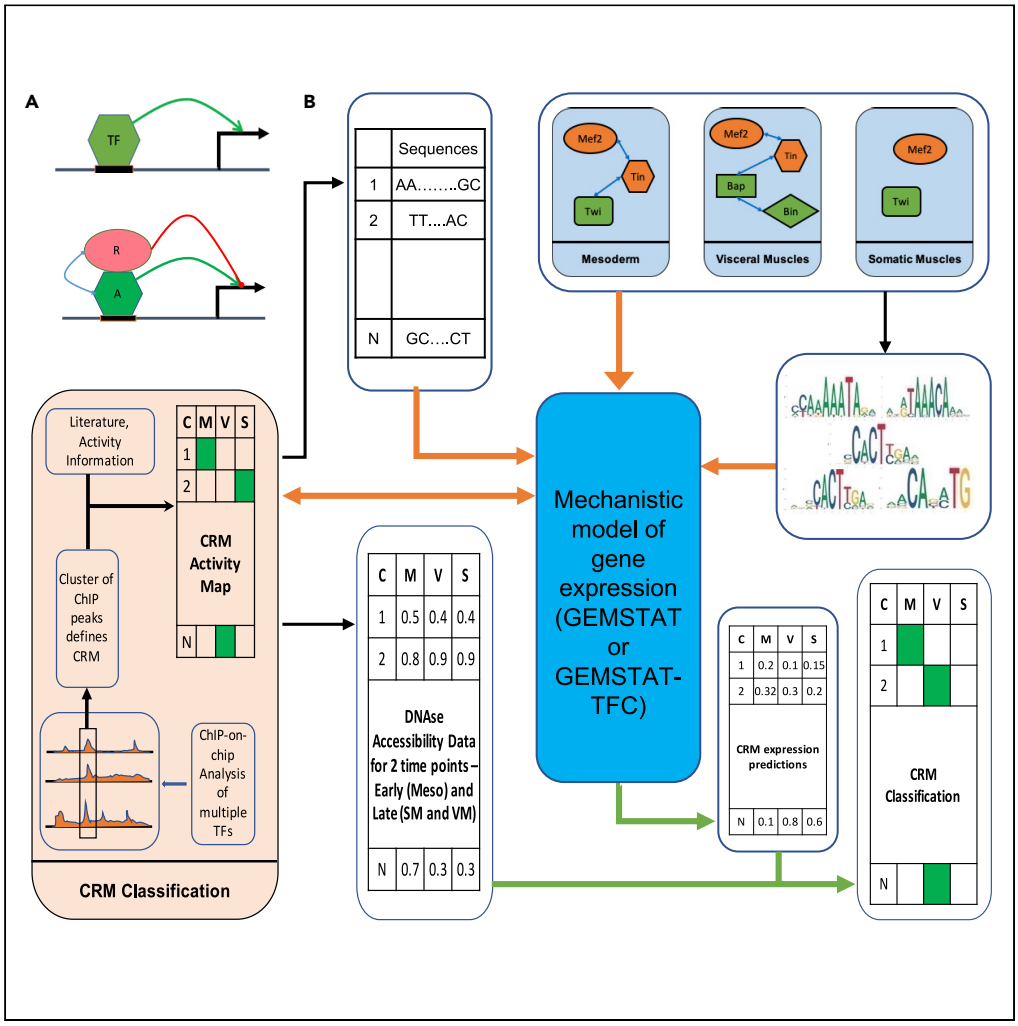
iScience

CellPress
OPEN ACCESS

Article

Thermodynamics-based modeling reveals regulatory effects of indirect transcription factor-DNA binding



Shounak Bhogale, Saurabh Sinha

sinhas@illinois.edu

Highlights

Inclusion of indirect DNA binding of transcription factor improves enhancer function prediction

Context specific activating or repressive roles of TFs

Indirect binding improves fits to experimental TF-DNA binding data

Role of Tinman depends on its DNA-binding mode (direct or indirect)

Bhogale & Sinha, iScience 25, 104152
May 20, 2022 © 2022
https://doi.org/10.1016/j.isci.2022.104152

iScience

CellPress
OPEN ACCESS

Article

Thermodynamics-based modeling reveals regulatory effects of indirect transcription factor-DNA binding

Shounak Bhogale[1] and Saurabh Sinha[1,2,3,4,5,*]

SUMMARY

Transcription factors (TFs) influence gene expression by binding to DNA, yet experimental data suggests that they also frequently bind regulatory DNA indirectly by interacting with other DNA-bound proteins. Here, we used a data modeling approach to test if such indirect binding by TFs plays a significant role in gene regulation. We first incorporated regulatory function of indirectly bound TFs into a thermodynamics-based model for predicting enhancer-driven expression from its sequence. We then fit the new model to a rich data set comprising hundreds of enhancers and their regulatory activities during mesoderm specification in Drosophila embryogenesis and showed that the newly incorporated mechanism results in significantly better agreement with data. In the process, we derived the first sequence-level model of this extensively characterized regulatory program. We further showed that allowing indirect binding of a TF explains its localization at enhancers more accurately than with direct binding only. Our model also provided a simple explanation of how a TF may switch between activating and repressive roles depending on context.

INTRODUCTION

Development of heterogeneous muscles layers from the mesoderm in a Drosophila embryo involves multiple transcription factors (TFs) working in tandem to regulate gene expression and control cell differentiation programs. Multiple signals from the ectoderm set up the differentiation process (Azpiazu et al., 1996; Mbodj et al., 2016), key TFs are then expressed in a successive manner at precise locations, allowing the mesodermal cells to differentiate into different muscle layers (Azpiazu et al., 1996; Zinzen et al., 2009). Here we study the formation of visceral muscles (VM) and somatic muscles (SM) from the mesoderm, focusing on the gene regulatory mechanisms underlying these processes. Our specific goal was to decipher how those mechanisms are encoded in cis-regulatory sequences associated with muscle specification.

The TFs Bagpipe (Bap), Biniou (Bin), Myocyte enhancer factor-2 (Mef2), Tinman (Tin), and Twi are key regulators of mesoderm development (Sandmann et al., 2006,2007; Jakobsen et al., 2007; Bonn and Furlong, 2008; Liu et al., 2009; Zinzen et al., 2009). These TFs act as high-level regulators throughout the mesoderm and its subsequent differentiated tissues, and their combinatorial regulatory action is known to be highly predictive of gene expression patterns in the muscle layers (Zinzen et al., 2009; Cusanovich et al., 2018). Such combinatorial regulation is mediated by sequences called enhancers that act as hubs for TF-DNA binding. Extensive prior work has resulted in large catalogs of enhancers (Zinzen et al., 2009; Cusanovich et al., 2018) with annotated regulatory activity in mesoderm, VM or SM. This allowed us to systematically analyze the TF binding site combinations that distinguish enhancers exhibiting one type of activity versus others, thus characterizing the cis-regulatory code (Yáñez-Cuna et al., 2013) of mesoderm development.

Our analytical approach was to fit thermodynamics-based models of enhancer function to predict each enhancer's regulatory activity in a muscle layer, based on the strengths of TF binding sites in the enhancer's sequence and information about TFs' expression in that layer. The baseline model of our choice is called GEMSTAT (He et al., 2010), a statistical thermodynamics-based model that we have previously used to model enhancers involved in specification of the anterior-posterior axis (Samee et al., 2017) and dorsoventral axis (Samee et al., 2015) in early Drosophila embryos. Put simply, the GEMSTAT model considers

[1]Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[2]Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[3]Carl R. Woese Institute of Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[4]Cancer Center at Illinois, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[5]Lead contact

*Correspondence: sinhas@illinois.edu

https://doi.org/10.1016/j.isci.2022.104152

the strength of each TF binding site along with the bound TF's interaction with the basal transcription machinery, to calculate the net regulatory output of the assortment of sites in an enhancer.

However, it has been speculated that TFs can exert their regulatory effects without actually binding to the DNA (Gordân et al., 2009; Shokri et al., 2019). Protein-protein interactions between TFs can help a TF "piggyback" on a DNA-bound TF, thus localizing at enhancers that have weak or no sites for the former. Evidence for such indirect binding of a TF to specific sites in the genome (where binding sites for its partner TF are present) has been reported previously from ChIP-seq data (Gordân et al., 2009; Author Anonymous, 2019). The possibility of TFs regulating gene expression through indirect (as well as direct) binding to enhancers has been discussed in the literature as the "TF Collective" model (Doitsidou et al., 2013; Uhl et al., 2016; Khoueiry et al., 2017), as a complementary view to the "billboard" and "enhanceosome" models (Arnosti and Kulkarni, 2005) of enhancer organization and function. Junion et al. (2012) provided strong evidence for the TF Collective model in mesoderm development, by examining regulatory activities of genomic segments bound by a collective of five cardiogenic TFs and noting less stringent sequence (motif) requirements therein, suggesting a significant role for cooperative as well as indirect TF-DNA binding in enhancer function. Khoueiry et al. (2017) used evolutionary analysis of TF ChIP peaks and sequence motifs in mesodermal enhancers to provide further support for the model. Here, we sought to quantify the extent to which the "TF Collective" phenomenon influences enhancer readout in mesoderm development. For this, we developed an extension of the GEMSTAT model that incorporates regulatory action of an indirectly bound TF, and assessed if this model, called GEMSTAT-TFC, better fits enhancer activity data than the baseline GEMSTAT model. This allowed us to determine the significance of the TF Collective phenomenon in a well-controlled quantitative manner.

We found that allowing indirectly bound TFs to exert regulatory influence leads to significantly more accurate predictions of enhancer activity, under various methods of assessment. For example, the accuracy of three-way classification into mesoderm, visceral muscle, and somatic muscle activity on a held-out data set of 69 enhancers increased from 62% to 74%. Accommodating indirect binding also yielded significantly better prediction of ChIP signal based on sequence, particularly for the TFs Twist and Tinman. Closer examination of the trained models revealed the possibility of a TF exhibiting "dual roles" (activating and repressive) (Bauer et al., 2010) in different cis- or trans- contexts without a switch in its biochemical function: for instance, the complex formed by a DNA-bound activator TF and a repressor TF "piggybacking" on it may have a net repressive effect mediated by the activator's binding site. Our models suggest such dual roles for Tinman in particular. In summary, this study presents the first sequence-to-expression modeling of mesoderm development enhancers and finds evidence of the TF Collective model and dual-role TFs in the organization of these enhancers.

RESULTS

A thermodynamics-based sequence-to-expression model that incorporates indirect binding of transcription factors

The new model proposed here, called GEMSTAT-TFC (TFC abbreviates Transcription Factor Collective (Doitsidou et al., 2013; Uhl et al., 2016; Khoueiry et al., 2017), is an extension of an existing sequence-to-expression model called GEMSTAT. GEMSTAT is a mechanistic model that predicts the quantitative expression driven by an enhancer in a given cellular context described by concentrations of relevant TFs. It uses motifs (position weight matrices or "PWM"s) representing TFs' binding preferences to estimate binding strength at every putative site in the enhancer and automatically learns each TF's regulatory strength and direction (activator or repressor) from given data. A key premise of GEMSTAT and in other thermodynamics-based models of enhancer function (He et al., 2010), is that a TF influences gene expression by direct binding to its cognate sites in enhancer DNA. The main idea behind a "TF collective" formalism is that TFs may also affect a gene's expression without direct DNA-binding, and thus without having any binding sites in the enhancer.

The GEMSTAT model considers all possible configurations of binding sites being bound (or not) by their cognate TFs. Thus, for example, an enhancer with a single binding site S_A, for TF "A", has two configurations, with the relative probability of the bound configuration σ_A being proportional to $W(\sigma_A) \times Q(\sigma_A)$, where $W(\sigma_A)$ mirrors TF-DNA interactions while $Q(\sigma_A)$ represents the interaction between a DNA-bound molecule of TF A and the basal transcriptional machinery (BTM). The term $W(\sigma_A)$ depends on concentration of TF A and on the interaction energy between TF A and its site S_A, estimated from the sequence of S_A and

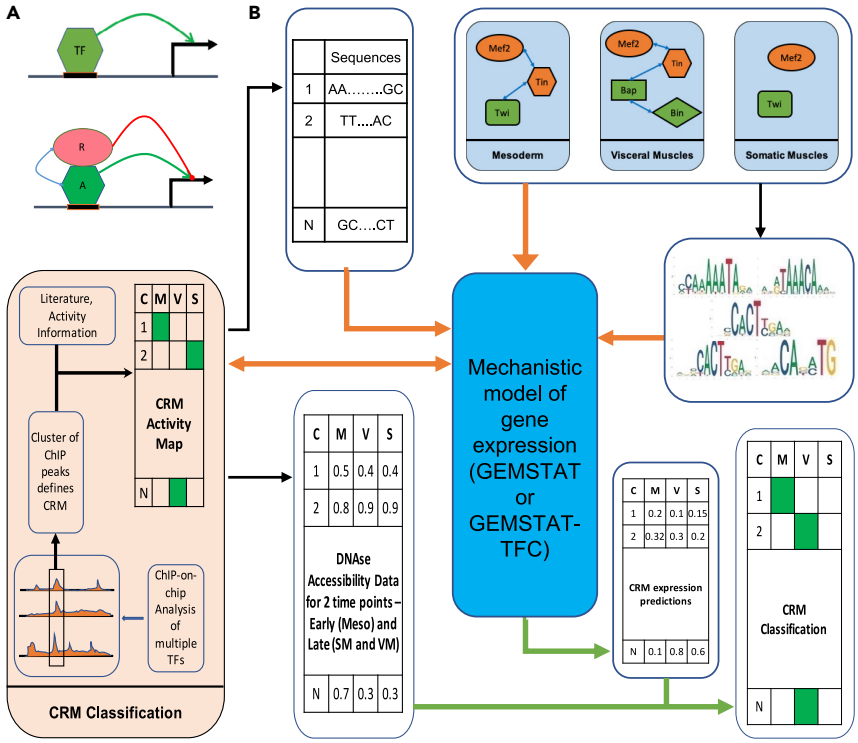


Figure 1. Model schematic

(A) Schematic illustrating difference between GEMSTAT model (upper) and GEMSTAT-TFC model (lower). TFC model allows for indirect DNA-binding of a TF at the site of another TF, leading to modulation of the latter (directly bound) TF's regulatory activity by the former (indirectly bound) TF.

(B) Schematic overview of model training and evaluation. (Left) Genomic regions having ChIP peaks for multiple TFs (from Bap, Bin, Mef2, Tin and Twi) were marked as CRMs, i.e., putative enhancers. Each CRM was assigned to one of three "layers" (classes) according to their experimentally reported activity in the three muscle layers – mesoderm (Meso), visceral muscle (VM), and somatic muscle (SM). The resulting activity map was provided to the model along with sequence of each CRM (top left). Information about relevant TFs was also provided; this includes muscle layer activities (present or absent) of each of five TFs and known/speculated protein-protein interactions among them (top, right), as well as PWMs characterizing their DNA-binding preferences (right, middle). The orange color represents the TFs that are known to have dual roles based on literature whereas green color represents known activators, but the model is not provided any information about the role of a TF. The model (GEMSTAT or GEMSTAT-TFC) scans each CRM for presence of TF binding sites using the PWMs and combines this "cis" information with "trans" information (presence of TF) to predict the expression level driven by the CRM in each muscle layer. DNA accessibility (rank normalized DNase I Hypersensitivity) scores of CRMs are used to modulate their model-predicted expression. Finally, the CRM is assigned to the layer with highest predicted expression value for that CRM.

the PWM of TF A. In the GEMSTAT-TFC model, a TF can "piggyback" on another TF that is already bound to the DNA, through protein-protein interaction (PPI) as shown in Figure 1A. The piggybacking TF can then modulate the net effect of the DNA-bound TF on a gene's expression level, as if it were directly bound to DNA. Thus, in the case of the single-site enhancer noted above, there are three possible configurations now, with the new configuration (say σ_{AR}) having a molecule of TF R bound via PPI to TF A, which in turn is bound to its site S_A. The relative probability of configuration σ_{AR} is of the form $W(\sigma_{AR}) \times Q(\sigma_{AR})$, as for σ_A, but now $W(\sigma_{AR})$ also depends on the concentration of TF R and the interaction energy between TF A and TF R, and $Q(\sigma_{AR})$ also reflects interaction between TF R and the BTM. GEMSTAT models the interactions between BTM and multiple bound TFs by adding their respective interaction energies. GEMSTAT-TFC inherits this feature and assumes additivity of the BTM-interaction energies of all directly or indirectly DNA-bound TFs. It allows one or more TFs to piggyback on to a DNA-bound TF but demands that only one TF may piggyback on a given TF molecule in a particular configuration. Also, we note that GEMSTAT allows cooperative interactions (via PPI) between two DNA-bound TFs if their respective bound sites are located proximally. GEMSTAT-TFC retains this feature and uses the same interaction parameter for a pair of TFs when modeling this scenario (both TFs DNA-bound) or the indirect binding scenario (one

TF DNA-bound). GEMSTAT trains two free parameters per TF (one representing TF-DNA binding strength and one for the TF's regulatory effect), and one free parameter for each pair of interacting TFs. GEMSTAT-TFC trains the same free parameters and only one additional global free parameter. See STAR Methods for details of the GEMSTAT-TFC model.

In GEMSTAT-TFC, as in GEMSTAT, a TF has a fixed biochemical effect on gene expression, i.e., the strength and direction (activator or repressor) is fixed, and learnable from data. However, the mathematical formulation of the GEMSTAT-TFC model implies a binding site for an activator (resp. repressor) may in fact have a repressive (resp. activating) net effect on gene expression. For example, if *A* is an activator but a repressor *R* can piggyback on a bound molecule of *A* (as in Figure 1A), then the effect of site S_A (of TF *A*) on gene expression includes both an activating contribution from the directly bound *A* and a repressive contribution from the indirectly bound *R*. As a result, mutagenizing site S_A may increase gene expression if the repressive contribution was larger in magnitude, making this appear to be a repressive site even though its cognate TF (A) is an activator in the model. Furthermore, if the latter condition is true, a knock-out of TF *A* not only precludes the activating effect of *A* but also the repressive effect of *R* (via the DNA-bound *A*), and thus may lead to an increase in gene expression, suggesting a net repressive role for *A* in this context, even though its intrinsic effect as an activator is unchanged. This is an illustration of how indirect binding in the GEMSTAT-TFC model may result in "dual roles" for TFs (See discussion).

GEMSTAT-TFC yields better fits to data on mesoderm specification

We used the GEMSTAT-TFC model to analyze available expression data on the well-studied gene regulatory network underlying mesoderm development in Drosophila (Figure 1B). Specifically, we studied a collection of cis-regulatory modules (CRMs) (Peng et al., 2019) defined based on ChIP-chip data for the major TFs of this network – Bap, Bin, Mef2, Tin and Twi – at multiple stages of embryonic development. We focused on 233 CRMs that are known to drive expression in one of the three muscle layers "mesoderm" ("Meso"), "visceral muscle" ("VM") and "somatic muscle" ("SM"), according to the CAD4 database (Cusanovich et al., 2018). Our modeling also utilized information about the five above-mentioned TFs known to be important to gene regulation in these three muscle layers: presence or absence of each TF in each muscle layer and known or speculated protein-protein interactions between pairs of TFs (Figure 1B, top right), as well as PWM motif of each TF (see STAR Methods, Figure S1). The model learns to map each CRM to a muscle layer. It does so by predicting (numeric) expression level driven by the CRM in each muscle layer, utilizing information on TF presence/absence in that layer, and assigns the CRM to the layer with highest predicted expression. This may also be viewed as a 3-way classification task.

We modeled the data with the GEMSTAT-TFC model and the GEMSTAT model separately, seeking to test the significance of the "TFC" component, which explicitly accounts for indirect TF binding. Recall that the GEMSTAT-TFC model has one additional parameter compared to GEMSTAT (16 versus 15), hence we used a cross-validation approach for model comparison, partitioning the 233 CRMs randomly into training and test sets in 4:1 ratio, repeating this process 100 times. (Random partitions were done in a way that retains the relative proportions of the three classes in training and test sets.) We noted GEMSTAT-TFC to be significantly more accurate than GEMSTAT overall (Figure 2A,"Total"), with an accuracy of 0.49 compared to 0.45 (p value 8.9E-11, paired Wilcoxon test). The improved accuracy is especially pronounced for the SM and VM classes (Figure 2A) whereas the two models have comparable performance for the Meso class.

Examination of hits and misses of models reveals learned mechanisms

For a better understanding of the models, we analyzed motif matches for the five TFs in the collection of CRMs. First, we counted motif presence (estimated number of binding sites) in CRMs of the three classes separately (Figure 2C) and observed statistically significant differences for some motifs. For example, CRMs of the VM class are enriched in the presence of the Bin motif compared to those of Meso and SM classes. This agrees with Bin being present exclusively in the VM layer, and both models utilizing it as an activator in this layer (Table S4). The Mef2 motif is enriched in CRMs of the SM class, but we did not notice any significant class enrichment for the other three TF motifs.

We next focused on CRMs that the models consistently succeeded or failed in classifying into their respective muscle layers, across the 100 iterations of random partitioning into training and test sets. We called a CRM a "hit" for the GEMSTAT (or-TFC) model if its activity class was correctly predicted by the model in at least 1/3 of the iterations where it appeared in the test set; otherwise, it was designated as a "miss". (The

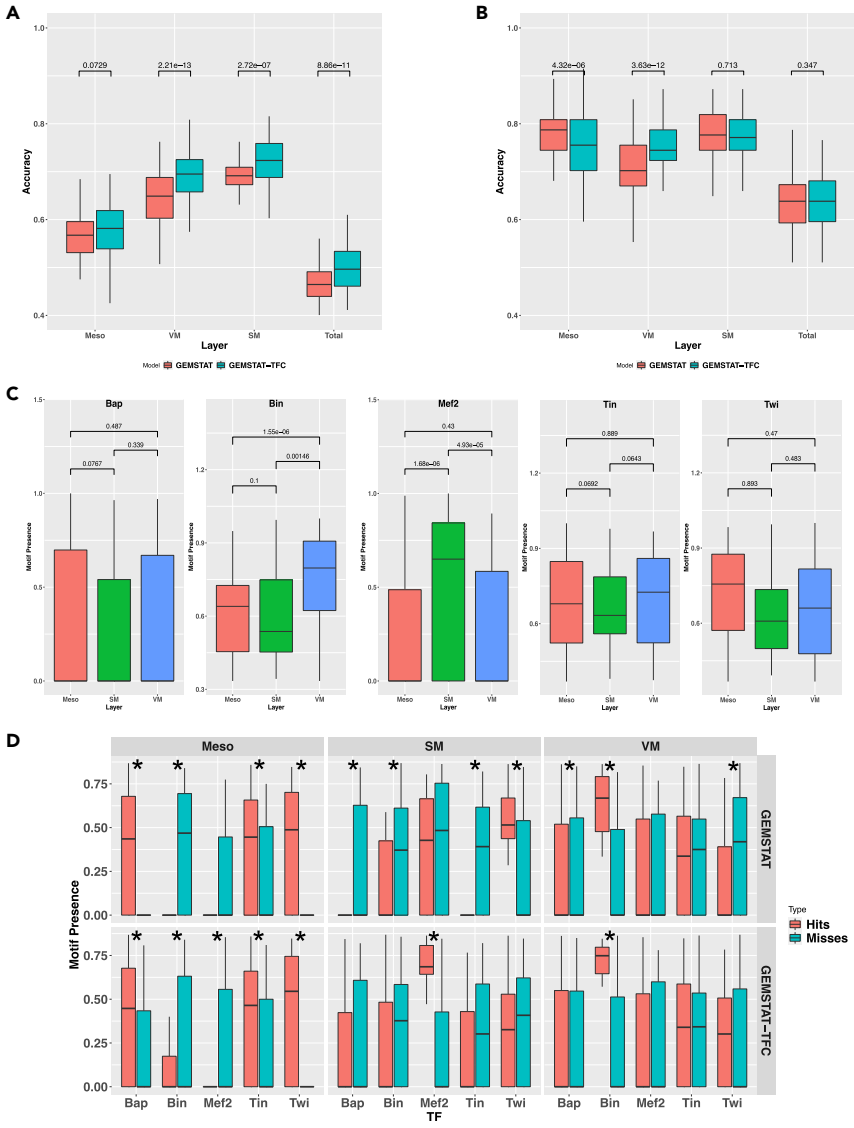


Figure 2. Model performance and features

(A and B) Performance comparison between the two models – GEMSTAT and GEMSTAT-TFC. CRMs were randomly partitioned into training set (80%) and test set (20%). Y-axis shows the test accuracy of the models, averaged across 100 repeats of random partitioning, training and testing, on the task of classifying CRM activity into one muscle layer versus the others. Total accuracy for the three-way classification task is also shown ("Total"). The two panels show accuracy comparisons for predictions without (A) and with (B) use of accessibility data. In both cases, the accuracy of GEMSTAT-TFC models (blue boxes) generally tends to be better than the GEMSTAT models (red boxes), as shown by p values of a paired Wilcoxon test applied to each pair of distributions.

(C) TFs exhibit enrichment of predicted binding sites in specific CRM classes. Each panel shows average binding site presence (LR scores after rank normalization) of a TF in CRMs assigned to each muscle layer – mesoderm ("Meso"), visceral muscle ("VM") and somatic muscle ("SM"). The TF Bin is known to be a key TF in visceral muscle and its binding sites are enriched in VM enhancers (blue box).

(D) Motifs of select TFs distinguish CRMs that are amenable to model-based activity prediction. CRMs were designated as "hits" or "misses" based on how frequently they were predicted correctly as test examples in the 100 repeats of random partitioning, training and testing of GEMSTAT or GEMSTAT-TFC models. Each TF's motif presence (rank normalized LR score) was recorded in CRMs of either class ("hits" and "misses") and compared between classes. The comparisons are shown separately for CRMs with known activity in each muscle layer (label at top of each plot) and for each model (label on right of each row of panels). The TFs that have significant differences in motif presence between "hit" and "miss" CRMs for a model are marked with a star for each muscle layer. For example, Bin is the key TF for the visceral muscle (VM), and CRMs that are not correctly predicted by the models have significantly lower motif presence of Bin than those that are correctly predicted. (This was observed with either model).

Table 1. Comparison of the predictive accuracy of different models, in leave-one-out cross-validation

Layer	Random	Size-aware Random	GEMSTAT	GEMSTAT-TFC	Accessibility	GEMSTAT(A)	GEMSTAT-TFC(A)	ChIP-based XGBoost
Meso vs Rest	0.52	0.51	0.56	0.57	0.82	0.79	0.76	0.81
VM vs Rest	0.57	0.60	0.64	0.67	0.69	0.71	0.75	0.81
SM vs Rest	0.57	0.59	0.69	0.73	0.67	0.75	0.77	0.83
3-way	0.33	0.35	0.45	0.49	0.59	0.62	0.64	–

Accuracy is shown separately for each muscle layer – mesoderm ("Meso"), visceral muscle ("VM"), and somatic muscle ("SM") – for the task of predicting if a CRM is active in that layer or not. Accuracy for the 3-way classification task is also shown ("3-way"). (Note that the data set only includes CRMs that are active in exactly one of the three muscle layers.) Two different random models were evaluated as baselines, one that predicts each class with 1/3 probability ("Random") and one that predicts classes with probability proportional to their representation in the data ("Size-aware random"). Listed accuracies for the random models are theoretical expectations. GEMSTAT and GEMSTAT-TFC models were evaluated without and with (denoted by "(A)") accessibility information. Predictions based on accessibility alone were also evaluated, for the task of classifying Meso CRMs versus VM/SM CRMs. A machine learning model that uses ChIP data for multiple TFs at multiple stages of development was also evaluated ("ChIP-based XGBoost"). This model is trained separately for the binary classification task for each layer, while GEMSTAT and GEMSTAT-TFC were trained globally (a single model for all layers).

threshold of 1/3 was used because the random chance of a prediction being correct is 1/3 because there are 3 possible predictions out of which only one could be correct.) Comparing motif presence in hits versus misses (Figure 2D) revealed additional insights into the trained models. For instance, "hit" CRMs of the VM class are significantly enriched for Bin motif presence compared to "miss" CRMs of this class. This is true for either of the models and is in agreement with the observation made above (Figure 2C) about Bin being utilized as an activator for the VM layer. Interestingly, the opposite trend is seen in Meso CRMs – the Bin motif is depleted in "hits" versus "misses" of this class. This suggests that the presence of Bin binding sites misleads the model to predict expression of true mesodermal CRMs in the VM layer, where that TF is present, rather than the Meso layer where the TF is absent.

As another example, the "hit" CRMs of the Meso class are enriched for Twi motif matches compared to "misses" of this class, for both models (Figure 2D). Indeed, the TF Twi is present in the mesoderm layer and both models learned an activating role for this TF (Supplement); that the models tend to fail on CRMs lacking Twi motif matches suggests an important role for Twi for this class. On the other hand, the Twi motif is relatively enriched in the "misses" of the SM class for the GEMSTAT model (Figure 2D); this suggests that the presence of this motif contributes to misclassification of certain SM CRMs into the Meso class. These observations remind us of the combinatorial and quantitative nature of the modeling challenge: The TF Twi is present in Meso as well as SM and its (putative) binding sites are present in CRMs of both layers, meaning that without consideration of the strength of Twi binding sites and the presence of other TFs' sites in a CRM it would be impossible to classify CRMs between these two layers.

Comparison of sequence-based predictions to those based on epigenomic data

Our next set of model evaluations relied on a "leave-one-out cross-validation" (LOOCV), where each iteration designates one CRM as test data and all remaining CRMs are used for model training. Once again, GEMSTAT-TFC exhibited higher predictive accuracy (0.49 versus 0.45, Table 1) than GEMSTAT, the improvement being most prominent on the SM and VM classes (the two smaller classes), and both models performing equally on the Meso class. Its accuracy was clearly greater than that of a random predictor that predicts each class with 1/3 probability (0.49 versus 0.33) or one that predicts each class with probability proportional to the class size in the data set (0.49 versus 0.35).

We next examined model performance in comparison with predictions based on epigenomic data. First, we utilized whole-embryo DNA accessibility data from stages 5–7 ("early") and 13–15 ("late") [9], using early accessibility data to predict mesoderm activity and late accessibility data to predict VM/SM activity. The available data do not allow us to discern between VM and SM classes (both muscle layers are formed in the later developmental stage), but they allowed accurate classification of Meso CRMs versus VM/SM CRMs (accuracy of 0.82, Table 1 and it achieves accuracies of 0.69 and 0.67 in VM vs Rest and SM versus Rest classification tasks respectively). Encouraged by this observation, and motivated by previous reports (Peng et al., 2015), we utilized the accessibility data to adjust the predictions of our sequence-based models. We adopted a simple heuristic, whereby the model-predicted expression in a layer is multiplied by the rank-normalized accessibility of the CRM in the appropriate stage (early for Meso and late for VM or SM). We refer

Table 2. Model accuracy on held-out data (69 unseen enhancers from CAD4 database, not overlapping with training CRMs)

Layer	GEMSTAT(A)	GEMSTAT-TFC(A)
Meso(33) vs. Rest(36)	0.75	0.78
VM(16) vs. Rest(53)	0.75	0.86
SM(20) vs. Rest(49)	0.73	0.84
3-way	0.62	0.74

Accuracy values are shown for the binary classification task for each muscle layer ("Meso","VM","SM") versus the other two, as well as overall ("3-way") accuracy. GEMSTAT-TFC correctly predicts the activity of 8 additional CRMs compared to GEMSTAT (51 true positives versus 43, out of a maximum of 69), and also makes 8 fewer false positive errors (18 compared to 26). The number in the brackets in the first column denotes the number of enhancers in the class.

to these prediction schemes as GEMSTAT(A) and GEMSTAT-TFC(A), indicating that the sequence-based model's prediction has been augmented with accessibility data ('A'). Note that accessibility data was not utilized during training of the GEMSTAT/TFC models. Thus, we adopted an admittedly simplistic scheme for incorporating accessibility data at a coarse resolution, and while more sophisticated approaches exist (Peng et al., 2015), we eschewed those here to make model training and comparisons easier. Despite their simplicity, these schemes were substantially more accurate than without accessibility data (Table 1). For instance, GEMSTAT-TFC(A) had an accuracy of 0.76 for in discerning Meso CRMs versus others, compared to 0.57 when not using accessibility data. It had a high accuracy (0.75, 0.77) in classifying VM (versus others) and SM (versus others) as well, even though accessibility data by itself cannot discern between these two classes. We also noted that the accuracy gap between GEMSTAT-TFC and GEMSTAT narrowed (0.64 versus 0.62 in 3-way classification) when both models were augmented with accessibility data. As accessibility data by itself cannot discern between the two differentiated classes, i.e., SM and VM, it achieves a 3-way classification accuracy of 0.59, lower than the 0.64 accuracy of GEMSTAT-TFC(A).

The mesoderm specification system has been extensively studied using TF-DNA binding profiles and previous work (Zinzen et al., 2009; Peng et al., 2019) utilized ChIP-chip data for the key TFs of this system (Twi, Bin, Bap, Tin, Mef2) at five temporal stages to classify CRMs into the three muscle layers. We therefore compared the predictive accuracy of the sequence-based models (augmented with accessibility data) to the high standard set by ChIP-based models. Table 1 (second last column) shows the accuracy of a state-of-the-art XGBoost model for the same set of CRMs and classes as in our study, as reported by (Peng et al., 2019). It is worth noting that these ChIP-based XGBoost predictions were made by three separate models, one trained for each layer, whereas our sequence-based predictions were based on a single model (augmented with early and late accessibility information). Despite this, GEMSTAT-TFC(A) accuracy was close to that of ChIP-based XGBoost predictions for the binary classification task of each muscle layer.

Finally, we evaluated (Table 2) the sequence-based models on a held-out set of 69 enhancers from the CAD4 database. The models were trained using the entire set of CRMs, and we ensured that no enhancer in the held-out set overlapped with any of the training CRMs. GEMSTAT-TFC(A) exhibited an accuracy of 0.74, slightly better than that seen in LOOCV cross-validation reported above, and substantially better than GEMSTAT(A) on the held-out enhancers.

Model learns context-specific activating or repressive roles of transcription factors

In a section above, we studied motif presence in CRMs whose activities were correctly predicted (hits versus misses) to gain insights about mechanisms learned by the two GEMSTAT models, including GEMSTAT-TFC (Figure 2D). One of the most useful aspects of a mechanistic model is the assignment of regulatory roles (activating or repressive) to TFs. The models have a free parameter per TF that reflects the biochemical effect of a bound molecule of that TF and can be used to infer the TF's regulatory role. This parameter (called "alpha") captures the energy of interaction between the bound TF and the BTM; a value greater than 1 (resp. less than 1) is equivalent to an activating (resp. repressive) effect. The learned values of the alpha parameter of all TFs are shown in Table S4. However, in the GEMSTAT-TFC model this does not reveal the full picture. Because of the indirect binding of TFs, the net effect of a TF on a particular CRM in a specific muscle layer may be contrary to its biochemical effect. Our next goal was to catalog the extent of this phenomenon.

GEMSTAT and GEMSTAT-TFC were trained on the entire set of 233 CRMs, and used to predict the expression of each of those CRMs after an "in silico knockdown" where a single TF's concentration was set to zero in a muscle layer. CRMs whose expression decrease (resp. increase) after a TF's knockdown were deemed as receiving a net activating (resp. repressive) input from that TF in the particular muscle layer and are shown as points below (resp. above) the diagonal in panels of Figure 3A). These plots reveal that the same TF may have "dual roles" in the GEMSTAT-TFC framework, i.e., an activating effect on one CRM in one layer and a repressive effect on another CRM in the same or different muscle layer. Tin and Mef2 provide examples of this "dual role" phenomenon: Mef2 has an activating role in the SM layer (Figure 3A, left) and a repressive role in the VM layer (Figure 3A, middle), whereas Tin has an activating role on some CRMs and represses other CRMs in the Mesoderm layer (Figure 3A, right). We examine this phenomenon in greater detail in a later section. Roles of each TF on all CRMs across all three muscle layers are shown in aggregate in Figure 3B, with green and red indicating relative frequencies of activation and repression, and yellow representing cases where a TF's knockdown made little or no difference to predicted expression. We note that Twi and Bin have activating roles in every (or nearly every) case where they have a significant effect on a CRM. These observations are consistent between GEMSTAT and GEMSTAT-TFC. However, the latter assigns dual roles to Tin and Mef2 (as seen in Figure 3A) and to a lesser extent to Bap, whereas GEMSTAT assigns a single role (activating for Tin and Mef2, and repressive for Bap). (It is notable that GEMSTAT assigns a dual role for Mef2, with a small fraction of CRMs receiving repressive input from this activator; this arises from cooperative DNA binding by Mef2 and Tin, see Figure S2).

We surveyed the literature for published evidence of regulatory roles learned by the models, cataloging eight cases where there is a direct evidence of regulation of a gene by a TF, in several cases via characterized enhancers (Table S5). We used in silico knockdowns as above to characterize the TF's role for each of these cases. For instance, both models assign an activating role for Bin on dpp enhancers (in VM) and Twi on mef2 enhancers (in SM) (Figure 3C), consistent with the literature (Cripps et al., 1998; Zaffran et al., 2001). The models disagree regarding the role of Tin on mef2 enhancers (in VM), with GEMSTAT assigning an activating role to the TF, consistent with the literature (Cripps et al., 1999).

In three other cases examined (Mef2 → β3-tubulin in SM (Damm et al., 1998), Tin → Tin in Meso (Xu et al., 1998), Twi → Tin in Meso (Yin et al., 1997)), the role assigned to the TF by the models was in agreement with the literature (Figure S3). In two cases, the GEMSTAT-TFC model does not predict a significant role for the TF for the gene, despite literature evidence for such a role. In these cases (Bap → β3-tubulin in VM (Zaffran and Frasch, 2002), Mef2 → Act57B in VM (Kelly et al., 2002)), the mediating enhancer was not precisely delineated in the literature, so it is possible that our in-silico knockdowns were not performed with the correct enhancer or its correct region. In summary, regulatory roles predicted by GEMSTAT and GEMSTAT-TFC were consistent with the literature for 7 and 5 of the 8 cases respectively, with the only discordant prediction being the assignment (by GEMSTAT-TFC) of a repressive role to Tin on the mef2 enhancer in the VM layer, where the literature suggests an activating role (Cripps et al., 1998). We discuss this misprediction in detail in a later section, as it points to an intriguing new insight.

Modeling of indirect binding improves fits to experimental TF-DNA binding data

The results above (Table 1) indicated that explicit modeling of indirect binding improves the prediction of CRM readout. We therefore asked if the same principle leads to improved prediction of TF localization on DNA, an important open challenge in bioinformatics (Cheng et al., 2013; Ghandi et al., 2014; Zhou and Troyanskaya, 2015; Xie et al., 2019). We tested this idea using ChIP data on genome-wide TF-DNA binding of five TFs – Bap, Bin, Mef2, Tin, and Twi – that are known to be key regulators of the mesoderm specification system. (These experimental profiles were also used above in the XGBoost-based scheme to predict expression.) ChIP peaks are commonly used markers of TF-DNA binding and are also believed to frequently represent indirect binding (Gordân et al., 2009; Shokri et al., 2019).

As a baseline method, we used each TF's position weight matrix (PWM) to score each of the 233 CRMs analyzed above, by adding the strength of every putative TF binding site in the CRM (following (He et al., 2012), and compared these scores to the ChIP signal at the enhancer using the Spearman's correlation coefficient. We incorporated DNA accessibility data into predictions, adjusting each CRM's PWM score by the rank-normalized accessibility score of the CRM (as above) in stage 10-11; the ChIP data were obtained from the same developmental stage. In contrast to the baseline prediction method, the "TFC" method additionally included "indirect sites" (Figure 4A) in estimating the TF-binding score of a

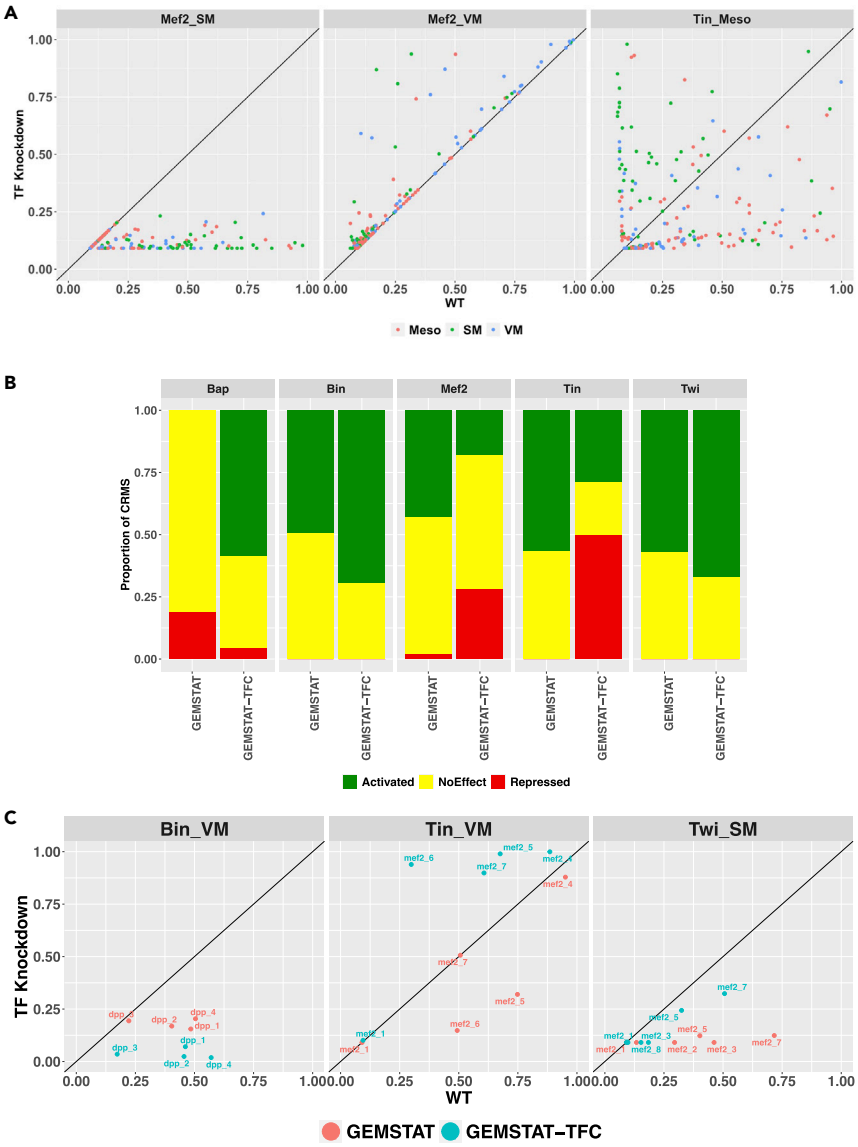


Figure 3. Predicted regulatory roles of TFs

(A) Scatterplots show CRM expression in a muscle layer, as predicted by the GEMSTAT-TFC model, under wild-type conditions ("WT", x axis) and upon *in silico* knockdown of a TF ("TF knockdown", y axis). Each point denotes a CRM and is color-coded according to the muscle layer where it is known to be active. Points above (respectively, below) diagonal indicate a repressive (respectively, activating) role for the TF. Predictions are shown for CRM expression in somatic muscle upon knockdown of Mef2 ("Mef2_SM"), in visceral muscle upon knockdown of Mef2 ("Mef2_VM") and in mesoderm upon knockdown of Tin ("Tin_Meso").

(B) For each TF, shown are the relative proportions of CRMs where the TF has an activating, repressive or neither role, as inferred by *in silico* knockdown of the TF, under the GEMSTAT and GEMSTAT-TFC models. Although some TFs, e.g., Bin, are predicted to have a predominantly activating role, there are also TFs, e.g., Mef2 and Tin, that are inferred to have dual roles – activating for some CRMs and repressive for others.

(C) Examples of activating and repressive roles of TFs in specific muscle layers. Each scatterplot shows the predicted effect of a TF (Bin, Tin, Twi) in a particular muscle layer (VM, VM and SM respectively), based on comparison of the CRM's model-predicted expression in wild-type condition and upon in-silico knockdown of the TF in that muscle layer. For each TF, results are shown for one or more CRMs associated with a gene that is reported in the literature as being regulated by that TF. Results are shown for both models (GEMSTAT-TFC in blue and GEMSTAT in red).

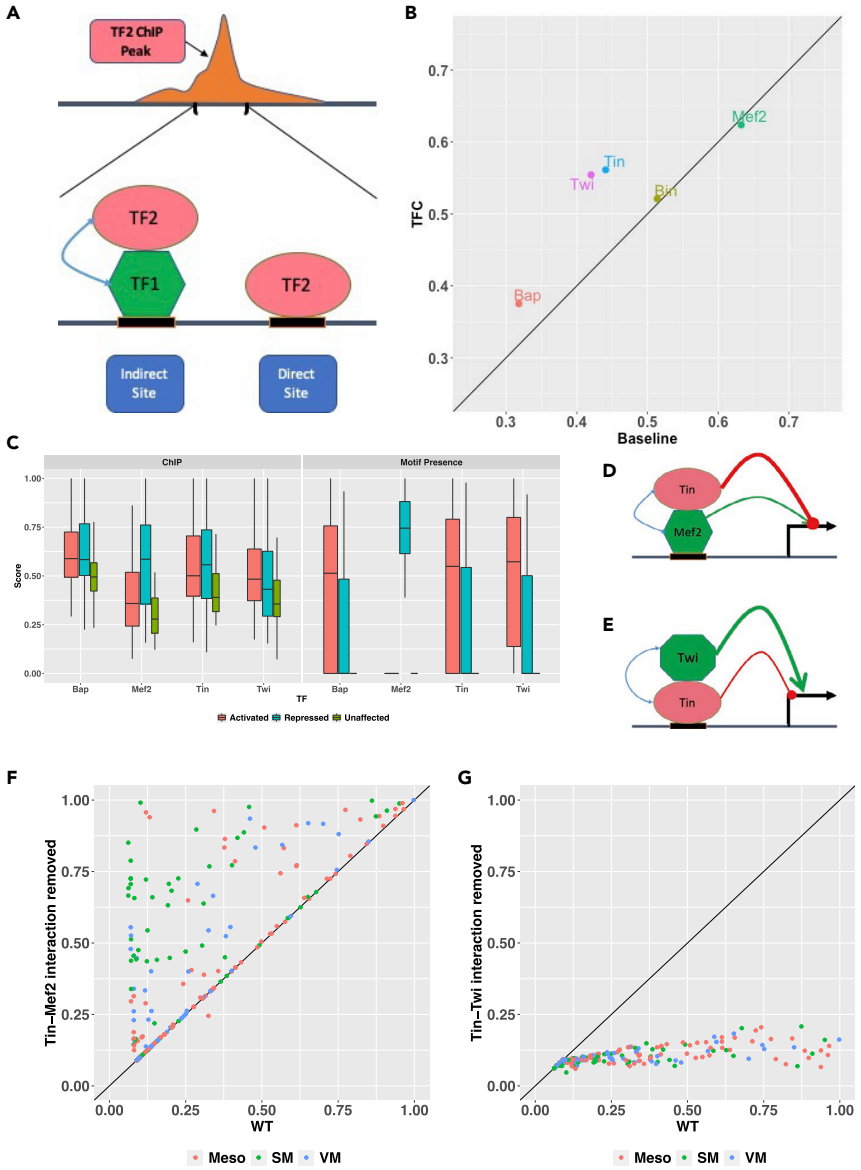


Figure 4. Model-based evidence of indirect binding

(A) In the GEMSTAT-TFC model, a TF may directly or indirectly (via another TF) bind DNA to regulate the associated gene. Thus, motif presence from this model's perspective (the "TFC method") is calculated as a sum of direct and indirect site strengths. Motif presence in the baseline method is simply the sum of direct site strengths.

(B) Spearman correlation between ChIP signal and motif presence in CRMs, for each of five TFs for which ChIP data are available, with motif presence being calculated as per either method (baseline and TFC). Each CRM's sequence-based motif presence score (rank normalized) was multiplied with its rank normalized accessibility score to get the final motif presence score of that CRM.

(C) TF binding strength as measured by ChIP data (left) and motif presence (right), for four different TFs, in CRMs divided into three classes based on GEMSTAT-TFC predictions: activated, repressed, or not regulated ("unaffected") by Tin in mesoderm. Box width represents size of (number of CRMs in) each class. For instance, CRMs predicted to be repressed by Tin have a significantly higher binding strength for Mef2, based on ChIP data and motif presence. On the other hand, CRMs activated by Tin have greater motif presence of Tin.

(D and E) A schematic that suggests how Tinman can have a dual role in Mesoderm, in a manner consistent with observations in (C).D: Tin can "piggyback" on to Mef2 through a protein-protein interaction learnt by the model, and thus indirectly bind DNA at Mef2 sites. It is also learnt to exert stronger repressive influence than Mef2's activating influence, so that the configuration shown results in overall repression, consistent with greater Mef2 presence in CRMs repressed by Tin. E: The model learns strong cooperative interaction between Twi and Tin, and Twi exerts stronger activating influence

Figure 4. *Continued*
than Tin's repressive influence. Thus, the shown configuration of Twi piggybacking on to a bound Tin molecule results in overall activating contribution, which is consistent with the greater motif presence of Tin in Tin-activated CRMs, even though this is counter-intuitive in light of Tin's repressive role learnt by the model.
(F) Effect of removal of Tin-Mef2 interaction on expression in mesoderm. Scatterplot shows model-predicted expression for each CRM in wild-type conditions (WT) versus upon removing the Tin-Mef2 interaction parameter in the model (Tin-Mef2 interaction removed).
(G) Effect of removal of Tin-Twi on expression in mesoderm.

CRM. In case of a typical binding site, the binding strength is estimated based only on the quality of match (likelihood score) between the TF's PWM and the putative site. Indirect sites, on the other hand, are putative sites of a partner TF and are scored based on (a) the match between the partner TF's PWM and the putative site and (b) the model-estimated strength of interaction between the partner TF and the TF assayed by ChIP. The TFC method scores a CRM by adding strengths of direct as well as indirect sites, and also incorporates DNA accessibility score, as in the baseline method. We found that the correlation between predicted scores and ChIP signal is conspicuously higher with the TFC method of predicting binding, for the TFs Twi and Tin (Figure 4B). The same trend was observed to a lesser extent for the TF Bap, and the two methods had comparable correlations with ChIP signal for the remaining two TFs. An example of why the TFC method is more accurate is that it predicts indirect Tin binding, via Mef2, to CRMs that lack Tin motif matches but harbor Mef2 binding sites, and if such indirect binding is indeed captured in the ChIP data. We further examine this possibility below.

The peculiar case of Tinman: a case study with GEMSTAT-TFC

The unique feature of the GEMSTAT-TFC model is that it allows for a TF to modulate the regulatory effect of a partner TF that is directly bound to DNA. This is also the basis of certain TFs exhibiting "dual roles" according to the model. The TF Tin provides an example of this phenomenon. In Figure 3A, we saw the dual role predicted for Tin in the mesoderm layer: it activates a subset of CRMs and represses other CRMs in the same layer. To understand this behavior in greater mechanistic detail, we partitioned the CRMs into three classes based on the net regulatory input from Tin –"Tin-activated" (expression reduced by at least 5% upon Tin knockdown in mesoderm), "Tin-repressed" (expression increased by at least 5% upon Tin knockdown in mesoderm) and the "Unaffected" class. We then examined the ChIP signal and motif presence of four TFs – Twi, Mef2, Bap and Tin itself – in these classes of CRMs (Figure 4C).

First, we noticed that the Tin-repressed class is enriched for Mef2 binding (ChIP) (p value 9.4E-8) as well as motif presence (p value 1.2E-29) compared to the Unaffected class; the enrichment is seen compared to the Tin-activated class as well. However, we did not see an enrichment of Tin binding in Tin-repressed versus Tin-activated CRMs, and the Tin motif is relatively under-represented (p value 6.2E-7) in the repressed class. This was surprising to us because the model learned Tin to have a repressive effect on its own (alpha parameter < 1, Table S4) and we expected Tin binding to be the most prominent feature of Tin-repressed CRMs. The enrichment we did see in Tin-repressed CRMs was that for Mef2 binding, even though Mef2 is learned to be a weak activator (alpha > 1) by the model (Table S4). Putting these observations together, we hypothesized that Tin's repressive role is mediated primarily by its indirect binding via (piggybacking on) Mef2 (Figure 4D) at the latter's binding sites. We suggest that the Tin-Mef2 complex has an overall repressive role because of Tin being stronger than that of Mef2, according to the model.

On the other hand, we found the Tin-activated class of CRMs to be enriched for present of the Tin motif (as noted above), which was unexpected for the same reason as above – that the model assigns a repressive biochemical role to this TF. This suggested that Tin forms a complex with a strong activator to exert its net activating role on these CRMs as per the model. Indeed, we observed the class to have relatively higher levels of Twi motif presence, and the model learned a strong activating role for Twi (Table S4), in agreement with the literature (Yin et al., 1997). We interpret these observations as suggesting that Tin exerts its activating effect (within the model) due to Twi piggybacking on the DNA-bound Tin and the Twi-Tin complex exhibiting a net activating role (Figure 4E).

To test the above hypotheses regarding mechanisms of Tin repression and activation in the trained GEMSTAT-TFC model, we nullified the parameter representing Tin-Mef2 interaction or Tin-Twi interaction respectively. Indeed, upon removal of Tin-Mef2 interaction the predicted expression of the Tin-repressed CRMs in mesoderm increases and on removing Tin-Twi interaction the Tin-activated CRMs exhibit lower

expression (Figures 4F and 4G), supporting the mechanistic hypotheses regarding Tin's dual role illustrated in Figures 4C and 4D. The hypotheses also explain why the Tin-repressed class has relatively low Tin motif presence compared to the Tin-activated class (noted above) even though the two classes have similar ChIP signal on average: despite the lack of Tin motif matches, this TF binds (indirectly) to CRMs of the Tin-repressed class via Mef2, whose motif is enriched in the class. Such indirect binding also suggests an explanation of the improved prediction of Tin binding (ChIP) by the TFC method (Figure 4B) that accounts for indirect binding sites.

DISCUSSION

The phenomenon of indirect binding between TF and DNA, mediated by a DNA-bound TF, has been discussed in various contexts in the literature (Gordân et al., 2009). Commonly, TF ChIP peaks are found to be lacking in footprints of the TF's direct binding in the form of motif matches (Guo and Gifford, 2017). Possible reasons for this may include inaccurate characterization of the TF's motif, presence of weak binding sites that fall below detection thresholds of motif matches but yield strong ChIP signal because of cooperative DNA binding, or indirect binding. There is also prior evidence that indirectly bound TFs within enhancers can contribute to gene regulation, especially if it is part of a "TF collective" that co-localizes on enhancers with loosely defined requirements on binding site content on those enhancers. These considerations motivated us to try to quantify the extent to which indirect TF-DNA binding influences expression levels driven by enhancers.

Our choice of the mesoderm development regulatory program was in part because of prior evidence of the TF collective phenomenon in this context (Junion et al., 2012; Khoueiry et al., 2017), and in part because spatio-temporal activity of enhancer in this program has been successfully predicted from TF-DNA binding profiles (Zinzen et al., 2009). Our work is, to the best of our knowledge, the first attempt at modeling those activities directly from enhancer sequences. We showed that the thermodynamics-based model GEMSTAT, along with a simple means to incorporate DNA accessibility information, can achieve an accuracy of 62% (Table 1) for 3-way classification of enhancer class (mesoderm, visceral muscle, or somatic muscle), a task for which the random baseline is an accuracy of ~33%. Furthermore, we implemented an important mechanistic extension to GEMSTAT by allowing indirectly bound TFs to exert regulatory influence, and the resulting GEMSTAT-TFC model improved the accuracy significantly, e.g., from 62% to 74% (Table 2) for the same 3-way classification task on held-out enhancers. The improvements were more pronounced for the VM and SM classes compared to that for the Mesoderm class. This is in part because GEMSTAT incorrectly assigns mesodermal expression to several VM/SM CRMs (due to Mef2 activation) whereas GEMSTAT-TFC avoids predicting such ectopic expression by utilizing Tin as a repressor in mesoderm.

Notably, our cross validation-based evaluations of a 2-way classification task (each muscle layer versus the other two, Table 1) revealed that GEMSTAT-TFC, based on sequence and accessibility data alone, can achieve accuracies in the range of 75%–77%, which is close to the 81%–83% accuracy achievable with a machine learning predictor using experimental (ChIP) data on the five TFs. The five TFs used in our models are the major (though not the only (Mbodj et al., 2016)) regulators of the studied enhancers, so the ChIP-based predictor using the same five TFs provides an important comparison point or upper bound for the sequence-based models.

It is worth noting that the difference between the performances of the two models becomes less pronounced once accessibility data is incorporated. This is in part because accessibility and TF-DNA binding are mutually influencing factors (Zaret and Carroll, 2011; Tsompana and Buck, 2014) and including experimental data on accessibility is expected to capture some of the regulatory events that a sequence-based model such as GEMSTAT or GEMSTAT-TFC aims to model; as such any advantage of a more mechanistically accurate model will be partly reduced when including accessibility data for both models. At the same time, we found the performance of GEMSTAT-TFC(A) to be better than that of GEMSTAT(A) on the held-out dataset of 69 enhancers from the CAD4 database (Table 2, 0.74 vs 0.62 3-way classification accuracy), supporting the advantage of the TFC model.

Examination of the GEMSTAT-TFC models revealed that regulation via indirect binding is a possible explanation of a TF may act as an activator at some enhancers while repressing others, without switching its underlying biochemical role. (In our models the biochemical role of a TF is captured by the alpha parameter, which is trained to a single value for each TF.) Such role-switching behavior has been reported in the

literature for some TFs, and mechanisms proposed have included structural alterations (Mo et al., 2004), binding site differences (Scully et al., 2000), separate protein domains for activation and repression, recruitment of co-activators or co-repressors, etcetera (Bauer et al., 2010; Boyle and Després, 2010). TF-TF interactions have also been suggested as a possible mechanism for dual role TFs, e.g., the activator Dorsal in *Drosophila. melanogaster* can act as a repressor for the gene *zen* because of physical interactions with a proximally bound repressor (Capicua) (Papagianni et al., 2018). Our analysis here suggests that TF-TF interactions may underlie dual roles of a TF even if both TFs are not DNA-bound. From the modeling perspective, this provides a simple and parsimonious explanation for dual roles, compared to a previous approach where each TF is accorded two tunable parameters, one for each direction of regulatory effect (Bauer et al., 2010).

The model's prediction of dual roles for Tin was intriguing to us. Specifically, the GEMSTAT-TFC model predicts Tin to activate mesodermal enhancers and inhibit SM enhancers in the mesoderm. Tin is well known for its activating role in the mesoderm (Xu et al., 1998; Cripps et al., 1999), but various reports of its repressive roles have also emerged (Zaffran and Frasch, 2005), without clear explanation of their mechanism. Moreover, the importance of Tin in regulating SM enhancers has also been documented (Liu et al., 2009). Closer inspection of the enhancers predicted to be activated by Tin revealed an important role for Twi in such activation whereas Tin-mediated repression was owed to indirect Tin-DNA binding via Mef2 rather than Tin repressing directly. This latter observation was also supported by ChIP data, as Mef2 binding was significantly enriched in (putative) Tin-repressed enhancers whereas the Tin motif was relatively infrequent in this class.

Our work represents a first step towards including the widely reported phenomenon of indirect TF-DNA binding into sequence-dependent models of gene regulation. This is a challenging task since sequence-dependence of the phenomenon is not characterized *a priori* through the TF's binding motif and must be discovered through the footprints of partner TFs (Guo and Gifford, 2017). At the same time, it is an important task because accurate understanding of the effects of indirect binding is necessary for mechanistic interpretation of non-coding polymorphisms. There are several interesting directions of future work that can make our proposed model more accurate without any major changes. The model currently allows a single TF to "piggyback" on a single DNA-bound TF in any given configuration. It should be worth exploring if relaxing this assumption and allowing larger complexes of interacting TFs to assemble at the binding site of any one of those TFs leads to greater explanatory power. Secondly, the current model assumes bidirectionality of indirect binding: if TF 'A' can exert regulatory influence through indirect binding via TF 'B' then 'B' can also regulate expression through indirect binding via 'A'. This assumption may not be true in practice, and future models can be used to test the assumption. Finally, it goes without saying that computational modeling of the nature presented here is meant only to generate plausible mechanistic hypotheses from the available data, and careful experimental validations are required for us to truly learn the subtle rules of cis-regulatory encoding.

Limitations of the study

This study is limited by insufficient knowledge and data availability for the regulatory network of interest. The models utilize information about the major regulators of the network, but not all known regulators. This choice was necessitated by data availability. The new model is designed to be a parsimonious implementation of the transcription factor collective phenomenon, motivated by statistical reasons rather than biochemical correctness. Evaluations of predicted transcription factor-enhancer relationships are limited by sparse direct knowledge of such relationships.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data collection
  - GEMSTAT
  - GEMSTAT-TFC

- Model training
- XGBoost model
- Model performance
- LR calculations

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104152.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (grant R35GM131819A to S.S.). We would also like to thank Charles Girardot for providing us DNA accessibility data. We are grateful to Bryan Lunt, Farzaneh Khajouei, and Pei-chen Peng for their assistance.

AUTHOR CONTRIBUTIONS

S.B. and S.S. conceived the study and designed the model. S.B. implemented the model and performed the analyses. Both authors contributed to the drafting of the manuscript and critical discussion of the results. Both authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 16, 2021
Revised: December 28, 2021
Accepted: March 21, 2022
Published: May 20, 2022

REFERENCES

Author Anonymous. (2019). 2 Chromatin patterns at transcription factor binding sites. Nature, 1. https://doi.org/10.1038/nature28171.

Arnosti, D.N., and Kulkarni, M.M. (2005). Transcriptional enhancers: intelligent enhanceosomes or flexible billboards? J. Cell Biochem. *94*, 890–898. https://doi.org/10.1002/jcb.20352.

Azpiazu, N., Lawrence, P.A., Vincent, J.P., and Frasch, M. (1996). Segmentation and specification of the Drosophila mesoderm. Genes Dev. *10*, 3183–3194. https://doi.org/10.1101/gad.10.24.3183.

Bauer, D.C., Buske, F.A., and Bailey, T.L. (2010). Dual-functioning transcription factors in the developmental gene network of Drosophila melanogaster. BMC Bioinformatics *11*, 366. https://doi.org/10.1186/1471-2105-11-366.

Berg, O.G., and von Hippel, P.H. (1987). Selection of DNA binding sites by regulatory proteins.Statistical-mechanical theory and application to operators and promoters. J. Mol. Biol. *193*, 723–750. https://doi.org/10.1016/0022-2836(87)90354-8.

Bonn, S., Zinzen, R.P., Girardot, C., Gustafson, E.H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyński, 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. https://doi.org/10.1038/ng.1064.

Bonn, S., and Furlong, E.E. (2008). cis-Regulatory networks during development: a view of Drosophila. Curr.Opin.Genet. Dev. *18*, 513–520. https://doi.org/10.1016/j.gde.2008.09.005.

Boyle, P., and Després, C. (2010). Dual-function transcription factors and their entourage. Plant Signal.Behav. *5*, 629–634.

Chen, T. and He, T.(n.d.)'xgboost: eXtreme Gradient Boosting', p. 4. https://cran.r-project.org/web/packages/xgboost/index.html.

Cheng, Q., Kazemian, M., Pham, H., Blatti, C., Celniker, S.E., Wolfe, S.A., Brodsky, M.H., and Sinha, S. (2013). Computational identification of diverse mechanisms underlying transcription factor-DNA occupancy. PLoS Genet. *9*, e1003571. https://doi.org/10.1371/journal.pgen.1003571.

Cripps, R.M., Black, B.L., Zhao, B., Lien, C.L., Schulz, R.A., and Olson, E.N. (1998). The myogenic regulatory gene Mef2 is a direct target for transcriptional activation by Twist during Drosophila myogenesis. Genes Dev. *12*, 422–434.

Cripps, R.M., Zhao, B., and Olson, E.N. (1999). Transcription of the myogenic regulatory gene Mef2 in cardiac, somatic, and visceral muscle cell lineages is regulated by a tinman-dependent core enhancer. Dev. Biol. *215*, 420–430. https://doi.org/10.1006/dbio.1999.9446.

Cusanovich, D.A., Reddington, J.P., Garfield, D.A., Daza, R.M., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H.A., Christiansen, L., Qiu, X., Steemers, F.J., et al. (2018). The cis-regulatory dynamics of embryonic development at single cell resolution. Nature *555*, 538–542. https://doi.org/10.1038/nature25981.

Damm, C., Wolk, A., Buttgereit, D., Löher, K., Wagner, E., Lilly, B., Olson, E.N., Hasenpusch-Theil, K., and Renkawitz-Pohl, R. (1998). Independent regulatory elements in the upstream region of the Drosophila beta 3 tubulin gene (beta Tub60D) guide expression in the dorsal vessel and the somatic muscles. Dev. Biol. *199*, 138–149. https://doi.org/10.1006/dbio.1998.8916.

Doitsidou, M., Flames, N., Topalidou, I., Abe, N., Felton, T., Remesal, L., Popovitchenko, T., Mann, R., Chalfie, M., and Hobert, O. (2013). A combinatorial regulatory signature controls terminal differentiation of the dopaminergic nervous system in C. elegans. Genes Dev. *27*, 1391–1405. https://doi.org/10.1101/gad.217224.113.

Gallo, S.M., Gerrard, D.T., Miner, D., Simich, M., Des Soye, B., Bergman, C.M., and Halfon, M.S. (2011). REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. Nucleic Acids Res. *39*, D118–D123. https://doi.org/10.1093/nar/gkq999.

Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M.A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput.Biol. *10*, e1003711. https://doi.org/10.1371/journal.pcbi.1003711.

Gordân, R., Hartemink, A.J., and Bulyk, M.L. (2009). Distinguishing direct versus indirect transcription factor–DNA interactions. Genome

Res. *19*, 2090–2100. https://doi.org/10.1101/gr.094144.109.

Guo, Y., and Gifford, D.K. (2017). Modular combinatorial binding among human trans-acting factors reveals direct and indirect factor binding. BMC Genomics *18*, 45. https://doi.org/10.1186/s12864-016-3434-3.

He, X., Samee, M.A., Blatti, C., and Sinha, S. (2010). Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. PLoS Comput.Biol. *6*, e1000935. https://doi.org/10.1371/journal.pcbi.1000935.

He, X., Duque, T.S., and Sinha, S. (2012). Evolutionary origins of transcription factor binding site clusters. Mol. Biol. Evol. *29*, 1059–1070. https://doi.org/10.1093/molbev/msr277.

Jakobsen, J.S., Braun, M., Astorga, J., Gustafson, E.H., Sandmann, T., Karzynski, M., Carlsson, P., and Furlong, E.E. (2007). Temporal ChIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network. Genes Dev. *21*, 2448–2460. https://doi.org/10.1101/gad.437607.

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. https://doi.org/10.1016/j.cell.2012.01.030.

Kelly, K.K., Meadows, S.M., and Cripps, R.M. (2002). Drosophila MEF2 is a direct regulator of Actin57B transcription in cardiac, skeletal, and visceral muscle lineages. Mech. Dev. *110*, 39–50. https://doi.org/10.1016/s0925-4773(01)00586-x.

Khoueiry, P., Girardot, C., Ciglar, L., Peng, P.C., Gustafson, E.H., Sinha, S., and Furlong, E.E. (2017). Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. Elife *6*, e28440. https://doi.org/10.7554/eLife.28440.

Kvon, E.Z., Kazmar, T., Stampfel, G., Yáñez-Cuna, J.O., Pagani, M., Schernhuber, K., Dickson, B.J., and Stark, A. (2014). Genome-scale functional characterization of Drosophila developmental enhancers *in vivo*. Nature *512*, 91–95. https://doi.org/10.1038/nature13395.

Liu, Y.H., Jakobsen, J.S., Valentin, G., Amarantos, I., Gilmour, D.T., and Furlong, E.E. (2009). A systematic analysis of Tinman function reveals Eya and JAK-STAT signaling as essential regulators of muscle development. Dev.Cell *16*, 280–291. https://doi.org/10.1016/j.devcel.2009.01.006.

Mbodj, A., Gustafson, E.H., Ciglar, L., Junion, G., Gonzalez, A., Girardot, C., Perrin, L., Furlong, E.E., and Thieffry, D. (2016). Qualitative dynamical modelling can formally explain mesoderm specification and predict novel developmental phenotypes. PLoS Comput.Biol. *12*, e1005073. https://doi.org/10.1371/journal.pcbi.1005073.

Mo, X., Kowenz-Leutz, E., Xu, H., and Leutz, A. (2004). Ras induces mediator complex exchange on C/EBP beta. Mol.Cell *13*, 241–250. https://doi.org/10.1016/S1097-2765(03)00521-5.

Papagianni, A., Forés, M., Shao, W., He, S., Koenecke, N., Andreu, M.J., Samper, N., Paroush, Z., González-Crespo, S., Zeitlinger, J., et al. (2018). Capicua controls Toll/IL-1 signaling targets independently of RTK regulation. Proc. Natl. Acad. Sci. U S A *115*, 1807–1812. https://doi.org/10.1073/pnas.1713930115.

Peng, P.C., Khoueiry, P., Girardot, C., Reddington, J.P., Garfield, D.A., Furlong, E.E.M., and Sinha, S. (2019). The role of chromatin accessibility in cis-regulatory evolution. Genome Biol. Evol. *11*, 1813–1828. https://doi.org/10.1093/gbe/evz103.

Peng, P.C., Hassan Samee, M.A., and Sinha, S. (2015). Incorporating chromatin accessibility data into sequence-to-expression modeling. Biophys. J. *108*, 1257–1267. https://doi.org/10.1016/j.bpj.2014.12.037.

Samee, M.A., Lim, B., Samper, N., Lu, H., Rushlow, C.A., Jiménez, G., Shvartsman, S.Y., and Sinha, S. (2015). A systematic ensemble approach to thermodynamic modeling of gene expression from sequence data. Cell Syst. *1*, 396–407. https://doi.org/10.1016/j.cels.2015.12.002.

Samee, M.A.H., Lydiard-Martin, T., Biette, K.M., Vincent, B.J., Bragdon, M.D., Eckenrode, K.B., Wunderlich, Z., Estrada, J., Sinha, S., and DePace, A.H. (2017). Quantitative measurement and thermodynamic modeling of fused enhancers support a two-tiered mechanism for interpreting regulatory DNA. Cell Rep. *21*, 236–245. https://doi.org/10.1016/j.celrep.2017.09.033.

Sandmann, T., Jensen, L.J., Jakobsen, J.S., Karzynski, M.M., Eichenlaub, M.P., Bork, P., and Furlong, E.E. (2006). A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. Dev.Cell *10*, 797–807. https://doi.org/10.1016/j.devcel.2006.04.009.

Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V., and Furlong, E.E. (2007). A core transcriptional network for early mesoderm development in Drosophila melanogaster. Genes Dev. *21*, 436–449. https://doi.org/10.1101/gad.1509007.

Scully, K.M., Jacobson, E.M., Jepsen, K., Lunyak, V., Viadiu, H., Carrière, C., Rose, D.W., Hooshmand, F., Aggarwal, A.K., and Rosenfeld, M.G. (2000). Allosteric effects of pit-1 DNA sites on long-term repression in cell type specification. Science *290*, 1127–1131. https://doi.org/10.1126/science.290.5494.1127.

Shea, M.A., and Ackers, G.K. (1985). The OR control system of bacteriophage lambda.A physical-chemical model for gene regulation. J. Mol. Biol. *181*, 211–230. https://doi.org/10.1016/0022-2836(85)90086-5.

Shokri, L., Inukai, S., Hafner, A., Weinand, K., Hens, K., Vedenko, A., Gisselbrecht, S.S., Dainese, R., Bischof, J., Furger, E., et al. (2019). A comprehensive Drosophila melanogaster transcription factor interactome. Cell Rep *27*, 955–e7. https://doi.org/10.1016/j.celrep.2019.03.071.

Tsompana, M., and Buck, M.J. (2014). Chromatin accessibility: a window into the genome. Epigenetics Chromatin *7*, 33. https://doi.org/10.1186/1756-8935-7-33.

Uhl, J.D., Zandvakili, A., and Gebelein, B. (2016). A hox transcription factor collective binds a highly conserved distal-less cis-regulatory module to generate robust transcriptional outcomes. PLoS Genet. *12*, e1005981. https://doi.org/10.1371/journal.pgen.1005981.

Xie, X., Hanson, C., and Sinha, S. (2019). Mechanistic interpretation of non-coding variants for discovering transcriptional regulators of drug response. BMC Biol. *17*, 62. https://doi.org/10.1186/s12915-019-0679-8.

Xu, X., Yin, Z., Hudson, J.B., Ferguson, E.L., and Frasch, M. (1998). Smad proteins act in combination with synergistic and antagonistic regulators to target Dpp responses to the Drosophila mesoderm. Genes Dev. *12*, 2354–2370. https://doi.org/10.1101/gad.12.15.2354.

Yáñez-Cuna, J.O., Kvon, E.Z., and Stark, A. (2013). Deciphering the transcriptional cis-regulatory code. Trends Genet. *29*, 11–22. https://doi.org/10.1016/j.tig.2012.09.007.

Yin, Z., Xu, X.L., and Frasch, M. (1997). Regulation of the twist target gene tinman by modular cis-regulatory elements during early mesoderm development. Development *124*, 4971–4982.

Zaffran, S., Küchler, A., Lee, H.H., and Frasch, M. (2001). Biniou (FoxF), a central component in a regulatory network controlling visceral mesoderm development and midgut morphogenesis in Drosophila. Genes Dev. *15*, 2900–2915. https://doi.org/10.1101/gad.917101.

Zaffran, S., and Frasch, M. (2002). The beta 3 tubulin gene is a direct target of bagpipe and biniou in the visceral mesoderm of Drosophila. Mech. Dev. *114*, 85–93. https://doi.org/10.1016/S0925-4773(02)00063-1.

Zaffran, S., and Frasch, M. (2005). The homeodomain of Tinman mediates homo- and heterodimerization of NK proteins. Biochem.Biophys. Res. Commun. *334*, 361–369. https://doi.org/10.1016/j.bbrc.2005.06.090.

Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. Genes Dev. *25*, 2227–2241. https://doi.org/10.1101/gad.176826.111.

Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934. https://doi.org/10.1038/nmeth.3547.

Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E.E. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature *462*, 65–70. https://doi.org/10.1038/nature08531.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Drosophila Genome Flybase Assembly 5 (dm3)	The FlyBase Consortium/Berkeley Drosophila Genome Project/Celera Genomics	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001215.2/
Software and algorithms		
GEMSTAT-TFC and GEMSTAT	This paper, He et al., (2010)	https://github.com/shounakbhogale/GEMSTAT-TFC
XGBOOST	Chen and He, n.d.	https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf
Other		
CAD4 database	Cusanovich et al., 2018	https://doi.org/10.1038/nature25981
Collection of 233 putative CRMs	Peng et al., 2019	https://doi.org/10.1093/gbe/evz103
Transcription Factor Expression Profile	Mbodj et al., 2016	https://doi.org/10.1371/journal.pcbi.1005073

RESOURCE AVAILABILITY

Lead contact

Further information and requests for data should be directed to and will be fulfilled by the lead contact, Dr. Saurabh Sinha (sinhas@illinois.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

All original code has been deposited at Github and is publicly available at https://github.com/shounakbhogale/GEMSTAT-TFC as of the date of publication.

METHOD DETAILS

Data collection

The set of 233 CRMs was collected from a previous study (Peng et al., 2019). These were based on ChIP-chip data on genome-wide TF binding for Bap, Bin, Mef2, Tin and Twi at multiple timepoints in *D. melanogaster* embryonic development (Zinzen et al., 2009), with CRMs being defined as the genomic regions where a cluster of multiple TF ChIP peaks was observed. The CRMs were classified into three classes – Mesoderm (Meso), Visceral Muscle (VM), and Somatic Muscle (SM) based on their activity cataloged in the CAD4 database (Cusanovich et al., 2018). (The CAD4 database compiles information from CAD (Bonn et al., 2012), RedFly enhancer database (Release 5) (Gallo et al., 2011), and data from Vienna tiling project (Kvon et al., 2014).) We also generated a subset of enhancers from CAD4 database that were active in exactly one of the 3 classes and did not overlap with the 233 CRMs. This subset was used as unseen test data for the models. Accessibility data at three different timepoints – stages 5–7, stages 10–11, and stages 13–15 – were also obtained from (Peng et al., 2019). Flybase Assembly 5 (dm3) was used to generate enhancer sequences. Each enhancer/CRM was assigned a unique accessibility score per timepoint, obtained by taking the maximum DNAse signal over the entire enhancer in the given timepoint. Information on each TF's presence (expression) or absence a layer was obtained from (Mbodj et al., 2016). A previous study (Khoueiry et al., 2017) suggests existence of protein-protein interactions involving Bap-Bin, Bap-Tin, Mef2-Tin, and Tin-Tw; we limited cooperative interactions to these TF pairs in training GEMSTAT and GEMSTAT-TFC models.

GEMSTAT

GEMSTAT (He et al., 2010) is a thermodynamics-based sequence-to-expression model based on ideas proposed by Shea and Ackers (1985). The model considers TF interactions with their cognate binding sites in an enhancer, with the transcription machinery, and with each other to predict the expression level driven by an enhancer. Its basic premise is that all regulation-related processes including TF-DNA binding, promoter-BTM (Basal Transcription Machinery) binding, and TF-BTM interaction occur under thermodynamic equilibrium. Further, it assumes that mRNA levels are directly proportional to the probability of the BTM being bound to the promoter. The model calculates this probability by considering every possible configuration of DNA-bound TFs and BTM and then calculates the total equilibrium probability of BTM-bound configurations. It further assumes that the equilibrium probability of a configuration σ, denoted by $p(\sigma)$ follows the Boltzmann distribution. Thus, $p(\sigma) = \exp(-\beta E(\sigma))/Z$ where $\beta = 1/k_B T$ (k_B is the Boltzmann constant and T is the temperature), $E(\sigma)$ denotes the energy of the configuration σ and Z is the partition function. $E(\sigma)$is calculated by adding the energies of individual interactions that occur in σ, which include the TF-DNA interactions and TF-BTM interactions. Because$p(\sigma)$ involves exponentiation of $E(\sigma)$, the additivity of energy terms amounts to a product of contributions from the interactions. The contribution of each TF-bound site S is given by $q(S) = \nu \cdot [TF] \cdot k(S_{opt}) \cdot \exp(-\beta \cdot \Delta E(S))$ where $[TF]$ is the concentration of the TF relative to some reference level ν, $k(S_{opt})$ is a free parameter representing the contribution of an optimal bound site S_{opt} and $\Delta E(S)$ denotes the mismatch energy of site S relative to S_{opt}. According to the theory by Berg and von Hippel (Berg and von Hippel, 1987), the mismatch energy of site S is related to the sequence of the site and the binding preferences of the TF by the equation $\beta \cdot \Delta E(S) = LLR(S_{opt}) - LLR(S)$ where $LLR(\cdot)$ is computed from the known position weight matrix (PWM) of the TF that binds at S and the background distribution of the nucleotides. For a given configuration σ, the unnormalized probability of the configuration when the BTM is absent is calculated by the taking the product of $q(S)$ overall occupied sites S in σ. When the BTM is present, the contribution of an individual site is augmented with TF-BTM interaction energy. Thus, $q(S)$ becomes $q(S) \cdot \alpha$ where α is the strength of interaction between BTM and the TF that binds at site S. (α>1for an activator and α<1 for a repressor, i.e., the bound TF makes the configuration more or less probable respectively.) If a TF pair A, B is pre-specified as exhibiting protein-protein interaction, a free parameter $\omega_{AB} > 0$ is introduced in the calculations, and $p(\sigma)$ is multiplied by this parameter ω_{AB} for every configuration σ where A and B are DNA-bound at proximally located sites. ($\omega_{AB} > 1$ indicates cooperative interaction and $\omega_{AB} < 1$ indicates antagonistic interaction; the model may learn $\omega_{AB} \sim 1$ if the data do not support such interaction.) In summary, the free parameters of the model include a TF-specific parameter for DNA-binding (equal to $\nu k(S_{opt})$), a TF-specific parameter for TF-BTM interaction (α), a TF-TF interaction parameter ω_{AB} for each TF pair (A, B) specified as interacting, and one additional parameter representing the basal rate of transcription. See (He et al., 2010) for further details.

GEMSTAT-TFC

In the GEMSTAT model, the cooperative interaction between two TFs occurs only when both the TFs are bound to DNA and the bound sites are located proximally. GEMSTAT-TFC model allows such TF-TF interaction even if only one of the TFs is bound to DNA and the other TF interacts with DNA indirectly through PPI with its DNA-bound partner. Specifically, we allow only one TF to "piggyback" on the DNA bound TF at a time (in a given configuration). Thus, $q(S)$ will also include terms representing interactions between the TF bound at site S and its partner TFs. So, $q_{tfc}(S) = q(S) \cdot (1 + \sum_{TF_i} \nu \cdot [TF_i] \cdot \omega_i)$ where TF_i is a partner TF and ω_i is the strength of PPI between the bound TF and TF_i. Note that the parameters used in defining the additional interactions, $\nu$and ω_i, are already part of the GEMSTAT model outlined above. However, in GEMSTAT, the concentration scaling factor ν is coupled with $k(S_{opt})$, so that $\nu k(S_{opt})$ is treated as a single free parameter. In GEMSTAT-TFC, on the other hand, we treat the concentration scaling factor ν as a separate trainable parameter (same for every TF), and $k(S_{opt})$ as the TF-specific parameter for DNA binding strength. Hence, the TFC model has one additional parameter compared to GEMSTAT.

Model training

Because our analysis includes five TFs and four cooperative interactions (Bap:Bin, Bap:Tin, Mef2-Tin, and Tin-Twi), GEMSTAT has 15 trainable parameters (5 × 2 + 4 + 1) and GEMSTAT-TFC has 16 trainable parameters. The model uses TF-specific position weight matrix (PWM) motifs to find cognate sites of the TF in an enhancer. We allow PWM matches in an enhancer with LLR score at least 0.5 times the LLR of the optimal site to be considered as binding sites. Model parameters were trained to minimize the Sum of Squared Errors (SSE) between model predictions and ground truth. Ground truth consisted of three values per

enhancer where the functional class of the enhancer had value 1 whereas the other two classes had value 0. (Note that all enhancers modeled were selected to be active in exactly one of the three classes.) The average time to train the model on 233 CRMs is about 5 min on a MacBook Air (2017) with 16gb RAM.

XGBoost model

In Peng et al. (2019), an XGBoost model in the mode logistic regression for binary classification was trained for each activity class (Meso, SM, and VM), where the positive set consists of CRMs of that particular class and the negative set comprises CRMs of the other two classes. Because we tackle a 3-way classification problem, we used prediction scores (continuous values) from the 3 models and assigned the CRM to the class with the maximum score. (This was done in a LOOCV setting.)

Model performance

GEMSTAT and GEMSTAT-TFC both predict continuous expression values. Our goal here was to classify the enhancers into one of the three classes (Meso, VM, SM). So, we assigned an enhancer to the class where it has the maximum predicted expression value. In case of a tie in predicted expression, we split the enhancer equally between the tied classes for the purpose of evaluation. For models that incorporate DNA accessibility, i.e., GEMSTAT(A) and GEMSTAT-TFC(A), predictions of expression class were made as follows. First, we multiplied the GEMSTAT or GEMSTAT-TFC prediction of expression value of the enhancer in a given muscle layer with the accessibility score of that enhancer in the layer. For this, we used early time point accessibility data for Meso and late time point accessibility for VM and SM classes; the accessibility score of an enhancer is the average accessibility of the entire enhancer in the appropriate time point, followed by a rank-normalization across all enhancers. Next, we assigned the enhancer to the class for which the accessibility-weighted expression value was the greatest among the three classes. Note that accessibility scores are class dependent (early timepoint accessibility for Meso and late timepoint accessibility for VM and SM class). The accuracy was then calculated by constructing a confusion matrix. It is worth noting that we did not use accessibility data during the training of GEMSTAT/TFC models, instead choosing to use enhancer-specific accessibility scores to up- or down-weight the predictions made by those models. A more integrative use of accessibility in the GEMSTAT framework was previously reported in (Peng et al., 2015), but we decided not to adopt that approach so as to simplify the training of GEMSTAT-TFC models and allow easier interpretations of model comparisons.

LR calculations

We used likelihood ratio (LR) score to estimate binding site strength for a TF in a DNA segment, following ChIP peaks (He et al., 2012). LR score of any segment, e.g., CRM or ChIP peak, for a particular TF is calculated by taking the sum of LR scores of all putative sites of the TF in that segment. The LR score of a site S is given by $\exp(LLR(S) - LLR(S_{opt}))$ where S_{opt} is the consensus site of the TF that binds at site S, and $LLR(s)$ denotes the commonly used log-likelihood ratio score of any site s. (Note: the sum is over putative sites defined by GEMSTAT, i.e., sites S with $LLR(S) \geq 0.5 \cdot LLR(S_{opt})$.) The LR score of a segment was rank-normalized for presentation purposes. Because the TFC model allows indirect binding of TFs, we modified the formula of a site's LR score to include the sites of partner TFs as well. Consider a TF A that piggybacks on TF B. Then the contribution of a TF B's site to the LR score for TF A will be given by $\exp(LLR(S_B) - LLR(S_{B,\,opt})) \cdot k(S_{B,\,opt}) \cdot \nu \cdot \omega(A,\ B)$ where $\omega(A,\ B)$ is the cooperative strength between TFs $A$and B, S_B denotes the binding site of TF B. Thus, the final score for TF A for enhancer e will be,

$$LR(e,\ A) = \sum_{S_A} \exp\left(LLR(S_A) - LLR\left(S_{A,\,opt}\right)\right) + \nu \cdot \sum_{B} k\left(S_{B,\,opt}\right) \cdot \omega(A,\ B) \cdot \left(\sum_{S_B} \exp\left(LLR(S_B) - LLR\left(S_{B,\,opt}\right)\right)\right)$$

where$B$ denotes all the partner TFs of A and S_{tf} denote all the sites of tf in the enhancer e.

The model dependent parameters in the above equation were obtained by training the model on the entire set of 233 putative CRMs. Further, we removed the contribution of a TF (direct or otherwise) in an enhancer if the said TF was absent in the functional class of that enhancer.