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,*], 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, TX 77030, USA

[3]Program in Developmental Biology, Baylor College of Medicine, Houston, TX 77030, USA

[4] Department of Pediatrics, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[5]Cardiovascular Research Institute, Baylor College of Medicine, Houston, TX 77030, USA

[6]Texas Heart Institute, Houston, TX 77030, USA

Supplemental Methods

Identification of loops from HiChIP data

We collected four HiChIP datasets: K562-YY1, HCT116-YY1, K562-H3K27ac, and GM12878-H3K27ac for prediction, the raw reads in fastq format were downloaded from GEO. We employed HiC-Pro (v2.11.1) (Servant et al. 2015) to align the paired-end reads and used hichipper (v0.7.7) (Lareau and Aryee 2018) to produce loops. Briefly, in the pipeline of HiC-Pro, the raw fastq HiChIP data were aligned to hg19 reference with “--very-sensitive” and “end-to-end” option. Then the alignment results as well as the restriction enzyme cut sites file (hg19 Mbol digest) were imported into hichipper, the chromatin loops were called by combining all the replicates and using all read density. We only considered the uniquely mapped reads with false discovery rate (FDR) < 0.05, the loops with at least 2 PETs supported and length larger than 5kb were retained. The vast majority of the loops (91.7% for K562-YY1,90.3% for HCT116-YY1, 96.3% for K562-H3K27ac, 95.9% for GM12878-H3K27ac) met above conditions and used for subsequent analysis.

Identification of regulatory elements for all the anchors

Firstly, we extracted anchors from loops, then we utilized ChromHMM (v1.21) (Ernst and Kellis 2017) annotations to identify the promoters and enhancers for the anchors from K562 and GM12878, and ENCODE Segway (Hoffman et al. 2012) for the anchors from HCT116 cell type. We also checked the transcribed activity of promoters by using the RNA-seq expression data from ENCODE (The ENCODE Project Consortium. 2012), if the FPKM> 0.5, then retained the promoter.

Super enhancer analysis

We employed the "-style super" option of findPeaks function in HOMER package (v4.10) (Heinz et al. 2010) to identify the super enhancer regions, the peaks found within 12.5 kb were merged together into large regions, the slope threshold was set to 1000.

GO analysis for enhancer anchors

GO analysis for enhancer anchors was taken by Metascape (v3.5) (Zhou et al. 2019) with the minimum overlap of 3, and minimum enrichment of 1.5.

Motif analysis for enhancer anchors

Firstly, the anchor regions from loops were overlapped and trimmed with the corresponding ATAC-seq peaks, then the trimmed anchors were used to detect motifs by HOMER package (v4.10) (Heinz et al. 2010) with size of 200, the transcription factors with $-\log_2$(p-value)> 100 were chosen. Then the RNA-seq expression data from ENCODE was scaled by FPKM, the chosen transcription factors were ranked by the gene expression.

Coincidence between different loop sets

We first used 500bp as a threshold to merge the nearby anchors into a valid anchor, if both of the valid anchors of one loop appear in another loop set, and there was an interaction between them, the loop is regarded as coincidence.

Multi-omics datasets processing

Multi-omics datasets were prepared for the prediction, including ChIP-seq/CUT&RUN, ATAC-

seq, in functional genomics, RRBS in epigenomics, and RNA-seq in transcriptomics. For ChIP-seq/CUT&RUN data, the peak files were downloaded from ENCODE if accessible, or used Bowtie2 to align the raw reads to reference with default settings, and uniquely mapping reads were used to identify enrichment regions. The sequence alignment was then transformed into platform independent data structure by makeTagdirectory package of HOMER (v4.10), and findPeaks package was used to detect peaks with False discovery rate (FDR) <0.001, for transcription factor datasets, the peak size was set to 200bp, for histone marker datasets, the peak size was set to 500bp. The chromatin accessibility was profiled by ATAC-seq data, we filtered out some uninformative reads after the alignment with mapQuality<30, and "isProperPair" was set to only retain the proper paired reads, then the reads aligned to chrM were removed, MACS2 (Zhang et al. 2008) was used to call peaks. The methylation profile of RRBS data were downloaded from ENCODE. The gene expression data were downloaded from ENCODE, or aligned the raw reads to reference by using STAR (2.7.2a) (Dobin et al. 2013) with 2 mismatches at most, the raw read counts were normalized by FPKM.

Features generation

For anchor type predictor (ATP), which is a minimal classifier, we trained the model with as few features as possible while ensuring the accuracy near optimal. Here, we only generated the features for the anchor regions. The R package GenomicRanges (Lawrence et al. 2013) was used to extract the profile values of specific regions. For ChIP-seq/CUT&RUN and ATAC-seq datasets, the peaks within anchor regions were extracted, and the mean value of peak signals were calculated as functional genomics features. For RRBS data, we calculated the methylation signals by multiplying methylation percentage by read counts for each profiled position within the anchor regions. The weighted mean values of methylation signals were used as epigenomics features. For RNA-seq data, we extracted the FPKM value of genes whose transcription start site locate within the anchors, then the mean FPKM value were used as transcriptomics features.

For Confidence predictor (CP), we constructed a powerful regressor to predict the score of loops accurately which required more features as input. We added left-flanking, in-between, right-flanking regions for each anchor pair to generate features, so there are five regions waiting for the feature generation including two anchor regions. The left-flanking regions were the 2kb

extension from the start site of left anchor, the in-between regions were the intermediate parts of anchor pairs, the right-flanking regions were the 5kb extension from the end site of right anchor. The features generation method for anchor regions is consistent with ATP. For left-flanking, in-between, and right-flanking regions, we calculated mean values as well as standard deviations for every region following the same method for ATP.

Training sample preparation

Four loop sets identified from HiChIP data were used as positive samples, the feature of each sample was generated as described above, and the annotations of regulatory elements for anchors were used as the target of samples, we only retained four types of targets for the prediction: promoter-enhancer, promoter-promoter, enhancer-enhancer, and none. The type of promoter-enhancer indicated one of the two anchors is promoter, and the other is enhancer, promoter-promoter and enhancer-enhancer indicated both of the anchors are promoters or enhancers. The type of none-none represented the loops are informative, including either of two anchors or both anchors are non-regulatory elements. Negative samples were produced by randomly selecting chromatin regions, avoiding ±2kb regions around TSS of any gene. The targets of negative samples were none-none, and the amount of negative sample was consistent with positive sample.

We combined positive and negative samples, and split the samples into 7:3 for training and testing, 5-fold cross validation was used in every training process.

Classifier selection for anchor type predictor (ATP)

We tested the F1 scores of four standard classifiers: LinearSVC, LogisticRegression, KNeighbors, and RandomForest in four HiChIP datasets, four classifiers were constructed by using scikit-learn (Pedregosa et al. 2012) with default parameters. The testing results is shown in Fig. **3A**, RandomForest outperformed the other classifiers, which was selected for the construction of ATP.

A hybrid Random Forest classifier based on multi-task framework

Random Forest is a powerful classifier which uses all the given features to perform the

prediction. As we need to train different classifiers for different HiChIP datasets, which is time consuming and tedious to select the important features to feed into model. Therefore, what we faced is how to select the most important features to minimize the input set while ensure the accuracy of prediction.

Multi-task learning is an approach which allows tasks training in parallel, and transforms information between related tasks, the inductive transformation would help each task learning better (Ruder 2017). Group LASSO is one of the sparse learning approaches, which utilizes the coefficients of features to construct the prediction model (Meier et al. 2008). In this study, we combined the feature selection ability of Group LASSO and the prediction power of Random Forest to construct a hybrid classifier. Then we built the hybrid Random Forest classifier on the framework of multi-task. Firstly, Group LASSO was used to explore the sparsity constraints of prediction, we defined the general classification task as $V_{\mathrm{i}} = m_i F_i$, *i* represents the number of sub-tasks, for the *i*-th task, V_i is the labels vector for the task, and m_i is the regression coefficient for *i*-th task, F_i is the feature matrix of task *i*, we assume there are N sub-tasks in total, M represents a N'*N matix, in which N' is the number of common features among all the tasks, the objective function is defined as

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

We applied the feature selection module before fitting Random Forest, the multi-task framework was implemented by scikit-learn. Then the hybrid classifier was integrated in ATP, we tested the performance of ATP in four HiChIP datasets by 5-fold cross validation, the importance of features was ranked for the prediction, which shows in Fig. 3E, and the F1 score performance of increasing feature numbers in Fig. 3C shows that ATP only needs 12 features as input to obtain near optimal.

Features correlation evaluation

Firstly, we used Pearson's method to calculate the correlation-based distance matrix for all the features, then applied hierarchical clustering on the matrix, which used the method of "average". The correlation heatmap was implemented by using R packages "stats" and "gplots" (v3.0.4).

ChIP-seq peaks density for YY1 and ELF1

The bed file of YY1 ChIP-seq peaks were used as the target regions, and the alignment file of YY1 and ELF1 in bigwig/bam format were used to plot the density heatmap, which was implemented by R package Genomation (v1.20.0) (Akalin et al. 2014).

ROC curve and precision-recall curve evaluation for ATP

Except for F1-score, we also utilized Receiver Operating Characteristic (ROC) metric and precision-recall curve to evaluate the classification quality of ATP. Regular ROC curves are used to evaluate the binary classification output, while our problem has four class labels. We therefore used the extension setting “Multiclass” in scikit-learn (v0.20.3) (Pedregosa et al. 2012) to draw the ROC curve for each label. In addition, two kinds of measures were used to evaluate the general classification: “micro-averaging” considers each kind of label in the indicator matrix as a binary classification problem; “macro-averaging” allocates equal weight to each label in the classification task.

Quantification of features in different regions

We gathered features for CP according to different loop-associated regions: two neighbors, two anchors and inter-anchor window. Then binned the features by the distance of 100bp, and the mean value and standard deviation of each region were calculated. We utilized z-score normalization to scale the signal of features, and then plotted the distribution of each region using boxplot from ggplot2 package (3.3.2) (Valero-Mora 2010).

An adaptable Gradient Boosted Regression Trees (GBRT) regressor

Gradient Boosted Regression Trees (GBRT) is a kind of inductively generated tree ensemble model, which trains a new tree against the negative gradient of loss function for each step. The motivation of GBRT is to combine multiple weak learners to generate a powerful 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 loos function for current model F_{m-1}, γ_m was step

length, which was calculated by line search

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

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 training sample, and the value of fraction also impacted the performance of regression. Therefore, it's 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 datasets, meanwhile automatically adapt to the unknown datasets input by users. To solve the problem, we developed an adaptable module for GBRT to generate different combination of parameters to fit the model iteratively, then selected the optimal one to train the dataset and performed prediction.

Evaluation of regression

We evaluated the performance of Confidence Predictor (CP) by calculating adjusted R-square value, Mean Absolute Error (MAE) and Root-Mean Squared Error (RMES). Adjusted R-square compares the explanatory power of regression models that contain different numbers of predictors, which is more objective than R-square to measure the multi-variable regression model. For R-square, the Sum of Squared Regression Error (RSS) and Sum of Squared Total Error were calculated, the calculation of adjusted R-square was based on R-square, which has been adjusted for the number of predictors in the model, and it is always lower than the R-square (Shieh 2008).

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$R^2_{adj} = 1 - (1 - R^2)\frac{N - 1}{N - M - 1}$$

MAE measures the average magnitude of errors in a set of predictions without considering the direction. RMSE is a quadratic scoring function which measures the average magnitude of error by calculating the square root of the average between prediction and actual observation. The scatter plot of the predicted values versus the real values indicate the accuracy of prediction

directly, the closer the slope of reference line to 1, the better the prediction.

Validation of predicted loops

We used the same method with “Coincidence between different loop sets” to identify the overlapping loops. The proportion of loop counts by distance was calculated by R package diffloop (v1.16.0) (Hnisz et al. 2016), and the differential analysis between loop sets was implemented by R package HiCcompare(v1.10.0) (Stansfield et al. 2018). The visualization of H3K27ac ChIP-seq track and interactions was implemented by WashU Epigenome (Li et al. 2019). For calculating the distribution of aggregated loop numbers around TSS, we first annotated the anchor regions by ChIPpeakAnno (v3.22.2) (Zhu et al. 2010), the distances between anchors and TSS were retrieved, and binned the loops by distances, then the number of loops for each bin was counted. The conformation plots of predicted loops were generated by Sushi (v1.26.0) (Phanstiel et al. 2014).

Conservation level evaluation for loops

The anchor regions of loops were overlapped and trimmed by the corresponding ATAC-seq peaks, then the phastCons60 scores for trimmed anchors were extracted by GenomicScores (v2.0.0) (Puigdevall and Castelo 2018).

Loop proportions within TAD boundary

The Hi-C TAD coordinates were predicted from Dixon et al. Nature 2012 (Dixon et al. 2012), we annotated the active states of TAD by integrating the ChIP-seq datasets of active histone marks (Matthews and Waxman 2018)

Genome Build Used in This Study

For human genomics analysis, we utilized reads mapped to the GRCh37 (hg19) genome for practical considerations. Firstly, part of the improvements made in the GRCh38 (hg20) genome build regard the incorporation of alternative sequences to adequately represent single nucleotide variability. The alignment strategies we utilized in this study would not be affected by these changes. The GRCh38 build also improves the mapping of centromeres, however,

centromeres possess a very low gene density and are generally transcribed at very low levels. Moreover, they contain large numbers of repetitive sequences (e.g. α-satellites), and so standard Next Generation Sequencing mapping pipelines are unlikely to align to these regions. Hence, they would not be enriched sources of enhancer-promoter contacts that would affect the analysis performed herein. Overall, we don't believe re-aligning reads to GRCh38 would significantly affect the conclusions made in this manuscript.

Published Datasets Used in This Study

ChIP-seq/CUT&RUN datasets: GSE29611 (The ENCODE Project Consortium. 2012), GSE35583 (Thurman et al. 2012), GSE31755 (The ENCODE Project Consortium. 2012), GSE127432 (The ENCODE Project Consortium. 2012), GSE32465 (Gertz et al. 2013), GSE30263 (Wang et al. 2012), GSE31477 (The ENCODE Project Consortium. 2012), GSE96253 (The ENCODE Project Consortium. 2012), GSE92075 (The ENCODE Project Consortium. 2012), GSE135286 (Xiao et al. 2019), GSE63255 (Sanij et al. 2015). ATAC-seq datasets: GSE108513 (Calviello et al. 2019), GSE47753 (Buenrostro et al. 2013), GSE10197 (The ENCODE Project Consortium. 2012) 5 (Kelso et al. 2017), GSE135286 (Xiao et al. 2019). RNA-seq da (The ENCODE Project Consortium. 2012) tasets: GSE88473, GSE90276 (The ENCODE Project Consortium. 2012), GSE33480, GSE72860 (Djebali et al. 2012). RR (The ENCODE Project Consortium. 2012) BS datasets: GSE27584 (The ENCODE Project Consortium. 2012), GSE27584, GSE27584 (The ENCODE Project Consortium. 2012).

References

Akalin A, Franke V, Vlahoviček K, Mason CE, Schübeler D. 2014. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinform Oxf Engl* **31**: 1127–9.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**: 1213–1218.

Calviello AK, Hirsekorn A, Wurmus R, Yusuf D, Ohler U. 2019. Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. *Genome Biol* **20**: 42.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**: 376.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489**: 101–108.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15–21.

Ernst J, Kellis M. 2017. Chromatin-state discovery and genome annotation with ChromHMM. *Nat Protoc* **12**: 2478.

Friedman JH. 2001. Greedy function approximation: A gradient boosting machine. *Ann Statistics* **29**: 1189–1232.

Friedman JH. 2002. Stochastic gradient boosting. *Comput Stat Data An* **38**: 367–378.

Gertz J, Savic D, Varley KE, Partridge EC, Safi A, Jain P, Cooper GM, Reddy TE, Crawford GE, Myers RM. 2013. Distinct Properties of Cell-Type-Specific and Shared Transcription Factor Binding Sites. *Mol Cell* **52**: 25–36.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38**: 576–589.

Hnisz D, Weintraub AS, Day DS, Valton A-L, Bak RO, Li CH, Goldmann J, Lajoie BR, Fan ZP, Sigova AA, et al. 2016. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**: 1454–1458.

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* **9**: 473.

Kelso TWR, Porter DK, Amaral ML, Shokhirev MN, Benner C, Hargreaves DC. 2017. Chromatin accessibility underlies synthetic lethality of SWI/SNF subunits in ARID1A-mutant cancers. *Elife* **6**: e30506.

Lareau CA, Aryee MJ. 2018. hichipper: a preprocessing pipeline for calling DNA loops from HiChIP data. *Nat Methods* **15**: 155.

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for Computing and Annotating Genomic Ranges. *Plos Comput Biol* **9**: e1003118.

Li D, Hsu S, Purushotham D, Sears RL, Wang T. 2019. WashU Epigenome Browser update 2019. *Nucleic Acids Res* **47**: W158–W165.

Matthews BJ, Waxman DJ. 2018. Computational prediction of CTCF/cohesin-based intra-TAD loops that insulate chromatin contacts and gene expression in mouse liver. *Elife* **7**: e34077.

Meier L, Geer SVD, Bühlmann P. 2008. The group LASSO for logistic regression. *J Royal Statistical Soc Ser B Statistical Methodol* **70**: 53–71.

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.

Phanstiel DH, Boyle AP, Araya CL, Snyder MP. 2014. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* **30**: 2808–2810.

Puigdevall P, Castelo R. 2018. GenomicScores: seamless access to genomewide position-specific scores from R and Bioconductor. *Bioinform Oxf Engl* **34**: 3208–3210.

Ruder S. 2017. An Overview of Multi-Task Learning in Deep Neural Networks.

Sanij E, Diesch J, Lesmana A, Poortinga G, Hein N, Lidgerwood G, Cameron DP, Ellul J, Goodall GJ, Wong LH, et al. 2015. A novel role for the Pol I transcription factor UBTF in maintaining genome stability through the regulation of highly transcribed Pol II genes. *Genome Res* **25**: 201–212.

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, Heard E, Dekker J, Barillot E. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**: 259.

Shieh G. 2008. Improved Shrinkage Estimation of Squared Multiple Correlation Coefficient and Squared Cross-Validity Coefficient. *Organ Res Methods* **11**: 387–407.

Stansfield JC, Cresswell KG, Vladimirov VI, Dozmorov MG. 2018. HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. *Bmc Bioinformatics* **19**: 279.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**: 75–82.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57.

Valero-Mora PM. 2010. ggplot2: Elegant Graphics for Data Analysis. *J Stat Softw* **35**.

Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, et al. 2012. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res* **22**: 1680–1688.

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. *Gene Dev* **33**: 1491–1505.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-seq (MACS). *Genome Biol* **9**: R137.

Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. 2019. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**: 1523.

Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, Green MR. 2010. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *Bmc Bioinformatics* **11**: 237.