1726–1739 *Nucleic Acids Research, 2015, Vol. 43, No. 3* *Published online 21 January 2015*
doi: 10.1093/nar/gkv011

Enhancer modeling uncovers transcriptional signatures of individual cardiac cell states in *Drosophila*

Brian W. Busser[1,*,†], Julian Haimovich[1,†], Di Huang[2], Ivan Ovcharenko[2,*] and Alan M. Michelson[1,*]

[1]Laboratory of Developmental Systems Biology, Genetics and Developmental Biology Center, Division of Intramural Research, National Heart Lung and Blood Institute, National Institutes of Health, Bethesda, MD 20892, USA and [2]Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20892, USA

Received November 3, 2014; Revised January 5, 2015; Accepted January 7, 2015

ABSTRACT

Here we used discriminative training methods to uncover the chromatin, transcription factor (TF) binding and sequence features of enhancers underlying gene expression in individual cardiac cells. We used machine learning with TF motifs and ChIP data for a core set of cardiogenic TFs and histone modifications to classify *Drosophila* cell-type-specific cardiac enhancer activity. We show that the classifier models can be used to predict cardiac cell subtype *cis*-regulatory activities. Associating the predicted enhancers with an expression atlas of cardiac genes further uncovered clusters of genes with transcription and function limited to individual cardiac cell subtypes. Further, the cell-specific enhancer models revealed chromatin, TF binding and sequence features that distinguish enhancer activities in distinct subsets of heart cells. Collectively, our results show that computational modeling combined with empirical testing provides a powerful platform to uncover the enhancers, TF motifs and gene expression profiles which characterize individual cardiac cell fates.

INTRODUCTION

An understanding of how individual cells acquire their unique fates and differentiate during organogenesis requires uncovering both their distinct gene expression profiles and the transcriptional enhancers that orchestrate the coordinated transcription of these co-expressed genes. Enhancers are non-coding stretches of DNA that respond to the combinatorial input of multiple classes of sequence-specific DNA binding transcription factor (TFs) to precisely control gene expression in the appropriate cells and at the correct time (1,2). Gene expression profiling, computational searches for TF binding motifs, histone mark modifications and *in vivo* TF binding have all been used to uncover transcriptional enhancers and gene expression patterns, as well as to build developmental transcriptional regulatory networks (3,4). Such studies often focus on whole organism or tissue-level analyses. However, to accurately describe the specification of individual cell fates requires knowledge of gene expression patterns and the regulatory mechanisms responsible for the activities of associated cell-type-specific enhancers.

The invertebrate heart is similar to the mammalian heart at the linear tube stage of differentiation, and many of the genetic mechanisms regulating the specification and differentiation of heart cells in *Drosophila* have similar functions in mammals (5). In fact, a core cardiac transcriptional network, composed of the inductive signaling pathways (Wnt, Fgf and Bmp) as well as TFs including Tinman (Tin or NKx2.5 in mammals), Pannier (Pnr or Gata4 in mammals), Tailup (or Isl-1 in mammals), Hand (or Hand-1/-2 in mammals) and the Dorsocross-1/-2/-3 genes (or Tbx5 in mammals) and H15/midline (or Tbx20 in mammals) is required for cardiogenesis in vertebrate and invertebrate species (5,6).

The cells comprising the *Drosophila* heart can be subdivided into two broad populations, the cardial cells (CCs) which express muscle genes and are contractile and the pericardial cells (PCs) whose functions are not as well described but are believed to act as nephrocytes (5). These cell types arise from segmentally-repeated clusters of cells in a portion of the early differentiating mesoderm called the cardiac mesoderm (CM). These cells ultimately arrange to form a

*To whom correspondence should be addressed. Tel: +1 301 451 8041; Fax: +1 301 496 9985; Email: michelsonam@nhlbi.nih.gov
Correspondence may also be addressed to Brian W. Busser. Tel: +1 301 496 3470; Fax: +1 301 480 5865; Email: busserbw@nhlbi.nih.gov
Correspondence may also be addressed to Ivan Ovcharenko. Tel: +1 301 435 8944; Fax: +1 301 480 2288; Email: ovcharen@nih.gov
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

linear tube of 6 CCs per hemi-segment that together form the lumen of the heart which, in turn, is surrounded by an outer layer of 10 PCs in each hemi-segment.

CCs and PCs can be further subdivided into individual identities based on differences in morphology, localization and gene expression patterns (5). The expression of several TFs have been shown to discriminate cardiac cell fates, with the NK homeodomain TF Tin expressed in the four posterior-most CCs, the COUP-TF Seven-up (Svp) expressed in the two anterior-most CCs, while the NK homeodomain TFs Ladybird early and Ladybird late (redundant genes hereafter referred to as Ladybird, Lb) are expressed in the two anterior-most Tin-expressing CCs. In addition, the 10 PCs per hemisegment can be subdivided into five discrete cell populations based on their localization and the expression pattern of Svp, Lb, Tin and the homeodomain TF Even-skipped (Eve). Thus, given that each cell is characterized by a discrete phenotypic identity, combined with the tractability of analyzing the relevant cell types in informative genetic backgrounds, the *Drosophila* heart provides a facile system to interrogate how individual cardiac cells acquire their unique identities.

Here we used an integrated machine learning approach to uncover the sequence motifs and epigenetic features that characterize the enhancers regulating gene expression in individual cardiac cells in *Drosophila* (Figure 1). To do so, we first built a gene expression atlas for hundreds of genes that are expressed in the heart. We next undertook a large-scale validation of previously characterized *Drosophila* cell-type-specific heart enhancers, which revealed enhancer activities restricted to distinct subpopulations of cardiac cells. These enhancers were used in a machine learning approach that integrated TF motifs with ChIP data for both TF binding and histone modifications, thereby uncovering both sequence and protein features which are predicted to discriminate specific cardiac cell identities and to reveal cell-specific enhancer activities. We validated these computational predictions using a large-scale analysis of predicted enhancers and sequence features in transgenic reporter assays. Finally, clustering the predicted enhancers from the individual cardiac classifiers associated with known cardiac genes uncovered previously uncharacterized functions of individual cardiac cells. In total, these results document the utility of computational modeling combined with empirical testing to uncover the enhancers, motifs and genes that characterize individual cardiac cell fates.

MATERIALS AND METHODS

Analysis of transgenic reporter constructs and embryo staining

Enhancer regions were either PCR-amplified or synthesized *in vitro* (Integrated DNA Technologies, Coralville, IA, USA), sequence-verified and then subcloned into the reporter vector pWattB-nlacZ (7–10). All constructs were targeted to attP40 (11) with phiC31-mediated integration (12), and homozygous viable insertion lines were obtained (Genetic Services, Inc., Sudbury, MA, USA). Whole-embryo immunohistochemistry and *in situ* hybridization followed standard protocols (8–10,13). The following antibodies were used: mouse anti-βgal (1:500, Promega, Madison, WI, USA), guinea pig anti-Zfh1 (1:1000, gift of J. Skeath) and rabbit anti-Tin (1:800, gift of M. Frasch).

Gene ontology (GO) analysis

Over-represented gene ontology (GO) categories were defined with FuncAssociate2.0 using standard parameters (14), and removal of the redundant GO terms from this list was performed with REVIGO using standard parameters (15).

Classifier training

The machine learning approach employed here was previously described (7,9,10,16). For each training sequence, 1000 unique controls were randomly sampled from the non-coding sequence of the *Drosophila melanogaster* (*dm3*) genome that were matched to their respective enhancer in GC-content and base pair length. Binding motifs were mapped by scanning the sequences using tfSearch from http://rvista.dcode.org/ (17), and the source code can be downloaded from http://www.dcode.org/tfSearch/tfSearch.tgz, with appropriate position weight matrices compiled from the TRANSFAC, JASPAR and UNIPROBE databases (18–20). Motif presence was normalized by generating a per base pair measurement of motif occurrences in each sequence by defining the number of motif occurrences divided by sequence length of a training or control sequence. The coordinates of genomic regions considered significantly bound by cardiogenic TFs and enriched for a particular histone modification were defined and mapped onto the enhancer and controls (21,22). The enhancer or control sequence was considered to contain the ChIP peak if at least 20% of the peak was present in that genomic region, and this was given a binary score of presence or absence of a given ChIP peak. The coordinates of genomic regions defined as bound by cardiogenic TFs was previously published using TileMap (22). The coordinates of genomic regions defined as enriched for a particular histone modification were identified using MACS (tsize = 36, shift size = 90, bw = 100, gsize = 135 000 000) by comparing to input sequence, and were required to be identified in the replicate ChIP-seq experiments (with at least 100 bp overlap) (21). For ease of reproducing the results, the training and control sequences as well as the motif IDs have been deposited here: http://dx.doi.org/10.6084/m9.figshare.1284097. We used a linear-kernel support vector machine (SVM) approach to discriminate training set sequences from controls. Classifier accuracy was validated using a 5-fold cross-validation strategy (20% of the training data was held out for testing). These SVM classifiers were used to predict similar regulatory activity across the *D. melanogaster* genome by using a sliding window of length 1000 bp and overlaps of 200 bp. Candidate regions were annotated to known heart genes by identifying its nearest neighbor along the genome. Hierarchical clustering was performed using the MATLAB clustergram function and normalized by using the built-in standardization function within the clustergram tool which transforms the standardized values so that the mean is 0 and the standard deviation is 1 (23).

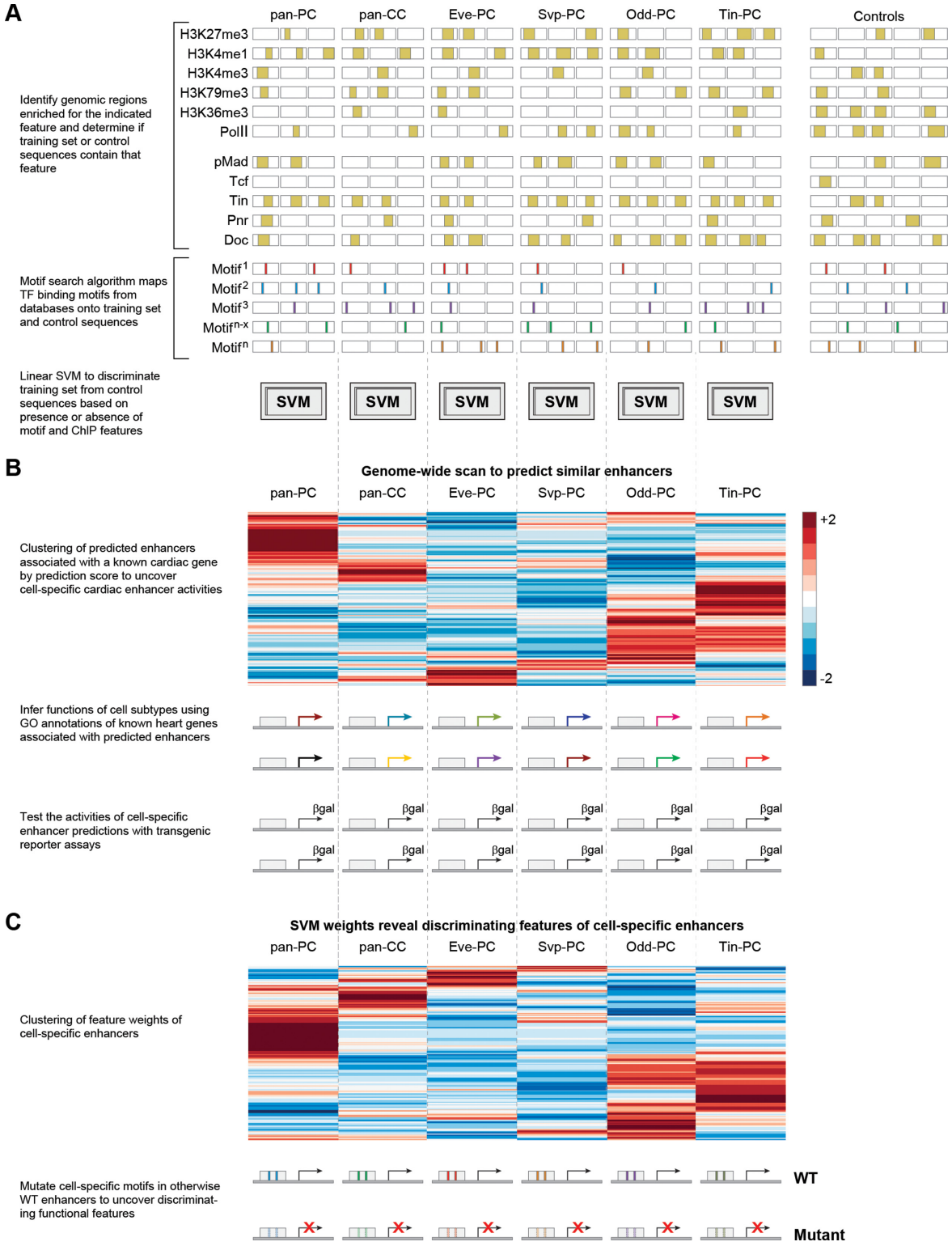

Figure 1. Schematic of the computational and experimental approaches used in this study. (**A**) Classifying cardiac cell subtype enhancer activity with histone marks, *invivo* TF binding and TF binding motifs. The presence of the indicated histone marks and *invivo* TF binding on different enhancer sequences is shown by yellow blocks, while the presence of TF binding motifs are represented by lines of different colors. SVMs are built to discriminate enhancer sequences for each cellular subtype from a set of control sequences. (**B**) Genome-wide scan to predict similar cell type-specific enhancers. Hierarchical clustering of the top-scoring predicted enhancers for each cell subtype classification associated with each known heart gene. A GO analysis is used on the associated genes within each cluster to infer functions of cardiac cell subtypes. Transgenic reporter assays are used to validate enhancer predictions. (**C**) Feature weights reveal discriminating features of cell type-specific enhancers. Hierarchical clustering of SVM feature weights reveals features that are enriched amongst one cardiac cell subtype and depleted or irrelevant to the others. The role of a particular sequence motif in discriminating cardiac cell subtype enhancer activities is tested via *cis* mutagenesis in transgenic reporter assays.

RESULTS AND DISCUSSION

A gene expression atlas of cardiac genes

We previously designed and applied a meta-analysis of gene expression profiles derived from purified mesodermal cells obtained from wild-type (WT) and informative mutants to characterize and predict gene activity in the *Drosophila* heart (24). In addition, recent studies have used chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) of numerous cardiac TFs to uncover the *cis* regulatory elements and genes which characterize the cardiac lineage (22,25). In order to compile a more comprehensive list of genes with confirmed expression in the *Drosophila* heart, we performed a large-scale validation of these predictions using whole embryo *in situ* hybridization (Supplementary Figure S1A). Out of 103 tested genes, we uncovered an additional 50 genes with previously uncharacterized expression in the CM and/or mature heart. Representative gene expression patterns are shown in Supplementary Figure S1A with a complete annotation of cardiac gene expression patterns documented in Supplementary Table S1. Combining these newly-identified cardiac genes with a complete curation of the literature reveals a total of 284 genes with verified expression in the heart (Supplementary Table S1).

We next used GO analysis (14) followed by the generation of a condensed summary of the list that was initially obtained by removing redundant GO terms (15). The purpose of this analysis was to uncover the functions of this large battery of cardiac genes. Indeed, the non-redundant GO terms revealed a diversity of functions for these genes, identifying both upstream (signaling, transcription, etc.) and downstream (adhesion, chemotaxis, metabolic processes, etc.) components of the heart gene regulatory network (Supplementary Figure S1B). In fact, a more detailed categorization revealed that 165 of these 284 genes are upstream components, with 82 of these being sequence-specific TFs (Supplementary Table S1). As there are presently only eight described cardiac cell subtypes (five PC and three CC), this shows that there are at least 10× the number of TFs than previously characterized cell states, suggesting that there is more extensive diversity in the combinations of TFs utilized to achieve specificity of cardiac gene expression than had been appreciated in prior studies. The diversity of TFs required to achieve cellular specificity of gene expression seems to be mirrored in the enhancers they regulate, as we found similar diversity in the combinations of motifs regulating and TFs binding myogenic enhancers (9,10). In total, this work uncovers a large battery of cardiac genes, and both the diversity of their inferred functions and the large number of TFs identified suggest that these genes are under complex combinatorial transcriptional regulation.

Defining large training sets of cell-type-specific cardiac enhancers

We next investigated the molecular mechanisms underlying the coordinate regulation of these heart genes. We previously characterized the motifs, enhancers and TFs that discriminate the two broad populations of the *Drosophila* heart, PCs and CCs (7). Here we sought to model enhancers with cardiac activity of individual cardiac cell states to gain insights into both the similarities and differences in sequence and chromatin features amongst the eight individual cardiac cell subtypes that are known to exist (see Figure 2A). To do so, we compiled a list of enhancers with previously reported activity in the *Drosophila* heart including those from our preceding study (7) and performed transgenic reporter assays to confirm and refine prior findings at the level of single cells of defined identities (Supplementary Table S2). To avoid the confounding effects of reporter variability due to insertion site, these reporters were inserted at a specific genomic locus that permits robust and reproducible activity in the mesoderm (7–10,13). We performed *in vivo* transgenic reporter assays with the 95 curated cardiac enhancer sequences and confirmed that 73 are active in the CM and/or heart, with the majority of the enhancer sequences with non-cardiac activity showing activity in the neighboring amnioserosa cells (Supplementary Table S2).

We next monitored the activity of these 73 cardiac reporters in the differentiated heart to compile training sets of enhancers with activity in the different cardiac cell subtypes. As the cells of the heart can be subdivided into individual identities based on morphological differences and the expression pattern of distinct TFs (Figure 2A), we used the expression of Tin, which marks a subset of CCs and PCs, and Zinc Finger Homeobox 1 (Zfh1) which labels all PCs, with anatomical and morphological differences of the cells to identify every distinct cardiac cell type (5). Representative results are shown in Figure 2B with a complete annotation of enhancer activities in Supplementary Table S2. Using these markers and monitoring reporter activity in the differentiated heart, we uncovered a set of enhancers with activity in all the PCs (22 total sequences, hereafter referred to as 'pan-PC') and/or all the CCs (33 total sequences, hereafter referred to as 'pan-CC'). Of these 73 cardiac reporters, we identified 6 to 7 enhancers with activity restricted to the subsets of the CCs (hereafter referred to as 'Tin-CC', 'Tin-Lb-CC' or 'Svp-CC') which is an insufficient quantity to serve as a training set for a machine learning analysis without over-fitting the data. We also identified many enhancer sequences with activity in the different PC subtypes of the heart, including the Svp-PCs, Odd-PCs and Eve-PCs (Supplementary Table S2). However, we were unable to individualize the activity of enhancer sequences in the Tin-alone or Tin-Lb-PCs, with only one enhancer sequence (that associated with the Lb genes) with activity restricted to the Tin-Lb PCs but not the Tin-alone PCs. As enhancer sequences are active in both of these cell types, we refer to this class as the 'Tin-PC' enhancers. In total, these results identified sets of enhancers with activities in different subsets of cardiac cells, including pan-PC, pan-CC, Eve-PC, Tin-PC, Odd-PC and Svp-PC.

Integrative modeling the enhancer activities of individual cardiac cells

Here we used a machine learning approach to uncover associated regulatory elements and the discriminating characteristics (sequence motifs and epigenetic features) that differentiate these individual heart cells. Previous work has shown that the distribution of epigenetic modifications of

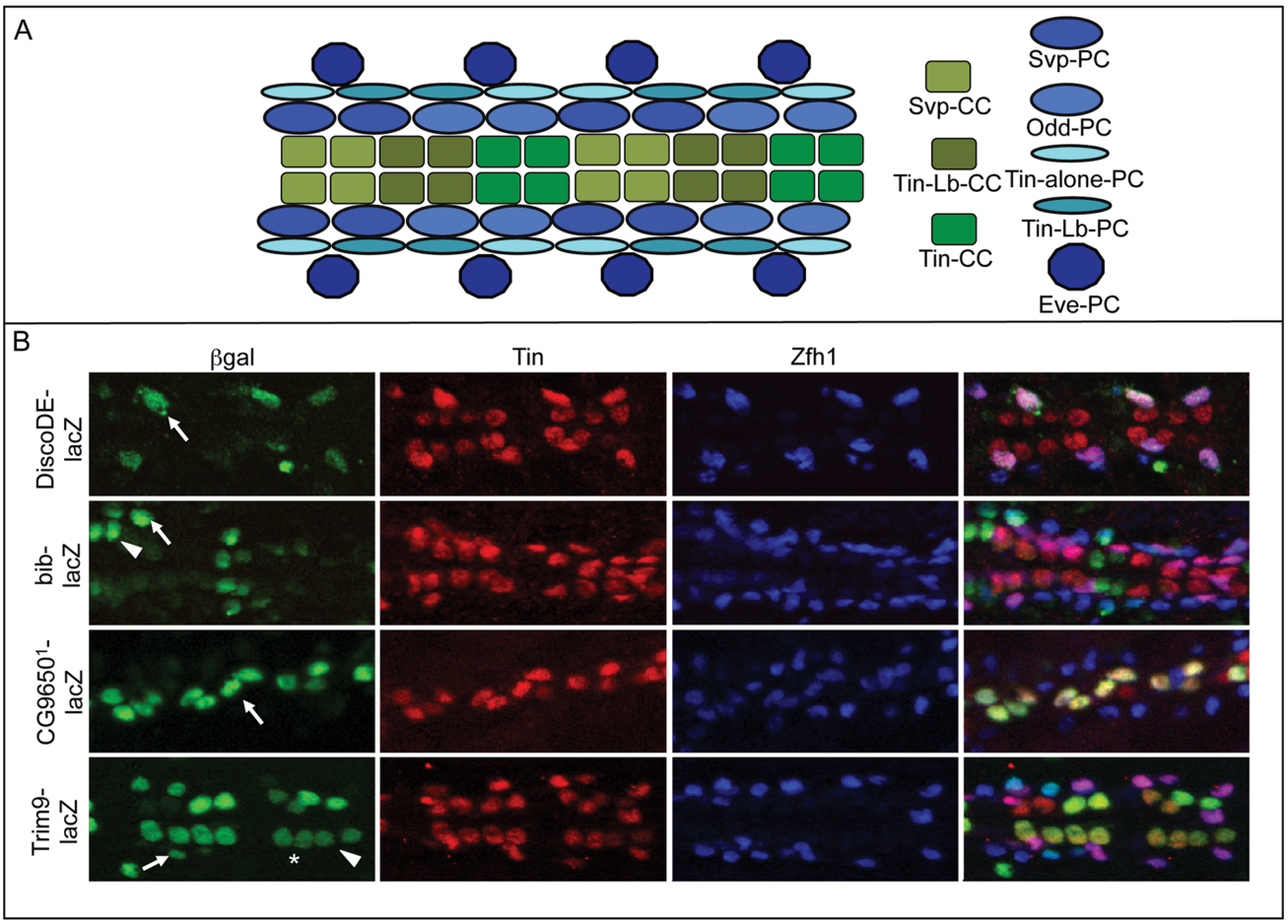

Figure 2. Individual cardiac cell states in the *Drosophila* heart. (**A**) Schematic of the previously identified distinct cardiac cell subtypes, as modified from Ward and Skeath (31). (**B**) Transgenic reporter assays reveal enhancer activities in distinct cardiac cell subtypes. Stage 16 embryos stained with βgal (green), Tin (red) and Zfh1 (blue). Arrow in DiscoDE-lacZ denotes βgal activity in the Eve-PCs. Arrow in bib-lacZ denotes activity in Svp-PC and arrowhead denotes Svp-CC. Arrow in CG9650[1]-lacZ denotes activity in Tin-alone-PCs and Tin-Lb-PCs. Arrow in Trim9-lacZ denotes activity in Odd-PC, arrowhead denotes Tin-CC and * denotes Tin-Lb-CC.

the histone proteins and *in vivo* binding profiles of relevant TFs can be used to predict *cis* regulatory elements and gene activity (3,4,7,9,10,26–29). Furlong *et al*. have recently described the distribution of a series of histone modifications in sorted mesodermal nuclei from *Drosophila* embryos at a developmental stage in which the cardiac precursor cells are being specified (21). In addition, Junion *et al*. (22) examined the *in vivo* binding sites of a series of conserved cardiogenic TFs at different developmental time points. These include the T-box TFs (Doc), the GATA4 ortholog Pnr, the Nkx2.5 ortholog Tin and the TFs downstream of the signaling pathways for Wnt (dTCF in *Drosophila*) and Bmp (phosphorylated Mad (pMad) in *Drosophila*). In addition to the aforementioned TFs and histone marks, we also included over 1000 binding motifs from available databases to identify sequence features critical for categorizing enhancer activities (7,9,10,16). The binding motifs and *in vivo* binding profiles for cardiogenic TFs and relevant histone modifications were mapped onto the training set and control sequences (see 'Materials and Methods' section) and a SVM was used to discriminate the training set from controls. To model cell-type-specific cardiac enhancer activity, we built separate SVM models for pan-PC, pan-CC, Eve-PC, Tin-PC, Svp-PC and Odd-PC sequences.

We initially attempted to classify the different cell subtypes against each other. However, this approach failed to discriminate the training set sequences from controls as the area under the receiver operator characteristic (AUC) curve values ranged from 0.46 to 0.67. This result is due to the overlap in the training set sequences, with most sequences showing activity in more than one cell type (Supplementary Table S2), which reflects a requirement for the gene products regulated by these enhancers in more than one cell type. To circumvent this issue, we built separate SVM models for training set sequences from GC and length-matched background sequence (see 'Materials and Methods' section for details). Here we observed reliable classification of cardiac cell subtype enhancers as the AUC curve varied for the separate classifiers from 0.96 to 0.99 (Figure 3A). In addition, enhancers predicted by these models are significantly associated with known heart genes (Figure 3B and C and Supplementary Table S3). Finally, we show that the enhancer predictions of cardiac cell classifications are cell-type-specific (Supplementary Figure S2). In total, these re-

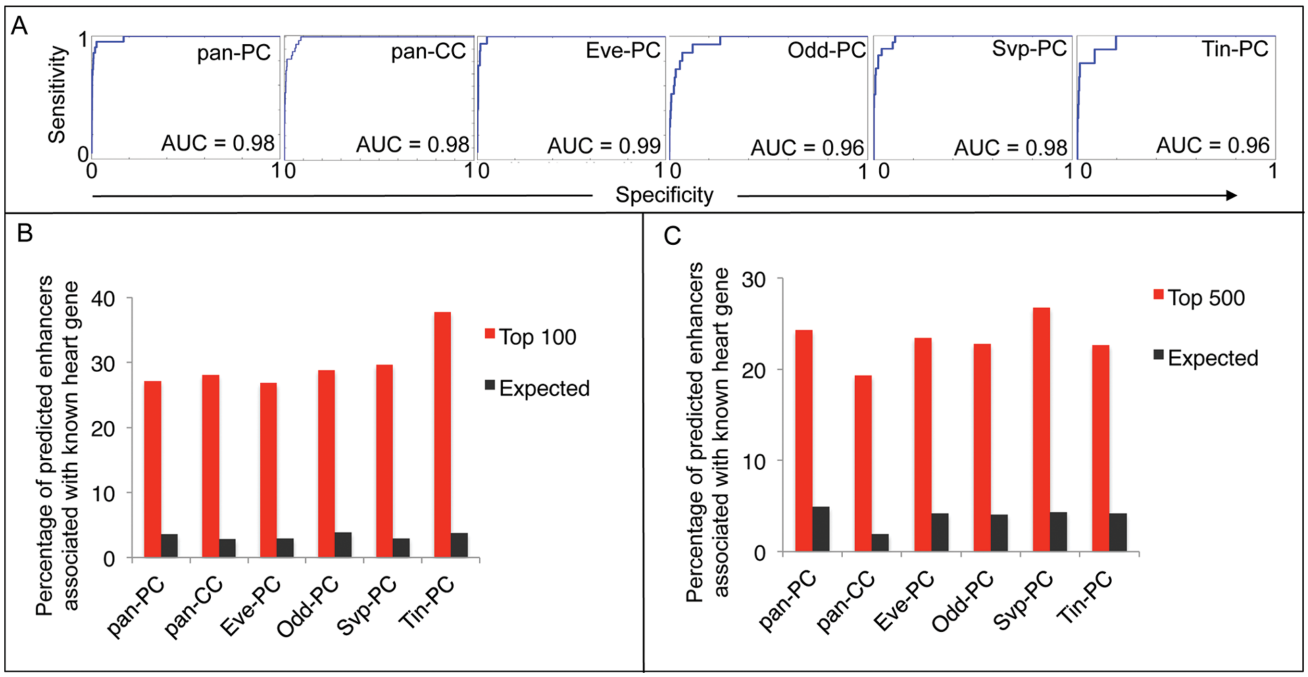

Figure 3. Cardiac cell subtype classifiers perform with high specificity and sensitivity. (**A**) ROC curves for classifying enhancers active in the indicated cardiac cell subtypes. Area under the ROC curve (AUC) is shown. Percentage of predicted enhancers in the neighborhood of known heart genes for the indicated classifications as determined for the top 100 (**B**) and top 500 (**C**) top-scoring enhancer predictions. Intronic sequences were associated with their host whereas intergenic sequences were associated with the closest gene. All comparisons to expected are significant as computed using the binomial test with the following *P*-values: pan-PC (top 100 = 3.17E-12 and top 500 = 1.3E-21), pan-CC (top 100 = 3.00E-9 and top 500 = 2.87E-12), Eve-PC (top 100 = 1.75E-08 and top 500 = 4.14E-21), Odd-PC (top 100 = 6.26E-10, top 500 = 1.98E-22), Svp-PC (top100 = 4.73E-10 and top 500 = 8.01E-26) and Tin-PC (top 100 = 6.04E-13 and top 500 = 4.24E-18).

sults confirm the generation of cardiac cell subtype-specific cardiac classifiers that can reliably discriminate the training set from controls.

Classifying cell-type-specific enhancer activities predicts cell-type-specific enhancer activity and gene function

We next asked if we could use the enhancer predictions from the individual cell-specific classifications to predict expression patterns of known cardiac genes, and to use these annotated gene expression patterns to uncover the functions of individual heart cells. To do so, we isolated the top-scoring cardiac cell subtype enhancer prediction from each classification for each gene with known heart expression. By focusing our analysis on genes with validated cardiac expression, we were able to confidently associate a predicted enhancer with *bona fide* transcriptional targets, findings that are not always available or included in such studies, often due to the lack of known expression patterns for candidate target genes. Underscoring the utility of this approach, 278 out of 284 heart genes (97.9%) were associated with a top-scoring predicted cell-specific cardiac enhancer (Supplementary Table S3). Out of these 278 heart genes associated with a predicted enhancer, 196 of these predictions were found within the introns of the heart gene (70.5%), increasing the confidence in its association with this transcriptional target. We used hierarchical clustering of the prediction scores to group related expression patterns, which uncovered distinct clusters of cell-specific cardiac gene expression (Figure 4A). This analysis revealed gene expression clusters specific for the individual cardiac cell subtypes and also for the pan-PC, pan-CC and all cardiac cell expression patterns.

With these expression clusters, we asked if we could infer functions associated with these individual cardiac cell subtypes. GO analysis for the genes within these expression clusters, followed by the removal of redundant terms, revealed functions for these gene expression clusters (Figure 4B). Genes associated with enhancers predicted to be active in all heart cells (pan-PC/pan-CC) were associated with developmental, signaling and transcriptional functions. This result is consistent with these genes playing a role in the upstream regulatory network that specifies the cardiac lineage. Furthermore, genes with predicted expression in all CCs (pan-CC) were enriched for myogenic functions including cell adhesion and the actin cytoskeleton which are expected functions for contractile cells. Interestingly, genes associated with pan-PC enhancers were associated with renal system development, which further supports their proposed role as insect nephrocytes (30).

This analysis also uncovered specialized functions for individual cardiac cell subtypes. For example, the Odd-PCs were enriched for chemotaxis and locomotion functions, suggesting these cells are responsive to migratory cues. Alternatively, in the anterior segments of the embryo, Odd is expressed in the PCs of the neighboring lymph gland which forms the adult blood cells and it is this population of cells which are responsive to migratory cues (31). Interestingly, the genes associated with enhancers with predicted activity in Tin-PCs are associated with development of endocrine

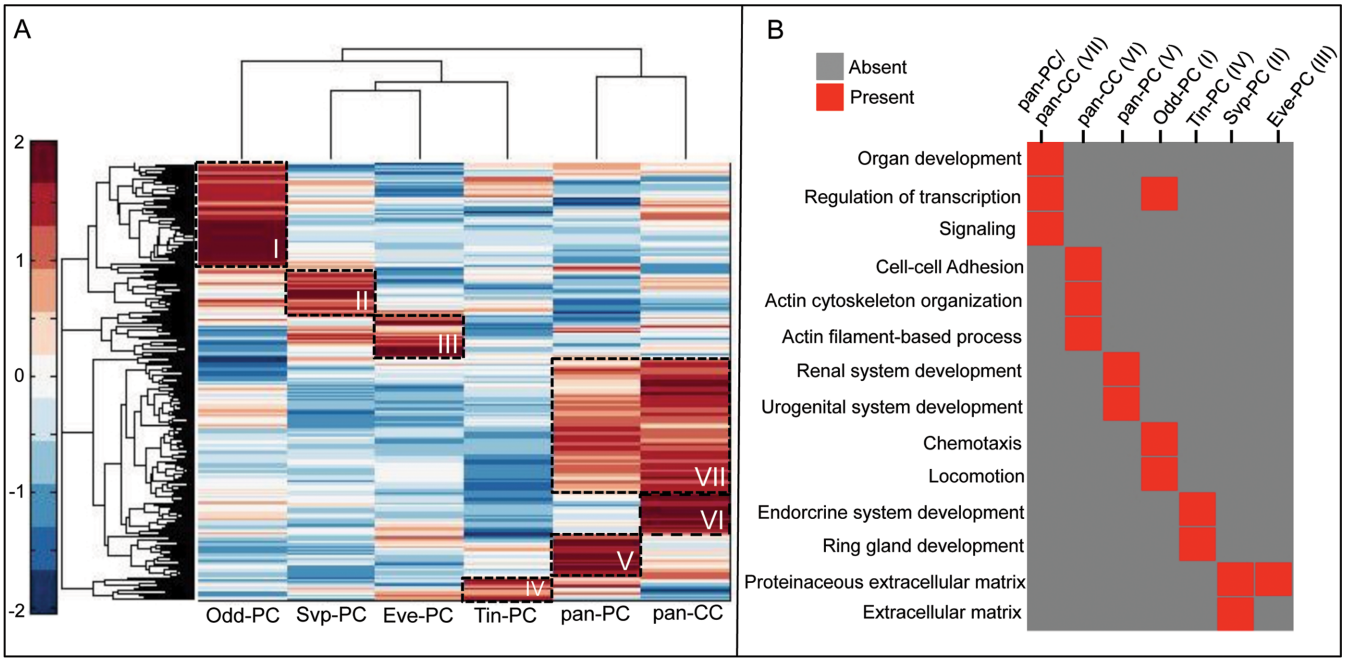

Figure 4. Classifying cell-type-specific enhancer activities predicts gene expression and function of individual cardiac cell subtypes (**A**) Hierarchical clustering of the top-scoring classifier-predicted enhancers associated with genes with known expression in the heart. Predictions scores are normalized by row maximum to better visualize differences. Each row represents a predicted enhancer associated with a known heart gene with the color of that row represents the prediction score. Clusters represent enhancers with top-scoring predictions for one classification versus the others, which are predicted to activate the neighboring heart gene in the indicated cell type(s). (**B**) Matrix shows the presence or absence of enriched GO categories associated with the heart genes from the indicated clusters. Genes from pan-PC/pan-CC (expression cluster VII), pan-CC (VI), pan-PC (V), Odd-PC (I), Tin-PC (IV), Svp-PC (II) and Eve-PC (III). Gene ontology analysis followed by REVIGO (to remove redundant GO terms) was used to reveal enriched GO terms (14,15).

functions (the ring gland in *Drosophila* is an endocrine organ). Since the physiological processes of filtration, secretion and reabsorption must be coordinated, this specialized endocrine role for Tin-PCs suggests these cells may act as a cellular relay mechanism between these components of the insect excretory system (30). Lastly, genes associated with enhancers with predicted activity in Eve-PCs and Svp-PCs specialize in the production of extracellular matrix components which is an essential aspect of proper filtration of the haemolymph (*Drosophila* blood) (30). In total, these results confirm that modeling cell-type-specific enhancer activities can be used to both confirm and identify previously uncharacterized functions of individual cardiac cells.

Functional assessment of enhancers predicted by the classifier

To test the *in vivo* transcriptional activities of the predicted enhancers, we used transgenic reporter assays inserted at specific genomic loci to test 47 enhancer predictions of varying scores in the cell-specific classifications (Supplementary Table S3). These results revealed that 46 of these 47 candidate enhancers were active reporters in the *Drosophila* embryo, with 19 of these 46 active reporters (41.3%) showing activity in the differentiated heart (Supplementary Table S3 and Figure 4). Analyses of cell-type-specific reporter activity uncovered a concordance between predicted and confirmed activity (Figure 5 and Supplementary Figure S3). For example, a predicted enhancer located within the first intron of CG5522 scores well in the pan-PC and pan-CC classifications and poorly in the classifications of individual cardiac cell subtypes (Figure 5A). Transgenic reporter assays confirm this result as this genomic region activates reporter expression in all PCs and CCs of the differentiated heart (Figure 5A). We have also used the distribution of prediction scores to reveal enhancers that are active in individual cardiac cells. For example, another enhancer prediction located within the first intron of the Dscam gene scores very well in the Eve-PC and Odd-PC classifications (Figure 5C). In agreement with these cell-specific predictions, we show that this enhancer prediction is active in these two cell types with additional activity in the Svp-PCs, thereby confirming the significant but slightly less robust Svp-PC prediction score (Figure 5C). Some successful enhancer predictions scored well in a cellular subtype classification as well as in the pan-PC and pan-CC classifications. It is possible that such regulatory elements may be composed of overlapping enhancer signatures, with one DNA segment regulating pan-PC and pan-CC activity while another DNA segment enhances transcription in a different cellular subtype. The transgenic reporter assays used to assay enhancer activity would be insensitive to detecting such minor differences in reporter activity due to *in vivo* perdurance of the reporter RNA and/or protein. In agreement with this possibility, we have previously uncovered multiple signatures in the enhancers regulating muscle founder cell gene expression (10). Taken together, these results show that the distribution of prediction scores for individual cardiac cell classifications can be used to predict enhancer activity in individual cardiac cell subtypes.

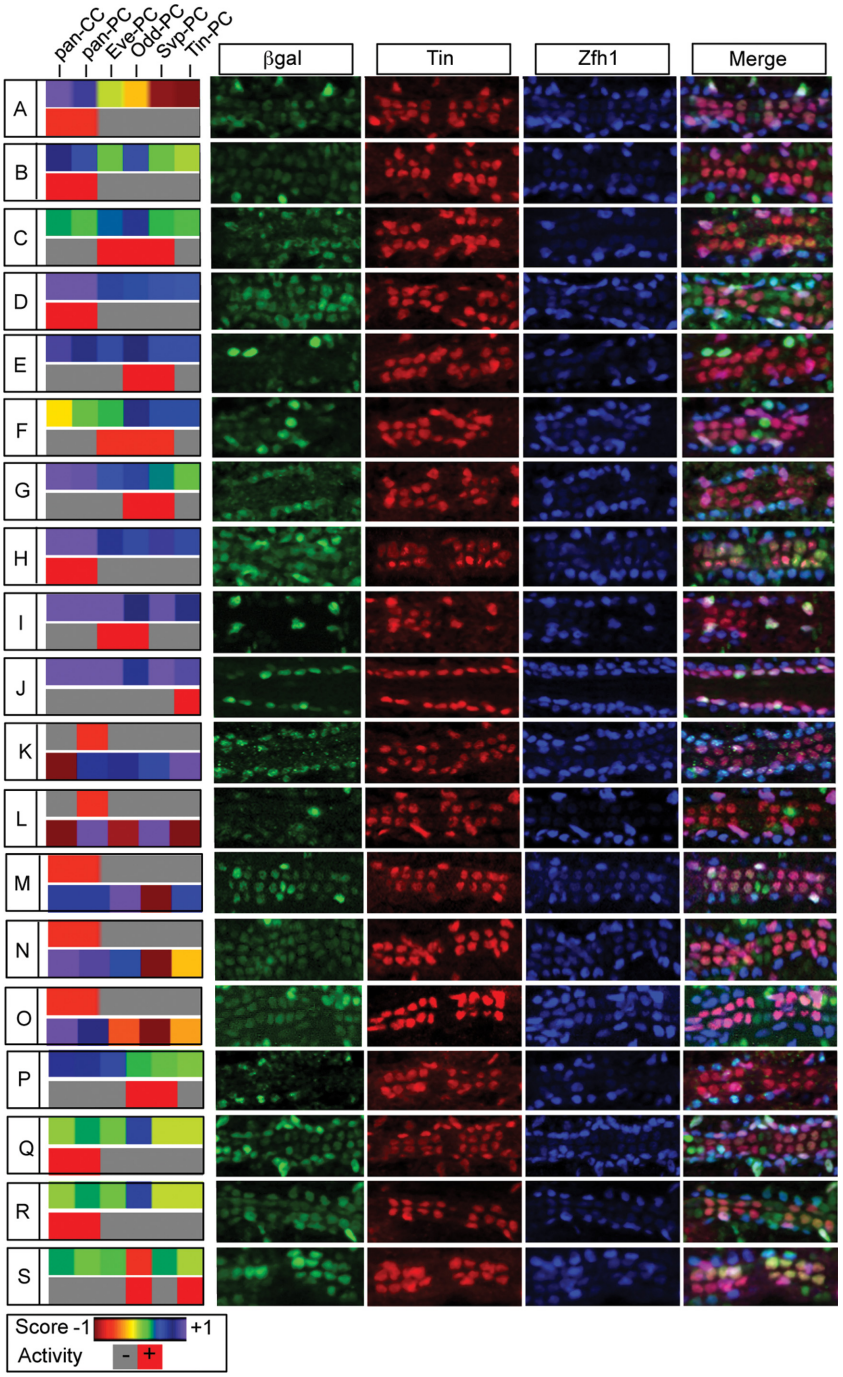

Figure 5. Validation of cardiac cell subtype-predicted enhancers. Transgenic reporter assays for predicted cardiac enhancers are shown. Prediction scores for the tested genomic region and validated enhancer activity are also indicated. For activity, red signifies βgal reporter expression in the indicated cell type(s) and gray reveals an absence of βgal expression. Transgenic reporter assays for genomic regions within the intron of CG5522 (**A**), CG8949 (**B**), Dscam (**C**), Ten-m (**D**), Akt1 (**E**), p130CAS (**F**), Otk (**G**), Tisl1 (**H**), jing (**I**), CG6234 (**J**), Cbx (**L**), phyl (**M**), Trx (**N**), S (**O**), Tsh (**Q**), Nhe2 (**R**) and CG8965 (**S**) as well as genomic regions spanning the transcriptional start sites of Aur (**K**) and CG15283 (**P**) and the intergenic sequence between tsh and CG11629 (L) are shown. A complete annotation of tested genomic coordinates and enhancer activities is shown in Supplementary Table S3. Stage 16 embryos were stained for βgal, Tin and Zfh1 to monitor expression in all cardiac cell subtypes.

TF features associated with cardiac cell subtype classification

To gain an understanding of the regulatory network required for specifying individual cardiac cell fates, we next assessed the sequence, TF binding and chromatin features critical for the classification of each subtype of heart cell included in our analyses (Figure 6). As features in the training set receive positive weights, those in the control set receive negative weights, and irrelevant features receive zero weight in linear SVMs, we examined the classification weights associated with the histone marks, TF binding and sequence features relevant to the previously delineated cell-specific regulatory models.

The *in vivo* binding of cardiogenic TFs was next examined as a feature at two developmental time points: (i) 4–6 h after egg laying, a time point in which the dorsal mesodermal derivatives—which includes the precursors of the CM—are specified; and (ii) 6–8 h after egg laying, a time point during which the more differentiated CM is specified (22). Tin, the Nkx2.5 ortholog in *Drosophila*, is first expressed in and required to specify the dorsal mesodermal derivatives, its expression and function then become restricted to the CM and later there is a confinement of Tin to subsets of cells comprising the mature heart (5). Pnr (the Gata4 ortholog in *Drosophila*) and Doc (Tbx4 ortholog in *Drosophila*) expression intersect with Tin in the CM, and both of these TFs are required for the differentiation of most cardiac cells (5). Finally, the overlap of signaling by Wnt (whose downstream effector in *Drosophila* is dTCF) and Bmp (whose downstream effector in *Drosophila* is phosphorylated Mad, pMad) is critical for specification of the CM (5).

Amongst the TFs examined, the greatest enrichment is seen with Tin at 6–8 h, which is consistent with the central role played by Tin in the cardiac transcriptional network in *Drosophila* (5) (Figure 6A). However, this interpretation should be considered with caution as the majority of heart enhancers in the training sets were identified based on the presence of Tin binding sites or *in vivo* binding. The larger positive classification weight at 6–8 h than at 4–6 h for Tin supports a more critical role for Tin binding to cardiac enhancers when the CM is specified.

Surprisingly, since Pnr has previously been shown to be a key regulator of cardiogenesis, the SVM weights reveal a minor role for the GATA TF Pnr binding in regulating cardiac enhancer activity. However, this finding is consistent with a recent report which failed to identify cardiac enhancers due to Pnr binding (22) and suggests either a non-enhancer role for such binding or an inability to accurately assess such enhancers with the transgenic reporter assays used in these studies. For example, as minimal promoters are used in transgenic reporter assays, this result could reflect a requirement for a certain promoter *in vivo* for enhancer activity driven by Pnr-dependent enhancers (32).

We note positive classification weights associated with pMad, Tcf and Doc amongst the different cell types. Interestingly, we found that differential SVM weights are associated with these TFs in the various cardiac subtype classifications. For example, Doc shows the greatest positive weight for the Eve-PC classification, and every newly-identified enhancer with Eve-PC activity is bound by Doc (Figure 5 and Supplementary Table S3). Furthermore, pMad demonstrates a greater SVM weight amongst the classifications of individual cardiac cell subtypes than amongst the pan-PC or pan-CC classifications. This outcome suggests that differential utilization of this signaling pathway may play a role in specifying individual cardiac cell fates. As 7 out of 11 pan-PC enhancers (63.6%) and 6 out of 8 individual cardiac cell subtype enhancers (75%) of newly-identified cardiac enhancers are bound by pMad, validation of this hypothesis requires further testing (Figure 5 and Supplementary Table S3). In conclusion, these data show that differential SVM weights of *in vivo* TF binding can be used to model cell-specific enhancer activities.

Chromatin features associated with cardiac cell subtype classification

As numerous studies have shown that the epigenetic modifications of the histone proteins can be used as predictors of *cis* regulatory element activity (3,4,21,26,33–35), we next examined the SVM weights for multiple histone mark modifications for each cardiac cell subtype classification identified in our analyses (Figure 6B). These histone modifications were examined at the 6–8 h developmental time point (a time at which the cardiac precursors are specified) from sorted mesodermal nuclei (21). Surprisingly, the strongest enrichment of any modification is tri-methylation of lysine 27 on histone 3 (H3K27me3) for all cardiac cell subtypes. We previously showed an enrichment of H3K27me3 on active mesodermal enhancers (8) which was in disagreement with a previous study that revealed a depletion of H3K27me3 on active mesodermal enhancers (21). As the polycomb complex which is associated with silent chromatin primarily trimethylates lysine 27 on histone 3 (36), the most likely explanation for these data is that they reflect the overall enhancer activity in a heterogenous rather than pure population of cells. Since the cells of the *Drosophila* heart only correspond to a tiny population of the entire mesoderm, and whole mesoderm was previously studied, the apparently inconsistent observation noted here suggests that the enhancer is repressed in the majority of the cells (non-heart mesodermal cells) and is active in the minority of cells examined (the fraction of the mesoderm which comprises the heart and its precursors). In agreement with this interpretation, the SVM weights for H3K27me3 are greater for the cardiac subpopulations than those with activity in all PCs or CCs in which a larger population of total cells would show signs of repression (Figure 6B). Furthermore, the enrichment for acetylation of lysine 27 on histone 3 (H3K27ac) on these same enhancers suggests that they are active in a subset of cells (Figure 6B). These results argue that an accurate interrogation of the epigenetic signatures of individual genomic loci requires isolating homogenous subpopulations of cells. This point is especially relevant when describing bivalent chromatin signatures which may reflect the presence of either a bivalent locus in a single cell or different epigenetic modifications in some but not all members of a more diverse cell population (37).

We also show that monomethylation of lysine 4 on histone 3 (H3K4me1) is positively weighted amongst all classifications, consistent with its description as an enhancer mark. In contrast, trimethylation of lysine 4 on histone 3

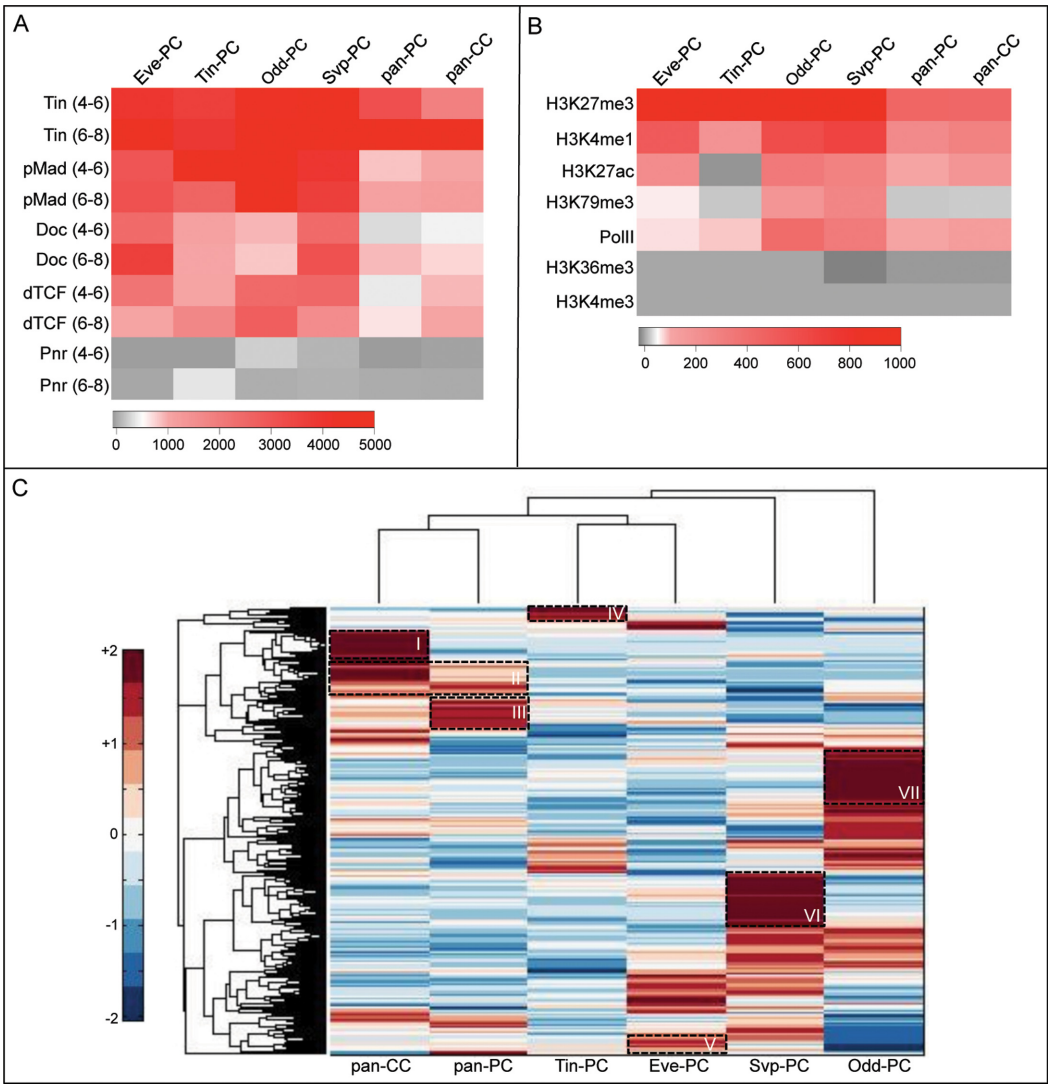

Figure 6. Chromatin and sequence features can be used as features to discriminate enhancer activities in cardiac cell subtypes. Matrix showing the distribution of SVM weights for TF ChIP signal (**A**) and histone marks (**B**) uncovered in the classification of individual cardiac cell subtypes. Positive weights indicate features enriched amongst the training set, negative weights are enriched amongst the controls and a zero weight indicates no enrichment amongst either set in linear SVMs. (**C**) Hierarchical clustering of SVM weights for the top 500 scoring motifs associated with each individual cardiac cell subtype classification. Each row represents a motif and the color of that row represents the SVM weight. Differential SVM weights for the individual classifications reveal clusters of sequence motifs enriched amongst pan-CC (I), pan-CC+pan-PC (II) pan-PC (III), Tin-PC (IV), Eve-PC (V), Svp-PC (VI) and Odd-PC (VII) versus the other classifications. SVM weights are normalized so that the standardized values have mean = 0 and standard deviation = 2 to better visualize differences.

(H3K4me3) and trimethylation of lysine 36 on histone 3 (H3K36me3) received either no weight or negative weights for all classifications, consistent with their description as marks of promoters and gene bodies, respectively (21,38). Surprisingly, the SVM weight for the active enhancer mark H3K27ac received no weight among Tin-PC enhancers (Figure 6B), which may be due to the fact that H3K27ac was seen to only mark two out of nine training set sequences (Supplementary Figure S4). This suggests that H3K27ac may not always associate with active enhancers in certain cell types. However, this interpretation should be regarded with caution as the training set was small for these cell types and two out of two newly-identified Tin-PC enhancers were marked by H3K27ac (Supplementary Table S3). Trimethylation of lysine 79 on histone 3 (H3K79me3) was positively associated with each cardiac cell subtype classification, a result that is in agreement with a recent study which observed H3K79me3 on a subset of developmental enhancers (21). Interestingly, H3K79me3 showed greater SVM weights associated with Svp-PC and Odd-PC classifications than with the other models, suggesting that these modifications may be differentially utilized amongst cardiac cell subtypes. A large-scale validation of enhancer activities will be required to test this hypothesis, although six out of seven (85.7%) newly-discovered enhancers with activity in Svp-PCs and/or Odd-PCs are marked by H3K79me3 while 7 out of 11 (63.6%) with pan-PC activity are marked by H3K79me3 (Figure 4 and Supplementary Table S3). In any event, such differential utilization of histone marks amongst cell types and regulatory elements may explain the incom-

plete association between a particular mark and a class of regulatory element (21,33–35). Furthermore, such a cell- or tissue-specific role for histone modifications likely explains the tissue-specific effects of loss-of-function mutations in histone-modifying enzymes (39). In total, these results uncover chromatin features that are enriched and that potentially discriminate among cardiac cell subtypes.

Sequence features associated with cardiac cell subtype classification

In order to identify DNA sequence similarities and differences amongst the cardiac cell subtype classifications, we examined the top 500 scoring sequence motifs amongst all classifications and used hierarchical clustering of their SVM weights to reveal clusters of similarly-acting regulatory motifs (Figure 6C). Similar to the clustering of enhancer activities, this analysis revealed motif clusters enriched amongst each cardiac cell subtype classification and depleted or irrelevant to the classification of the other cardiac cells (see clusters I, III, IV, V, VI and VIII in Figure 6C). In addition, this analysis revealed motifs relevant for activity in all cardiac cells (cluster II in Figure 6C). The identification of cell-type-specific clusters suggests a role for these motifs in mediating particular patterns of gene expression that are specific for different subsets cardiac cells.

Predicted sequence features discriminate individual cardiac cell fates

The preceding section identified sequence features that potentially discriminate enhancer activity in individual cardiac cells. In order to test this hypothesis, we first identified sequence features that were positively weighted within a cell subtype classification(s) and that were depleted or irrelevant for the other cardiac subtype models. We then used *cis* mutagenesis of a selected fraction of these sequence motifs in transgenic reporter assays to monitor the effects of their targeted removal from otherwise WT enhancers. For this purpose, we analyzed the activity of five separate motifs, each of which is predicted to discriminate regulatory element activity within subtypes of cardiac cells: V$ZF5_01, V$ETS_Q4, VTEF_01, VEVI1_06 and V$MTF_01 (Figure 7 and Supplementary Table S4).

The WT *mib1* enhancer ($mib1^{WT}$) is active in the Odd-PCs (Figure 7A) and contains two V$ZF5_01 motifs. This motif has a high positive weight within the Odd-PC classification, suggesting that it plays a critical role in Odd-PC enhancer activity (Figure 7A). In agreement with this hypothesis, mutagenesis of the V$ZF5_01 motifs in the *mib1* enhancer ($mib1^{ZF5}$) leads to a loss of reporter expression in Odd-PCs (Figure 7A).

We previously documented an essential role for Ets binding sites in enhancers with activity in Eve-PCs (40,41). We now extend this observation by showing that V$ETS_Q4 motifs are heavily weighted in the Eve-PC classification, and that the two V$ETS_Q4 motifs in the *Doc1* enhancer are critical for activity in Eve-PCs (Figure 7B). Interestingly, the V$ETS_Q4 motif is derived from binding sites for the ETS1 TF, whose ortholog in *Drosophila* is Pointed (Pnt) (42). In prior studies we also showed that Pnt was critical in *trans* for enhancer activity in Eve-PCs (40), a finding which further establishes that motif enrichment in enhancers can be used to reveal cell-type-specific TFs (10,24).

The V$TEF_01 motif is positively weighted amongst the Eve-PC and Odd-PC classification (Figure 7C), suggesting that it contributes a critical function to Eve-PC and Odd-PC enhancer activities. We now show that mutagenesis of the two V$TEF_01 motifs in the *CG13822* enhancer ($CG13822^{TEF}$) leads to a loss of reporter expression in Odd-PCs and de-repression into Eve-PCs. The V$TEF_01 motif is recognized by thyrotroph embryonic factor, which is a member of the proline and acidic amino acid-rich (PAR) subfamily of basic region/leucine zipper TFs, whose closest *Drosophila* ortholog is Par domain protein 1 (Pdp1) (42). The functional role of V$TEF_01 motifs in the *CG13822* enhancer suggests a role for *Pdp1* in cardiogenesis. In support of this hypothesis, a previous functional genomic screen uncovered a role for *Pdp1* in patterning the fly heart (43). Thus, both *cis* and *trans* tests of Pdp1 function are consistent with each other in establishing a key role for this TF in *Drosophila* cardiogenesis.

Finally, we used the SVM weights enriched amongst pan-PC and pan-CC classifications to uncover features that are essential for activity in all heart cells. The SVM weights for V$MTF1_01 and V$EVI1_06 motifs are positive amongst classifications of pan-PC and pan-CC enhancers (Figure 7D). The WT *sty* enhancer (sty^{WT}) is active in all PCs and CCs. Mutagenesis of the one V$EVI1_06 motif ($sty^{EVI}$) or the one V$MTF1_01 motif (sty^{MTF}) in the *sty* enhancer abrogates enhancer activity in the majority of PCs and CCs (Figure 7D), suggesting a critical role for these motifs in regulating enhancer activity in all heart cells. V$MTF1_01 is recognized by Metal regulatory factor 1 (MTF1) in vertebrates and V$EVI1_06 is recognized by EVI-1 (also known as MECOM and PRDM3) whose *Drosophila* orthologs correspond to MTF1 and hamlet (ham), respectively (42). The present identification and characterization of these TFs makes them excellent candidates for regulating cardiogenesis in *Drosophila*. In support of this model, targeted depletion of *ham* in the dorsal mesoderm using RNAi causes abnormalities in cardiogenesis (B.W.B. and A.M.M., unpublished observations).

CONCLUSIONS

The distribution of histone marks, *in vivo* TF binding, and the presence of TF binding motifs have all been exploited to reveal the enhancers that govern gene expression (3,4). Here we combined all three of these approaches using discriminative machine learning methods on a training set of enhancers with activity in distinct subtypes of cardiac cells to model cell-type-specific enhancer activity in the *Drosophila* heart. Using this approach, we uncovered sequence, chromatin and TF binding features that appear to underlie enhancer activity in individual cardiac cells. From these findings, we hypothesize that such features potentially discriminate the unique enhancer specificities of single cardiac cells, which we empirically confirm for a series of sequence motifs in regulating appropriate patterns of cardiac enhancer activity. Finally, by associating a cardiac gene expression atlas with the predicted enhancers from each cell subtype clas-

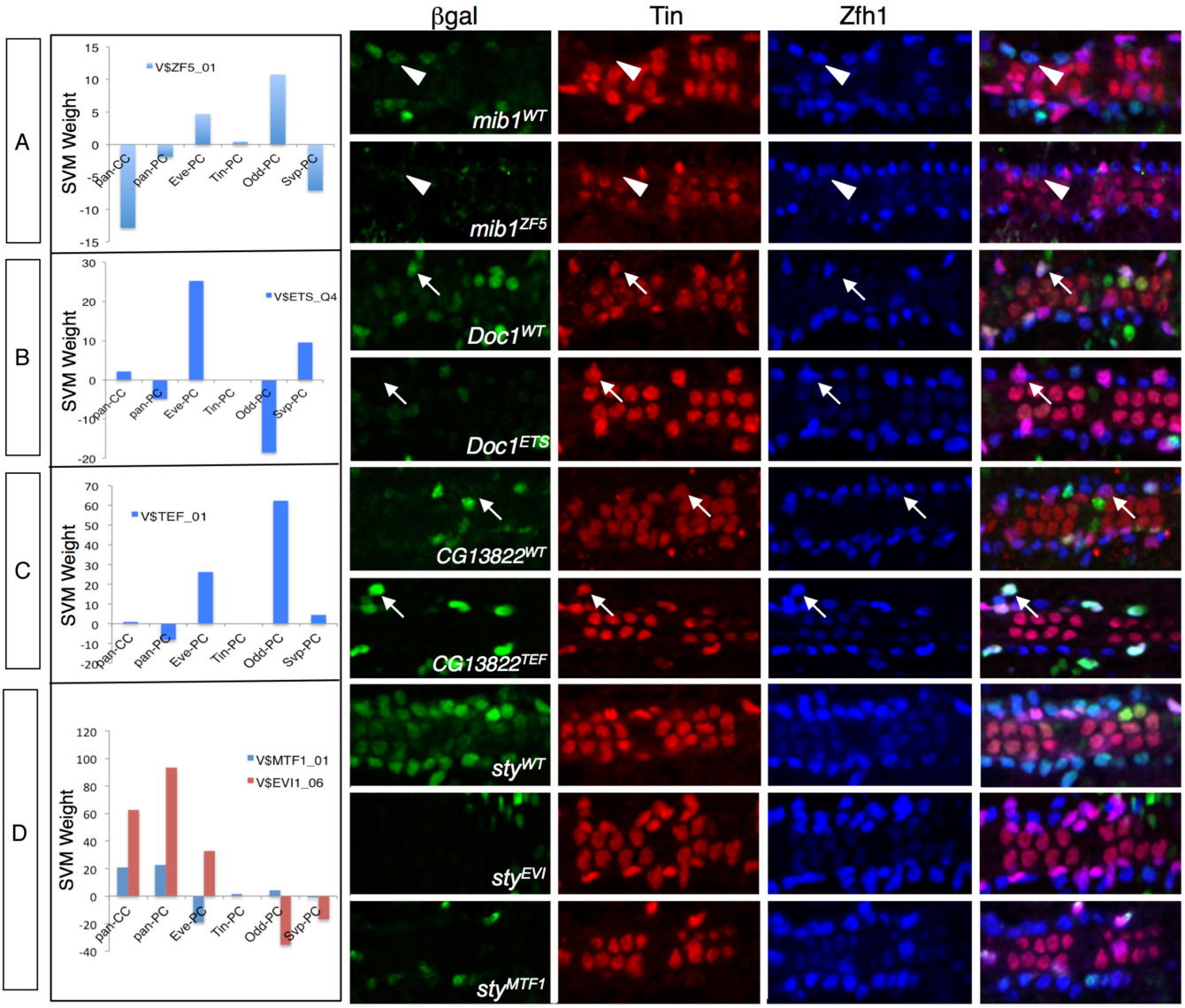

Figure 7. Differential enrichment of motifs uncovers sequence features which discriminate cardiac cell subtypes. SVM weights for each cardiac cell subtype classification for the indicated motifs are shown. Stage 16 embryos of the indicated genotypes were stained for βgal (green), Tin (red) and Zfh1 (blue). (**A**) SVM weights for V$TEF_01 shows enrichment amongst Eve-PC and Odd-PC classifications. βgal driven by the WT *CG13822* enhancer ($CG13822^{WT}$) is not expressed in Eve-PCs (arrow) but is expressed in Odd-PCs (not shown). Mutagenesis of V$TEF_01 motifs in the *CG13822* ($CG13822^{TEF.01}$) enhancer leads to de-repression of the reporter into Eve-PCs (arrowhead) and loss of reporter in Odd-PCs. (**B**) SVM weights for V$ETS_Q4 show enrichment amongst Eve-PC and Svp-PC classifications. The WT *Doc1* enhancer ($Doc1^{WT}$) is active in Eve-PCs (arrow) while mutagenesis of the V$Ets_Q4 motifs in the *Doc1* enhancer ($Doc1^{Ets.Q4}$) leads to a loss of reporter in Eve-PCs (arrow). (**C**) SVM weights for V$ZF5_01 show enrichment amongst Odd-PC and Eve-PC classifications. βgal driven by the WT *mib1* enhancer ($mib1^{WT}$) is expressed in the Odd-PCs (arrow) while is extinguished when the V$ZF5_01 motifs are mutated in the *mib1* enhancer ($mib1^{ZF5}$). (**D**) SVM weights for V$MTF_01 and V$EVI1_06 show are enriched amongst the pan-CC and pan-PC classifications. The WT *sty* enhancer (sty^{WT}) drives βgal expression in all PCs and CCs. Mutagenesis of the V$EVI_06 motifs in the *sty* enhancer (sty^{EVI}) leads to a loss of βgal reporter in all PCs and CCs whereas mutagenesis of V$MTF1_01 motifs in the *sty* enhancer (sty^{MTF1}) abrogates βgal expression in the majority of PCs and CCs.

sification, we uncovered previously unknown functions of individual cells of the *Drosophila* heart. Collectively, these results document the utility of computational modeling of enhancers to uncover the sequence motifs, chromatin and TF binding patterns as well as the gene expression profiles and functions of individual cells within the overall cardiac lineage.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank J. Skeath, M. Frasch and N. Perrimon for providing fly strains and antibodies and C. Sonnenbrot and E. Cozart for technical assistance.

FUNDING

National Heart Lung and Blood Institute Division of Intramural Research (to A.M.); Intramural Research Program of the National Institutes of Health, National Library of Medicine (to I.O.). Funding for open access charge: National Institutes of Health.
Conflict of interest statement. None declared.

REFERENCES

1. Busser,B.W., Bulyk,M.L. and Michelson,A.M. (2008) Toward a systems-level understanding of developmental regulatory networks. *Curr. Opin. Genet. Dev.*, **18**, 521–529.
2. Davidson,E. (2006) *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press, London.
3. Shlyueva,D., Stampfel,G. and Stark,A. (2014) Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.*, **15**, 272–286.
4. Spitz,F. and Furlong,E.E. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, **13**, 613–626.
5. Bodmer,R. and Frasch,M. (2010) *Development and Aging of the Drosophila Heart*. Academic Press, London.
6. Olson,E.N. (2006) Gene regulatory networks in the evolution and development of the heart. *Science*, **313**, 1922–1927.
7. Ahmad,S.M., Busser,B.W., Huang,D., Cozart,E.J., Michaud,S., Zhu,X., Jeffries,N., Aboukhalil,A., Bulyk,M.L., Ovcharenko,I. *et al.* (2014) Machine learning classification of cell-specific cardiac enhancers uncovers developmental subnetworks regulating progenitor cell division and cell fate specification. *Development*, **141**, 878–888.
8. Busser,B.W., Gisselbrecht,S.S., Shokri,L., Tansey,T.R., Gamble,C.E., Bulyk,M.L. and Michelson,A.M. (2013) Contribution of distinct homeodomain DNA binding specificities to *Drosophila* embryonic mesodermal cell-specific gene expression programs. *PLoS One*, **8**, e69385.
9. Busser,B.W., Huang,D., Rogacki,K.R., Lane,E.A., Shokri,L., Ni,T., Gamble,C.E., Gisselbrecht,S.S., Zhu,J., Bulyk,M.L. *et al.* (2012) Integrative analysis of the zinc finger transcription factor Lame duck in the *Drosophila* myogenic gene regulatory network. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 20768–20773.
10. Busser,B.W., Taher,L., Kim,Y., Tansey,T., Bloom,M.J., Ovcharenko,I. and Michelson,A.M. (2012) A machine learning approach for identifying novel cell type-specific transcriptional regulators of myogenesis. *PLoS Genet.*, **8**, e1002531.
11. Markstein,M., Pitsouli,C., Villalta,C., Celniker,S.E. and Perrimon,N. (2008) Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. *Nat. Genet.*, **40**, 476–483.
12. Groth,A.C., Fish,M., Nusse,R. and Calos,M.P. (2004) Construction of transgenic *Drosophila* by using the site-specific integrase from phage phiC31. *Genetics*, **166**, 1775–1782.
13. Busser,B.W., Shokri,L., Jaeger,S.A., Gisselbrecht,S.S., Singhania,A., Berger,M.F., Zhou,B., Bulyk,M.L. and Michelson,A.M. (2012) Molecular mechanism underlying the regulatory specificity of a *Drosophila* homeodomain protein that specifies myoblast identity. *Development*, **139**, 1164–1174.
14. Berriz,G.F., Beaver,J.E., Cenik,C., Tasan,M. and Roth,F.P. (2009) Next generation software for functional trend analysis. *Bioinformatics*, **25**, 3043–3044.
15. Supek,F., Bosnjak,M., Skunca,N. and Smuc,T. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.
16. Narlikar,L., Sakabe,N.J., Blanski,A.A., Arimura,F.E., Westlund,J.M., Nobrega,M.A. and Ovcharenko,I. (2010) Genome-wide discovery of human heart enhancers. *Genome Res.*, **20**, 381–392.
17. Loots,G.G. and Ovcharenko,I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W217–W221.
18. Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhäuser,R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.
19. Sandelin,A., Alkema,W., Engström,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91-D94 .
20. Robasky,K. and Bulyk,M.L. (2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **39**, D124–D128.
21. Bonn,S., Zinzen,R.P., Girardot,C., Gustafson,E.H., Perez-Gonzalez,A., Delhomme,N., Ghavi-Helm,Y., Wilczynski,B., Riddell,A. and Furlong,E.E. (2012) Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.*, **44**, 148–156.
22. Junion,G., Spivakov,M., Girardot,C., Braun,M., Gustafson,E.H., Birney,E. and Furlong,E.E. (2012) A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell*, **148**, 473–486.
23. (2013) *MATLAB Version 8.1.0.604*. The MathWorks Inc., Natick, MA.
24. Ahmad,S.M., Tansey,T.R., Busser,B.W., Nolte,M.T., Jeffries,N., Gisselbrecht,S.S., Rusan,N.M. and Michelson,A.M. (2012) Two forkhead transcription factors regulate the division of cardiac progenitor cells by a polo-dependent pathway. *Dev. Cell*, **23**, 97–111.
25. Jin,H., Stojnic,R., Adryan,B., Ozdemir,A., Stathopoulos,A. and Frasch,M. (2013) Genome-wide screens for in vivo tinman binding sites identify cardiac enhancers with diverse functional architectures. *PLoS Genet.*, **9**, e1003195.
26. Dong,X., Greven,M.C., Kundaje,A., Djebali,S., Brown,J.B., Cheng,C., Gingeras,T.R., Gerstein,M., Guigo,R., Birney,E. *et al.* (2012) Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.*, **13**, R53.
27. Wilczynski,B., Liu,Y.H., Yeo,Z.X. and Furlong,E.E. (2012) Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLoS Comput. Biol.*, **8**, e1002798.
28. Gorkin,D.U., Lee,D., Reed,X., Fletez-Brant,C., Bessling,S.L., Loftus,S.K., Beer,M.A., Pavan,W.J. and McCallion,A.S. (2012) Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. *Genome Res.*, **22**, 2290–2301.
29. Kantorovitz,M.R., Kazemian,M., Kinston,S., Miranda-Saavedra,D., Zhu,Q., Robinson,G.E., Gottgens,B., Halfon,M.S. and Sinha,S. (2009) Motif-blind, genome-wide discovery of cis-regulatory modules in *Drosophila* and mouse. *Dev. Cell*, **17**, 568–579.
30. Denholm,B. and Skaer,H. (2009) Bringing together components of the fly renal system. *Curr. Opin. Genet. Dev.*, **19**, 526–532.
31. Ward,E.J. and Skeath,J.B. (2000) Characterization of a novel subset of cardiac cells and their progenitors in the *Drosophila* embryo. *Development*, **127**, 4959–4969.
32. Marsman,J. and Horsfield,J.A. (2012) Long distance relationships: enhancer-promoter communication and dynamic gene transcription. *Biochim. Biophys. Acta*, **1819**, 1217–1227.
33. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
34. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
35. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
36. Margueron,R. and Reinberg,D. (2011) The Polycomb complex PRC2 and its mark in life. *Nature*, **469**, 343–349.
37. Voigt,P., Tee,W.W. and Reinberg,D. (2013) A double take on bivalent promoters. *Genes Dev.*, **27**, 1318–1338.
38. Zhou,V.W., Goren,A. and Bernstein,B.E. (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.*, **12**, 7–18.
39. Lin,W. and Dent,S.Y. (2006) Functions of histone-modifying enzymes in development. *Curr. Opin. Genet. Dev.*, **16**, 137–142.

40. Halfon,M.S., Carmena,A., Gisselbrecht,S., Sackerson,C.M., Jiménez,F., Baylies,M.K. and Michelson,A.M. (2000) Ras pathway specificity is determined by the integration of multiple signal-activated and tissue-restricted transcription factors. *Cell*, **103**, 63–74.
41. Halfon,M.S., Grad,Y., Church,G.M. and Michelson,A.M. (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.*, **12**, 1019–1028.
42. Hu,Y., Flockhart,I., Vinayagam,A., Bergwitz,C., Berger,B., Perrimon,N. and Mohr,S.E. (2011) An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics* **12**, 357.
43. Kim,Y.O., Park,S.J., Balaban,R.S., Nirenberg,M. and Kim,Y. (2004) A functional genomic screen for cardiogenic genes using RNA interference in developing *Drosophila* embryos. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 159–164.