Published online 5 August 2011 *Nucleic Acids Research, 2011, Vol. 39, No. 22* ***9463–9472***
doi:10.1093/nar/gkr621

Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison

Majid Kazemian[1], Qiyun Zhu[2], Marc S. Halfon[2,3,]* and Saurabh Sinha[1,4,]*

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, [2]Department of Biological Sciences, [3]Department of Biochemistry and NY State Center of Excellence in Bioinformatics & Life Sciences, SUNY-University at Buffalo, NY 14260 and [4]Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Received April 27, 2011; Revised July 12, 2011; Accepted July 16, 2011

ABSTRACT

Despite recent advances in experimental approaches for identifying transcriptional *cis*-regulatory modules (CRMs, 'enhancers'), direct empirical discovery of CRMs for all genes in all cell types and environmental conditions is likely to remain an elusive goal. Effective methods for computational CRM discovery are thus a critically needed complement to empirical approaches. However, existing computational methods that search for clusters of putative binding sites are ineffective if the relevant TFs and/or their binding specificities are unknown. Here, we provide a significantly improved method for 'motif-blind' CRM discovery that does not depend on knowledge or accurate prediction of TF-binding motifs and is effective when limited knowledge of functional CRMs is available to 'supervise' the search. We propose a new statistical method, based on 'Interpolated Markov Models', for motif-blind, genome-wide CRM discovery. It captures the statistical profile of variable length words in known CRMs of a regulatory network and finds candidate CRMs that match this profile. The method also uses orthologs of the known CRMs from closely related genomes. We perform *in silico* evaluation of predicted CRMs by assessing whether their neighboring genes are enriched for the expected expression patterns. This assessment uses a novel statistical test that extends the widely used Hypergeometric test of gene set enrichment to account for variability in intergenic lengths. We find that the new CRM prediction method is superior to existing methods. Finally, we experimentally validate 12 new CRM predictions by examining their regulatory activity *in vivo* in *Drosophila*; 10 of the tested CRMs were found to be functional, while 6 of the top 7 predictions showed the expected activity patterns. We make our program available as downloadable source code, and as a plugin for a genome browser installed on our servers.

INTRODUCTION

Temporal and spatial regulation of gene expression results from the interaction of transcription factors (TFs) with specific *cis*-regulatory DNA sequences. These regulatory sequences are typically organized in a modular fashion with each module containing one or more binding sites for a specific combination of TFs (1). A '*cis*-regulatory module' (CRM) can thus be defined as a collection of TF-binding sites that function jointly to regulate a discrete aspect of a gene's expression pattern. CRMs are generally described as being short (<1000 bp) contiguous stretches of DNA, located few to hundreds of kilobases away from their associated gene, and are found five-prime, three-prime and within introns of the gene. Davidson (2) estimates that there may be as many as 10-fold more CRMs than genes, yet the vast majority of CRMs have not been identified. The problem of CRM discovery is complicated by the fact that unlike protein-coding regions, which have recognizable sequence features such as open reading frames and codon-usage biases, no similar properties are known for CRMs; and unlike promoters, which by definition lie immediately five-prime to the gene and which contain a limited number of well-defined sequence motifs (3), CRMs are constrained neither in location nor by motif composition.

*To whom correspondence should be addressed. Tel: +1 716 829 3126; Fax: +1 716 829 2725; Email: mshalfon@buffalo.edu
Correspondence may also be addressed to Saurabh Sinha. Tel: +1 217 333 3233; Fax: +1 217 265 6494; Email: sinhas@illinois.edu

The traditional approach to CRM identification involves a tedious process of testing many sequence fragments for regulatory activity in a reporter gene assay. Recently developed genomic techniques have led to relatively rapid screens for potential gene-regulatory regions; such techniques include chromatin immunoprecipitation coupled to genomic tiling arrays (ChIP-chip) (4) and ultra high-throughput sequencing (ChIP-Seq) (5). Despite their great promise, these empirical methods have limitations: it is currently infeasible to assay all tissue types under all conditions, and potential CRMs may be missed by these techniques. In recent years, computational methods have provided an attractive alternative for module identification. Computational methods (6–11) typically start with a small set of relevant TFs and their binding motifs, and search the genome for clusters of putative binding sites (motif matches). However, this approach is ineffective if the relevant TFs and/or their binding specificities are unknown. Motif databases for *Drosophila* currently catalog only a fraction of the estimated number of TFs (12–14). Intense efforts are being made to characterize TF-binding specificities in mouse and fly (15,16), but these efforts are labor intensive and expensive, and the problem of sparse motif knowledge may thus persist for scientists studying organisms other than human, mouse and fruit fly. Moreover, for the majority of regulatory systems, some or most of the relevant TFs have not yet been identified. For these reasons, the application of motif-based computational methods has been limited to a few well-understood biological systems, where the necessary prior knowledge is available. Specialized computational tools to discover motifs from the training data may alleviate the problem, but the modest success rate of motif finding programs (especially for metazoan genomes), as suggested by past surveys (17–19), casts doubts upon the prospect of CRM discovery based on computational motif finding.

This article examines CRM prediction in the common scenario where knowledge of the relevant TFs and/or their binding specificities (motifs) is missing. In particular, we ask the following question: suppose a small set of modules participating in a transcriptional network is known *a priori*. Can we use such information as 'training data' to guide the search for other modules in that network? We call this task the 'supervised CRM prediction problem'. The success of a CRM prediction may be defined as the ability of the predicted sequence to drive a discrete temporal and/or spatial expression pattern. An additional, stricter criterion for success is that the predicted CRM recapitulates the common expression pattern of the training CRMs.

Here, we undertake supervised CRM discovery without assuming motif knowledge or the ability to accurately discover the relevant motifs *ab initio*. In a previous publication (20), we proposed various statistics to capture the sequence similarity (due to shared TF-binding sites) between a candidate CRM and the training set of modules. Our study included previously reported approaches to the problem (21,22) as well as some new statistical techniques. We demonstrated that these techniques could then be used effectively to discover CRMs in *Drosophila* and human genomes. Narlikar and colleagues (23) used a similar approach to predict human heart enhancers. Here, we advance the state-of-the-art in solving this problem by developing a new statistical approach that improves significantly upon previous methods. This new score is based on 'interpolated Markov models' (IMM) and the use of multi-species comparison. An IMM can be thought of as a combination of Markov chains of varying orders, and considers the frequencies of words of variable length in learning a generative probabilistic model from the training CRMs. We train an IMM on the training set of CRMs from *D. melanogaster* as well as their orthologs from other *Drosophila* species, and use the likelihood ratio of this model and a suitable null model as the score of a candidate CRM. The new method is shown to be superior to existing techniques in terms of predictive accuracy, which is assessed using a new statistical test that extends the widely used Hypergeometric test to correct for an important bias present in this test as applied to our setting.

We experimentally validated 12 new CRM predictions by examining their regulatory activity *in vivo*, and found 10 of these to be functional. Moreover, we observed that the new IMM score can effectively discriminate CRMs whose regulatory activity matches that of the training set from sequences that either do not drive expression or drive an expression pattern different from that of the training set.

The new CRM prediction method and evaluation strategy developed here are made publicly available at http://veda.cs.uiuc.edu/scrm-2/. The method can also be applied on the *Drosophila* genome through an online genome browser interface (http://veda.cs.uiuc.edu/gs, plugin 'Supervised CRM Discovery').

MATERIALS AND METHODS

Data sets

The CRM training sets were obtained from (24). The corresponding expression gene sets were taken from Supplementary Table S10 in (20) (see Methods in Supplementary Data for more details). Both training sets and expression gene sets are publically available at: http://veda.cs.uiuc.edu/scrm-2/.

IMM-based score

We implemented the IMM introduced by (25) for finding microbial genes. Let $S = \{s_1, s_2,\ldots, s_N\}$ be a sequence of length N. For an nth-order IMM, the conditional probability of generating s_i given its preceding context (of n characters), is:

$$\begin{aligned}\mathrm{IMM_n}(s_i|s_{i-1},\ldots,s_{i-n}) &= \lambda_n(s_{i-1},\ldots,s_{i-n})Pr(s_i|s_{i-1},\ldots,s_{i-n})\\ &+(1-\lambda_n(s_{i-1},\ldots,s_{i-n}))IMM_{n-1}(s_i|s_{i-1},\ldots,s_{i-n+1})\end{aligned} \qquad \mathbf{(1)}$$

Here, $\mathrm{Pr}(s_i|s_{i-1},\ldots,s_{i-n})$ is proportional to the frequency of the word $(s_{i-n},\ldots,s_{i-1},s_i)$ in training data and $\lambda_n(s_{i-1},\ldots,s_{i-n})$ is a 'mixture weight' that depends on the number of occurrences of the word $(s_{i-n},\ldots,s_{i-1})$ in the training data [$\lambda_0 = 1$; see (25) for details]. Thus, $\mathrm{IMM_n}(s_i|s_{i-1},\ldots,s_{i-n})$ is modeled as a mixture over n

different probability distributions on s_i, each of which is conditioned on a context of a different length (0, 1,..., n).

We score every candidate CRM S using the log likelihood score defined as:

$$\text{score}(S) = log \frac{p(S|\text{IMM}_5^+)}{p(S|\text{IMM}_5^-)} \quad \textbf{(2)}$$

where IMM_5^+ and IMM_5^- are two 5th-order IMMs trained on positive and negative sequences, respectively. The training set was used as positive sequences (both strands were used for training the model while ignoring the statistical dependencies between the strands), while the negative sequences used were regions of non-coding genomic sequence with G/C content similar to the native flanks of the CRMs. The training data set is publicly available at: http://veda.cs.uiuc.edu/scrm-2/

Multi-species scoring scheme

Given a training set of CRMs from *D. melanogaster*, we obtained orthologous sequences of each CRM from 10 other *Drosophila* species using the lift Over program from the UCSC Genome Browser database, and treated the resulting set of CRMs as the sequence data on which the model (IMM or HexMCD) was trained (see Methods in Supplementary Data). Fewer than 4% of the CRMs were missing an ortholog in one or more species, which eliminates any potential bias toward CRMs for which more species are alignable.

Locus length-aware Hypergeometric test

A significance test of overlap between two gene sets is traditionally performed with the Hypergeometric test. However, this test assumes that every gene is equally likely to be selected *a priori*. This assumption is not true for the 'predicted gene set' in our setting, which is the set of genes nearest to predicted CRMs, since gene loci vary in length across the genome (26). We confirmed that such variation in locus lengths is true for our data sets (Supplementary Figure S1A). We designed the 'locus length-aware Hypergeometric test' (LLHT) to correct for this bias. Our key idea is to assign to each gene a prior probability of being sampled, in proportion to its locus length. Other than this modification, the LLHT calculations are identical to the Hypergeometric test. We compared evaluation *P*-values obtained using LLHT with those based on the Hypergeometric test, and found the former to be noticeably more conservative for most of the 33 data sets (Supplementary Figure S1B). We also undertook a simulation-based experiment to assess the specificity of the two tests, and found that the Hypergeometric test is highly susceptible to false positives when the gene set has atypical locus lengths, while the LLHT retains its high specificity in this scenario (Supplementary Table S1).

Drosophila reporter constructs and transgenic analysis

Genomic sequences were generated by PCR and confirmed by sequencing; genome coordinates for predicted and tested CRMs are provided in Supplementary Table S2. The putative CRM sequences were subcloned into plasmid pattBnucGFPs, a ϕC31-enabled *Drosophila* transformation vector containing EGFP under the control of a minimal hsp70 promoter (details available on request). Transgenic flies were produced by Genetic Services Inc. (Cambridge, MA, USA) by injection into line attP2. Homozygous transgenic embryos were collected, fixed and stained with antibodies to GFP (Abcam, ab290) using standard methods. Muscle expression was visualized using anti-Tropomyosin (Babraham, MAC 141). DIC microscopy was performed using a Zeiss Axioplan microscope with a Retiga-EXi camera (Qimaging) and Openlab software (Improvision). Fluorescent images were acquired using a Leica SP2 confocal microscope. All images were color corrected and contrast adjusted in Adobe Photoshop.

RESULTS AND DISCUSSION

The problem

Suppose we are given a set *T* of CRMs (called the 'training set') that are known to drive similar expression patterns. For example, these could be CRMs that regulate genes with early mesodermal expression or genes that are involved in wing development. The task is to search the genome for other CRMs that drive an expression pattern similar to those in the training set *T*. A successful method to perform this task will capture the *cis*-regulatory signatures of TFs regulating *T*, and search genome wide for occurrences of those signatures. A predicted CRM can be validated by testing its function in a reporter gene assay and judging if its expression pattern is similar to that of the training set. We refer to the task just defined as the 'supervised CRM prediction' problem. Note that the problem is not completely specified in this formulation, since we do not precisely lay down what 'similarity of expression pattern' means. Instead, we allow the domain specialist to determine which characteristics of an expression pattern are used to define *T* and to judge the success of a prediction. One simple way to define *T* (in *Drosophila*) is to search the REDfly CRM database (27) with specific expression terms, such as 'mesoderm' or 'wing', thereby retrieving CRMs that have been shown to drive expression in the mesoderm or wing of the fly, respectively.

It is possible that a predicted CRM upon experimental testing shows regulatory activity in agreement with a neighboring gene's expression pattern, but this regulatory activity does not match that of the training set *T*. This is possible if TFs regulating the training set pleiotropically participate in a different regulatory network; the CRM prediction method may correctly learn the *cis*-regulatory signatures of such TFs and correctly identify a CRM that has those signatures, but the CRM may belong to the second network and hence show activity that does not match the training set. In cognizance of the above possibility, we define a second alternative criterion for a successful prediction: a CRM that drives an expression pattern matching that of a neighboring gene. In our experimental evaluations, we will examine success rates for prediction using both criteria introduced here.

A new scoring scheme for supervised CRM prediction

An obvious approach to supervised CRM prediction is to scan the genomic sequence with a shifting window of fixed size, and score the window for similarity to the training set of CRMs. The crucial aspect of this approach is the scoring system for matching a candidate module (‘test CRM’) to the training set. Different scoring systems have been proposed in the past, as reviewed and evaluated in our previous work (20). All of these scores have been based on comparing the frequencies of short words (henceforth called ‘k-mers’) between the test CRM and the training set. This is a reasonable strategy, since the regulatory function of CRMs is believed to be encoded in the types and numbers of transcription factor-binding sites. Specifically, binding sites that occur repeatedly in the training CRMs ought to cause some level of overrepresentation of specific k-mers, and the occurrence of such words in the test CRM should then indicate functional similarity. The kinds of scores that have been explored in this context include Markov chain-based scores (22), dot product-based scores (28), scores based on Poisson statistics (29) and others (30,31). For statistical convenience, these scores have been based on frequencies of k-mers of a fixed length (chosen from the range 5 to 8). However, in reality, the recognition motifs of different TFs are of varying lengths and binding sites of the same TF are known to display some amount of variability in sequence. For these reasons, the use of fixed length k-mers in the statistical scores may be problematic. While longer k-mers might reflect the binding site pattern more accurately, they may not have the statistical support necessary for robust testing, due to the variability among sites. Moreover, a CRM comprises binding sites of different TFs, with varying lengths, further undermining the use of fixed length k-mers in capturing the overall binding site composition of the training set. One natural solution to this problem is to start building a model (of the training set) with the shortest k-mers and attempt to include longer words whenever there are ‘enough’ occurrences of the word in the training data. This is the key idea in the statistical model we adopt here: a *k*th order IMM (25), which is a mixture of Markov models of all orders up to order k. As in a fixed order Markov chain, the IMM computes the probability of generating a sequence by multiplying the probabilities of each character (nucleotide) given the characters before it. However, instead of conditioning on a fixed number (say, *k*) of previous characters, the probability is calculated separately by looking at 1, 2,. . . *k* previous characters, and a linear combination of these probabilities is the multiplicative factor contributed by the current character (See Materials and methods section).

We train two 5^{th} order IMMs, one on the training set (‘positive’ model) and one on suitably chosen non-CRM or ‘background’ sequences (‘negative’ model), and score every candidate sequence by calculating the (log)-likelihood ratio between two models. A positive score means that the candidate sequence is more likely to have been generated by (i.e. is more similar to) the positive model. This score is henceforth called the IMM score.

Exploiting multi-species sequence data

Functional binding sites are known to be under evolutionary constraints, a fact that can be used to guide the search for potential CRMs (32). As a simple implementation of this idea, we trained the IMM using training CRMs from *D. melanogaster* as well as their orthologs from 10 other *Drosophila* genomes. The log-likelihood ratio score derived from the model(s) trained in this manner is henceforth called the multi-species IMM (msIMM) score. Note that this approach does not use alignment information except when determining the boundaries of the orthologs of a training CRM. That is, it does not rely upon alignments of orthologous CRMs, and thus it is free from artifacts of potential errors in such alignments. Evolutionary comparison in this manner also provides a natural way to ‘smooth’ the counts of k-mers, as explained next. Binding sites are known to be variable, yet standard Markov chains as well as the IMM use counts of exact matches to a k-mer to quantify the importance of that k-mer in the generative model. Previous methods have attempted to address this issue by allowing for a small number (e.g. 1) of mismatches to the word in counting its occurrence—a strategy that is rather simplistic in its modeling of binding site variability. Examining many orthologs of a CRM provides a natural way to capture the variability of binding sites, to be used in training a model.

Evaluation methodology

Given a set of CRM predictions genome wide, the definitive test of their accuracy would be to determine experimentally if each CRM drives an expression pattern in the expected tissue and/or developmental stage. Since such tests are typically not performed *en masse* (due to resource limitations), we need an *in silico* evaluation strategy. In our previous work, we proposed the following strategy, which exploits existing databases of expression pattern annotation across thousands of genes in *D. melanogaster* (33) [BDGP (http://www.fruitfly.org)]. It is possible to obtain, for any given training set *T*, a corresponding set G_E of genes with expression patterns similar to those of *T*. Successful CRM prediction (by both criteria described in ‘Introduction’ section) implies that CRMs predicted using the training set *T* regulate genes with similar expression patterns, i.e. the predictions would be overrepresented in the control regions of genes in G_{E}. In other words, if we consider the CRM predictions for the training set *T*, and denote the set of ‘corresponding’ genes (e.g. closest neighbors) as G_{T}, we could use the statistical significance of the overlap between G_{T} (the ‘predicted gene set’) and G_{E} (the ‘expression gene set’) as an assessment of prediction accuracy (Figure 1A). This approach to evaluating CRM prediction has been used by several groups previously (20,22,32,34) and is a natural option to adopt when the existing knowledge of functional CRMs in a regulatory network is sparse, yet gene expression databases are available (see Methods in Supplementary Data, Notes 1 and 2 for more details). We evaluated the msIMM score in this manner, using the LLHT for statistical assessment of enrichment. This is a generalization of the widely used Hypergeometric test, which we designed

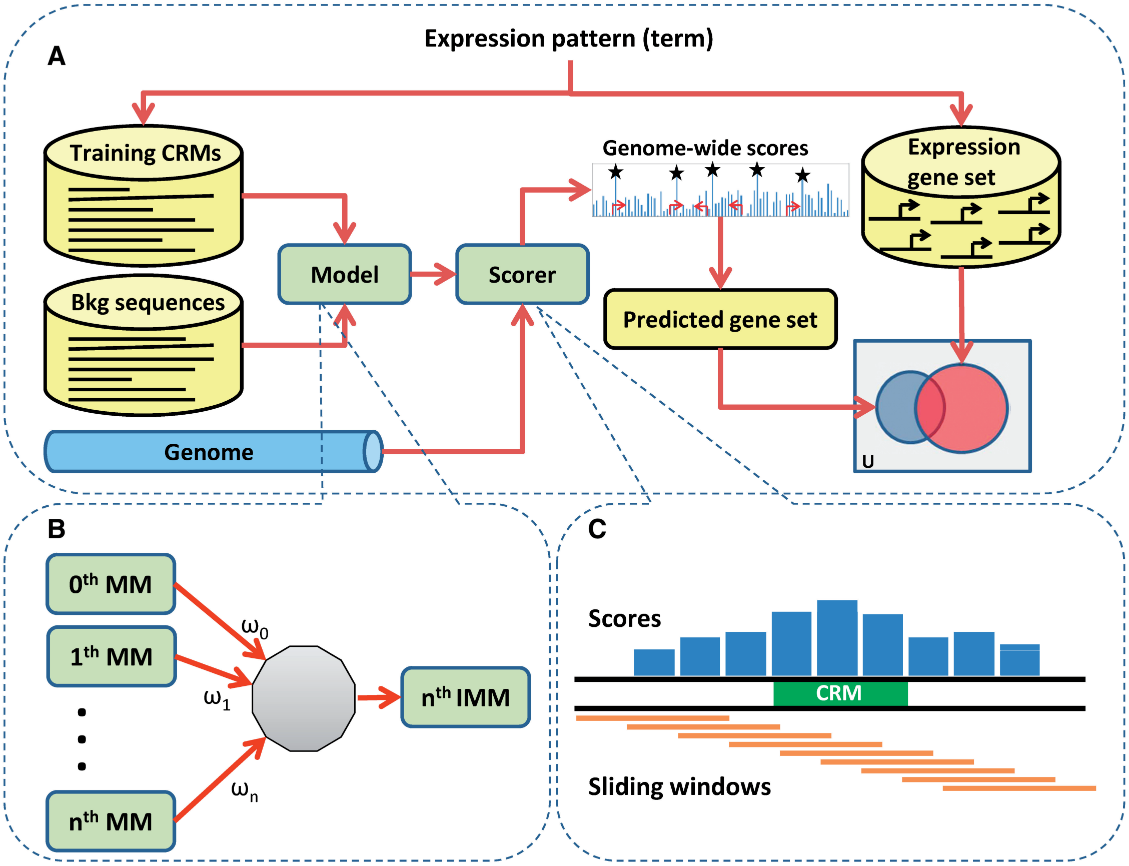


Figure 1. **(A)** General scheme for evaluation of a CRM discovery method. We first select a set of genes from BDGP (or FlyBase) with expression patterns commensurate with those of the CRMs in a data set. We next take the modules predicted for that data set and extract their nearest neighboring genes ('predicted gene set'). (In this step, we ignore predicted modules that overlap with any training CRMs.) Finally, we perform a Hypergeometric test of enrichment between the expression gene set and the predicted gene set. **(B)** IMM of order n is a mixture of Markov models up to order n. '0th MM', '1th MM',..., 'nth MM' denote Markov models of order 0, 1,..., n, respectively. $\omega_0, \ldots, \omega_n$ are the mixture weights. **(C)** Scorer scores every windows of length 500 bp with 50 bp shift across the entire genome. In this toy example, the score of each sliding window (orange lines) is shown as blue bar.

specifically for assessing the significance of overlap between two gene sets (see 'Materials and Methods' section; Materials in the Supplementary Data). We used 33 training sets and their corresponding expression gene sets, collectively called 'data sets', from (20), based on the REDfly (35), FlyBase (33) and BDGP (36) databases (see 'Materials and Methods' section). Figure 1 describes the complete pipeline of supervised CRM prediction by the IMM score and the evaluation strategy.

Evaluation results

The methods evaluated included the single and multi-species versions of the IMM score, the three best performing methods from (20)—HexMCD [derived from (22)], PAC-rc [derived from (29)] and HexYMF-s200-rc, as well as a motif-based approach called Clover-ClusterBuster (20)—that uses 'Clover' (37) for motif selection and 'ClusterBuster' (8) for scanning with selected motifs (see Methods in Supplementary Data for more details)—and a newer alignment-free method from (38) (henceforth called AJTO). We also implemented a multi-species version of HexMCD (called msHexMCD), as we had done for the IMM score. The comparison between IMM and HexMCD is of special interest as both are based on a Markov chain formulation, but HexMCD uses a fixed order (of 5), in contrast to the interpolated order approach of IMM. We obtained the 'predicted gene set' (G_T, of size 200) computed a P-value (LLHT based) for each method—data set pair, henceforth called the 'evaluation P-value'. We then counted how many data sets (out of 33 in our test-bed) yielded an evaluation P-value better than a threshold, and varied this threshold. The results are shown in Figure 2A.

- The msIMM score was found to be superior to all other methods by this criterion. For example, msIMM had an evaluation P-value better than E-5 on eight data sets, compared to the second best (non-IMM) method msHexMCD for which the corresponding number was five data sets. At E-10, the corresponding numbers were four data sets for msIMM, one for msHexMCD and IMM and zero for all other methods.
- The msIMM score yields clearly better results than the single species IMM score, indicating the advantage of using multi-species data in the training phase.

This trend was also seen when comparing the multi-species version of HexMCD to the corresponding single-species score used in (20).

- The single species IMM score was found to be superior to the single species HexMCD score and to the other single species scores (PAC-rc, HexYMF-s200-rc), revealing the advantage of using variable length words to model CRMs.

The AJTO method led to no data sets with evaluation P-value < 0.001. Since we have relatively little experience as users of this software (38), we chose to exclude it from further comparisons, to avoid making any biased inferences about its accuracy.

To provide a more detailed view of the above comparisons, we plotted the overlap between the expression gene set and the top k predicted genes, as a function of k (Figure 2B; Supplementary Figure S2A and S2B), for one particular data set (complete results in Supplementary Figure S2A). While most methods are comparable to each other in the high specificity range (e.g. ≤75 predictions), the msIMM method distinguishes itself from the others in the top 100–200 predictions. (Note that the set G_T is of size 200 for each method, for fair comparison above.) We note, however, that this particular pattern of improved performance is not common to all data sets (Supplementary Figure S2A).

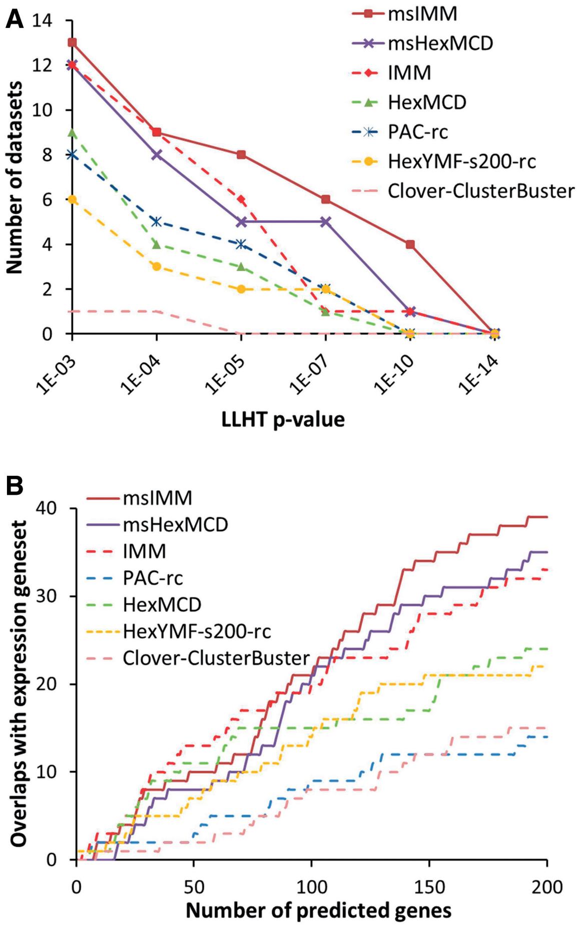


Figure 2. Evaluation of methods. For each method, shown is (**A**) the number of data sets for which the evaluation P-value is significant at different LLHT P-value thresholds. For clarity, all single species methods are formatted as dashed lines and the two multi-species methods are shown as solid line. (**B**) Y-axis shows the number of overlaps between the expression gene set and the top k predicted genes for 'imaginal_disc.2' data set. (See Supplementary Figure S2A and S2B for all other data sets.).

We also compared the evaluation P-value of msIMM to the P-value of the best method reported in (20), for all 15 data sets where this evaluation was performed in (20). These are listed in Table 1. (See Supplementary Table S3 for evaluation P-value of all methods in all data sets.) In every data set, the msIMM P-value is better than or similar to that of the competing method, with the most dramatic improvements observed in the data sets 'cns.1' (E-5 to E-9), 'imaginal_disc.1' (E-6 to E-11) and 'ventral_ectoderm.1' (E-4 to E-12).

Table 1. Comparison between evaluation P-values of msIMM and the best method from (20), for the 15 data sets that were reported on by (20)

Data set	Universe size	Expression set size	Kantorovitz *et al.* (2009)	msIMM
blastoderm.1	5506	206	2E-10	5E-13
cardiac_mesoderm.1	5506	162	3E-01	5E-02
cns.1	5506	839	3E-05	6E-09
eye.1	13 155	73	3E-01	6E-02
imaginal_disc.1	13 155	312	3E-06	9E-11
mesectoderm.1	5506	93	3E-01	2E-03
mesoderm.1	5506	764	3E-05	6E-04
neuroectoderm.1	13 155	59	1E-04	1E-04
pns.1	5506	96	8E-02	1E-02
somatic_muscle.1	13 155	76	8E-01	3E-02
ventral_ectoderm.1	5506	326	8E-04	2E-12
visceral_mesoderm.1	5506	133	1E-02	5E-03
ectoderm.2	5506	834	4E-03	1E-05
neuronal.2	5506	66	3E-03	8E-02
wing.2	13 155	135	2E-03	3E-06

The second and third columns show the total number of genes in the expression data source and the size of expression gene set for each data set, respectively (the size of the predicted gene set is 200 for all data sets). The lowest (best) P-value for each pair is shaded. Note that the P-values reported here are LLHT P-values, which are different from the standard Hypergeometric P-values shown in Table 2 of (20).

In the above presentation, a method's performance is assessed based on the overlap between the predicted gene set (200 genes with the strongest predicted CRMs in their control regions) and the expression gene set (see Methods in Supplementary Data, Note 1). We also examined the entire distribution of a method's score for each gene in the expression gene set and compared it to the score distribution from a random collection of genomic segments (with lengths matching the gene control regions). These two distributions were compared using a Wilcoxon rank sum test, and the resulting P-value indicates how well the particular score discriminates the expression gene set (which should have CRMs) from the randomly chosen segments (which should only have coincidental occurrences of CRMs). The P-values are shown for each data set and each method in Table 2. It is clear from this evaluation that the msIMM

Table 2. Discrimination between control regions of an expression gene set and random sequences of matching lengths

Data set	CRM set size	Expression set size	Clover-ClusterBuster	PAC-rc	HexYMF-s200-rc	HexMCD	msHexMCD	IMM	msIMM
adult_mesoderm.1	5	56	0.0731	9E-06	0.0003	0.0004	7E-06	6E-06	0.0002
amnioserosa.1	7	126	0.0002	6E-10	6E-08	5E-17	2E-10	6E-12	3E-15
blastoderm.1	77	206	6E-20	4E-34	2E-31	7E-44	1E-42	1E-50	9E-51
cardiac_mesoderm.1	8	162	0.0001	1E-05	1E-05	4E-19	7E-13	5E-20	2E-22
cns.1	34	839	7E-30	7E-17	1E-12	6E-54	5E-41	2E-82	2E-63
dorsal_ectoderm.1	8	640	6E-29	9E-28	4E-17	4E-56	7E-47	3E-58	1E-67
ectoderm.1	37	109	5E-12	2E-19	1E-11	1E-18	1E-17	6E-19	8E-19
endoderm.1	16	195	6E-06	4E-13	4E-09	7E-27	3E-20	5E-24	4E-22
eye.1	6	73	8E-05	0.0012	0.016	2E-07	3E-05	5E-06	1E-07
fat_body.1	5	61	1	0.0014	0.0025	3E-09	5E-08	6E-08	3E-08
female_gonad.1	10	237	0.0015	0.0545	0.0098	7E-16	5E-07	2E-32	8E-39
glia.1	7	44	0.0031	5E-07	2E-05	2E-07	5E-06	4E-06	4E-06
imaginal_disc.1	47	312	8E-13	7E-27	7E-24	4E-25	9E-28	2E-39	1E-40
male_gonad.1	8	237	1	0.3967	0.2561	7E-16	5E-07	2E-32	8E-39
malpighian_tubules.1	4	9	0.0491	0.0003	0.0459	0.2022	0.0169	0.0093	0.0354
mesectoderm.1	5	93	2E-05	8E-10	1E-07	5E-17	3E-16	1E-15	8E-15
mesoderm.1	16	764	0.1972	0.0178	0.1324	3E-47	7E-30	3E-126	4E-142
neuroectoderm.1	7	59	2E-06	9E-12	2E-12	9E-07	2E-15	4E-10	6E-09
pns.1	24	96	0.0003	3E-07	2E-05	6E-08	4E-07	1E-10	8E-10
salivary_gland.1	6	123	0.0663	0.0084	0.003	6E-10	5E-05	4E-06	1E-10
somatic_muscle.1	12	76	0.0023	2E-07	0.0002	8E-11	3E-09	5E-11	5E-12
tracheal_system.1	9	355	6E-12	1E-21	3E-19	4E-25	3E-16	3E-32	2E-25
ventral_ectoderm.1	12	326	2E-22	1E-14	1E-12	5E-40	2E-33	1E-42	1E-47
visceral_mesoderm.1	12	133	2E-06	5E-08	1E-05	1E-18	7E-15	7E-16	1E-18
ectoderm.2	51	834	2E-43	2E-43	2E-29	4E-77	4E-62	1E-98	3E-83
eye.2	18	153	2E-05	4E-13	1E-11	3E-15	3E-16	8E-17	3E-19
imaginal_disc.2	12	312	1	8E-17	2E-17	9E-26	8E-28	9E-30	2E-36
mesoderm.2	45	764	3E-07	0.0006	0.059	3E-47	3E-30	7E-95	1E-66
neuronal.2	54	66	0.0002	0.0053	0.0187	4E-05	0.0003	8E-05	3E-05
reproductive_system.2	21	421	5E-09	8E-05	1E-04	4E-21	1E-12	2E-35	2E-26
wing.2	33	135	2E-05	3E-14	4E-11	4E-15	8E-15	4E-19	3E-19
adult.3	34	394	2E-13	2E-22	4E-18	2E-30	1E-25	6E-33	3E-30
larva.3	69	743	3E-37	4E-61	3E-53	7E-78	2E-68	3E-85	2E-80
Number of 'wins'			0	2	0	5	1	9	16

The second and third columns show the number of training CRMs and the size of expression gene set, respectively, for each data set. Scores of genes in an expression gene set were compared to scores of a collection of randomly chosen genomic regions. The score of a sequence is the maximum score in that region, under a CRM prediction scheme. For each gene in the expression gene set, 50 random genomic segments of length equal to the gene's territory length were included in the random collection. The last seven columns show the *P*-values (Wilcoxon rank sum test) of such a comparison for each data set and for each method. The best *P*-value for each data set is shaded. The last row indicates the number of times that a method is superior (smallest *P*-value).

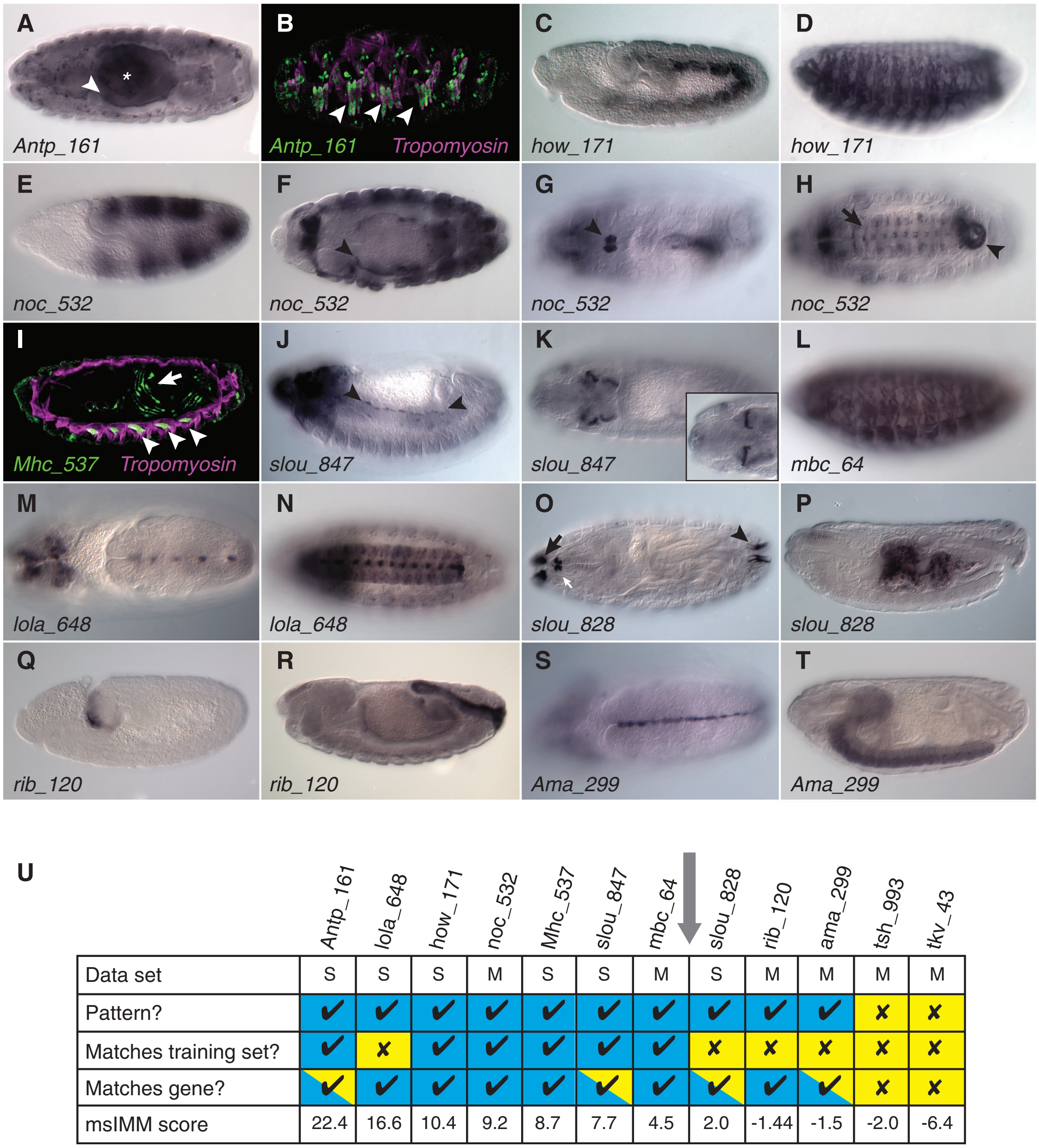
score is best able to discriminate between the expression gene set and the random set, for the most number of data sets.

Experimental validation

We chose 12 putative CRMs predicted using the 'mesoderm' and 'somatic muscle' training sets (six for each), and tested them for regulatory activity *in vivo*. These two data sets were chosen partly because their evaluation *P*-values were significant but not among the best. (This is especially true for the 'somatic muscle' set; evaluation $P = 0.03$.) Unlike (20), where the tested CRMs were from the data set with the strongest evaluation *P*-value ('blastoderm', $P = 5E\text{-}13$), our goal here was to investigate the effectiveness of our methods in what seem to be more challenging data sets. We also note that these two data sets are not exclusive from other data sets (e.g. 'visceral mesoderm') in terms of their defining expression patterns, which creates additional difficulties for supervised CRM discovery. The 12 candidate CRMs were selected by the following criteria:

(1) they are located near genes in the respective expression gene set;
(2) they collectively fall across the spectrum of msIMM scores, allowing us to examine (through direct experimental assays) the accuracy of this new scoring scheme not just at its highest scoring predictions but also at the medium and low scoring predictions; and
(3) each candidate scores well by either msIMM or PAC-rc, i.e. a candidate with a relatively low msIMM score (<0) has some supporting evidence in favor of it being a potential CRM.

Figure 3 summarizes the results of our experiments. Of the 12 tested sequences, 10 (83%) were found to drive reporter gene expression in a tissue- or stage-specific pattern (see Methods in Supplementary Data, Note 3). The two failed predictions (corresponding to genes *tkv* and *tsh*) were the two lowest scoring predictions of msIMM (in this set).



	Antp_161	lola_648	how_171	noc_532	Mhc_537	slou_847	mbc_64	slou_828	rib_120	ama_299	tsh_993	tkv_43
Data set	S	S	S	M	S	S	M	S	M	M	M	M
Pattern?	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✘	✘
Matches training set?	✔	✘	✔	✔	✔	✔	✔	✘	✘	✘	✘	✘
Matches gene?	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✘	✘
msIMM score	22.4	16.6	10.4	9.2	8.7	7.7	4.5	2.0	-1.44	-1.5	-2.0	-6.4

Figure 3. *In vivo* validation of 12 Drosophila CRM predictions. Transgenic embryos were stained with antibodies to GFP to detect reporter gene expression. Embryos are oriented anterior to the left. Panels A, F, G, K, M, O and S are dorsal views; H and N are ventral views; the remainder are lateral views with dorsal to the top. 'Pattern-specific' CRMs are shown in panels A–L and non-pattern specific CRMs in panels M–T. (**A**) The *Antp_161* CRM drives expression in several tissues, including the visceral mesoderm (arrowhead) and the non-mesodermal midgut (asterisk). The endogenous *Antp* gene is expressed in a much more restricted manner, suggesting that the reporter gene expression is either ectopic or that the CRM is associated with a different gene of undetermined identity. (**B**) Co-labeling for Tropomyosin (magenta) shows GFP expression (green) in somatic muscles, in particular the lateral transverse muscle fibers (arrowheads). (**C** and **D**) The *how_171* CRM drives expression in the mesoderm in both mid-stage (C) and late-stage (D) embryos, consistent with *how* gene expression. (**E–H**) The *noc_532* CRM is active in many *noc*-positive tissues throughout embryogenesis. Pictured is metameric expression in both ectoderm and mesoderm at stage 9 (E), in the visceral mesoderm (F, arrowhead), in the mesodermally-derived lymph glands (G, arrowhead), the ventral nerve cord (H, arrow) and the hindgut (H, arrowhead). (**I**) *Mhc_537* regulates reporter gene expression (green) in a subset of *Mhc*-positive mesodermal cells including longitudinal visceral muscles (arrow) and several ventral oblique somatic muscle fibers (arrowheads). (**J**) Longitudinal visceral muscle precursors express GFP under the control of the *slou_847* CRM (arrowheads), cells not positive for endogenous *slou* expression. Expression is also observed in the supraesophogeal ganglion (**K**) and ventral nerve cord (data not shown), consistent with known *slou* expression patterns. Inset shows a more dorsal view of the anterior portion of the embryo in the main panel. (**L**) *mbc_64* drives expression in the *mbc*-positive mesoderm, pictured here at stage 16. (**M** and **N**). Expression driven by the *lola_648* CRM is confined to the central nervous system in both mid-stage (M) and late-stage (N) embryos, consistent with *lola* expression. *lola_648* overlaps the independently discovered CRM40 of (40). (**O**) The *slou-828* CRM regulates reporter gene expression in tissues that are not *slou*-positive including cells in the antenno-maxillary complex (black arrow), the posterior spiracles (arrowhead) and cells in the anterior and posterior portions of the foregut (white arrow and data not shown). (**P**) *slou_828*-controlled expression is also observed in the midgut, consistent with normal *slou* expression. (**Q** and **R**) The *rib_120* CRM drives expression throughout hindgut development, part of the normal *rib* expression pattern. (**S** and **T**) Reporter gene expression regulated by the *ama_299* CRM is restricted to the central nervous system. Although *ama* is not expressed in the CNS of late-stage embryos (41), earlier expression in the ventral midline beginning at stage 8 (data not shown) is consistent with *ama* expression at that stage.

(continued)

We then examined the two specific success criteria we laid down in the beginning ('Introduction' section). Due to the non-exclusive and overlapping nature of the expression patterns found within each of the two training sets, we considered any mesodermal gene expression as matching the training pattern. Of the 12 tested sequences (50%), 6 showed reporter gene expression in one or more mesodermal tissues (Figure 3A–L). Remarkably, the msIMM score was able to predict success by this criterion with 90% accuracy. In particular, a threshold value of the msIMM score was able to separate the pattern-specific CRMs from the non-specific CRMs (or non-CRMs) with only one false positive error (Figure 3U). Of the 12 predicted CRMs, 10 (83%) drive an expression pattern matching at least in part that of the endogenous expression of the nearest gene.

Visualization tool

In addition to making the source code directly available for download, we have developed a plugin for Gbrowse (39), called 'Supervised CRM discovery', through which the user can select a set of related CRMs as their starting point and browse a genomic region of interest for other putative CRMs with similar functionality. The plugin can be accessed via: http://veda.cs.uiuc.edu/gs.

CONCLUSION

We have presented a new statistical method for genome-wide prediction of CRMs, beginning with a collection of known modules in the regulatory network of interest. This problem of 'supervised CRM prediction' was visited in our previous work (20), where we established the necessary benchmarks and evaluation methodology, and also compiled a suite of pre-existing and novel methods to solve the problem. The new method we propose now is shown to be superior to the best of the methods considered in (20), hence advancing the state of the art on this important topic.

Our work makes three additional regulatory networks—central nervous system, imaginal disc and ventral ectoderm—amenable to supervised CRM prediction. Our experiments also contribute 10 new CRMs, involved in development of early mesoderm and of somatic muscle, to the literature.

Our new method, along with our previously compiled techniques (20), represents a flexible, accurate way of identifying regulatory sequences, especially for less fully described processes for which we lack the knowledge of relevant transcription factors. It can prove to be an important *in silico* adjunct to and extension of empirical CRM discovery approaches. Furthermore, it has the potential to lead to novel CRM predictions in experimentally less well-characterized arthropods such as mosquito, wasp, beetle and honeybee, while exploiting existing collections of CRMs in *Drosophila*.

We have also made a key improvement to the evaluation methodology for genome-wide CRM prediction methods, by proposing the locus LLHT. This new test is shown to lead to superior specificity (lower false positive rate) when enrichment tests are performed with gene sets of unusually large locus lengths. The Hypergeometric test of significance of overlap between gene sets forms the basis of many tools used in genomics, and the common phenomenon of locus length variability therefore casts doubts on the statistical inferences from these tools (26). Our modification to the Hypergeometric test (LLHT) thus has the potential of widespread usability in contexts beyond CRM discovery.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: National Institutes of Health (1R01GM085233-01 to S.S. and M.S.H.).

Conflict of interest statement. None declared.

REFERENCES

1. Davidson,E.H. (2001) *Genomic Regulatory Systems*, 1 edn. Academic Press, San Diego.
2. Davidson,E.H. (2006) *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*, 1st edn. Academic Press, Burlington, MA.
3. Xi,H., Yu,Y., Fu,Y., Foley,J., Halees,A. and Weng,Z. (2007) Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Res.*, **17**, 798–806.
4. 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.
5. 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.
6. 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.
7. Halfon,M.S., Grad,Y., Church,G.M. and Michelson,A.M. (2002) Computation-based discovery of related transcriptional regulatory

Figure 3. Continued
(U) Summary of results and msIMM scores for each predicted CRM. Names are based on the closest gene to each predicted module and do not necessarily reflect the actual gene regulated by the CRM. 'Data set' indicates the data set from which the CRM was predicted ('S' and 'M' for *somatic muscle* and *mesoderm* respectively). 'Pattern?' indicates whether the tested sequence drives a spatial and/or temporal expression pattern. 'Matches training set?' and 'Matches gene?' indicate whether the expression pattern agrees with that of the training set or the nearest gene, respectively. Check marks/blue coloring denote a positive result, crosses and yellow coloring a negative result. Mixed blue and yellow coloring is used for cases where both endogenous and ectopic gene expression patterns are observed. 'Score' shows msIMM scores for each tested CRM. The gray arrow points to the best decision stump on msIMM scores in terms of predicting pattern-specific CRMs.

modules and motifs using an experimentally validated combinatorial model. *Genome Res.*, **12**, 1019–1028.

8. 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.
9. Sinha,S., van Nimwegen,E. and Siggia,E.D. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19(Suppl. 1)**, i292–i301.
10. Philippakis,A.A., He,F.S. and Bulyk,M.L. (2005) Modulefinder: a tool for computational discovery of cis regulatory modules. *Pac. Symp. Biocomput.*, **10**, 519–530.
11. Donaldson,I.J., Chapman,M. and Gottgens,B. (2005) TFBScluster: a resource for the characterization of transcriptional regulatory networks. *Bioinformatics*, **21**, 3058–3059.
12. 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.
13. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
14. Bergman,C.M., Carlson,J.W. and Celniker,S.E. (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics*, **21**, 1747–1749.
15. 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.
16. Zhu,L.J., Christensen,R.G., Kazemian,M., Hull,C.J., Enuameh,M.S., Basciotta,M.D., Brasefield,J.A., Zhu,C., Asriyan,Y., Lapointe,D.S. *et al.* (2011) FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res.*, **39**, D111–D117.
17. 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.
18. Hu,J., Li,B. and Kihara,D. (2005) Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res.*, **33**, 4899–4913.
19. Sandve,G.K., Abul,O., Walseng,V. and Drablos,F. (2007) Improved benchmarks for computational motif discovery. *BMC Bioinformatics*, **8**, 193.
20. Kantorovitz,M.R., Kazemian,M., Kinston,S., Miranda-Saavedra,D., Zhu,Q., Robinson,G.E., Gottgens,B., Halfon,M.S. and Sinha,S. (2009) Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse. *Dev. Cell*, **17**, 568–579.
21. Chan,B.Y. and Kibler,D. (2005) Using hexamers to predict cis-regulatory motifs in Drosophila. *BMC Bioinformatics*, **6**, 262.
22. 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.
23. Narlikar,L., Sakabe,N.J., Blanski,A.A., Arimura,F.E., Westlund,J.M., Nobrega,M.A. and Ovcharenko,I. (2010) Genome-wide discovery of human heart enhancers. *Genome Res.*, **20**, 381–392.
24. 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.
25. Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**, 544–548.
26. Taher,L. and Ovcharenko,I. (2009) Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. *Bioinformatics*, **25**, 578–584.
27. Gallo,S.M., Gerrard,D.T., Miner,D., Simich,M., Des Soye,B., Bergman,C.M. and Halfon,M.S. (2011) REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res.*, **39**, D118–D123.
28. Kantorovitz,M.R., Robinson,G.E. and Sinha,S. (2007) A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, **23**, i249–i255.
29. van Helden,J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics*, **20**, 399–406.
30. Vinga,S. and Almeida,J. (2003) Alignment-free sequence comparison-a review. *Bioinformatics*, **19**, 513–523.
31. Leung,G. and Eisen,M.B. (2009) Identifying cis-regulatory sequences by word profile similarity. *PLoS ONE*, **4**, e6901.
32. Sinha,S., Schroeder,M.D., Unnerstall,U., Gaul,U. and Siggia,E.D. (2004) Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics*, **5**, 129.
33. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.* (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
34. 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.
35. 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.
36. Tomancak,P., Berman,B.P., Beaton,A., Weiszmann,R., Kwan,E., Hartenstein,V., Celniker,S.E. and Rubin,G.M. (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.*, **8**, R145.
37. 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.
38. Arunachalam,M., Jayasurya,K., Tomancak,P. and Ohler,U. (2010) An alignment-free method to identify candidate orthologous enhancers in multiple Drosophila genomes. *Bioinformatics*, **26**, 2109–2115.
39. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
40. Rouault,H., Mazouni,K., Couturier,L., Hakim,V. and Schweisguth,F. (2010) Genome-wide identification of cis-regulatory motifs and modules underlying gene coregulation using statistics and phylogeny. *Proc. Natl Acad. Sci. USA*, **107**, 14615–14620.
41. Fremion,F., Darboux,I., Diano,M., Hipeau-Jacquotte,R., Seeger,M.A. and Piovant,M. (2000) Amalgam is a ligand for the transmembrane receptor neurotactin and is required for neurotactin-mediated cell adhesion and axon fasciculation in Drosophila. *EMBO J.*, **19**, 4463–4472.