



## Evidence of widespread, independent sequence signature for transcription factor cobinding

Manqi Zhou, Hongyang Li, Xueqing Wang, et al.

*Genome Res.* 2021 31: 265-278 originally published online December 10, 2020

Access the most recent version at doi:[10.1101/gr.267310.120](https://doi.org/10.1101/gr.267310.120)

---

**References** This article cites 53 articles, 6 of which can be accessed free at:  
<http://genome.cshlp.org/content/31/2/265.full.html#ref-list-1>

**Creative Commons License** This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).



---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

## Method

# Evidence of widespread, independent sequence signature for transcription factor cobinding

Manqi Zhou,<sup>1</sup> Hongyang Li,<sup>1</sup> Xueqing Wang, and Yuanfang Guan

*Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA*

Transcription factors (TFs) are the vocabulary that genomes use to regulate gene expression and phenotypes. The interactions among TFs enrich this vocabulary and orchestrate diverse biological processes. Although simple models identify open chromatin and the presence of TF motifs as the two major contributors to TF binding patterns, it remains elusive what contributes to the *in vivo* TF cobinding landscape. In this study, we developed a machine learning algorithm to explore the contributors of the cobinding patterns. The algorithm substantially outperforms the state-of-the-field models for TF cobinding prediction. Game theory–based feature importance analysis reveals that, for most of the TF pairs we studied, independent motif sequences contribute one or more of the two TFs under investigation to their cobinding patterns. Such independent motif sequences include, but are not limited to, transcription initiation–related proteins and known TF complexes. We found the motif sequence signatures and the TFs are rarely mutual, corroborating a hierarchical and directional organization of the regulatory network and refuting the possibility of artifacts caused by shared sequence similarity with the TFs under investigation. We modeled such regulatory language with directed graphs, which reveal shared, global factors that are related to many binding and cobinding patterns.

[Supplemental material is available for this article.]

Regulatory biomolecules cooperatively decode the human genome and ultimately orchestrate a variety of tissue-specific cellular processes. One of the key challenges for understanding the mechanisms underlying gene expression and human diseases is how to analyze and delineate the genome-wide landscape of multiple characteristic biochemical signatures. An important aspect is the regulation of transcription factors (TFs), which bind to DNAs and drive the expression or suppression of genes. TFs do not work alone; they cooperate and interact with each other and thus create complex transcription patterns of the genome (Lemon and Tjian 2000; Perez-Pinera et al. 2013; Li et al. 2014; Ang et al. 2016).

Studying the collaborative patterns of TFs is a much more complicated problem than characterizing the contributors of a single TF binding event, for which open chromatin and the presence of the TF motifs have been recognized as the key determinants (Pique-Regi et al. 2011; Neph et al. 2012; Tsompana and Buck 2014). Naïvely, one might deduce that the main contributors to the cobinding patterns of a TF pair should be the presence of the two TF motifs and open chromatin. Yet, empirically many open questions remain to be investigated: are there other independent sequence signatures contributing to TF cobinding? If such sequence signatures do exist, are they generic for a TF when it is paired with any other TFs, or are they unique to a specific TF pair? Are there global, important sequence signatures affecting many TF cobinding patterns? We attempt to answer these questions with machine learning and game theoretical approach–based feature analysis.

The Encyclopedia of DNA Elements (ENCODE) and the NIH Roadmap Projects have provided invaluable resources of the *in vivo* biochemical signatures, including chromatin accessibility and TF binding. Many exciting computational approaches have

been developed to predict transcription factor binding sites (TFBSs) and to study what controls the binding patterns (Bulyk 2003; Blanchette et al. 2006; Kumar and Bucher 2016; Chen et al. 2017; Quach and Furey 2017). However, only a few models are developed for the prediction of TF cobinding patterns, and most of them focus on the pairing with a single motif (e.g., CTCF [Liu et al. 2016]). Although pioneering works have made substantial contributions, there is an urgent need to develop a generalizable model for the prediction of TF-cobinding and identify key contributors for this process.

This study aims at addressing these questions from a machine learning perspective. Towards this goal, we developed an algorithm that focuses on dissecting the predictive elements for genome-wide TF cobinding across diverse cell types based on chromatin accessibility, TF binding motifs, and gene locations. What is unique to this model is that not only the motifs of the TF pairs under investigation but also the motifs of other seemingly irrelevant TFs are considered in building the model. This allows us to further adopt game theory–based approaches to identify independent contributors to the binding patterns of a TF pair. Additionally, we created regulatory graphs based on this feature importance analysis.

## Results

### Designing a machine learning model to dissect individual contributors that predict TF cobinding

Our goal is to dissect the individual contributors to TF cobinding, which include motif locations of the TF pair, other TF sequence signatures, chromatin accessibility, and gene locations. Our approach is to first establish a predictive model to predict TF

<sup>1</sup>These authors contributed equally to this work.

Corresponding author: [gyuanfan@umich.edu](mailto:gyuanfan@umich.edu)

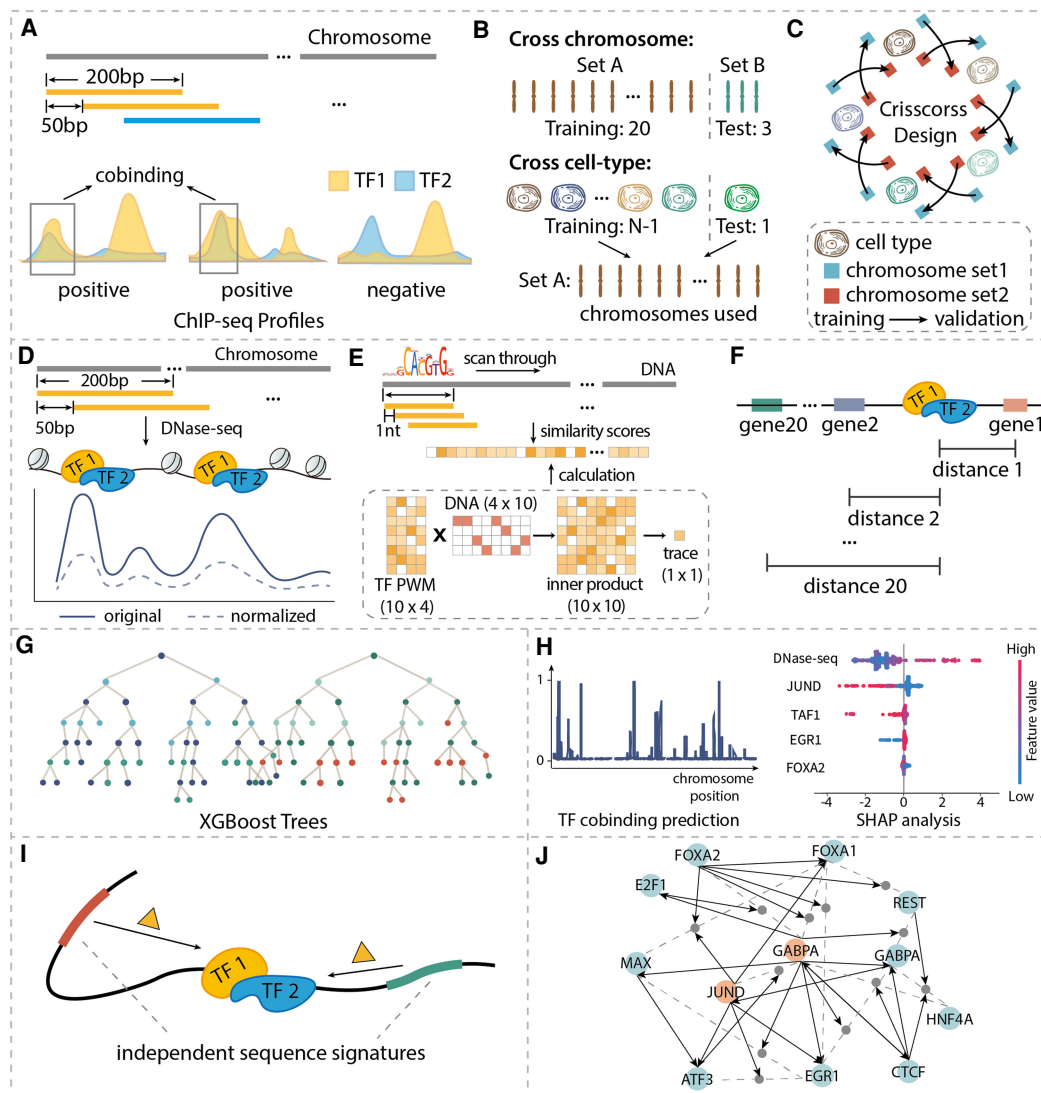
Article published online before print. Article, supplemental material, and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.267310.120>.

© 2021 Zhou et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <http://genome.cshlp.org/site/misc/terms.xhtml>). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

cobinding using features extracted from the above data types and then use game theoretical feature analysis to identify independent contributions of each feature.

To this end, we first created the gold standard cobinding data set for training the model. Following one of the mainstream conventions in the TF binding field (Levy and Hannehalli 2002; Valouev et al. 2008; Yu et al. 2016), chromosomes are segmented every 200 bp, with a moving step of 50 bp (Fig. 1A). If a 200-bp region is bound by two transcription factors at the same time in a specific cell line as examined by ChIP-seq experiments, it is defined as a positive example. Otherwise, the region is considered as a negative

example (Fig. 1A). We carried out the experiments on 13 commonly seen and well-studied TFs (ATF3, CTCF, E2F1, EGR1, FOXA1, FOXA2, GABPA, HNF4A, JUND, MAX, NANOG, REST, and TAF) using the ENCODE data set (Supplemental Tables S1, S2). The TFs were chosen by two criteria: belonging to different families grouped by binding site similarity as defined by sequence homology to a previously characterized DNA-binding domain (Supplemental Table S1; Lambert et al. 2018); and having sufficient ChIP-seq data across ENCODE cell lines for subsequent analysis of TF-TF interactions. A total of 228 TF cobinding profiles involving 56 different TF-TF pairs were prepared to train and evaluate the model.



**Figure 1.** Machine learning–based method for identifying independent sequence signatures for TF binding and cobinding. (A) Preparing gold standard co-occupancy profiles from the ChIP-seq data of single TF binding profiles. For each 200-bp genomic interval of interest, it was labeled as ‘cobound’ if the labels of both TFs at this interval were ‘bound.’ (B) Train-test data partition for cross-chromosome and cross-cell type cases. (C) Crisscross design of model training and validating. Chromosomes were split into two sets. The blue and red squares represent the chromosome set 1 and set 2, respectively. The starting point of the arrow represents the training data set, and the endpoint of the arrow represents the validation data set. (D) DNase-seq-based features. Quantile normalization was applied to the original reads to eliminate experimental biases. In the plot, the dashed line is the values after quantile normalization and the solid line is the original reads. (E) DNA sequence and motif-based features. (F) Distance-to-gene features. Top 20 closest distances to proximal genes are calculated according to GENCODE annotation. (G) Illustration of tree-based machine learning models. (H) Predictions are continuous values between 0 and 1, which were generated by averaging results from all built models, and SHAP analysis (figure is illustrative only) was used to calculate feature importance in predicting TF cobinding. (I) Independent sequence signatures affect TF cobinding. (J) Illustration of TF cobinding effector network. Dashed lines connect TF pairs, and solid arrows connect TF-specific effectors with TF- or TF pair–specific effectors with TF pairs.

We sought to evaluate the model performance from two aspects: (1) cross-cell type predictions, that is, when the binding pattern is known in some cell lines and we have corresponding open chromatin status in other cell lines sharing the same genomic sequence; (2) predictions on new chromosomes, that is, when we move the model learned from one set of chromosomes to others. The former is useful for imputing a large amount of cobinding patterns across cell lines. The latter would be useful for studying binding patterns associated with SNPs or moving to other species. Toward this goal, we partitioned the chromosomes (Chr) into the training set (Chr 2, Chr 3, Chr 4, Chr 5, Chr 6, Chr 7, Chr 9, Chr 10, Chr 11, Chr 12, Chr 13, Chr 14, Chr 15, Chr 16, Chr 17, Chr 18, Chr 19, Chr 20, Chr 22, and Chr X), which are used to evaluate model performance across cell types in Case 1, and the test set (Chr 1, Chr 8, and Chr 21), which are used to evaluate model performance across chromosomes in Case 2 (Fig. 1B). This partition was made following several traditions in the TF-binding field (Keilwagen et al. 2019; Li et al. 2019). We randomly selected one cell line as the test cell line for each TF-TF pair and ensured that the number of test pairs in each cell line did not differ too much after the random selection.

TF co-occupancy is a much less explored field than single TF binding. With very limited prior experience in feature engineering, we consider it logical to adapt the feature engineering methods based on the state-of-the-field in the TF binding field. Open chromatin and motifs are two widely accepted factors that affect TF binding (Zhou and Liu 2004; Pique-Regi et al. 2011; Schmidt et al. 2017). Proximity to genes is another important factor, with some disagreement between studies (Chen et al. 2020). For the completion of the study, we considered all three of these. Based on several previous studies (Kelley et al. 2016; Li and Guan 2019a; Li et al. 2019; Kelley 2020), we first used long-range DNase-seq-based features to capture open chromatin long-distance information. Moving windows of 200 bp in size, with a step of 50 bp until 1500 bp, are included as features (Fig. 1D). Additionally, the variability of open chromatin signals was suggested to be informative for TF binding (Pique-Regi et al. 2011; Davie et al. 2015; Schmidt et al. 2017; Li et al. 2019). Thus, we used the maximum, minimum, average DNase signals in each of the 200-bp regions as additional features. To capture and correct for the DNA position biases, we also used the difference between the DNase value and the average value across all cell lines. Then, we scanned TF motifs along DNA sequences and used the top four sequence similarity scores in each 200-bp region as the motif-based features. Again, we took motifs that are even not within 200 bp under consideration (up to -750 to 750 bp) in order to study the effect of other TF signatures in cobinding (Fig. 1E). Finally, we used the closest distances to the 20 proximal genes as the distance-to-gene features, as previous studies suggested that TF binding is closely related to the positioning of open reading frames (Fig. 1F; Clements et al. 2007; Maienschein-Cline et al. 2012; Ezer et al. 2014). This resulted in a total of 526 features, for each TF-TF-cell line combination, which were nested trained with XGBoost (Fig. 1C,G; see Methods), a classical classifier used for TF binding predictions to build models for TF cobinding.

### Robust performance in predicting the rare events of TF cobinding across cell types and in previously unseen genome sequences

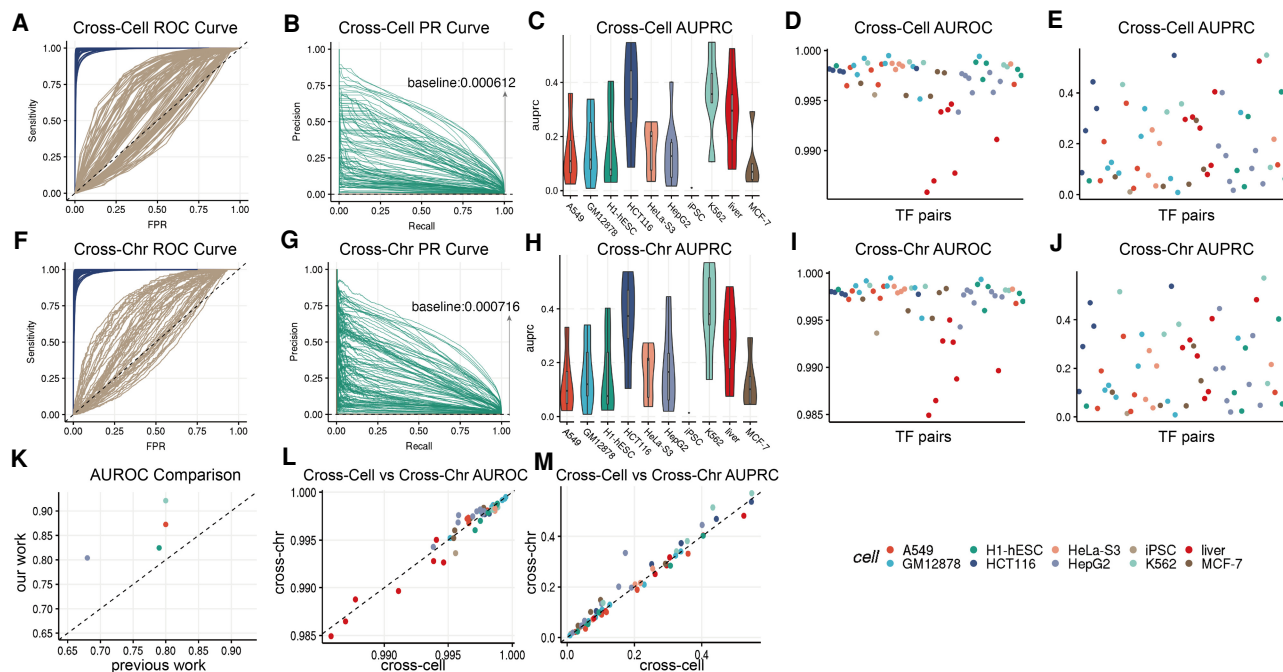
This algorithm demonstrated strong predictive performance when evaluated on the testing data set covering 56 TF pairs. We first calculated the area under the receiver operating characteristic curve

(AUROC) to evaluate the model performance (Fig. 2A,D; Supplemental Table S3). The median AUROC of cross-cell line predictions is 0.998 in 56 TF pairs, and the range of our prediction AUROCs is between 0.986 and 0.999. In addition to AUROC, we further calculated the area under the precision-recall curve (AUPRC) (Fig. 2B,C,E; Supplemental Table S3; Davis and Goadrich 2006). Whereas the AUROC baseline is consistently 0.5, the AUPRC baseline is the proportion of positive examples, which is 0.000504 in our data, because TF co-occupancy is extremely rare in the human genome. The median of AUPRC was 278 times over the baseline across 56 TF pairs.

The model represents a substantial improvement over the previous work, whose test set was limited to CTCF-associated TF pairs but could nevertheless serve as a benchmark for this study (Liu et al. 2016). In this benchmark work, Liu et al. focused on distinguishing between TF-TF cobinding and single TF binding events. Although our model was not directly designed for this purpose, it can still predict such events, as single-binding events were also used as negatives when we constructed the gold standard. In the test set cell line-TF pair combination, we selected the evaluation chunks limited to the single-bound and cobound intervals, for the CTCF-associated pairs shared between our study and their study on four cell lines (A549, H1-hECs, HepG2, and K562). Compared with the AUROC values of 0.80, 0.79, 0.68, and 0.80 in the previous research, our method reached AUROC values of 0.8724, 0.8245, 0.8039, and 0.9211, and AUPRC 0.2367, 0.2365, 0.0955, and 0.2438, respectively, on this evaluation (Fig. 2K). This improvement in performance might come from the extraction of maximal, minimal, and variance of the DNase features and long-range features, which are additional in this study. Of note, differentiating cobinding and single-binding events is a much more challenging task, explaining why these AUROCs are comparably lower than the global AUROCs, and the model presented in this paper was not designed for that purpose.

We also benchmarked with two existing software: TACO and coTraCTE (Jankowski et al. 2014; van Bömmel et al. 2018). Both are unsupervised methods. Both TACO and coTraCTE make predictions for a subset of the chromosome and do not consider the regions that are not open. Thus, for a comprehensive comparison, we compared both on their restricted subset, as well as the whole genome by assuming all nonpredicted sites are zero. For the restricted subset of the chromosomes, TACO's AUROC is 0.49867 and AUPRC is 0.01103; for the method presented in this study on the same region, the values are 0.857 and 0.061. For coTraCTE, its performance was AUROC=0.68450 and AUPRC=0.08259. The comparison highlights the importance of supervised learning in cobinding prediction.

Next, we tested the model on its ability to make predictions on the three held-out chromosomes (Chr 1, 8, and 21) on test cell lines. We obtained AUROCs with a range from 0.985 to 0.9995, with an average value of 0.997 (Fig. 2F,I; Supplemental Table S4). For the AUPRCs, the improvement of our predictions over the baseline ranges from 155 $\times$  to 1759 $\times$  (Fig. 2G,H,J; Supplemental Table S4). Compared with cross-cell line prediction, the cross-chromosome performance can reach equally high AUROC values (difference less than 0.001) and even better AUPRCs. Compared with single-bound examples, no-bound examples were easier to separate (Supplemental Table S5). For single-bound, the average AUROC and the average AUPRC are 0.817 and 0.336 for cross-cell type evaluation, and 0.820 and 0.348 for cross-chromosome evaluation. For no-bound, the



**Figure 2.** Cross-cell type and cross-chromosome evaluation of TF cobinding predictions. (A) The ROC curves of predicting cross-cell type TF cobinding. The brown lines are ROC curves of the logistic regression model. (B) The PR curves of predicting cross-cell type TF cobinding. The brown lines are PR curves of the logistic regression model. The dashed line with AUPRC of 0.000612 is the average AUPRC of the logistic regression. (C) The AUPRCs of cross-cell type predictions of 56 TF-TF pairs in 10 cell types. (D,E) The cross-cell type AUROCs and AUPRCs of 56 TF-TF pairs. Different colors indicate different cell types. All cell lines represented in the figure are testing cell lines. (F–J) Represent cross-chromosome evaluations. (K) AUROC comparison of our method (*y*-axis) and a previous work (*x*-axis) (Liu et al. 2016). The cross-cell type predictions of ATF3-CTCF on cell lines A549, H1-hESC, HepG2, and K562 are compared. (L) The AUROC and (M) the AUPRC comparison of cross-cell type (*x*-axis) and cross-chromosome (*y*-axis) predictions.

average AUROC and the average AUPRC are 0.998 and 0.411 for cross-cell type evaluation, and 0.998 and 0.403 for cross-chromosome evaluation.

Of note, predicting cobinding in previously unseen genome sequences is often considered a much more difficult task than making predictions for a chunk of sequence that already occurs in the training set (i.e., cross-cell line predictions). We compared the performance of these two situations (Fig. 2L,M) and found a strong correlation, indicating that the performance could depend on intrinsic difficulties in the TF pair rather than the sequence. In fact, the performance is almost indistinguishable compared to cross-cell line predictions.

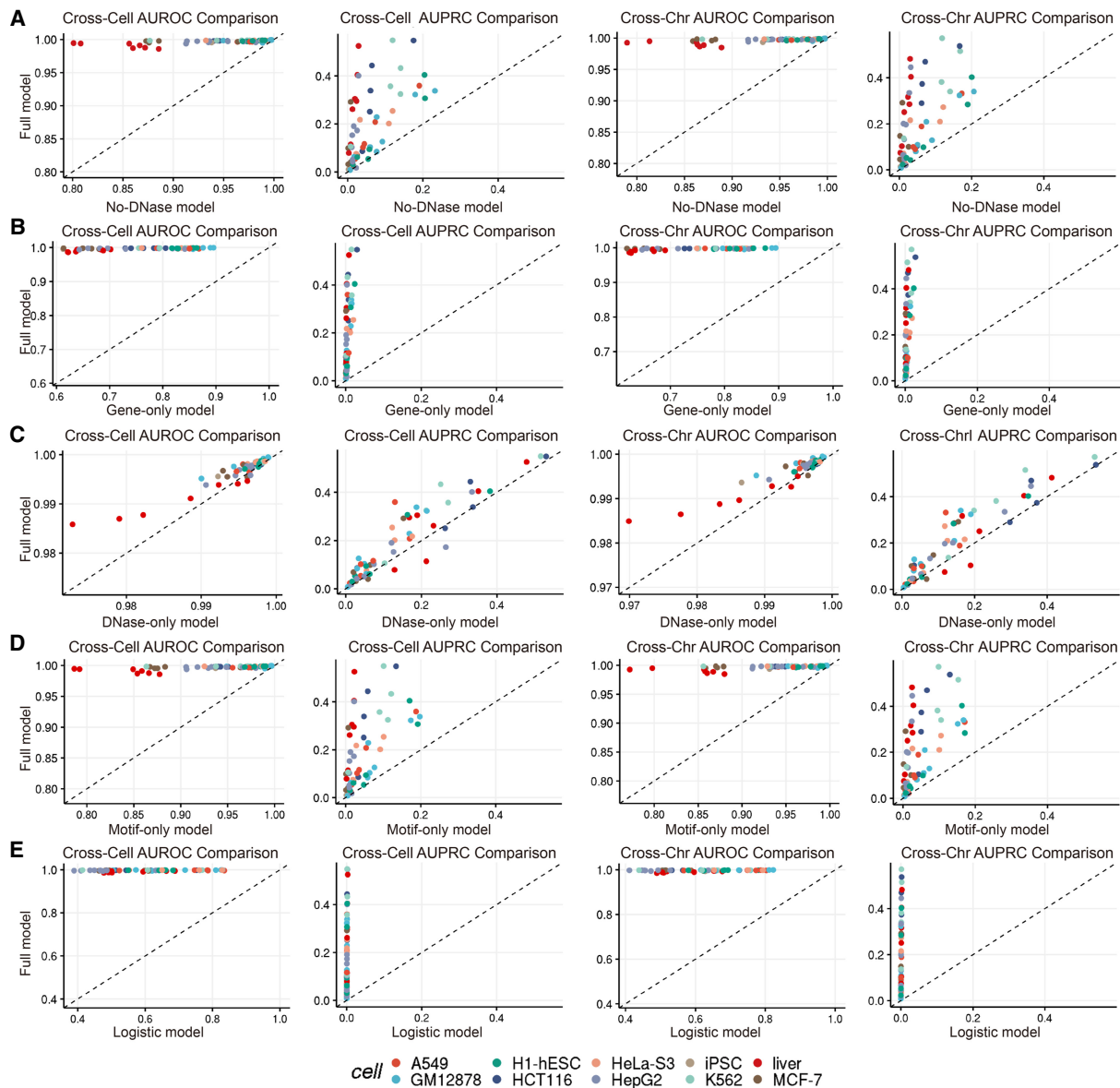
To evaluate the individual contribution of each type of feature, we separately evaluated model performance by including only DNase, only sequence motif, and only distance to genes, as well as a model with motif and gene distance (but excluding DNase) (Fig. 3). DNase is by far the most predictive feature, as it reflects chromatin status and achieves a cross-cell type AUROC of 0.995 and AUPRC of 0.149. The cross-chromosome AUROC is 0.995 and AUPRC is 0.148 (Fig. 3C). Removing DNase, but retaining the other two, we obtained the cross-cell type AUROC of 0.9450 and AUPRC of 0.0573. The cross-chromosome AUROC is 0.943 and AUPRC is 0.0553 (Fig. 3A). Motif-based features make the second most important contribution: the cross-cell type AUROC is 0.941 and AUPRC is 0.0492. The cross-chromosome AUROC is 0.940 and AUPRC is 0.0470 (Fig. 3D). Gene distance-based features are the least important: the cross-cell type AUROC is 0.773 and AUPRC is 0.00577. The cross-chromosome AUROC is 0.7620 and AUPRC is 0.00572 (Fig. 3B).

We also benchmarked with logistic regression and compared the model performance (Fig. 3E). For the logistic regression model with full features, the cross-cell type AUROC is 0.613 and AUPRC is 0.000612. The cross-chromosome AUROC is 0.622 and AUPRC is 0.000716. The adoption of the XGBoost model contributes to the improvement of the performance.

The above-described algorithm has practical application to differentiate TF cobinding status and predict TF cobinding sites, especially given the large amount of missingness in the ENCODE database. The predictions made by this algorithm are capable of differentiating TF cobinding versus single TF binding events and, certainly, versus nonbinding events, as has been demonstrated in the performance analysis in previous paragraphs (AUROCs). Here, we used some intervals from JUND-MAX in cell line K562 in the cross-cell line experiment to illustrate this point (Fig. 4). We were able to distinguish these three cases with the prediction values of 0.9938 for the cobinding case (Fig. 4A), 0.8174 for the single-binding case (Fig. 4B), and 0.1851 for the no-binding case (Fig. 4C). These results indicate that the algorithm leverages both TF motif sequences and chromosome accessibility to predict TF cobinding.

### Game theoretical approach-based analysis suggests widespread, strong, and independent sequence signatures related to TF cobinding patterns

The above model is unique in that not only the motif presence of the TF pair under investigation is used as a predictive feature but also the seemingly irrelevant motifs of different TFs. This allows

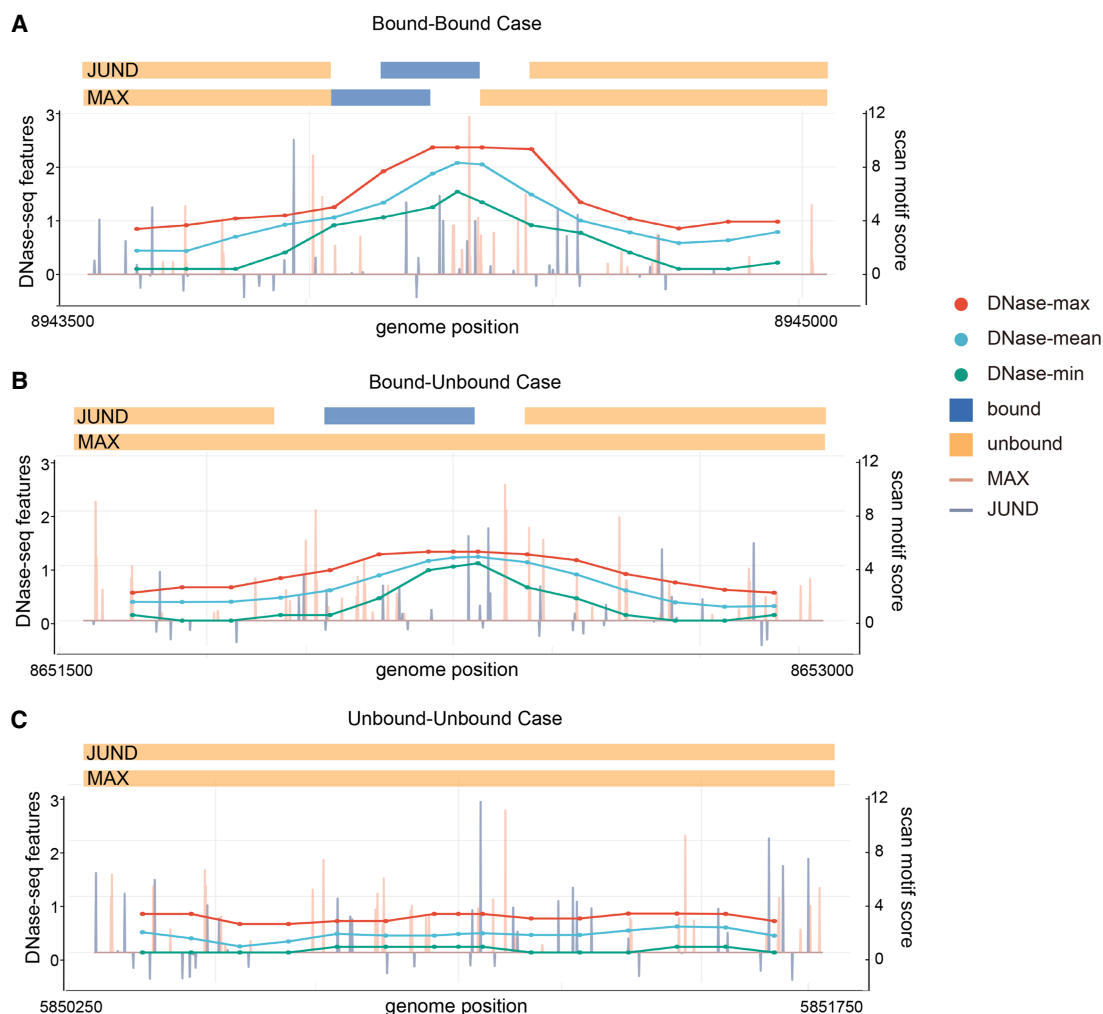


**Figure 3.** Performance of models with different features. (A) Comparison of performance between model without DNase-based feature and model with full features. (B) Comparison of performance between model with gene distance–based feature and model with full features. (C) Comparison of performance between model with DNase-based feature and model with full features. (D) Comparison of performance between model with sequence motif–based feature and model with full features. (E) Comparison of performance between logistic regression model with full features and XGBoost model with full features.

us to investigate the determinants of transcription factor co-occupancy beyond what is already known: open chromatin and the presence of motifs. A technical challenge in finding independent contribution is addressed by a recent advance in game theory application: an improved SHapley Additive exPlanations (SHAP) analysis (Fig. 1H,I; Shapley 1988; Lundberg et al. 2018). Mimicking the process of finding out the contribution of football players in a game, this analysis assigns the independent contribution of each of the features considering the context and the existence of other features. This is more appropriate than direct correlation analysis where correlations do not conclude direct contribution but could be a consequence of shared patterns with another important feature.

First, we analyzed the 526 features encompassing open chromatin (DNase), TF motif presence, other TF sequence signature, and gene location (Fig. 5A,B; Supplemental Fig. S1). We found that DNase and the presence of TF motifs under investigation are globally the most important factors, and the proximity to gene location plays a small role. Additionally, the DNase and TF features that are closer to the center of the 200-bp chunk would be more important (Fig. 5C–E).

To study the contribution of independent sequence signatures, for each TF pair, we calculated the SHAP value for each feature (Fig. 5F; Supplemental Figs. S2, S3). The original SHAP values for all motif features are a  $416 \times 56$  matrix. The 416 rows come from 13 TFs multiplied by 32, because each TF has a total



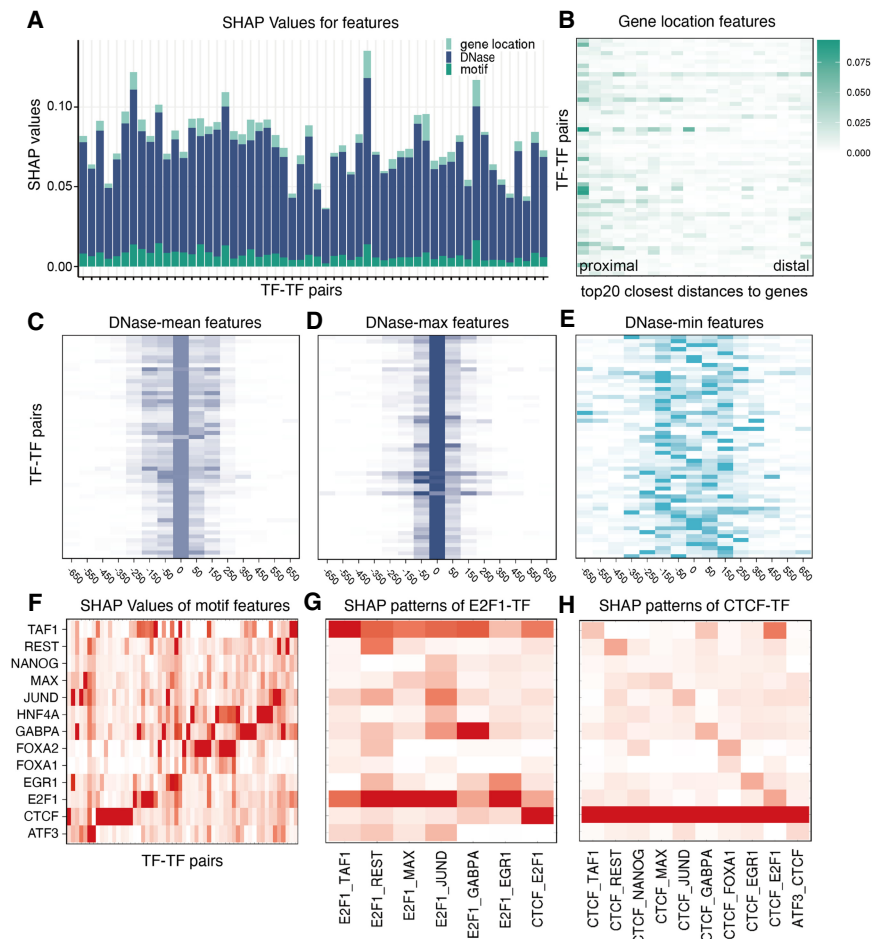
**Figure 4.** Examples of predicting TF-TF cobinding and un-cobinding events. The comparison of TF binding signals from ChIP-seq (the blue/orange horizontal bar on the *top*), DNase-seq-based features (the red/blue/green scatter points and lines in the *middle*), and TF motif hits (the blue/orange vertical bars on the *bottom*). (A) A cobound (Bound-Bound) case of the JUND-MAX pair in Chromosome 2 between position 8944200 and 8944400 in K562. The red, blue, and green dots represent the maximum, mean, and minimum DNase-seq values, respectively, in a 200-bp interval. The two bars on the *top* highlight the binding locations MAX and JUND. Our prediction for this interval is ‘cobound’ (0.9938). (B) An un-cobound (Bound-Unbound) case in Chromosome 2 between position 8652400 and 8652600. Specifically, JUND is bound while MAX is unbound in this interval. The prediction for this interval is ‘un-cobound’ (0.8174). (C) An un-cobound (Unbound-Unbound) case in Chromosome 2 between 5850900 and 5851100. Both JUND and MAX are unbound. The prediction for this interval is ‘un-cobound’ (0.1851).

number of 32 features, depending on their distance to the center of the chunk under investigation. We took the mean of those 32 features for 13 TFs, respectively, and scaled the SHAP values to 0–1 for each cell line-TF pair combination. An alternative by taking the maximum values was also investigated, and the patterns are very similar to taking the mean (Supplemental Fig. S4).

The first most prominent observation is that proximity to an open chromosome region and the appearance of motifs of the TF pair are the most important features. For example, for pairs tied with E2F1, E2F1 and TAF1 are the two most important features (Fig. 5G). For the effect of TAF1, overall it is a positive indicator for all E2F1-related pairs (Supplemental Fig. S3). Similarly, for pairs tied with CTCF, CTCF is the most important feature (Figs. 5H, 6E). For the DNase-mean and DNase-max features, the central interval is more important than neighboring intervals when predicting TF co-occupancy events. Additionally, proximity to genes only contributes to about half of the binding patterns of TF pairs.

However, we also found other irrelevant TF sequence signatures as significant contributors. In the majority (62.5%) of the TF pairs we studied, independent motif sequences contribute one or more TFs under investigation to the TF cobinding patterns. In seven cases among the 56 TF pairs we studied, there exist independent sequence signatures that contributed more than both TFs under investigation. Excluding cases related to FOXA1-FOXA2, which share great sequence similarity, there remain three cases (Table 2; Supplemental Fig. S5). Moreover, we found that, for seven pairs tied with ATF3, the average contribution of JUND is the highest among those 13 TFs; it is marginally higher than ATF3 itself. This is a strong piece of evidence of the existence of an independent TF-specific effector.

As mentioned above, some of these signatures could be a result of shared sequences or complementary sequences with the two TFs under study. For example, FOXA2 is an effector for pairs tied with FOXA1, and FOXA1 and FOXA2 share similar motifs.



**Figure 5.** Game theoretical approach-based analysis of determinants of TF cobinding. (A) SHAP values of DNase-seq-based features (blue), TF motif-based features (dark green), and distance-to-gene features (light green). DNase-seq-based features play the most important role in all TF-TF pairs. (B) The spatial distribution of distance-to-gene features. Columns correspond to the top 20 closest distances to proximal genes and rows are the 56 TF-TF pairs. Distances to proximal genes are more important than those to distal genes. (C–E) Spatial distributions of DNase-seq-based features. In the heat map, each row corresponds to a TF-TF pair and each column corresponds to one 200-bp genomic interval. Zero represents the feature calculated from the target interval, and  $\pm$  represents the upstream/downstream neighboring intervals. (F) The spatial distribution of TF motif-based features. Each column is a TF-TF pair and each row is the motif-based features averaging at one TF. The motifs-based features for TFs that are the component of the TF-TF pair have higher contributions. In addition to the component TF, other TF sequences are also predictive. (G) The SHAP value pattern of TF pairs tied with E2F1. For most pairs, E2F1 motif-based features have the highest SHAP values. For E2F1-TAF1, both E2F1 and TAF1 have higher SHAP values than other TFs. (H) The SHAP value pattern of pairs tied with CTCF. For most pairs, CTCF motif-based features have the highest SHAP values. Other TF motifs like NANOG and TAF1 also play an important role.

However, many effectors and their TFs do not directly show sequence similarity. For example, GABPA is an effector for pairs EGR1-TF (Fig. 6C), but the motif sequences of GABPA and EGR1 do not share many similarities, nor do they form a protein complex. Thus, some of these sequence signatures seem to speak a language other than forming complexes to facilitate certain shared processes. To investigate the contribution of cobinding in this feature importance analysis, we calculated two types of peak fractions: fractions in which effectors are bound with TF pairs in the same bin; and fractions in which effectors are bound with TF pairs within 300 bp (Fig. 6I; Supplemental Table S6) because a single binding event will result in consecutive sets of bins annotated to 'bound.' The above two criteria are, in fact, quite representative

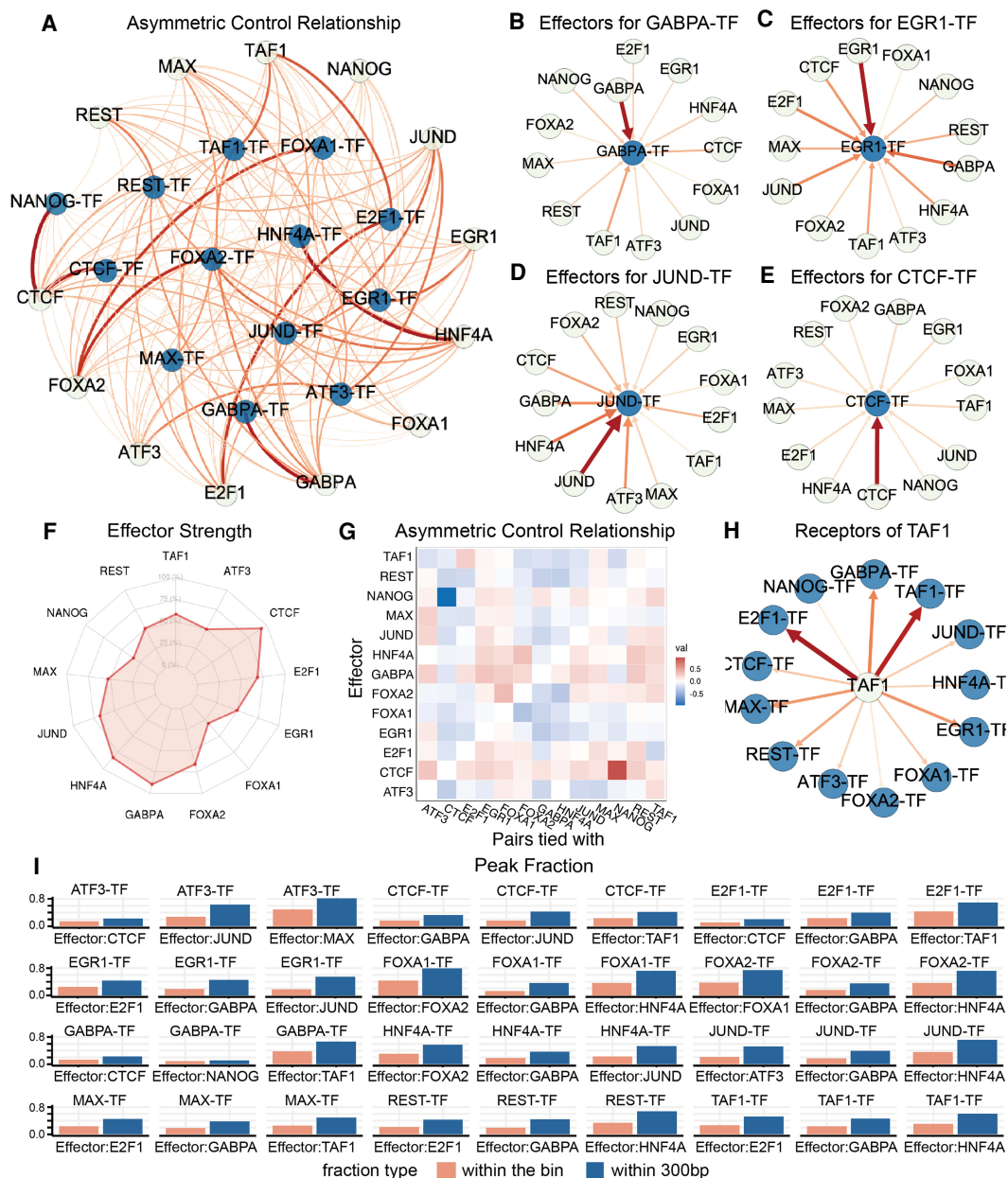
and generous in considering the contribution of cobinding. The average fraction of peaks that the top effector binds in the same bin with the TF-TF pair is 0.248. The average fraction of peaks in which the top effector binds with TF pairs within 300 bp is 0.490. For example, for ATF3-GABPA and effector MAX, the fraction that MAX bound with ATF3-GABPA in the same bin is 0.624, and the fraction that MAX bound with the pair within 300 bp is 0.893. For ATF3-TAF1 and effector CTCF, the fraction that CTCF bound with ATF3-TAF1 is 0.124, and the fraction that CTCF bound with the pair within 300 bp is 0.198. This indicates that combining is an important mechanism, resulting in the observed signatures. However, at the same time, even for the top effectors, the binding only contributes to an estimated 25% to, at most, 50% of the times, indicating that it only partially explains these sequence signatures. As a matter of fact, 27.06% of relationships between a sequence signature and a TF pair are negative, suggesting that there exist other mechanisms beyond binding.

#### Independent sequence signatures can be categorized into generic, TF-specific, and TF pair-specific

Such independent contributors of sequence signatures reflect, to some extent, known biology and could be roughly grouped into three categories, which are not necessarily mutually exclusive. First, the sequence signature is related to many TFs, for example, GABPA is related to 10 TFs (CTCF-TF, E2F1-TF, EGR1-TF, FOXA1-TF, FOXA2-TF, HNF4A-TF, JUND-TF, MAX-TF, REST-TF, and TAF1-TF) in this study, and JUND is related to four TFs (ATF3-TF, EGR1-TF, CTCF-TF, and HNF4A-TF) (Table 1). Certainly, in this case, JUND has been found to directly interact with ATF3 (Chu et al. 1994). TAF1 is an effector for pairs

CTCF-TF, E2F1-TF, GABPA-TF, and MAX-TF (Table 1). Transcription initiation factor TFIID subunit 1, TAF1, appears in this global effector list, as the initiation of transcription by RNA polymerase II requires the coordination of TFIID and the binding of this complex at the promoter region (Bieniossek et al. 2013). Of note, even though TAF1 is a global effector, its strength differs for different TFs, with the strongest in TAF1, E2F1, and GABPA (Fig. 6H). We term such effectors as generic effectors. Potential important ones are GABPA, JUND, TAF1, and HNF4A.

Second, for the sequence signature that is related to a particular TF, no matter what other TF we are studying, we term it a TF-specific effector (Table 1). To identify a TF-specific effector, we first calculated the mean of absolute SHAP values of each




**Figure 6.** Asymmetric regulatory relationships of TF-TF pairs in controlling cobinding. (A) In this network, each node represents a TF or a TF-TF pair. Each edge that connects a TF and a pair represents the contribution of that TF to the pair when predicting cobinding events. The width and color of an edge represent the strength of the contribution. (B–E) Directed regulatory networks of TF-TF pairs effectors. In each network, the blue node represents a TF-TF pair and the light color node represents a TF motif signature. The arrow direction represents the contribution of a TF to a TF-TF pair. The width and color of an edge represent the strength of the contribution. In B, GABPA, TAF1, and CTCF are the three most effective effectors for pairs tied with GABPA. In C, JUND, TAF1, and EGR1 are the most effective effectors for pairs tied with EGR1. In D, JUND is the most dominant effector for pairs tied with JUND. In E, CTCF is the most dominant effector for pairs tied with CTCF. The four examples show that the contribution of different TF effectors and both pair-relative and pair-irrelative TFs affect cobinding. (F) Radar plot of effectors strength averaging over pairs. Among the 13 TFs, CTCF, JUND, and TAF1 are the three most contributive effectors compared with other TFs. (G) Asymmetric control heat map. Each cell represents the difference between (1) the contribution of TF1 to pairs tied with TF2 and (2) the contribution of TF2 to pairs tied with TF1. Positive and negative values are shown in red and blue, respectively. (H) Contribution of TAF1 is different when predicting the cobinding of pairs tied with different TFs. (I) Peak fractions of TF-specific effectors binding with TF pairs. Red bars: the fraction of the effector binding in the bins under investigation; blue bars: the fraction of the effector binding within 300 bp of the bins under investigation.

effector on pairs that tied with the specific TF. Then, we selected the top three effectors, excluded the specific TF itself, and identified the remaining ones as TF-specific effectors. For example, HNF4A is a TF-specific effector for pairs tied with JUND (Fig. 6D), and TAF1 is a TF-specific effector for pairs tied with GABPA (Fig.

6B). TFs sharing strong sequence similarity will come out as an artifact of such TF-specific effector, for example, FOXA2 for FOXA1 and vice versa. Third, the sequence signature that is related to a particular TF pair, in the sense that it is stronger than any of the TFs of the pair under study, is termed as a TF pair-specific effector

**Table 1.** TF-specific effectors, which are related to cobinding patterns involving a specific TF, as identified for all TFs with more than two samples available; the TF under investigation is excluded in the table

Pairs tied with	TF-specific effector			
 ATF3	 JUND	 MAX	 CTCF	
 CTCF	 TAF1	 GABPA	 JUND	
 E2F1	 TAF1	 GABPA	 CTCF	
 EGR1	 GABPA	 E2F1	 JUND	
 FOXA1	 FOXA2	 HNF4A	 GABPA	
 FOXA2	 HNF4A	 FOXA1	 GABPA	
 GABPA	 TAF1	 CTCF	 NANOG	
 HNF4A	 JUND	 GABPA	 FOXA2	
 JUND	 HNF4A	 GABPA	 ATF3	
 MAX	 GABPA	 E2F1	 TAF1	
 REST	 GABPA	 HNF4A	 E2F1	
 TAF1	 GABPA	 E2F1	 HNF4A	




(Table 2). For example, HNF4A is the TF pair-specific effector for FOXA1-REST. We acknowledge this is an extremely stringent criterion, and if we relax it, many more TF pair-specific effectors may arise, and the full spectrum is provided in Supplemental Figures S2 and S3.

We found TF-specific and TF pair-specific effectors tend to be tissue-specific. FOXA2 is a TF that activates liver-specific gene expression (Tuteja et al. 2008). The cobinding of FOXA2-TAF1 is linked to the appearance of another tissue-specific transcription factor, HNF4A, which is critical for liver development (Babeu and Boudreau 2014; Thakur et al. 2019). Such tissue specificity seems to be elegantly orchestrated through sequence combination, in that the HNF4A motif comes out as an important effector for FOXA1, FOXA2, and TAF1. Similarly, FOXA2 does not appear as an effector for transcription factors other than FOXA1, which shares sequence identity with FOXA2.

We acknowledge one potential limitation in the above analysis: regions bound by a large number of transcription factors

might confound meaningful signals. To examine this effect, we repeated the above analysis by removing the HOT sites (High-Occupancy-Target sites) from the analysis. The HOT sites are defined by counting the number of binding TFs in each 200-bp bin and cutting off using the 99th percentile threshold to define the HOT sites (Wreczycka et al. 2019), which was deemed in line with previous studies (Gerstein et al. 2010). SHAP patterns excluding HOT sites do not differ substantially from the ones including the HOT sites (Supplemental Figs. S6–S13). The correlations between SHAP values with and without HOT sites range from 0.9939 to 0.9995. For example, for cell line K562, GABPA is the most dominant effector for JUND-TF pairs. Similarly, for cell line HepG2, FOXA2 is the most dominant effector for FOXA1-TF pairs. After removing the HOT sites, GABPA remains the top feature of the JUND-TF pair in K562, and FOXA2 also remains the top feature of FOXA1-TF pairs in HepG2. This result indicates the HOT sites had minimal impact on the feature importance analysis.

**Table 2.** TF pair-specific effectors, defined as TFs whose sum of absolute SHAP values is larger than both of the TFs of the pair; FOXA1 and FOXA2 are mutually excluded in this analysis due to high sequence similarity

TF pairs	TF pair-specific effector
ATF3-EGR1	 JUND
FOXA1-REST	 HNF4A
JUND-TAF1	 GABPA

### Independent sequence signatures for TF binding are rarely mutual but mostly directional

For both TF-specific effector and TF pair-specific effector, we observe striking cases where the independent TF sequence signature weights more than the original TFs in predicting cobinding. This motivated us to create a directed graph that represents such contribution relationships (Figs. 1J, 6). First, from the regulatory graph, we observe that the control relationship is rarely symmetric, with the exception of TFs that share strong sequence similarity, for example, FOXA1 and FOXA2. Although the motif signature of TF-A is important for another pair TF-B and TF-C, it is not necessary that TF-B and TF-C will be a strong signature of pairs involving TF-A (Fig. 6A,G). This refutes the possibility of significant artifacts caused by sequence signatures sharing sequence similarity with the TFs under investigation and supports the directional regulatory relationship.

Second, we observe sequence motifs that appear to be hubs in affecting many TFs (Fig. 6A,F). In the network (Fig. 6A), the width and color of each edge connecting a TF and a TF-TF pair represent the contribution of that TF motif features to the TF-TF pair. The weight value of an edge connecting TF-A and TF-B-TF is calculated by averaging the SHAP feature importance of all pairs tied with TF-B for TF-A's motif features. The strength of each sequence motif is calculated by averaging weights connected to that motif in the network (Fig. 6F). From the plot, CTCF, GABPA, and HNF4A are the three most contributive effectors (Fig. 6F), and GABPA, HNF4A, TAF1, and JUND affect many TF-TF pairs in this network under investigation (Table 1), whereas others control more specific and fewer TFs and TF pairs. Overall, this graph supports a hierarchical organization of the regulatory network that is orchestrated by a few sequence signatures and rewires into more specific functions and specificity through combinations of more specific sequence signatures.

## Discussion

Unlike single TF binding, the study of cobinding is a less explored area. Although chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) has been applied to the combination among 10 cell lines and 13 transcription factors in the ENCODE database, experimentally measuring the genome-wide co-occupancy of every TF pair in every cell type is practically infeasible. Thus, characterizing the TF cobinding patterns has been an open scientific question for years. Existing software such as TACO and coTraCTE also used DNase-seq and TF motifs to predict TF-TF co-occurrence (Jankowski et al. 2014; van Bömmel et al. 2018).

TACO modeled TF motif complex enrichment based on cell type-specific open chromatin regions from the ENCODE DNase-seq data set. Similarly, coTraCTE first defined cell type-specific DNase Hypersensitive Sites (DHSs) through the *t*-statistic measure. Then, the TRanscription factor Affinity Prediction method (TRAP) (Roeder et al. 2007) was used to estimate the TF binding sites among the DHSs based on TF motifs. In this work, we further leveraged the large-scale TF ChIP-seq data to build machine learning models for TF-TF co-occurrence detection. Consistent with previous studies, we also found that open chromatin and TF motifs play central roles.

Specifically, we developed machine learning methods to dissect the contributors to TF cobinding. Our focus is on dissecting the independent sequence contributors. The design of the algorithm thus included features of other TFs irrelevant to the ones under investigation. This experiment resulted in the finding of independent sequence signatures that contribute as strongly as or even more strongly than the TFs under investigation. This analysis also allowed us to construct a directed graph of how these sequence signatures affect each other.

Such sequence signatures can be roughly divided into three categories: TF pair-specific, TF-specific, and TF-generic. For example, TAF1 is an important signature for a range of transcription factor binding patterns. On the other hand, TF-specific and TF pair-specific sequence signatures tend to involve TFs controlling more specific biological processes. For example, the cobinding patterns of FOXA2, a TF that activates liver-specific gene expression, are affected by HNF4A, another liver-specific TF. The result from SHAP analysis and network analysis overall suggests a hierarchy of organization of sequence signatures that coordinates to fulfill global functionality of transcription by global effectors as well as tissue specificity by the interplay of binding sites of tissue-specific TFs. Analysis of feature importance could be affected by a variety of factors such as the frequency of occurrence of a feature and the base learners. Upscaled and more comprehensive studies covering these areas could be carried out in the future.

The human genome encodes for a limited number of 1500 transcription factors (Ignatieva et al. 2015), but they are sufficient to facilitate the formation of cellular diversity and development dynamics. One key is the combinatorial interplays of multiple TFs. This study identifies the sequence signatures by using a machine learning approach and statistical analysis and contributes to our understanding of the rules that drive such interplay.

## Methods

### Data sets for model building and testing

In this study, we used the DNase-seq and ChIP-seq data from the ENCODE Project, covering 13 TFs (ATF3, CTCF, E2F1, EGR1, FOXA1, FOXA2, GABPA, HNF4A, JUND, MAX, NANOG, REST, and TAF1) in 10 cell types (A549, GM12878, H1-hESC, HCT116, HeLa-S3, HepG2, K562, MCF-7, iPSC, liver). The ChIP-seq data were downloaded from <https://www.synapse.org/#!Synapse:syn6181337> (conservative peaks) and the ENCODE Project website (<https://www.encodeproject.org/>). The accession IDs are listed in Supplemental Table S1. The DNase-seq data were downloaded from <https://www.synapse.org/#!Synapse:syn6176232>. The data were processed following the standard ENCODE analysis pipeline (Data Processing Pipelines 2020, <https://www.encodeproject.org/pipelines/>, accessed October 5, 2020). The reference human genome is GRCh37. The conclusions in this manuscript would not be significantly affected if using GRCh38 as the reference genome. GRCh38 has improvements over GRCh37 such as annotation of

the centromere regions and the addition of alternate loci, but these differences should not greatly impact the patterns of transcription factor cobinding. In particular, the DNase-seq data provided the cell type-specific chromatin accessibility, which was highly associated with TF binding events, and we used the filtered alignment files of DNase-seq. The ChIP-seq data provided the single TF binding sites across various cell types. For each 200-bp interval sliding every 50 bp in the human genome, we defined three types of binding labels—"bound" (B); "unbound" (U); and "ambiguous" (A)—by overlapping 200-bp intervals with peaks generated from SPP peak caller (Kharchenko et al. 2008) at the irreproducible discovery rate (IDR) cutoff of 5% (Li et al. 2011; Bionetworks S. Synapse | Sage Bionetworks, <https://www.synapse.org/#!Synapse:syn6131484/wiki/402033>, accessed November 1, 2019). We further defined a "cobound" genomic bin of a TF-TF pair when the single TF ChIP-seq labels from both TFs were "bound," serving as the positive examples. When only one TF ChIP-seq label is "bound," or none of the TFs is bound, the interval is considered as a negative example.

### Data set partition for cross-cell type and cross-chromosome training and testing

In this study, we used a total of 228 TF co-occupancy profiles involving 56 different TF-TF pairs. For each TF-TF pair, we randomly held out one cell type for testing and the other cell types for model building. In this way, the performance of our models was evaluated in a cross-cell type manner (Supplemental Fig. S14; Supplemental Table S2). To build a robust and generalizable model for predicting TF co-occupancy, we designed a "crisscross" strategy to exploit the training data and avoid overfitting (Fig. 1C). In particular, we partitioned the cell types and chromosomes into the training set for model building and the validation set for hyperparameter tuning. This is because we used XGBoost, an iteration-based machine learning model, which required the validation-based early stopping. First, the 20 training chromosomes were randomly split into set1 (Chr 2, Chr 4, Chr 6, Chr 7, Chr 12, Chr 13, Chr 15, Chr 16, Chr 17, Chr 20, Chr X) and set2 (Chr 3, Chr 5, Chr 9, Chr 10, Chr 11, Chr 14, Chr 18, Chr 19, Chr 22). The model was trained on one chromosome set and validated on the other set to avoid overfitting on chromosomes. Second, for a TF-TF pair with  $N$  training cell types,  $2N$  XGBoost models would be built. The first model was trained on chromosome set1 in cell type 1, and validated on chromosome set2 in cell type 2. Similarly, the  $k$ th model was trained on chromosome set1 in cell type  $k$  and validated on chromosome set2 in cell type  $k+1$ . The  $N$ th model was trained on chromosome set1 in cell type  $N$ , and validated on chromosome set2 in cell type 1. In this way, we trained the model in one cell type but validated it against a different cell type, which improved the generalizability of our method in unseen cell types. After that, we switched chromosome set1 and set2 by training a model on chromosome set2 in cell type  $k$  and validating on chromosome set1 in cell type  $k+1$ . In this way, we obtained another  $N$  models. Finally, predictions from the  $2N$  models were averaged as the final prediction.

An additional test was carried out for cross-chromosome evaluation. In this test, different from cross-cell type evaluation which tested on the 20 training chromosomes, we tested on the three left out chromosomes (Chr 1, Chr 8, Chr 21). Of note, the two approaches of cross-validation were aimed to assess model transferability in different scenarios. The follow-up feature analysis focused on cross-cell prediction models.

The distributions of single TFBSs and co-occupied TFBSs were investigated (Supplemental Table S7; Supplemental Figs. S15–S17). To validate whether a TF-TF pair was significantly co-occupied, we applied a nonparametric paired Wilcoxon signed-rank (Wilcoxon 1945) test to all 56 TF pairs after pairwise comparison selection.

The overall distributions of co-occupancy counts and  $P$ -values (Supplemental Table S8; Supplemental Figs. S18, S19) are shown in Supplemental Figure S19, C–E. The results showed that all pairs detected were significantly cooperative (all  $P$ -value < 0.0001, with family-wise error rate 0.01), which also corresponds to literature reports. In particular, the ratio of observed co-occupancy between FOXA1 and FOXA2 was significantly higher than other TF pairs, which may be attributed to the similarity in their motifs (Kulakovskiy et al. 2018) and regulatory functions (van der Sluis et al. 2008; Li et al. 2012). We also observed a high ratio of co-occupancy between HNF4A and FOXA2 (Supplemental Fig. S19B,E). In fact, the molecular interaction between them has been reported previously (Wallerman et al. 2009).

### Feature extraction

A total of 90 DNase-seq-based features, 416 TF motif-based features, and 20 distance-to-gene features were used in our model. To remove the potential batch effect, the original DNase-seq filtered alignment was quantile-normalized before extracting DNase-seq-based features. For each 200-bp genomic bin, the maximum, minimum, and mean DNase-seq values were calculated as 3M-DNase features, and  $\Delta$ maximum,  $\Delta$ minimum, and  $\Delta$ mean DNase-seq values were extracted as  $\Delta$ 3M-DNase features (Li et al. 2019). The  $\Delta$ 3M-DNase features were the difference between the DNase-seq signal of a specific cell line and the average signals of all 10 cell lines used in this study. We further considered 14 upstream and downstream neighboring bins to extract the corresponding 3M-DNase and  $\Delta$ 3M-DNase features, resulting in a total of  $90 = (3 + 3) \times 15$  DNase-seq-based features. This neighboring information is also reported to improve predictive performance in recent studies.

In addition to the DNase-seq-based features, we also integrated TF motif features of all 13 TFs in the models. Specifically, for each TF, the position weight matrix (PWM) (Stormo et al. 1982) can be represented by an  $a$ -by-4 matrix  $M$ ,

$$\begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ \vdots & \vdots & \vdots & \vdots \\ m_{a1} & m_{a2} & m_{a3} & m_{a4} \end{bmatrix}$$

where  $a$  is the length of the TF motif. Similarly, DNA sequence can be one-hot encoded and represented by a 4-by- $b$  matrix  $X$ ,

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1b} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2b} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3b} \\ x_{41} & x_{42} & x_{43} & \dots & x_{4b} \end{bmatrix}$$

where  $b$  is the number of nucleotides in the human genome and four rows represent four types of nucleotides A/C/G/T. Each item in  $X$  was a binary value and for each column, or nucleotide position, the row where  $x = 1$  corresponded to the nucleotide type in that position. By multiplying  $M$  and  $X$ , we can obtain an  $a$ -by- $b$  matrix  $Y$ ,

$$\begin{bmatrix} y_{11} & y_{12} & \dots & y_{1a} & y_{1(a+1)} & \dots & y_{1b} \\ y_{21} & y_{22} & \dots & y_{2a} & y_{2(a+1)} & \dots & y_{2b} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{a1} & y_{a2} & \dots & y_{aa} & y_{a(a+1)} & \dots & y_{ab} \end{bmatrix}$$

trace s1
trace s2
...

where  $a$  is the length of the TF motif and  $b$  is the length of a chromosome. In order to obtain the TF motif match score for each nucleotide position, we calculated the trace of the submatrix  $Y[, i:(i+a-1)]$  for  $i = 1, 2, \dots, b-a+1$ . For example, when  $i = 1$ , the trace of the submatrix within the red rectangle was calculated as  $s_1$ . Similarly, when  $i = 2$ , the trace of the submatrix within the blue rectangle was calculated as  $s_2$ . Therefore, for each TF, obtained a 1-by- $(b-a+1)$  score vector  $S = [s_1 \ s_2 \ s_3 \ \dots \ s_{b-a+1}]$ . This score vector  $S$  was further padded with zeros at both ends to the length  $a$ . Then, we extracted bin-level motif-based features from this vector  $S$ . For each 200-bp interval, the top three scores among that 200 positions were saved. Similarly, the top three scores of the product between TF PWM and the reverse complement of the genome sequence were calculated. Then, the top four of the two sets of top three scores were considered as motif features. Similar to the DNase-seq-based features, we also considered eight neighboring intervals and generated a total of  $32 = 4 \times 8$  motif features for each TF. To consider the potential interactions of all TFs, we used the 32 motif features from all 13 TFs under consideration in this study. In addition to the DNase-seq-based and motif-based features, we further extracted the distance-to-gene features. Specifically, for each 200-bp interval, we calculated the top 20 closest distances to proximal genes based on GENCODE annotation (Li et al. 2019).

### Using XGBoost to model nonlinear interactions between features and to facilitate subsequent game theory-based feature importance analysis

The tree-based XGBoost models (Chen and Guestrin 2016) were trained on a total of 526 ( $90 + 416 + 20$ ) features mentioned above. The parameters of XGBoost are: (1) the step size shrinkage is 1; (2) the maximum depth of a tree is 7; (3) the minimum sum of weights in a child is 5; and (4) the minimum loss reduction is 0.1. We used cross-entropy loss, as the label is binary, specifically:  $-[y \times \log(y_{\text{pred}}) + (1-y) \times \log(1-y_{\text{pred}})]$ , where  $y$  represents the true label and  $y_{\text{pred}}$  represents the prediction value. The tree-based models can learn efficiently the nonlinear interactions between features and are robust to noise and outliers in data (Li et al. 2018; Li and Guan 2019b). Furthermore, we set the maximum number of iterations to 1000 and applied a validation-based early stopping strategy in a crisscross fashion, across both cell types and chromosomes. Final predictions were obtained by averaging the results of all models.

### Evaluating predictive performance with AUROC and AUPRC

To evaluate the predictive performance of our model, we used the area under the receiver operating characteristic curve and the area under the precision recall curve as the primary scoring metrics (Davis and Goadrich 2006). For each cutoff in the AUROC and AUPRC curves, true positive (TP) was the number of correctly predicted cobound bins, false positive (FP) was the number of predicted cobound bins whose true labels were un-cobound, true negative (TN) was the number of correctly predicted un-cobound bins, and false negative (FN) was the number of predicted un-cobound bins whose true labels were cobound.

### Benchmark comparison

We benchmarked the performance of TF cobinding prediction using methods TACO and coTRaCTE. Briefly, these two methods focus on open chromatin regions and estimate TF cobinding enrichment based on motif information. They assign a  $P$ -value for each genomic region under consideration. For TACO, we defined the prediction score by multiplying the  $P$ -values of two TFs

in the same region. For coTRaCTE, we first transformed the original  $P$ -value of a TF pair into  $-\log_{10}(P\text{-value})$ . Then, we rescaled them to the range between zero and one and used one minus the rescaled value as the prediction score.

### SHAP analysis to identify independent sequence signatures

We used SHAP (Shapley 1988; Lundberg et al. 2018) to interpret the output of our model and reveal independent sequence signatures' contributions to cobinding events. The typical feature importance calculation in the gradient boosting model is based on the improvement of each attribute in the performance (Chen and Guestrin 2016; Xia et al. 2017). In contrast, SHAP is based on the magnitude of feature attributions (Molnar 2019). We used SHAP feature importance, which is measured as the mean absolute Shapley values, in our analysis.

For our model, we calculated a  $526 \times 56$  matrix containing SHAP values for 526 features and 56 pairs. For a specific TF or TF-TF pair, we calculated the relative feature importance by analyzing the corresponding submatrix. Additionally, to reveal the control relationship between sequence motif features and TF-TF pairs, we assigned SHAP values to a directed regulatory network, and the edges with different weights were indicated by widths. In the network, each node represents a TF or a TF-TF pair and each edge represents the SHAP feature importance of that sequence feature to the TF-TF pair. For an edge connecting 'TF1' and 'TF2-TF', the weight was calculated by averaging SHAP values of the TF1 motif feature predicting all pairs tied with TF2. For example, there are seven pairs tied with FOXA2 (FOXA2-TAF1, FOXA2-REST, FOXA2-AMX, FOXA2-JUND, FOXA2-HNF4A, FOXA2-GABPA, and FOXA1-FOXA2) (Supplemental Fig. S2); we only selected a submatrix of the corresponding seven columns from the complete SHAP value matrix and calculated the average absolute SHAP value for each TF as the weights of the edges connecting the nodes TF' and 'FOXA2-TF'.

### Software availability

The source code of the analysis is available as Supplemental Code and at GitHub ([https://github.com/GuanLab/Sequence\\_Analysis\\_for\\_TF-co-binding](https://github.com/GuanLab/Sequence_Analysis_for_TF-co-binding)).

### Competing interest statement

The authors declare no competing interests.

### Acknowledgments

This work is supported by the National Science Foundation (NSF#1452656) and the National Institutes of Health (R35-GM133346).

*Author contributions:* Y.G. directed the study; M.Z. and H.L. performed the experiments; M.Z. and X.W. created the figures; M.Z., Y.G., and H.L. wrote the manuscript. All authors have read and approved the manuscript.

### References

- Ang Y-S, Rivas RN, Ribeiro AJS, Srivas R, Rivera J, Stone NR, Pratt K, Mohamed TMA, Fu J-D, Spencer CI, et al. 2016. Disease model of GATA4 mutation reveals transcription factor cooperativity in human cardiogenesis. *Cell* **167**: 1734–1749.e22. doi:10.1016/j.cell.2016.11.033
- Babeu J-P, Boudreau F. 2014. Hepatocyte nuclear factor 4- $\alpha$  involvement in liver and intestinal inflammatory networks. *World J Gastroenterol* **20**: 22–30. doi:10.3748/wjg.v20.i1.22
- Bieniossek C, Papai G, Schaffitzel C, Garzoni F, Chaillet M, Scheer E, Papadopoulos P, Tora L, Schultz P, Berger I. 2013. The architecture of

- human general transcription factor TFIID core complex. *Nature* **493**: 699–702. doi:10.1038/nature11791
- Blanchette M, Bataille AR, Chen X, Poitras C, Laganière J, Lefebvre C, Deblouis G, Giguère V, Ferretti V, Bergeron D, et al. 2006. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res* **16**: 656–668. doi:10.1101/gr.4866006
- Bulyk ML. 2003. Computational prediction of transcription-factor binding site locations. *Genome Biol* **5**: 201. doi:10.1186/gb-2003-5-1-201
- Chen T, Guestrin C. 2016. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. San Francisco. ACM Press, New York.
- Chen X, Yu B, Carriero N, Silva C, Bonneau R. 2017. Mocap: large-scale inference of transcription factor binding sites from chromatin accessibility. *Nucleic Acids Res* **45**: 4315–4329. doi:10.1093/nar/gkx174
- Chen C-H, Zheng R, Tokheim C, Dong X, Fan J, Wan C, Tang Q, Brown M, Liu JS, Meyer CA, et al. 2020. Determinants of transcription factor regulatory range. *Nat Commun* **11**: 2472. doi:10.1038/s41467-020-16106-x
- Chu HM, Tan Y, Kobierski LA, Balsam LB, Comb MJ. 1994. Activating transcription factor-3 stimulates 3',5'-cyclic adenosine monophosphate-dependent gene expression. *Mol Endocrinol* **8**: 59–68. doi:10.1210/mend.8.1.8152431
- Clements M, van Someren EP, Knijnenburg TA, Reinders MJT. 2007. Integration of known transcription factor binding site information and gene expression data to advance from co-expression to co-regulation. *Genomics Proteomics Bioinformatics* **5**: 86–101. doi:10.1016/S1672-0229(07)60019-9
- Davie K, Jacobs J, Atkins M, Potier D, Christiaens V, Halder G, Aerts S. 2015. Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. *PLoS Genet* **11**: e1004994. doi:10.1371/journal.pgen.1004994
- Davis J, Goadrich M. 2006. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240. Pittsburgh. ACM, New York.
- Ezer D, Zabet NR, Adryan B. 2014. Homotypic clusters of transcription factor binding sites: a model system for understanding the physical mechanics of gene expression. *Comput Struct Biotechnol J* **10**: 63–69. doi:10.1016/j.csbj.2014.07.005
- Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**: 1775–1787. doi:10.1126/science.1196914
- Ignatieva EV, Levitsky VG, Kolchanov NA. 2015. Human genes encoding transcription factors and chromatin-modifying proteins have low levels of promoter polymorphism: a study of 1000 Genomes Project data. *Int J Genomics Proteomics* **2015**: 260159. doi:10.1155/2015/260159
- Jankowski A, Prabhakar S, Tiuryn J. 2014. TACO: a general-purpose tool for predicting cell-type-specific transcription factor dimers. *BMC Genomics* **15**: 208. doi:10.1186/1471-2164-15-208
- Keilwagen J, Posch S, Grau J. 2019. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol* **20**: 9. doi:10.1186/s13059-018-1614-y
- Kelley DR. 2020. Cross-species regulatory sequence activity prediction. *PLoS Comput Biol* doi:10.1101/660563
- Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26**: 990–999. doi:10.1101/gr.200535.115
- Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* **26**: 1351–1359. doi:10.1038/nbt.1508
- Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, Medvedeva YA, Magana-Mora A, Bajic VB, Papatsenko DA, et al. 2018. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* **46**: D252–D259. doi:10.1093/nar/gkx1106
- Kumar S, Bucher P. 2016. Predicting transcription factor site occupancy using DNA sequence intrinsic and cell-type specific chromatin features. *BMC Bioinformatics* **17**: 4. doi:10.1186/s12859-015-0846-z
- Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The human transcription factors. *Cell* **175**: 598–599. doi:10.1016/j.cell.2018.09.045
- Lemon B, Tjian R. 2000. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev* **14**: 2551–2569. doi:10.1101/gad.831000
- Levy S, Hannehalli S. 2002. Identification of transcription factor binding sites in the human genome sequence. *Mamm Genome* **13**: 510–514. doi:10.1007/s00335-002-2175-6
- Li H, Guan Y. 2019a. Leopard: fast decoding cell type-specific transcription factor binding landscape at single-nucleotide resolution. bioRxiv doi:10.1101/856823
- Li H, Guan Y. 2019b. Machine learning empowers phosphoproteome prediction in cancers. *Bioinformatics* **36**: 859–864. doi:10.1093/bioinformatics/btz639
- Li Q, Brown JB, Huang H, Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5**: 1752–1779. doi:10.1214/11-aos466
- Li Z, Tuteja G, Schug J, Kaestner KH. 2012. Foxa1 and Foxa2 are essential for sexual dimorphism in liver cancer. *Cell* **148**: 72–83. doi:10.1016/j.cell.2011.11.026
- Li P, Spolski R, Liao W, Leonard WJ. 2014. Complex interactions of transcription factors in mediating cytokine biology in T cells. *Immunol Rev* **261**: 141–156. doi:10.1111/immr.12199
- Li H, Panwar B, Omenn GS, Guan Y. 2018. Accurate prediction of personalized olfactory perception from large-scale chemoinformatic features. *Gigascience* **7**: 1–11. doi:10.1093/gigascience/gix127
- Li H, Quang D, Guan Y. 2019. Anchor: trans-cell type prediction of transcription factor binding sites. *Genome Res* **29**: 281–292. doi:10.1101/gr.237156.118
- Liu L, Zhao W, Zhou X. 2016. Modeling co-occupancy of transcription factors using chromatin features. *Nucleic Acids Res* **44**: e49. doi:10.1093/nar/gkv1281
- Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK-W, Newman S-F, Kim J, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* **2**: 749–760. doi:10.1038/s41551-018-0304-0
- Maienschein-Cline M, Zhou J, White KP, Sciammas R, Dinner AR. 2012. Discovering transcription factor regulatory targets using gene expression and binding data. *Bioinformatics* **28**: 206–213. doi:10.1093/bioinformatics/btr628
- Molnar C. 2019. *Interpretable machine learning*. Leanpub, n.p. Available at <https://christophm.github.io/interpretable-ml-book/>.
- Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al. 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**: 83–90. doi:10.1038/nature11212
- Perez-Pinera P, Oustersout DG, Brunger JM, Farin AM, Glass KA, Guilak F, Crawford GE, Hartemink AJ, Gersbach CA. 2013. Synergistic and tunable human gene activation by combinations of synthetic transcription factors. *Nat Methods* **10**: 239–242. doi:10.1038/nmeth.2361
- Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* **21**: 447–455. doi:10.1101/gr.112623.110
- Quach B, Furey TS. 2017. DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. *Bioinformatics* **33**: 956–963. doi:10.1093/bioinformatics/btw740
- Roider HG, Kanhere A, Manke T, Vingron M. 2007. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* **23**: 134–141. doi:10.1093/bioinformatics/btl565
- Schmidt F, Gasparoni N, Gasparoni G, Gianmoena K, Cadenas C, Polansky JK, Ebert P, Nordström K, Barann M, Sinha A, et al. 2017. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res* **45**: 54–66. doi:10.1093/nar/gkw1061
- Shapley LS. 1988. A value for n-person games. In *The Shapley value* (ed. Roth AE), pp. 31–40. Cambridge Univ. Press, Cambridge. doi:10.1017/cbo9780511528446.003
- Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. 1982. Use of the “Perceptron” algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res* **10**: 2997–3011. doi:10.1093/nar/10.9.2997
- Thakur A, Wong JCH, Wang EY, Lotto J, Kim D, Cheng J-C, Mingay M, Cullum R, Moudgil V, Ahmed N, et al. 2019. Hepatocyte nuclear factor 4- $\alpha$  is essential for the active epigenetic state at enhancers in mouse liver. *Hepatology* **70**: 1360–1376. doi:10.1002/hep.30631
- Tsompana M, Buck MJ. 2014. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* **7**: 33. doi:10.1186/1756-8935-7-33
- Tuteja G, Jensen ST, White P, Kaestner KH. 2008. Cis-regulatory modules in the mammalian liver: composition depends on strength of Foxa2 consensus site. *Nucleic Acids Res* **36**: 4149–4157. doi:10.1093/nar/gkn366
- Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5**: 829–834. doi:10.1038/nmeth.1246
- van Bömmel A, Love MI, Chung H-R, Vingron M. 2018. coTraCTE predicts co-occurring transcription factors within cell-type specific enhancers. *PLoS Comput Biol* **14**: e1006372. doi:10.1371/journal.pcbi.1006372

- van der Sluis M, Vincent A, Bouma J, Korteland-Van Male A, van Goudoever JB, Renes IB, Van Seuningen I. 2008. Forkhead box transcription factors Foxa1 and Foxa2 are important regulators of Muc2 mucin expression in intestinal epithelial cells. *Biochem Biophys Res Commun* **369**: 1108–1113. doi:10.1016/j.bbrc.2008.02.158
- Wallerman O, Motallebipour M, Enroth S, Patra K, Bysani MSR, Komorowski J, Wadelius C. 2009. Molecular interactions between HNF4a, FOXA2 and GABP identified at regulatory DNA elements through ChIP-sequencing. *Nucleic Acids Res* **37**: 7498–7508. doi:10.1093/nar/gkp823
- Wilcoxon F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* **1**: 80. doi:10.2307/3001968
- Wreczycka K, Franke V, Uyar B, Wurmus R, Bulut S, Tursun B, Akalin A. 2019. HOT or not: examining the basis of high-occupancy target regions. *Nucleic Acids Res* **47**: 5735–5745. doi:10.1093/nar/gkz460
- Xia Y, Liu C, Li Y, Liu N. 2017. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. *Expert Syst Appl* **78**: 225–241. doi:10.1016/j.eswa.2017.02.017
- Yu C-P, Lin J-J, Li W-H. 2016. Positional distribution of transcription factor binding sites in *Arabidopsis thaliana*. *Sci Rep* **6**: 25164. doi:10.1038/srep25164
- Zhou Q, Liu JS. 2004. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* **20**: 909–916. doi:10.1093/bioinformatics/bth006

Received June 12, 2020; accepted in revised form December 3, 2020.