Predicting unrecognized enhancer-mediated genome topology by an ensemble machine learning model

Li Tang,[1,2] Matthew C. Hill,[3] Jun Wang,[4] Jianxin Wang,[1] James F. Martin,[2,3,5,6] and Min Li[1]

[1]*Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China;* [2]*Department of Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, Texas 77030, USA;* [3]*Program in Developmental Biology, Baylor College of Medicine, Houston, Texas 77030, USA;* [4]*Department of Pediatrics, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA;* [5]*Cardiovascular Research Institute, Baylor College of Medicine, Houston, Texas 77030, USA;* [6]*Texas Heart Institute, Houston, Texas 77030, USA*

Transcriptional enhancers commonly work over long genomic distances to precisely regulate spatiotemporal gene expression patterns. Dissecting the promoters physically contacted by these distal regulatory elements is essential for understanding developmental processes as well as the role of disease-associated risk variants. Modern proximity-ligation assays, like HiChIP and ChIA-PET, facilitate the accurate identification of long-range contacts between enhancers and promoters. However, these assays are technically challenging, expensive, and time-consuming, making it difficult to investigate enhancer topologies, especially in uncharacterized cell types. To overcome these shortcomings, we therefore designed LoopPredictor, an ensemble machine learning model, to predict genome topology for cell types which lack long-range contact maps. To enrich for functional enhancer-promoter loops over common structural genomic contacts, we trained LoopPredictor with both H3K27ac and YY1 HiChIP data. Moreover, the integration of several related multi-omics features facilitated identifying and annotating the predicted loops. LoopPredictor is able to efficiently identify cell type–specific enhancer-mediated loops, and promoter–promoter interactions, with a modest feature input requirement. Comparable to experimentally generated H3K27ac HiChIP data, we found that LoopPredictor was able to identify functional enhancer loops. Furthermore, to explore the cross-species prediction capability of LoopPredictor, we fed mouse multi-omics features into a model trained on human data and found that the predicted enhancer loops outputs were highly conserved. LoopPredictor enables the dissection of cell type–specific long-range gene regulation and can accelerate the identification of distal disease-associated risk variants.

[Supplemental material is available for this article.]

Developmental gene regulatory networks rely on *cis*-regulatory elements, like enhancers, to drive gene expression patterns in both space and time in a cell type–specific fashion. Enhancer evolution also plays an important role in driving morphological divergence (Prescott et al. 2015). Moreover, enhancers play a role in maintaining cell identity and responding to external stimuli, like injury and infection. Many enhancers work over long genomic distances through the formation of topological loops to promoters to regulate gene expression (Levine 2010). Moreover, the majority of identified disease-associated genetic variants uncovered through genome-wide association studies (GWAS) reside in noncoding intergenic regions that often can be ascribed enhancer activity. Hence, identifying the promoters looped to these variants in a cell type–specific manner is important for determining their pathological roles (Pennacchio et al. 2006; Zinzen et al. 2009; The ENCODE Project Consortium 2012).

In the past decade, high-throughput-based Chromosome Conformation Capture (3C) techniques have been developed to understand genome architecture (Dekker et al. 2002). High-throughput Chromosome Conformation Capture (Hi-C) (Rao et al. 2014) identifies physical genomic interactions in a genome-wide fashion but requires deep sequencing to achieve high resolution, which is costly and difficult to apply on a large-scale. Chromatin Interaction Analysis with Paired-End Tag sequencing (ChIA-PET) aims to detect the specific long-range interactions associated with a protein of interest (Fullwood et al. 2009). However, ChIA-PET requires a large number of cells as input (Fullwood et al. 2009). Recently, HiChIP, a protein-centric chromatin conformation method was developed, which requires lower input and also achieves a larger number of conformation-informative reads compared to traditional ChIA-PET protocols (Mumbach et al. 2016). HiChIP has been used to produce contact data for a number of key chromatin binding factors, including YY1, and cohesion (Mumbach et al. 2016, 2017; Weintraub et al. 2017). H3K27ac, an active enhancer- and promoter-associated histone mark, distinguishes active enhancers from inactivate enhancers (Heintzman et al. 2007; Creyghton et al. 2010; Rada-Iglesias et al. 2011). In addition, H3K27ac HiChIP data identifies high-confidence functional enhancer-promoter interactions (Mumbach et al. 2017). Similarly, YY1 binds to active enhancers and promoter–proximal elements and acts as a structural regulator of enhancer-promoter interactions to facilitate gene expression, making it a suitable marker for identifying distal acting enhancer-promoter

Corresponding authors: limin@mail.csu.edu.cn, jfmartin@bcm.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.264606.120.

pairs (Weintraub et al. 2017). However, HiChIP, like other 3C methodologies, still requires specialized reagents, equipment, and high-depth sequencing, making it difficult to perform on a large scale. We therefore constructed an ensemble machine learning model, LoopPredictor, to predict enhancer-mediated loops in a genome-wide fashion across different cell lines and species.

Results

Identifying active enhancer-promoter loops with H3K27ac and YY1 HiChIP

We analyzed H3K27ac HiChIP loops from K562 cells and found that the majority of loops were enhancer-mediated (for the identification of HiChIP loops and annotation of loop anchors, see Supplemental Methods), with a small percentage of promoter-promoter interactions (Fig. 1A). Next, we called super-enhancers from K562 H3K27ac ChIP-seq data and found that super-enhancers account for 5.6% of all enhancers that overlap H3K27ac HiChIP anchors (Fig. 1B; Supplemental Methods). Gene Ontology (GO) analysis of these super-enhancer anchors indicated that they contribute to the cell identity of K562 cells, confirming the suitable quality of this data set (Fig. 1C). We then carried out motif analysis on H3K27ac HiChIP loop anchors (Fig. 1D). The results indicated that the YY1 motif is significantly enriched ($-\log_2[P\text{-value}] < 100$) in loop anchors and is also highly expressed in K562 cells. We hypothesized that H3K27ac and YY1 co-occupied enhancer-promoter loops should overlap favorably

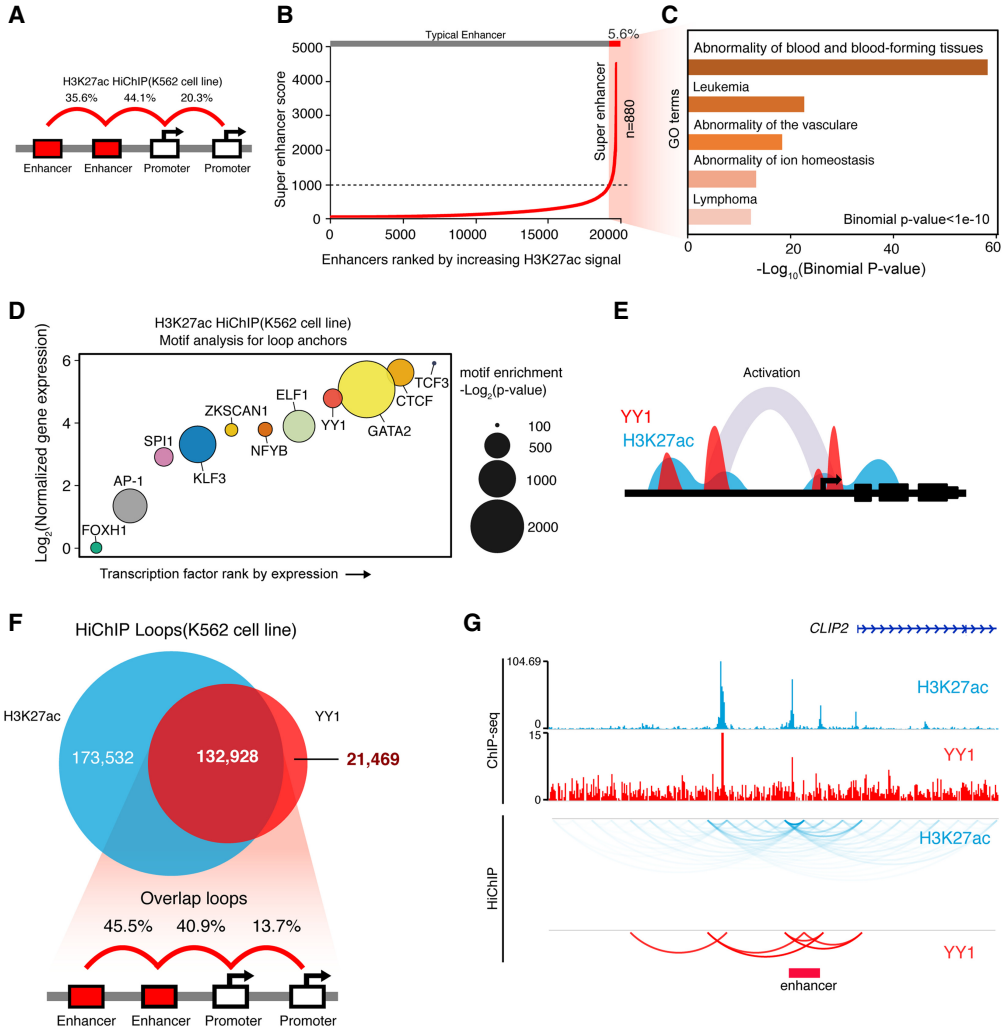

Figure 1. H3K27ac and YY1 HiChIP demarcate active enhancer loops. (*A*) Proportion of annotated loop types for K562-H3K27ac HiChIP data. Each loop identified with an FDR < 0.05 and pair-end tag number ≥ 2. ChromHMM was used to annotate the anchors, with only enhancer and promoter type anchors being retained. The majority of loops are enhancer-mediated (79.7%). (*B*) Super-enhancer plot for anchors in K562-H3K27ac HiChIP data. Slope threshold was set at 1000. A total of 880 super-enhancers were found, which accounted for 5.6% of all enhancers. Super-enhancer signal derived from H3K27ac ChIP-seq data. (*C*) GO analysis for super-enhancer anchors in K562-H3K27ac HiChIP data (binomial *P*-value < 1×10^{-10}). (*D*) De novo motif enrichment analysis on K562-H3K27ac loops. Transcription factors ranked by normalized gene expression. The size of each point indicates the motif enrichment *P*-value. The transcription factors with high motif enrichment ($-\log_2[P\text{-value}] > 100$) and gene expression are shown. The YY1 motif was significantly enriched ($-\log_2[P\text{-value}] = 306$). (*E*) Diagram depicting the putative co-enrichment of H3K27ac and YY1. (*F*) *Top*: Venn diagram showing the intersect of K562-YY1 and K562-H3K27ac HiChIP data sets. *Bottom*: Proportion of each annotated loop category from the overlapping H3K37ac and YY1 HiChIP loops ($n = 132{,}928$). The majority of the overlapping loops were enhancer-associated (86.4%). (*G*) Genome browser tracks showing H3K27ac and YY1 ChIP-seq signals and topological interactions.

and represent a high-confidence set of active distal enhancers (Fig. 1E). Indeed, the majority of YY1 HiChIP loops (86%) overlapped with H3K27ac loops. Moreover, the overlapping set of distal interactions were primarily enhancer-mediated loops (Fig. 1F). For example, the *CLIP2* locus showed similar H3K27ac and YY1 ChIP-seq profiles with both H3K27ac and YY1 enriched topological interactions present between distal elements and the promoter (Fig. 1G). Together, these results indicated that H3K27ac and YY1 HiChIP data could be combined to comprehensively characterize the full suite of highly active enhancer-promoter pairs present in a cell type of interest.

An ensemble machine learning model to predict enhancer-mediated loops

To overcome the shortcomings associated with large-scale HiChIP, Hi-C, and ChIA-PET experiments, we developed an ensemble machine learning model, named LoopPredictor, to predict enhancer-mediated loops from multi-omics features (Fig. 2A). The algorithm core of LoopPredictor was constructed, which consists of two components: Anchor type Predictor (ATP) and Confidence Predictor (CP). ATP is a minimal classifier, based on Random Forest (RF) (Breiman 2001) and multitask frameworks, that uses a minimum number of features to get optimal prediction power, and then identifies the possible conformation between genomic regions. For ATP, HiChIP loops are analyzed with ChromHMM (Ernst and Kellis 2017) to annotate the chromatin state of each anchor. The annotations of anchors were regarded as targets. We collected a variety of multi-omics data for the feature generator, which used a standard scaler for normalization and batch effect removal (for the procedures relating to multi-omics data sets and feature generation, see Supplemental Methods). After training, data sets of interest are used as input (e.g., H3K27ac ChIP-seq peaks) into the ATP model to generate the possible anchor pairs and predict the

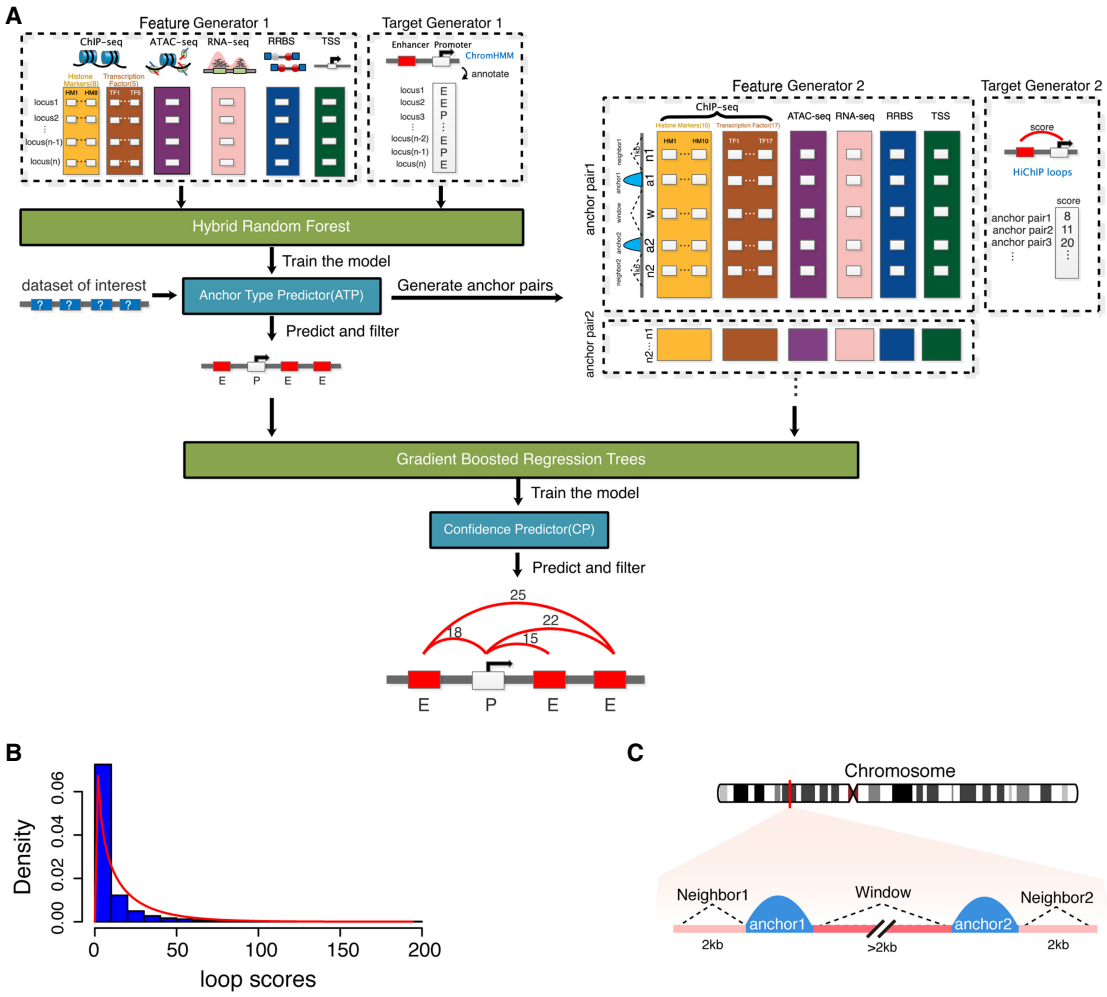

Figure 2. LoopPredictor, an ensemble machine learning model. (*A*) The LoopPredictor algorithm. H3K27ac and YY1 HiChIP data sets and multi-omics features (e.g., ChIP-seq, RNA-seq, ATAC-seq, and RRBS) were first processed and then integrated to train the model. Targets were defined from the extracted HiChIP anchors via ChromHMM annotation. Next, the trained model and the functional genomics data of interest get put through Anchor type Predictor (ATP), which then identifies the putative topological interactions existing between active genomic regions. The anchor type output from ATP, the newly generated features, and targets of anchor pairs (enhancers and promoters) then get imported into our Confidence Predictor (CP) following Gradient Boosted Regression Trees (GBRTs)-based training. Finally, CP assigns a confidence metric to each predicted chromatin loop, which can be utilized for the filtering of the final LoopPredictor output (Supplemental Methods). (*B*) Distribution of general HiChIP loop scores after merging four HiChIP data sets (K562-H3K27ac, K562-YY1, HCT116-YY1, and GM12878-H3K27ac). (*C*) Diagram depicting the loop-associated regions used to gather features in Confidence Predictor.

type of anchors for loop prediction (for the preparation of training sample, see Supplemental Methods).

The second component of the algorithm core is CP, which is a powerful regressor based on Gradient Boosted Regression Trees (GBRT). The possible conformation and the corresponding loop type generated by ATP were imported into CP to predict the confidence level as a loop score. We found that the scores of loops obey a gamma distribution (Fig. 2B), so we used the density function to identify high-confidence loops and then normalized the scores for target generator. Next, we integrated more features for CP from different genomic scales, including the flanking regions of two anchors, the distance between two anchors (window), and the neighboring regions outside of each anchor (Fig. 2C; Supplemental Methods). The anchor type from ATP and the newly generated features and targets of anchor pairs were imported into GBRT for training. Finally, we used CP to predict chromatin loops with scores indicating the confidence of the topological interaction.

The performance of Anchor type Predictor

Before determining which classifier to be used in ATP, we tested the F1-score of four standard classifiers: LinearSVC (Pedregosa et al. 2012), LogisticRegression (Pedregosa et al. 2012; Schmidt et al. 2017), KNeighbors (Pedregosa et al. 2012), and Random Forest (Breiman 2001; Pedregosa et al. 2012). The classifiers were trained with four different HiChIP data sets (K562-YY1, K562-H3K27ac, HCT116-YY1, and GM12878-H3K27ac). The evaluation of the F1-score and precision-recall rate showed that Random Forest outperforms other standard methods (Fig. 3A; Supplemental Fig. S1). Similarly, the Receiver Operating Characteristic (ROC) curves (Fan et al. 2006) of these four classifiers in K562-YY1 HiChIP data indicated that RF achieved the best results in the prediction of different kinds of loops (promoter–enhancer, promoter–promoter, enhancer–enhancer) (Fig. 3B) (for ROC curve evaluation, see Supplemental Methods). Therefore, we integrated RF into ATP to generate the possible anchor pairs and predict the type of anchors.

To obtain the minimum input for optimal performance of ATP, we tested the F1-score with an increasing number of features (total $n = 24$). A feature is a multi-omics data set (e.g., H3K4me1 ChIP-seq). The performance with 12 features was close to optimal, so we chose this number of features to use as input for RF (Fig. 3C). To determine the correlation between features, Pearson's correlation and hierarchical clustering analysis were used for all 24 features (Fig. 3D; Supplemental Methods). We found that ELF1 ChIP-seq was highly correlated with YY1. ELF1 is a lymphoid transcription factor known to regulate the expression of MEIS1 (in K562 cells), another transcriptional master-regulator associated with leukemic hematopoiesis whose motif was found to be enriched in H3K27ac anchors (Xiang et al. 2010). Analysis of YY1 and ELF1 ChIP-seq data confirmed the colocalization of the two factors (Fig. 3E,F; Supplemental Methods). Thus, the correlations uncovered here are likely to be functionally relevant. The feature correlations of all HiChIP data sets were similarly clustered (Supplemental Fig. S2). We then investigated the importance of our top 12 ranking features from the K562-YY1 HiChIP data set using a fivefold cross-validation procedure. The data showed that the H3K27ac ChIP-seq signal within two anchors was the most important feature, followed by the distance from anchors to transcription start sites (TSSs) and chromatin accessibility (Fig. 3G). The feature importance of other HiChIP data sets was similar to K562-YY1 HiChIP (Supplemental Fig. S2).

The performance of Confidence Predictor

To characterize the general profile of all the features in different loop-associated regions, we quantified each individual feature signal on a *z*-score normalized scale (Jain et al. 2005) (for the quantification of features, see Supplemental Methods). The feature signal of the inter-anchor window region was highest and most variable, while the outer neighbor regions presented the lowest intensity (Fig. 4A). Here, we trained CP with four individual HiChIP data sets and four integrated data sets. To interpret the contribution and correlation of features in the prediction, we calculated an importance score for each feature with fivefold cross-validation, and then filtered the features whose importance score was greater than 0.001 for Pearson's correlation analysis and hierarchical clustering (Fig. 4B; Supplemental Fig. S3). The data showed that features are correlated well by loop-associated regions, and the size of window was the most important factor for the prediction, which was consistent across the cell lines analyzed.

To evaluate the performance of CP, we calculated the adjusted *R*-square value and mean absolute error (MAE) for different prediction cases, and then assessed actual values versus predicted observations (Fig. 4C; Supplemental Methods). For the prediction of four individual data sets, CP achieved an adjusted *R*-square from 0.72 to 0.77, while for the integrated data sets, the adjusted *R*-square values of CP were all larger than 0.85. Moreover, the integration of K562* (YY1 + H3K27ac), GM12878, and HCT116 outperformed the others (Fig. 4C). Specifically, the distribution of actual loop scores and predicted loop scores are consistent (Supplemental Fig. S4A–D). These results suggest that the integrated data sets are more favorable for the training of CP.

We next assembled the ATP and CP modules together into an adaptable model, which were trained with the integrated HiChIP data sets. Next, the adaptable model and multi-omics features from different cell types were fed into LoopPredictor to predict enhancer-mediated interactions. One concern with utilizing our multicellular multi-omics trained adaptive model is a loss of cell type–specific loops and a potential enrichment for common regulatory genomic interactions. To evaluate the performance of our adaptive training model for predicting cell type–specific observations, we fed multi-omics features from three different cell lines separately (Fig. 4D). We identified thousands of unique loops for each input cell line. To determine the regulatory characteristics of these unique loops, we extracted the predicted anchors and overlapped them with cell line-specific accessible chromatin peaks (ATAC-seq and DNase-seq) (Fig. 4E). The highest enriched motifs from these three trimmed anchor sets were extracted and ranked by gene expression to produce a list of cell line-specific transcription factors (Fig. 4F). Indeed, we identified GATA1 activity in K562 cells, enrichment for IRF factors in GM1212878 loops, and NRF2 binding in HCT116 cells, consistent with the literature (Huang et al. 2004; Li et al. 2009; Mariani et al. 2017). Hence, our comprehensive adaptive training model is a powerful tool for predicting cell type–specific enhancer-promoter loops.

Functional validation of predicted enhancer-mediated interactions

To investigate the degree to which the predicted loops output from LoopPredictor matched with experimentally measured enhancer-promoter loops, we compared the predicted loops from K562 cells with H3K27ac HiChIP performed in K562 cells. Only 4.1% of all loops were K562-H3K27ac HiChIP-specific (Fig. 5A). To investigate the distribution of loops by distance, we binned the loops by 100-

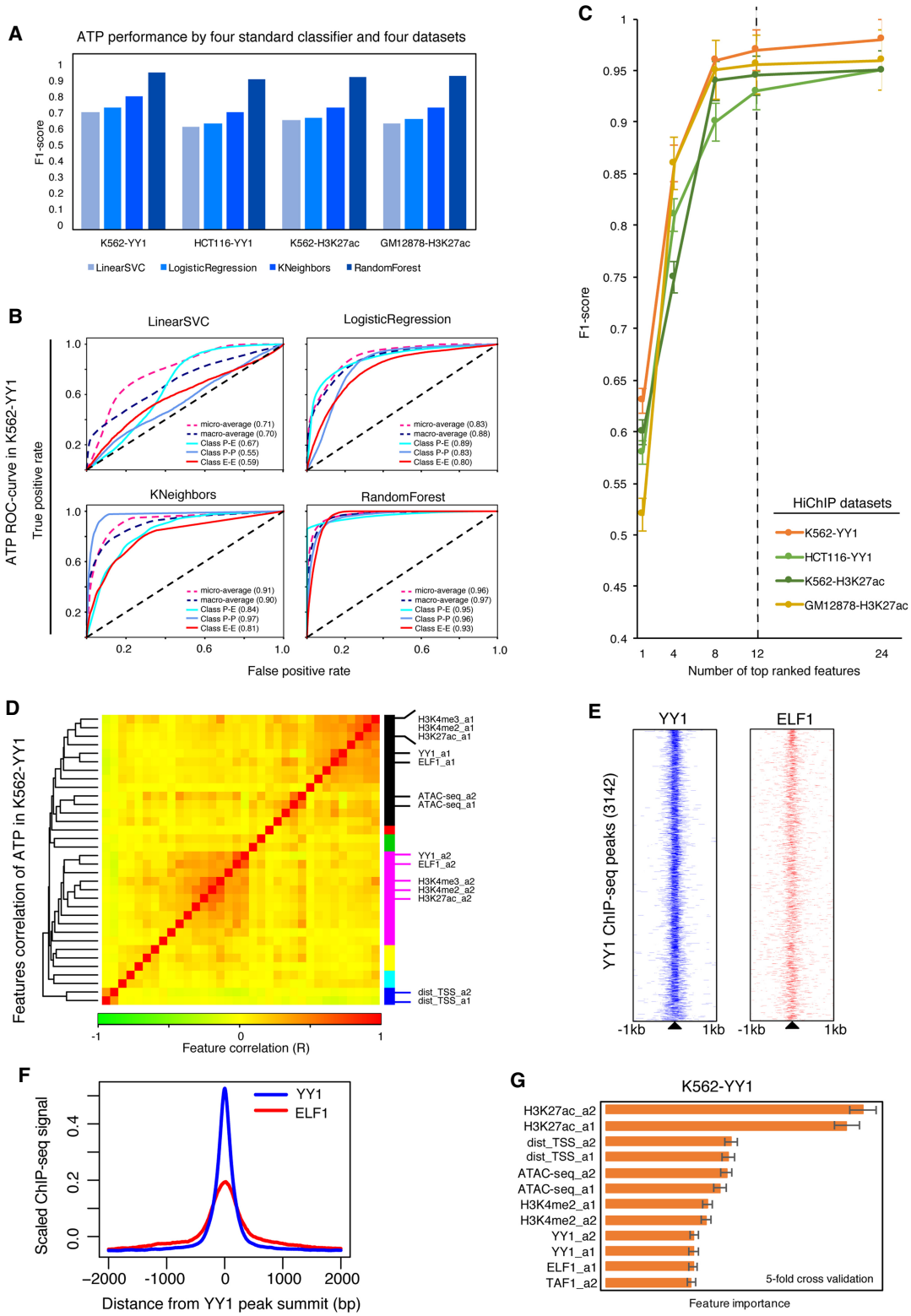

Figure 3. The performance of Anchor type Predictor. (*A*) F1-score of ATP evaluated using four standard classifiers across four HiChIP data sets. (*B*) ROC curves of ATP in K562-YY1 HiChIP data (Supplemental Methods). (*C*) The F1-score performance of ATP in four HiChIP data sets with increasing number of top-ranked features by fivefold cross-validation. The vertical line indicates point at which ATP achieved close to optimal performance with a modest input of 12 features. (*D*) Pearson's correlation combined with hierarchical clustering for K562 cell features. The colored bars on the *right* side indicate the identity of each hierarchical cluster. Colored bars at the *bottom* mark the feature correlation coefficient, R. (*E*) Heat map displaying YY1 and ELF1 ChIP-seq signals across YY1 peaks ($n = 3142$). (*F*) Comparison of YY1 and ELF1 ChIP-seq peaks signals by distance from YY1 peak summit. (*G*) Feature importance for the top 12 features in K562-YY1 HiChIP data set with fivefold cross-validation. Error bars indicate the standard deviation.

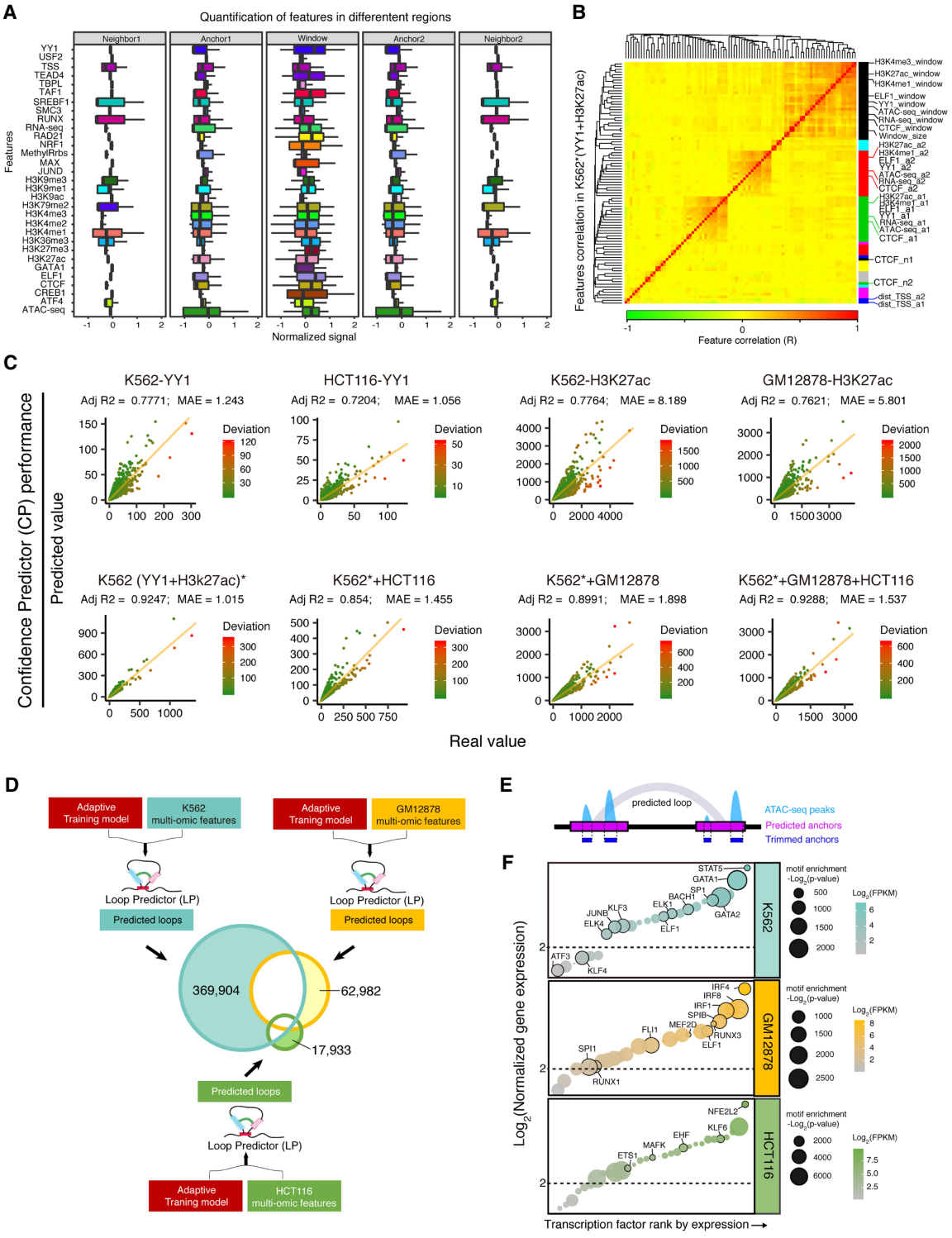

Figure 4. The performance of Confidence Predictor. (*A*) Quantification of features across different loop-associated regions. The signals of features were normalized by *z*-score. (*B*) Pearson's correlation and hierarchical clustering for the features of K562*(YY1 + H3K27ac). The colored bars on the *right* side indicate hierarchical clusters. Colored bars at the *bottom* demarcate the feature correlation coefficient, R. (*C*) Prediction performance evaluation for CP in four individual data sets (*upper*) and four integrated data sets (*lower*). (*D*) Evaluation of LoopPredictor for identifying cell type–specific loops. Our adaptive model and the multi-omics features derived from three individual cell lines were fed into LoopPredictor to perform predictions. The results from each cell type were marked by different colors in the Venn diagram. The number of predicted cell type–specific (K562, GM12878, HCT116) loops were 369,904, 62,982, and 17,933, respectively. (*E*) Diagram for trimming anchors with ATAC-seq peaks to identify loop binding transcription factors. (*F*) Transcription factor enrichment analysis for cell type–specific predicted loops. Cell type–specific loops identified in *D*. Transcription factors were ranked by normalized gene expression. The size of each point indicates the motif enrichment *P*-value. The color of each point codes for the normalized expression of the indicated transcription factor. The threshold of motif enrichment was $-\log_2(P\text{-value}) > 500$, and the threshold for gene expression was set to $\log_2(\text{FPKM}) > 2$.

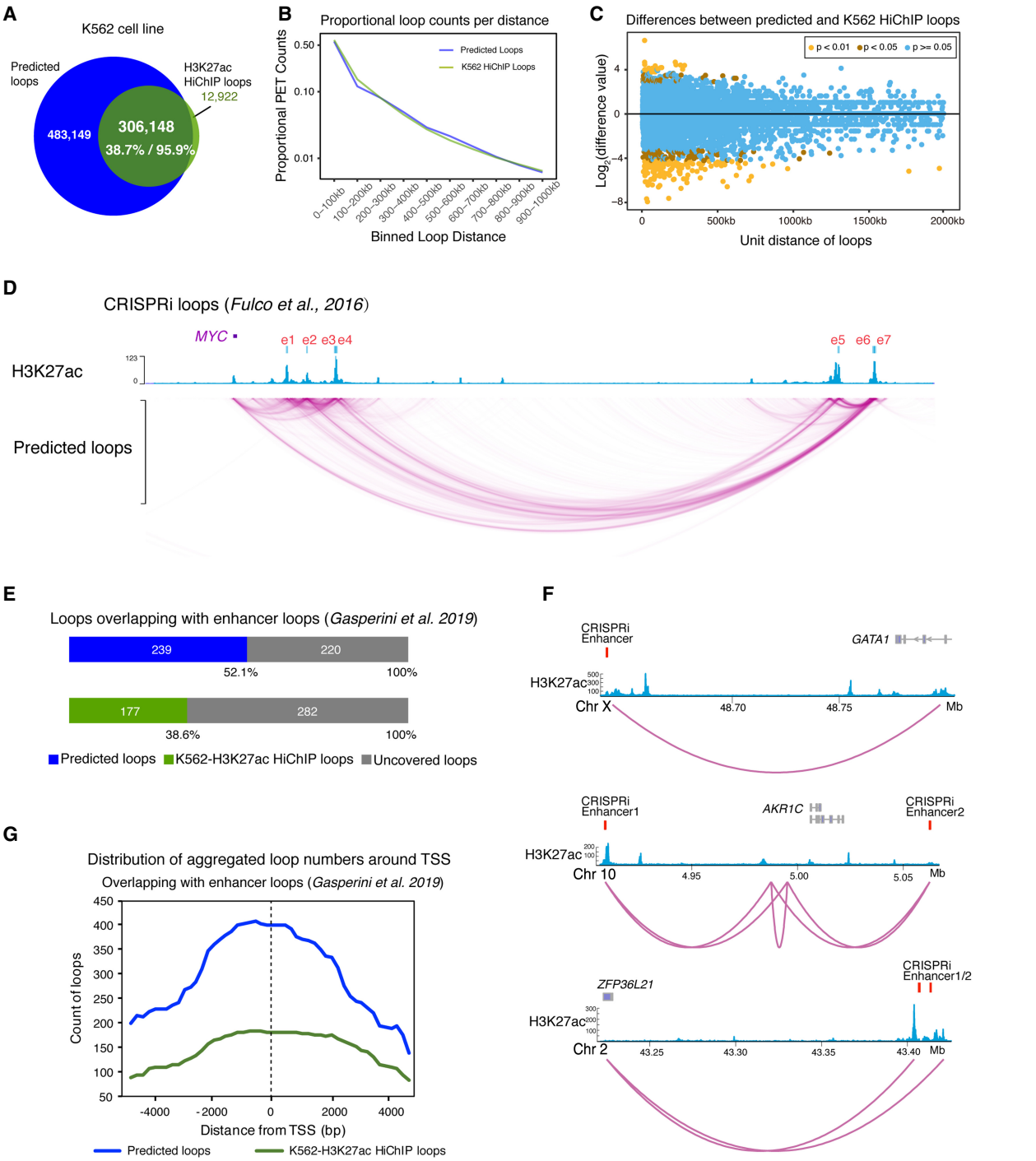

Figure 5. Functional validation of predicted enhancer-mediated interactions. (*A*) Venn diagram showing the overlap of K562 predicted loops and K562-H3K27ac HiChIP loops. Overall, 306,148 loops were detected by both LoopPredictor and HiChIP experiments, which accounted for 95.9% of the HiChIP loops; 4.1% (12,922 loops) were HiChIP-specific loops. (*B*) Proportional loop counts per distance for predicted and K562-H3K27ac HiChIP loops. Loops were binned by 100 kb to calculate the proportion. (*C*) Differences between predicted and K562-H3K27ac HiChIP loops. Loops with a *P*-value ≥ 0.05 (blue dots) were classified as nonsignificant, loops with a *P*-value < 0.05 (brown dots) were labeled significant, and differences with a *P*-value < 0.01 (yellow dots) were marked highly significant. The vast majority of loops showed no significant differences between the two sets of loops. (*D*) Validation of predicted loops by focused CRISPRi integration (Fulco et al. 2016). Seven previously validated *MYC* enhancers with strong H3K27ac ChIP-seq signals (blue track) were annotated as e1 through e7 (red track). The predicted loops contacted these published CRISPRi loops. (*E*) Validation of predicted loops by high-throughput CRISPRi screening integration (Gasperini et al. 2019). A total of 459 high-confidence gene-enhancer loop pairs were overlapped with predicted loops as well as H3K27ac loops. In total, 52% of the high-confidence loop pairs were identified by LoopPredictor. Only 38% of these high-confidence loop pairs were recovered from K562-H3K27ac HiChIP loops. (*F*) Genome browser tracks for validated enhancer loops identified in *E*. Validated CRISPRi high-confidence pairs are shown in red; H3K27ac ChIP-seq in blue; predicted loops in purple. (*G*) Promoter contacts of CRISPRi loops for predicted loops and H3K27ac HiChIP loops. Distribution of aggregated loop numbers by distance of loops from *E*. The distribution was calculated by ± 4-kb distance from TSS.

kb windows. The resulting loop proportions were consistent between the predicted universal interactions and the observed H3K27ac HiChIP interactions (Fig. 5B). Differential analysis of loop scores also indicated that most loops were not significantly different (Fig. 5C). Moreover, we found that predicted loops output from LoopPredictor behaved similarly with respect to topologically associated domain (TAD) boundaries when compared to HiChIP-derived loops (Supplemental Fig. S4E; Supplemental Methods).

We next wanted to validate our predicted loops by comparing them to functionally validated enhancer-promoter pairs identified in K562 cells. Previously, 7 *MYC* enhancers were identified via a systematic CRISPR interference (CRISPRi) screen, which were annotated as e1 through e7 (Fulco et al. 2016). We found that the predicted loops output from LoopPredictor proximal to *MYC* were in accordance with these published loops (Fig. 5D). Recently, 664 enhancer-gene loops were identified from a large-scale multiplex enhancer-gene pair screening effort in K562 cells (Gasperini et al. 2019). From this study, we identified the high-confidence enhancer-gene loops (n = 470 pairs) for comparison with our predicted K562 loops. Fifty-two percent of the functionally validated enhancer loops overlapped with our predicted loops, compared to just 38% overlap with K562-H3K27ac HiChIP loops (Fig. 5E,F; (Mumbach et al. 2017). Within these overlapping loops, the predicted observations had stronger enrichment of loop counts proximal to the TSSs compared to H3K27ac HiChIP loops (Fig. 5G). Hence, LoopPredictor is capable of predicting functional enhancer-gene loops with high sensitivity.

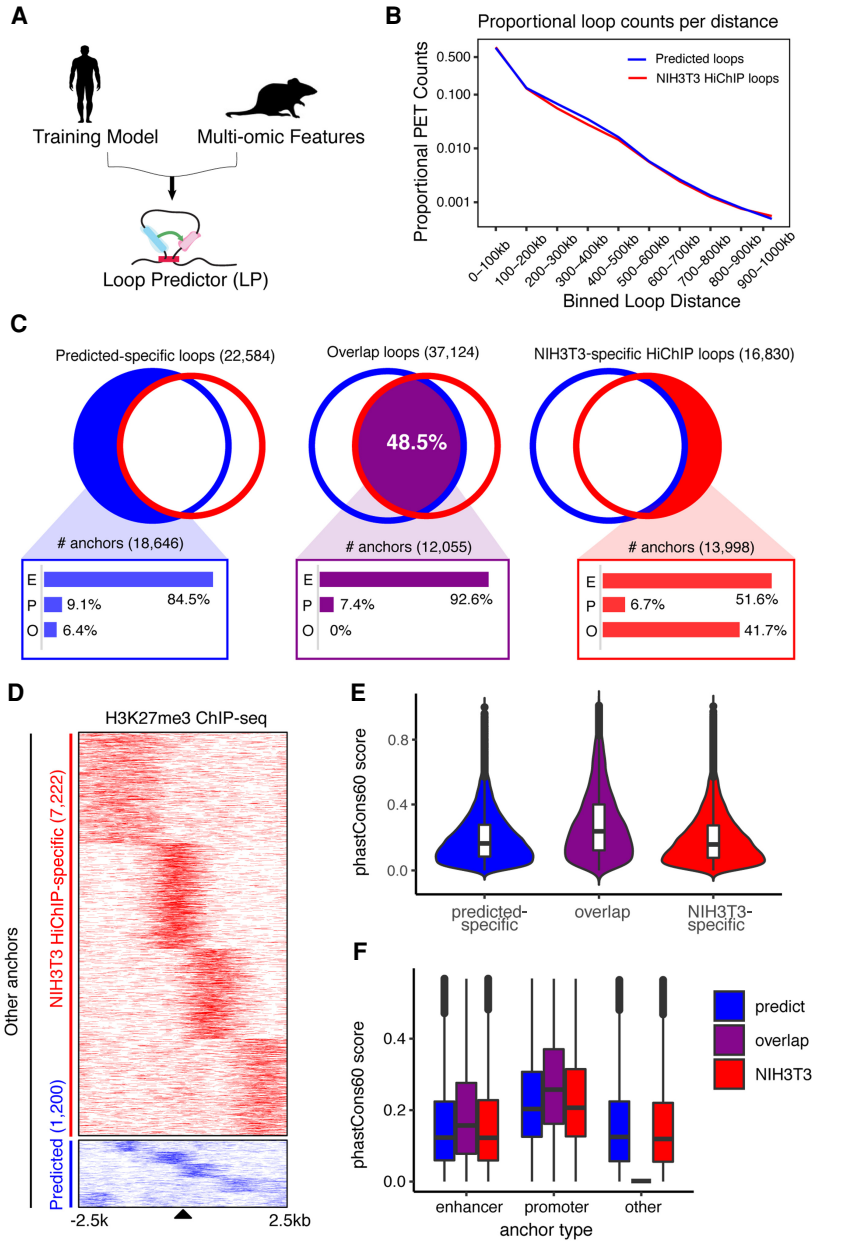

Figure 6. Cross-species long-range chromatin loop predictions. (*A*) Diagram for cross-species chromatin predictions. The adaptive training model implemented previously and trained on human cell lines was combined with multi-omics data from murine NIH3T3 myofibroblasts as input for LoopPredictor. (*B*) Proportional loop counts per distance for predicted and NIH3T3 HiChIP loops. Loops were binned by 100 kb to calculate the proportion. (*C*) Venn diagrams depicting the amounts of predicted-specific loops (22,584), overlapping loops (37,124), and NIH3T3-specific HiChIP loops (16,830). The common loops accounted for 48.5% of the total loops. The bar plots shown *below* indicate the composition of anchor types from each loop set. (E) Enhancer type, (P) promoter type (P), (O) other type. (*D*) Heat map for H3K27me3 ChIP-seq signal over O-type anchors in NIH3T3 HiChIP-specific and predicted-specific loops found in *C*. The number of O-type anchors in NIH3T3 HiChIP-specific and predicted-specific sets are 7222 and 1200, respectively. ChIP-seq signals were visualized across a 5-kb window. (*E*) Violin plots of conservation scores across anchor types. Anchor regions derived from *C*. (*F*) Boxplot for conservation scores by anchor type derived from *C*.

Predicting chromatin interactions in a model organism

LoopPredictor can predict the topological interactions for any cell type which lacks 3D genomic information and, because it is trained on highly conserved mammalian gene regulatory features, it should also be able to predict enhancer-promoter interactions for other mammalian species (Cheng et al. 2014). We gathered multi-omics features from the murine NIH3T3 myofibroblasts cell line and the aforementioned adaptable model trained on human cell lines to feed into LoopPredictor (Fig. 6A). After the prediction, we obtained 59,708 loops as output, the proportional loop counts by distance showed a high coincidence between the predicted loops and published NIH3T3 HiChIP loops (Fig. 6B; Xiao et al. 2019), and the differential analysis showed that most of the loops had no significant difference (Supplemental Fig. S4F,G). To interpret the component

differences between predicted loops and NIH3T3 HiChIP loops, we annotated the anchor types by ChIP-seq data and the distance from anchors to TSSs (Fig. 6C). The anchors were classified into three types: Enhancers (E), Promoters (P), and Other (O). Annotation results showed that enhancer anchors accounted for the majority of anchors in all loop sets. In addition, the overlapping HiChIP and predicted loop anchors were comprised entirely of enhancer and promoters, while there were 41.7% O-type loops in the NIH3T3-specific HiChIP loops, which was higher than the LoopPredictor-specific loops (6.4%) (Fig. 6C). We hypothesized that NIH3T3-specific non-enhancer-promoter anchors may lie within inactive heterochromatic regions, so we analyzed NIH3T3 H3K27me3 CUT&RUN data over these O-type anchors (Fig. 6D). NIH3T3 HiChIP-specific and LoopPredictor-specific O-type loops both displayed high H3K27me3 signals, suggesting that these anchors primarily lie in inactive genomic regions. Thus, the loops output from LoopPredictor are primarily active loops with decreased heterochromatin composition compared to H3K27ac HiChIP-specific loops.

As we conducted a cross-species comparison, we next investigated the degree of conservation between human and mouse over the LoopPredictor-specific, NIH3T3 HiChIP-specific, and overlapping loops (Fig. 6E). The mean conservation score of overlapping loop anchors was greatest among the three groups (Fig. 6E) (for sequence conservation analysis, see Supplemental Methods). Moreover, the promoter anchors were the most conserved anchor type (Fig. 6F). This is consistent with previous large-scale regulatory studies, which found that the binding of orthologous transcription factors to promoters is more highly conserved than binding to distal regulatory sequences (Cheng et al. 2014). Thus, LoopPredictor is capable of performing cross-species loop predictions with improved sensitivity over running H3K27ac HiChIP alone.

LoopPredictor compares favorably with other loop prediction tools

With the rapid increases in genomics data, several computational methods have been developed for predicting enhancer-gene interactions. TargetFinder is a reliable machine learning-based method for reconstructing topological loops (Whalen et al. 2016). We compared LoopPredictor with TargetFinder, using the same data sets. The results showed LoopPredictor achieved lower recall scores but higher accuracy and precision scores (Supplemental Fig. S5A). A newer algorithm, known as 3DPredictor, predicts chromatin interaction frequencies using gene expression and CTCF-binding information (Belokopytova et al. 2020). To compare LoopPredictor with 3DPredictor fairly, we tested the performance of two different parametric modes of CP; the regression results showed that the ensemble CP achieved the highest adjusted *R*-square value, and MAE (Supplemental Fig. S5B). To further evaluate the robustness of LoopPredictor, we trained LoopPredictor and 3DPredictor with the same histone mark ChIP-seq and chromatin accessibility data sets derived from H9 human embryonic stem cells. Next, we wanted to assess ability of each set of predicted loops to identify active embryonic regulatory elements. We then overlaid the predicted loops with 32,353 highly active embryonic stem cell enhancers identified via ChIP-STARR-seq (Barakat et al. 2018) and found that LoopPredictor could predict more embryonic enhancer-mediated loops (39.3%) than 3DPredictor (10.5%) (Supplemental Fig. S5C).

Discussion

Enhancer-mediated interactions play important roles in gene expression, evolution, disease, and development. Currently, it is a significant challenge to investigate the genome topology for all cell types across species. Therefore, we generated LoopPredictor, an ensemble machine learning model to predict the genome topology for any cell type which lacks a 3D profile. LoopPredictor incorporates H3K27ac/YY1 HiChIP data sets and an assembly of multi-omics features to learn active long-range enhancer-mediated looping characteristics. Users need only to provide multi-omics data sets as input, alongside our adaptive training model, into LoopPredictor to generate a list of predicted loops ranked by confidence and comprehensively annotated.

The adaptive training model we incorporated for predicting loops was generated using many multi-omics data sets derived from several distinct cell lines. Despite the diverse cellular input underlying our model, we were able to isolate cell type–specific gene regulatory networks from among three different cancer cell lines. The vast complexity of diverse cancers is well-known. Efforts in the field of cancer genomics have already been directed toward incorporating multi-omics data sets and whole-genome sequencing to help predict clinical phenotypes for the purposes of personalized medicine (Wang et al. 2015). Knowledge of the gene regulatory networks active in a certain cancer cell would inform clinicians about drug resistance, genomic instability, EMT status, and/or metastatic properties. Future work incorporating such approaches as the Cancer Hallmark Network with LoopPredictor may improve our understanding of the gene regulatory networks driving tumorigenesis and improve future personalized medical analysis.

Sophisticated computational methods, like TargetFinder, incorporate diverse multi-omics features to predict enhancer-promoter interactions with improved accuracy compared to using only the closest gene (Whalen et al. 2016; Hong et al. 2019; Moore et al. 2020). Indeed, the predictive importance of genomics features output from TargetFinder were similar when directly compared with LoopPredictor. One important distinction between TargetFinder and LoopPredictor is the use of Hi-C data as opposed to H3K27ac- and YY1-HiChIP data, respectively. The majority of loops identified from Hi-C data are thought to be large structural CTCF and cohesion anchored loops (Mumbach et al. 2016). However, the HiChIP loops incorporated into LoopPredictor are enriched for YY1- and H3K27ac-bound regions and should thus have a higher proportion of active enhancer-mediated loops and subsequently a decreased amount of background CTCF-associated interactions (Mumbach et al. 2016). The differences in 3C-based technologies incorporated may explain why LoopPredictor's near-optimal performance was achieved with 12 features, while TargetFinder required approximately 16. Further, our model offers several other advantages over existing predictive tools, including the ability to detect enhancer-enhancer, enhancer-promoter, and promoter-promoter interactions on a genome-wide scale.

The ensemble model implemented by LoopPredictor consists of two core components, ATP and CP. For the classification step performed by ATP, we tested performance with the F1-score for different cell lines and different classifiers, and the results indicated that Random Forest performed the best in all data sets. We also found that HiChIP data quality is crucial for the optimal performance of LoopPredictor, as HiChIP data sets with greater ChIP efficiencies produced the best results. The classification capability of ATP was optimal with an input of 24 features; however, 12 features

is recommended for standard use. In addition, while any type of multi-omics data can be used as features, LoopPredictor performs best when the input data sets are more representative of active transcription (e.g., RNA-seq, ATAC-seq, H3K27ac, H3K4me2, H3K4me1, and transcription factor ChIP-seq). Input feature series with a higher composition of heterochromatin-associated data sets will not perform as well with our adaptive model (Supplemental Fig. S6A,B). The normalized quantification results of features at different regions indicated the inter-anchor window region was most informative, as mentioned above, while the signal derived from the outer anchor regions was lowest. Moreover, most of the features were correlated well by window or anchor regions, while the features of two neighbors were scattered with no obvious association. The enrichment for feature signal between anchors is somewhat to be expected since the genomic regions between an active regulatory loop may contain more active enhancers, bound transcription factors, and highly accessible gene promoters. In contrast, genomic regions which are outside of anchors will more likely contain heterochromatic regions and 3′ genic regions. Future studies aimed at dissecting the components of active enhancer-promoter loops could benefit from performing a similar analysis and assessing these regions individually.

LoopPredictor has the ability to identify functional 3D enhancer loops. Here, we found that, after training our adaptive model and inputting several multi-omics features derived from K562 cells, LoopPredictor predicted loops that were highly consistent with a published set of H3K27ac HiChIP loops derived from K562 cells. From this predicted loop set, we found an overlap for distal regulatory interactions between the *MYC* locus and seven enhancers which have been previously validated via CRISPRi screening (Fulco et al. 2016). Moreover, a high-throughput gene-enhancer pair screen performed in K562 cells identified several hundred high-confidence enhancer pairs which overlapped more favorably with the predicted loops compared to the experimental H3K27ac HiChIP loops (Mumbach et al. 2017). An explanation for these differences may be that the predicted loops, which are the result of a set of several combined functional genomic features and HiChIP loops, may be more sensitive than HiChIP alone. Our findings suggest that the predictive power achieved through incorporating multi-omics data with HiChIP loops is able to overcome dropouts of enhancer interactions from HiChIP data sets, which may be due to technical shortcomings of the assay, like GC content, length of interaction, sequencing depth, or chromatin composition. Alternatively, the result may be attributable to basic probability given that there were a greater number of observed predicted loops than experimental HiChIP loops.

The wide-spread availability of high-throughput, low-input, and low-cost multi-omics profiling technologies has increased the number of cell type–specific functional genomics data sets. Hence, there is a burgeoning need for tools to predict meaningful distal regulatory features with cell type–specific accuracy. LoopPredictor makes it theoretically possible to predict active regulatory topologies with high accuracy and sensitivity in all cell types that lack topological data.

Methods

Classifier selection for Anchor type Predictor

We tested the F1 scores of four standard classifiers: LinearSVC, LogisticRegression, KNeighbors, and RandomForest in four HiChIP data sets; four classifiers were constructed by using scikit-learn (Pedregosa et al. 2012) with default parameters. RandomForest outperformed the other classifiers and was selected for the construction of ATP.

A hybrid Random Forest classifier based on multitask framework

In this study, we combined the feature selection ability of Group LASSO and the prediction power of Random Forest to construct a hybrid classifier (Berzal et al. 2004). Then, we built the hybrid Random Forest classifier on the framework of multitask. First, Group LASSO was used to explore the sparsity constraints of prediction; we defined the general classification task as $V_i = m_i F_i$; i represents the number of subtasks. For the i-th task, V_i is the labels vector for the task, m_i is the regression coefficient for i-th task, and F_i is the feature matrix of task i. We assume there are N subtasks in total, and M represents a N′*N matrix, in which N′ is the number of common features among all the tasks; the objective function is defined as

$$\hat{M} = \sum_{i=1}^{N} \| V_i - m_i F_i \|_2^2 + \lambda \| M \|_{1/2}.$$

We applied the feature selection module before fitting Random Forest, and the multitask framework was implemented by scikit-learn. Then, the hybrid classifier was integrated in ATP, and we tested the performance of ATP in four HiChIP data sets by fivefold cross-validation.

An adaptable Gradient Boosted Regression Trees regressor

The additive model of GBRT was built in greedy function (Friedman 2001).

$$F_m(x) = F_{m-1}(x) + \nu \gamma_m h_m(x).$$

The tree newly added in each step was represented by h_m, which tried to minimize the loss L, and GBRT used a type of negative gradient loss function for the current model F_{m-1}; γ_m was step length, which was calculated by line search

$$\gamma_m = arg\ \min = \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right),$$

and υ was used to scale step length, which called learning rate; learning rate impacted the training error cooperating with the number of weak learners. In addition, GBRT considered the strategy of stochastic gradient boosting (Friedman 2002), which combined gradient boosting with bagging; for each iteration, GBRT trained the base model on a fraction of the training sample, and the value of the fraction also impacted the performance of regression. Therefore, it is crucial to determine the combination of learning rate, weak learner number, and subsample fraction. Our problem is how to tune the model parameters for four different HiChIP data sets, while automatically adapting to the unknown data sets input by users. To solve the problem, we developed an adaptable module for GBRT to generate different combinations of parameters to fit the model iteratively, then selected the optimal one to train the data set and perform prediction.

Software availability

The source code of LoopPredictor and its execution instruction are available at GitHub (https://github.com/bioinfomaticsCSU/LoopPredictor). We have also made the source code available in the Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by grants from the National Natural Science Foundation of China No. 61732009 (M.L.), the 111 Project No. B18059 (M.L.), Hunan Provincial Science and Technology Program 2019CB1007 (M.L.), the National Institutes of Health DE023177, HL127717, HL130804, and HL118761 (J.F.M.) and F31HL136065 (M.C.H.), Vivian L. Smith Foundation (J.F.M.), State of Texas funding (J.F.M.), and Foundation LeDucq Transatlantic Networks of Excellence in Cardiovascular Research No. 14CVD01, "Defining the genomic topology of atrial fibrillation" (J.F.M.).

Author contributions: Conceptualization, L.T. and M.C.H.; methodology, L.T., M.C.H., and M.L.; investigation, L.T. and M.C.H.; writing—original draft, M.C.H. and L.T.; writing—review and editing, L.T., M.C.H., M.L., J.W., and J.F.M.; funding acquisition, M.L., J.F.M., and M.C.H.; resources, J.F.M. and M.L.; supervision, M.L., JX.W., and J.F.M.; visualization, M.C.H. and L.T.; data curation, L.T. and M.C.H.

References

Barakat TS, Halbritter F, Zhang M, Rendeiro AF, Perenthaler E, Bock C, Chambers I. 2018. Functional dissection of the enhancer repertoire in human embryonic stem cells. *Cell Stem Cell* **23:** 276–288.e8. doi:10.1016/j.stem.2018.06.014

Belokopytova PS, Nuriddinov MA, Mozheiko EA, Fishman D, Fishman V. 2020. Quantitative prediction of enhancer–promoter interactions. *Genome Res* **30:** 72–84. doi:10.1101/gr.249367.119

Berzal F, Cubero J-C, Sánchez D, Serrano JM. 2004. ART: a hybrid classification model. *Mach Learn* **54:** 67–92. doi:10.1023/B:MACH.0000008085.22487.a6

Breiman L. 2001. Random forests. *Mach Learn* **45:** 5–32. doi:10.1023/A:1010933404324

Cheng Y, Ma Z, Kim B-H, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J, et al. 2014. Principles of regulatory information conservation between mouse and human. *Nature* **515:** 371–375. doi:10.1038/nature13985

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci* **107:** 21931–21936. doi:10.1073/pnas.1016071107

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* **295:** 1306–1311. doi:10.1126/science.1067799

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74. doi:10.1038/nature11247

Ernst J, Kellis M. 2017. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12:** 2478–2492. doi:10.1038/nprot.2017.124

Fan J, Upadhye S, Worster A. 2006. Understanding receiver operating characteristic (ROC) curves. *CJEM* **8:** 19–20. doi:10.1017/S1481803500013336

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. *Ann Statistics* **29:** 1189–1232. doi:10.1214/aos/1013203451

Friedman JH. 2002. Stochastic gradient boosting. *Comput Stat Data An* **38:** 367–378. doi:10.1016/S0167-9473(01)00065-2

Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B, Lander ES, Engreitz JM. 2016. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **354:** 769–773. doi:10.1126/science.aag2445

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. 2009. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462:** 58–64. doi:10.1038/nature08497

Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A, Schreiber J, Noble WS, et al. 2019. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176:** 377–390.e19. doi:10.1016/j.cell.2018.11.029

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Calcar SV, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39:** 311–318. doi:10.1038/ng1966

Hong Z, Zeng X, Wei L, Liu X. 2019. Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinform Oxf Engl* **36:** 1037–1043. doi:10.1093/bioinformatics/btz694

Huang D-Y, Kuo Y-Y, Lai J-S, Suzuki Y, Sugano S, Chang Z-F. 2004. GATA-1 and NF-Y cooperate to mediate erythroid-specific transcription of *Gfi*-1B gene. *Nucleic Acids Res* **32:** 3935–3946. doi:10.1093/nar/gkh719

Jain A, Nandakumar K, Ross A. 2005. Score normalization in multimodal biometric systems. *Pattern Recogn* **38:** 2270–2285. doi:10.1016/j.patcog.2005.01.012

Levine M. 2010. Transcriptional enhancers in animal development and evolution. *Curr Biol* **20:** R754–R763. doi:10.1016/j.cub.2010.06.070

Li C-Q, Kim MY, Godoy LC, Thiantanawat A, Trudel LJ, Wogan GN. 2009. Nitric oxide activation of Keap1/Nrf2 signaling in human colon carcinoma cells. *Proc Natl Acad Sci* **106:** 14547–14551. doi:10.1073/pnas.0907539106

Mariani L, Weinand K, Vedenko A, Barrera LA, Bulyk ML. 2017. Identification of human lineage-specific transcriptional coregulators enabled by a glossary of binding modules and tunable genomic backgrounds. *Cell Syst* **5:** 187–201.e7. doi:10.1016/j.cels.2017.06.015

Moore JE, Pratt HE, Purcaro MJ, Weng Z. 2020. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol* **21:** 17. doi:10.1186/s13059-019-1924-8

Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13:** 919–922. doi:10.1038/nmeth.3999

Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML, Rubin AJ, Granja JM, Kazane KR, et al. 2017. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet* **49:** 1602–1612. doi:10.1038/ng.3963

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G, et al. 2012. Scikit-learn: machine learning in Python. *J Mach Learn Res* **12:** 2825–2830.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. *In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444:** 499–502. doi:10.1038/nature05295

Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, Gage FH, Swigut T, Wysocka J. 2015. Enhancer divergence and *cis*-regulatory evolution in the human and chimp neural crest. *Cell* **163:** 68–83. doi:10.1016/j.cell.2015.08.036

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470:** 279–283. doi:10.1038/nature09692

Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159:** 1665–1680. doi:10.1016/j.cell.2014.11.021

Schmidt M, Roux NL, Bach F. 2017. Minimizing finite sums with the stochastic average gradient. *Math Program* **162:** 83–112. doi:10.1007/s10107-016-1030-6

Wang E, Zaman N, Mcgee S, Milanese J-S, Masoudi-Nejad A, O'Connor-McCourt M. 2015. Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Semin Cancer Biol* **30:** 4–12. doi:10.1016/j.semcancer.2014.04.002

Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, Abraham BJ, Cohen MA, Nabet B, Buckley DL, et al. 2017. YY1 is a structural regulator of enhancer-promoter loops. *Cell* **171:** 1573–1588.e28. doi:10.1016/j.cell.2017.11.008

Whalen S, Truty RM, Pollard KS. 2016. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* **48:** 488–496. doi:10.1038/ng.3539

Xiang P, Lo C, Argiropoulos B, Lai CB, Rouhi A, Imren S, Jiang X, Mager D, Humphries RK. 2010. Identification of E74-like factor 1 (ELF1) as a transcriptional regulator of the Hox cofactor MEIS1. *Exp Hematol* **38:** 798–808.e2. doi:10.1016/j.exphem.2010.06.006

Xiao Y, Hill MC, Li L, Deshmukh V, Martin TJ, Wang J, Martin JF. 2019. Hippo pathway deletion in adult resting cardiac fibroblasts initiates a cell state transition with spontaneous and self-sustaining fibrosis. *Genes Dev* **33:** 1491–1505. doi:10.1101/gad.329763.119

Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EEM. 2009. Combinatorial binding predicts spatio-temporal *cis*-regulatory activity. *Nature* **462:** 65–70. doi:10.1038/nature08531

Received April 11, 2020; accepted in revised form October 2, 2020.