 776–784.
13. Frith, M. C., Hansen, U. & Weng, Z. (2001) *Bioinformatics* **17,** 878–889.
14. Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. (2000) *Nature (London)* **403,** 564–567.
15. Bailey, T. L. & Elkan, C. (1994) *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2,** 28–36.
16. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.* (2000) *Science* **287,** 2185–2195.
17. Hertz, G. Z. & Stormo, G. D. (1999) *Bioinformatics* **15,** 563–577.
18. Thummel, C. S., Boulet, A. M. & Lipshitz, H. D. (1988) *Gene* **74,** 445–456.
19. Small, S., Blair, A. & Levine, M. (1992) *EMBO J.* **11,** 4047–4057.
20. Kosman, D. & Small, S. (1997) *Development (Cambridge, U.K.)* **124,** 1343–1354.
21. Niessing, D., Rivera-Pomar, R., La Rosee, A., Hader, T., Schock, F., Purnell, B. A. & Jackle, H. (1997) *J. Cell. Physiol.* **173,** 162–167.
22. St Johnston, D. & Nusslein-Volhard, C. (1992) *Cell* **68,** 201–219.
23. Macdonald, P. M. & Struhl, G. (1986) *Nature (London)* **324,** 537–545.
24. Mlodzik, M. & Gehring, W. J. (1987) *Cell* **48,** 465–478.
25. Rivera-Pomar, R., Lu, X., Perrimon, N., Taubert, H. & Jackle, H. (1995) *Nature (London)* **376,** 253–256.
26. Sauer, F., Rivera-Pomar, R., Hoch, M. & Jackle, H. (1996) *Philos. Trans. R. Soc. London B* **351,** 579–587.
27. Berg, O. G. & von Hippel, P. H. (1987) *J. Mol. Biol.* **193,** 723–750.
28. Stormo, G. D. (2000) *Bioinformatics* **16,** 16–23.
29. Stormo, G. D. & Fields, D. S. (1998) *Trends Biochem. Sci.* **23,** 109–113.
30. Kraut, R. & Levine, M. (1991) *Development (Cambridge, U.K.)* **111,** 601–609.
31. Isshiki, T., Pearson, B., Holbrook, S. & Doe, C. Q. (2001) *Cell* **106,** 511–521.
32. Krivan, W. & Wasserman, W. W. (2001) *Genome Res.* **11,** 1559–1566.
33. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., *et al.* (2001) *Nucleic Acids Res.* **29,** 281–283.
34. Pennacchio, L. A. & Rubin, E. M. (2001) *Nat. Rev. Genet.* **2,** 100–109.
35. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. (2000) *Nat. Genet.* **26,** 225–228.
36. Arnosti, D. N., Barolo, S., Levine, M. & Small, S. (1996) *Development (Cambridge, U.K.)* **122,** 205–214.
37. Mohler, J., Eldon, E. D. & Pirrotta, V. (1989) *EMBO J.* **8,** 1539–1548.
38. Kraut, R. & Levine, M. (1991) *Development (Cambridge, U.K.)* **111,** 611–621.
39. Eldon, E. D. & Pirrotta, V. (1991) *Development (Cambridge, U.K.)* **111,** 367–378.
40. Capovilla, M., Eldon, E. D. & Pirrotta, V. (1992) *Development (Cambridge, U.K.)* **114,** 99–112.
41. Finkelstein, R. & Perrimon, N. (1990) *Nature (London)* **346,** 485–488.
42. Gao, Q., Wang, Y. & Finkelstein, R. (1996) *Mech. Dev.* **56,** 3–15.
43. Gao, Q. & Finkelstein, R. (1998) *Development (Cambridge, U.K.)* **125,** 4185–4193.
44. Wimmer, E. A., Jackle, H., Pfeifle, C. & Cohen, S. M. (1993) *Nature (London)* **366,** 690–694.
45. Wimmer, E. A., Simpson-Brose, M., Cohen, S. M., Desplan, C. & Jackle, H. (1995) *Mech. Dev.* **53,** 235–245.
46. Cockerill, K. A., Billin, A. N. & Poole, S. J. (1993) *Mech. Dev.* **41,** 139–153.
47. Martinez-Arias, A., Ingham, P. W., Scott, M. P. & Akam, M. E. (1987) *Development (Cambridge, U.K.)* **100,** 673–683.
48. Jack, T. & McGinnis, W. (1990) *EMBO J.* **9,** 1187–1198.
49. Reinitz, J. & Levine, M. (1990) *Dev. Biol.* **140,** 57–72.
50. Harding, K. & Levine, M. (1988) *EMBO J.* **7,** 205–214.
51. Dearolf, C. R., Topol, J. & Parker, C. S. (1989) *Genes Dev.* **3,** 384–398.
52. Dearolf, C. R., Topol, J. & Parker, C. S. (1989) *Nature (London)* **341,** 340–343.
53. Topol, J., Dearolf, C. R., Prakash, K. & Parker, C. S. (1991) *Genes Dev.* **5,** 855–867.
54. Coulter, D. E., Swaykus, E. A., Beran-Koehn, M. A., Goldberg, D., Wieschaus, E. & Schedl, P. (1990) *EMBO J.* **9,** 3795–3804.
55. Lehmann, M., Siegmund, T., Lintermann, K. G. & Korge, G. (1998) *J. Biol. Chem.* **273,** 28504–28509.