Genome-wide discovery of human heart enhancers

Leelavati Narlikar,[1] Noboru J. Sakabe,[2] Alexander A. Blanski,[2] Fabio E. Arimura,[2] John M. Westlund,[2] Marcelo A. Nobrega,[2,3] and Ivan Ovcharenko[1,3]

[1]*Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (NIH), Bethesda, Maryland 20894, USA;* [2]*Department of Human Genetics, The University of Chicago, Chicago, Illinois 60637, USA*

The various organogenic programs deployed during embryonic development rely on the precise expression of a multitude of genes in time and space. Identifying the *cis*-regulatory elements responsible for this tightly orchestrated regulation of gene expression is an essential step in understanding the genetic pathways involved in development. We describe a strategy to systematically identify tissue-specific *cis*-regulatory elements that share combinations of sequence motifs. Using heart development as an experimental framework, we employed a combination of Gibbs sampling and linear regression to build a classifier that identifies heart enhancers based on the presence and/or absence of various sequence features, including known and putative transcription factor (TF) binding specificities. In distinguishing heart enhancers from a large pool of random noncoding sequences, the performance of our classifier is vastly superior to four commonly used methods, with an accuracy reaching 92% in cross-validation. Furthermore, most of the binding specificities learned by our method resemble the specificities of TFs widely recognized as key players in heart development and differentiation, such as SRF, MEF2, ETS1, SMAD, and GATA. Using our classifier as a predictor, a genome-wide scan identified over 40,000 novel human heart enhancers. Although the classifier used no gene expression information, these novel enhancers are strongly associated with genes expressed in the heart. Finally, in vivo tests of our predictions in mouse and zebrafish achieved a validation rate of 62%, significantly higher than what is expected by chance. These results support the existence of underlying *cis*-regulatory codes dictating tissue-specific transcription in mammalian genomes and validate our enhancer classifier strategy as a method to uncover these regulatory codes.

[Supplemental material is available online at http://www.genome.org.]

The regulatory apparatus of a vertebrate gene typically consists of a proximal promoter and multiple transcriptional regulatory elements (enhancers and silencers) (Maston et al. 2006). These regulatory elements are often distant from the promoter, with the separation reaching millions of nucleotides (Lettice et al. 2003; Nobrega et al. 2003) and sometimes acting over intermediate genes (Loots et al. 2000; Tschopp et al. 2009). Cases of *trans*-acting regulatory elements capable of activating promoters on adjacent chromosomes have also been previously described (Lomvardas et al. 2006).

Identification of regulatory elements has always been a challenge, as it implies locating a small (typically a few hundred base pair) segment embedded in a large segment of otherwise anonymous sequence. Comparative sequence analysis has been instrumental in facilitating identification of regulatory elements that have been deeply conserved through evolution, with many deeply conserved noncoding sequences shown to act as enhancers in experimental models (Woolfe et al. 2005; Pennacchio et al. 2006). Enhancers identified using solely the increased degree of evolutionary conservation display notable variation in spatial expression patterns (Woolfe et al. 2005; Pennacchio et al. 2006), as conservation is blind to the functional identity of enhancers. Furthermore, the complex expression patterns of genes are often achieved through the concerted action of multiple enhancers in a single locus (Nobrega et al. 2003; de la Calle-Mustienes et al. 2005). As a result, predicting the precise temporal and spatial pattern regulated by each enhancer in a locus becomes a significant challenge.

With the recent advances in various high-throughput experimental platforms, genome-wide profiling of activation and repression signatures of proximal and distant regulatory elements has become feasible (Robertson et al. 2007). For example, using chromatin immunoprecipitation with massively parallel sequencing (ChIP-seq) technology, 55,000 tissue-specific putative enhancers have been reported in the human genome for different cell types, with only 5400 of them overlapping across different cell lines (Heintzman et al. 2009). ChIP-seq experiments targeting EP300 (also known as p300), a transcriptional coactivator common to enhancers and promoters, have been very accurate in localizing developmental enhancers in mice, displaying a remarkable 80% specificity in in vivo validation assays (Visel et al. 2009). Microarray gene expression data have also been essential in validating predicted enhancers by matching enhancer tissue-specificity with the expression pattern of closely positioned genes (Pennacchio et al. 2007; De Val et al. 2008; Visel et al. 2009) and have served as a template of tissue-specific signals extracted from the promoters of coexpressed genes (Sharan et al. 2004; Waleev et al. 2006).

Another strategy of identifying regulatory elements is the use of computational tools that scan DNA in search of certain sequence-based signatures. The computational identification of tissue-specific enhancers has mainly relied on detection of clusters of binding sites specific to activator proteins for which appropriate tissue specificity has been previously reported (Thompson et al. 2004; Segal et al. 2008). In *Drosophila*, the homotypic structure of enhancers, with multiple transcription factor binding sites (TFBSs)

[3]**Corresponding authors.**
E-mail ovcharei@ncbi.nlm.nih.gov.
E-mail nobrega@uchicago.edu.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.098657.109. Freely available online through the *Genome Research* Open Access option.

of the same transcription factor (TF), has been especially helpful in locating enhancers (Lifanov et al. 2003). In humans, heterotypic clusters of binding sites have been successfully used to scan the genome sequence to predict tissue-specific enhancers using a previously characterized TFBS enhancer structure (Hallikas et al. 2006; De Val et al. 2008).

Identification of the enhancers partaking in the heart developmental program has always been of particular interest. Heart development proceeds during embryogenesis through a series of intertwined genetic programs, which are precisely orchestrated in time and space by transcriptional factors and chromatin regulators that activate and/or suppress downstream gene targets. Deviations from these finely tuned events lead to congenital heart diseases (CHD), afflicting almost 1% of live births and considered the number one cause of neonatal deaths (Bruneau 2008).

Identifying the *cis*-regulatory motifs that dictate the expression of genes involved in heart development would provide a substantial contribution to the appreciation of what genetic pathways partake in this process. It would also provide the genomic coordinates of functional noncoding sequences that, if mutated, might lead to abnormal gene expression and result in CHDs. Therefore, strategies to map *cis*-regulatory modules (CRMs) of genes involved in heart development and dissecting their components has become the focus of major efforts in genomics. Several previous computational studies have addressed the regulatory code of aspects of cardiovascular development, but most of them relied heavily on the previous knowledge of associated TFBS configurations in cardiovascular enhancers (Sun et al. 2006; De Val et al. 2008). Each study targeted a different enhancer structure as bait in the genome scan, probably mimicking different developmental processes in such an approach. To establish a generalized classification mechanism capable of highlighting heart enhancers within the pool of noncoding DNA in the human genome, we propose combining footprints of multiple regulatory pathways into a model that selects key discriminatory TF binding sites within heart enhancers and scores their combinatorial occurrence.

In this study, we utilized the largest available data set of heart enhancers with expression validated in vivo using either mouse or zebrafish embryos to train a discriminatory computational model. We identified 41 TF binding specificities from the TRANSFAC and JASPAR databases and five additional binding specificities retrieved using a modified Gibbs sampling approach to be strongly associated with heart enhancers. The analysis of sequence composition of heart enhancers displayed evidence for 50 of them having similar sequence composition, thus being amenable to classification using a single set of classification rules with 92% accuracy, which is significantly higher than using other currently available methods. We identified 41,930 potential heart enhancers across the human genome using the developed classification rules. Using an in vivo transgenic zebrafish reporter assay, we tested 26 predicted elements, and validated 62% as heart enhancers.

Results

Building the heart enhancer classifier

A large set of experimentally validated heart enhancers is necessary to build an accurate sequence-based classifier of sequence signatures that discriminate heart enhancers from the remaining vast amount of noncoding DNA in the human genome. Being interested in the heart regulatory program at embryonic development, we focused our search exclusively on heart enhancers active during heart development and differentiation. First, we compiled a data set of 30 enhancers from the literature, for which heart activity has been recorded in vivo using reporter assays in various vertebrate model organisms (Supplemental Table S1). Next, we added 14 heart enhancers identified through the random scan of enhancer activity of deeply conserved regions in the human genome (Pennacchio et al. 2006). Finally, we performed an enhancer screen of noncoding elements from several selected heart gene loci using zebrafish enhancer assays (see Methods) that produced an additional 33 heart enhancers (M Nobrega, unpubl.). In the input set of 77 embryonic heart enhancers, 35% of the sequences were proximal to the transcription start site (within 2 kb), 27% were intronic, and the remaining 38% were distal intergenic. All these elements are conserved between human and mouse genomes; 42% are also conserved between human and chicken genomes (with >70% identity, the default option of the ECR Browser; Ovcharenko et al. 2004).

We used a multi-tiered approach consisting of de novo motif discovery, Markov sequence feature characterization, known motif mapping, and feature selection based on regression to build a classifier that distinguishes heart enhancers from other noncoding regions. The classifier was based on three different sets of features derived from the sequences. The first set of features was drawn from binding specificities of vertebrate TFs compiled from the TRANSFAC (Wingender et al. 2001) and JASPAR (Sandelin et al. 2004) databases. To account for the binding of TFs for which binding specificity has not been recorded in TRANSFAC/JASPAR, we employed PRIORITY (Narlikar et al. 2007)—a tool based on Gibbs sampling—to search for de novo motifs enriched in the set of heart enhancers. Putative binding site occurrences derived from these motifs made up the second set of features. Finally, we note that different functional genomic regions often have specific sequence features that could be captured using Markov models. For example, a first-order Markov model, which models the probability of observing a nucleotide based on the previous nucleotide, has been used successfully to detect splice sites (Zhang and Marr 1993; Salzberg 1997), third-order Markov models have been used to detect protein-coding genes (DeCaprio et al. 2007), and higher-order Markov models have been used to model intergenic regions of different organisms (Thijs et al. 2001) and in meta-genome analyses (McHardy et al. 2007). To detect whether heart enhancers display Markovian sequence signatures, our third set of features contained Markov models of all orders between zero and five.

We had a total of 727 features and expected many of these features to have a minor impact on classification, since not all TFs are presumably active during heart development. Moreover, TFs are likely to differ in relative contribution to the heart regulatory program, with only a subset causing a major impact. We built the classifier using LASSO (Tibshirani 1996), which learns a linear formula comprising a small subset of relevant features with appropriate weights. LASSO has previously been used in genomic contexts to distinguish nucleosomal DNA from nucleosome-depleted DNA while simultaneously learning relevant sequence features (Lee et al. 2007) and to select biomarkers based on gene expression data (Ghosh and Chinnaiyan 2005). We assumed no prior knowledge of TFs active in the heart, with the goal being to discover them using the feature weights learned by the classifier.

It is important to note that not all motifs in TRANSFAC, JASPAR, and the de novo set are equivalent in terms of their information content. In other words, some motifs in our set of features are more "stringent" (they match few DNA regions) while others are more "relaxed" (they match many DNA regions). This

can be explained by the varying specificities of TFs, and also because some of these motifs are built from limited TF binding sites. While the stringent motifs can result in identification of fewer true binding sites, the relaxed motifs can result in many false positives. Since LASSO is based on the discriminatory power of the features, both these kind of motifs are expected not to be selected and hence not affect the performance of the classifier.

The overall generality of a classifier can be best assessed when tested on an independent set. For this purpose, we conducted a standard fivefold cross-validation (CV) procedure where the classifier is trained on four-fifths of the training data and tested independently on the remaining one-fifth of the training data. As our control set, we drew a random sample of 1000 noncoding human sequences with a similar sequence-length distribution as the enhancers. The classifier achieved an average sensitivity of 77% at a false-positive rate of 50% on the five distinct test sets. While this implies that the classifier can distinguish signal from noise (a noninformative classifier would have achieved only 50% sensitivity), it does so at the cost of many false predictions.

Extracting a homogeneous heart enhancer set

Our set of heart enhancers is largely heterogeneous, comprising data from experiments conducted at different laboratories employing various animal models and experimental conditions, including different developmental stages. Furthermore, the heart itself contains several distinct cell types, including cardiac myocytes, smooth muscle, endothelial cells, and fibroblasts that are most probably controlled by different genetic pathways. In other words, heart enhancers are not likely to be bound by the same set of TFs or have sequence similarities if they are active in different cellular contexts. Therefore, learning a common "heart enhancer formula" from enhancers acting under such broad contexts based solely on sequence features is difficult, if at all possible.

In an effort to reduce the sequence heterogeneity of the 77 heart enhancers, we focused on selecting a large subset of these sequences sharing homogeneous sequence features. Toward this end, we repeated the fivefold CV procedure 20 times using different random splits of both positive and negative sequence sets into training and test sets, and tested the validation of the classifier on each individual element 20 times separately. From the results of these CVs, we assessed the number of times each sequence was predicted as positive: This frequency effectively measured the similarity of each element to the rest of the sequence set according to the classifier (Fig. 1A). As seen from the shift toward higher frequencies, the majority of the sequences were consistently predicted correctly. We grouped these sequences, predicted as positives more than half the time, as our new positive set. This new set, which contains 50 sequences, represents a cluster of homogeneous sequences within our data set. This homogeneity is based on the sequence features considered in the classification. Upon further analysis of this subset, other features such as conservation and GC content did not show up as significantly different from the rest of the 27 sequences.

All further computations and analyses were performed on this homogeneous set. It is critical to note that reducing the size of our positive set might hurt the sensitivity of our classifier, but our objective was to make fewer false predictions.

Accurate classification of heart enhancers

We retrained the classifier with the selected homogeneous set as our positive set and used a separate randomly generated control set of 1000 sequences with matching GC content and length distribution as the negative set. This different control set was used to ensure that the selection of the homogeneous set described in the earlier section was not biased toward any specific features of the previous control set. We conducted a similar CV analysis, but with 10 folds to account for the reduction in the size of the data set, to judge the performance of the new classifier. We compared its performance with four state-of-the-art methods that detect CRMs: CisModule (Zhou and Wong 2004), Cluster-Buster (Frith et al. 2003), MSCAN (Frith et al. 2003; Alkema et al. 2004), and Stubb (Sinha et al. 2006) (Fig. 1B). Our method outperformed all others in terms of both sensitivity and specificity. For example, at 0% false-positive rate, our method achieved a sensitivity of 30%, while all other methods achieved a sensitivity of at most 8%. The area under the receiver operating characteristic (ROC) curve—often used as a measure of performance or prediction accuracy, with 1.0 and 0.5 values depicting ideal and random classification, respectively—was 0.92 for our method, 0.72 for Cluster-Buster, 0.71 for Stubb, 0.64 for MSCAN, and 0.52 for CisModule.

To confirm that our method was not at an advantage due to the positive set being chosen based on results of our classifier (as described in the earlier section), we repeated the multiple fivefold CV and selection procedure using Cluster-Buster (the closest competitor to our method). Cluster-Buster produced an area under the ROC curve of only 0.77 for its newly selected homogeneous set. This number is only slightly higher than the 0.72 it achieved when the selection was done using our classifier, and much lower than 0.92 obtained with our classifier.

Selected features have been previously implicated in heart development

In LASSO linear regression implementation, features irrelevant to the classification receive zero weight, while those associated with the signal and control set receive positive and negative weights, respectively. Out of the original 727 features, only 45 (6.2%) were assigned nonzero weights, suggesting their relevance to predicting heart enhancers. They contained 30 known TF binding specificities from TRANSFAC/JASPAR (excluding redundant binding definitions; these correspond to four TFs), five de novo motifs, and two Markov features (Supplemental Table S2). Binding specificities of MEF2 and SRF obtained the maximum positive weight: Both of these TFs are known to play an important role in heart development (Edmondson et al. 1994; Sepulveda et al. 2002), with their knockout in mouse displaying severe cardiovascular abnormalities (Bi et al. 1999; Miano et al. 2004). Additionally, 18 out of the total 26 TFs with a positive weight have been previously shown to be active in the heart (Fig. 2A; Supplemental Table S2). These 18 features contribute to >65% of the total positive weight. Four motif-based features were selected with a negative weight, implying that the presence of these features in a DNA sequence reduces the likelihood of them acting as a heart enhancer. For three of them, IRF, FREAC2, and ZF5, TFs believed to recognize these motifs have been reported to act as repressors or are not implicated in heart development (Yokoro et al. 1998; Aitola et al. 2000; Sherry 2002), supporting the notion that their negative weights are indicative of their absence from gene regulation in the heart. It is important to note that the classifier learns motifs or sequence patterns that are either over- or underrepresented in the heart enhancer set. One must be cautious in ascribing specific TFs that bind these motifs since some motifs are known to be recognized by multiple TFs. Further experimental work confirming the identities of TFs binding these regions is warranted.

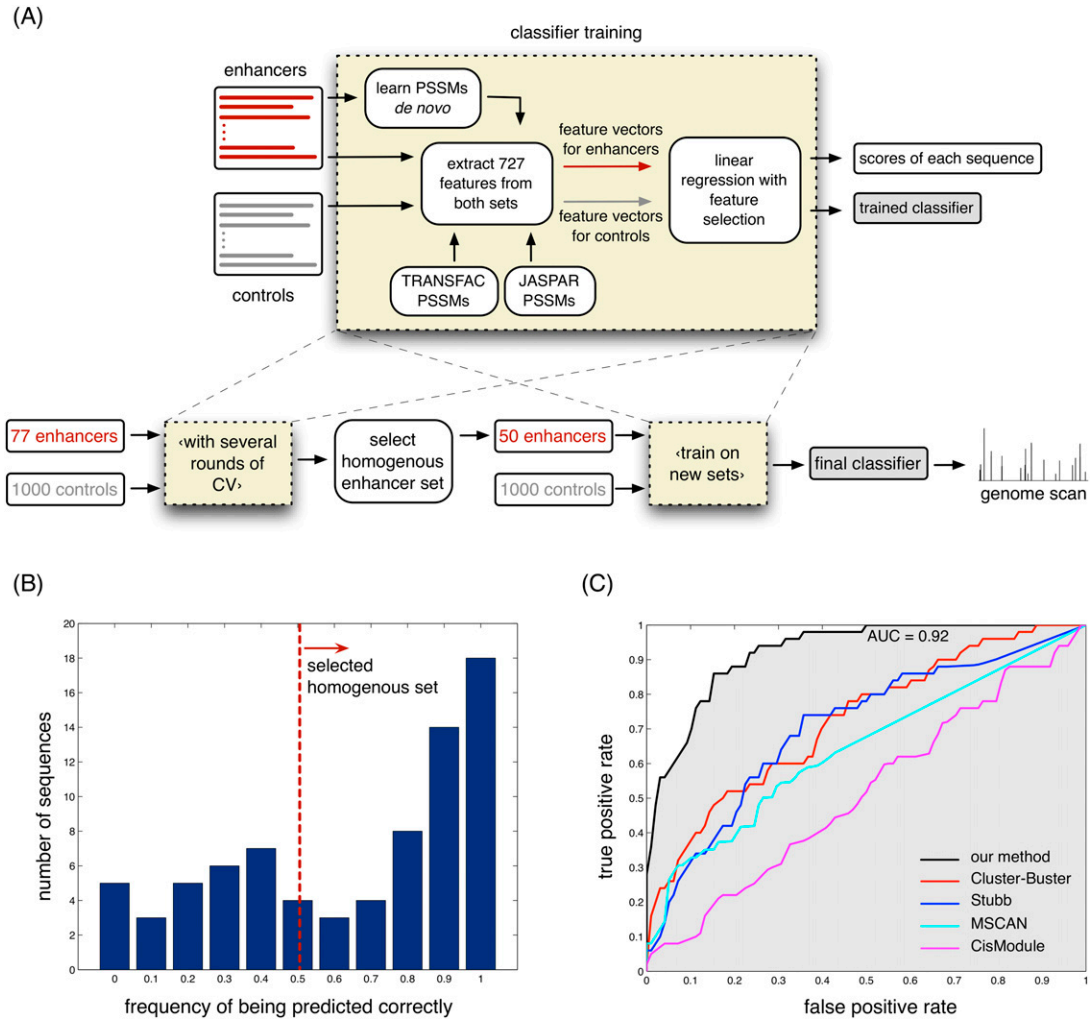


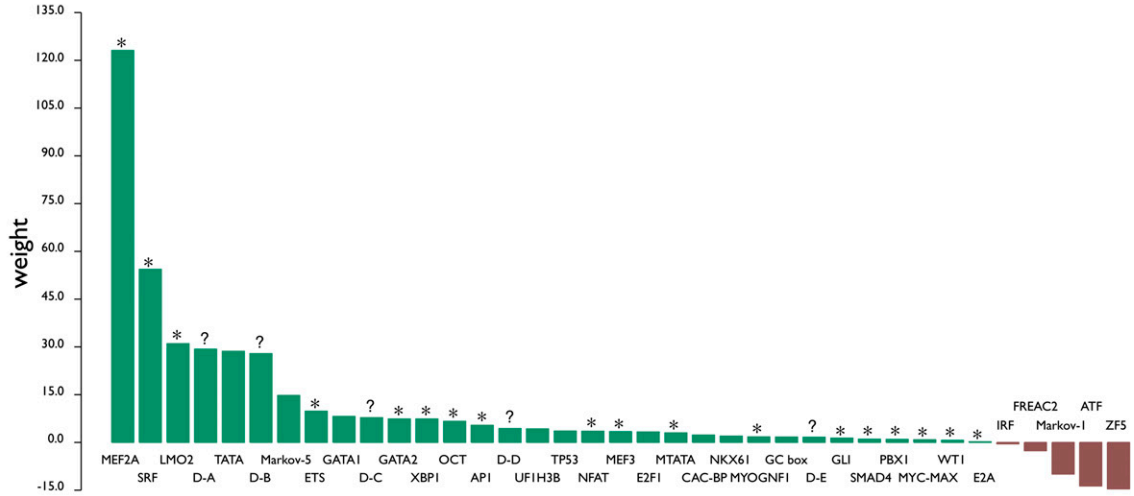
Figure 1. (*A*) Overview of the methodology. The yellow box shows the main classifier that takes as input two sets of sequences: enhancers and controls. The classifier is used first to select a homogenous set of enhancers and then used again to classify between the selected set and control sequences. (*B*) Distribution of positive sequences predicted correctly. Almost one-third of the sequences are predicted consistently (>50% of the time) as positives (red dotted line). Sequences to the *right* of the line were considered homogenous. (*C*) ROC curve for five different methods on selected homogeneous sets. Performance details of our method and of four state-of-the-art methods are shown here. The maximum area under the ROC curve is achieved by our method (0.92) (shaded in gray).

In addition to known motifs, five of the 20 de novo motifs—WRATAASG, TWTAAMNAGS, ARRGNNWKCG, GYTYMCWNTT, and CCNKCCCCYS[4]—were considered important by the classifier. As these motifs represent binding specificities not profiled in either the TRANSFAC or JASPAR databases, we investigated their similarity to binding similarities of families of TFs using the STAMP tool (Mahony and Benos 2007) (Fig. 2B). These motifs resemble the binding specificity of TFs in the families of LMO2, MEF2, ETS, and SP1—transcription factors known to play roles in heart development (Iida et al. 1999; Flesch 2001; Zhu et al. 2005; Pham et al. 2007; Gratzinger et al. 2009). The fact that these novel motifs were chosen in addition to the database motifs can be explained by one of the three scenarios: (1) specificities of the TFs binding the database motifs may not have been characterized precisely; (2) these motifs could be recognized by novel TFs, the specificities of which have not been characterized so far; or (3) the database TFs, specificities of which are characterized using in vitro methods, may have slightly different specificities in the heart due to various cofactors in vivo. Indeed, studies have shown examples of the same TF having slight differences in binding preferences in different tissues (Andres et al. 1995). In conclusion, our analysis predicted these motifs to be characteristic to heart enhancers, but follow-up studies are necessary to identify TFs that bind these sequences.

To assess the importance of the Markov features and their contribution to the classifier performance, we retrained the classifier without them, which yielded an area under the ROC curve of 0.86, a nonnegligible 7% reduction in accuracy. Additionally, only one Markov model, the fifth order, was chosen with a high positive weight. As fifth-order Markov models are based on 6-mers, which is the typical length of a TF binding site, it is likely that these Markov models capture certain dependencies within binding sites that other motif related features cannot.

[4]IUPAC codes are used to depict degenerate positions.

(A)



(B)

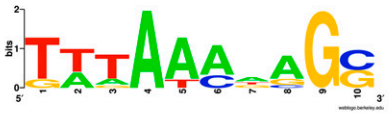
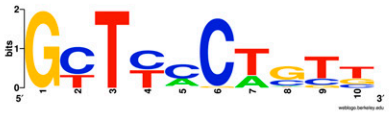
Name	Motif logo	Closest match according to STAMP (p-value)	Fraction of sequences with ≥ 1 match
A		LMO2 (2.6×10^{-8})	
B		MEF2 (5.0×10^{-8})	
C		ETS (1.6×10^{-4})	
D		ETS (8.3×10^{-7})	
E		SP1 (2.8×10^{-7})	

Figure 2. (*A*) Feature weights. (Green) Positive weights learned by the classifier; (brown) negative weights. Motif features of the same TF are clubbed together. The names of the features are listed near the baseline of the graph. (*) Features known previously to be implicated in heart activity or heart development; (?) de novo motifs A–E. (*B*) The five de novo motifs with positive weights. STAMP (Mahony and Benos 2007) was used to predict de novo motif associations with binding specificities of TF families from TRANSFAC and JASPAR. The top match with its *P*-value is shown. The last column indicates the fraction of sequences in the enhancer and the control set containing a match to each de novo motif.

Heart enhancers in the human genome

We next used the classifier as a genome-wide predictor of human heart enhancers. We selected conserved noncoding elements (CNEs) with at least 70% identity across the human and mouse genomes (Ovcharenko et al. 2004) and fragmented them into overlapping windows of size 150 bp, a length consistent with previously known enhancer elements (Sinha and Fuchs 2001; de Souza et al. 2005). A total of 730,000 CNEs were longer than 150 bp, with an average distance of 4463 bp between neighboring CNEs. Each CNE was assigned the score of the highest scoring fragment within it. The higher the score, the more likely the CNE will act as a heart enhancer according to the developed classifier.

A total of 42,000 sequences were assigned a positive score by the classifier, and these sequences were selected as putative heart enhancers (for a complete list of predictions, refer to http://www.dcode.org/echo). Not surprisingly, the scores of the known heart enhancers used in training were skewed toward the top-scoring CNEs (Fig. 3A). While the putative heart enhancers appear to be scattered across the human genome, they are slightly enriched near transcription start sites (TSSs) (6.0% vs. 5.5% in length-matched controls, *P*-value = 0.01). This is not unexpected since our classifier was trained on several promoter elements in addition to distant sequences. However, 95% of putative heart enhancers were found to reside distantly (at least 2 kb away) from the nearest TSS, suggesting that our classifier is not biased toward proximal elements. A negative score was reported for 688,000 CNEs, and these were used as controls for all further analyses.

Predicted heart enhancers are associated with genes expressed in the heart

Since our method predicts putative enhancers throughout the genome, it would be expected that these predicted enhancers map in the vicinity of genes expressed in the developing heart. To test this, we used gene expression microarray data from 79 human tissues (Su et al. 2002), including heart. The top 100 highly expressed genes, based on the microarray intensities for each tissue (normalized across the panel of these tissues), were used to compile 79 tissue-specific gene sets. We measured the fraction of high-scoring CNEs, i.e., putative heart enhancers, within the locus of each gene and computed the mean fraction within each gene set. On average, such a gene set should contain 10.1% high-scoring CNEs in its set of loci. Amongst all 79 tissues, the heart gene set contained the highest percentage of high-scoring CNEs (16.0%), resulting in more than a 1.5-fold enrichment over the expectation (Fig. 3B). This indicates that our method finds enhancers that are active specifically in the heart (Fig. 3C). The tissue with the second highest percentage of high-scoring CNEs was the skeletal muscle, which displayed a 1.3-fold enrichment over the expectation. A close examination of the skeletal muscle and the heart gene sets revealed that almost one-third of the genes are common in both. Additionally, several TFs are active in both these tissues during development (Edmondson et al. 1994; Schulz and Yutzey 2004). Interestingly, as we included less highly expressed genes to compile our gene sets, the enrichment of the heart gene set went down (Supplemental Fig. S1), further supporting the association of predicted elements with the level of gene expression in the heart. The significant association of our predictions with heart genes is notable considering that the classifier was not directly learned from any kind of gene expression data, but solely from sequence features within experimentally validated enhancers.

In vivo validation

The ultimate test for the ability of our predictor to accurately reveal heart enhancers is to experimentally demonstrate, in vivo, their enhancer properties. Toward that end, we selected 26 elements predicted to behave as putative heart enhancers and an additional set of 20 elements that were predicted not to behave as heart enhancers and tested them in zebrafish. The putative heart enhancers had an average score of 0.85, and were scattered across the full spectrum of predicted enhancers. Importantly, we tried to ensure that the putative enhancers tested included elements in loci of genes that were not highly expressed in the heart. In doing so we decreased the probability of hitting heart enhancers by chance just because they map in the vicinity of genes expressed in the heart. In choosing the 20 elements as putative negative controls for our in vivo validation, we did the opposite, deliberately picking sequences that flank genes with well-established expression in the heart, thus enriching our set of negative controls with elements with a higher likelihood of containing heart enhancer properties. All elements tested are evolutionarily conserved at least among mammals.

We used a zebrafish-based in vivo reporter assay (Kawakami et al. 2004) to test the enhancer properties of our selected set of elements. Zebrafish embryos are transparent, greatly facilitating the direct visualization of reporter gene expression in vivo throughout embryonic development, making it a preferred model system to test putative enhancer sequences of unknown spatial and temporal specificities.

Each element was PCR-amplified from the human genome and cloned in a eGFP reporter cassette, driven by a minimal c-fos promoter (Fisher et al. 2006a,b). This cassette contains two tol2 transposon sites, for the rapid and efficient integration of the transgene (Kawakami et al. 2004). Each construct was coinjected with tol2 transposase in 100–200 one-cell stage zebrafish embryos. Previous reports have demonstrated that the patterns observed in mosaic G_0 fish are reproduced in the germline transmission to G_1 (Fisher et al. 2006a); therefore, we evaluated enhancer properties in G_0 fish embryos.

Of the 26 predicted enhancers, 16 (62%) displayed reproducible and consistent expression in the heart (Table 1; Fig. 4; Supplemental Fig. S2). Of the set of 20 negative sequences, two (10%) displayed heart enhancer properties throughout zebrafish embryogenesis (Table 2). We assayed developing G_0 embryos daily, for 4–7 d, for GFP expression in the heart. For a construct to be classified as a heart enhancer we required a minimum of 20% of developing fish expressing GFP in the heart, comparable with rates previously reported (McGaughey et al. 2009). Though we observe variability in rates of expressing fish between constructs and variable patterns of expression among embryos transgenic for the same construct, there is generally a clear distinction between constructs that are classified as enhancers and ones that are not (Fig. 4B).

While the in vivo zebrafish enhancer assay is a valuable experimental model, it can be intrinsically limited in its ability to read out mammalian regulatory sequences. The large phylogenetic separation between fish and humans results in lineage-specific biological properties that may lead to both false-positive and false-negative results in our transgenic assay. To specifically address the concern of false-positive results in the zebrafish assay, we cross-validated, in transgenic mice, enhancers that we uncovered originally in zebrafish. For these experiments, we selected the four heart enhancers with the lowest score in our classifier, among the 16 enhancers that we discovered in fish (Table 1). This ensures that

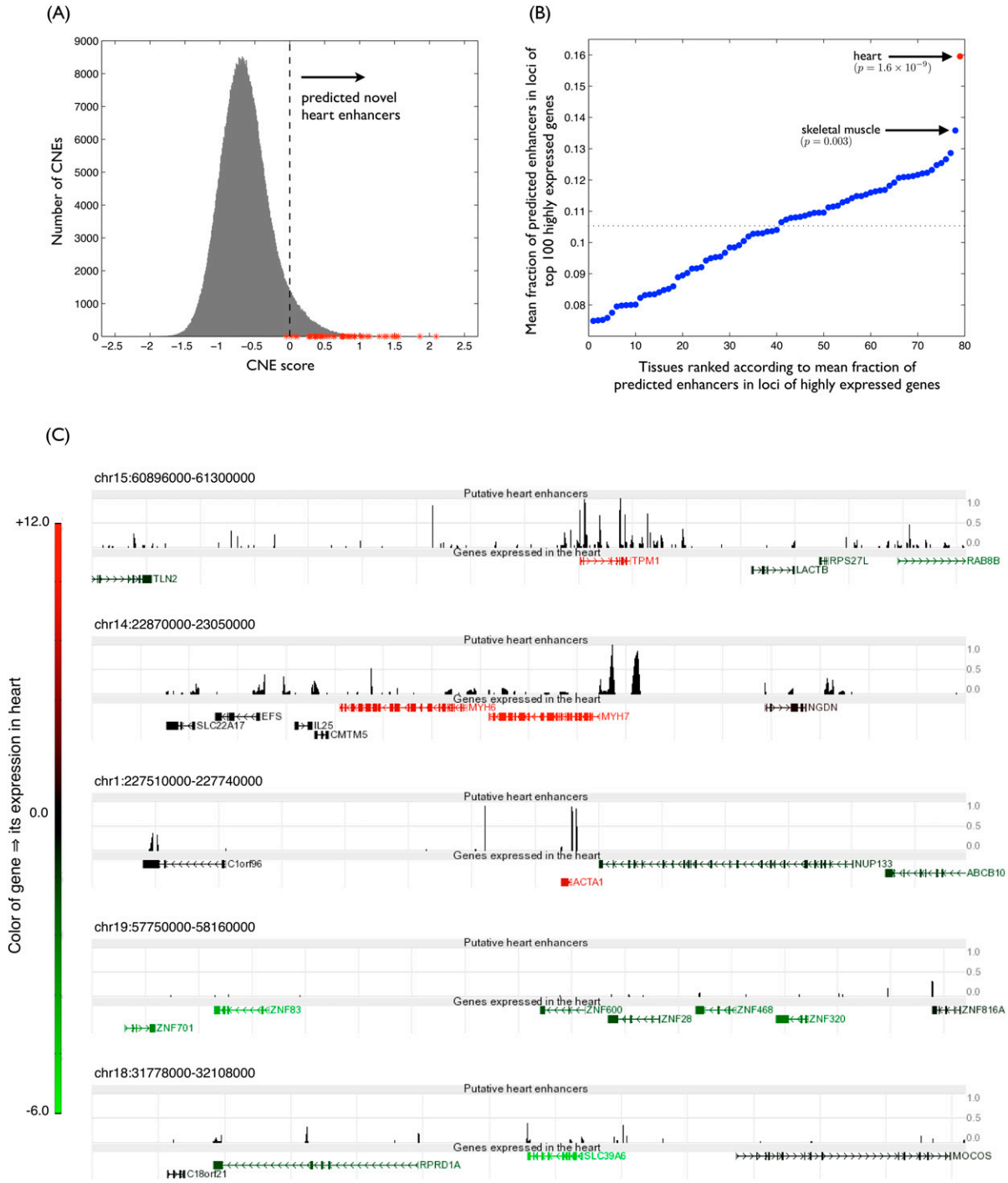


Figure 3. (*A*) Distribution of heart scores of CNEs. Scores assigned by the classifier for all tested CNEs are shown here. We use zero as a cutoff (Methods) for putative enhancers (dotted line). (Red) Scores of the training enhancer set. (*B*) Mean fraction of high-scoring CNEs in loci of genes highly expressed in each tissue. Tissues are sorted based on the mean fraction of putative heart enhancers in their loci. *P*-values were computed using a rank sum test; heart tissue had the most significant *P*-value of 1.6×10^{-9}. (*C*) (Black peaks) Snapshots of genome-wide view of predictions near genes. The score returned by the classifier is transformed to lie between 0 and 1, with numbers >0.5 indicating the occurrence of a putative heart enhancer. The color and shade of the gene transcript depict the type and level of gene expression, respectively: (red) genes highly expressed in the heart; (green) repressed genes. Genes highly expressed in the heart have typically more enhancers in their loci (*top three* genomic regions), while genes repressed or not expressed in the heart have fewer predictions in their loci. (All elements in the training set are excluded in these figures.)

we cross-validated sequences that were representative of average predicted heart enhancers by our classifier. Using the Gateway cloning system, we shuttled the four selected elements tested in fish into a *hsp68-lacZ* reporter cassette and injected them in fertilized mouse pro-nuclei. We generated two to six transgenic embryos for each construct, each representing an independent

Table 1. In vivo testing of *26* putative heart enhancer elements in zebrafish transgenics

No.	Human element tested (hg18)	Closest gene	Score	Heart expression of element
1	chr1:198308267–198308840	NR5A2	1.48	Positive
2	chr6:98988308–98988882	POU3F2	1.43	Negative
3	chr14-104259460–104259664	ADSSL1	1.35	Negative
4	chr12:88227744–88228315	DUSP6	1.31	Negative
5	chr20-31474451–31474982	SNTA1	1.15	Negative
6	chr6:99009085–99010799	POU3F2	1.13	Negative
7	chr15:65850899–65851174	MAP2K5	1.12	Positive
8	chr17:2064996–2065844	SMG6	1.10	Positive
9	chr4:85393274–85393635	NKX6-1	1.09	Positive
10	chr1:26922178–26923419	ARID1A	1.01	Positive
11	chr15:66780022–66780209	CORO2B	1.00	Positive
12	chr16:52908154–52909087	IRX3	0.94	Negative
13	chr9:99303863-99304144	TDRD7	0.92	Negative
14	chr13:108905001–108905276	MYO16	0.92	Positive
15	chr2:119239216–119239469	EN1	0.92	Negative
16	chr12:123569012–123569279	NCOR2	0.91	Negative
17	chr12:25430328–25430548	IFLTD1	0.90	Negative
18	chr1:50349222–50349402	ELAVL4	0.79	Positive
19	chr4:7835327–7835487	AFAP1	0.75	Positive
20	chr7:90761390–90763395	FZD1	0.59	Positive
21	chr7:69313052–69313811	AUTS2	0.56	Positive
22	chr19-50518697–50518939	CKM	0.31	Positive
23[a]	chr10-29208945–29209165	BAMBI	0.30	Positive
24[a]	chr8-30389248–30389688	RBPMS	0.13	Positive
25[a]	chr5-172109493–172109842	LOC54492-DUSP1	0.06	Positive
26[a]	chr15-32876189–32876660	ACTC1	0.03	Positive

Elements are ranked according to their enhancer predictive score in our classifier.
[a]Elements also tested in mouse transgenic embryos (shown in Fig. 4).

random genomic integration event. Embryos were assayed for *lacZ* activity at embryonic day (E)11.0–E12.5. All four constructs led to robust, reproducible reporter expression in mouse hearts (Fig. 4A), corroborating the results obtained in zebrafish. These results support the use of zebrafish as an experimental system to test mammalian sequences for putative cardiac enhancer function.

Generalization of the classifier to other cell types

The methodology of classifying tissue-specific enhancers described here is by no means restricted to the heart tissue. Indeed, apart from the training data set, there is no heart-related information used in the building of the classifier. To evaluate the utility of our method in other organisms and other cell types, we analyzed the recently published EP300 ChIP-seq data (Visel et al. 2009). This data set contains EP300-enriched DNA regions in three mouse cell types: limb, midbrain, and forebrain. Since EP300 is a transcriptional coactivator, we applied our method on these three tissue sets. For each tissue, we generated 2500 control sequences by sampling from the mouse noncoding regions, ensuring the samples had the same GC content and length distribution as the respective tissue set. The three independent classifiers demonstrate remarkable accuracy in a fivefold cross-validation: The areas under the ROC curve for limb, midbrain, and forebrain tissues sets are, respectively, 0.86, 0.82, and 0.92. These numbers are significantly better than those produced by other programs, which never cross the 0.75 mark in any tissue (Supplemental Fig. S3).

Discussion

In this study we developed a sequence classifier that distinguishes enhancers active in heart development and differentiation using a combination of binding information of TFs, overrepresented short sequence motifs, and Markov models. By using cross-validation we observed that the classifier is powerful in revealing the heart-related functionality of genomic DNA sequences while making few false-positive calls. The final features, selected in an automated manner from a large set of initial features, are also substantiated by literature evidence of TFs active in the heart. Additionally, the 42,000 predicted heart enhancers in the human genome are supported by an independent large-scale experiment that profiled gene expression. Most importantly, the validation of our predictions containing a set of long-range intronic and intergenic elements in a zebrafish in vivo reporter assay demonstrated a success rate of 62%. Furthermore, our tool was able to correctly rule out evolutionarily conserved sequences bracketing genes expressed in the developing heart as heart enhancers. Combined with a high accuracy, this will likely provide an important prioritization tool in determining which elements will be validated in costly, laborious functional assays.

The cardiac enhancers that our classifier predicts are constrained by the biased data set of cardiac enhancers that we used as a training set. It is likely that a number of heart enhancers in the human genome are not predicted by our classifier. Furthermore, a large fraction of the 42,000 heart enhancers predicted by our classifier likely does not drive gene expression exclusively in the heart, but also in other tissues and organs. For example, of the four cardiac enhancers that we validated in mice, three also gave rise to distinct spatial domains of expression outside the heart (Fig. 4A). This spatial heterogeneity reflects the modular architecture of enhancers, each of which comprises a collection of TFBSs that define multiple combinations of spatial specificities. What our classifier routinely identifies as a heart enhancer is a combination of TFBSs that together correspond to but a fraction of a DNA element that contains other TFBSs with distinct specificities. Future studies using training sets that correspond to highly specialized subsets of enhancers will likely uncover a different collection of predicted enhancers, with distinct spatial specificities in the heart.

Variants of linear and logistic regression have been used before to detect promoters and other regulatory elements. One of the early works that detected tissue-specific CRMs employed logistic regression to identify elements active in muscle (Wasserman and Fickett 1998) and liver (Krivan and Wasserman 2001). However, in that study, prior knowledge was required of TFs active in the tissue of interest as well as their binding specificities. A notable aspect of our approach is that our method is applicable even when no information regarding active TFs is known. In addition, since we also look for de novo motifs, we hope to learn motifs of TFs not yet characterized, but relevant to the tissue of interest. Although we have applied our method to heart development, the framework is applicable to any tissue where experimental data are available to train the model. Indeed, we showed that the learning approach

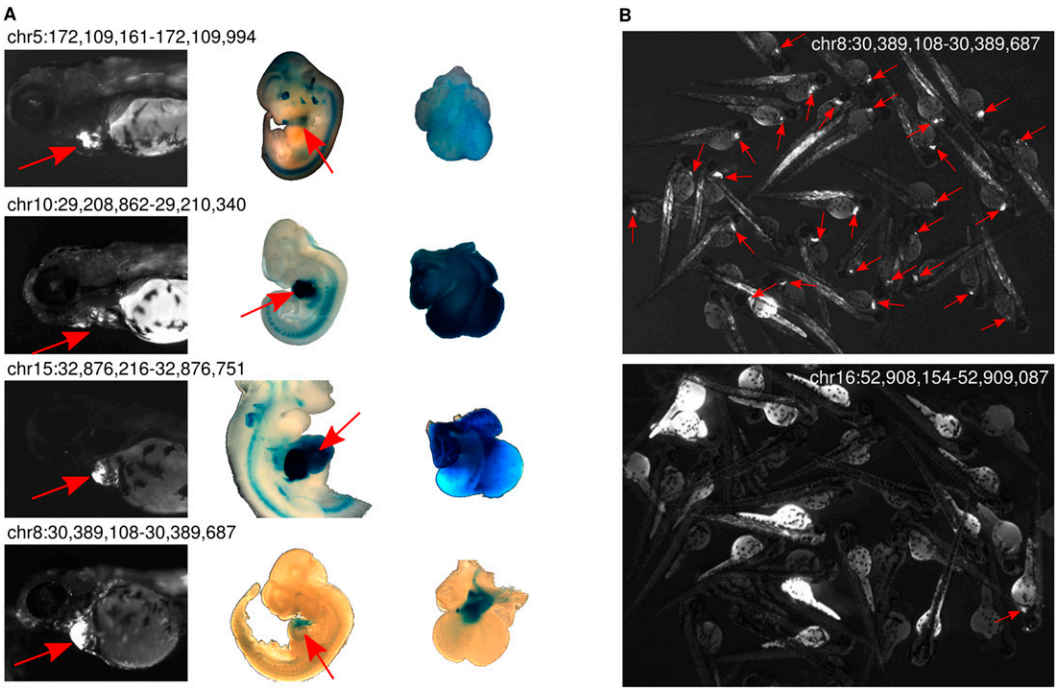


Figure 4. Experimental validation of predicted heart enhancers. (*A*) Four predicted heart enhancers driving expression of the reporter genes GFP in transgenic zebrafish (*first* column) and *lacZ* in transgenic mouse embryos (*second* column). (Red arrows) Expression of the reporter genes in the heart. (*Third* column) Reporter expression in dissected hearts for each of the constructs shown. Coordinates are hg18. (*B*) Positive and negative predicted heart enhancers identified in zebrafish transgenics. (*Top* image) Transgenic zebrafish displaying GFP expression in the heart driven by a predicted heart enhancer (positive). (*Bottom* image) Another predicted heart enhancer that did not drive reporter gene expression in the heart (negative).

can be successfully applied to ChIP-seq data to accurately classify enhancers active in limb, midbrain, and forebrain.

Our method is inherently generalizable to handle additional features during the learning process. For example, specific histone modifications have been shown to be enriched or depleted at enhancers (Heintzman et al. 2009). Such data, if available for heart tissues, may be added to the model to further improve the precision of the predictions. Other sources of information such as nucleosome occupancy and TF concentration can also be inserted as features when such data become available. Large-scale efforts to annotate the noncoding fraction of the human genome, such as the ENCODE project, rely on a multiplicity of corroborating lines of evidence to discern a biologically active sequence from an inactive one. Our study outlines a feature-based approach for the annotation of functional, tissue-specific, *cis*-regulatory sequences.

Table 2. In vivo testing of negatively scoring regions

No.	Human element tested (hg18)	Closest gene	Score	Heart expression of element
1	chr14:20554707–20555594	NRG2	−0.10	Negative
2	chr18:3210192–3210342	MYOM1	−0.11	Negative
3	chr1:234869454–234870011	HEATR1	−0.13	Positive
4	chrX:15243649–15243826	ASB11	−0.30	Negative
5	chr1:234915291–234915491	ACTN2	−0.32	Negative
6	chr19:50003308-50003625	BCAM	−0.51	Negative
7	chr15:83146046–83146997	ZNF592	−0.55	Negative
8	chr19:41334392–41334619	COX7A1	−0.62	Negative
9	chr11:19186361–19186843	CSRP3	−0.62	Negative
10	chr8:67243189–67243557	TRIM55	−0.68	Negative
11	chr1:204753459–204754124	RASSF5	−0.68	Negative
12	chr22:49354295–49354799	CHKB	−0.70	Negative
13	chr6:30956511–30956971	DDR1	−0.74	Negative
14	chr3:8762554–8762850	CAV3	−0.77	Negative
15	chr6:118975818–118976308	PLN	−0.78	Negative
16	chr1:158425305–158425516	CASQ1	−0.82	Negative
17	chr2:26767583–26768078	KCNK3	−0.85	Negative
18	chr15:88448268–88448421	IDH2	−0.88	Negative
19	chr17:34138517–34138739	MLLT6	−0.92	Positive
20	chr11:246852–247079	PSMD13	−1.00	Negative

All elements were found to be inactive in the heart.

Methods

Heart enhancers and control sequences

From a literature search, 30 human, mouse, and rat sequences shown to drive expression in the embryonic heart were compiled and their human orthologs were identified. Another 14 sequences that were listed as positive in the heart in the VISTA enhancer browser were added to the list (Visel et al. 2007). An additional 33 sequences that were shown to drive heart expression in our lab based on random scans of heart gene loci and predictions from our earlier method (Pennacchio et al. 2007) resulted in a total of 77 sequences referred to as the heart enhancer set. These sequences range in length from 120 to 1985 bp (average length 570 bp) and in GC content from 31% to 74%

(average GC content 50%). The control set was compiled by drawing 1000 sequences with a similar GC content as the enhancer set to avoid any GC-related bias (35% of our enhancers were from the promoter regions that are known to be GC-rich). We also ensured that the lengths of the control sequences were similar to the sequences in the enhancer set. This was done in the following manner: For each sequence in the enhancer set, a region of the same length was drawn from the noncoding regions of the human genome and selected if the GC content of that region was within a 2% difference from the GC content of the original enhancer sequence. This was repeated several times for each enhancer sequence until 1000 control sequences were generated. All analyses were done using the NCBI Build 36.1 assembly of the human genome.

Identification of CNEs

Human–mouse alignments generated by the ECR Browser (Ovcharenko et al. 2004) that are at least 150 bp long with >70% identity were compiled. These alignments were subsequently filtered out for overlapping coding exons of RefSeq genes, resulting in a data set of 729,781 CNEs. These CNEs were fragmented into windows of 150 bp with an overlap of 10 bp. Each CNE was assigned the score of its highest-scoring fragment. A total of 41,930 CNEs with a positive score was demarcated as putative heart enhancers.

Gene expression data

GNF Novartis Atlas2 tissue-specific gene expression (Su et al. 2002) was extracted from the gnfAtlas2 table from the UCSC genome browser and mapped to its respective RefSeq genes. This included expression profiles of 16,047 genes in at least one of 79 human tissues. 15,118 of these genes had at least one CNE in their loci, where a locus of a gene was defined as the noncoding region between the 3′ end of the gene immediately upstream of it and the 5′ end of the gene immediately downstream of it. For each gene, the fraction of high-scoring CNEs in its locus was computed. For each tissue, genes were sorted based on their normalized expression value. The mean fraction of high-scoring CNEs was computed for the top 100 highly expressed genes for each tissue and used to plot Figure 3B.

Classifier training

From each sequence in the enhancer set *E* and the control set *C*, three distinct sets of sequence features were extracted:

Set I. Number of matches to known vertebrate transcription factor binding specificities. Motifs of all vertebrate TFs characterized in TRANSFAC and JASPAR were compiled to get a total of 701 motifs. MAST (Bailey and Gribskov 1998) was used to compute the number of matches per base pair to each motif in the sequence, giving 701 features.

Set II. Number of matches to novel PWMs. PRIORITY (Narlikar et al. 2007), a motif discovery program, was used to learn the top 20 overrepresented motifs in the enhancer set. MAST was used to compute the number of matches per base pair to each novel motif in the sequence, giving 20 more features.

Set III. Log likelihoods from Markov models. Markov models of orders 0–5 from the enhancer set were learned based on its nucleotide frequencies. Similar models were also learned from the control set. The sequence under consideration was scored by the likelihood ratio of the two models for each order, giving six additional features.

The LASSO linear regression method (Tibshirani 1996) was used to find features relevant for distinguishing between the two classes E and C. This method models the class (+1 for E and −1 for C) of each sequence as a linear combination of features, and learns the optimal weights associated with each feature. It imposes a constraint on the absolute norm of the weights during the estimation, thereby producing a sparse solution in the feature space. This reduces the possibility of the model being overfitted to the training data, especially since it is expected that most features will be irrelevant for classification. The constraint bound on the weights was estimated using cross-validation within the training data.

A standard fivefold CV procedure was used to assess the accuracy of the classifier when applied to the original set of 77 enhancers, while a 10-fold CV procedure was used to assess the accuracy of the classifier when applied to the smaller homogenous enhancer set. Note that while the first set of features is independent of sequences in the training set, other features are not: Features based on de novo motifs are learned from the enhancer set, while Markov features are learned from both the enhancer and the control set. In each fold of the CV procedure, the motifs and the Markov models were computed based only on the training data, thereby ensuring that the test data were completely unseen before the predictions were made.

The source code of the classifier, the list of CNEs and heart enhancer predictions, along with ECR Browser and UCSC Genome Browser tracks are available at http://www.dcode.org/echo.

Comparison with other methods

Four freely available state-of-the-art methods that detect CRMs, CisModule (Zhou and Wong 2004), Cluster-Buster (Frith et al. 2003), MSCAN (Alkema et al. 2004), and Stubb (Sinha et al. 2006), were used for comparison. Each program was run with its default settings. For each CV fold, the free parameters of CisModule, Cluster-Buster, and Stubb were fitted using the training set. Since MSCAN uses its precomputed parameter values, it was run as is on the test sets in each fold.

EP300 data

Genomic regions enriched for EP300 in mouse forebrain, midbrain, and limb tissues were extracted from Supplemental Tables 2–4 of Visel et al. (2009). Sequences unique to each tissue were retained resulting in a total of 2052 forebrain sequences, 278 midbrain sequences, and 1949 limb sequences. A control set for each tissue was compiled by randomly sampling 2500 sequences from noncoding, nonrepetitive regions as in the case of the heart sequences. A fivefold cross-validation was performed for each of the three classifiers; the same folds were also assessed using the four state-of-the-art CRM detectors mentioned earlier.

In vivo validation

Zebrafish were raised and bred in accordance with standard conditions (Kimmel et al. 1995; Whitlock and Westerfield 2000). Embryos were obtained from natural crosses, incubated at 28.5°C, and staged (Warga and Kimmel 1990). Sequences of interest were amplified with specific attB-containing primers and cloned into a donor vector (pDONR 221) of the Gateway cloning system (Invitrogen). Plasmid DNAs for microinjection were purified on QIAprep Mini-prep (Qiagen) spin columns. Transposase RNA was transcribed in vitro using the mMessage mMachine Sp6 kit (Ambion). Injection solutions were made with 25 ng/mL transposase RNA and 15–25 ng/mL circular plasmid in water. DNA was injected into the yolk of wild-type embryos at the two-cell stage. At least 100 embryos were injected for each element.

The DNA fragments tested in mouse were cloned in to a *lacZ* reporter construct and injected in mouse fertilized pro-nuclei

nuclei as previously described (Pennacchio et al. 2006). Six independent mouse transgenic lines harboring this construct were obtained. Embryos were examined at E11.0 and E12.5 for heart enhancer properties.

Acknowledgments

We thank Ryan Vo for helping us compile the set of known heart enhancers, Andrew McCallion for providing the tol2-EGFP plasmid, and Robert Ho for generously sharing his zebrafish facility and injection bays with us. This research was supported by the Intramural Research Program of the NIH, National Library of Medicine (I.O.), and by grants HL088393 and HG004428 (M.A.N).

References

Aitola M, Carlsson P, Mahlapuu M, Enerback S, Pelto-Huikko M. 2000. Forkhead transcription factor FoxF2 is expressed in mesodermal tissues involved in epithelio-mesenchymal interactions. *Dev Dyn* **218:** 136–149.

Alkema WB, Johansson O, Lagergren J, Wasserman WW. 2004. MSCAN: Identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res* **32:** W195–W198.

Andres V, Cervera M, Mahdavi V. 1995. Determination of the consensus binding site for MEF2 expressed in muscle and brain reveals tissue-specific sequence constraints. *J Biol Chem* **270:** 23246–23249.

Bailey TL, Gribskov M. 1998. Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics* **14:** 48–54.

Bi W, Drake CJ, Schwarz JJ. 1999. The transcription factor MEF2C-null mouse exhibits complex vascular malformations and reduced cardiac expression of angiopoietin 1 and VEGF. *Dev Biol* **211:** 255–267.

Bruneau BG. 2008. The developmental genetics of congenital heart disease. *Nature* **451:** 943–948.

DeCaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, Galagan JE. 2007. Conrad: Gene prediction using conditional random fields. *Genome Res* **17:** 1389–1398.

de la Calle-Mustienes E, Feijoo CG, Manzanares M, Tena JJ, Rodriguez-Seguel E, Letizia A, Allende ML, Gomez-Skarmeta JL. 2005. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res* **15:** 1061–1072.

de Souza FS, Santangelo AM, Bumaschny V, Avale ME, Smart JL, Low MJ, Rubinstein M. 2005. Identification of neuronal enhancers of the proopiomelanocortin gene by transgenic mouse analysis and phylogenetic footprinting. *Mol Cell Biol* **25:** 3076–3086.

De Val S, Chi NC, Meadows SM, Minovitsky S, Anderson JP, Harris IS, Ehlers ML, Agarwal P, Visel A, Xu SM, et al. 2008. Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. *Cell* **135:** 1053–1064.

Edmondson DG, Lyons GE, Martin JF, Olson EN. 1994. *Mef2* gene expression marks the cardiac and skeletal muscle lineages during mouse embryogenesis. *Development* **120:** 1251–1263.

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006a. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312:** 276–279.

Fisher S, Grice EA, Vinton RM, Bessling SL, Urasaki A, Kawakami K, McCallion AS. 2006b. Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. *Nat Protoc* **1:** 1297–1305.

Flesch M. 2001. On the trail of cardiac specific transcription factors. *Cardiovasc Res* **50:** 3–6.

Frith MC, Li MC, Weng Z. 2003. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res* **31:** 3666–3668.

Ghosh D, Chinnaiyan AM. 2005. Classification and selection of biomarkers in genomic data using LASSO. *J Biomed Biotechnol* **2005:** 147–154.

Gratzinger D, Zhao S, West R, Rouse RV, Vogel H, Gil EC, Levy R, Lossos IS, Natkunam Y. 2009. The transcription factor LMO2 is a robust marker of vascular endothelium and vascular neoplasms and selected other entities. *Am J Clin Pathol* **131:** 264–278.

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

Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459:** 108–112.

Iida K, Hidaka K, Takeuchi M, Nakayama M, Yutani C, Mukai T, Morisaki T. 1999. Expression of MEF2 genes during human cardiac development. *Tohoku J Exp Med* **187:** 15–23.

Kawakami K, Takeda H, Kawakami N, Kobayashi M, Matsuda N, Mishina M. 2004. A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. *Dev Cell* **7:** 133–144.

Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. 1995. Stages of embryonic development of the zebrafish. *Dev Dyn* **203:** 253–310.

Krivan W, Wasserman WW. 2001. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res* **11:** 1559–1566.

Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. 2007. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* **39:** 1235–1244.

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12:** 1725–1735.

Lifanov AP, Makeev VJ, Nazina AG, Papatsenko DA. 2003. Homotypic regulatory clusters in *Drosophila*. *Genome Res* **13:** 579–588.

Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, Axel R. 2006. Interchromosomal interactions and olfactory receptor choice. *Cell* **126:** 403–413.

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288:** 136–140.

Mahony S, Benos PV. 2007. STAMP: A web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* **35:** W253–W258.

Maston GA, Evans SK, Green MR. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7:** 29–59.

McGaughey DM, Stine ZE, Huynh JL, Vinton RM, McCallion AS. 2009. Asymmetrical distribution of non-conserved regulatory sequences at PHOX2B is reflected at the ENCODE loci and illuminates a possible genome-wide trend. *BMC Genomics* **10:** 8. doi: 10.1186/1471-2164-10-8.

McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. 2007. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* **4:** 63–72.

Miano JM, Ramanan N, Georger MA, de Mesy Bentley KL, Emerson RL, Balza RO Jr, Xiao Q , Weiler H, Ginty DD, Misra RP. 2004. Restricted inactivation of serum response factor to the cardiovascular system. *Proc Natl Acad Sci* **101:** 17132–17137.

Narlikar L, Gordan R, Hartemink AJ. 2007. A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput Biol* **3:** e215. doi: 10.1371/journal.pcbi.0030215.

Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. 2003. Scanning human gene deserts for long-range enhancers. *Science* **302:** 413.

Ovcharenko I, Nobrega MA, Loots GG, Stubbs L. 2004. ECR Browser: A tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res* **32:** W280–W286.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444:** 499–502.

Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I. 2007. Predicting tissue-specific enhancers in the human genome. *Genome Res* **17:** 201–211.

Pham VN, Lawson ND, Mugford JW, Dye L, Castranova D, Lo B, Weinstein BM. 2007. Combinatorial function of ETS transcription factors in the developing vasculature. *Dev Biol* **303:** 772–783.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4:** 651–657.

Salzberg SL. 1997. A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput Appl Biosci* **13:** 365–376.

Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. 2004. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* **32:** D91–D94.

Schulz RA, Yutzey KE. 2004. Calcineurin signaling and NFAT activation in cardiovascular and skeletal muscle development. *Dev Biol* **266:** 1–16.

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. 2008. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451:** 535–540.

Sepulveda JL, Vlahopoulos S, Iyer D, Belaguli N, Schwartz RJ. 2002. Combinatorial expression of GATA4, Nkx2-5, and serum response factor directs early cardiac gene activity. *J Biol Chem* **277:** 25775–25782.

Sharan R, Ben-Hur A, Loots GG, Ovcharenko I. 2004. CREME: *Cis*-Regulatory Module Explorer for the human genome. *Nucleic Acids Res* **32:** W253–W256.

Sherry B. 2002. The role of interferon regulatory factors in the cardiac response to viral infection. *Viral Immunol* **15:** 17–28.

Sinha S, Fuchs E. 2001. Identification and dissection of an enhancer controlling epithelial gene expression in skin. *Proc Natl Acad Sci* **98:** 2455–2460.
Sinha S, Liang Y, Siggia E. 2006. Stubb: A program for discovery and analysis of *cis*-regulatory modules. *Nucleic Acids Res* **34:** W555–W559.
Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci* **99:** 4465–4470.
Sun Q, Chen G, Streb JW, Long X, Yang Y, Stoeckert CJ Jr, Miano JM. 2006. Defining the mammalian CArGome. *Genome Res* **16:** 197–207.
Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y. 2001. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17:** 1113–1122.
Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE. 2004. Decoding human regulatory circuits. *Genome Res* **14:** 1967–1974.
Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B Methodol* **58:** 267–288.
Tschopp P, Tarchini B, Spitz F, Zakany J, Duboule D. 2009. Uncoupling time and space in the collinear regulation of Hox genes. *PLoS Genet* **5:** e1000398. doi: 10.1371/journal.pgen.1000398.
Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res* **35:** D88–D92.
Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, 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.
Waleev T, Shtokalo D, Konovalova T, Voss N, Cheremushkin E, Stegmaier P, Kel-Margoulis O, Wingender E, Kel A. 2006. Composite Module Analyst: Identification of transcription factor binding site combinations using genetic algorithm. *Nucleic Acids Res* **34:** W541–W545.
Warga RM, Kimmel CB. 1990. Cell movements during epiboly and gastrulation in zebrafish. *Development* **108:** 569–580.
Wasserman WW, Fickett JW. 1998. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* **278:** 167–181.
Whitlock KE, Westerfield M. 2000. The olfactory placodes of the zebrafish form by convergence of cellular fields at the edge of the neural plate. *Development* **127:** 3645–3653.
Wingender E, Chen X, Fricke E, Geffers R, Hehl R, Liebich I, Krull M, Matys V, Michael H, Ohnhauser R, et al. 2001. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res* **29:** 281–283.
Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, Vavouri T, Smith SF, North P, Callaway H, Kelly K, et al. 2005. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* **3:** e7. doi: 10.1371/journal.pbio.0030007.
Yokoro K, Yanagidani A, Obata T, Yamamoto S, Numoto M. 1998. Genomic cloning and characterization of the mouse POZ/zinc-finger protein ZF5. *Biochem Biophys Res Commun* **246:** 668–674.
Zhang MQ, Marr TG. 1993. A weight array method for splicing signal analysis. *Comput Appl Biosci* **9:** 499–509.
Zhou Q, Wong WH. 2004. CisModule: De novo discovery of *cis*-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci* **101:** 12114–12119.
Zhu H, Traver D, Davidson AJ, Dibiase A, Thisse C, Thisse B, Nimer S, Zon LI. 2005. Regulation of the *lmo2* promoter during hematopoietic and vascular development in zebrafish. *Dev Biol* **281:** 256–269.

Received July 19, 2009; accepted in revised form January 8, 2010.