Modeling transcription factor combinatorics in promoters and enhancers

Jimmy Vandel [1,2], Océane Cassan [1,2], Sophie Lèbre [1,3,†], Charles-Henri Lecellier [1,4,†,*], Laurent Bréhélin [1,2,†*]

[1] Institut de Biologie Computationnelle, Montpellier, France , [2] LIRMM, CNRS, Univ. Montpellier, Montpellier, France , [3] IMAG, CNRS, Univ. Montpellier, Montpellier, France , [4] IGMM, CNRS, Univ. Montpellier, Montpellier, France , [†]These authors contributed equally to this work

ABSTRACT

We propose a new approach (TFcoop) that takes into account cooperation between transcription factors (TFs) for predicting TF binding sites. For a given a TF, TFcoop bases its prediction upon the binding affinity of the target TF as well as any other TF identified as cooperating with this TF. The set of cooperating TFs and the model parameters are learned from ChIP-seq data of the target TF. We used TFcoop to investigate the TF combinations involved in the binding of 106 different TFs on 41 different cell types and in four different regulatory regions: promoters of mRNAs, lncRNAs and pri-miRNAs, and enhancers. Our experiments show that the approach is accurate and outperforms simple PWM methods. Moreover, analysis of the learned models sheds light on important properties of TF combinations. First, for a given TF and region, we show that TF combinations governing the binding of the target TF are similar for the different cell-types. Second, for a given TF, we observe that TF combinations are different between promoters and enhancers, but similar for promoters of distinct gene classes (mRNAs, lncRNAs and miRNAs). Analysis of the TFs cooperating with the different targets show over-representation of pioneer TFs and a clear preference for TFs with binding motif composition similar to that of the target. Lastly, our models accurately distinguish promoters into classes associated with specific biological processes.

INTRODUCTION

Transcription factors (TFs) are regulatory proteins that bind DNA to activate or repress target gene transcription. TFs play a central role in controlling biological processes, and are often mis-regulated in diseases (22). Technological developments over the last decade have allowed the characterization of binding preferences for many transcription factors both *in vitro* (4, 16) and *in vivo* (15). These analyses have revealed that a given TF usually binds similar short nucleic acid sequences that are thought to be specific to this TF, and that are conserved along evolution (5). In addition to enabling characterization of the sequence specificity of TF binding sites (TFBS), *in vivo* approaches such as ChIP-seq also have the potential to precisely identify the position of these binding sites genome-wide, in a particular biological condition (cell type or treatment). While consortiums such as ENCODE (40) have generated hundred of ChIP-seq datasets for different TFs under different conditions, it is not possible to provide data for every TFs in every possible biological condition. Therefore, accurate computational approaches are needed to complement experimental results. Also, a biological explanation is missing: knowing where a TF binds in the genome does not explain why it binds there. Understanding TF binding involves developing biophysical or mathematical models able to accurately predict TFBSs.

Traditionally, TFBSs have been modeled by position weight matrices (PWMs) (43). These models have the benefit of being simple, easy to visualize, and they can be deduced from *in vitro* and *in vivo* experiments. As a result, several databases such as JASPAR (25), HOCOMOCO (21), and Transfac (44), propose position frequency matrices (PFM, which can then be transformed in PWMs) for hundred of TFs. These PWMs can be used to scan sequences and identify TFBSs using tools such as FIMO (12) or MOODS (20). However, to a certain extent, PWMs lack sensitivity and accuracy. While the majority of TFs seem to bind to a unique TFBS motif, *in vivo* and *in vitro* approaches have revealed that some TFs recognize different motifs (17, 32). In that case, one single PWM is not sufficient to identify all potential binding sites of a given TF. Moreover, while a PWM usually identifies thousands

*To whom correspondence should be addressed. Emails: brehelin@lirmm.fr and charles.lecellier@igmm.cnrs.fr

of potential binding sites for a given TF in the genome (45), ChIP-seq analyses have revealed that only a fraction of those sites are effectively bound (19). There may be different reasons for this discrepancy between predictions and experiments. First, PWMs implicitly assume that the positions within a TFBS independently contribute to binding affinity. Several approaches have thus been proposed to account for positional dependencies within the TFBS (see for example (27, 47)). Other studies have focused on the TFBS genomic environment, revealing that TFs seems to have a preferential nucleotide content in the flanking positions of their core binding sites (9, 23).

Second, beyond the primary nucleotide sequence, other structural or epigenetic data may also affect TF binding. For example, it is thought that TFs use DNA shape features (e.g., minor groove width and rotational parameters such as helix twist, propeller twist and roll) to distinguish binding sites with similar DNA sequences (36). The contributions of base and shape composition for TFBS recognition vary across TF families, with some TFs influenced mainly by base composition, and some other TFs influenced mostly by DNA shape (36). Some attempts have thus been made to integrate DNA shapes information with PWMs (26). Other studies have investigated the link between TF binding and epigenetic marks, showing that many TFs bind regions associated with specific histone marks (10). However, it remains unclear whether these chromatin states are a cause or a consequence of TF binding (14). Similarly, ChIP-seq experiments also revealed that most TFBSs fall within highly accessible (*i.e.*, nucleosome-depleted) DNA regions (41). Consequently, several studies have proposed to supplement PWM information with DNA accessibility data to identify the active TFBS in a given cell type (24, 31, 38). However, as for other epigenetic marks, DNA accessibility can also be a consequence of TF binding rather than its cause. For example, Sherwood et al. (39) used DNase-seq data to distinguish between "pioneer" TFs, which open chromatin dynamically, and "settler" TFs, which only bind already-opened chromatin.

Competition and cooperation between TFs (combinatorics) can also impact the binding capability of a given TF. As reviewed in Morgunova et al. (30), multiple mechanisms can lead to TF cooperation. In its simplest form, cooperation involves direct TF-TF interactions before any DNA binding. But cooperation can also be mediated through DNA, either with DNA providing additional stability to a TF-TF interaction (18), or even without any direct protein-protein interaction. Different mechanisms are possible for the later. For example, the binding of one TF may alter the DNA shape in a way that increases the binding affinity of another TF (30). Another system is the pioneer/settler hierarchy described above, with settler TFs binding DNA only if adequate pioneer TFs have already bound to open the chromatin (39). Lastly, other authors have hypothesized a non-hierarchical cooperative system, with multiple concomitant TF bindings mediated by nucleosomes (29). This is related to the "billboard" system proposed for enhancers (3).

Recently, deep learning approaches such as DeepSea (33, 48) have been proposed for predicting epigenetic marks (including TF binding sites) from raw data sequences. These approaches show higher prediction accuracy than PWM-based methods, but the biological interpretation of the learned neural network is not straightforward. Moreover, approaches such as DeepSea involve a very high number of parameters and hence require high amount of learning data to work.

In this paper, we propose a new approach (TFcoop) for modeling TFBSs taking into account the cooperation between TFs. More formally, for a given cell type, regulatory region (for example 500bp around the TSS of a particular gene), and TF, we aim to predict whether this TF binds the considered sequence in the given cell type. Our predictor is a logistic model based on a linear combination of two kinds of variables: i) the binding affinity (*i.e.* PWM affinity score), of the TF of interest as well as any other TFs identified as cooperating with the target TF; and ii) the nucleotide composition of the sequence. The set of cooperating TFs and the model parameters are learned from ChIP-seq data of the target TF. This approach thus allows us to take into account the potential presence of cooperating TFs to predict the presence or absence of the target TF. Another advantage is that it allows us to consider all available PWMs for a given TF, and therefore to handle alternative binding-site motifs. Lastly, the learned model can be readily analyzed and directly yields a list of potentially cooperating TFs. Variable selection (*i.e.* identification of cooperating TFs) is done via lasso penalization (42). Learning can be done using a moderate amount of data, which allows us to learn specific models for different types of regulatory sequences.

Using ChIP-seq data from the ENCODE project, we used TFcoop to investigate the TF combination involved in the binding of 106 different TFs on 41 different cell types and in four different regulatory regions: promoters of mRNAs, long non-coding (lnc)RNAs and microRNAs, and enhancers (2, 7, 11, 13). Our experiments show

that this approach outperforms simple PWM methods, with accuracy and precision close to that of DeepSea (48). Moreover, analysis of the learned models sheds light on important properties of TF combinations. First, for a given TF and region, we show that TF combinations governing the presence/absence of the target TF are similar for the different cell-types. Second, for a given TF, we observe that TF combinations are different between promoters and enhancers, but similar for all promoters of all gene classes (mRNAs, lncRNAs, and miRNAs). Analysis of the composition of TFs cooperating with the different targets show over-representation of pioneer TFs (39), especially in promoters. We also observed that cooperating TFs are enriched for TFs whose binding is weakened by methylation (46). Lastly, our models can accurately distinguish promoters into classes associated with specific biological processes.

MATERIALS AND METHODS

Data

Promoter, enhancer, long non-coding RNA and microRNA sequences. We predicted TF binding in both human promoters and enhancers. For promoters, sequences spanning $\pm$ 500bp around starts (i.e. most upstream TSS) of protein-coding genes, long non-coding RNAs and microRNAs were considered. Starts of coding and lncRNA genes were obtained from the hg19 FANTOM CAGE Associated Transcriptome (CAT) annotation (11, 13). Starts of microRNA genes (primary microRNAs, pri-miRNAs) were from (7). For enhancers, sequences spanning $\pm$ 500bp around the mid-positions of FANTOM-defined enhancers (2) were used. Lastly, our sequence datasets are composed of 20,845 protein coding genes, 1,250 pri-microRNAs, 23,887 lncRNAs, and 38,553 enhancer sequences.

Nucleotide and dinucleotide features. For each of these sequences, we computed nucleotide and dinucleotide relative frequencies as the occurrence number in the sequence divided by sequence length. Frequencies were computed in accordance with the rule of DNA reverse complement. For nucleotides, we computed the frequency of A/T and G/C. Similarly, frequencies of reverse complement dinucleotides (*e.g.* ApG and CpT) were computed together. This results in a total of 12 features (2 nucleotides and 10 dinucleotides).

PWM. We used vertebrate TF PFMs from JASPAR (25), including all existing versions of each PFM, resulting in a set of 638 PFMs with 118 alternative versions. PFMs were transformed into PWMs as described in Wasserman and Sandelin (43). PWM scores used by TFcoop for a given sequence were computed as described in (43), keeping the maximal score obtained in any position of the sequence.

ChIP-seq data. We collected ChIP-seq data from the ENCODE project (40) for human immortalized cell lines, tissues, and primary cells. Experiments were selected when the targeted TF were identified by a PWM in JASPAR. Thus we studied 409 ChIP-seq experiments for 106 distinct TFs and 41 different cell types. The most represented TF is CTCF with 69 experiments, while 88% of the experiments are designed from immortalized cell lines (mainly GM12878, HepG2 and K562). The detailed list of all used experiments is given in Supplementary materials. For each ChIP-seq experiment, regulatory sequences were classified as positive or negative for the corresponding ChIP targeted TF. We used Bedtools v2.25.0 (34) to detect intersection between ChIP-seq binding sites and regulatory sequences (both mapped to the hg19 genome). Each sequence that intersects at least one ChIP-seq binding region was classified as a positive sequence. The remaining sequences formed a negative set. The number of positive sequences varies greatly between experiments and sequence types. Mean and standard deviation numbers of positive sequences are respectively $2661(\pm 1997)$ for mRNAs, $1699(\pm 1151)$ for lncRNAs, $216(\pm 176)$ for microRNAs, and $1516(\pm 1214)$ for enhancers.

Expression data. To control the effect of expression in our analyses, we used ENCODE CAGE data restricted to 41 cell lines. The expression per cell line was calculated as the mean of the expression observed in all corresponding replicates. For microRNAs, we used the small RNA-seq ENCODE expression data collected for 3,043 mature microRNAs in 37 cell lines (corresponding to 403 ChIP-seq experiments). The expression of microRNA genes (i.e. pri-microRNAs) was calculated as the sum of the expression of the corresponding mature microRNAs.

Logistic model

We propose a logistic model to predict the regulatory sequences bound by a specific TF. Contrary to classical approaches, we not only consider the score of the PWM associated with the target TF, but also the scores of all other available PWMs. The main idea behind this is to unveil the TF interactions required for effective binding of

the target TF. We also integrate in our model the nucleotide and dinucleotide compositions of the sequences, as the environment of TFBSs are thought to play major role in binding affinity (9, 23).

For each ChIP-Seq experiment, we learn different models to predict sequences bound by the target TF in four regulatory regions (promoters of mRNA, lncRNA and pri-miRNA, and enhancers). For a given experiment and regulatory region, our model aims to predict response variable y_s by the linear expression

$$\alpha + \sum_{m \in Motifs} \beta_m \times Score_{m,s} + \sum_{n \in Nucl} \beta_n \times Rate_{n,s} + \varepsilon_s,$$

where y_s is the boolean response variable representing the TF binding on the given sequence s ($y_s = 1$ for TF binding, 0 otherwise); $Score_{m,s}$ is the score of motif m on sequence s; $Rate_{n,s}$ is the frequency of (di)nucleotide n in sequence s; α is a constant; β_m and β_n are the regression coefficient associated with motif m and (di)nucleotide n, respectively; and ε_s is the error associated with sequence s. $Motifs$ and $Nucl$ sets respectively contain 638 JASPAR PWMs and 12 (di)nucleotide frequencies.

To perform variable selection (*i.e.* identifying cooperating TFs), we used the lasso regression minimising the prediction error within a constraint over $l1$-norm of β (42). The weight of the lasso penalty is chosen by cross-validation by minimising the prediction error with the R package $glmnet$ (35). As the response variable is boolean, we used a logistic regression giving an estimation of the probability to be bound for each sequence. We evaluate the performance of the model using 10-fold cross validation. In each validation loop, 90% of sequences (training data) are used to learn the β parameters and the remaining 10% (test data) are used to evaluate the predictive performance of the model.

Alternative approaches

We compared the predictive accuracy of our model to three other approaches.

Best hit approach. The traditional way to identify TF binding sites consists in scanning a sequence and scoring the corresponding PWM at each position. Positions with a score above a predefined threshold are considered as potential TFBS. A sequence is then considered as bound if it contains at least one potential TFBS.

TRAP score. An alternative approach proposed by Roider et al. (37) is based on a biophysically inspired model that estimates the number of bound TF molecules for a given sequence. In this model, the whole sequence is considered to define a global affinity measure, which enables us to detect low affinity bindings as described in (**?**). We use the R package $tRap$ (35) to compute the affinity score of the 638 PWMs for all sequences. As proposed in (37), we use default values for the two parameters ($R_0(width)$, $\lambda = 0.7$).

DNA shape. In addition to PWMs, Mathelier et al. (26) considered 4 DNA shapes to increase binding site identification: helix twist, minor groove width, propeller twist, and DNA roll. The 2^{nd} order values of these DNA shapes are also used to capture dependencies between adjacent positions. Thus, each sequence is characterized by the best hit score for a given PWM plus the 1^{st} and 2^{nd} DNA shape order values at the best hit position. The approach based on gradient boosting classifier requires a first training step with foreground (bound) and background (unbound) sequences to learn classification rules. Then the classifier is applied to the set of test sequences. We used the same 10-fold cross-validation scheme that we used in our approach. We applied two modifications to speed-up the method, which was designed for smaller sequences. First, in the PWM optimization step of the training phase, we reduced the sequences to $\pm$ 50bp around the position with highest ChIP-Seq peak for positive sequences and to $\pm$ 50bp around a random position for negative sequences. Second, after this first step we also reduced sequences used to train and test the classifiers to $\pm$ 50bp around the position for which the (optimized) PWM gets the best score.

DeepSEA. Zhou and Troyanskaya (48) proposed a deep learning approach for predicting the binding of chromatin proteins and histone marks from DNA sequences with single-nucleotide sensitivity. Their deep convolutional network takes 1000bp genomic sequences as input and predicts the states associated with several chromatin marks in different tissues. We used the predictions provided by DeepSEA server (http://deepsea.princeton.edu/). Namely, coordinates of the analyzed promoter and enhancer sequences were provided to the server, and the predictions associated with each sequence were retrieved. Only the predictions related to the ChIP-seq data we used in our analyses were considered (*i.e.* 214 ChIP-seq data in total).

RESULTS

Computational approach

Given a target TF, the TFcoop method identifies the TFBS combination that is indicative of the TF presence in a regulatory region. We first considered the promoter region of all mRNAs (defined as the 1000bp centered around gene start). TFcoop is based on a logistic model that predicts the presence of the target TF in a particular promoter using two kinds of variables: PWM affinity scores and (di)nucleotide frequencies. For each promoter sequence, we computed the affinity score of the 638 JASPAR PWMs (redundant vertebrate collection), and the frequency of every mono- and dinucleotide in the promoter. These variables were then used to train a logistic model that aims to predict the outcome of a particular ChIP-seq experiment in mRNA promoters. Namely, every promoter sequence with a ChIP-seq peak is considered as a positive example, while the other sequences are considered as negative examples (see below). In the experiments below, we used 409 ChIP-seq datasets from ENCODE and different models. Each model targets one TF and one cell type. Given a ChIP-seq experiment, the learning process involves selecting the PWMs and (di)nucleotides that can help discriminate between positive and negative sequences, and estimate the model parameters that minimize prediction error. Note that the learning algorithm can select any predictive variable including the PWM of the target TF. See Material and methods for more details on the data and logistic model.

As explained above, positive sequences are promoters overlapping a ChIP-seq peak in the considered ChIP-seq experiment. We used two different procedures for selecting the positive and negative examples. Each procedure actually defines a different prediction problem. In the first case, we kept all positive sequences, and randomly selected the same number of negative sequences among all sequences that do not overlap a ChIP-seq peak. In the second case, we used an additional dataset that measures gene expression in the same cell type as the ChIP-seq data. We then selected all positive sequences with non zero expression level and randomly selected the same number of negative sequences among all sequences that do not overlap a ChIP-seq peak but that have a similar expression level as the selected positive sequences. Hence, in this case (hereafter called the expression-controlled case), we learn a model that predicts the binding of a target TF in a promoter knowing that the corresponding gene is expressed. On the contrary, in the first case we learn a model that predicts the binding without knowledge about gene expression.

Classification accuracy and model specificity

We ran TFcoop on the 409 ChIP-seq datasets and for the two classification cases. The accuracy of each model was assessed by cross-validation by measuring the area under the Receiver Operating Curve (ROC). For comparison, we also measured the accuracy of the classical approach that discriminates between positive and negative sequences using only the affinity score of the PWM associated with the target TF. In addition, we estimated the accuracy of the TRAP method, which uses a biophysically inspired model to compute PWM affinity (37) and that of the approach proposed in (26), which integrates DNA shape information with PWMs. As shown in Figure 1Accuracy on mRNA promoters. (a) An example of ROC curves obtained on ChIP-seq targeting TF USF1 in cell-type A549. (b) Violin plots of the area under the ROC curves obtained in the 409 ChIP-seq. Best hit (red), TRAP (blue), DNAshape (green), TFcoop with no expression control (purple), and TFcoop with expression control (orange). ROC curves for Best hit, TRAP and DNAshape were computed in the non expression-controlled casefigure.1 and Supp. Figures 1 and 2, TFcoop outperforms these PWM-based approaches on many TFs. Next, we ran TFcoop using the TRAP scoring approach instead of the standard PWM scoring method but did not observe better results (data not shown), despite the fact that TRAP slightly outperforms the standard method when used standalone (Figure 1(b)Subfigure 1(b)subfigure.1.2 and Supp. Figure 1a). We also ran TFcoop with tri- and quadri-nucleotide frequencies in addition to di-nucleotide frequencies. Although a consistent AUC improvement was observed, the increase was very slight most of the time (Supp. Figure 3). Lastly, we compared TFcoop accuracy to that of the deep learning approach DeepSea (48) and observed very close results (see Supp. Figure 4).

We then sought to take advantage of the relative redundancy of target TFs in the set of 409 ChIP-seq experiments to investigate the specificity of the learned models. Namely, we compared pairs of models learned from ChIP-seq experiments targeting (i) the same TF in the same cell-type, (ii) the same TF in different cell-types, (iii) different TFs in the same cell-type, and (iv) different TFs in different cell-types. In these analyses, we used the model learned on one ChIP-seq experiment A to predict the outcome of another ChIP-seq experiment B, and we compared the results to those obtained with the model directly learned on B. More precisely, we measured the difference of the Area under the ROC Curves (AUC) between the model learned on A and applied on B and the model learned and applied on B. As shown in Figure 2Model specificity on mRNA promoters. Distribution of AUC differences obtained when using a model learned on a first ChIP-seq experiment to predict the outcome of a second ChIP-seq

experiment. Different pairs of ChIP-seq experiments were used: experiments on the same TF and same cell type (red), experiments on the same TF but different cell type (yellow), experiments on different TFs but same cell type (light blue), and experiments on different TFs and different cell types (blue). For each pair of ChIP-seq experiment A-B, we measured the difference between the AUC achieved on A using the model learned on A, and the AUC achieved on A using the model learned on B. AUC differences were measured on the non expression-controlled case (a) and on the expression-controlled case (b)figure.2, models learned on the same TF (whether or not on the same cell-type) have overall smaller AUC differences than models learned on different TFs.

We then analyzed cell and TF specificity more precisely. Cell specificity refers to the ability of a model learned on one TF and in one cell type to predict the outcome of the same TF in another cell type. Oppositely, TF specificity refers to the ability of a model learned on one TF in one cell type to predict the outcome of another TF in the same cell type. Cell and TF specificities were evaluated by the shift between the associated distributions of AUC differences in Figure 2Model specificity on mRNA promoters. Distribution of AUC differences obtained when using a model learned on a first ChIP-seq experiment to predict the outcome of a second ChIP-seq experiment. Different pairs of ChIP-seq experiments were used: experiments on the same TF and same cell type (red), experiments on the same TF but different cell type (yellow), experiments on different TFs but same cell type (light blue), and experiments on different TFs and different cell types (blue). For each pair of ChIP-seq experiment A-B, we measured the difference between the AUC achieved on A using the model learned on A, and the AUC achieved on A using the model learned on B. AUC differences were measured on the non expression-controlled case (a) and on the expression-controlled case (b)figure.2: cell specificity was assessed by the shift between red and yellow distributions, while TF specificity was assessed by the shift between red and light blue distributions. We used a standard t-test to measure that shift. Low p-values indicate high distribution shifts (hence high cell/TF specificity), while high p-values indicate low shifts (hence low specificity). Our results indicate very low cell specificity (p-values 0.91 and 0.95 in the non-controlled and expression-controlled cases, respectively) and high TF specificity ($1 \cdot 10^{-61}$ and $3 \cdot 10^{-83}$). The fact that the TF specificity is slightly higher in the expression-controlled case suggests that part of the TF combinations that help discriminate between bound and unbound sequences is common to several TFs in the non-controlled case. It is indeed known that the majority of ChIP-seq peaks are found in open and expressed promoters (41). Thus, most positive examples are associated with open chromatin marks. However, in the non-expression-controlled case, a large part of the negative examples are in closed chromatin and are therefore likely associated with other chromatin marks. As a result, in this case, TFcoop also learns the TFBS signature that helps differentiate between these chromatin marks. Oppositely, in the expression-controlled case, the positive and negative examples have similar chromatin states, and TFcoop unveils the TFBS signature specific to the target TF.

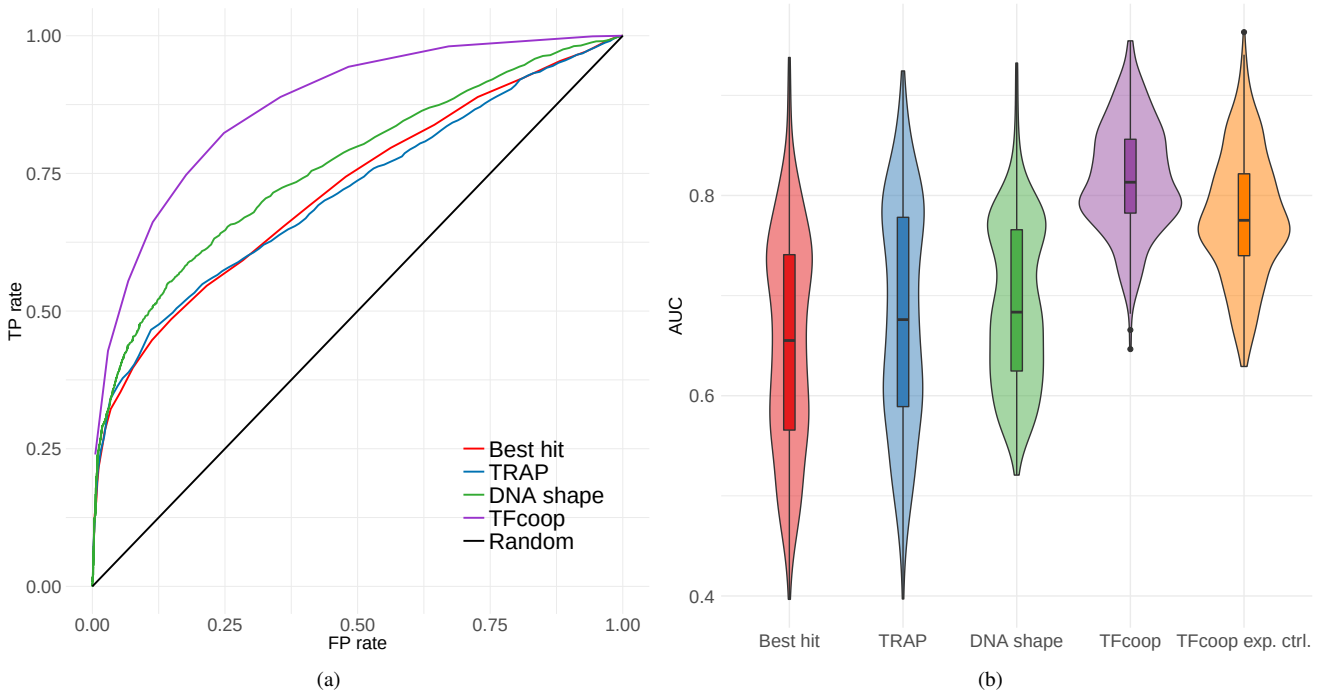

Figure 1. Accuracy on mRNA promoters. (a) An example of ROC curves obtained on ChIP-seq targeting TF USF1 in cell-type A549. (b) Violin plots of the area under the ROC curves obtained in the 409 ChIP-seq. Best hit (red), TRAP (blue), DNAshape (green), TFcoop with no expression control (purple), and TFcoop with expression control (orange). ROC curves for Best hit, TRAP and DNAshape were computed in the non expression-controlled case.

Finally the low cell specificity means that the general rules governing TFBS combination in promoters do not dramatically change from one tissue to another. This is important in practice because it enables us to use a model learned on a specific ChIP-seq experiment to predict TBFSs of the same TF in another cell type.

Analysis of TFBS combinations in promoters

We next analyzed the different variables (PWM scores and (di)nucleotide frequencies) that were selected in the 409 learned models. Overall, 95% of the variables correspond to PWM scores. Although only 5% of the selected variables are (di)nucleotide frequencies, almost all models include at least one of these features (see Supp. Figure 7).

We then looked at the presence of the PWM associated with the target TF in its model. As mentioned earlier, the learning algorithm does not use any prior knowledge and can select the variables that best help predict the ChIP-seq experiment without necessarily selecting the PWM of the target TF. In fact, our analysis shows that, for 75% of the models, at least one version of the target PWM was selected. However, it is important to note that similar PWMs tend to have correlated scores. Hence, another PWM may be selected instead of the target. To overcome this bias, we also considered all PWMs similar to the target PWM. We used Pearson correlation between PWM scores in all promoters to measure similarity and set a threshold value of 0.75 to define the list of similar PWMs. With this threshold, 90% models include the target or a similar PWM.

Next, following the analyses of Levo et al. (23) and Dror et al. (9) we used our models to investigate the link between the nucleotide composition of the target PWM and that of the TFBS flanking region. First, we did not observe a significant link between target PWM composition and the (di)nucleotide variables that were selected in the models (Kolmogorov-Smirnov test p-val=0.448; see Supp. Figure 14). However, the (di)nucleotide composition of target PWM exhibited strong resemblance to that of the other selected PWMs (see Figure 3Pearson correlation between nucleotide composition of the target PWM and the mean composition of selected PWMs (with positive and negative coefficients in red and blue, respectively) in 409 models. Grey: correlation achieved by randomly selecting the same number of PWMs for each modelfigure.3). Specifically, except for dinucleotide TpT, the nucleotide and dinucleotide frequencies of the target PWM were strongly correlated with that of the PWMs selected with a positive coefficient, but not with those selected with a negative coeficient. This is in accordance with the findings of Dror et al. (9), who show that TFBS flanking regions often have similar nucleotide composition as the the TFBS.

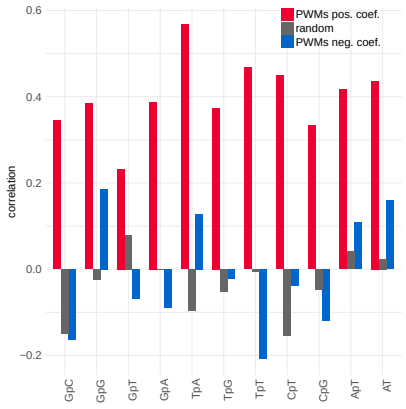

Figure 3. Pearson correlation between nucleotide composition of the target PWM and the mean composition of selected PWMs (with positive and negative coefficients in red and blue, respectively) in 409 models. Grey: correlation achieved by randomly selecting the same number of PWMs for each model.

We next evaluated the possibility of clustering the 409 learned models using the selected variables. As shown in Supp. Figure 5, the models can be partitioned in a few different classes. In Supp. Figure 5 models were clustered in 5 classes with a k-means algorithm. Supp. Figure 7 reports the most used variables in these different classes. We can first observe that, in agreement with our analysis of model specificity, the models associated with the same TF tend to cluster together. For example, the 4^{th} class of our clustering is exclusively composed of CTCF models. This clustering seems to be essentially driven by the nucleotide composition of the PWMs belonging to the models (see Supp. Figure 15a). Note that we did not observe any enrichment for the classical TF structural families (bHLH, Zinc finger, ...) in the different classes (data not shown).

Pioneer TFs are thought to play an important role in transcription by binding to condensed chromatin and enhancing the recruitment of other TFs (39). As shown in Figure 4Pioneer TF distribution of selected PWMs in the different models. We kept one model for each target PWM to avoid bias due to over-representation of the same PWM in certain classes. Grey represents the distribution of all PWMs associated with a family in Sherwood et al. (39) (159 over 520 non-redundant PWMs)figure.4 and Supp. Figure 7, pioneer factors clearly are the most represented TFs in the selected variables of all models (regardless of the model class), whereas they represent less than 14% of all TFs. These findings are in agreement with their activity: pioneer TFs occupy previously closed chromatin and, once bound, allow other TFs to bind nearby (39). Hence the binding of a given TF requires the prior binding of at least one pioneer TF. We also observed that TFs whose binding is weakened by methylation (46) are enriched in all models (Supp. Figure 16a). This result may explain how CpG methylation can negatively regulate the binding of a given TF *in vivo* while methylation of its specific binding site has a neutral or positive effect *in vitro* (46): regardless of the methylation status on its binding site, the binding of a TF can also be influenced *in vivo* by the sensitivity of its partners to CpG methylation.

TFBS combinations in lncRNA and pri-miRNA promoters

We then ran the same analyses on the promoters of lncRNAs and pri-miRNAs using the same set of ChIP-seq experiments. Results are globally consistent with what we observed on mRNA promoters (see Figure **??** for the expression-controlled case). Overall, models show good accuracy and specificity on lncRNAs. Models are less accurate and have lower specificity for pri-miRNAs but this likely results from the very low number of positive examples available for these genes in each ChIP-seq experiment (Supp. Figure 13), which impedes both the learning of the models and estimation of their accuracy.

Next we sought to compare the models learned on mRNA promoters to the models learned on lncRNA and pri-miRNA promoters. For this, we interchanged the models learned on the same ChIP-seq experiment, *i.e.* we used the model learned on mRNA promoters to predict the outcome on lncRNA and pri-miRNA promoters. One striking fact illustrated by Figure **??** is that models learned on mRNA promoters and those learned on lncRNA promoters are almost perfectly interchangeable. This means that the TFBS rules governing the binding of a specific TF in a promoter are similar for both types of genes. We obtained consistent results when we used the models learned on mRNAs to predict the ChIP-seq outcomes on pri-miRNA promoters (Figure **??**). Accuracy is even better than that obtained by models directly learned on pri-miRNA promoters, illustrating the fact that the poor performance

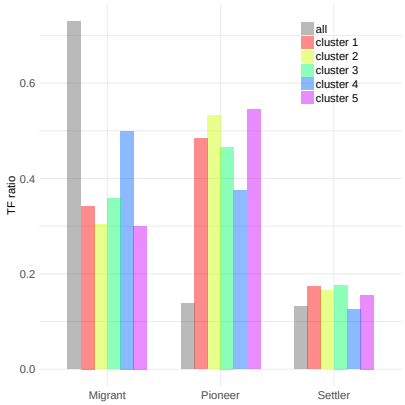

Figure 4. Pioneer TF distribution of selected PWMs in the different models. We kept one model for each target PWM to avoid bias due to over-representation of the same PWM in certain classes. Grey represents the distribution of all PWMs associated with a family in Sherwood et al. (39) (159 over 520 non-redundant PWMs).

achieved on pri-miRNA promoters likely results from the small number of learning examples available for these genes.

TFBS combinations in enhancers

We next applied the same approach on 38,554 enhancers defined by the FANTOM consortium (2). We used the same ChIP-seq experiments as for the promoters. All enhancer sequences overlapping a ChIP-seq peak in the considered ChIP-seq experiment were considered as positive examples. As for promoters, we used two strategies to select the negative examples: in a first case we did not apply any control on the expression of the negative enhancers, while in a second case, we used CAGE expression data to ensure that negative enhancers have globally the same expression levels as positive enhancers.

As observed for promoters, TFcoop outperforms classical PWM-based approaches on many TFs (see Figure **??**) and achieves results close to that of DeepSea (48) (Supp. Figure 4). However, analysis of model specificity reveals somewhat different results from that observed for promoters. Globally, models have good TF specificity: models learned on the same TF have more similar prediction accuracy than models learned on different TFs. However, in contrast to promoters, cell specificity is high in the non-controlled case (p-value $2 \cdot 10^{-45}$, see peak shift in Figure **??**), although much lower in the expression-controlled case (p-value $1.6 \cdot 10^{-12}$). Additionally, TF specificity seems slightly higher in the expression-controlled case than in the non-controlled case (p-values $1.7 \cdot 10^{-102}$ *vs.* $1. \cdot 10^{-114}$). This is in accordance with our hypothesis formulated for promoters, that part of the TF combination learned by TFcoop in the non-controlled case actually differentiates between close and open chromatin marks. Moreover, this also seems to indicate that these TF combinations are cell-type specific, while the remaining combinations seems more general (as illustrated by the $1.6 \cdot 10^{-12}$ p-value measured on the expression-controlled case). The fact that cell-type specificity is more apparent for enhancers than for promoters in the non expression-controlled case ($2 \cdot 10^{-45}$ for enhancers *vs.* 0.91 for promoters) is in accordance with the lowest ubiquity of enhancers (2) and the fact that, contrary to promoters, most of enhancers are expressed in a cell-specific manner (as illustrated in Supp. Figure 9).

We next analyzed the different TFBS combinations of the enhancer models. As for promoters, we observed that the selected PWMs tends to have similar (di)nucleotide composition as the target PWM (Figure. **??**). Moreover, models can also be partitioned in a few different classes according to the selected variables (Supp. Figures 11 and 12). These classes mostly correspond to the nucleotide composition of the target and selected PWMs (Supp. Figure 18). Pioneer TFs are also over-represented in the selected PWMs but surprisingly less so than for promoters (Figure **??** and Supp. Figure 12).

Next we sought to compare the models learned on enhancers to the models learned on promoters. We used the models learned in the expression-controled cases. First, we can observe that these models have globally similar prediction accuracy (see Figure **??**). However, a pairwise comparison of the enhancer and promoter models learned on each ChIP-seq experiment shows that the prediction accuracy is only moderately correlated (see Supp. Figure 10, Pearson correlation 0.33). Moreover, if we interchange the two models learned on the same ChIP-seq experiment, we observe that the model learned on promoters is generally not as good on enhancers as it is on

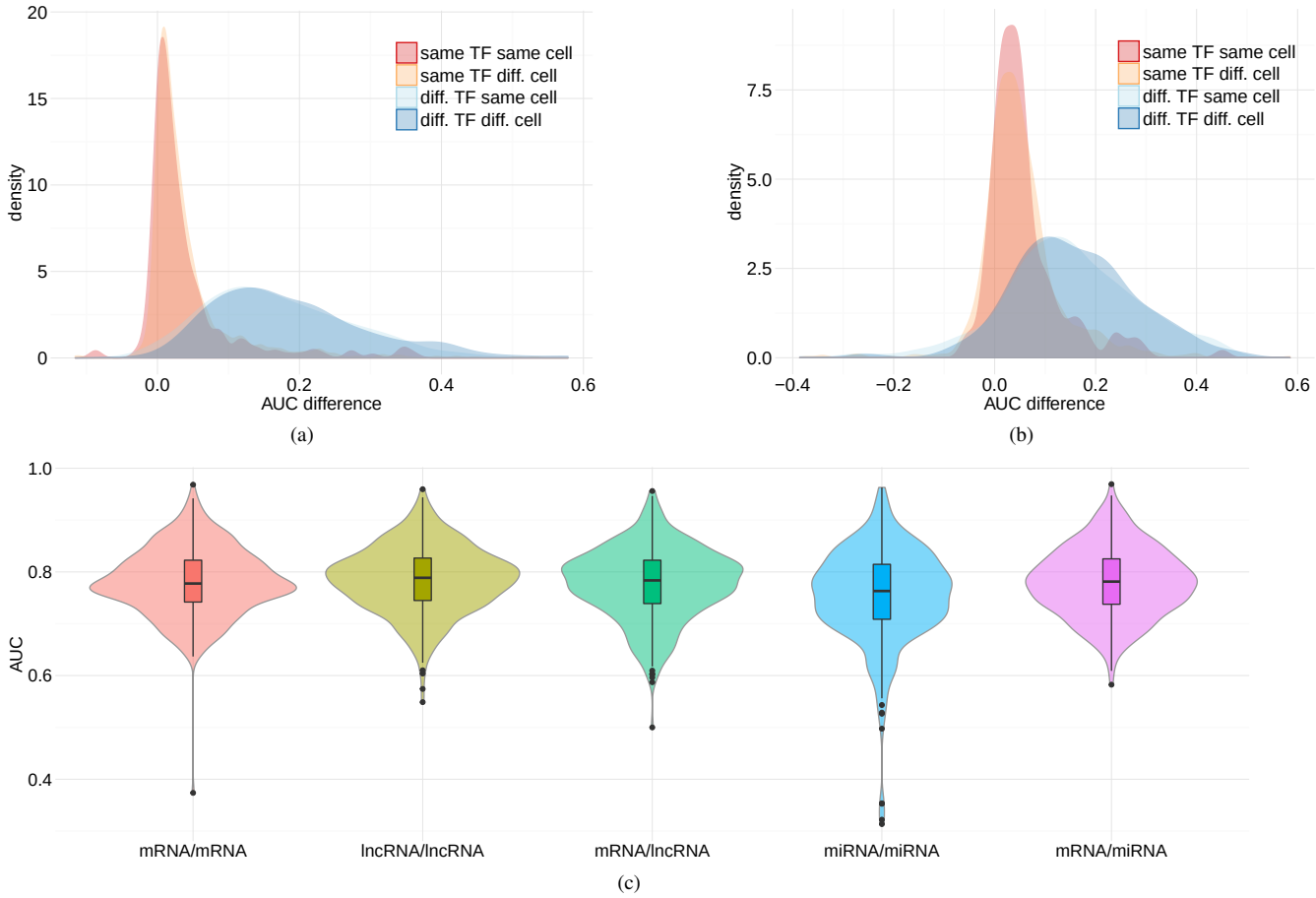

Figure 5. Specificity and accuracy on lncRNA and pri-miRNA promoters. Top: Model specificity on promoters of lncRNA (a) and pri-miRNAs (b). See legend of Figure 2Model specificity on mRNA promoters. Distribution of AUC differences obtained when using a model learned on a first ChIP-seq experiment to predict the outcome of a second ChIP-seq experiment. Different pairs of ChIP-seq experiments were used: experiments on the same TF and same cell type (red), experiments on the same TF but different cell type (yellow), experiments on different TFs but same cell type (light blue), and experiments on different TFs and different cell types (blue). For each pair of ChIP-seq experiment A-B, we measured the difference between the AUC achieved on A using the model learned on A, and the AUC achieved on A using the model learned on B. AUC differences were measured on the non expression-controlled case (a) and on the expression-controlled case (b)figure.2 for details. Bottom: Promoter models are interchangeable. For each ChIP-seq experiment, we computed the AUC of the model learned and applied on mRNAs (pink), learned and applied on lncRNAs (yellow-green), learned and applied on pri-miRNAs (blue), learned on mRNAs and applied to lncRNAs (green), learned on mRNAs and applied to pri-miRNAs (purple).

promoters and *vice-versa* (Figure **??**). Hence, while the rules learned on enhancers (promoters) in a given cell type are valid for enhancers (promoters) of other cell types, they do not apply to promoters (enhancers) of the same cell type. Note that AUCs of models learned on promoters and applied to enhancers are greater than that of models learned on enhancers and applied to promoters (Figure **??**). This result might be explained by the existence of promoters able to exert enhancer functions (6, 8). Note that, conversely, the FANTOM definition of enhancers precludes potential promoter functions (2).

Using TFcoop score for describing regulatory sequences

We next explored whether TFcoop scores could be used to provide meaningful descriptions of regulatory sequences. This was assessed in two ways. First, we used the TFcoop models to cluster mRNA promoters and searched for over-represented gene ontology (GO) terms in the inferred clusters. We randomly selected one model for each TF, and used the 106 selected models to score the 20,846 mRNA promoter sequences. Each promoter sequence was then described by a vector of length 106. We next ran a k-means algorithm to partition the promoters into 5 different clusters, and we searched for over-represented GO terms in each cluster. For comparison, we ran the same procedure using two other ways to describe the promoter sequences: the classical PWM scores of the same 106 selected TFs (so promoters are also described by vectors of length 106), and the (di)nucleotide frequencies of the promoters (vector of length 12). Globally, the same GO terms appear to be over-represented in the different gene clusters and the three different clusterings: defense response, immune system process, cell cycle,

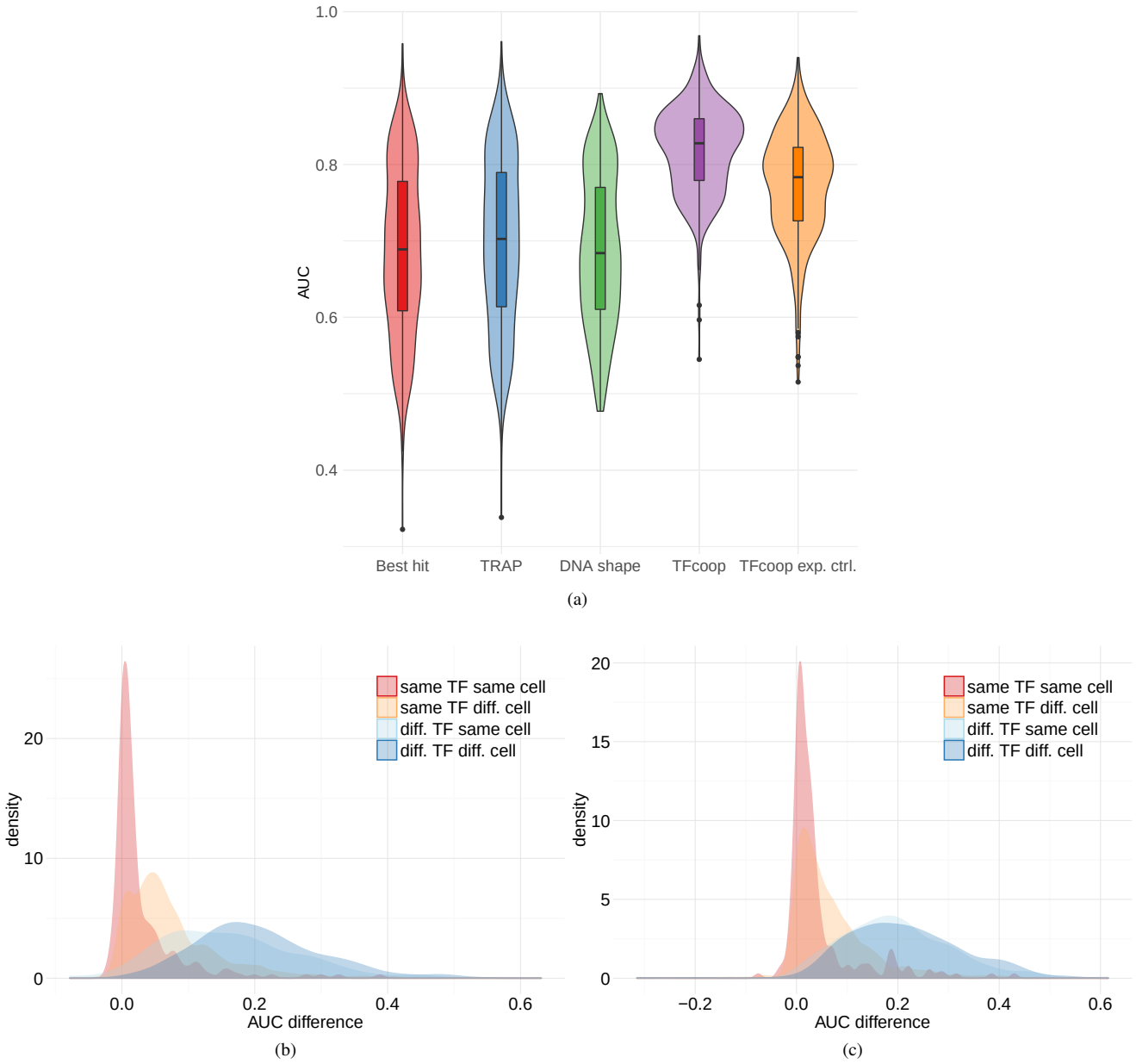

Figure 6. Model accuracy and specificity on enhancers. See legend of Figure 1Accuracy on mRNA promoters. (a) An example of ROC curves obtained on ChIP-seq targeting TF USF1 in cell-type A549. (b) Violin plots of the area under the ROC curves obtained in the 409 ChIP-seq. Best hit (red), TRAP (blue), DNAshape (green), TFcoop with no expression control (purple), and TFcoop with expression control (orange). ROC curves for Best hit, TRAP and DNAshape were computed in the non expression-controlled casefigure.1 and Figure 2Model specificity on mRNA promoters. Distribution of AUC differences obtained when using a model learned on a first ChIP-seq experiment to predict the outcome of a second ChIP-seq experiment. Different pairs of ChIP-seq experiments were used: experiments on the same TF and same cell type (red), experiments on the same TF but different cell type (yellow), experiments on different TFs but same cell type (light blue), and experiments on different TFs and different cell types (blue). For each pair of ChIP-seq experiment A-B, we measured the difference between the AUC achieved on A using the model learned on A, and the AUC achieved on A using the model learned on B. AUC differences were measured on the non expression-controlled case (a) and on the expression-controlled case (b)figure.2 for details.

metabolic process, and developmental process. We noticed that the p-values obtained with the TFcoop scores were invariably better than the two others. To avoid any clustering bias, we repeated the k-means clusterings several times, with various numbers of clusters. Namely, for each approach we ran 3 clusterings for each number of clusters ranging between 3 and 10 (resulting in 24 different clusterings for each approach) and computed over-representation p-values for the 5 GO terms in each cluster. As shown in Figure **??**, the TFcoop scores substantially and systematically outperform the other scoring functions, indicating that the classification obtained with this score is more accurate to functionally annotate promoters than the others.

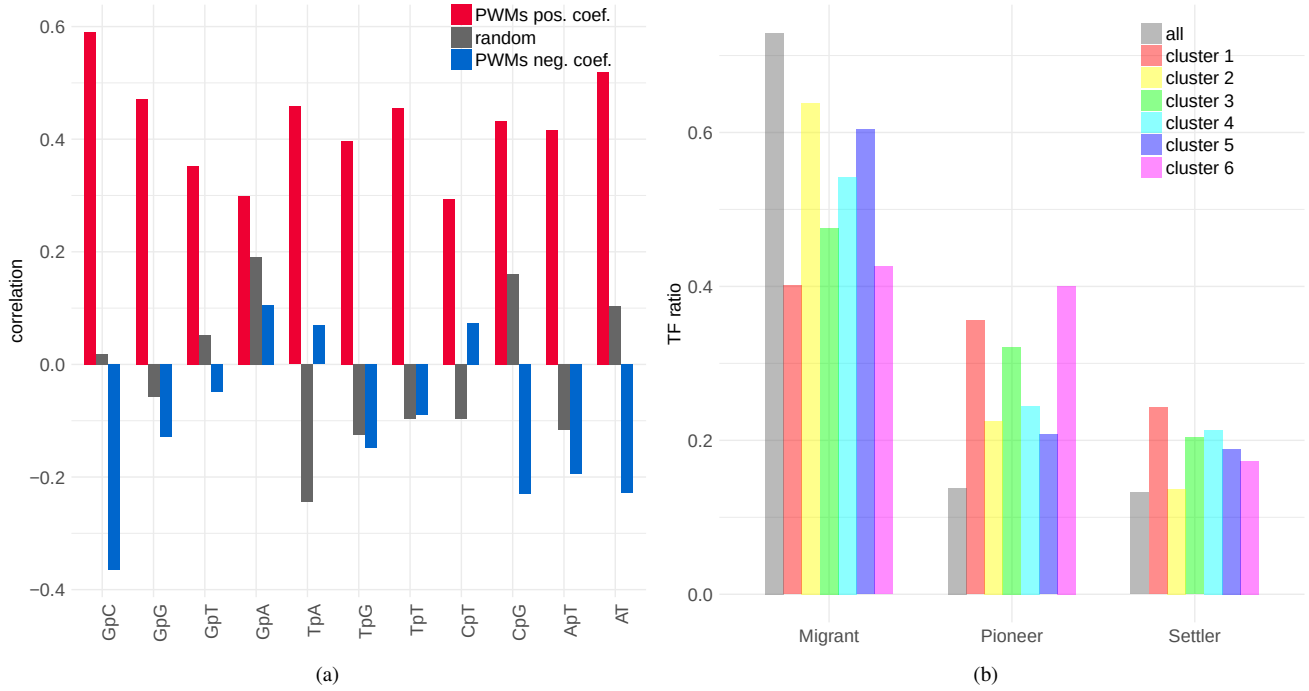

Figure 7. Selected PWMs in enhancers. (a) Pearson correlation between nucleotide composition of the target PWM and the mean composition of selected PWMs (see legend of Figure 3Pearson correlation between nucleotide composition of the target PWM and the mean composition of selected PWMs (with positive and negative coefficients in red and blue, respectively) in 409 models. Grey: correlation achieved by randomly selecting the same number of PWMs for each modelfigure.3) (b) Pioneer TF distribution in selected PWMs (same legend as Figure 4Pioneer TF distribution of selected PWMs in the different models. We kept one model for each target PWM to avoid bias due to over-representation of the same PWM in certain classes. Grey represents the distribution of all PWMs associated with a family in Sherwood et al. (39) (159 over 520 non-redundant PWMs)figure.4).

Next, we used the TFcoop models to discriminate between mRNA promoters and enhancers. We randomly split the sets of promoters and enhancers in training and test sets, and learned a K-nearest neighbor (KNN) classifier for discriminating between promoter and enhancer sequences on the basis of scores of the TFcoop models learned on promoters. As above, we also used the classical PWM scores of the same 106 selected TFs and (di)nucleotide frequencies of the sequences. We resumed the procedure with a number of neighbors (K) varying between 1 and 20, and computed the number of errors obtained by each approach on the test set (Figure **??**). Here again, TFcoop description outperforms other description methods, with an error rate around 2% for TFcoop *vs.* 15% and 25% for the other approaches. This result confirms the existence of DNA features distinguishing enhancers from mRNA promoters (1, 2) and identifies TF combinations as potent classifiers.

Identifying TFs responsible for gene expression change

As a final test, we sought to use TFcoop to identify the TFs responsible for gene expression change in various gene expression experiments. For this, we used the compendium collected by Meng et al. (28). The interest of this compendium is that each data corresponds to a particular TF for which the activity has been modified (repressed or enhanced), hence the TF responsible for deregulation (hereafter called as the "responsible TF") is known. In each experiment, we selected the top 500 genes with the highest positive log fold change, and computed the difference of score distribution of the responsible TF in the top 500 promoters and in all other promoters with a Kolmogorov-Smirnov test. This was done using the classical PWM scoring function and with the TFcoop scores. Of the 21 experiments, 5 responsible TFs achieved enrichment p-values below 1% with the classical PWM scoring function, while this number rises to 13 with the TFcoop score (see Supp. Table 1).

One striking fact however is that numerous TFs (not solely the responsible TF) appear to be also enriched in the top 500 promoters (Supp. Table 1). Note that this effect is not restricted to the TFcoop scoring. The classical PWM scoring method also has numerous enriched TFs on the experiments for which it yields good p-values on the responsible TF. There can be different explanations for this effect. First, modifying the activity of the responsible TF may induce a cascade of activations/repressions of other TFs. Second, if two TFs A and B often bind together in promoters, they may share a high number of target genes. In this case, TF B may appear as over-represented in the promoters of genes deregulated by TF A, even if TF B is not itself deregulated. This provides us with an

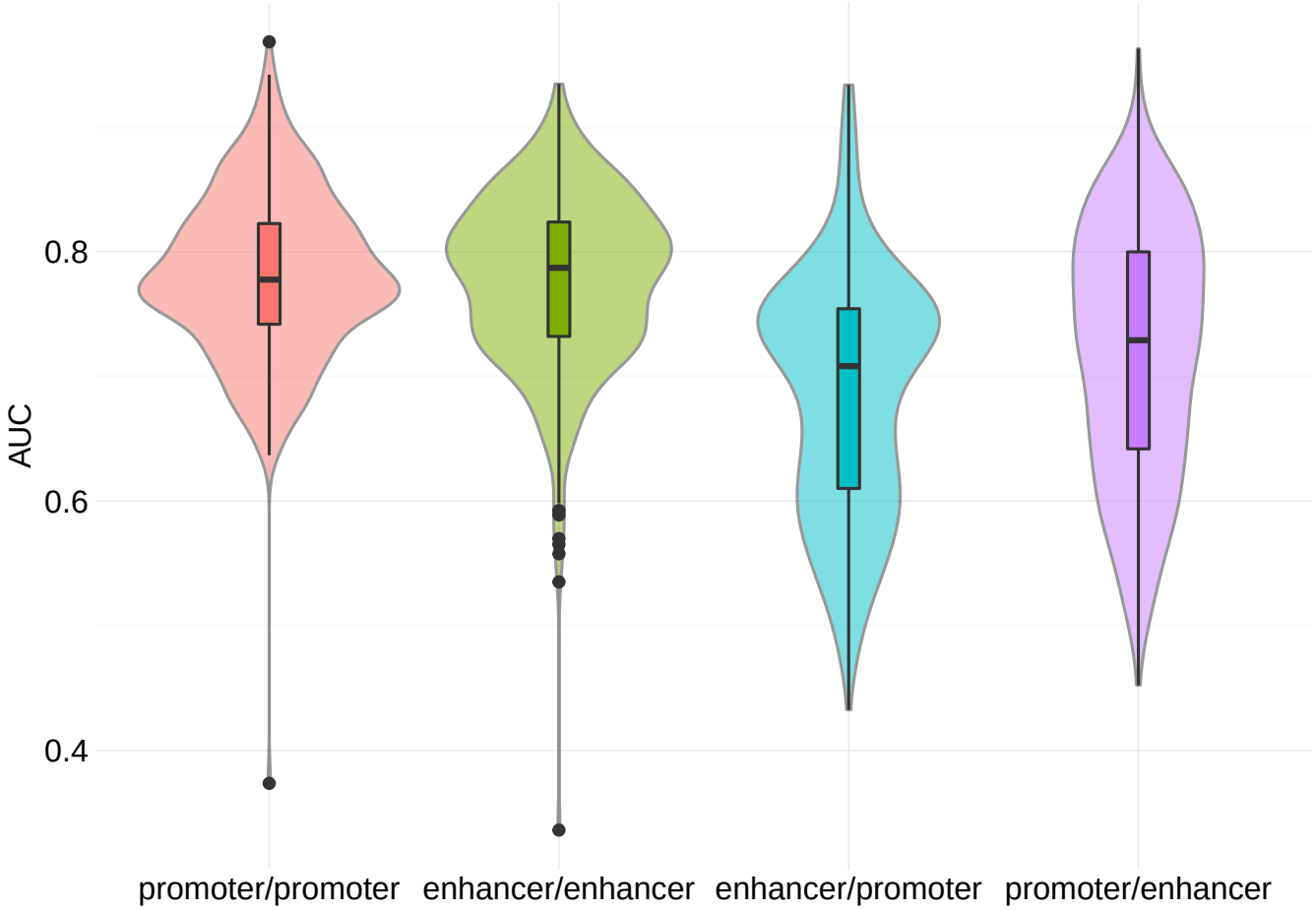

Figure 8. AUCs obtained in mRNA promoter and enhancer models. For each ChIP-seq experiment we computed the AUC of the model learned and applied on the promoters (red), learned and applied on the enhancers (green), learned on enhancers and applied to promoters (blue), and learned on promoters and applied to enhancers (purple).

interesting way to assess our models. Namely, when this appends (and if our models are meaningful) then TF A should be present among the selected variables of the TF B model. For each experiment, we therefore enumerated all TFs enriched in the top 500 promoters and checked whether the responsible TF was present in their models. We used a Fisher exact test to assess whether this appends more often than expected in the different experiments (Supp. Table 1). Of the 18 testable experiments, 13 yield a p-value below 5%, indicating that the responsible TF is often involved in the TF combinations associated with the TFs enriched in the top 500 promoters.

DISCUSSION

In this paper we proposed a method that takes TF combinations into account to predict whether a target TF binds a given regulatory sequence or not. Our approach is based on a logistic model learned from ChIP-seq experiments on the target TF. Once learned, the model can be used to predict the TF binding affinity on any other sequence of the same type (promoter or enhancer). Cross-validation study showed that the approach is effective and outperforms classical approaches on many TFs. It is important to note that TFcoop combinations do not necessarily reflect just cooperation, but also competition. For instance, a TF A competing with a TF B may be useful to predict the binding of B and would thus appear in the TF B model while A and B do not cooperate.

We distinguished two prediction problems associated with two situations, depending whether the aim is to predict binding in any promoter/enhancer or solely in expressed promoters/enhancers. For expressed promoters/enhancers, our experiments showed that the learned models have high TF specificity and quite low cell-type specificity. On the other hand, for the problem of expressed and not expressed promoters/enhancers binding, the learned models are less TF specific and more cell-type specific (especially for enhancers). These results are in accordance with a two-level model of gene regulation: (i) cell-type specific level that deposits specific chromatin

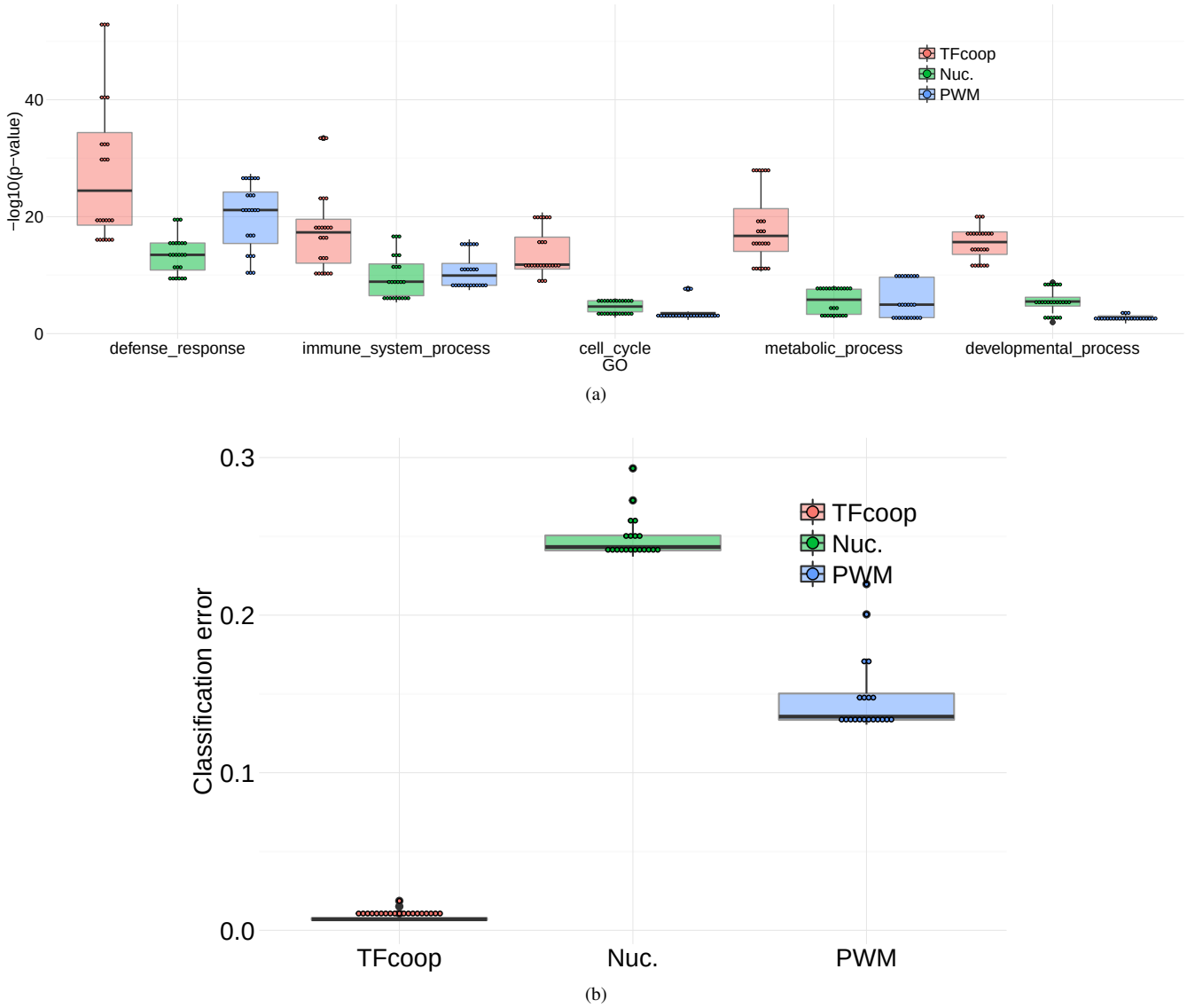

Figure 9. Using TFcoop scores for describing regulatory sequences. (a) GO term enrichment obtained with different promoter descriptions. Promoters were described using three different representations—TFcoop scores (red), (di)nucleotide frequencies (green), classical PWM scores (blue)— and then partitioned several times with different k-means and different class numbers (see main text). For each clustering we identified the best p-value (Fisher exact test) associated with 5 GO terms ("defense response", "immune system process", "cell cycle", "metabolic process", "developmental process") in any cluster. (b) Classification errors achieved with KNN classifiers discriminating between promoter and enhancer sequences. Boxplots describe the errors obtained using TFcoop scores (red), (di)nucleotide frequencies (green), and the classical PWM scores (blue), using different number of neighbors (K).

marks on the genome, and (ii) non, or poorly, cell-type specific level that regulates TF binding in all DNA regions associated with appropriate marks.

An important property highlighted by our models is that rules governing TF combinations are very similar in the promoters of the three gene types analyzed (mRNA, pri-miRNA and lncRNA), but different between promoters and enhancers. This is further confirmed by our experiments for discriminating between promoter and enhancer sequences showing that scores produced by TFcoop models allow accurate classification between the two types of sequences. Our results thus argue for a prominent role of transcription factor binding as the fundamental determinant of regulatory activity able to distinguish enhancers and promoters (1). Furthermore, as promoters and enhancers produce different RNA molecules (1, 2), our results also suggest that the production of enhancer RNAs (eRNAs) on one hand, and that of mRNAs, lncRNAs, and miRNAs on the other hand, requires a specific and distinct subset of TFs.

Our approach could be improved in several ways. A quite straightforward improvement would be to use the DNAshape score developed by Mathelier et al. (26) instead of the classical PWM score. This could improve TFcoop accuracy for several TFs, especially for TFs such as CTCF for which TFcoop fails to outperform classical PWM scoring. More profoundly, one drawback of TFcoop is that the logistic model enables us to learn only a single TF combination for each target TF. However, we can imagine that certain TFs may be associated with two or more different TF combinations depending on the promoter/enhancer they bind. A solution for this would be to learn a discrimination function based on several logistic models instead of a single one.

SUPPLEMENTARY DATA

Supplementary data are available at: http://www.lirmm.fr/∼brehelin/TFcoop. Models, data and R code (R Markdown file) for reproducing some of experiments described in the paper are available at the same address.

ACKNOWLEDGEMENTS

We thank Anthony Mathelier and Wyeth Wasserman for insightful discussions and suggestions. We are indebted to researchers around the globe who generated experimental data and made them freely available. This work was supported by funding from CNRS, *Plan d'Investissement d'Avenir* #ANR-11-BINF-0002 *Institut de Biologie Computationnelle* (young investigator grant to C-H.L. and post-doctoral fellowship to J.V.), Labex NUMEV (post-doctoral fellowship to J.V.), INSERM-ITMO Cancer project "LIONS" BIO2015-04, and CNRS International Associated Laboratory "miREGEN".

Conflict of interest statement. None declared.

REFERENCES

1. R. Andersson. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *Bioessays*, 37(3):314–323, Mar 2015.
2. R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. J?rgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Muller, A. R. Forrest, P. Carninci, M. Rehli, A. Sandelin, A. R. Forrest, H. Kawaji, M. Rehli, J. K. Baillie, M. J. de Hoon, V. Haberle, T. Lassmann, I. V. Kulakovskiy, M. Lizio, M. Itoh, R. Andersson, C. J. Mungall, T. F. Meehan, S. Schmeier, N. Bertin, M. J?rgensen, E. Dimont, E. Arner, C. Schmid, U. Schaefer, Y. A. Medvedeva, C. Plessy, M. Vitezic, J. Severin, C. A. Semple, Y. Ishizu, R. S. Young, M. Francescatto, I. Alam, D. Albanese, G. M. Altschuler, T. Arakawa, J. A. Archer, P. Arner, M. Babina, S. Rennie, P. J. Balwierz, A. G. Beckhouse, S. Pradhan-Bhatt, J. A. Blake, A. Blumenthal, B. Bodega, A. Bonetti, J. Briggs, F. Brombacher, A. M. Burroughs, A. Califano, C. V. Cannistracti, D. Carbajo, Y. Chen, M. Chierici, Y. Ciani, H. C. Clevers, E. Dalla, C. A. Davis, M. Detmar, A. D. Diehl, T. Dohi, F. Drabl?s, A. S. Edge, M. Edinger, K. Ekwall, M. Endoh, H. Enomoto, M. Fagiolini, L. Fairbairn, H. Fang, M. C. Farach-Carson, G. J. Faulkner, A. V. Favorov, M. E. Fisher, M. C. Frith, R. Fujita, S. Fukuda, C. Furlanello, M. Furuno, J. Furusawa, T. B. Geijtenbeek, A. P. Gibson, T. Gingeras, D. Goldowitz, J. Gough, S. Guhl, R. Guler, S. Gustincich, T. J. Ha, M. Hamaguchi, M. Hara, M. Harbers, J. Harshbarger, A. Hasegawa, Y. Hasegawa, T. Hashimoto, M. Herlyn, K. J. Hitchens, S. J. Ho Sui, O. M. Hofman, I. Hoof, F. Hori, L. Huminiecki, K. Iida, T. Ikawa, B. R. Jankovic, H. Jia, A. Joshi, G. Jurman, B. Kaczkowski, C. Kai, K. Kaida, A. Kaiho, K. Kajiyama, M. Kanamori-Katayama, A. S. Kasianov, T. Kasukawa, S. Katayama, S. Kato, S. Kawaguchi, H. Kawamoto, Y. I. Kawamura, T. Kawashima, J. S. Kempfle, T. J. Kenna, J. Kere, L. M. Khachigian, T. Kitamura, S. P. Klinken, A. J. Knox, M. Kojima, S. Kojima, N. Kondo, H. Koseki, S. Koyasu, S. Krampitz, A. Kubosaki, A. T. Kwon, J. F. Laros, W. Lee, A. Lennartsson, K. Li, B. Lilje, L. Lipovich, A. Mackay-Sim, R. Manabe, J. C. Mar, B. Marchand, A. Mathelier, N. Mejhert, A. Meynert, Y. Mizuno, D. A. de Lima Morais, H. Morikawa, M. Morimoto, K. Moro, E. Motakis, H. Motohashi, C. L. Mummery, M. Murata, S. Nagao-Sato, Y. Nakachi, F. Nakahara, T. Nakamura, Y. Nakamura, K. Nakazato, E. van Nimwegen, N. Ninomiya, H. Nishiyori, S. Noma, T. Nozaki, S. Ogishima, N. Ohkura, H. Ohmiya, H. Ohno, M. Onshima, M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. A. Ovchinnikov, A. Pain, R. Passier, M. Patrikakis, H. Persson, S. Piazza, J. G. Prendergast, O. J. Rackham, J. A. Ramilowski, M. Rashid, T. Ravasi, P. Rizzu, M. Roncador, S. Roy, M. B. Rye, E. Saijyo, A. Sajantila, A. Saka, S. Sakaguchi, M. Sakai, H. Sato, H. Satoh, S. Savvi, A. Saxena, C. Schneider, E. A. Schultes, G. G. Schultz-Tanzil, A. Schwegmann, T. Sengstag, G. Sheng, H. Shimoji, Y. Shimoni, J. W. Shin, C. Simon, D. Sugiyama, T. Sugiyama, M. Suzuki, N. Suzuki, R. K. Swoboda, P. A. 't Hoen, M. Tagami, N. Takahashi, J. Takai, H. Tanaka, H. Tatsukawa, Z. Tatum, M. Thompson, H. Toyoda, T. Toyodo, E. Valen, M. van de Wetering, L. M. van den Berg, R. Verardo, D. Vijayan, I. E. Vorontsov, W. W. Wasserman, S. Watanabe, C. A. Wells, L. N. Winteringham, E. Wolvetang, E. J. Wood, Y. Yamaguchi, M. Yamamoto, M. Yoneda, Y. Yonekura, S. Yoshida, S. E. Zabierowski, P. G. Zhang, X. Zhao, S. Zucchelli, K. M. Summers, H. Suzuki, C. O. Daub, J. Kawai, P. Heutink, W. Hide, T. C. Freeman, B. Lenhard, V. B. Bajic, M. S. Taylor, V. J. Makeev, A. Sandelin, D. A. Hume, P. Carninci, and Y. Hayashizaki. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, Mar 2014.
3. David N. Arnosti and Meghana M. Kulkarni. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of Cellular Biochemistry*, 94(5):890–898, April 2005.
4. Michael F. Berger, Anthony A. Philippakis, Aaron M. Qureshi, Fangxue S. He, Preston W. Estep, and Martha L. Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, 24(11):1429–1435, November 2006. 00467.
5. Alan P. Boyle, Carlos L. Araya, Cathleen Brdlik, Philip Cayting, Chao Cheng, Yong Cheng, Kathryn Gardner, LaDeana W. Hillier, Judith Janette, Lixia Jiang, Dionna Kasper, Trupti Kawli, Pouya Kheradpour, Anshul Kundaje, Jingyi Jessica Li, Lijia Ma, Wei Niu, E. Jay Rehm, Joel Rozowsky, Matthew Slattery, Rebecca Spokony, Robert Terrell, Dionne Vafeados, Daifeng Wang, Peter Weisdepp, Yi-Chieh Wu, Dan Xie, Koon-Kiu Yan, Elise A. Feingold, Peter J. Good, Michael J. Pazin, Haiyan Huang, Peter J. Bickel, Steven E. Brenner, Valerie Reinke, Robert H. Waterston, Mark Gerstein, Kevin P. White, Manolis Kellis, and Michael Snyder. Comparative analysis of regulatory information and circuits across distant species. *Nature*, 512(7515):453–456, August 2014.
6. L. T. M. Dao, A. O. Galindo-Albarran, J. A. Castro-Mondragon, C. Andrieu-Soler, A. Medina-Rivera, C. Souaid, G. Charbonnier, A. Griffon, L. Vanhille, T. Stephen, J. Alomairi, D. Martin, M. Torres, N. Fernandez, E. Soler, J. van Helden, D. Puthier, and S. Spicuglia. Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.*, Jun 2017.
7. D. de Rie, I. Abugessaisa, T. Alam, E. Arner, P. Arner, H. Ashoor, G. Astrom, M. Babina, N. Bertin, A. M. Burroughs, A. J. Carlisle, C. O. Daub, M. Detmar, R. Deviatiiarov, A. Fort, C. Gebhard, D. Goldowitz, S. Guhl, T. J. Ha, J. Harshbarger, A. Hasegawa, K. Hashimoto, M. Herlyn, P. Heutink, K. J. Hitchens, C. C. Hon, E. Huang, Y. Ishizu, C. Kai, T. Kasukawa, P. Klinken, T. Lassmann, C. H. Lecellier, W. Lee, M. Lizio, V. Makeev, A. Mathelier, Y. A. Medvedeva, N. Mejhert, C. J. Mungall, S. Noma, M. Ohshima, M. Okada-Hatakeyama, H. Persson, P. Rizzu, F. Roudnicky, P. S?trom, H. Sato, J. Severin, J. W. Shin, R. K. Swoboda, H. Tarui, H. Toyoda, K. Vitting-Seerup, L. Winteringham, Y. Yamaguchi, K. Yasuzawa, M. Yoneda, N. Yumoto, S. Zabierowski, P. G. Zhang, C. A. Wells, K. M. Summers, H. Kawaji, A. Sandelin, M. Rehli, Y. Hayashizaki, P. Carninci, A. R. R. Forrest, and M. J. L. de Hoon. An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nat. Biotechnol.*, 35(9):872–878, Sep 2017.
8. Y. Diao, R. Fang, B. Li, Z. Meng, J. Yu, Y. Qiu, K. C. Lin, H. Huang, T. Liu, R. J. Marina, I. Jung, Y. Shen, K. L. Guan, and B. Ren. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods*, 14(6):629–635, Jun 2017.
9. Iris Dror, Tamar Golan, Carmit Levy, Remo Rohs, and Yael Mandel-Gutfreund. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Research*, 25(9):1268–1280, January 2015.
10. Jason Ernst and Manolis Kellis. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Research*, 23(7):1142–1154, July 2013.
11. A. R. Forrest, H. Kawaji, M. Rehli, J. K. Baillie, M. J. de Hoon, V. Haberle, T. Lassmann, I. V. Kulakovskiy, M. Lizio, M. Itoh, R. Andersson, C. J. Mungall, T. F. Meehan, S. Schmeier, N. Bertin, M. J?rgensen, E. Dimont, E. Arner, C. Schmidl, U. Schaefer, Y. A. Medvedeva, C. Plessy, M. Vitezic, J. Severin, C. Semple, Y. Ishizu, R. S. Young, M. Francescatto, I. Alam, D. Albanese, G. M. Altschuler, T. Arakawa, J. A. Archer, P. Arner, M. Babina, S. Rennie, P. J. Balwierz, A. G. Beckhouse, S. Pradhan-Bhatt, J. A. Blake, A. Blumenthal, B. Bodega, A. Bonetti, J. Briggs, F. Brombacher, A. M. Burroughs, A. Califano, C. V. Cannistraci, D. Carbajo, Y. Chen, M. Chierici, Y. Ciani, H. C. Clevers, E. Dalla, C. A. Davis, M. Detmar, A. D. Diehl, T. Dohi, F. Drabl?s, A. S. Edge, M. Edinger, K. Ekwall, M. Endoh, H. Enomoto, M. Fagiolini, L. Fairbairn, H. Fang, M. C. Farach-Carson, G. J. Faulkner, A. V. Favorov, M. E. Fisher, M. C. Frith, R. Fujita, S. Fukuda, C. Furlanello, M. Furino, J. Furusawa, T. B. Geijtenbeek, A. P. Gibson, T. Gingeras, D. Goldowitz, J. Gough, S. Guhl, R. Guler, S. Gustincich, T. J. Ha, M. Hamaguchi, M. Hara, M. Harbers, J. Harshbarger, A. Hasegawa, Y. Hasegawa, T. Hashimoto, M. Herlyn, K. J. Hitchens, S. J. Ho Sui, O. M. Hofmann, I. Hoof, F. Hori, L. Huminiecki, K. Iida, T. Ikawa, B. R. Jankovic, H. Jia, A. Joshi, G. Jurman, B. Kaczkowski, C. Kai, K. Kaida, A. Kaiho, K. Kajiyama, M. Kanamori-Katayama, A. S. Kasianov, T. Kasukawa, S. Katayama, S. Kato, S. Kawaguchi, H. Kawamoto, Y. I. Kawamura, T. Kawashima, J. S. Kempfle, T. J. Kenna, J. Kere, L. M. Khachigian, T. Kitamura, S. P. Klinken, A. J. Knox, M. Kojima, S. Kojima, N. Kondo, H. Koseki, S. Koyasu, S. Krampitz, A. Kubosaki, A. T. Kwon, J. F. Laros, W. Lee, A. Lennartsson, K. Li, B. Lilje, L. Lipovich, A. Mackay-Sim, R. Manabe, J. C. Mar, B. Marchand, A. Mathelier, N. Mejhert, A. Meynert, Y. Mizuno, D. A. de Lima Morais, H. Morikawa, M. Morimoto, K. Moro, E. Motakis, H. Motohashi, C. L. Mummery, M. Murata, S. Nagao-Sato, Y. Nakachi, F. Nakahara, T. Nakamura, Y. Nakamura, K. Nakazato, E. van Nimwegen, N. Ninomiya, H. Nishiyori, S. Noma, S. Noma, T. Noazaki, S. Ogishima, N. Ohkura, H. Ohimiya, H. Ohno, M. Ohshima,

M. Okada-Hatakeyama, Y. Okazaki, V. Orlando, D. A. Ovchinnikov, A. Pain, R. Passier, M. Patrikakis, H. Persson, S. Piazza, J. G. Prendergast, O. J. Rackham, J. A. Ramilowski, M. Rashid, T. Ravasi, P. Rizzu, M. Roncador, S. Roy, M. B. Rye, E. Saijyo, A. Sajantila, A. Saka, S. Sakaguchi, M. Sakai, H. Sato, S. Savvi, A. Saxena, C. Schneider, E. A. Schultes, G. G. Schulze-Tanzil, A. Schwegmann, T. Sengstag, G. Sheng, H. Shimoji, Y. Shimoni, J. W. Shin, C. Simon, D. Sugiyama, T. Sugiyama, M. Suzuki, N. Suzuki, R. K. Swoboda, P. A. 't Hoen, M. Tagami, N. Takahashi, J. Takai, H. Tanaka, H. Tatsukawa, Z. Tatum, M. Thompson, H. Toyodo, T. Toyoda, E. Valen, M. van de Wetering, L. M. van den Berg, R. Verado, D. Vijayan, I. E. Vorontsov, W. W. Wasserman, S. Watanabe, C. A. Wells, L. N. Winteringham, E. Wolvetang, E. J. Wood, Y. Yamaguchi, M. Yamamoto, M. Yoneda, Y. Yonekura, S. Yoshida, S. E. Zabierowski, P. G. Zhang, X. Zhao, S. Zucchelli, K. M. Summers, H. Suzuki, C. O. Daub, J. Kawai, P. Heutink, W. Hide, T. C. Freeman, B. Lenhard, V. B. Bajic, M. S. Taylor, V. J. Makeev, A. Sandelin, D. A. Hume, P. Carninci, and Y. Hayashizaki. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470, Mar 2014.

12. Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, April 2011.
13. Chung-Chau Hon, Jordan A. Ramilowski, Jayson Harshbarger, Nicolas Bertin, Owen J. L. Rackham, Julian Gough, Elena Denisenko, Sebastian Schmeier, Thomas M. Poulsen, Jessica Severin, Marina Lizio, Hideya Kawaji, Takeya Kasukawa, Masayoshi Itoh, A. Maxwell Burroughs, Shohei Noma, Sarah Djebali, Tanvir Alam, Yulia A. Medvedeva, Alison C. Testa, Leonard Lipovich, Chi-Wai Yip, Imad Abugessaisa, Mickal Mendez, Akira Hasegawa, Dave Tang, Timo Lassmann, Peter Heutink, Magda Babina, Christine A. Wells, Soichi Kojima, Yukio Nakamura, Harukazu Suzuki, Carsten O. Daub, Michiel J. L. de Hoon, Erik Arner, Yoshihide Hayashizaki, Piero Carninci, and Alistair R. R. Forrest. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*, 543(7644):199–204, 2017. 00009.
14. ukasz Huminiecki and Jarosaw Horbaczuk. Can We Predict Gene Expression by Understanding Proximal Promoter Architecture? *Trends in Biotechnology*, 0(0), April 2017.
15. David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, 316(5830):1497–1502, June 2007. 02068.
16. Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M. Vaquerizas, Jian Yan, Mikko J. SillanpŁŁ, Martin Bonke, Kimmo Palin, Shaheynoor Talukder, Timothy R. Hughes, Nicholas M. Luscombe, Esko Ukkonen, and Jussi Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, 20(6):861–873, June 2010. 00245.
17. Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M. Vaquerizas, Renaud Vincentelli, Nicholas M. Luscombe, Timothy R. Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. DNA-Binding Specificities of Human Transcription Factors. *Cell*, 152(12):327–339, January 2013.
18. Arttu Jolma, Yimeng Yin, Kazuhiro R. Nitta, Kashyap Dave, Alexander Popov, Minna Taipale, Martin Enge, Teemu Kivioja, Ekaterina Morgunova, and Jussi Taipale. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527(7578):384–388, November 2015.
19. Tommy Kaplan, Xiao-Yong Li, Peter J. Sabo, Sean Thomas, John A. Stamatoyannopoulos, Mark D. Biggin, and Michael B. Eisen. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS genetics*, 7(2):e1001290, February 2011.
20. Janne Korhonen, Petri Martinmki, Cinzia Pizzi, Pasi Rastas, and Esko Ukkonen. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics*, 25(23):3181–3182, December 2009.
21. Ivan V. Kulakovskiy, Ilya E. Vorontsov, Ivan S. Yevshin, Anastasiia V. Soboleva, Artem S. Kasianov, Haitham Ashoor, Wail Ba-alawi, Vladimir B. Bajic, Yulia A. Medvedeva, Fedor A. Kolpakov, and Vsevolod J. Makeev. HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Research*, 44(D1):D116–D125, January 2016.
22. Tong Ihn Lee and Richard A. Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, March 2013. 00297.
23. Michal Levo, Einat Zalckvar, Eilon Sharon, Ana Carolina Dantas Machado, Yael Kalma, Maya Lotam-Pompan, Adina Weinberger, Zohar Yakhini, Remo Rohs, and Eran Segal. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research*, 25(7):1018–1029, January 2015.
24. Xiao-Yong Li, Sean Thomas, Peter J. Sabo, Michael B. Eisen, John A. Stamatoyannopoulos, and Mark D. Biggin. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biology*, 12(4):R34, 2011.
25. A. Mathelier, O. Fornes, D. J. Arenillas, C. Y. Chen, G. Denay, J. Lee, W. Shi, C. Shyr, G. Tan, R. Worsley-Hunt, A. W. Zhang, F. Parcy, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 44(D1):D110–115, Jan 2016.
26. A. Mathelier, B. Xin, T. P. Chiu, L. Yang, R. Rohs, and W. W. Wasserman. DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Syst*, 3(3):278–286, Sep 2016.
27. Anthony Mathelier and Wyeth W. Wasserman. The next generation of transcription factor binding site prediction. *PLoS computational biology*, 9(9):e1003214, 2013.
28. Guofeng Meng, Axel Mosig, and Martin Vingron. A computational evaluation of over-representation of regulatory motifs in the promoter regions of differentially expressed genes. *BMC Bioinformatics*, 11(1):267, 2010.
29. Leonid A. Mirny. Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(52):22534–22539, December 2010.
30. Ekaterina Morgunova and Jussi Taipale. Structural perspective of cooperative transcription factor binding. *Current Opinion in Structural Biology*, 47:1–8, December 2017.
31. Anirudh Natarajan, Galip Grkan Yardimci, Nathan C. Sheffield, Gregory E. Crawford, and Uwe Ohler. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Research*, 22(9):1711–1722, September 2012.
32. Kazuhiro R. Nitta, Arttu Jolma, Yimeng Yin, Ekaterina Morgunova, Teemu Kivioja, Junaid Akhtar, Korneel Hens, Jarkko Toivonen, Bart Deplancke, Eileen E. M. Furlong, and Jussi Taipale. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *eLife*, 4, March 2015.
33. Daniel Quang and Xiaohui Xie. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, 44(11):e107–e107, June 2016.
34. A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, Mar 2010.
35. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.
36. Remo Rohs, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. The role of DNA shape in protein-DNA recognition. *Nature*, 461(7268):1248–1253, October 2009.
37. H. G. Roider, A. Kanhere, T. Manke, and M. Vingron. Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141, Jan 2007.
38. Florian Schmidt, Nina Gasparoni, Gilles Gasparoni, Kathrin Gianmoena, Cristina Cadenas, Julia K. Polansky, Peter Ebert, Karl Nordstrm, Matthias Barann, Anupam Sinha, Sebastian Frhler, Jieyi Xiong, Azim DehghaniAmirabad, Fatemeh BehjatiArdakani, Barbara Hutter, Gideon Zipprich, Brbel Felder, Jrgen Eils, Benedikt Brors, Wei Chen, Jan G. Hengstler, Alf Hamann, Thomas Lengauer, Philip Rosenstiel, Jrn Walter, and Marcel H. Schulz. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Research*, 45(1):54–66, January 2017.

39. Richard I. Sherwood, Tatsunori Hashimoto, Charles W. O'Donnell, Sophia Lewis, Amira A. Barkal, John Peter van Hoff, Vivek Karun, Tommi Jaakkola, and David K. Gifford. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology*, 32(2):171–178, February 2014.
40. C. A. Sloan, E. T. Chan, J. M. Davidson, V. S. Malladi, J. S. Strattan, B. C. Hitz, I. Gabdank, A. K. Narayanan, M. Ho, B. T. Lee, L. D. Rowe, T. R. Dreszer, G. Roe, N. R. Podduturi, F. Tanaka, E. L. Hong, and J. M. Cherry. ENCODE data at the ENCODE portal. *Nucleic Acids Res.*, 44(D1):D726–732, Jan 2016.
41. Robert E. Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T. Maurano, Eric Haugen, Nathan C. Sheffield, Andrew B. Stergachis, Hao Wang, Benjamin Vernot, Kavita Garg, Sam John, Richard Sandstrom, Daniel Bates, Lisa Boatman, Theresa K. Canfield, Morgan Diegel, Douglas Dunn, Abigail K. Ebersol, Tristan Frum, Erika Giste, Audra K. Johnson, Ericka M. Johnson, Tanya Kutyavin, Bryan Lajoie, Bum-Kyu Lee, Kristen Lee, Darin London, Dimitra Lotakis, Shane Neph, Fidencio Neri, Eric D. Nguyen, Hongzhu Qu, Alex P. Reynolds, Vaughn Roach, Alexias Safi, Minerva E. Sanchez, Amartya Sanyal, Anthony Shafer, Jeremy M. Simon, Lingyun Song, Shinny Vong, Molly Weaver, Yongqi Yan, Zhancheng Zhang, Zhuzhu Zhang, Boris Lenhard, Muneesh Tewari, Michael O. Dorschner, R. Scott Hansen, Patrick A. Navas, George Stamatoyannopoulos, Vishwanath R. Iyer, Jason D. Lieb, Shamil R. Sunyaev, Joshua M. Akey, Peter J. Sabo, Rajinder Kaul, Terrence S. Furey, Job Dekker, Gregory E. Crawford, and John A. Stamatoyannopoulos. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, September 2012.
42. Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
43. Wyeth W. Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, April 2004.
44. E. Wingender, P. Dietze, H. Karas, and R. Knppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Research*, 24(1):238–241, January 1996.
45. Zeba Wunderlich and Leonid A. Mirny. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in genetics: TIG*, 25(10):434–440, October 2009.
46. Y. Yin, E. Morgunova, A. Jolma, E. Kaasinen, B. Sahu, S. Khund-Sayeed, P. K. Das, T. Kivioja, K. Dave, F. Zhong, K. R. Nitta, M. Taipale, A. Popov, P. A. Ginno, S. Domcke, J. Yan, D. Schubeler, C. Vinson, and J. Taipale. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science*, 356(6337), May 2017.
47. Yue Zhao, Shuxiang Ruan, Manishi Pandey, and Gary D. Stormo. Improved models for transcription factor binding site identification using nonindependent interactions. *Genetics*, 191(3):781–790, July 2012.
48. Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10):931–934, October 2015.