Motif-Blind, Genome-Wide Discovery of *cis*-Regulatory Modules in *Drosophila* and Mouse

Miriam R. Kantorovitz,[1,8] Majid Kazemian,[2,8] Sarah Kinston,[5] Diego Miranda-Saavedra,[5] Qiyun Zhu,[6] Gene E. Robinson,[3,4] Berthold Göttgens,[5,*] Marc S. Halfon,[6,7,*] and Saurabh Sinha[2,3,*]

[1]Department of Mathematics
[2]Department of Computer Science
[3]Department of Entomology
[4]Institute for Genomic Biology
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[5]Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 0XY, UK
[6]Department of Biological Sciences
[7]Department of Biochemistry
State University of New York at Buffalo, Buffalo, NY 14260, USA
[8]These authors contributed equally to this work
*Correspondence: bg200@cam.ac.uk (B.G.), mshalfon@buffalo.edu (M.S.H.), sinhas@illinois.edu (S.S.)
DOI 10.1016/j.devcel.2009.09.002

SUMMARY

We present new approaches to *cis*-regulatory module (CRM) discovery in the common scenario where relevant transcription factors and/or motifs are unknown. Beginning with a small list of CRMs mediating a common gene expression pattern, we search genome-wide for CRMs with similar functionality, using new statistical scores and without requiring known motifs or accurate motif discovery. We cross-validate our predictions on 31 regulatory networks in *Drosophila* and through correlations with gene expression data. Five predicted modules tested using an in vivo reporter gene assay all show tissue-specific regulatory activity. We also demonstrate our methods' ability to predict mammalian tissue-specific enhancers. Finally, we predict human CRMs that regulate early blood and cardiovascular development. In vivo transgenic mouse analysis of two predicted CRMs demonstrates that both have appropriate enhancer activity. Overall, 7/7 predictions were validated successfully in vivo, demonstrating the effectiveness of our approach for insect and mammalian genomes.

INTRODUCTION

In metazoans, much of transcriptional regulation is mediated by *cis*-regulatory modules (CRMs; also "modules" or "enhancers") that form the building blocks of gene regulatory networks (Carroll et al., 2001). CRM identification formerly had been possible only through a dedicated empirical approach of testing sequence fragments for regulatory activity in a reporter gene assay. The genomics era has led to the development of new genome-wide techniques to screen for potential gene-regulatory regions, such as chromatin immunoprecipitation coupled to genomic tiling arrays (ChIP-chip) (Li et al., 2008) or ultra-high-throughput sequencing (ChIP-Seq) (Visel et al., 2009). Although they show great promise, even these empirical methods do not identify regulatory elements or predict their tissue-specific activity with complete accuracy. Importantly, the fact that it is impossible to assay all tissue types under all conditions means that potential CRMs will be missed by these techniques.

In recent years, computational methods have provided an attractive complementary approach for module identification. However, their effectiveness has been limited to a few well-understood biological systems, where prior knowledge of the requisite TFs and their binding sites could be exploited. This paper examines CRM prediction in the much more common scenario where knowledge of the relevant TFs and/or their binding specificities (motifs) is missing. We tackle here a common variant of this problem: "Suppose a small set of modules participating in a transcriptional subnetwork is known a priori. The task is to use such information as 'training data' to guide the search for other modules in that subnetwork." We call this task the "supervised CRM prediction problem." The term subnetwork refers here to a group of genes that are coordinately expressed as a result of having common regulatory inputs. We define a successful CRM prediction as identification of a sequence that drives an expression pattern commensurate with a nearby gene's endogenous expression. Depending on how specific the expression patterns of the training CRMs are, a successfully predicted CRM may recapitulate their common expression pattern, or may not.

Computational methods for CRM prediction (Berman et al., 2002; Halfon et al., 2002; Frith et al., 2003; Sinha et al., 2003; Philippakis et al., 2005; Donaldson et al., 2005) typically scan the genome for clusters of putative binding sites that are defined by sequence similarity to known motifs. This approach may fail on at least two grounds. First, motif information is sparse as of now; e.g., motif databases for *Drosophila* catalog only ~12% of the estimated number of TFs. Intense efforts are being made to characterize the binding specificities of all TFs in mouse (Berger et al., 2006) and fruit fly (Noyes et al., 2008) and may in the long term alleviate the problem. However, these efforts are labor-intensive and relatively expensive, and the problem may

thus persist for scientists studying organisms other than human, mouse, and fruit fly. A second, more serious problem facing CRM discovery stems from the fact that most computational tools need prior knowledge of the TFs relevant to the specific regulatory network of interest. For less-studied regulatory systems, such knowledge may not be available. Admittedly, even if the relevant TFs and/or their motifs are unknown, computational motif finding tools may be used to discover position-weight-matrix (PWM) motifs from the training data. However, the modest success rate of motif-finding programs, as suggested by a recent survey (Tompa et al., 2005), casts doubts upon the prospect of CRM discovery based on computational motif finding.

Here, we address simultaneously both problems by undertaking supervised CRM discovery in the absence of motif knowledge and without relying upon accurate motif finding. We propose and examine various statistics to capture the functional similarity (due to shared binding sites) between a candidate CRM and the given set of modules. These statistics belong to the realm of "alignment-free" sequence comparison, since the similarity to be detected is not due to orthology. The statistics are based on frequencies of short words, akin to many motif-finding programs, but without the usual objective of finding the most specific (biochemically accurate) characterization of the TF's binding sites. All new methods developed here are made publicly available as source code at http://veda.cs.uiuc.edu/scrm/index.htm.

Previous attempts at solving the supervised CRM prediction problem (Chan and Kibler, 2005; Grad et al., 2004; Nazina and Papatsenko, 2003) have been primarily tested on a single data set, the anterior-posterior patterning subnetwork in *D. melanogaster*. We now evaluate two existing and seven new scoring schemes on 31 data sets in *Drosophila* and 8 data sets in mammals, and we perform in vivo validation in both species. In our previous work (Ivan et al., 2008), we proposed computational methods for CRM discovery without prior knowledge of motifs or modules, where the search was constrained to regions around coregulated genes. Here, we relax that constraint and enable genome-wide search by leveraging the prior knowledge of related CRMs where available. The methods of Ivan et al. (2008) are not applicable in this setting. An unsupervised version of our problem was also addressed in Rajewsky et al. (2002) through the use of Poisson statistics on short word counts.

We performed extensive cross-validation tests with our scores on 31 data sets representing a broad spectrum of regulatory subnetworks in *D. melanogaster*, exploiting known modules cataloged in the REDfly database (Halfon et al., 2008). Our tests established the feasibility of supervised CRM prediction for about half of the examined data sets and also identified data sets that are not amenable to our scores. We then predicted modules genome-wide for each amenable regulatory subnetwork and found their neighboring genes to be highly enriched for the expected expression patterns. We filtered our predicted module collection based on gene expression data, producing a high-confidence set of putative CRMs belonging to a regulatory subnetwork. We tested five predicted modules in vivo and found each of the five to drive reporter gene expression that recapitulates aspects of the endogenous gene expression (although not always in the expected pattern).

Assessment of the supervised prediction pipeline on 8 data sets in mammals, comprising 244 tissue-specific enhancers, led to ~60% of the enhancers being recovered. We finally applied this pipeline to predict CRMs with roles in mammalian blood and cardiovascular development. In vivo validation in transgenic mice allowed us to demonstrate successful identification of two regulatory regions with the predicted activity and demonstrates the extensibility of our computational approach beyond *Drosophila*.

RESULTS

Scoring Schemes

Given a genomic region in which to search, each of our CRM prediction schemes scans the sequence with a shifting window of fixed size, and scores the window for similarity to a (given) training set of CRMs. Thus, the crucial component of each of these schemes is its distinct scoring system for matching a candidate module ("test CRM") to the set of known "training modules." These scoring schemes are described next at a general level; details are available in Experimental Procedures and Table S11 (available online).

(1) *Markov chain-based score*: The "HexMCD" score trains separate generative models (5th order Markov chains) for training modules and background sequences, and quantifies which model matches the test sequence better. This score was originally proposed by Grad et al. (2004).

(2) *Dot product-based scores, with statistical significance estimation ("D2z")*: These scores are based on the dot-product of k-mer frequency distributions of training and test sequences. Importantly, the scores are made to reflect the statistical significance of this dot-product, by analytical computation of "z scores" under suitable null models, hence the name "D2z." For technical reasons described in Experimental Procedures, we had to develop and implement this analytical calculation differently from our previous work (Ivan et al., 2008; Kantorovitz et al., 2007) on the D2z score. We developed three scoring schemes in this category (Experimental Procedures).

(3) *Selection of representative k-mers, followed by Poisson statistics*: In the new score called "*P*oisson *a*dditive *c*onditional" or "PAC," words that are even weakly associated with the training CRMs are identified and their additive contributions to the similarity score are based on overrepresentation of those words in the test sequence, relative to background sequences, calculated using Poisson statistics (Experimental Procedures).

(4) *Selection of representative k-mers, followed by weighted sum of counts*: The two scores in this category are generically defined as $\sum_{w \in W} \sigma(w)n(w)$, where $n(w)$ is the number of occurrences of word w in the test sequence, $\sigma(w)$ is a weight reflecting its association with the training modules, and the set W comprises the top ranking words based on $\sigma(w)$. In the "HexDiff" score of Chan and Kibler (2005), reimplemented here, the weight of a word is the ratio of its frequency in training and background sequences. In the "HexYMF" scheme we have designed, the weight of w is the "z score" (Experimental Procedures) of the count of w in the training CRMs,.

(5) *Motif database-driven score*: We developed a program, called Stubb-MDB (Stubb based on *m*otif *d*atabase), that begins

with a large compendium of experimentally validated motifs (Matys et al., 2003; Noyes et al., 2008; Halfon et al., 2008), determines the motifs that are relevant to the regulatory subnetwork of interest, and runs the Stubb program (Sinha et al., 2003) with these short-listed motifs to score the test sequence for matches to the motifs. This motif-based approach provides a useful point of comparison to the motif-blind approaches outlined above (1–4).

Comparison of Different Scoring Schemes and Characterization of Data Sets

We utilized a modified version of the CRM prediction benchmark developed in our previous work (Ivan et al., 2008). The new benchmark consists of 31 data sets, each data set comprising a collection of bona fide CRMs that mediate gene expression patterns with some level of commonality. We employed "leave-one-out cross-validation" (LOOCV) to assess the relative performance of our scoring schemes (see Experimental Procedures). Figure 1A summarizes the results of our cross-validation tests, with asterisks indicating cases where the CRM-level sensitivity was statistically significant ($p \leq 0.05$; see Supplemental Experimental Procedures). In such cases, we say that the method is "successful" on the data set. We immediately note that the best motif-blind methods (HexMCD, HexDiff-rc, HexYMF-s200-rc) succeed on close to half of the data sets. In contrast, the method that makes use of a motif database (Stubb-MDB-rc) succeeds on fewer data sets. We also tested an alternative motif-based pipeline—that of using "Clover" (Frith et al., 2004) for motif selection and "ClusterBuster" (Frith et al., 2003) for scanning with selected motifs—and found success levels to be comparable to that of Stubb-MDB (Table S12). Nucleotide-level sensitivity measurements show similar trends (Figure S1). Repeating the cross-validation exercise four times (see Experimental Procedures), we examined which methods succeed on a specific data set in all four LOOCV "instantiations" (Table S3). We find that 15 of the 31 data sets have at least one method (among the eight listed) for which this consistency requirement is met (Figure 1D). We designate these as the "amenable" data sets, i.e., the CRMs in such a data set have the extent and kind of sequence similarity that we need for supervised prediction. We find that for data sets neuroectoderm and blastoderm, all shown methods succeed consistently, suggesting that these two regulatory subnetworks are among the easiest ones for supervised CRM prediction. A closer look at the top 200 words used by HexYMF for the blastoderm data set reveals matches to 6 of the canonical motifs for this network (Figure S4). We also find examples, such as cardiac-mesoderm, eye.1, and mesectoderm, where only one of the methods is a consistent performer (white cells in Figure 1D), indicating that the common sequence features of the CRMs in each of these regulatory subnetworks may be harder to capture.

Figures 1B and 1C show one-on-one comparison of the eight methods. We find the top four methods—HexMCD, PAC-rc, HexYMF-s200-rc, and HexDiff-rc—to be competitive with one another. The emerging theme however is that CRMs in different data sets are best predicted by different scoring schemes. For instance, the "D2z-cond-weights" score is the only method consistently successful on the "somatic muscle" data set (Figure 1D), even though in the final tally over all data sets (Figures 1B and 1C)

it is the least successful score. It is natural to ask if a combination of two or three methods can yield better predictive performance on average. We chose the methods HexYMF-s200-rc, HexDiff-rc, and PAC-rc to build what we call a "fusion" method—which scores a candidate sequence by the product of the scores from these three methods. Figures 1B and 1C show that this naive combination method is indeed better overall than any individual method. It has consistent performance on ten data sets (Figure 1D), the most among all methods. It is also the superior method in one-on-one comparisons (Figure 1B).

Finally, we attempted to find characteristics of data sets that allow strong performance. We found a significant correlation between prediction accuracy on a data set and (1) the extent of homotypic clustering of short words in the training set (Figure S2A, $p = 0.009$), (2) the GC content of the training set (Figure S2B, $p = 0.005$), and (3) the extent of nucleotide-level conservation with orthologous sequence (Figure S2C, $p = 0.007$). We did not however see a significant correlation between the number of training CRMs and the performance on a data set.

Multispecies Comparisons Improve Predictions

We obtained the orthologs of the training CRMs from each of two other *Drosophila* species (*D. ananassae* and *D. pseudoobscura*), and predicted CRMs in *D. melanogaster* sequence with each of these training sets. Scores from these runs were then averaged to give us a multispecies score profile. Cross-validation on all 31 data sets showed that the multispecies version of HexMCD substantially outperformed its single species version (Figure 1E; Table S3).

Role of the Training Set

We asked if the success of our supervised CRM prediction methods is due to an ability of our scoring schemes to learn the characteristics of the specific regulatory subnetwork, or an ability to infer general sequence characteristics of CRMs. We constructed for each data set a "random training set" consisting of REDfly CRMs from outside of that regulatory subnetwork, and repeated the above evaluation. Of the 16 data sets on which the "fusion" method performs significantly well (Figure 1A), 9 data sets perform worse when using the random training set, meaning that the CRM predictions are specific to the particular regulatory subnetwork. On the other hand, the performance on seven data sets, though significantly strong, is comparable to that with training sets comprising randomly selected CRMs, suggestive of a less specific type of CRM prediction based on sequence characteristics of CRMs in general (Table 1). Interestingly, the data sets that give rise to specific prediction are typically smaller than the nonspecific ones ($p < 0.02$, Table 1), and their expression patterns, as measured by the number of individual anatomy terms defining the pattern (Table S7), are more specific ($p < 0.002$, Table 1). We also assessed the value of the training set by executing an unsupervised CRM prediction method ("CSam," from our previous work [Ivan et al., 2008]) on each of the 31 data sets. The method was successful on only one of the data sets (data not shown), indicating a major advantage conferred by the use of the training set, and also pointing out that our modified benchmark is substantially "harder" than that of Ivan et al. (2008) (Supplemental Experimental Procedures).
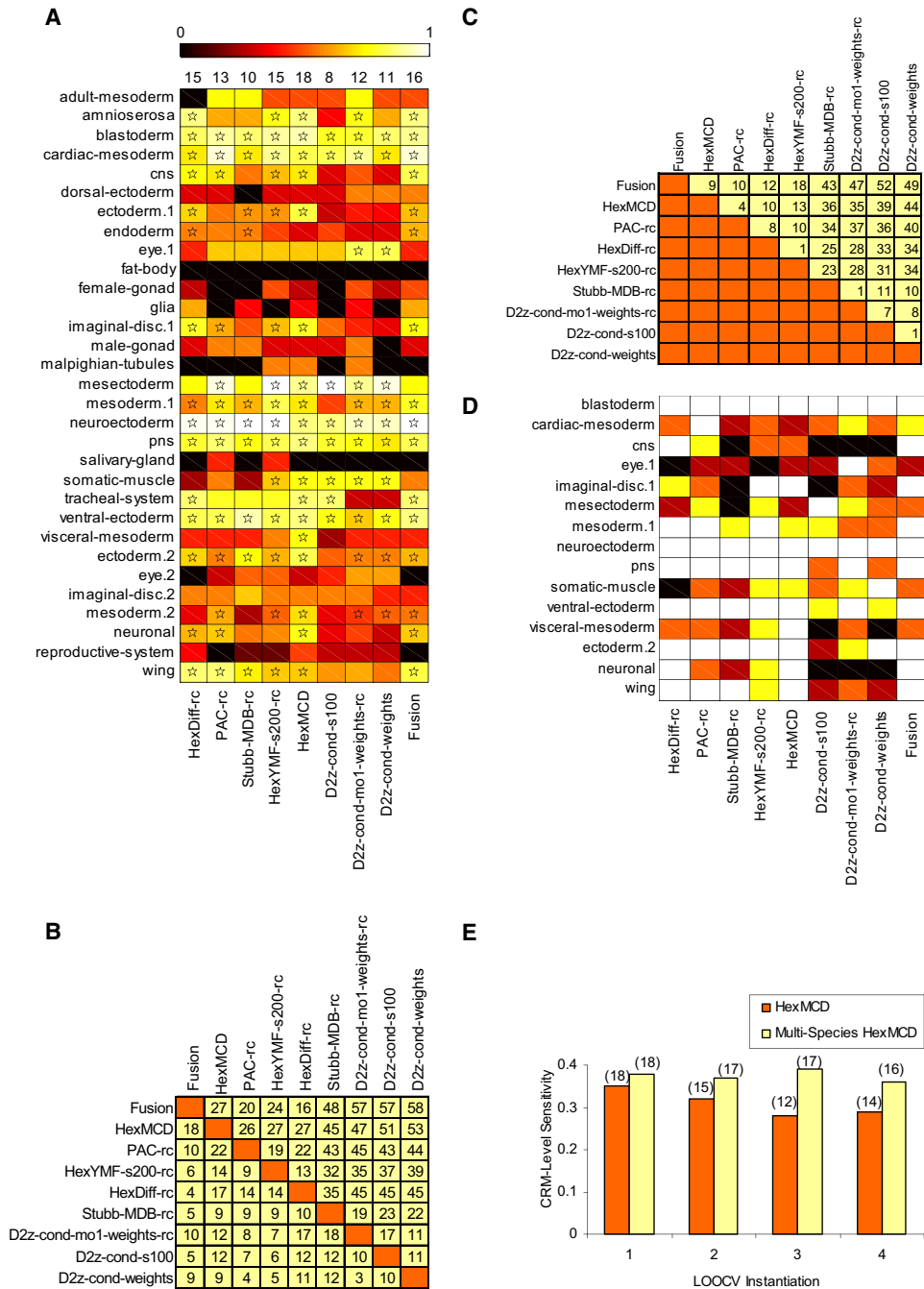
Figure 1. Assessment of Methods

(A) LOOCV performance of each of 9 different methods on 31 data sets in the benchmark. Color accents represent CRM-level sensitivity on a scale of 0 to 1, and cases with an empirical p value ≤ 0.05 are marked by asterisks. The top row shows the number of data sets amenable to supervised prediction by each method. One-on-one comparison of methods: For each pair of methods M1 (row) and M2 (column), (B) the "wins" of M1 versus M2 (i.e., the number of data sets on which CRM-level sensitivity of M1 was greater than that of M2 by at least 10% of data set size).

(C) The difference between the wins of M1 versus M2 and the wins of M2 versus M1 in CRM-level sensitivity.

(D) Fifteen data sets on which at least one method succeeds in all four instantiations. Color indicates the number of instantiations (out of four) on which the performance was significant ($p \leq 0.05$): white = 4, yellow = 3, orange = 2, brown = 1, black = 0.

(E) Comparison of single species and multispecies versions of HexMCD. For each each LOOCV instantiation, the average CRM-level sensitivity over all data sets (*y* axis) and number of amenable data sets (number above each bar) are compared between the two methods.

Table 1. Categorization of Data Sets Amenable to the Fusion Method into Specific and Nonspecific

Data Set	"Real" Training Set	"Random" Training Set	Performance	n (Data Set)	n (Anatomy Terms)
Blastoderm	0.45	0.23	specific	77	1
Wing	0.34	0.29	specific	33	7
Endoderm	0.22	0.15	specific	16	13
Mesoderm.1	0.32	0.18	specific	16	2
Ventral_ectoderm	0.34	0.19	specific	12	4
Trachea1 system	0.32	0.15	specific	9	7
Cardiac-mesoderm	0.47	0.07	specific	8	9
Amnioserosa	0.30	0	specific	7	3
Neuroectoderm	0.76	0.14	specific	7	2
Mean (SD)	**0.39 (0.16)**	**0.16 (0.08)**		**22.25 (23.63)**	**4.38 (2.92)**
Neuronal	0.23	0.25	nonspecific	54	31
Ectoderm.2	0.18	0.20	nonspecific	51	16
Imaginal disc.1	0.25	0.25	nonspecific	47	14
Mesoderm.2	0.20	0.18	nonspecific	45	47
Ectoderm.1	0.19	0.17	nonspecific	37	4
CNS	0.33	0.32	nonspecific	34	21
PNS	0.3	0.3	nonspecific	24	15
Mean (SD)	**0.24 (0.06)**	**0.24 (0.06)**		**41.71 (10.58)**	**20.12 (13.27)**
p (specific < nonspecific)				**<0.013**	**<0.002**

"Specific" (column 4) indicates that nucleotide-level sensitivity in LOOCV tests was better (≥ 0.05) when the proper training set was used (column 2) than when a randomized training set was used (column 3). Columns 5 and 6 list the size of data sets and number of anatomy terms used in defining them, respectively. Values for the specific and nonspecific data sets were compared using a one-tailed Wilcoxon rank-sum test. See also Figure S3.

Genome-Wide Prediction of *Drosophila* CRMs

For each of the 15 amenable data sets, taking all known modules as training data, we scanned the noncoding genome of *D. melanogaster* for the highest scoring modules, using the "fusion" method as well as the best individual scoring scheme from the cross-validation results for that data set. We then used gene expression databases (FlyBase, BDGP) to test the quality of these genome-wide predictions. First, we defined a set of genes ("expression gene set") with expression patterns commensurate with those of the CRMs in the data set (see Experimental Procedures). We next took the modules predicted for that data set and extracted their nearest neighboring genes ("predicted gene set"). Finally, we performed hypergeometric tests of enrichment between the expression gene set and the predicted gene set (Table 2). Of the 15 data sets, 8 had an enrichment p value $\leq 10^{-10}$, and 14 had p value ≤ 0.01 either with fusion or the best individual scoring scheme. (Given that 15 × 2 tests were done, 13 are significant at p < 0.05 with Bonferroni correction.) The strongest enrichments were observed for the data sets blastoderm, imaginal disc, and wing. In a negative control experiment, we "predicted" random genes for each data set and performed the same test for enrichment; the enrichment p value was not significant on any data set (Table 2). Thus, intersecting the predictions with gene expression data provides strong support for our genome-wide CRM predictions. Furthermore, taking this intersection gives us a high confidence set of module predictions for each of the 15 data sets, which are made available in an online interface with "Genome Browser" integration (Figure S6). This interface also lists associations with motifs from existing databases, for each data set.

In Vivo Validation of Predictions for the Blastoderm Data Set

The blastoderm data set performed strongly in the cross-validation exercises using each of the scoring methods and had the strongest gene expression pattern enrichment in the genome-wide search. We therefore focused our follow-up efforts on this high-performing data set. Our compendium lists 113 modules as being related to the training CRMs in this data set (see Experimental Procedures and Table S2). Thirty-three of these top modules belong to the training set itself, while an additional 26 are known modules that were not included in the training set (Table S2). The remaining 54 (113 – 33 – 26) modules are novel CRM predictions (also see Table S13).

Although the expression pattern enrichment results and successful prediction of known modules not included in the training set suggest that a large fraction of our predictions are correct, only in vivo testing of the predicted elements can confirm that they have regulatory function. We therefore chose five putative modules, located near the genes *edl*, *srp*, *odd*, *SoxN*, and *cas*, for in vivo validation using a GFP-based reporter construct in transgenic *Drosophila* embryos. (None of these was predicted by the motif-based "Ahab" program in Schroeder et al. [2004].) Remarkably, all five tested constructs showed reporter gene expression in a pattern consistent with the expression of their predicted associated gene (Figure 2). Interestingly, although the five modules were selected using the blastoderm data set as training data, and all five associated genes are expressed during the blastoderm stage of development, only two of the identified CRMs, those for *edl* and *srp*, appear to regulate blastoderm-stage gene expression (Figures 2A, 2B, 2G, and 2H).

Table 2. Genome-Wide CRM Predictions for the 15 Amenable Data Sets, Using the Fusion Method and the Best Individual Method from LOOCV Tests

Data Set	m	N	FUSION		LOOCV		Best LOOCV Method	Random	
			P_F	K_F	P_L	K_L		P_R	K_R
Blastoderm	208	5572	**6.7E−24**	45	**6.7E−28**	49	HexMCD	0.48	8
Cardiac_mesoderm	248	5572	5.4E−01	9	8.9E−01	6	PAC-rc	0.95	5
CNS	846	5572	8.2E−07	57	**1.6E−11**	68	HexDiff-rc	0.27	34
Eye.1	76	14149	2.9E−01	2	4.5E−03	5	D2z-cond-mo1-weights-rc	1.00	0
Imaginal_disc.1	318	14149	**1.9E−19**	33	**1.9E−19**	33	HexMCD	0.83	3
Mesectoderm	93	5572	**7.4E−13**	22	1.8E−03	10	D2z-cond-s100	0.86	2
Mesoderm.1	768	5572	1.5E−02	39	1.3E−03	47	PAC-rc	0.74	25
Neuroectoderm	60	14149	**1.9E−12**	13	1.1E−08	10	HexYMF-s200-rc	1.00	0
PNS	97	5572	8.0E−06	14	3.8E−05	13	PAC-rc	0.27	5
Somatic_muscle	76	14149	9.3E−09	11	2.9E−01	2	D2z-cond-weights	1.00	0
Ventral_ectoderm	328	5572	**7.0E−13**	41	**5.9E−11**	38	PAC-rc	0.13	16
Visceral_mesoderm	134	5572	8.3E−05	15	5.1E−06	17	HexMCD	0.11	8
Ectoderm.2	841	5572	**3.8E−12**	69	1.1E−10	66	HexMCD	0.26	34
Neuronal	66	5572	5.4E−04	9	1.1E−04	10	HexMCD	0.91	1
Wing	30	14149	**5.3E−15**	20	**9.9E−11**	16	HexDiff-rc	0.86	1

The hypergeometric p values of enrichment (P_F,P_L,P_R) between the predicted CRMs for a data set and the corresponding expression-based gene sets are shown, along with the size of the intersection (K_F,K_L,K_R). Column m = size of expression gene set, Column N = total number of genes in the expression data source. The column "Random" is the result of a negative control, where CRM predictions were replaced by random locations.

Both of these also regulate expression at later stages in development. In the case of *edl*, this later expression is a subset of the complete expression pattern of the gene (Figures 2C–2F), consistent with the notion of modular gene regulation through the action of multiple CRMs for each gene (Arnone and Davidson, 1997). Expression driven by the *srp* CRM more comprehensively covers the range of endogenous *srp* expression, in addition to driving apparently ectopic reporter gene expression in the midgut of later stage embryos (Figures 2I and 2J, and data not shown). This may represent actual ectopic expression, but may also simply be perdurance of GFP expression in tissues that developed from earlier *srp*-expressing progenitors. The *SoxN* and *cas* CRMs closely recapitulate subsets of the central nervous system expression of their respective associated genes from mid-embryogenesis on (Figures 2M–2T), while the *odd* CRM recapitulates native *odd* expression in a number of tissues including mid-stage mesodermal progenitor cells and late-stage fat body (Figures 2K and 2L, and data not shown). Ectopic reporter gene expression is observed in cardioblasts (data not shown). This may again merely represent perdurance of the GFP reporter in a subset of mesodermal cells, or may indicate improper activity of the CRM in descendants of a common cell lineage (a subset of Odd-negative cardioblasts are sibling cells to Odd-positive pericardial cells, arising from a common progenitor [Ward and Skeath, 2000]). Detailed experimental analysis will be needed to distinguish between these possibilities.

Application to Data Sets of Tissue-Specific Enhancers from Mouse
Encouraged by these strong validation results, we also sought to determine whether our methods would work effectively for prediction of mammalian CRMs. We constructed eight new data sets, comprising a total of 244 CRMs known to drive expression

in specific tissues in mouse (Pennacchio et al., 2006). We performed leave-one-out cross-validation as in the *Drosophila* analysis above, with each of the seven motif-blind methods. Seven of the eight data sets were amenable to consistent prediction by at least one method, over ten LOOCV instantiations (Experimental Procedures; Figure 3; Table S3); as observed previously, no one method proved universally superior. (Also see results on cross-validation with native flanks [Figure S5; Table S5].) The accuracy is especially noteworthy on the forebrain, hindbrain rhombencephalon, and neural tube data sets, being at 75%–85% at the CRM level, and 58%–76% at the nucleotide level (Figure 3; Table S6). Overall, roughly 60% of the full complement of enhancers tested were recovered by supervised prediction in a cross-validation setting, making a strong case for the generalizability of the approach.

Application to Early Blood and Cardiovascular Development in Human/Mouse
To follow up on these results in a genome-wide manner, we turned to a data set of ten enhancers known to function in the developing blood and vasculature (Supplemental Experimental Procedures). We used the two top-scoring methods from LOOCV, HexYMF-s200-rc (54% sensitivity) and PAC-rc (48% sensitivity), to predict CRMs in the human genome. For efficiency, the search was limited to "evolutionarily conserved regions" (ECRs) (Loots and Ovcharenko, 2007) based on human-mouse conservation. The top 1000 CRM predictions (of each method) were then assessed for enrichment for a set of genes known to be differentially expressed in blood stem cells. Predicted CRMs with a neighboring gene in this set were counted, and they were found to be highly statistically significant when using either method (z scores > 10 and > 6, respectively, see Figure 3C).
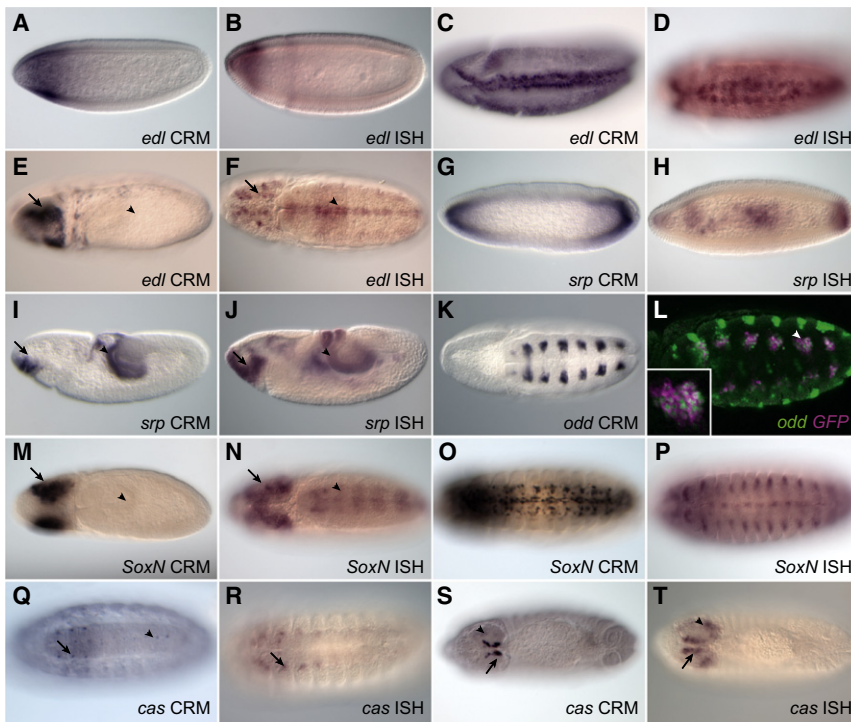
Figure 2. In Vivo Validation of *Drosophila* CRM Predictions

Each predicted CRM was used to drive a GFP reporter gene in transgenic embryos. Expression was visualized using antibodies to GFP and compared to the endogenous expression of the putative associated gene as determined by whole-mount in situ hybridization to mRNA (in the case of *odd*, Odd antiserum). All embryos are shown with anterior to the left. (A) The *edl* CRM drives expression in the anterior of blastoderm-stage embryos, consistent with endogenous *edl* expression (B). At embryonic stage 10, reporter gene expression (C) mimics *edl* gene expression (D) at the ventral midline. By stage 11, reporter gene expression is no longer seen at the ventral midline despite continued endogenous gene expression at this site (arrowheads in E and F); however, reporter gene expression is maintained in the developing brain (arrows in E and F). The *srp* CRM reporter gene (G) is expressed similarly to *srp* mRNA (H) in the blastoderm as well as in stage 8 embryos in the developing posterior endoderm ([I and J], arrowheads) and anterior endoderm ([I and J], arrows). The *odd* CRM recapitulates *odd* gene expression in mesodermal progenitors ([K], magenta in L) but not in the ecotderm ([L]; anti-Odd is in green and is nuclear, GFP expression driven by the Odd CRM reporter is in magenta and primarily cytoplasmic). The inset in (L) shows a close up of the mesodermal cluster marked by the arrowhead. (M–P) The *SoxN* CRM drives reporter gene expression in a subset of the endogenous mRNA pattern in the central nervous system. At stage 9, reporter gene expression is visible in the brain ([M], arrow) but not the remainder of the central nervous system ([M], arrowhead; compare with N). By stage 14, additional nervous system expression is apparent (O), consistent with endogenous *SoxN* (P). (Q–T) *cas* CRM reporter gene expression faithfully reproduces *cas* expression in the late embryonic ventral nerve cord and brain. Nerve cord expression of both the reporter gene and the endogenous mRNA is most prominent in a set of lateral cells in the thoracic segments (arrows in Q and R); additional reporter gene expression in midline cells ([Q], arrowhead) is likely perdurance from a slightly earlier stage of expression (data not shown). Brain expression of the reporter gene, like that of *cas* mRNA, is strongest in a group of medial cells ([S and T], arrows), with weaker expression laterally ([S and T], arrowheads).

All ten training CRMs contain consensus binding sites for the Ets and GATA families of transcription factors. We therefore tested the top 1000 CRM predictions for the presence of GATA and Ets motifs (Figure 3D). We found 234 of the top 1000 predictions from PAC-rc to have the Ets motif; comparing this to a random expectation of 59, we see a four-fold enrichment and a z score of >45. Similar enrichment for the GATA motif is observed (z > 13) in the PAC-rc predictions and for both motifs in the HexYMF-200-rc predictions (z > 7, Figure 3D).

We constructed a high-confidence CRM prediction set for each method by taking the 1000 top predictions, requiring that a neighboring gene be in the blood stem-cell gene set, and specifying that either the Ets or GATA motif be present. This led to 75 distinct predicted CRMs based on HexYMF-200-rc and 114 based on PAC-rc (Table S4).

In Vivo Validation in Mouse

To demonstrate that the above approach was indeed able to identify regulatory modules and predict their in vivo biological activity, we generated lacZ reporter constructs for two predicted intronic elements for testing in transgenic mouse embryos. Neither of the two respective gene loci has previously been implicated in either blood or cardiovascular development. The first gene (*ebf3*) encodes a little-known paralog of early B cell factor *ebf1*, a helix-loop-helix transcription factor important during early B lymphocyte development (Busslinger et al., 2000). (The

predicted CRM is ranked at 41 by HexYMF-s200-rc and 63 by PAC-rc.) The second gene (*c1orf164*) corresponds to an uncharacterized open reading frame on chromosome 1 predicted to encode a ring-finger-domain-containing protein. (The predicted CRM is ranked 12 by HexYMF-s200-rc and 23 by PAC-rc.)

Multiple transgenic founders were generated for each construct, and day E11.5 embryos were collected and stained for reporter gene activity. As exemplified by a representative *lyl1* promoter transgenic embryo, the ten elements used as training data all show tissue specific enhancer activity by driving expression in blood vessels, the heart, and developing blood cells in the fetal liver (Figure 4). The transgenic embryos generated with the *ebf3* and *c1orf164* candidate enhancers reproducibly showed transgene expression in two and three (respectively) of these three tissues, thus demonstrating that the computational screen not only led to the identification of bona fide transcriptional enhancers, but more importantly was able to predict the tissue-specific activity of these elements.

DISCUSSION

The problem of supervised CRM prediction is easily motivated when we consider that the interactions in a regulatory subnetwork involve up to hundreds of genes (Davidson, 2006) and a relatively small set of TFs. Clearly, there is a significant layer of combinatorial regulation in between, implemented through
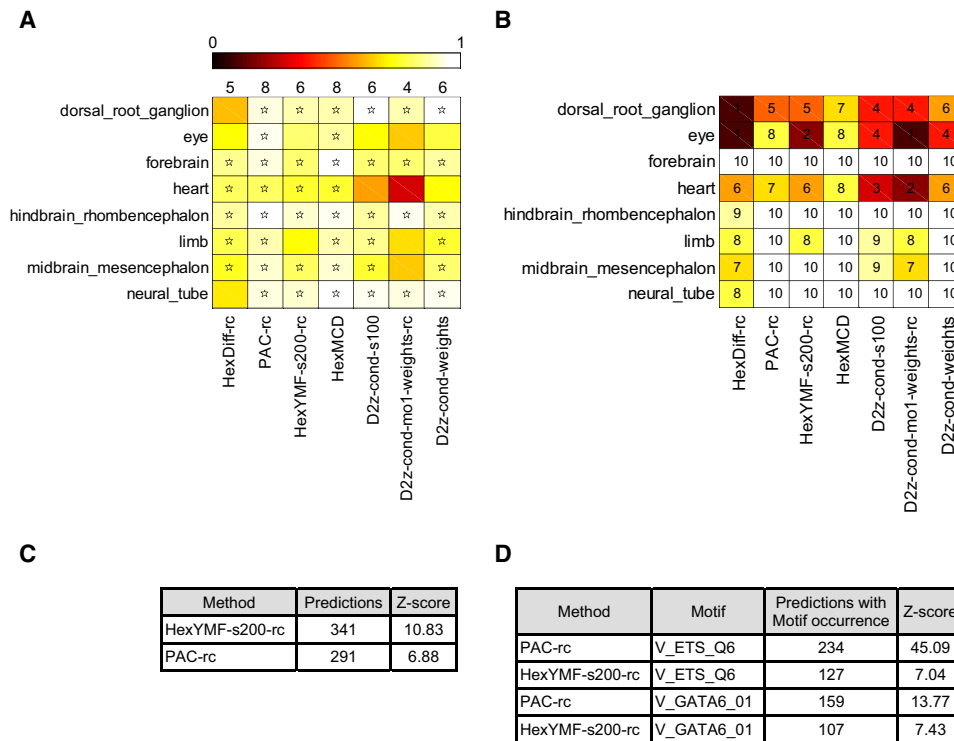
Figure 3. *In silico* Validation on Mammalian Enhancers

(A) LOOCV performance of each of seven different motif-blind methods on eight data sets in the benchmark. The format is identical to that of Figure 1A.

(B) Consistency of a method's accuracy on each data set, over ten instantiations of LOOCV. Colors and numbers indicate the number of instantiations on which the performance was significant (at $p \leq 0.05$).

(C) The set of nearest genes for the top 1000 CRM predictions for early blood and cardiovascular development in human/mouse was intersected with a set of 7035 genes differentially expressed in blood stem cells in mouse (Miranda-Saavedra et al., 2009). The intersection size ("Predictions") was significantly above chance expectation ("z score"; mean 204, standard deviation 13, estimated by simulations).

(D) Predicted CRMs were scanned for significant occurrence of Ets-related and GATA motifs from TRANSFAC; counts of motif-containing predictions are in column 3, and significance estimates of these counts, based on random sampling of CRMs, are in column 4.

cis-regulatory modules. We thus expect tens to hundreds of modules that share some degree of similarity in their binding site content, and we should be able to predict most of these given a representative subset. Such an initial set of modules typically will be obtained through reporter gene assays, computational methods, or high-throughput technologies such as ChIP-chip. Our algorithms will then leverage the initial set to provide much greater coverage of the regulatory subnetwork. We have demonstrated here that our methods apply similarly to both the *Drosophila* and mouse genomes with high accuracy.

It is important to contrast our method with the "tissue-specific CRM prediction" approaches undertaken in Chen and Blanchette (2007), Hallikas et al. (2006), Pennacchio et al. (2007), Smith et al. (2007), and Yu et al. (2007) for the human genome. All of these methods rely upon a large collection of vertebrate TF motifs (from TRANSFAC [Matys et al., 2003] and/or JASPAR [Bryne et al., 2008]), which is their main point of difference from our approach. Moreover, their problem formulation and data assumptions are distinct from ours: their strategy hinges on large-scale gene expression data across a large spectrum of tissues (Chen and Blanchette, 2007; Pennacchio et al., 2007; Smith et al., 2007; Yu et al., 2007) or on knowledge of TFs mediating the specific transcriptional response (Hallikas et al., 2006).

In many cases, such information may not be available. We require instead the prior knowledge of some of the CRMs involved in a particular regulatory subnetwork (which could be a tissue-specific subnetwork). The "EDGI" program of Sosinsky et al. (2007) has a similar objective to ours, i.e., CRM prediction without motif knowledge, relying instead on interspecies conservation and clustering of binding sites, but the only published tests of this method have been on the A/P patterning network in *Drosophila*, precluding statements about its broader applicability. The most definitive test of any method is its ability to predict CRMs that function in vivo. Current success rates based on in vivo validation top out at approximately 80% in *Drosophila* and 70% in mouse (Table S8). Although continued validation using a larger number of predictions drawn from a broader selection of data sets is still required, the 100% true-positive rate we have achieved so far in both fly and mouse is highly encouraging and at a minimum puts our method on a par with the top existing approaches.

Recent advances in genome-scale empirical methods represent a promising new means for CRM discovery and will provide an important complement to, although not a replacement for, computational methods. For example, Visel et al. (2009) have described a ChIP-seq-based study in which CRM sequences
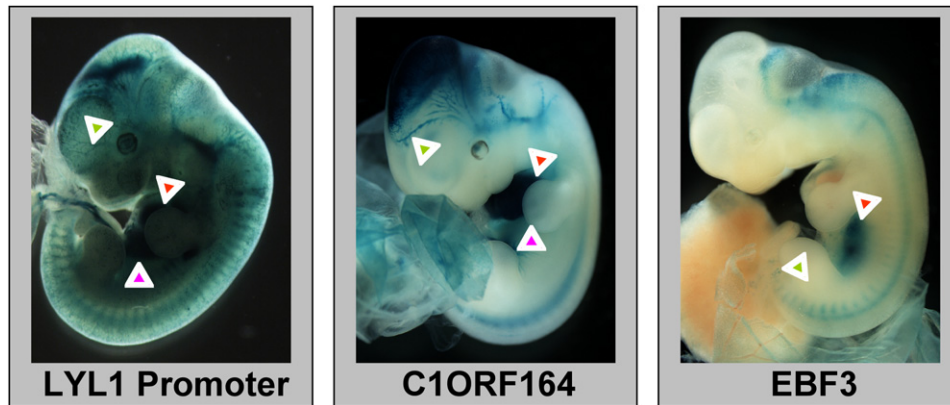
Figure 4. Validation of Predicted CRMs for In Vivo Activity in Blood and Cardiovascular Tissues Using Transgenic Mouse Assays
LacZ reporter constructs were microinjected into mouse embryos and assayed by whole-mount staining of midgestation F0 embryos. Shown to the left is a transgenic embryo with the *lyl1* promoter construct showing transgene expression in heart (red arrowhead), fetal liver (purple arrowhead), and vessels (green arrowhead). The middle and right hand panels show representative transgenic embryos with reporter constructs for the predicted CRMs in exon 3 of *c1orf164* and exon 6 of *ebf3*, respectively. Arrowheads indicate staining in heart, fetal liver, and vessels.

were enriched through chromatin immunoprecipitation of the common enhancer-binding protein p300 using RNA isolated from specific tissues. At present, such approaches require significant amounts of biological material, which represents a particular problem when studying stem cell systems or early developmental programs. Employing a computational strategy not only circumvents the need for pure cell populations but also has the potential to provide information on *cis*-regulatory elements operating in all cell types. Moreover, while tissue-specific p300-directed ChIP-Seq can reveal that two modules are active in the same tissue, it makes no predictions as to whether they may be related in terms of their control mechanisms or the specific subnetworks in which they participate. Importantly, our methods should be easily adaptable for assessing which empirically identified CRMs are functioning through related mechanisms, and data from empirical methods will therefore provide valuable input to the computational discovery approaches we have outlined here. Consequently, the approach developed here represents a widely applicable strategy for deciphering transcriptional regulatory networks across a wide range of model systems.

Methodology

We recommend use of the cross-validation step as a quick method to assess whether a set of modules is amenable to computational prediction and, if so, which scores are good at capturing the essence of these regulatory sequences. If the user finds that his or her data set does not show consistently significant performance with any of our methods, he or she should not proceed further with our pipeline. Although the absolute performance values in cross-validation do not pertain to genome-wide prediction accuracy, the empirical p values do give us an idea of whether the supervised prediction scheme is feasible for the data set.

Our CRM prediction pipeline encapsulates a broad variety of scoring schemes to capture the essential features of functionally related modules. We implemented two previously reported scoring schemes (HexDiff and HexMCD), designed five novel motif-blind scoring schemes (HexYMF, PAC, and three variants

of D2z), and examined statistical issues related to them (such as normalization for background composition). Our results (Figures 1B and 1C) show that each of the scoring schemes explored has its merits, and there is no universally superior method. At the same time, some general trends may guide us in our search for better scores. For instance, counting words on both strands improved performance of most scores (see Table S1). Using a subset of 100–200 6-mers rather than all k-mers was another beneficial choice which presumably increases the signal by removing words unrelated to the true motifs in the data. It is also clear that motif-blind approaches are competitive with or better than the motif-based Stubb-MDB (or Clover-Cluster-Buster). In fact, the data sets on which Stubb-MDB is the "best performing" method (neuroectoderm, ventral ectoderm, and ectoderm.2) are those on which almost all methods are successful. We believe this will in general be the case for motif-based methods, although a comprehensive test of existing approaches (e.g., Philippakis et al., 2005) has not yet been conducted.

We have used slightly different pipelines for genome-wide prediction in *Drosophila* and human—the fusion method and the best LOOCV method were deployed in *Drosophila*, while the two best LOOCV methods were used in the human scans. This is meant to show that the underlying statistical scoring schemes may be used in a variety of ways that can be decided by the user. Also, the vertebrate analysis filters the CRM predictions for presence of either the GATA or the Ets motif, while the *Drosophila* analysis does not impose any motif filter. Again, this demonstrates the flexibility of including optional filters, which could potentially aid in refining the tissue specificity of predictions, based on the user's prior knowledge of the biological system. Note that even in the vertebrate case the motif filters were used postsearch to prioritize the results; the search itself, like the *Drosophila* search, was conducted in a fully motif-blind fashion.

Characterization of Data Sets

The strong correlations seen between performance and simple properties of data sets—size, GC content, etc. (Figure S2A–S2C)—are suggestive of a possible source of "contamination"

in the training sets. Experimentally identified modules are not always the minimal sequence required for the specific function and may include nonregulatory sequences flanking the functionally important core(s). Since CRMs in general demonstrate a higher degree of evolutionary sequence conservation, GC content, and, in some cases, homotypic motif clustering than their flanking regions (Li et al., 2007), data sets with relatively lower values of these variables may be the ones with greater "contamination," which may partially explain their relatively low amenability to prediction.

Genome-Wide Predictions

One may define a successful prediction as either a module with expression consistent with the training set, or as a module capturing some aspect of its associated gene's expression pattern. While we consider the latter condition to be necessary, the former condition depends largely on how tightly defined the training set is and on the nature of the TFs that act via the training CRMs. These TFs will frequently also regulate other biological processes and may lead our supervised prediction framework to report modules related to those processes. For instance, of the five tested *Drosophila* modules, only two (*edl*, *srp*) mediated the predicted blastoderm gene expression, but three drove expression in the brain and midline of the central nervous system (*edl*, *cas*, *SoxN*); suggestively, many of the gap and pair-rule genes that regulate gene expression in the blastoderm also act during nervous system development. It is also possible that a training set spans a very broad spectrum of expression patterns, making it hard to learn the *cis*-regulatory commonalities from them. In such cases, the supervised prediction may learn generic characteristics of CRMs and predict successfully, but without specificity, or may fail completely. This may explain why smaller data sets with more tightly defined expression characteristics led to more specific CRM detection in our LOOCV experiments.

EXPERIMENTAL PROCEDURES

New Scoring Schemes
HexYMF
The z score of a word is calculated by the YMF program (Sinha and Tompa, 2000), based on its count in training CRMs, and the mean and standard deviation are calculated from a third-order Markov chain trained on background sequences. In our implementation, called "HexYMF-s200-rc," W comprises the 200 top-ranking 6-mers by z score (hence "-s200"), and $n(w)$ denotes the count of w on both strands of a sequence.
Poisson Additive Conditional
This score is defined as:

$$\frac{1}{|W|}\sum_{w \in W} F(\lambda(w), n(w) - 1),$$

where $F(\lambda, x)$ is the cumulative Poisson distribution function, $\lambda(w)$ is the expected count of w in the test sequence, and $n(w)$ is its observed count. W is the set of the most overrepresented words in training CRMs, as defined above. Note that $1 - F(\lambda(w), n(w) - 1)$ represents the p value of the observed count of w. This score considers the words that are most associated with the training set, and then examines how overrepresented each of these words is in the test sequence, relative to the assumed background. The implementation is called "PAC-rc" because word counts on both strands are considered in identifying the set W.
D2z Score
The "D2 statistic" is the number of k-mer matches between two given sequences, and the "D2z score" introduced in our earlier work (Kantorovitz

et al., 2007) computes the statistical significance (z score) of this number. Here, we cannot use the calculations from Kantorovitz et al. (2007) (also used in Ivan et al., 2008) because the model under which the z score was computed is inapplicable in the supervised prediction setting. In the null model of Kantorovitz et al. (2007), both sequences are random sequences, while in our setting the training sequence S is known (treated as a fixed sequence) and only the test sequence T is random (in the null model). We show how to analytically calculate the z score for D2 under this setting and call it conditional-D2z. In addition, we develop the following variations of the conditional D2z score. (All derivations are provided in Supplemental Experimental Procedures.)

(1) Subsets of words: This restricts the summation to k-mers $w \in W$ defined above.
(2) Weighted summation: Here, the D2 score is redefined to be a weighted dot product, with z(w) (see HexYMF) as weights; i.e., we redefine D2 as $\sum_{w \in W} z(w) n_S(w) n_T(w)$ and compute its z score.
(3) Reverse complement counting: We extended the D2z score of Kantorovitz et al. (2007) to count words on both strands while ignoring statistical dependencies between the strands.

The three variants of D2z score that are used in our final pipeline include "D2z-cond-mo1-weights-rc," "D2z-cond-weights," and "D2z-cond-s100" ("-cond" for conditional z score, "-mo1" for first-order Markov background, "-weights" for weighted summation, "-rc" for counting words on both strands, and "-s100" for subset of top 100 words).
Stubb-MDB
Given a set of experimentally characterized motifs (PWMs) and a set S of training CRMs, the first step determines the relevance of each motif M to S, as follows:
(1) Calculate the "log likelihood ratio score" of M for each sequence s in S as:

$$LLR(s, M) = \log \frac{Pr(s|HMM(M, p_{free}))}{Pr(s|HMM(M, p_{global}))},$$

where HMM(M, p) is a two-state zeroth order HMM (Sinha et al., 2003), with motif transition probability p and background transition probability $1 - p$. The value p_{global} in the denominator is the maximum likelihood value of this parameter learned from genomic background, while p_{free} in the numerator is a free parameter trained on sequence s.
(2) Calculate the empirical p value of LLR(s,M) based on scores of equal length windows in genomic background.
(3) Declare M to be relevant to S if more than 10% of sequences s in S have a p value (LLR(s,M)) ≤ 0.05.
The second step is to filter the relevant motifs for redundancies based on relative entropy between two PWMs. The final step is to scan the test sequence for CRMs using Stubb (Sinha et al., 2003), using the top ten relevant, nonredundant motifs.

Evaluation of CRM Prediction by Cross-Validation
If there are n CRMs in a data set, the cross-validation was done in n "folds." In each fold, one CRM is the test data, and the remaining $n-1$ CRMs are training data. The former is embedded in a 10 kbp long noncoding genomic sequence with G/C content similar to the native flank of the CRM, thus creating the "test sequence." For each data set, the average length L of known modules was computed, and the task in each fold was to predict a module of length L in the test sequence. The highest scoring window by each scoring scheme was treated as the prediction of that scheme. We obtain L-length predictions in all n test sequences, which are then evaluated based on overlap with the embedded CRMs. We use two measures of accuracy: (1) "nucleotide level sensitivity," determined by the base-pair overlap, and (2) "CRM level sensitivity," determined by the number of folds where the true and predicted CRMs overlap by 100 bp or more. Due to this experimental design, the sensitivity values are equal to the respective precision (PPV) values. We further compute empirical p values of the sensitivity, as in Ivan et al. (2008). Note that the test CRM is embedded in a randomly chosen noncoding region, instead of being kept in its native flank, in order to maximize the odds that there are no other related modules in a test sequence. In different "instantiations" of cross-validation, the randomly chosen flanking sequences are different.

Finally, we note that the p value threshold of 0.05 used in designating success on a data set is meant only as a guide to choose methods, and not for biological discovery. As such, no multiple hypothesis correction is applied here.

Relationship between Performance and Data Set Properties
We measured the relative extent of homotypic clustering in a CRM by its "FTT-Z rank" (Li et al., 2007), which computes the FTT (fluffy-tail-test) score (Abnizova et al., 2005) of the sequence, calculates its empirical z score, and ranks the CRM by this FTT-Z score. Evolutionary conservation with *D. pseudoobscura* was calculated as in Li et al. (2007).

Genome-Wide Scans and Validation
The *D. melanogaster* genome (Release 4.3) was masked for exons and for short tandem repeats (Benson, 1999), and the CRM prediction pipeline was run with 500 bp windows. For each data set, nearest neighboring genes of the top scoring modules genome-wide were extracted until we had a list of 200 distinct genes. This list was tested for enrichment for the corresponding gene set defined from expression data, defined by specifying the appropriate sets of anatomical terms at BDGP (http://www.fruitfly.org/) or at FlyBase (http://flybase.org/) (Table S10). For the blastoderm data set, we further considered the subset of the 200 predicted genes that have an anterior-posterior (A/P) pattern and obtained 137 modules, of which 113 were within 10 kbp of the proximal gene.

Drosophila Transgenic Analysis
Genomic sequences were generated by PCR (Table S9) and subcloned into an EGFP reporter vector (details available on request). Transgenic flies were created by injection into line ϕX-96E (Bischof et al., 2007). Homozygous transgenic embryos were collected, fixed, and stained with antibodies to GFP (mAb JL-8, Clontech) using standard methods. Labeling for Odd expression used guinea pig anti-Odd (Ward and Skeath, 2000). Probes for in situ hybridization were generated using clones from the *Drosophila* Gene Collection (Stapleton et al., 2002).

Mouse Reporter Constructs and Transgenic Analysis
Reporter constructs were generated using primers listed in Table S9 and confirmed by sequencing. Detailed information on reporter constructs is available on request. Plasmids were linearized, and founder transgenic embryos were produced by pronuclear injection as described (Landry et al., 2005). Embryos were harvested at E11.5 and analyzed as described (Pimanda et al., 2006).

Mammalian Data Sets for Cross-Validation
Tissue-specific human enhancers were downloaded from the Vista Enhancer Browser (http://enhancer.lbl.gov/), and enhancers driving expression in a specific tissue were considered. These were grouped by the associated tissue, and groups with size ≥5 were treated as data sets (covering 244 enhancers overall). Cross-validation was performed by "planting" each CRM in 10 kbp random noncoding sequence (Figure 3 and Table S6), as well as in local context (Figure S5 and Table S5). A data set was considered amenable to a method if the sensitivity p value was ≤0.05 on at least eight of ten instantiations of LOOCV.

SUPPLEMENTAL DATA

Supplemental Data include Supplemental Experimental Procedures, six figures, and thirteen tables and can be found with this article online at http://www.cell.com/developmental-cell/supplemental/S1534-5807(09)00384-0.

ACKNOWLEDGMENTS

REFERENCES

Abnizova, I., te Boekhorst, R., Walter, K., and Gilks, W.R. (2005). Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the Drosophila genome: the fluffy-tail test. BMC Bioinformatics 6, 109.

Arnone, M.I., and Davidson, E.H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. Development 124, 1851–1864.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. 24, 1429–1435.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl. Acad. Sci. USA 99, 757–762.

Bischof, J., Maeda, R.K., Hediger, M., Karch, F., and Basler, K. (2007). An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. Proc. Natl. Acad. Sci. USA 104, 3312–3317.

Bryne, J.C., Valen, E., Tang, M.H., Marstrand, T., Winther, O., da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. 36, D102–D106.

Busslinger, M., Nutt, S.L., and Rolink, A.G. (2000). Lineage commitment in lymphopoiesis. Curr. Opin. Immunol. 12, 151–158.

Carroll, S.B., Grenier, J.K., and Weatherbee, S.D. (2001). From DNA to Diversity. Molecular Genetics and the Evolution of Animal Design (Massachusetts: Blackwell Science).

Chan, B.Y., and Kibler, D. (2005). Using hexamers to predict cis-regulatory motifs in Drosophila. BMC Bioinformatics 6, 262.

Chen, X., and Blanchette, M. (2007). Prediction of tissue-specific cis-regulatory modules using Bayesian networks and regression trees. BMC Bioinformatics 8 (Suppl 10), S2.

Davidson, E.H. (2006). The Regulatory Genome: Gene Regulatory Networks in Development and Evolution (Burlington, MA: Academic Press).

Donaldson, I.J., Chapman, M., and Gottgens, B. (2005). TFBScluster: a resource for the characterization of transcriptional regulatory networks. Bioinformatics 21, 3058–3059.

Frith, M.C., Li, M.C., and Weng, Z. (2003). Cluster-Buster: finding dense clusters of motifs in DNA sequences. Nucleic Acids Res. 31, 3666–3668.

Frith, M.C., Fu, Y., Yu, L., Chen, J.F., Hansen, U., and Weng, Z. (2004). Detection of functional DNA motifs via statistical over-representation. Nucleic Acids Res. 32, 1372–1381.

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

Halfon, M.S., Gallo, S.M., and Bergman, C.M. (2008). REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. Nucleic Acids Res. 36, D594–D598.

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

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

Ivan, A., Halfon, M.S., and Sinha, S. (2008). Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs. Genome Biol. *9*, R22.

Kantorovitz, M.R., Robinson, G.E., and Sinha, S. (2007). A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics *23*, i249–i255.

Landry, J.R., Kinston, S., Knezevic, K., Donaldson, I.J., Green, A.R., and Gottgens, B. (2005). Fli1, Elf1, and Ets1 regulate the proximal promoter of the LMO2 gene in endothelial cells. Blood *106*, 2680–2687.

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

Li, X.Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C.L., et al. (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol. *6*, e27.

Loots, G., and Ovcharenko, I. (2007). ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. Bioinformatics *23*, 122–124.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. *31*, 374–378.

Miranda-Saavedra, D., De, S., Trotter, M.W., Teichmann, S.A., and Gottgens, B. (2009). BloodExpress: a database of gene expression in mouse haematopoiesis. Nucleic Acids Res. *37*, D873–D879.

Nazina, A.G., and Papatsenko, D.A. (2003). Statistical extraction of Drosophila cis-regulatory modules using exhaustive assessment of local word frequency. BMC Bioinformatics *4*, 65.

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

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499–502.

Pennacchio, L.A., Loots, G.G., Nobrega, M.A., and Ovcharenko, I. (2007). Predicting tissue-specific enhancers in the human genome. Genome Res. *17*, 201–211.

Philippakis, A.A., He, F.S., and Bulyk, M.L. (2005). Modulefinder: a tool for computational discovery of cis regulatory modules. Pac. Symp. Biocomput., 519–530.

Pimanda, J.E., Chan, W.Y., Donaldson, I.J., Bowen, M., Green, A.R., and Gottgens, B. (2006). Endoglin expression in the endothelium is regulated by Fli-1, Erg, and Elf-1 acting on the promoter and a −8-kb enhancer. Blood *107*, 4737–4745.

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

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

Sinha, S., and Tompa, M. (2000). A statistical method for finding transcription factor binding sites. Proc. Int. Conf. Intell. Syst. Mol. Biol. *8*, 344–354.

Sinha, S., van Nimwegen, E., and Siggia, E.D. (2003). A probabilistic method to detect regulatory modules. Bioinformatics *19* (Suppl 1), i292–i301.

Smith, A.D., Sumazin, P., and Zhang, M.Q. (2007). Tissue-specific regulatory elements in mammalian promoters. Mol. Syst. Biol. *3*, 73.

Sosinsky, A., Honig, B., Mann, R.S., and Califano, A. (2007). Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting. Proc. Natl. Acad. Sci. USA *104*, 6305–6310.

Stapleton, M., Liao, G., Brokstein, P., Hong, L., Carninci, P., Shiraki, T., Hayashizaki, Y., Champe, M., Pacleb, J., Wan, K., et al. (2002). The Drosophila gene collection: identification of putative full-length cDNAs for 70% of D. melanogaster genes. Genome Res. *12*, 1294–1300.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotechnol. *23*, 137–144.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Ward, E.J., and Skeath, J.B. (2000). Characterization of a novel subset of cardiac cells and their progenitors in the Drosophila embryo. Development *127*, 4959–4969.

Yu, X., Lin, J., Zack, D.J., and Qian, J. (2007). Identification of tissue-specific cis-regulatory modules based on interactions between transcription factors. BMC Bioinformatics *8*, 437.