Discovering transcriptional regulatory regions in *Drosophila* by a nonalignment method for phylogenetic footprinting

Alona Sosinsky*†‡, Barry Honig*†‡§, Richard S. Mann†, and Andrea Califano‡¶

*Howard Hughes Medical Institute, †Department of Biochemistry and Molecular Biophysics, ¶Department of Biomedical Informatics, and ‡Center for Computational Biology and Bioinformatics, Columbia University, 1130 Nicholas Avenue, New York, NY 10032

Contributed by Barry Honig, February 22, 2007 (sent for review October 13, 2006)

The functional annotation of the nonprotein-coding DNA of eukaryotic genomes is a problem of central importance. Phylogenetic footprinting methods, which attempt to identify functional regulatory regions by comparing orthologous genomic sequences of evolutionarily related species, have shown promising results. The main advantage of this class of approaches is that they do not require any knowledge of the regulating transcription factors. Here we describe a method called Enhancer Detection using only Genomic Information (EDGI), which integrates a traditional motif-discovery algorithm with a local permutation-clustering algorithm. Together, they can identify large regulatory elements (e.g., enhancers) as evolutionarily conserved order-independent clusters of short conserved motifs. We show that EDGI can distinguish between established sets of known enhancers and nonenhancers with 88% accuracy, rivaling predictions by methods that rely on the knowledge of the regulating transcription factors and their DNA-binding specificities. We tested EDGI's performance on a set of *Drosophila* genomes. Our results demonstrate that comparative genomic analysis of multiple closely related species has substantial power to identify key functional elements without additional biological knowledge.

enhancer discovery | regulatory sequence evolution | regulatory sequence prediction | transcription regulation | comparative genomics

Eukaryotic genomes comprise both protein- and noncoding DNA. Noncoding DNA harbors a variety of regulatory elements, called cis-regulatory modules (CRMs) or enhancers, which contain binding sites for specific combinations of transcription factors (TFs) that interact in coordinated fashion. By binding specific sets of TFs, CRMs integrate this information to direct the spatial and temporal expression pattern of nearby gene(s). The discovery and characterization of such regulatory modules are critical for understanding how genomes ultimately control all aspects of cellular behavior, including the orchestration of animal development and cellular homeostasis.

CRMs are typically 500–1,000 bp in length and can be located far (>50 kb in some cases) from the start of transcription of the regulated gene, making their detection in higher eukaryotic genomes particularly challenging. Because CRMs contain specific binding sites for sets of coregulating TFs, a common approach for their identification relies on the discovery of tight clusters of candidate TF-binding sites in genomic sequences (1–4). TF-binding sites are often represented as position weight matrices (PWMs), where the probability of finding a specific nucleotide at each position within a binding site is estimated from a training set of known experimentally validated binding sites. PWMs can be used to rapidly search for relatively tight genomic regions (500–1,000 bp), with a statistically significant concentration (or co-occurrence) of specific candidate-binding sites.

The main limitation to the widespread application of methods based on PWMs is the requirement for prior knowledge of the TFs that may act in a coordinated fashion within a specific enhancer; an additional limitation is the absence of high-quality binding data for many TFs, such that a reasonably specific PWM may be generated. Fortunately, the evolutionary conservation of regulatory sequences, compared with nonfunctional DNA, provides an additional rationale for CRM discovery, without knowledge of the binding sites. Evolutionary conservation has been used to develop a set of computational approaches termed "phylogenetic footprinting" for the systematic discovery of CRMs from sets of orthologous sequences (5). In these methods cross-species conservations in noncoding regions, or footprints, are treated as likely candidates for regulatory regions. The conservation arises from the presence of a large number of short functional TF-binding sites that are preferentially conserved throughout evolution. These may be used as "anchors" for longer (and thus statistically significant) sequence alignments. A number of alignment methods have been applied to define the boundaries of regulatory sequences. Gapped global pairwise alignments with following filtering for conserved sequences (e.g., $x\%$ conservation over at least y nucleotides) led to discoveries of several new enhancers (6–8). However, this method failed to discover some enhancers because of the lack of conservation over an extended region (9, 10). Another study shows that some enhancers can be discovered as clusters of short (10-bp) conserved motifs revealed from the pairwise alignment of genomic sequences from two *Drosophila* species (11). Methods for local alignments, ungapped Bayes block aligner (12), and gapped Waterman–Eggert (13), with subsequent integration of scores for multiple suboptimal alignments, have resulted in an enrichment of experimentally identified regulatory sites among most evolutionary conserved noncoding genomic sequences. However, the method based on Waterman–Eggert local alignments (13) shows only partial success in identifying upstream promoters from the Eukaryote Promoter Database.

The number of false-positive predictions due to residual sequence similarity in the pairwise alignments can be potentially reduced by assessing conservation across multiple related genomes. Given the independent divergence of each lineage after separation from a common ancestor, sequence conservation across multiple species is less likely to show residual similarities in nonselected regions than in the pairwise alignment. A pairwise alignment strategy was adopted for the analysis of ortholo-

Author contributions: A.S., B.H., R.S.M., and A.C. designed research; A.S., B.H., R.S.M., and A.C. performed research; A.S., B.H., R.S.M., and A.C. analyzed data; and A.S., B.H., R.S.M., and A.C. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: TF, transcription factor; CRM, cis-regulatory module; EDGI, enhancer detection using only genomic information; LPC, local permutation cluster; SCM, short evolutionarily conserved motif; SCS, sequence conservation score; SCSM, SCS metric; PCM, percent coverage metric; PWM, position weight matrices.

§To whom correspondence should be addressed. E-mail: bh6@columbia.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0701614104/DC1.

GENETICS

gous sequences from three mammalian species (human, mouse, and dog) (14). It was suggested that enhancers can be identified as regions of intersections of sequence conservation from multiple pairwise alignments (11, 14). Recent studies (15–17) attempted to identify regulatory motifs directly from multiple sequence alignments. Although this method successfully identified functional TF-binding sites in yeast intergenic regions (≈500-bp sequences) and mammalian promoter regions (1-kb sequences), it has not been applied to the discovery of distant enhancers in long eukaryotic sequences (30- to 100-kb sequences), probably in part because of difficulties in generating multiple alignments for long sequences.

In general, phylogenetic footprinting methods based on sequence alignments are not yet able to accurately predict functional CRMs. Recent results provide some insights into the issue. Although the evolutionary forces that shape CRM sequences are not well understood, it is thought that stabilizing selection can maintain CRM function for long periods of evolutionary time while at the same time allowing mutational turnover of functionally important binding sites (18). Therefore, although the same set of TFs may continue to regulate a given CRM over long evolutionary distances, specific TF-binding sites may be lost, gained, or rearranged relative to other TF-binding sites (18, 19). Such behavior makes the recognition of CRMs by sequence alignment very difficult, because it enforces a linear ordering of the aligned sequences.

Allowing binding sites to rearrange among species could be satisfied in principle by one of several motif discovery algorithms (20). Functional binding sites were successfully identified in bacterial promoters by Gibbs sampler (21) or in known mammalian enhancers by FootPrinter (22). However, these approaches are effective only when relatively short DNA regions are considered (<1 kb), because the number of nonfunctional conserved motifs increases dramatically in longer sequences.

In this paper, we introduce a CRM identification method, which is effective in the analysis of long genomic sequences (30 kb to >100 kb), allowing whole genes in higher eukaryotic genomes to be analyzed. The method relies on the evolutionary conservation of the CRM binding sites, without requiring any knowledge of the regulating TFs or their DNA-binding sites. We have named this method Enhancer Detection using only Genomic Information (EDGI). Unlike existing methods, this approach allows the discovery of CRMs as order-independent local permutation clusters (LPCs), closely related to permutation patterns (23), of short evolutionarily conserved motifs (SCMs). In such LPCs, individual motifs may vary both in position, multiplicity (number of time that the same motif occurs), and order between each orthologous sequence, as long as the same motifs are present within a window of predefined maximum length in each sequence. EDGI's ability to classify enhancers and nonenhancers is comparable to that of methods that rely on the knowledge of the regulating TFs and their DNA-binding profiles, whereas EDGI does not require this type of information.

Methods

Sequence Selection. EDGI is used to analyze orthologous noncoding sequences from the genomes of several closely related species. It can analyze full genes, including introns as well as Δx kb upstream and downstream of transcription start and stop sites, based on a specific genome annotation. In this paper, Δx = 15 kb, which is expected to harbor most *Drosophila*'s CRMs. However, longer or shorter regions may be selected depending on the specific organism or gene being analyzed. Several features were excluded from this analysis, including: (*i*) all exons, (*ii*) interspersed DNA repeat sequences (masked by Repeat Masker), and (*iii*) low complexity regions (di- and trinucleotide repeats that are longer than minimum motif length; <10% differences from perfect repeats are allowed). Noncoding exons may also be included into analysis.

SCM Discovery. In the first step, SCM, present at least once in each of the input orthologous sequences, are discovered by the SPLASH pattern discovery algorithm (24) (Fig. 1*B*). SPLASH discovers SCMs as conserved rigid motifs represented as regular expressions, where positions conserved across all sequences are represented by the corresponding nucleic acid symbol, and variable sites are represented by a "." character (e.g., AC.G.T-TA.T). SPLASH allows SCMs to be located anywhere within the input sequences, including multiple times within the same sequence. The choice of SPLASH parameters determines the class of motifs discovered by the algorithm. Critical parameters, in this case, are the minimum sequence support (minimum number of sequences, q, that must contain the motif), the motif density (minimum number of matching characters, k, required over a given window length, w), and the total length of the motif (i.e., the minimum number, l, of matching base pairs). For instance, a motif such as AC.T.AGGTA.T occurring in seven sequences would satisfy the density constraint of ($k = 6$, $w = 8$), because for each set of eight consecutive positions, six or more are fully defined. It also has a length $l = 9$, because it contains nine fully defined base pairs and a sequence support $q = 7$. Given the degenerate nature of TF-binding sites, a relatively loose density constraint of six exact matches for each 8-bp window ($k = 6, w = 8$) is recommended and was used in this paper. These parameters were chosen based on the observed changes of TF-binding sites in orthologous enhancers from multiple *Drosophila* species (18, 19). However, different SPLASH parameters can be selected depending on the specific set of analyzed organisms. Several values of l, ranging from $l = 15$ to $l = 21$ bp, were tested. For typical gene size (30 to >100 kb), smaller values (i.e., $l < 15$) produced such a large number of SCMs that the subsequent permutation clustering step could not be completed within a reasonable timeframe. Although the current minimum motif length (15 bp) is larger than a typical monomer TF-binding site, binding sites for cooperative TFs can overlap or reside very close to each other and form longer stretches of conserved sequences. Finally, given the relatively small number of analyzed species (seven genomes), we required that the discovered motifs be conserved within each sequence (i.e., support $q = 7$). As more genomes become available, q may be set to a value smaller than the total number of sequences without loss of generality. This improvement would prevent motif variability across genomes from reducing the approach sensitivity.

LPC Discovery. Given a set **M** of SCMs discovered from the set of orthologous sequences for single gene, we define an LPC as a subset $\{m_i\} \in \mathbf{M}$ of motifs, which are all contained within a region of predefined maximum cluster length, L, in each of the orthologous sequences from that same set (Fig. 1*C*). Note that this definition is independent of motif order and multiplicity. Thus, as long as they occur within L nucleotides, $\{m_1, m_{15}, m_8\}$ and $\{m_8, m_{15}, m_8, m_1\}$ are both valid instances of the $\{m_1, m_8, m_{15}\}$ canonical LPC (where motifs are ordered by their index value). The L constraint differentiates an LPC from a traditional permutation pattern (23), which does not restrict the motifs to occur within a predefined length region. EDGI considers strictly "maximal" LPCs, i.e., those such that the addition of any additional motif from **M** would further reduce the number of the LPC's occurrences. Given a set of SPLASH discovered SCMs, discovery of maximal LPC over the set of orthologous sequences is performed using the PromoClust algorithm [see supporting information (SI) *Text* for algorithm description]. PromoClust algorithm allows LPCs to be located anywhere within the input sequences.

We tested different values of the cluster window, ranging from

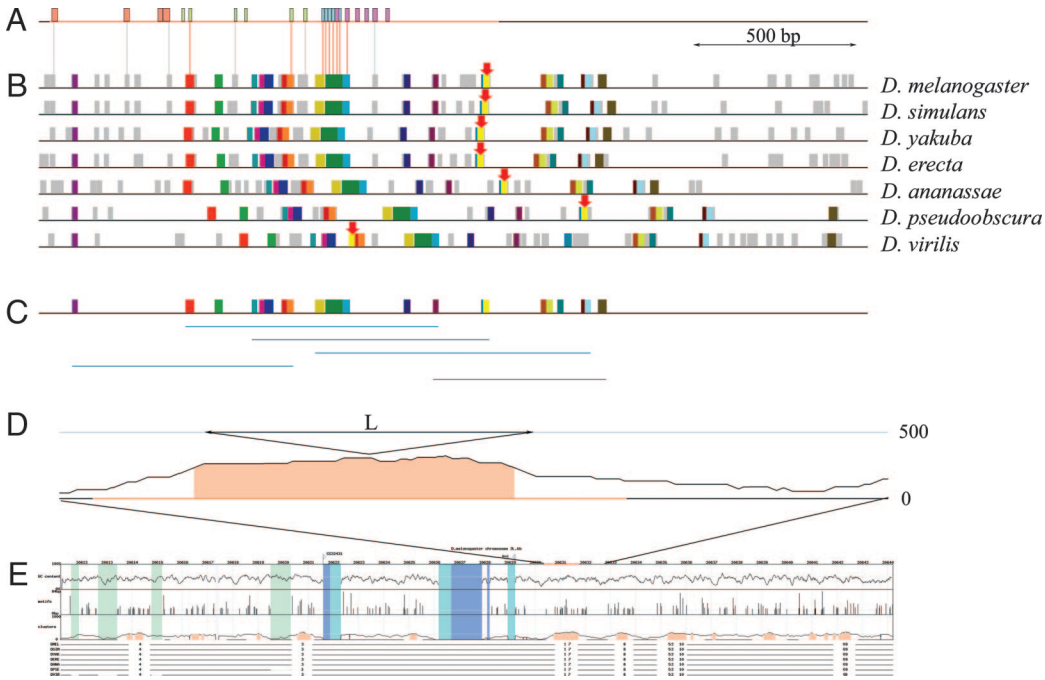

Fig. 1. Overview of EDGI algorithm and output for gene *knirps*. (*A*) Experimentally identified upstream enhancer for *knirps* gene (red horizontal line) and functional binding sites for *tailless* (orange boxes), *hunchback* (green boxes), *bicoid* (blue boxes), and *caudal* (purple boxes) TFs (30, 31). The horizontal line corresponds to the *D. melanogaster* genomic sequence from chromosome 3L (chromosomal coordinates 20630341–20632841 according to Release 4 of *D. melanogaster* genome). (*B*) EDGI-discovered SCMs in orthologous sequences from seven *Drosophila* species. The *D. melanogaster* region is identical to and aligned with the sequence shown in *A*. Colored boxes represent SCMs that contribute to LPCs [EDGI parameters (l = 15, L = 1,000)], and gray boxes represent SCMs that do not contribute to any LPC. Red lines connect functional binding sites from line A with EDGI-discovered SCMs that contribute to LPCs if SCM overlaps with at least 50% of binding site sequence. Gray lines connect functional binding sites with EDGI-discovered SCMs that do not contribute to any LPC. Example of reshuffled SCMs within LPC: motif GAC.TT.TGACTTTT.A.C (yellow box with red arrow) overlap with motif TG.G.TGACTTTTTGACTTT.C (blue box) in most species, whereas in *D. virilis* the orthologous sequence motif GAC.TT.TGACTTTT.A.C was translocated toward the 5′ terminal. (*C*) Location of LPCs (blue lines) in *D. melanogaster* sequence. Top line reproduces top line from *B*. (*D*) Graph of the SCS along the *D. melanogaster* sequence with nucleotide positions with highest SCS covering 15% of the noncoding *knirps* sequences shaded in orange. The orange line represents the predicted enhancer (LPCs overlapping with the selected nucleotide positions). (*E*) EDGI output with parameters [l = 15, L = 1,000]. The top line illustrates the gene position and structure according to Release 4 of the *D. melanogaster* genome (27). Transcription start sites are indicated with dark blue flags. Protein coding regions are shown in blue and untranslated exonic regions are shown in light blue. Gaps in pairwise alignments between *D. melanogaster* and other *Drosophila* species are shown in green and low-complexity regions (di- and trinucleotide repeats) and interspersed repeats (from Repeat Masker) are shown in gray. Known enhancers from the test set (not necessary all known enhancers) are shown by the red line along the *D. melanogaster* sequence. In the first data line, the local GC content is plotted using a window of 100 bp. In the second data line, the position of the discovered conserved motifs along the *D. melanogaster* sequence is shown by vertical lines; the height of the line corresponds to the length of the motif (scale is on the left). Note that only motifs that are included within an LPC are shown. The third data line shows a graph of the conservation score along the *D. melanogaster* sequence. Below this graph are the aligned sequences for *Drosophila* species used in the analysis. Numbers locate positions for top 10 LPCs. Gaps correspond to gaps in the pairwise alignments between *D. melanogaster* and these *Drosophila* species. A known enhancer overlaps with the two top-scoring LPCs identified by EDGI (LPCs 1 and 7) as well as with highest-scoring nucleotide positions covering 15% of the noncoding *knirps* sequences (orange picks).

L = 300 to L = 1,500 bp, based on the expected eukaryotic CRM size. The resulting LPCs were sorted by the total number of nucleotides overlapping with individual motifs comprising LPC. Only LPCs that contain a number of conserved positions greater than a predefined threshold $t_{\min} = 2l$ were retained. Thus LPCs comprising at least two SCMs or single SCM of an equivalent length were reported. Finally, we retained only those SCMs that contributed to an LPC (Fig. 1*C*). This step dramatically reduced the number of SCMs.

Enhancer Discovery. After all LPCs were discovered, each position in the input sequences was assigned a sequence conservation score (SCS), which is equal to the number of nucleotides in an SCMs within the region of size L, centered at this position (Fig. 1*D*). The advantage of this heuristic score is that motifs in overlapping LPCs collectively contribute to the score. We explored two metrics to test EDGI's ability to identify enhancers. In the first one [SCM metric (SCSM)], we selected nucleotide positions with a SCS above θ, with θ a predefined cutoff in the range 1–500. LPCs that overlap with selected positions are predicted to identify enhancers. In the second metric [percent coverage metric (PCM)], we selected the nucleotide positions with highest SCS covering ξ% of the input *Drosophila melanogaster* sequence, with ξ% a predefined cutoff in the range 1–100%. As was shown arlier (9), identifying the most-conserved noncoding nucleotide positions instead of setting a fixed conservation cutoff may provide a more efficient regulatory element selection due to the regional variation in evolutionary rates (25). This variation can affect the conservation of functional elements and SCSs for enhancers can vary dramatically from one gene (genomic region) to another.

Results

Evaluation of EDGI Performance. EDGI's ability to discriminate between real enhancers and nonenhancers was tested on a set of previously characterized functional and nonfunctional CRMs (26) (SI Table 2). All putative CRMs, both those validated as enhancers and those that failed validation, were initially identified as regions of the *D. melanogaster* genome containing an unusually high density of putative binding sites for five TFs

GENETICS

Table 1. Accuracy of CRM prediction by different methods

Method	Parameters l and L	Area under RP curve	Optimal accuracy	Maximal recall, $P = 1$	Maximal precision, $R = 1$
EDGI SCSM	15/300	0.88	0.83 (25)	0.27 (125)	—
	15/600	0.94	**0.88 (100)**	0.13 (230)	**0.83 (100)**
	15/1,000	0.96	**0.88 (155)**	0.47 (240)	**0.83 (155)**
	15/1,500	0.92	0.83 (210)	0.27 (365)	0.79 (160)
	18/300	0.83	0.77 (40)	0.47 (75)	—
	18/600	**0.97**	0.85 (75)	**0.67 (115)**	0.71 (50)
	18/1,000	0.96	0.85 (100)	0.53 (220)	0.79 (85)
	18/1,500	0.92	0.85 (160)	0.13 (335)	0.79 (140)
	21/300	0.57	0.69 (25)	0.33 (55)	—
	21/600	0.83	0.77 (65)	0.53 (95)	—
	21/1,000	0.94	0.85 (125)	0.33 (180)	0.68 (20)
	21/1,500	0.92	0.81 (145)	0.20 (255)	0.71 (55)
EDGI PCM	15/300	0.89	0.79 (25%)	0.40 (2%)	—
	15/600	0.95	**0.88 (12%)**	0.27 (1%)	**0.79 (25%)**
	15/1,000	**0.97**	**0.88 (14%)**	0.47 (2%)	**0.79 (20%)**
	15/1,500	0.93	0.79 (8%)	0.27 (1%)	**0.71 (25%)**
	18/300	0.83	0.77 (14%)	0.60 (14%)	—
	18/600	**0.97**	0.85 (25%)	**0.67 (10%)**	**0.75 (30%)**
	18/1,000	**0.97**	**0.88 (20%)**	0.47 (2%)	**0.71 (40%)**
	18/1,500	0.75	0.81 (30%)	—	**0.75 (30%)**
	21/300	0.59	0.73 (16%)	0.53 (16%)	—
	21/600	0.83	0.77 (14%)	0.60 (16%)	—
	21/1,000	0.94	0.85 (20%)	0.27 (2%)	0.65 (70%)
	21/1,500	0.93	0.77 (20%)	0.20 (1%)	0.71 (50%0
Multiple pairwise alignments	300	0.95	**0.85 (35%)**	0.27 (18%)	**0.71 (45%)**
	600	0.92	0.77 (25%)	0.27 (8%)	0.68 (50%)
	1,000	0.93	0.77 (20%)	0.27 (8%)	**0.71 (50%)**
	1,500	**0.96**	0.81 (20%)	**0.67 (20%)**	0.68 (45%)
Target Explorer		**0.98**	**0.92 (8)**	**0.60 (12)**	**0.88 (8)**

Numbers in parentheses show cutoff that produce indicated values for accuracy, recall, and precision. Bold indicates best results for each method.

involved in embryonic patterning: Bicoid, Hunchback, Kruppel, Knirps, and Caudal. Those that drove reporter gene expression in transgenic *D. melanogaster* embryos were considered validated enhancers, while the others (nonenhancers) did not show any activity in this assay (26). We excluded PCE8008, PCE8014, and PCE8026 from the set of nonenhancers because of their remote location from the transcription start site of the closest gene (>15 kb), which would prevent EDGI analysis given our choice $\Delta x =$ 15 kb.

Genomic regions for the genes regulated by (or proximal to) these elements were analyzed by using EDGI. *D. melanogaster* sequences from Release 4 (27) were selected as described in *Methods* section. Orthologous sequences were retrieved from the genomes of six additional *Drosophila* species (*Drosophila simulans*, *Drosophila yakuba*, *Drosophila erecta*, *Drosophila ananassae*, *Drosophila pseudoobscura*, and *Drosophila virilis*) following pairwise alignments with the *D. melanogaster* sequence (28). EDGI's performance was tested with several combinations of parameters by varying minimal motif length ($l = 15$, $l = 18$, and $l = 21$ bp) and maximal cluster window size ($L = 300$, $L = 600$, $L = 1{,}000$, and $L = 1{,}500$ bp). For several genes, we were unable to run the LPC discovery algorithm when the minimal motif length $l < 15$ bp was selected because of the prohibitive number of discovered SCMs. For similar reasons, EDGI failed to complete the analysis for two genes from the set of nonenhancers (CG32835 and *pum*) when $l = 15$ bp was used. A sequence from the test set was considered "discovered by EDGI as an enhancer" if at least 50% of its length overlapped with EDGI-predicted enhancers (see overlap between the red horizontal bar representing functional enhancer in Fig. 1*A* and the orange horizontal bar representing EDGI-discovered enhancer in Fig. 1*D*).

EDGI's ability to discriminate enhancers from nonenhancers at different cutoff values was calculated and plotted using recall-precision curves (SI Fig. 2 and Table 1) by varying respectively the θ and ξ% thresholds. Recall (R) was calculated as the ratio of enhancers that were discovered by EDGI to all enhancers in the test set. Precision (P) was calculated as the ratio of enhancers that were discovered by EDGI to all discovered enhancers and nonenhancers. Finally, Accuracy (A) was calculated as the ratio of correct predictions (enhancers discovered by EDGI and nonenhancers not discovered by EDGI) to the total number of enhancers and nonenhancers in the test set.

Using the SCSM, an optimal accuracy, $A_{\mathrm{SCSM}} = 88\%$, was produced (Table 1). This corresponded to three incorrectly predicted nonenhancers [parameters ($l = 15$, $L = 600$) and ($l = 15$, $L = 1{,}000$)]. Using the PCM, we obtained the same optimal accuracy $A_{\mathrm{PCM}} = 88\%$, corresponding to one missed enhancer and two incorrectly predicted nonenhancers [parameters ($l = 15$, $L = 600$), ($l = 15$, $L = 1{,}000$), and ($l = 18$, $L = 1{,}000$)]. We also compared the areas under the recall-precision curves for these two methods. This value illustrates how far the real curve is from the ideal one with only correct predictions. EDGI predictions using either type of scoring metrics produced curves with the same area coverage of 97% (for clustering parameters and cutoff values, see Table 1), indicating an excellent overall performance. Additionally, both metrics correctly predicted the same number of enhancers (67%) with no incorrect nonenhancer predictions (maximal R with $P = 1$). The SCSM predicted all enhancers with

a slightly lower false-positive rate than the PCM ($P_{SCSM} = 83\%$ vs. $P_{PCM} = 79\%$ with $R = 1$). Overall results were very stable with respect to the choice of parameters l and L.

In summary, the two metrics perform very similarly. However, using PCM appears to be slightly less sensitive to clustering parameters. To optimize accuracy for *Drosophila* genome analysis, the recommended EDGI parameters are [$l = 15, L = 1{,}000$], using $\xi\% = 14\%$ threshold for PCM or $\theta = 150$ threshold for SCSM. The results for this test set are available online at http://luna.bioc.columbia.edu/EDGI/test_set.

A key advantage of using a nonalignment method such as EDGI for CRM discovery is that the order of SCMs (putative binding sites) within LPCs (putative enhancers) can differ in different species, as found in functional enhancers (18). Consistently, almost all enhancers from the test set (with only one exception) were discovered as LPCs with reshuffled SCMs (see an example in Fig. 1*B*).

EDGI Comparison with Other Methods. EDGI results were compared with those obtained using Target Explorer (4). Like several other method (1–3), Target Explorer searches for clusters of putative binding sites for TFs with known binding specificities. We used the same training sets of binding sites for Bicoid, Hunchback, Kruppel, Knirps, and Caudal, as described (26). PWMs were produced for each TF. All CRM sequences from the enhancer and nonenhancer test sets for each *Drosophila* species were searched for binding site matches with a P value $<10^{-3}$. As shown (26), the density of conserved, but not necessarily aligned, binding sites allowed an almost perfect discrimination between the enhancer and nonenhancer sets. Therefore, the number of conserved sites for each TF in a putative CRM was equal to the minimum number of sites for that factor in each examined *Drosophila* species. Conserved binding sites for each TF were summarized and their density was calculated. For each cutoff value, recall and precision were calculated and plotted (SI Fig. 2 and Table 1). Using a minimal density of eight conserved sites per kilobase, this method discovered all enhancers and two incorrectly predicted nonenhancers, corresponding to an accuracy $A = 92\%$ (EDGI made three incorrect predictions with an accuracy $A = 88\%$). Interestingly, EDGI correctly discovered more enhancers with zero incorrect nonenhancer predictions than Target Explorer ($R = 67\%$ vs. $R = 60\%$ with $P = 1$). However, Target Explorer correctly discovered all enhancers with a lower rate of incorrect nonenhancer predictions than EDGI ($P = 88\%$ vs. $P = 83\%$ with $R = 1$). In summary, EDGI was comparable to Target Explorer in discovering validated enhancers from the test set. However, EDGI required no prior knowledge of the regulating TFs or their binding specificities.

We further compared EDGI to a more traditional phylogenetic footprinting method based on sequence alignment. Pairwise alignments between *D. melanogaster* and other *Drosophila* species are visualized by Vista plots (28). They allow a semiquantitative analysis of *Drosophila* sequence conservation (SI Fig. 3*A*). Pink areas on the plot mark short noncoding aligned regions of ≥10 bp with ≥90% nucleotide identity. These parameters were provided by the Bergman *et al.* study (11), which suggested that clusters of short conserved noncoding regions extracted from pairwise alignments can be used to predict CRMs. Importantly, this method recognizes only aligned conserved regions between *D. melanogaster* and one other species. The authors attempted to extend this method to multiple pairwise alignments between *D. melanogaster* and other *Drosophila* species by finding intersections of clusters of conserved regions among multiple pairwise alignments. SI Fig. 3 illustrates a comparison between the Vista plot (SI Fig. 3*A*) and EDGI (SI Fig. 3*B*) outputs. Vista fails to discover an enhancer in the gene *giant* (*gt*), because the density of short conserved regions in the enhancer sequence is similar to the density in other *gt* genomic sequences. In contrast, this enhancer overlaps with the top LPCs identified by EDGI and with noncoding *gt* sequences with the highest SCS.

Because the method suggested by Bergman is qualitative in nature, we developed a more quantitative variant so that its accuracy on the test set could be measured. We analyzed each pairwise alignment in turn and selected all short ungapped aligned noncoding regions of 10 bp with ≥90% nucleotide identity. Then for each pairwise alignment each position in the *D. melanogaster* sequence was assigned a conservation score equal to the number of base pair covered by these conserved regions within a window ranging from $L = 300$ to $L = 1{,}500$ bp and centered at this position. Highest-scoring positions covering $\xi\%$ of the input *D. melanogaster* sequence were then selected for each alignment. Stretches of highest-scoring positions were extended to include all conserved regions that contributed to these scores (conserved domains). Finally, only positions that were included in conserved domains in all pairwise alignments were considered globally conserved. Again, enhancers and nonenhancers in the test set were considered discovered if at least 50% of their length overlapped with globally conserved positions. The most accurate predictions were made with $L = 300$ bp: two enhancers were missed, whereas two nonenhancers were incorrectly discovered (four incorrect predictions vs. three incorrect predictions for EDGI) (Table 1). This alignment method achieved the best accuracy $A = 85\%$ using $\xi\% = 35\%$ threshold. However, EDGI showed an optimal accuracy $A = 88\%$ using $\xi\% = 12$–14% threshold. Thus, to achieve accuracy comparable with EDGI, the sequence alignment based method predicts more than one-third of the input DNA sequence as a candidate enhancer. We suggest that this difference is a significant advantage of EDGI.

Additionally, the area under the recall-precision curves was higher, and the number of incorrect predictions with all enhancers discovered was lower for EDGI (Table 1). In summary, EDGI was better at distinguishing between the sets of enhancers and nonenhancers, producing significantly fewer DNA sequences required for further biological validation than a method based on combined multiple pairwise sequence conservation.

EDGI Web Server. We developed a Web server that integrates the above steps for identifying putative CRMs (http://luna.bioc.columbia.edu/EDGI). The output consists of two parts: a text file that lists the sequences of the discovered, clustered motifs and their positions in the *Drosophila* genomes and a graphical display of EDGI's results (Fig. 1*E*).

Discussion

We have presented a powerful approach for CRM discovery in eukaryotic genomes, which significantly extends the range of analysis of current methods. EDGI does not require any prior knowledge of the TFs or their DNA-binding specificities. It is based only on the assumption that regulating TFs as well as their binding sites are conserved in orthologous enhancers of evolutionarily related species. Unlike traditional comparative genomics methods (6–13), EDGI identifies conserved short DNA motifs without requiring sequence alignment. Thus, EDGI cannot miss SCMs due to the lack of conservation over an extended region. EDGI introduces the use of LPCs to discover regulatory modules where motifs are conserved across several orthologous genomes, but their order and number may vary. Thus, unlike sequence alignments methods, EDGI can discover CRMs in which evolutionary pressures have conserved the overall set of regulatory inputs but altered the spacing and order of their corresponding binding sites (Fig. 1*B*). EDGI scores predicted CRMs based on the total length of the conserved motifs in a

permutation cluster. Thus, random background conservation cannot influence the score as in sequence alignment methods.

EDGI is the quantitative method that allows the systematic discovery of enhancers from the genomes of multiple *Drosophila* species. It gives us a unique opportunity to compare orthologous sequences from multiple organisms with an evolutionary distance of <40 million years. Unlike genes for evolutionarily distant organisms (human–mouse, human–fish), orthologous genes from closely related species are most likely to retain the same expression regulation mechanism, involving a very similar set of TFs in each species. However, given their short evolutionary separation, there is relatively little sequence divergence between each pair of *Drosophila* species. Such a high sequence conservation rate obscures the difference between functional and passively conserved elements as revealed by pairwise alignment methods. However, because each species diverged independently after separation from a common ancestor, the phylogenetic distances covered are effectively additive. Thus a comparison among multiple *Drosophila* species effectively reflects a much longer period than the actual time of divergence from a common ancestor. Another way to view this advantage is that passively conserved (i.e., nonselected) sequences are likely to be different for each pairwise species comparison. Hence, orthologous genomic sequences from multiple *Drosophila* species are likely to share similar sets of conserved TF-binding sites, and conservation across multiple species is more likely to correspond to functional regions. Our results, as well as sequence comparisons of numerous primate species (29), demonstrate that orthologous sequences from multiple closely related species are sufficient to identify important regions of conservation that encode functional elements without additional biological knowledge.

Consistently, we find that EDGI's ability to distinguish between enhancers and nonenhancers is comparable to Target Explorer (4), a method that requires full prior knowledge of the regulating TFs and their binding sites (Table 1). Importantly, EDGI predicted CRMs with higher accuracy than the method based on multiple pairwise sequence alignments (11) (Table 1) and with much smaller amounts of predicted regulatory sequences requiring biological validation.

Currently, we are validating previously undescribed LPCs predicted by EDGI using several biological assays. Using a standard reporter gene assay, preliminary tests suggest that at least one-third of EDGI-discovered LPCs show enhancer activity *in vivo* (data not shown). Some of the failures may be due to limitations in the reporter gene assay that was used or to the limited developmental stages or tissues that were examined. Several methods are also in progress to extend EDGI's ability to discover biologically active elements. For example, we are combining EDGI and Target Explorer to discover LPCs that also contain a user-defined set of binding sites for TFs thought to regulate gene expression at a particular developmental stage or in specific tissues. Because false-positive predictions for these two methods only partially overlap (data not shown), combining these two approaches can further increase the accuracy of CRM prediction. EDGI can also be modified to allow the analysis of genes that are coexpressed, based on gene expression profiling, and thus may be controlled by regulatory modules that share inputs.

The current limitation for the minimum motif length of 15 bp does not allow comprehensive discovery of monomer TF-binding sites within EDGI-predicted enhancers. As can be seen on Fig. 1 *A* and *B*, only about one-half of the functional binding sites in the upstream enhancer for gene *knirps* overlap with EDGI-discovered SCMs that contribute to LPCs. We are planning to develop a recursive version of EDGI, to allow the further dissection of each LPC into its individual conserved motifs of ≈8 bp, thus providing direct insight into the combinatorial control mechanism integrated by a CRM.

Finally, although EDGI was used for CRM discovery in *Drosophila*, it should be readily adaptable to discover regulatory modules in other organisms, including mammals and other vertebrates.

We thank Oliver Hobert for careful reading of and critical remarks on the paper. This work was supported by National Institutes of Health Grants GM054510 and GM074105 (to R.S.M.) and U54 CA121852-02 (to A.C.).

1. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB (2002) *Proc Natl Acad Sci USA* 99:757–762.
2. Markstein M, Markstein P, Markstein V, Levine MS (2002) *Proc Natl Acad Sci USA* 99:763–768.
3. Rajewsky N, Vergassola M, Gaul U, Siggia ED (2002) *BMC Bioinformatics* 3:30.
4. Sosinsky A, Bonin CP, Mann RS, Honig B (2003) *Nucleic Acids Res* 31:3589–3592.
5. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT (1988) *J Mol Biol* 203:439–455.
6. Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, Brenner S (1995) *Proc Natl Acad Sci USA* 92:1684–1688.
7. Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA (2000) *Science* 288:136–140.
8. Gottgens B, Barton LM, Gilbert JG, Bench AJ, Sanchez MJ, Bahn S, Mistry S, Grafham D, McMurray A, Vaudin M, *et al.* (2000) *Nat Biotechnol* 18:181–186.
9. Flint J, Tufarelli C, Peden J, Clark K, Daniels RJ, Hardison R, Miller W, Philipsen S, Tan-Un KC, McMorrow T, *et al.* (2001) *Hum Mol Genet* 10:371–382.
10. Grad YH, Roth FP, Halfon MS, Church GM (2004) *Bioinformatics* 20:2738–2750.
11. Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, *et al.* (2002) *Genome Biol* 3:RESEARCH0086.
12. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE (2000) *Nat Genet* 26:225–228.
13. Dieterich C, Cusack B, Wang H, Rateitschak K, Krause A, Vingron M (2002) *Bioinformatics* 18(Suppl 2):S84–S90.
14. Dubchak I, Brudno M, Loots GG, Pachter L, Mayor C, Rubin EM, Frazer KA (2000) *Genome Res* 10:1304–1306.
15. Prakash A, Tompa M (2005) *Nat Biotechnol* 23:1249–1256.
16. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) *Nature* 423:241–254.
17. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M (2003) *Science* 301:71–76.
18. Ludwig MZ, Bergman C, Patel NH, Kreitman M (2000) *Nature* 403:564–567.
19. Dermitzakis ET, Bergman CM, Clark AG (2003) *Mol Biol Evol* 20:703–714.
20. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, *et al.* (2005) *Nat Biotechnol* 23:137–144.
21. McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE (2001) *Nucleic Acids Res* 29:774–782.
22. Blanchette M, Tompa M (2002) *Genome Res* 12:739–748.
23. Eres R, Landau GM, Parida L (2004) *J Comput Biol* 11:1050–1060.
24. Califano A (2000) *Bioinformatics* 16:341–357.
25. Hardison RC, Roskin KM, Yang S, Diekhans M, Kent WJ, Weber R, Elnitski L, Li J, O'Connor M, Kolbe D, *et al.* (2003) *Genome Res* 13:13–26.
26. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE (2004) *Genome Biol* 5:R61.
27. Grumbling G, Strelets V (2006) *Nucleic Acids Res* 34:D484–D48.
28. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) *Nucleic Acids Res* 32:W273–W279.
29. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM (2003) *Science* 299:1391–1394.
30. Pankratz MJ, Busch M, Hoch M, Seifert E, Jackle H (1992) *Science* 255:986–989.
31. Rivera-Pomar R, Lu X, Perrimon N, Taubert H, Jackle H (1995) *Nature* 376:253–256.