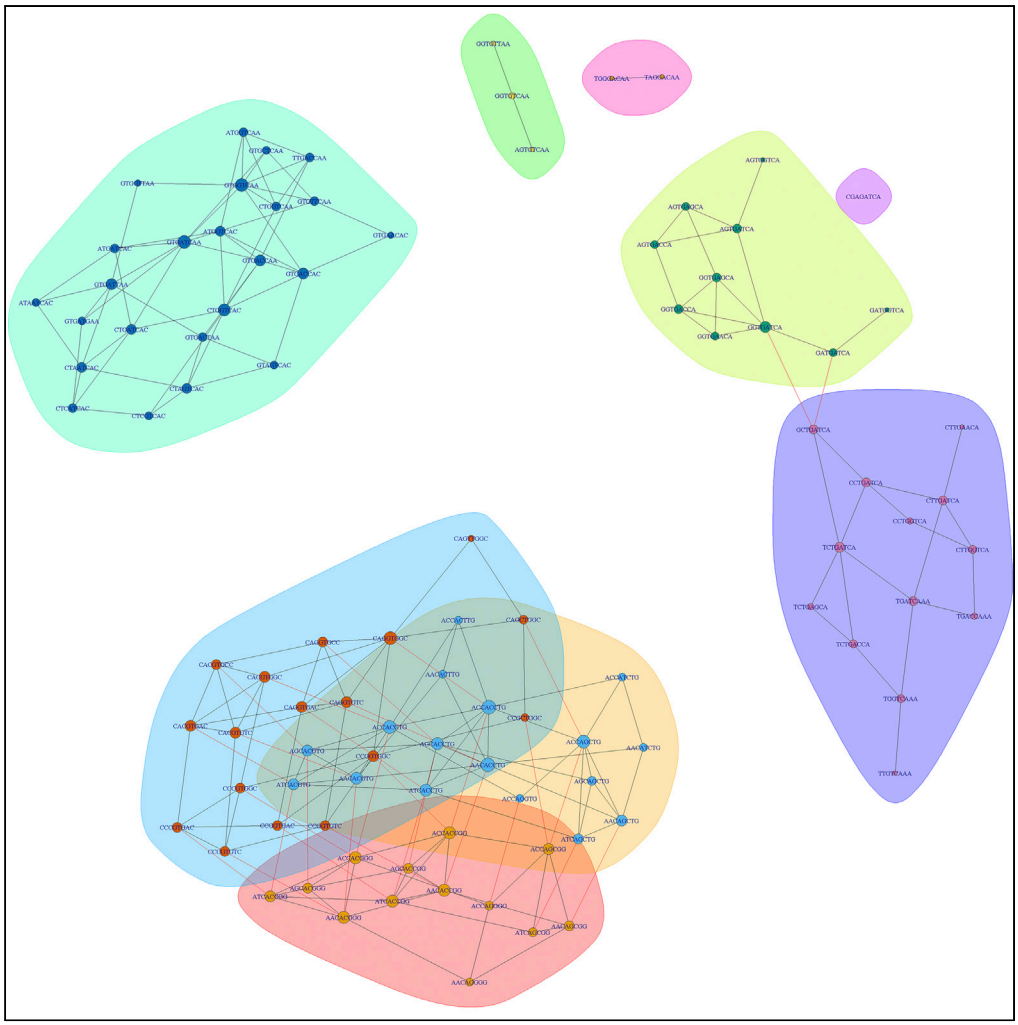
iScience

CellPress

Article

DNA Motif Recognition Modeling from Protein Sequences



Ka-Chun Wong

kc.w@cityu.edu.hk

HIGHLIGHTS

DNA motif modeling from protein is fundamental for understanding gene regulation

A framework is proposed at the highest possible sequence resolution for the first time

It is validated on millions of k-mer intensities from 92 proteins across 5 families

It can prioritize the unobserved regulatory single nucleotide variants on DNA motifs

Wong, iScience 7, 198–211
September 28, 2018
https://doi.org/10.1016/j.isci.2018.09.003

iScience

CellPress

Article

DNA Motif Recognition Modeling from Protein Sequences

Ka-Chun Wong[1,2,*]

SUMMARY

Although the existing works on DNA motif discovery on DNA sequences are plethoric, mechanistic knowledge to infer DNA motifs from protein sequences across multiple DNA-binding domain families without conducting any wet-lab experiments is still lacking. Therefore, the k-spectrum recognition modeling is proposed to address the issues at the highest possible resolutions. The k-spectrum model can capture DNA motif patterns from protein sequences at the resolution in which local sequence context and nucleotide dependency can be taken into account completely. Multiple evaluation metrics are adopted and measured on millions of k-mer binding intensities from 92 proteins across 5 DNA-binding families (i.e., bHLH, bZIP, ETS, Forkhead, and Homeodomain), demonstrating its competitive edges. In addition, it not only can contribute to DNA motif recognition modeling but also can help prioritize the observed or even unobserved binding of single nucleotide variants on transcription factor binding sites in a genome-wide manner.

INTRODUCTION

According to a robust estimation (Li and Biggin, 2015), 73% of protein expression is regulated by gene transcription. Such a percentage is substantially higher than the other steps such as translation (8%), protein degradation (8%), and mRNA degradation (11%). A recent study also indicated that most individuals have unique repertoires of gene transcription activities, which can contribute to phenotypic variations (Barrera et al., 2016) and thus the difficulties in developing personalized medicine. Therefore, understanding gene transcription forms the important basis for personalized medicine development. Especially, the protein-DNA binding interactions between transcription factors (TFs) and transcription factor binding sites (TFBSs) are the essential components in eukaryotic gene transcription where TFs bind to TFBSs in a sequence-specific manner as evidenced by the study that 17.5%–19.5% of the top expression quantitative trait loci are overlapped with the annotated TFBSs (1000 Genomes Project Consortium et al., 2015).

Therefore, substantial efforts have been made into elucidating the DNA-binding specificity patterns (TFBS patterns) of TFs, which are the essential proteins for gene transcription. TFs bind onto specific DNA sites (TFBSs) on regulatory regions (e.g., promoters and enhancers), controlling when and where each gene is transcribed. Given its central importance, the existing protein-DNA binding structures have been analyzed to decipher the protein-DNA binding interactions between TFs and TFBSs (Luscombe et al., 2001) on specific TF families (e.g., zinc fingers [Krishna et al., 2003]). Bonding and force types, bending of the DNA (Jones et al., 1999), TF conservation, and mutations (Luscombe and Thornton, 2002) have been discovered as the key factors that can determine the binding amino acid residues on the TF side (Jones et al., 2003). In addition, many have tried to generalize the binding rules (i.e., the one-to-one mapping between amino acids from TFs and nucleotides from TFBSs). Unfortunately, despite the efforts (Mandel-Gutfreund et al., 1995; Mandel-Gutfreund and Margalit, 1998; Luscombe and Thornton, 2002), it is suggested that there is not any general binding rule across different TF families (Sarai and Kono, 2005).

To address this, high-throughput biotechnologies have been developed such as chromatin immunoprecipitation (ChIP) sequencing technologies (e.g., ChIP-Chip, ChIP-seq, and ChIP-exo), ChIA-PET, HT-SELEX, CHAMP, and ORGANIC (Jung et al., 2017; Kasinathan et al., 2014). In particular, protein binding microarray (PBM) has been developed as a high-throughput technology that can discover the *in vitro* TFBSs (Barrera et al., 2016). Each PBM run can measure binding intensities of a given TF to all possible DNA k-mers ($k \geq 8$).

With these technologies, international projects (e.g., ENCODE, GTEx, and FANTOM) have been successfully launched (ENCODE Project Consortium, 2012; GTEx Consortium, 2015; Forrest et al., 2014),

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong

[2]Lead Contact

*Correspondence: kc.w@cityu.edu.hk

https://doi.org/10.1016/j.isci.2018.09.003

accumulating massive TFBS data that we can study for human TFs on a genome-wide scale; for instance, the sequencing data from the ENCODE consortium has enabled the systematic genome-wide discovery and characterization of human TFBSs (Kheradpour and Kellis, 2014). In addition, Jolma et al. have also characterized the DNA-binding specificity landscape of human TFs for large-scale TFBS data (Jolma et al., 2013).

The *de facto* modeling of TFBS is usually in the form of either solid consensus sequences or position weight matrices (PWMs). Nonetheless, it is well known that nucleotide dependencies and indel operations exist (Tomovic and Oakeley, 2007). Therefore, Wong et al. have proposed to adopt probabilistic modeling to tackle these challenges (Wong et al., 2013). Transcription Factor Flexible Model has also been proposed by Mathelier and Wasserman to solve these issues with variable sequence lengths (Mathelier and Wasserman, 2013). Recently, deep convolutional neural networks have also been applied for large-scale modeling (Alipanahi et al., 2015). Therefore, TFBS modeling is still an active but central challenge in nucleic acids research.

However, these studies are limited to the DNA side. In recent years, given the exponential increase in data, people started to realize that we can predict DNA motifs from the protein side; for instance, Pelossof et al. have successfully adopted the PBM data to predict the DNA-binding affinities of different Homeodomain proteins (Pelossof et al., 2015). Gupta et al. have also proposed a random forest recognition model to predict DNA motif matrices from the protein sequences of the C2H2 zinc finger family (Gupta et al., 2014). Unfortunately, both studies are limited to a single DNA-binding family. In addition, its modeling methodology is either limited to k-mer independence assumption or PWM assumption. The actual modeling performance across multiple DNA-binding families remain speculative. Therefore, in this study, an approach that can take into account sequence context and nucleotide dependency for k-mer dependence modeling at the highest possible resolution is proposed.

DNA Motif Recognition Modeling

For this study, the proposed framework is summarized in Figure 1. The major step descriptions can be referenced from the figure caption. Further details are provided in the following subsections.

Data Sources

The DNA-binding-family-specific recognition models are retrieved from the previous study (Wong et al., 2015). The training and testing k-spectrum data (i.e., e-scores), if available, are retrieved from UniPROBE in October 2017 (Barrera et al., 2016). In particular, the top five DNA-binding domain families with the largest available PBM data are selected and tabulated in Table S1. For each family, leave-one-out cross-validations (LOOCVs) are conducted for fair evaluations. The official Pfam web server is chosen in October 2017 (Finn et al., 2016). The ground truth DNA motif matrices are based on CIS-BP (v1.02) in human (Weirauch et al., 2014).

DNA k-Spectrum Recognition Model Training

DNA k-spectrum refers to the complete quantitative modeling on m DNA k-mers $\{kmer_1, kmer_2, \ldots, kmer_M\}$ of length k. The DNA kmers can be selected based on the DNA-binding intensity ranking from PBM experiments.

Although progress has been made in DNA motif recognition modeling, we lack *in silico* DNA motif recognition models that can achieve such a high resolution for the TF of interest across multiple DNA-binding families. Therefore, we aim at developing the first-of-its-kind *in silico* DNA k-spectrum recognition model that can fully capture the DNA-binding specificities of TFs at a very high resolution across multiple DNA-binding families. As the intermediate step, we have already developed a unique model $Model_{NAR}$, which can predict specificity-determining residue-nucleotide interactions between TFBSs and TFs with known annotations from protein-DNA binding sequences with the help of random forest model training on the existing protein-DNA binding complex structures from Protein Data Bank (PDB) (Wong et al., 2015).

Briefly, given a TF whose DNA-binding specificity we would like to find, we can run the existing protein domain annotation programs (e.g., InterPro and PFam) to annotate its DNA-binding domains (DBDs) on its protein sequence. For the DBD of interest (D), we collect T known PBM k-spectrum profiles

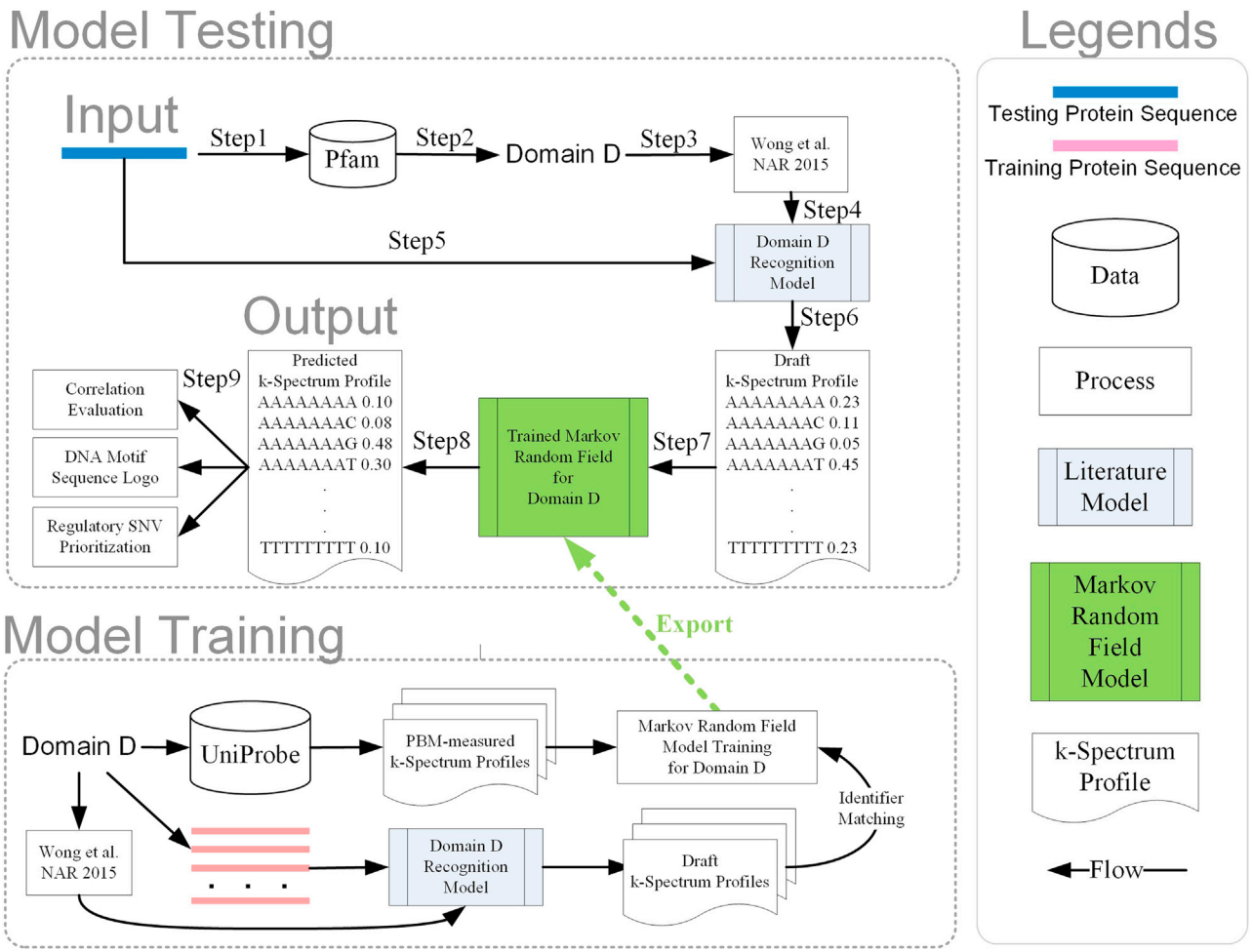


Figure 1. Proposed Framework Overview

(Step 1) The input protein sequence is given for elucidating its DNA-binding specificities. (Step 2) The input protein sequence is scanned with Pfam to retrieve its DNA-binding domain (*D*) annotation. (Step 3) Based on *D*, the corresponding recognition model can be retrieved from the previous study (Wong, 2015). (Steps 4 and 5) The domain (*D*) annotation and its sequence are fed into the retrieved model. (Step 6) A draft DNA k-spectrum is generated from the model. (Step 7) The draft DNA k-spectrum is forwarded into the Markov random field model trained for detailed refinement. (Step 8) The predicted DNA k-spectrum profile is generated from the Markov random field model. (Step 9) Evaluations and applications are conducted under leave-one-out cross-validations (LOOCVs).

$\{PP_D^{(t)} | \forall\, t \leq T \in \mathbb{N}\}$ from UniProbe where t denotes the DBD instance index for model training or building. In each profile $PP_D^{(t)}$, we have the binding intensities of all DNA k-mers of interest:

$$PP_D^{(t)} = \left\{ P_{kmer_1}^{(t)}, P_{kmer_2}^{(t)}, \ldots, P_{kmer_M}^{(t)} \right\}$$

where $P_{kmer_a}^{(t)}$ denotes the actual binding intensity of $kmer_a$ as measured by PBM. In addition, we can also collect the corresponding k-spectrum recognition profile $RP_D^{(t)}$ from the model $Model_{NAR}$ using DNA motif recognition modeling techniques for each DBD instance based on its protein sequence (Wong et al., 2015):

$$RP_D^{(t)} = \left\{ I_{kmer_1}^{(t)}, I_{kmer_2}^{(t)}, \ldots, I_{kmer_M}^{(t)} \right\}$$

where $I_{kmer_a}^{(t)}$ denotes the binding intensity of $kmer_a$ as recognized by $Model_{NAR}$. As a result, our objective is to build a mathematical model that can predict $PP_D^{(t)}$ from $RP_D^{(t)}$, given the T known instances of the DBD of interest (D). Statistically, it is a sparse multi-task regression problem where the predictor variables are $RP_D^{(t)} = \{I_{kmer_1}^{(t)}, I_{kmer_2}^{(t)}, \ldots, I_{kmer_M}^{(t)}\}$ and the response variables are $PP_D^{(t)} = \{P_{kmer_1}^{(t)}, P_{kmer_2}^{(t)}, \ldots, P_{kmer_M}^{(t)}\}$. However, such a formulation does not take into account the sequence neighborhood among different k-mers.

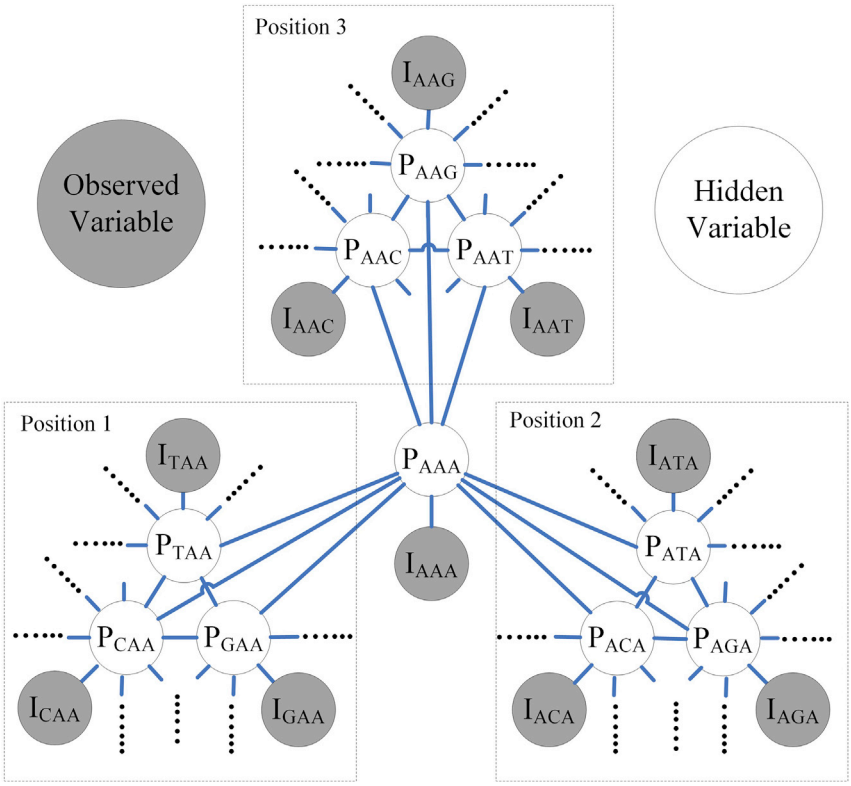


Figure 2. A Markov Random Field Example of the Sequence Neighborhood *NB* ($kmer_a$ = *AAA*) with One Substitution Error

Reverse complements are not shown for illustrative purposes.

Therefore, a Markov Random Field (MRF) model $Model_{MRF} = \{(f_{ab}, g_a) | \forall a, b \leq M \in \mathbb{N}\}$ is proposed to capture the sequence neighborhood information for sparse multi-task regression. Mathematically, the energy function E of $Model_{MRF}$ is given by:

$$E = \sum_{t=1}^{T} \left(\sum_{a=1}^{M} g_a \left(P_{kmer_a}^{(t)}, I_{kmer_a}^{(t)} \right) + \sum_{a=1}^{M} \sum_{b \in NB(a)} f_{ab} \left(P_{kmer_a}^{(t)}, P_{kmer_b}^{(t)} \right) \right)$$

where T is the number of DBD instances, M is the number of DNA k-mers, g_a and f_{ab} are the MRF clique potential functions, and $NB(a)$ is the sequence neighborhood of $kmer_a$, which can be defined in different settings. An example of the sequence neighborhood $NB(kmer_a = AAA)$ with one substitution error is illustrated in Figure 2. To ensure its function convexity, the classic least square error estimation formulation is adopted: $g_a(P_{kmer_a}^{(t)}, I_{kmer_a}^{(t)}) = (P_{kmer_a}^{(t)} - s_a I_{kmer_a}^{(t)} - i_a)^2$ and $f_{ab}(P_{kmer_a}^{(t)}, P_{kmer_b}^{(t)}) = (P_{kmer_a}^{(t)} - P_{kmer_b}^{(t)})^2$. Given the training data $\{PP_D^{(t)}\}$ and $\{RP_D^{(t)}\}$, we can take partial derivatives to the energy function E with respect to the regression parameters $\{s_a\}$ and $\{i_a\}$, resulting in the typical ordinary least aquares estimations for the model training of $Model_{MRF}$.

DNA k-Spectrum Recognition Model Testing

For each DBD of interest (D), its known PBM k-spectrum profiles $\{PP_D^{(t)} | \forall t \leq T \in \mathbb{N}\}$ have already been collected from UniProbe where t denotes the DBD instance index for model training on the top M k-mers (e-scores). On the other hand, its corresponding k-spectrum recognition profiles $\{RP_D^{(t)} | \forall t \leq T \in \mathbb{N}\}$ have also been generated via the model $Model_{NAR}$ (Wong et al., 2015). To test the model $Model_{MRF}$, LOOCVs are conducted on those DBD instances. For each left-out DBD instance indexed by t, given its k-spectrum recognition profile $RP_D^{(t)}$, we have to compute for its predicted PBM k-spectrum profile $PP_D^{(t)\prime} = \{P_{kmer_1}^{(t)\prime}, P_{kmer_2}^{(t)\prime}, \ldots, P_{kmer_M}^{(t)\prime}\}$ via the model $Model_{MRF}$, which has been trained on the other instances. Unfortunately, we do not have any closed form solution without any distribution assumption. Therefore,

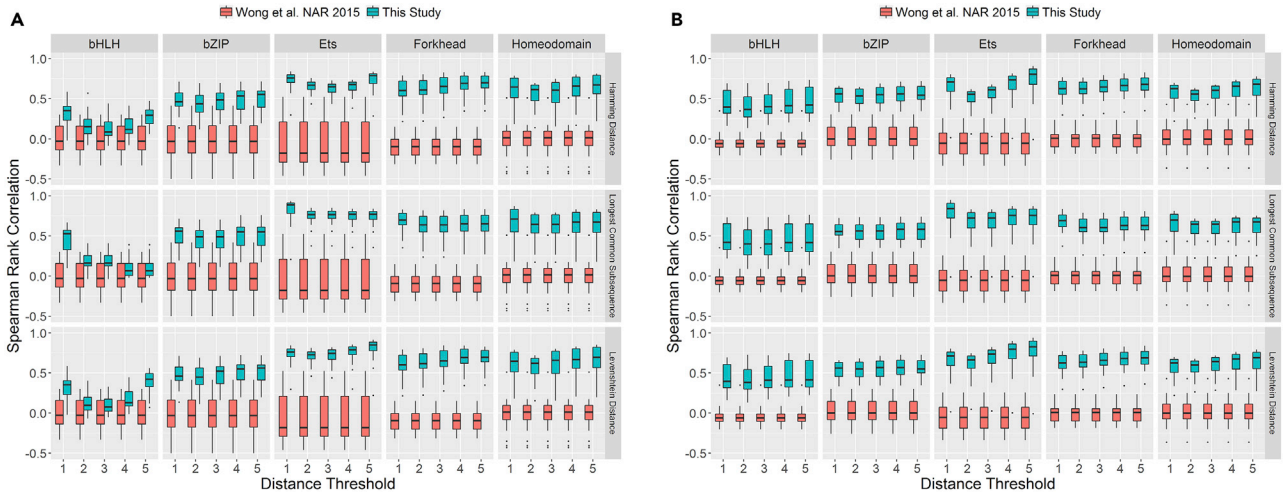


Figure 3. DNA-Binding Intensity Correlation Comparisons under Different k-mer Neighborhood Settings

Box plots on the Spearman rank correlations between the actual binding intensities of k-mers and the predicted binding intensities of k-mers using the previous method (Wong et al., 2015), and the current method denoted in red and blue colors, respectively.

(A) Number of top k-mers (M) = 100.

(B) Number of top k-mers (M) = 1,000.

the iterative formula is obtained by taking partial derivatives to the energy function E with respect to each $P^{(t)}_{kmer_a}$ for all $kmer_a$ (i.e., $\forall a \leq M \in \mathbb{N}$):

$$P^{(t)}_{kmer_a} = \frac{1}{1+|NB(a)|}\left(\left(s_a I^{(t)}_{kmer_a} + i_a\right) + \sum_{b \in NB(a)} P^{(t)}_{kmer_b}\right)$$

Such a formulation has the linear complexity $O(|NB(a)|)$. Convexity is also guaranteed because of the least square error estimation setting. Therefore, it can be easily iterated from random initialization until convergence or maximal number of iterations, resulting in the overall linear complexity $O(IM|NB(a)|)$ where I is the actual number of iterations (i.e., less than 50 iterations in most cases). At the implementation level, we can also combine the equations for all $kmer_a$ into a system of linear equations and solve it via the matrix inverse operation if sufficient memory is available.

DNA k-mer Neighborhood Modeling

Given the central importance of the DNA k-mer neighborhood modeling (*NB*), different possible definitions are parameterized and compared in this study. In particular, three distance metrics are proposed for measuring k-mer sequence similarity: Hamming distance (Norouzi et al., 2012), longest common subsequence (LCS) distance (Paterson and Dančík, 1994), and Levenshtein distance (also known as edit distance) (Yujian and Bo, 2007). Notably, LCS distance is a special case of Levenshtein distance. To extensively cover all possible scenarios, the distance thresholds from one to five are enumerated for each distance metric, resulting in 15 different definitions of DNA k-mer neighborhood *NB*.

RESULTS

DNA-Binding Intensity Correlation

Given the linear model complexities, for each DBD (*D*), the aforementioned LOOCV procedures are conducted to obtain the predicted PBM k-spectrum profiles $\{PP^{(t)\prime}_D | \forall t \leq T \in \mathbb{N}\}$ and compared with the actual PBM k-spectrum profiles $\{PP^{(t)}_D | \forall t \leq T \in \mathbb{N}\}$ where $k = 8$, following previous studies (Weirauch et al., 2013; Wong et al., 2013; Zhao and Stormo, 2011; Chen et al., 2007). Spearman rank correlation coefficients are computed for evaluations as depicted in Figure 3. It can be observed that the proposed approach performs better than the previous approach in capturing DNA-binding intensities (Wong et al., 2015); for instance, for most of the cases, it can achieve the correlation values well above 0.5 and up to 0.9. However, the proposed approach cannot work very well with the bHLH and bZIP families partly due to their dimerization nature. The bHLH family is especially the worst because its TFs are known for its structural dimer flexibility as

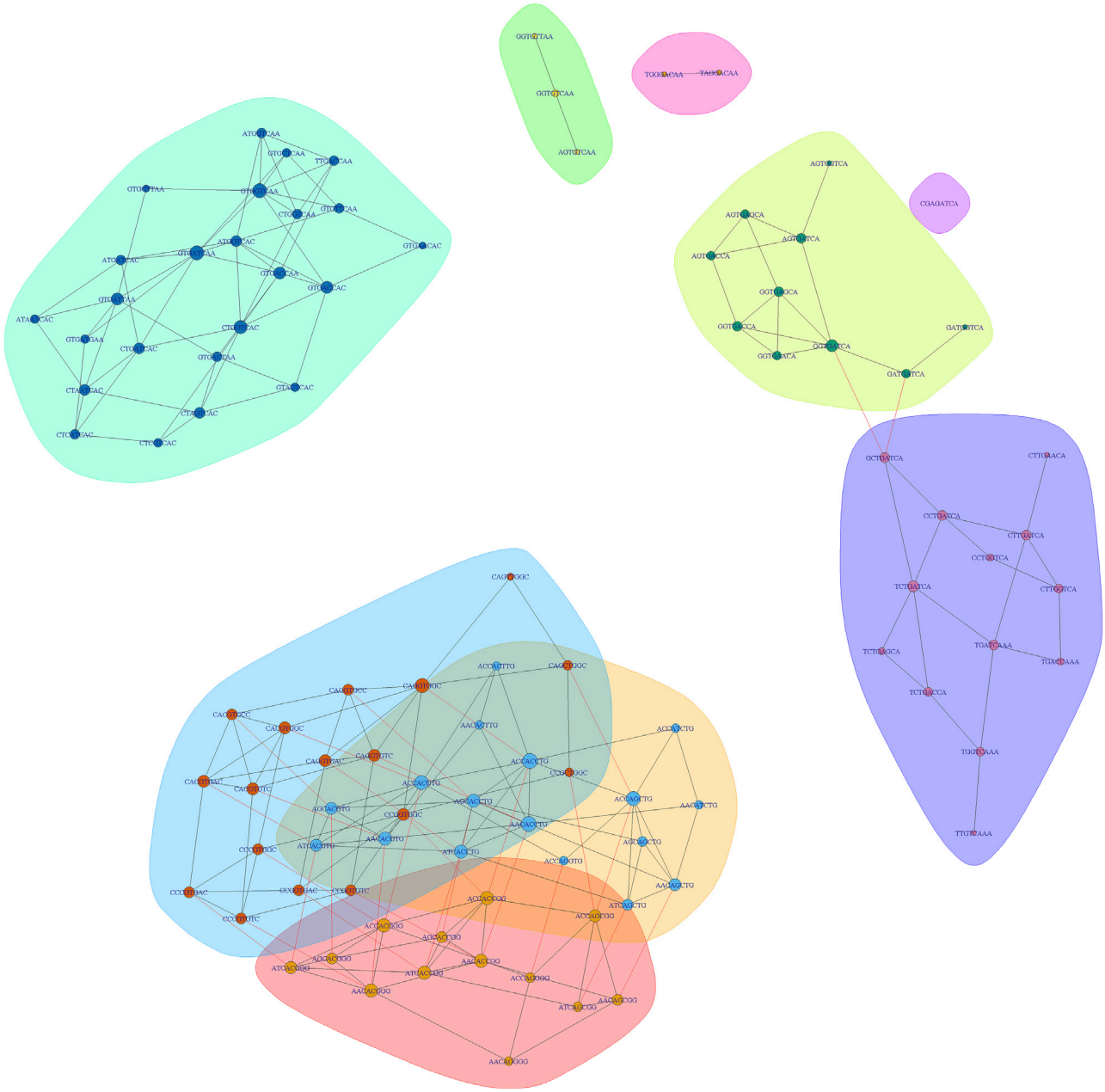


Figure 4. Markov Random Field Network of the Hidden Variables for the bHLH Family When 100 Top k-mers Are Selected (*M*) for Modeling

Hamming distance threshold of one is adopted for sequence neighborhood connections while reverse complements are considered equivalent. The edge betweenness community detection method has been adopted to segregate the Markov network into different communities for modularity maximization (i.e., "cluster_edge_betweenness" function in R), as highlighted in different colors. The node sizes are proportional to the node degrees.

well as loose DNA-binding contacts, leading to diverse DNA motif instances that are difficult to be captured (Ellenberger, 1994).

To investigate it further, the MRF networks of the hidden variables (i.e., the predicted PBM k-spectrum profile variables) of the five DNA-binding families when the number of top k-mers (M) is 100 are plotted in Figures 4 and S1–S4. The corresponding performance values are visualized in Figure 3. The edge betweenness community detection method is applied to segregate each Markov network into different communities for modularity maximization (i.e., the "cluster_edge_betweenness" function in R).

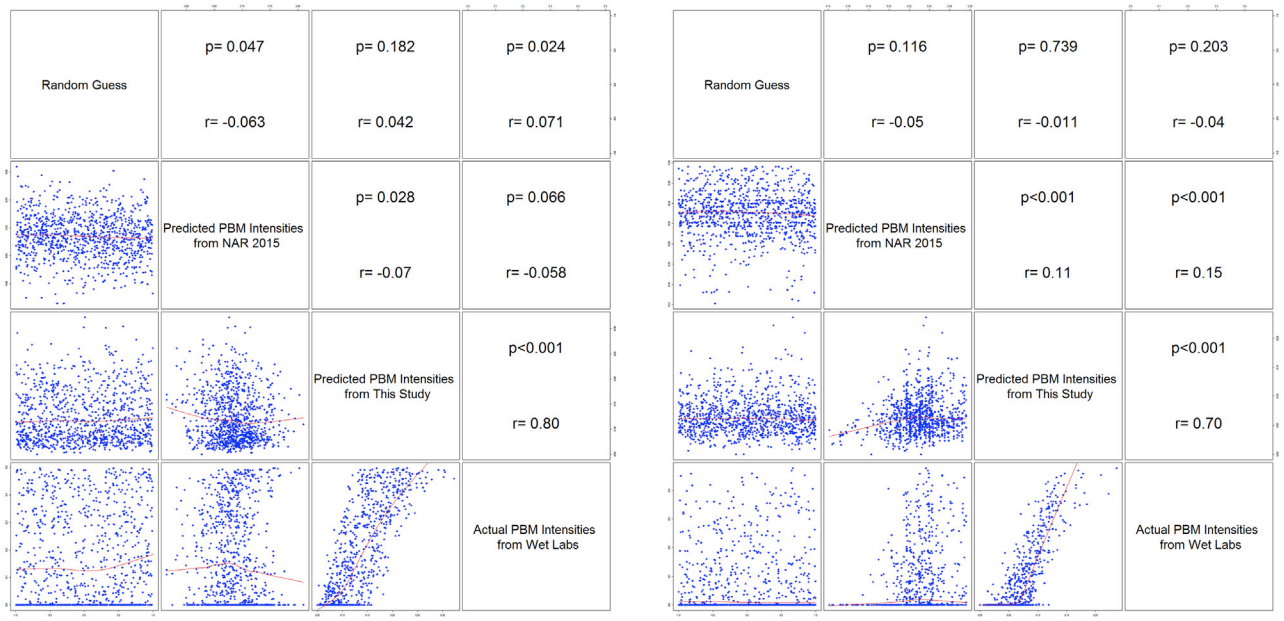


Figure 5. Examples of Left-One-Out Cross-Validation Predictions on 1,000 Top k-mers Whose Neighborhood Is Defined Based on the Hamming Distance Threshold of One

Each dot represents one k-mer. The evaluations are based on Spearman rank correlations (*r*). Random initialization shows the initial guess before the iterations. The figures are drawn using R, where the red curves are local polynomial regression fittings with $\alpha = 2/3$ and the p values are computed using algorithm AS 89 (Best and Roberts, 1975). Additional examples are depicted in Figures S6–S8 for illustration purposes.

(A) ERG from ETS family.

(B) PROP1 from Homeodomain family.

Interestingly, based on the community structures, we can observe several characteristics and reasons for the aforementioned family-specific performance difference. (1) The bHLH family has multiple disjoint communities that impose difficulties in synergizing the sequence neighborhood information for binding intensity correlation modeling. (2) The bZIP and Homeodomain families have multiple singleton communities that can isolate the related k-mers from other communities for network modeling. (3) The ETS and Forkhead families have strongly overlapping communities that can boost sequence neighborhood information for DNA motif recognition modeling.

Based on these observations, we may wonder whether we can improve the modeling performance by increasing the number of DNA k-mers. In this study, the number of top DNA k-mers *M* is varied among {100,200,300,500,1,000} as depicted in Figure S5. Interestingly, it can be observed that the proposed approach can be improved with the increasing number of top k-mers, regardless of the DNA-binding family type. It indicates that our MRF modeling can be scaled for improvement once rich k-mer information is made available. Two left-out examples are shown in Figure 5. The first example concerns the DNA-binding specificities of ERG, which is an oncogene related to hematopoiesis, whereas the second example belongs to PROP1, which is responsible for hormone regulation. Additional examples are depicted in Figures S6–S8.

DNA Motif Pattern Recognition

Given those top k-mers ranked, consistent with the previous study by Zhao and Stormo (Zhao and Stormo, 2011), we pick the top 25 scoring k-mers and compare them with the previous studies' top 25 k-mers (Wong et al., 2015). In particular, for each left-out DBD instance, the literature procedures are followed to align the top 25 scoring k-mers and build DNA motif matrices, which are then compared with the ground truth DNA motif matrices in CIS-BP (v1.02) as indexed by the protein names (Weirauch et al., 2014). To quantify the DNA motif matrix pattern similarities, the previously published motif matrix distance is adopted (Wong

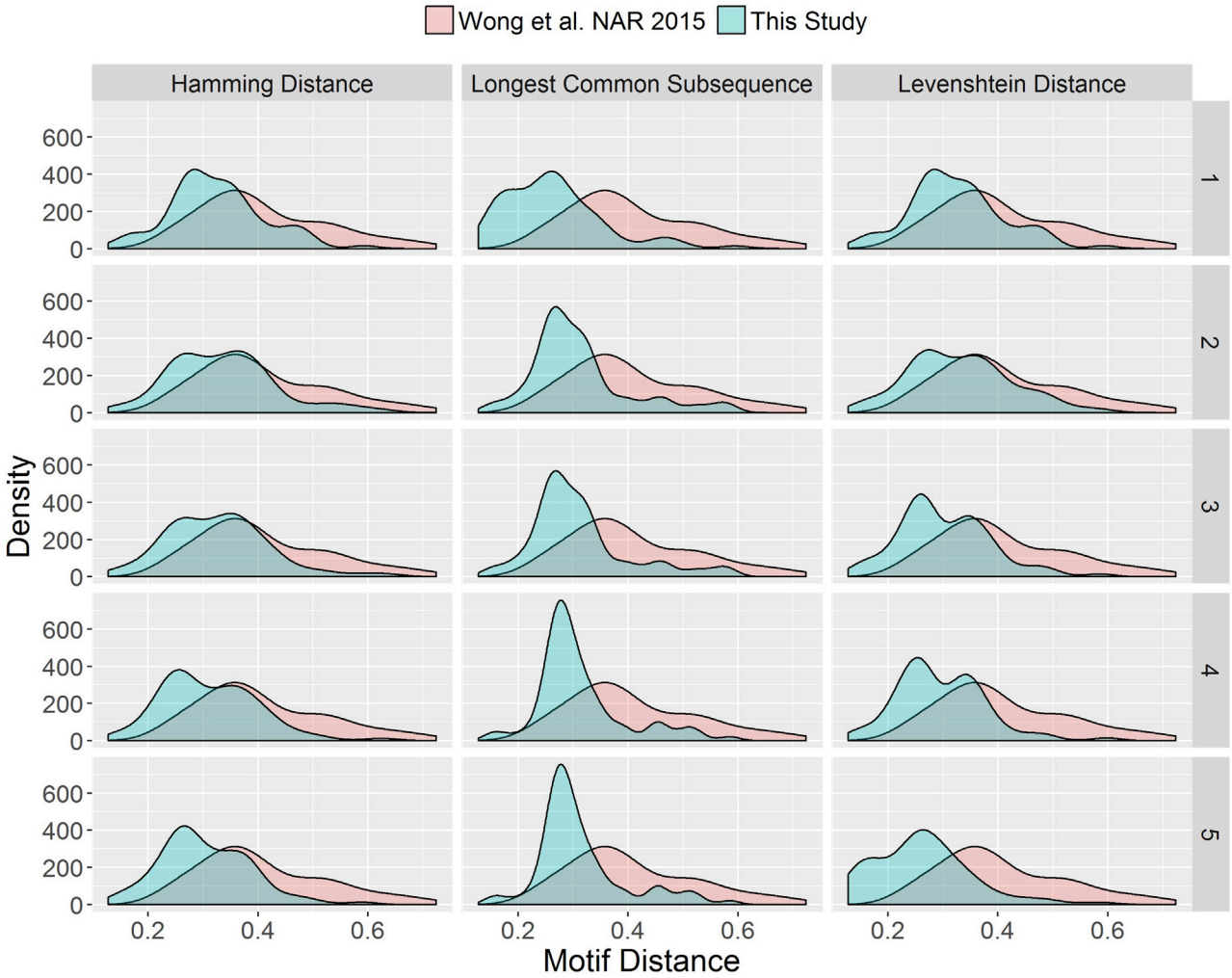


Figure 6. Complete DNA Motif Pattern Recognition Performance Comparisons under Different k-mer Neighborhood Settings

Distribution plots on the DNA motif matrix distances (Wong et al., 2013) between the actual DNA motif matrices in CIS-BP (v1.02) (Weirauch et al., 2014) and the predicted DNA motif matrices using the previous method (Wong et al., 2015) and the current method denoted in red and blue colors, respectively The horizontal axis denotes different neighborhood distance metrics, whereas the vertical axis denotes different distance thresholds.

et al., 2013). All resultant DNA motif matrices are compared with the ground truth DNA motif matrices in CIS-BP (v1.02) (Weirauch et al., 2014). The results are depicted in Figure 6.

From the figure, it can be observed that the motif distance distribution of our generated motifs are skewed more toward the left side (zero side) than the previous approach (Wong et al., 2015); it indicates that the DNA motif matrices generated by our approach are more similar to the ground truth DNA motif matrices than those generated by the previous approach (Wong et al., 2015). In particular, if the k-mer neighborhood is defined using the LCS distance, the DNA motif matrices are consistently similar to the ground truth across different thresholds. It is important as the ground truth DNA motif matrices in CIS-BP (v1.02) are built on both *in vivo* and *in vitro* technologies such as ChIP-seq, HT-SELEX, and PBM. It can be observed that the DNA motif matrices generated in this study not only can capture the *in vitro* DNA-binding specificities but also have the potential to infer *in vivo* DNA-binding specificities. Examples are visualized in Figure 7. Additional examples are depicted in Figures S9–S12 for illustration purposes.

Although the proposed approach has been extensively benchmarked across five DNA-binding families (i.e., bHLH, bZIP, ETS, Forkhead, and Homeodomain), one may wonder about its performance compared

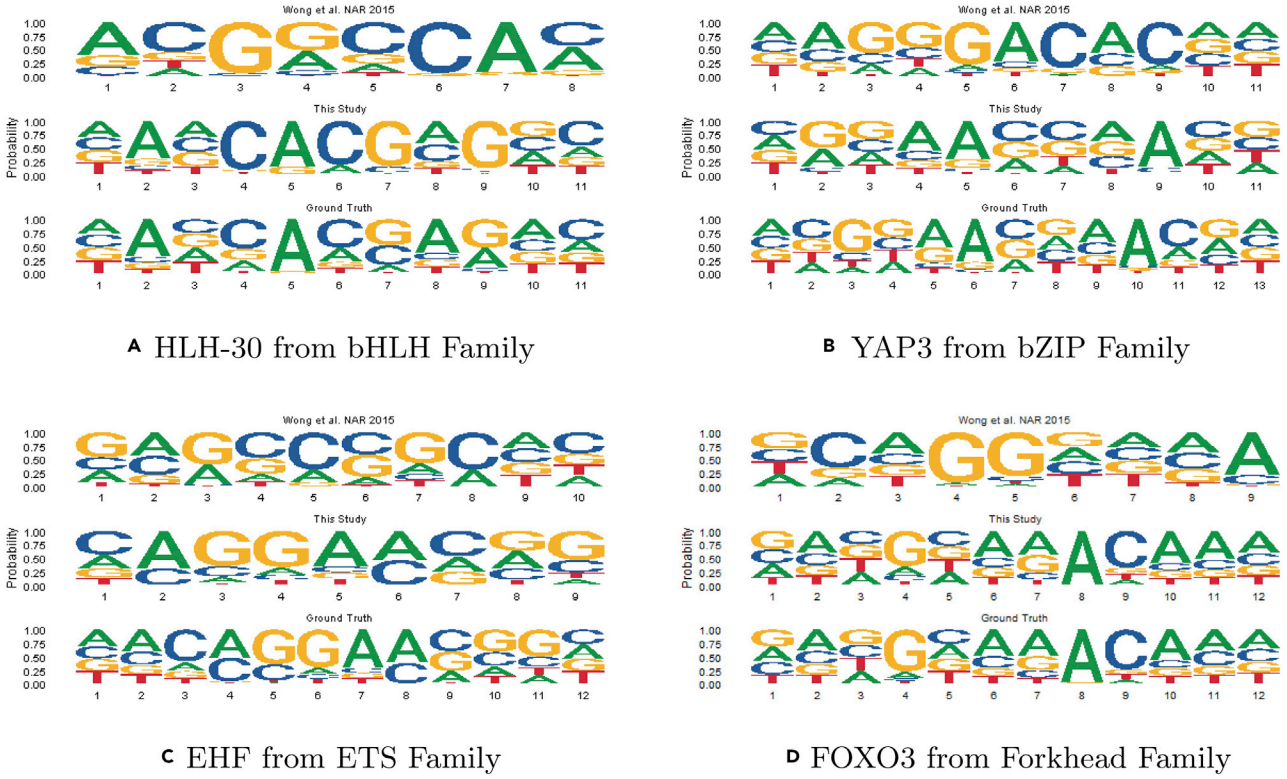


Figure 7. Examples of DNA Motif Matrices Generated by the Previous Approach and Our Approach and the Actual DNA Motif Matrices as Measured using PBM

The first two motif matrices are based on the left-one-out cross-validation predictions on 1,000 top k-mers whose neighborhood is defined based on the Hamming distance threshold of one. All settings strictly follow the protocols established by Zhao and Stormo (Zhao and Stormo, 2011). Additional examples are depicted in Figures S9–S11 for illustration purposes.

(A) HLH-30 from bHLH family.

(B) YAP3 from bZIP family.

(C) EHF from ETS family.

(D) FOXO3 from Forkhead family.

with the existing motif recognition modeling approaches. Therefore, the existing works on DNA motif elucidation are surveyed. Unfortunately, most of the existing works are based on verified DNA sequences (e.g., ChIP-seq, HiTS-FLIP, and DNase hypersensitivity data) (Wang et al., 2014; Khamis et al., 2018; Dai et al., 2017). It is not fair to do the comparisons. On the other hand, the most related work is limited to a specific DNA-binding family (i.e., homeodomain) (Pelossof et al., 2015). Therefore, the performance comparisons are conducted on the Homeodomain motif recognition tasks here.

Pelossof et al. have proposed an affinity regression approach to infer DNA motifs from Homeodomain protein sequences (Pelossof et al., 2015). It is applied to generate and compare their DNA motifs against our DNA motifs. All experimental settings follow the standard setting by Zhao and Stormo (Zhao and Stormo, 2011). The results are depicted in Figure 8. It can be observed that our approach can generate the Homeodomain motifs more similar to the CIS-BP motifs than the other approaches, demonstrating its competitive edges. The comparison is significant as the Pelossof method already has 163 Homeodomain motif data for model building, whereas our approach only has 19 Homeodomain motif data for model building under LOOCVs. Some of the motif examples are depicted in Figure S13 for illustration purposes. Its success can be attributed to the k-mer community segregation ability of the underlying MRF modeling as exemplified in Figure S4, where you can see that the Homeodomain top k-mers are segregated into different communities, consistent with the existing knowledge that the Homeodomain family members have independently evolved into different subtypes for its DNA-binding specificity over the past years (Berger et al., 2008).

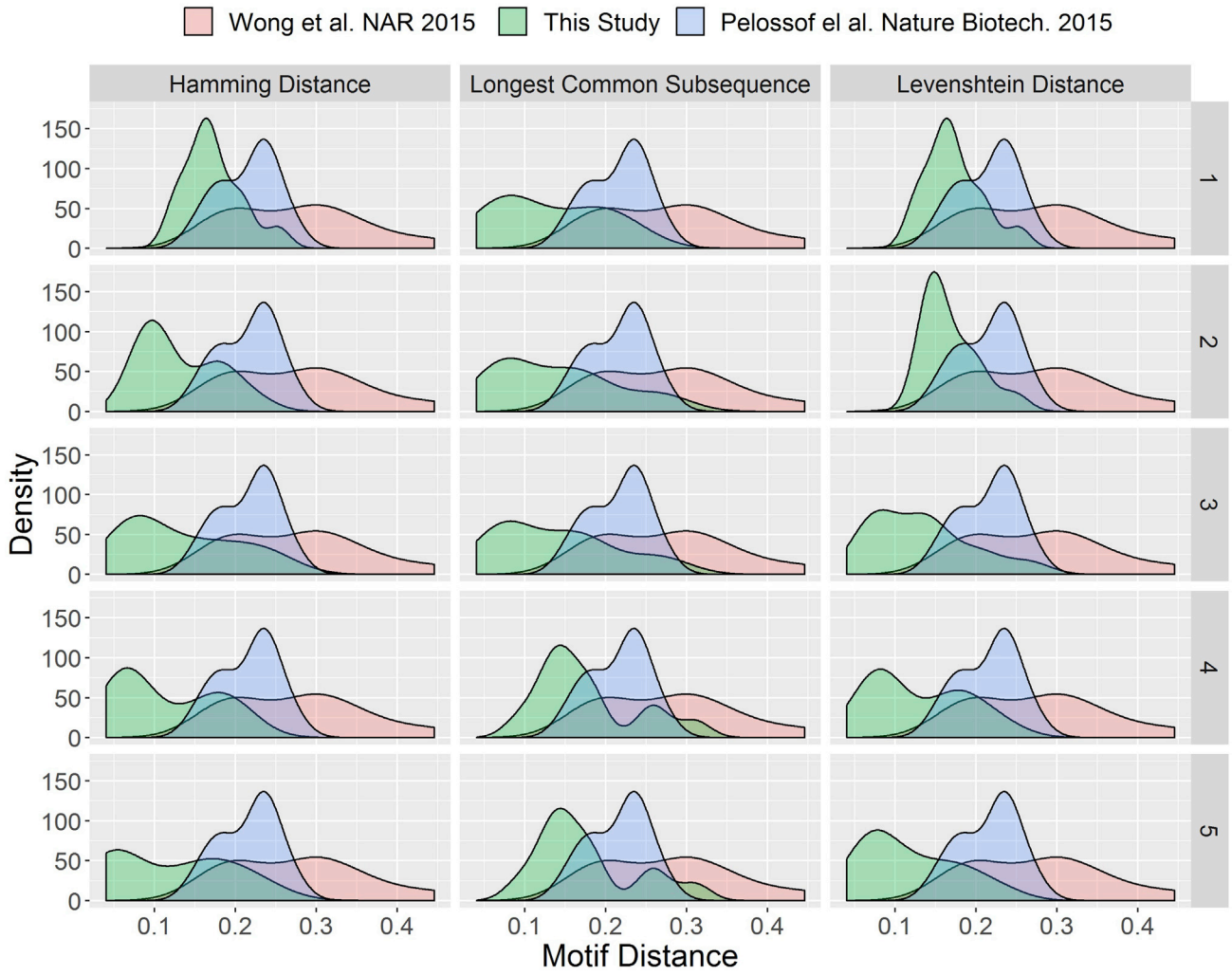


Figure 8. Homeodomain DNA Motif Pattern Recognition Performance Comparisons under Different k-mer Neighborhood Settings

Distribution plots on the Homeodomain DNA motif matrix distances (Wong et al., 2013) between the actual DNA motif matrices in CIS-BP (v1.02) (Weirauch et al., 2014) and the predicted DNA motif matrices using the previous method (Wong et al., 2015), Pelossof method (Pelossof et al., 2015), and the current method denoted in red, blue, and green colors, respectively. The horizontal axis denotes different neighborhood distance metrics, whereas the vertical axis denotes different distance thresholds of the current method.

Single Nucleotide Variants on DNA Motifs

In the previous studies, significant efforts have been made and relied on DNA motif matrices for regulatory single nucleotide variant (rSNV) prioritization (Macintyre et al., 2010; Guo et al., 2013; Zeng et al., 2016) with additional genomic information such as DNase accessibility and chromatin features (Shi et al., 2016; Li et al., 2016) on TFBSs; for instance, atSNP was proposed to address the rSNV prioritization challenge with ENCODE and JASPAR motif matrices (Zuo et al., 2015). SNP2TFBS was also built as an online resource that can bridge the annotation gap between rSNV and the classic SNV databases based on existing motif matrices automatically (Kumar et al., 2017). Unfortunately, motif matrices are well known for their position independence assumption (Benos et al., 2002). Although several improvements in DNA motif modeling have been proposed to address the issues (Mathelier and Wasserman, 2013; Wong et al., 2013), the combinatorial space of the sequence context in DNA motif modeling remains computationally expensive to be modeled around the rSNVs.

The current study offers a novel opportunity for us to quantify the effects of rSNVs on DNA motifs, thanks to its high-resolution k-spectrum modeling methodology. Given an rSNV on DNA motif instances known to be bound by a family-specific TF, we can apply our model to compute the score difference between the

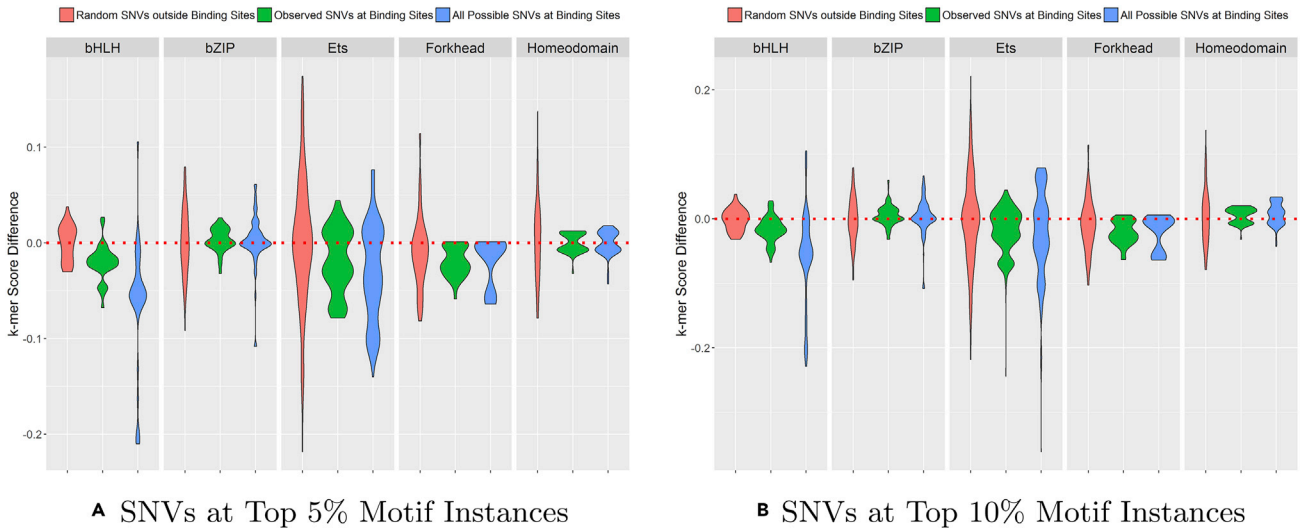


Figure 9. Violin Plots on the k-mer Score Difference Distributions of the SNVs at Top-Ranked Motif Instances via Family-Specific Recognition Modeling

The observed SNVs are retrieved from the clinically verified dataset, ClinVar (version 20171029), whereas the DNA motif instances are ranked by TFBSTools (Tan and Lenhard, 2016). Other settings are visualized in Figure S14.

(A) SNVs at top 5% motif instances.

(B) SNVs at top 10% motif instances.

reference sequence and its variant; for example, at Chr3: 38591950 (on Human Assembly GRCh37), an SNV (rs972777761) has been reported where the reference adenine (A) is changed to guanine (G) (or thymine [T] to cytosine [C] on the opposite strand). At the same time, we have also detected a surrounding motif instance (AAGGAAGTG) as bound by the ELF1 protein from the ETS family between Chr3:38591945 and Chr3:38591953. We can enumerate the reference k-mers and alternate k-mers surrounding the SNV, for instance, the 8-mers AGGAAGTG and AGGAGGTG, and compute their score difference using the ETS family model in our study. Since the scores represent k-mer binding intensities, the score difference and its sign could quantify the DNA-binding effect of the candidate SNV for mechanistic prioritization. In this case, our model returns −0.15, which indicates that the SNV could disrupt the DNA-binding affinity of ELF1 and thus its downstream regulation.

To provide genome-wide insights, we have adopted the DNA motifs generated in our previous section and scanned the humane genome (GRCh37) using TFBSTools (Tan and Lenhard, 2016). In particular, we have selected the advanced genome-wide phylogenetic footprinting function (i.e., "searchPairBSgenome" in R) to improve the scanning performance by chaining the human genome (GRCh37) over the mouse genome (mm10), taking advantage of evolutionary conservation (Wasserman and Sandelin, 2004). Once scanned, we overlapped those motif instances with the clinically verified SNVs from ClinVar (version 20171029), resulting in 303,666 motif instances overlapping with known SNVs (Fisher's exact test p value < 0.001). For each DNA-binding family, we focus on the top DNA motif instances as ranked by TFBSTools (Tan and Lenhard, 2016). The family-specific results with different top ranks are depicted in Figure 9. For completeness, additional results with other parameter settings are depicted in Figure S14.

Interestingly, it can be observed that different DNA-binding families have strikingly different score difference distributions of observed SNVs from the clinically verified dataset, ClinVar (version 20171029). The bHLH, ETS, and Forkhead families have the score difference distributions skewed toward negativity. It indicates that the SNVs on the DNA motif instances bound by the TFs from those three families usually act by disrupting the corresponding protein-DNA binding interactions. In contrast, we observed the opposite trend for the bZIP family. The trend is even more complicated for the Homeodomain family as we can observe bimodal distributions for the SNVs on the DNA motif instances bound by its TFs, regardless of the DNA motif instance ranks as observed from Figures 9 and S14. On the other hand, we also observe that the clinically verified SNVs are more neutral than the possible SNVs on the DNA motif instances which are yet to be observed, consistent with the previous finding that negative selection pressures have been observed on DNA motif instances (Vorontsov et al., 2016).

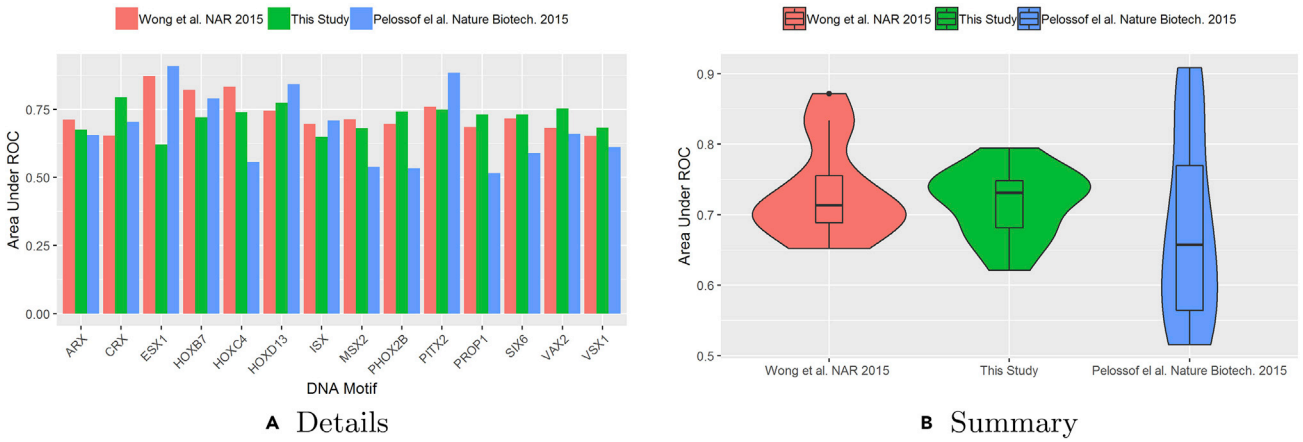


Figure 10. Single Nucleotide Variants (SNV) Prioritization on DNA Motifs according to ClinVar and TFBSTools

SNV prioritization performance comparisons based on area under receiver operating characteristics (ROCs) on Homeodomain motif rSNVs using the previous method (Wong et al., 2015), Pelossof method (Pelossof et al., 2015), and the current method denoted in red, blue, and green colors, respectively.

(A) Bar chart details.

(B) Violin and boxplot summary.

To support the aforementioned insights, the SNVs overlapping with the Homeodomain motif instances are extracted, resulting in 126,924 SNVs clinically verified in ClinVar (version 20171029). On the other hand, the previous setting is followed to generate an equal number of random SNVs as the control. Therefore, it constitutes a two-class SNV dataset (i.e., ClinVar versus control) for rSNV prioritization benchmark comparisons. The previous methods and the affinity regression approach (Pelossof et al., 2015) are run on those SNVs for prioritization. The k-mer frequencies and its differences are computed to ascertain the significance of each rSNV on those Homeodomain motif instances. The resultant areas under receiver operating characteristics values on different motifs are depicted and summarized in Figure 10. It is not surprising that different approaches have their own competitive edges on different DNA motifs for rSNV prioritization. Therefore, we seek to summarize them in the second figure, where we can observe that our approach is slightly better than the previous approaches as a whole. However, we note that we cannot ascertain any statistical significance using hypothesis testing (i.e., t test and rank-sum test) here. It indicates that our approach is comparable to the existing state-of-the-art approaches including the previous method published in Nucleic Acids Research and the affinity regression method published in Nature Biotechnology.

DISCUSSION

DNA motif recognition modeling offers opportunities for us to infer DNA motifs from protein sequences. Such an approach is not only applicable in the cases in which the direct evidences are unavailable but also holds promise for us to understand the DNA-binding specificities from the first principle on the protein side.

The proposed approach directly addressed such issues at the unprecedented resolution based on the k-spectrum MRF modeling. It has been extensively benchmarked on millions of k-mer binding intensities from 92 TFs across 5 DNA-binding families bHLH, bZIP, ETS, Forkhead, and Homeodomain, as tabulated in Table S1.

The DNA-binding intensity correlation results demonstrate that the proposed approach is robust against different numbers of top k-mers. In particular, it can be scaled and keeps improving with increasing k-mers. The DNA motif pattern recognition results also reveal that it can capture not only the *in vitro* patterns but also the *in vivo* patterns in CIS-BP (v1.02). Last, the DNA motif patterns have been overlapped with the clinically verified SNVs, revealing genome-wide insights into the DNA-binding mechanisms across five DNA-binding families. Thanks to the underlying formulation, the models also have the potential to predict the deleteriousness of unobserved SNVs for the DNA-binding specificities of TFs. It is especially important to uncover unobserved deleterious SNVs as the current studies estimated that, even if we have 500,000 sequenced individuals, we can only observe 12% of all possible SNVs under the protein-coding variant subset (Zou et al., 2016). Therefore, our approaches should be promising based on the first principle in the near future.

As the future works, one may be interested in integrating the existing DNA shape data into the modeling process (Yang et al., 2013). However, it is subject to data availability as well as reliability since the current DNA shape data are mostly computationally predicted (Zhou et al., 2013) and may not be applicable to our DNA-binding specificity studies (Rossi et al., 2017).

Limitation of Study

The current study is limited to five DNA-binding families: bHLH, bZIP, ETS, Forkhead, and Homeodomain because of data availability. In the future, it should be extended to other families such as zinc finger. In addition, the study can be benchmarked with longer k-mers than the current ones. The Markov assumption here can be investigated further under different Markov orders.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Transparent Methods, 14 figures, 1 table, and 1 data file and can be found with this article online at https://doi.org/10.1016/j.isci.2018.09.003.

ACKNOWLEDGMENTS

The author would like to thank the UniPROBE, CIS-BP, and Pfam communities for making their data publicly available. The work described in this article was substantially supported by three grants from the Research Grants Council of the Hong Kong Special Administrative Region [CityU 21200816], [CityU 11203217], and [CityU 11200218]. The donation support of the Titan Xp GPU from the NVIDIA Corporation is appreciated.

AUTHOR CONTRIBUTIONS

K.W. conceived, designed, and implemented the study. K.W. wrote the manuscript.

DECLARATION OF INTERESTS

The author declares no competing interests.

Received: July 5, 2018
Revised: August 8, 2018
Accepted: September 4, 2018
Published: September 28, 2018

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., and McCarthy, S. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.

Barrera, L.A., Vedenko, A., Kurland, J.V., Rogers, J.M., Gisselbrecht, S.S., Rossin, E.J., Woodard, J., Mariani, L., Kock, K.H., Inukai, S., et al. (2016). Survey of variation in human transcription factors reveals prevalent DNA binding changes. Science *351*, 1450–1454.

Benos, P.V., Bulyk, M.L., and Stormo, G.D. (2002). Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Res. *30*, 4442–4451.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266–1276.

Best, D., and Roberts, D. (1975). Algorithm as 89: the upper tail probabilities of spearman's rho. J. R. Stat. Soc. Ser. C Appl. Stat. *24*, 377–379.

Chen, X., Hughes, T.R., and Morris, Q. (2007). RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. Bioinformatics *23*, i72–79.

Dai, H., Umarov, R., Kuwahara, H., Li, Y., Song, L., and Gao, X. (2017). Sequence2Vec: a novel embedding approach for modeling transcription factor binding affinity landscape. Bioinformatics *33*, 3575–3583.

Ellenberger, T. (1994). Getting a grip on DNA recognition: structures of the basic region leucine zipper, and the basic region helix-loop-helix DNA-binding domains. Curr. Opin. Struct. Biol. *4*, 12–21.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44*, D279–D285.

Forrest, A.R., Kawaji, H., Rehli, M., Baillie, J.K., De Hoon, M.J., Haberle, V., Lassmann, T., Kulakovskiy, I.V., Lizio, M., et al. (2014). A promoter-level mammalian expression atlas. Nature *507*, 462.

GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Guo, L., Du, Y., Chang, S., Zhang, K., and Wang, J. (2013). rSNPBase: a database for curated regulatory snps. Nucleic Acids Res. *42*, D1033–D1039.

Gupta, A., Christensen, R.G., Bell, H.A., Goodwin, M., Patel, R.Y., Pandey, M., Enuameh, M.S., Rayla, A.L., Zhu, C., Thibodeau-Beganny, S., et al. (2014). An improved predictive recognition model for cys2-his2 zinc finger proteins. Nucleic Acids Res. *42*, 4800–4812.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Jones, S., van Heyningen, P., Berman, H.M., and Thornton, J.M. (1999). Protein-DNA interactions: a structural analysis. J. Mol. Biol. *287*, 877–896.

Jones, S., Shanahan, H.P., Berman, H.M., and Thornton, J.M. (2003). Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res. *31*, 7189–7198.

Jung, C., Hawkins, J.A., Jones, S.K., Xiao, Y., Rybarski, J.R., Dillard, K.E., Hussmann, J., Saifuddin, F.A., Savran, C.A., Ellington, A.D., et al. (2017). Massively parallel biophysical analysis of CRISPR-Cas complexes on next generation sequencing chips. Cell *170*, 35–47.

Kasinathan, S., Orsi, G.A., Zentner, G.E., Ahmad, K., and Henikoff, S. (2014). High-resolution mapping of transcription factor binding sites on native chromatin. Nat. Methods *11*, 203–209.

Khamis, A.M., Motwalli, O., Oliva, R., Jankovic, B.R., Medvedeva, Y.A., Ashoor, H., Essack, M., Gao, X., and Bajic, V.B. (2018). A novel method for improved accuracy of transcription factor binding site prediction. Nucleic Acids Res. *46*, e72.

Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. *42*, 2976–2987.

Krishna, S.S., Majumdar, I., and Grishin, N.V. (2003). Structural classification of zinc fingers: survey and summary. Nucleic Acids Res. *31*, 532–550.

Kumar, S., Ambrosini, G., and Bucher, P. (2017). SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. Nucleic Acids Res. *45*, D139–D144.

Li, J.J., and Biggin, M.D. (2015). Gene expression. Statistics requantitates the central dogma. Science *347*, 1066–1067.

Li, M.J., Pan, Z., Liu, Z., Wu, J., Wang, P., Zhu, Y., Xu, F., Xia, Z., Sham, P.C., Kocher, J.A., et al. (2016). Predicting regulatory variants with composite statistic. Bioinformatics *32*, 2729–2736.

Luscombe, N.M., and Thornton, J.M. (2002). Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J. Mol. Biol. *320*, 991–1009.

Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. *29*, 2860–2874.

Macintyre, G., Bailey, J., Haviv, I., and Kowalczyk, A. (2010). is-rSNP: a novel technique for in silico regulatory SNP detection. Bioinformatics *26*, i524–530.

Mandel-Gutfreund, Y., and Margalit, H. (1998). Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. Nucleic Acids Res. *26*, 2306–2312.

Mandel-Gutfreund, Y., Schueler, O., and Margalit, H. (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. J. Mol. Biol. *253*, 370–382.

Mathelier, A., and Wasserman, W.W. (2013). The next generation of transcription factor binding site prediction. PLoS Comput. Biol. *9*, e1003214.

Norouzi, M., Fleet, D.J., and Salakhutdinov, R.R. (2012). Hamming distance metric learning. In Advances in Neural Information Processing Systems, pp. 1061–1069, http://papers.nips.cc/paper/4808-hamming-distance-metric-learning.

Paterson, M. and Dančík, V. (1994). Longest Common Subsequences. Mathematical Foundations of Computer Science 1994, pp. 127–142, https://link.springer.com/chapter/10.1007/3-540-58338-6_63.

Pelossof, R., Singh, I., Yang, J.L., Weirauch, M.T., Hughes, T.R., and Leslie, C.S. (2015). Affinity regression predicts the recognition code of nucleic acid-binding proteins. Nat. Biotechnol. *33*, 1242–1249.

Rossi, M.J., Lai, W.K., and Pugh, B.F. (2017). Correspondence: DNA shape is insufficient to explain binding. Nat. Commun. *8*, 15643.

Sarai, A., and Kono, H. (2005). Protein-DNA recognition patterns and predictions. Annu. Rev. Biophys. Biomol. Struct. *34*, 379–398.

Shi, W., Fornes, O., Mathelier, A., and Wasserman, W.W. (2016). Evaluating the impact of single nucleotide variants on transcription factor binding. Nucleic Acids Res. *44*, 10106–10116.

Tan, G., and Lenhard, B. (2016). TFBSTools: an r/bioconductor package for transcription factor binding site analysis. Bioinformatics *32*, 1555–1556.

Tomovic, A., and Oakeley, E.J. (2007). Position dependencies in transcription factor binding sites. Bioinformatics *23*, 933–941.

Vorontsov, I.E., Khimulya, G., Lukianova, E.N., Nikolaeva, D.D., Eliseeva, I.A., Kulakovskiy, I.V., and Makeev, V.J. (2016). Negative selection maintains transcription factor binding motifs in human cancer. BMC Genomics *17*, 395.

Wang, X., Kuwahara, H., and Gao, X. (2014). Modeling DNA affinity landscape through two-round support vector regression with weighted degree kernels. BMC Syst. Biol. *8* (*Suppl 5*), S5.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. *5*, 276–287.

Weirauch, M.T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T.R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., et al. (2013). Evaluation of methods for modeling transcription factor sequence specificity. Nat. Biotechnol. *31*, 126–134.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443.

Wong, K.C. (2015). Computational Methods for Learning and Predicting the DNA Binding Specificities of Transcription Factors, PhD thesis (University of Toronto).

Wong, K.C., Chan, T.M., Peng, C., Li, Y., and Zhang, Z. (2013). DNA motif elucidation using belief propagation. Nucleic Acids Res. *41*, e153.

Wong, K.C., Li, Y., Peng, C., Moses, A.M., and Zhang, Z. (2015). Computational learning on specificity-determining residue-nucleotide interactions. Nucleic Acids Res. *43*, 10180–10189.

Yang, L., Zhou, T., Dror, I., Mathelier, A., Wasserman, W.W., Gordân, R., and Rohs, R. (2013). TFBS shape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Res. *42*, D148–D155.

Yujian, L., and Bo, L. (2007). A normalized levenshtein distance metric. IEEE Trans. Pattern Anal. Mach. Intell. *29*, 1091–1095.

Zeng, H., Hashimoto, T., Kang, D.D., and Gifford, D.K. (2016). GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. Bioinformatics *32*, 490–496.

Zhao, Y., and Stormo, G.D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nat. Biotechnol. *29*, 480–483.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R., and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic Acids Res. *41*, W56–W62.

Zou, J., Valiant, G., Valiant, P., Karczewski, K., Chan, S.O., Samocha, K., Lek, M., Sunyaev, S., Daly, M., and MacArthur, D. (2016). Quantifying the unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. Nat. Commun. *7*, 13293.

Zuo, C., Shin, S., and Keleş, S. (2015). atSNP: transcription factor binding affinity testing for regulatory SNP detection. Bioinformatics *31*, 3353–3355.

ISCI, Volume *7*

Supplemental Information

DNA Motif Recognition Modeling from Protein Sequences

Ka-Chun Wong

S1 Supplementary Information

S1.1 Transparent Methods

The computer programs for generating the draft k-spectrum profile are implemented in JAVA. The computer programs for the Markov random field modeling are implemented in R. The library and package dependencies are explicitly stated in the programs. The source code and data are freely available for open reproducibility at Supplemental Data or `http://bioinfo.cs.cityu.edu.hk/MotifMRF.htm`.

S1.2 Supplemental Data

Supplemental Data is available online.

Data S1: Source code and example data, Related to Figure 1.
The programs for k-spectrum profiling are implemented in JAVA while the programs for Markov random field modeling are implemented in R.

S1.3 Supplemental Figures and Tables

Supplemental Figures and Tables are listed below.

Family	UniPROBE ID	Protein Name	Family	UniPROBE ID	Protein Name
bHLH	UP00050	BHLHB2	ETS	UP00418	ETV6
bHLH	UP00324	TYE7	ETS	UP00419	SPIC
bHLH	UP00332	PHO4	ETS	UP00420	ELK3
bHLH	UP00356	RTG3	ETS	UP00421	ETV1
bHLH	UP00357	HLH-11	ETS	UP00422	ETV3
bHLH	UP00358	HLH-10	ETS	UP01363	ELF1
bHLH	UP00364	HLH-27	Forkhead	UP00025	FOXK1
bHLH	UP00365	HLH-30	Forkhead	UP00039	FOXJ3
bHLH	UP00367	MXL-3	Forkhead	UP00041	FOXJ1
bHLH	UP00368	HLH-29	Forkhead	UP00061	FOXL1
bHLH	UP00370	HLH-26	Forkhead	UP00073	FOXA2
bHLH	UP00379	HLH-3	Forkhead	UP00303	FHL1
bHLH	UP00383	REF-1	Forkhead	UP00312	FKH2
bHLH	UP00384	HLH-2	Forkhead	UP00353	FKH1
bHLH	UP00386	HLH-25	Forkhead	UP00521	FOXN2
bHLH	UP00387	HLH-1	Forkhead	UP00522	FOXN4
bZIP	UP00020	ATF1	Forkhead	UP00523	FOXR1
bZIP	UP00285	GCN4	Forkhead	UP00526	FOXN1
bZIP	UP00316	YAP6	Forkhead	UP00528	FOXM1
bZIP	UP00327	YAP1	Forkhead	UP00589	FOXC1
bZIP	UP00426	JUN	Forkhead	UP01365	FOXA3
bZIP	UP00453	CAD1	Forkhead	UP01366	FOXB1
bZIP	UP00454	CIN5	Forkhead	UP01367	FOXC2
bZIP	UP00455	CST6	Forkhead	UP01368	FOXG1
bZIP	UP00457	HAC1	Forkhead	UP01369	FOXJ2
bZIP	UP00464	SKO1	Forkhead	UP01371	FOXO3
bZIP	UP00473	YAP3	Homeodomain	UP00584	ARX
bZIP	UP01354	ATF3	Homeodomain	UP00586	CRX
bZIP	UP01356	CEBPA	Homeodomain	UP00588	ESX1
bZIP	UP01357	CEBPB	Homeodomain	UP00594	HESX1
bZIP	UP01359	DBP	Homeodomain	UP00595	HOXB7
bZIP	UP01402	TEF	Homeodomain	UP00596	HOXC4
bZIP	UP01403	XBP1	Homeodomain	UP00597	HOXD13
ETS	UP00015	EHF	Homeodomain	UP00598	ISX
ETS	UP00038	SPDEF	Homeodomain	UP00603	MSX2
ETS	UP00085	SFPI1	Homeodomain	UP00604	NKX2-5
ETS	UP00090	ELF3	Homeodomain	UP00605	NKX2-8
ETS	UP00404	ELF2	Homeodomain	UP00613	PBX4
ETS	UP00409	ELF5	Homeodomain	UP00614	PHOX2B
ETS	UP00410	ELK1	Homeodomain	UP00615	PITX2
ETS	UP00411	ERG	Homeodomain	UP00619	PROP1
ETS	UP00412	ETV5	Homeodomain	UP00620	SIX6
ETS	UP00413	ELF4	Homeodomain	UP00623	VAX2
ETS	UP00414	ETS1	Homeodomain	UP00624	VENTX
ETS	UP00416	FLI1	Homeodomain	UP00625	VSX1
ETS	UP00417	ETV4	Homeodomain	UP00626	VSX2

Table S1 List of Protein Binding Microarray (PBM) data as available and retrieved from UniPROBE in Oct 2017, related to Figure 1.
The family column indicates the DNA-binding domain families.

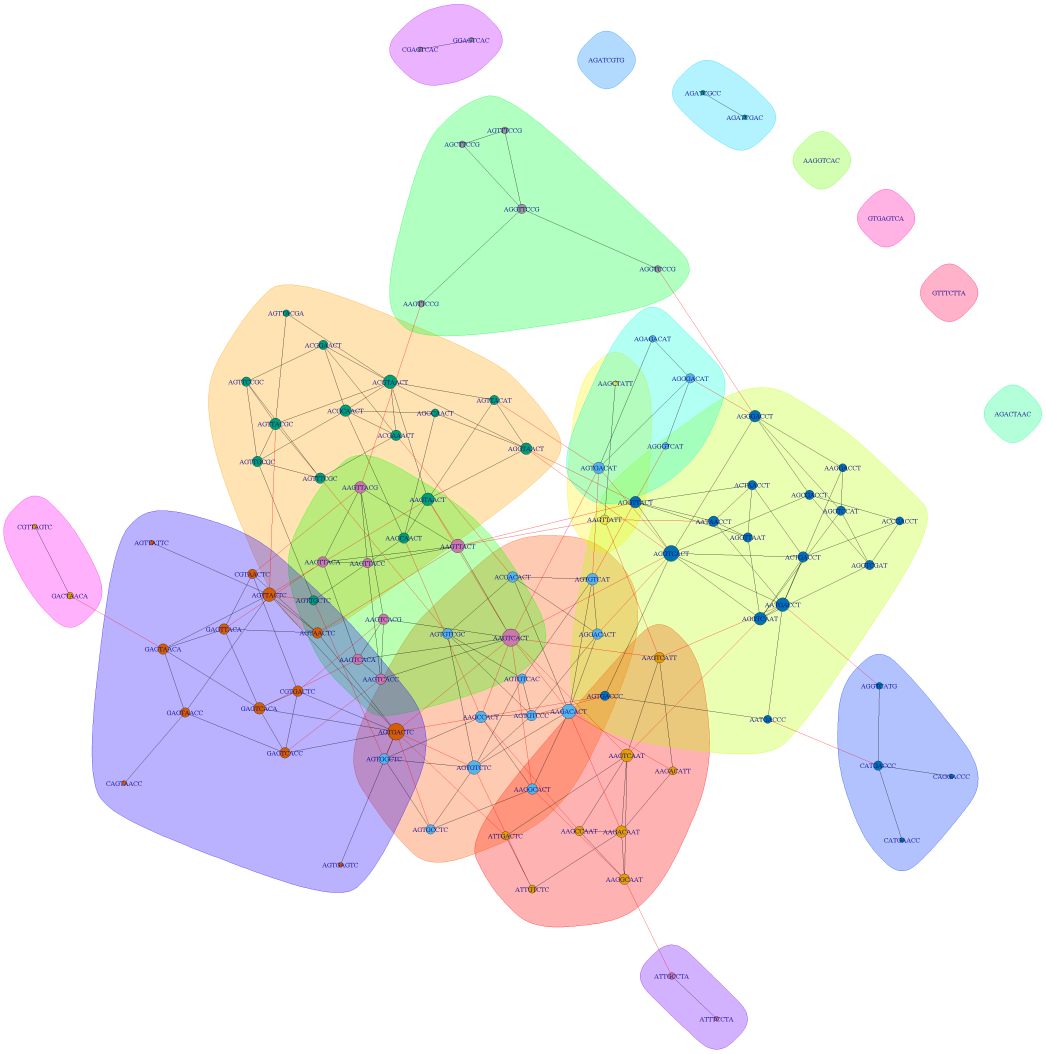


Figure S1 Markov random field network of the hidden variables for the bZIP family when 100 top k-mers are selected (M) for modeling, related to Figure 4

Hamming distance threshold of one is adopted for sequence neighborhood connections while reverse complements are considered equivalent. The edge betweenness community detection method has been adopted to segregate the Markov network into different communities for modularity maximization (i.e. 'cluster_edge_betweenness' function in R) as highlighted in different colours. The node sizes are proportional to the node degrees.

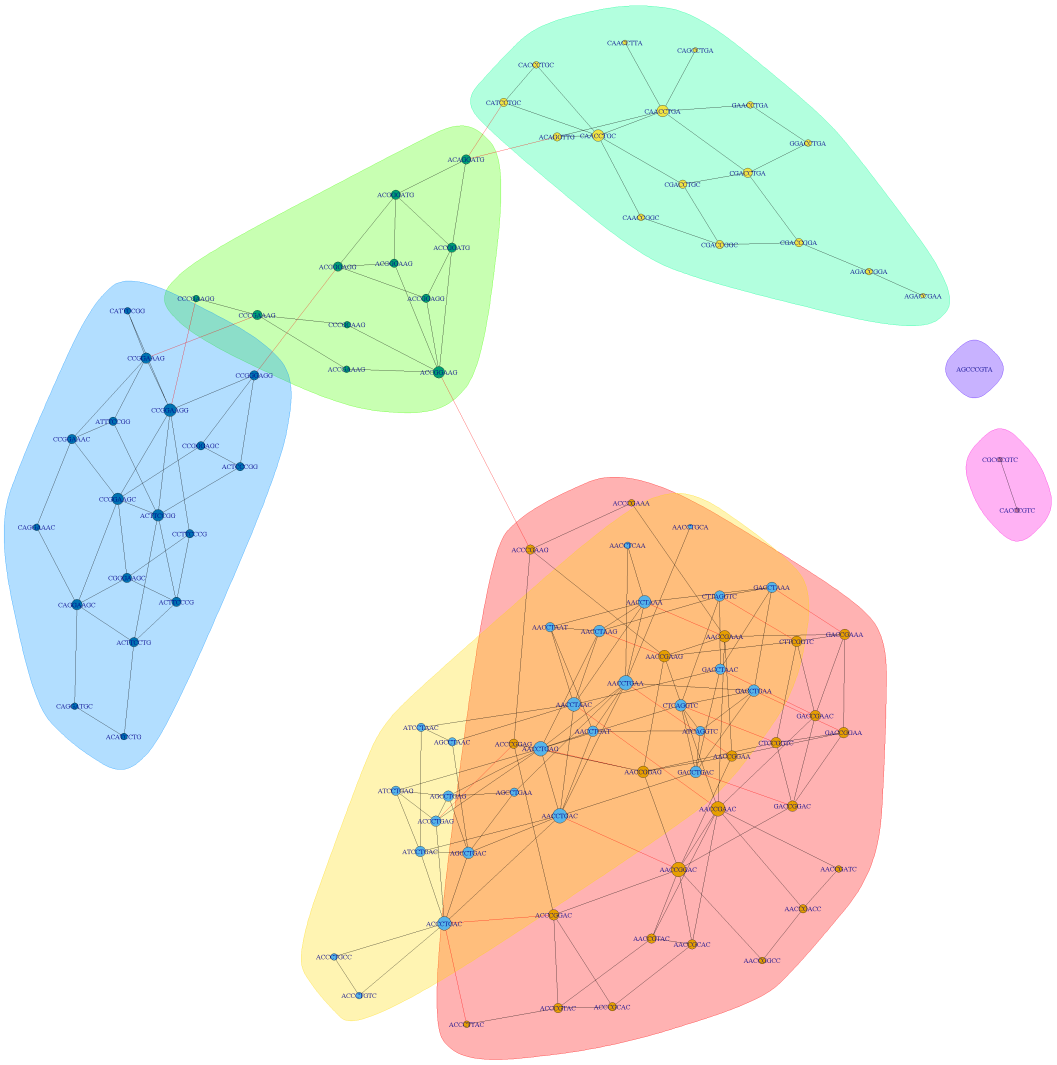


Figure S2 Markov random field network of the hidden variables for the Ets family when 100 top k-mers are selected (M) for modeling, related to Figure 4

Hamming distance threshold of one is adopted for sequence neighborhood connections while reverse complements are considered equivalent. The edge betweenness community detection method has been adopted to segregate the Markov network into different communities for modularity maximization (i.e. 'cluster_edge_betweenness' function in R) as highlighted in different colours. The node sizes are proportional to the node degrees.

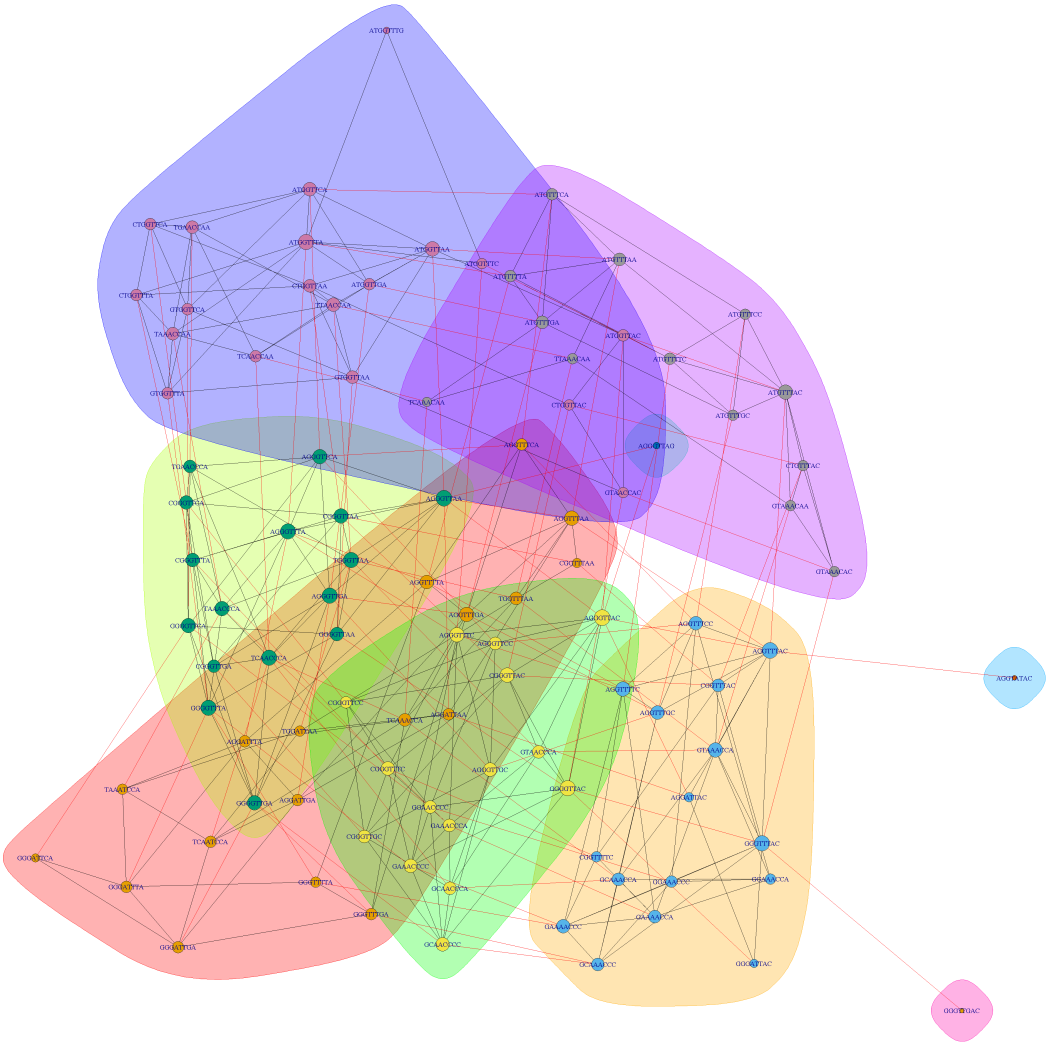


Figure S3 Markov random field network of the hidden variables for the Forkhead family when 100 top k-mers are selected (M) for modeling, related to Figure 4

Hamming distance threshold of one is adopted for sequence neighborhood connections while reverse complements are considered equivalent. The edge betweenness community detection method has been adopted to segregate the Markov network into different communities for modularity maximization (i.e. 'cluster_edge_betweenness' function in R) as highlighted in different colours. The node sizes are proportional to the node degrees.

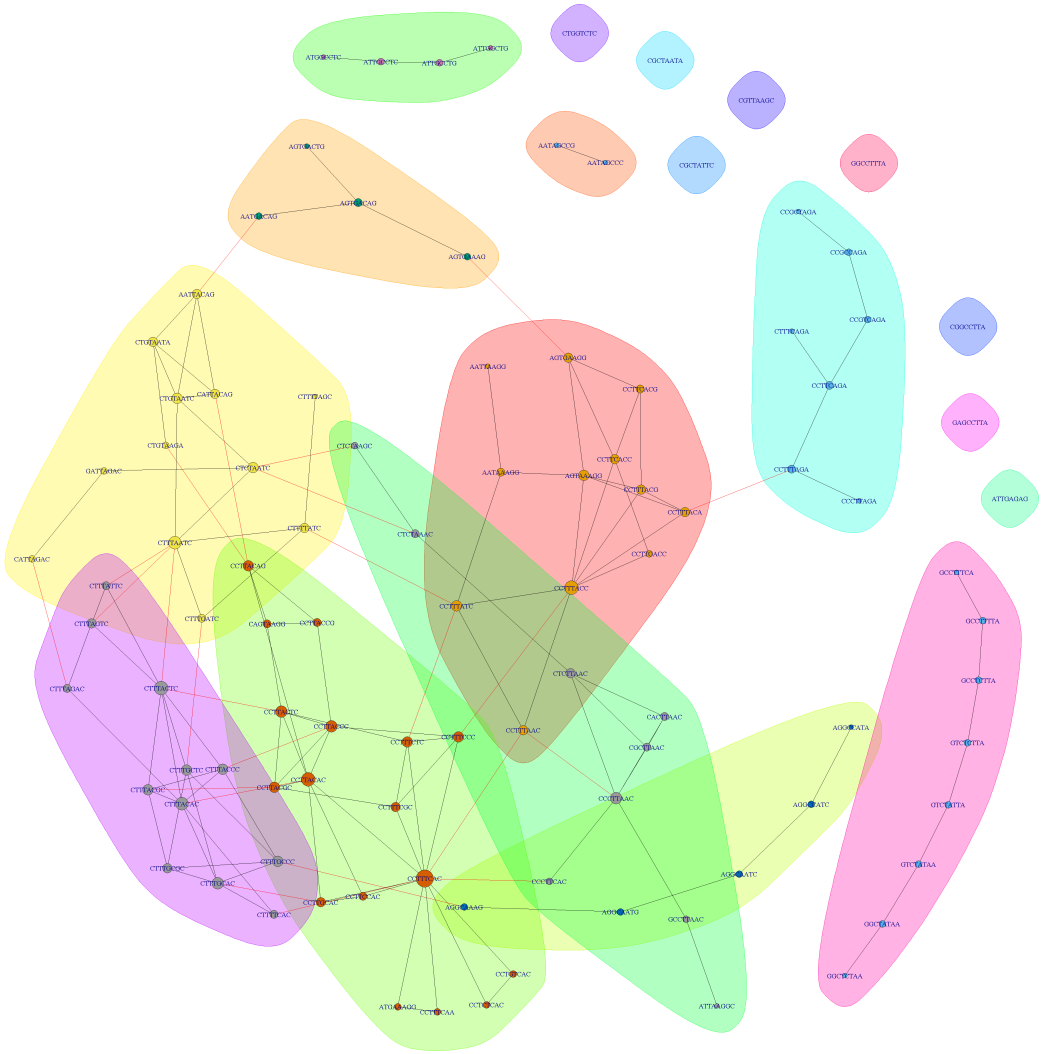


Figure S4 Markov random field network of the hidden variables for the Homeodomain family when 100 top k-mers are selected (M) for modeling, related to Figure 4

Hamming distance threshold of one is adopted for sequence neighborhood connections while reverse complements are considered equivalent. The edge betweenness community detection method has been adopted to segregate the Markov network into different communities for modularity maximization (i.e. 'cluster_edge_betweenness' function in R) as highlighted in different colours. The node sizes are proportional to the node degrees.

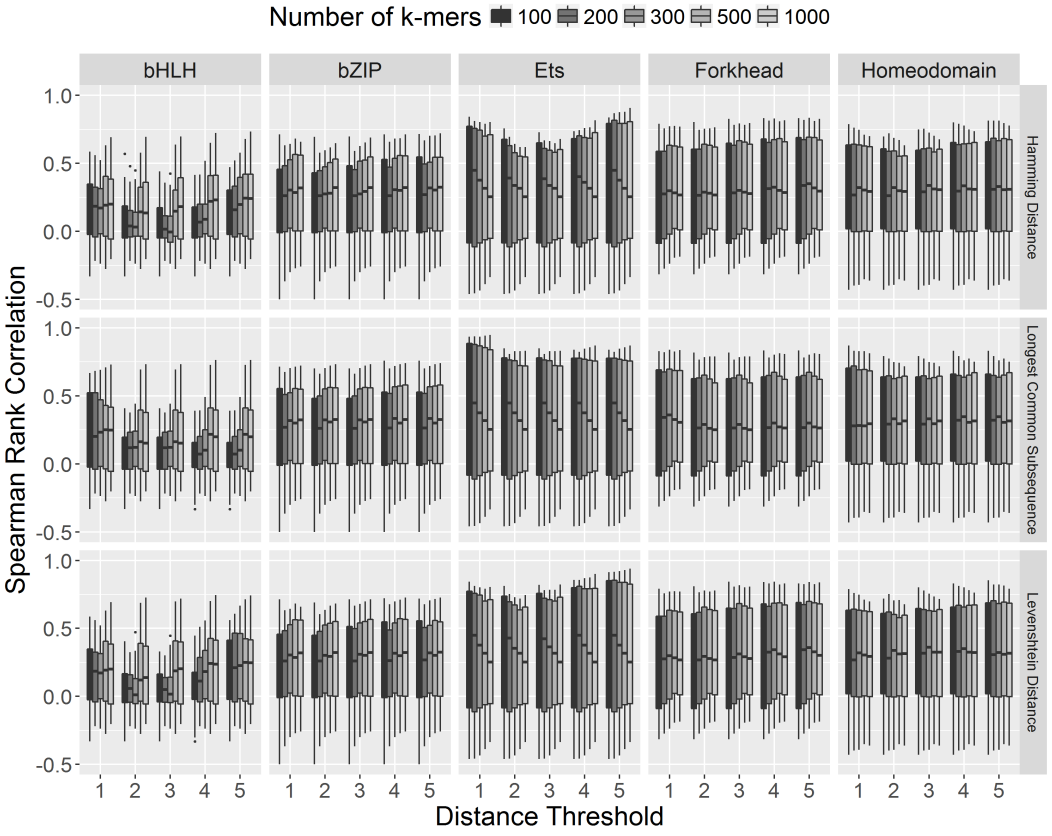
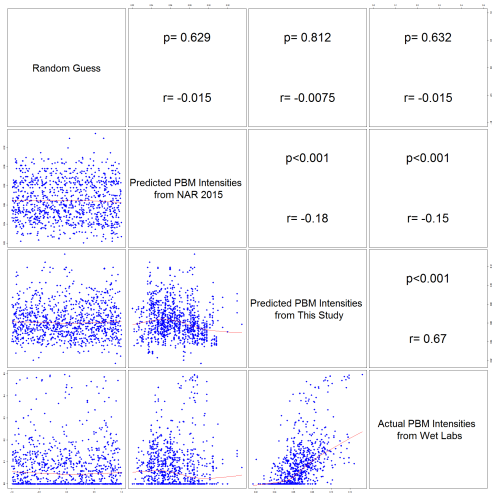
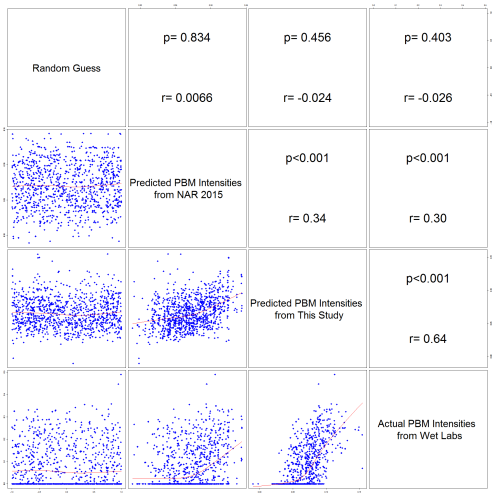


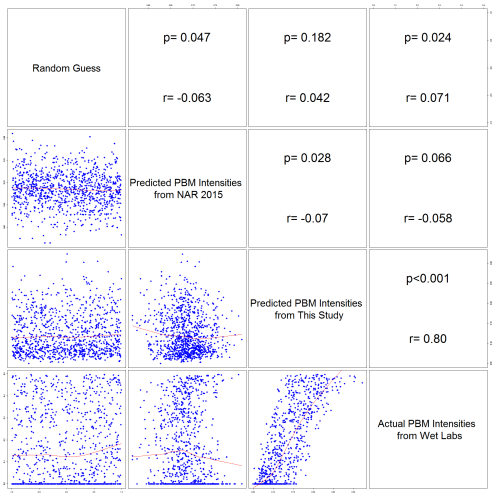
Figure S5 Box plots on the Spearman rank correlations between the actual binding intensities of k-mers and the predicted binding intensities of k-mers against different number of k-mers using the current method, related to Figure 2



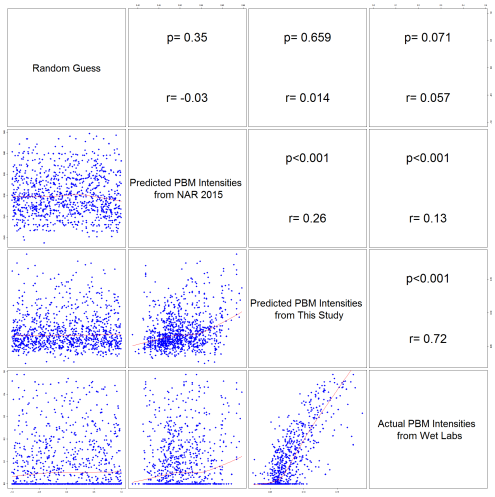
(a) HLH-25 from bHLH Family



(b) CST6 from bZIP Family



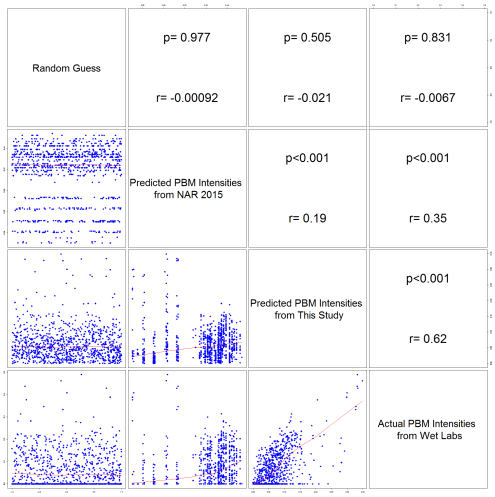
(c) ERG from ETS Family



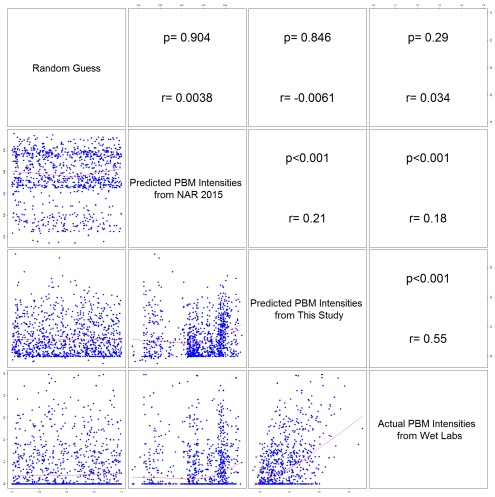
(d) FOXJ1 from Forkhead Family

Figure S6 Examples of left-one-out cross-validation predictions on 1000 top k-mers which neighborhood is defined based on the Hamming distance threshold of one, related to Figure 5

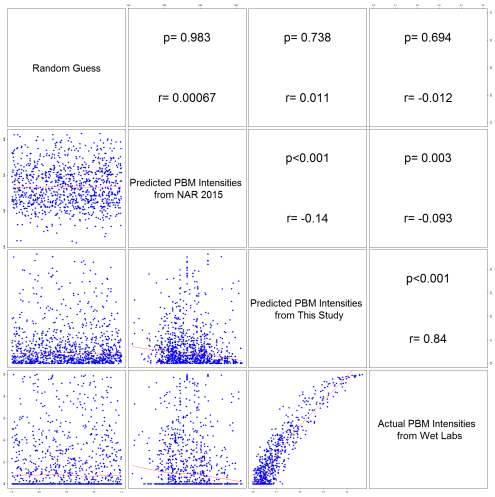
Each dot represents one k-mer. The evaluations are based on Spearman rank correlations (r). Random initialization shows the initial guess before the iterations. The figures are drawn using R where the red curves are local polynomial regression fittings with $\alpha = 2/3$ and the p-values are computed using algorithm AS 89 (Best and Roberts, 1975).



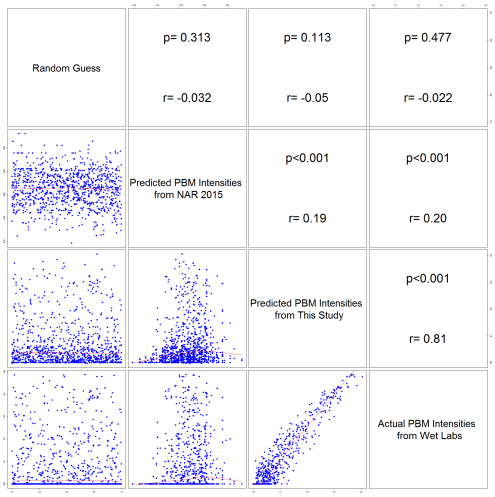
(a) HLH-26 from bHLH Family



(b) JUN from bZIP Family



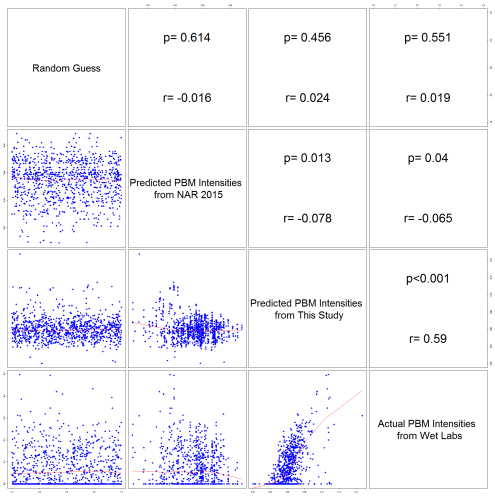
(c) FOXG1 from Forkhead Family



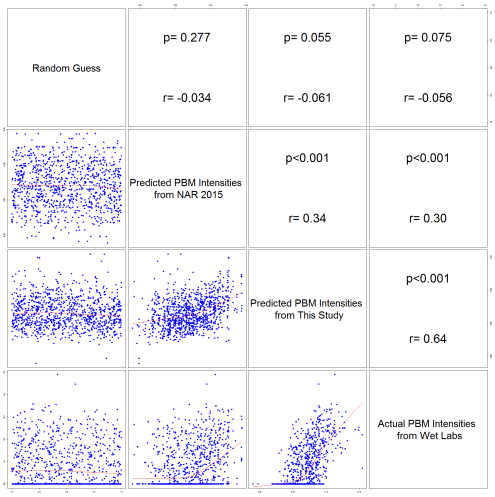
(d) VSX2 from Homeodomain Family

Figure S7 Examples of left-one-out cross-validation predictions on 1000 top k-mers which neighborhood is defined based on the longest common subsequence (LCS) distance threshold of one, related to Figure 5

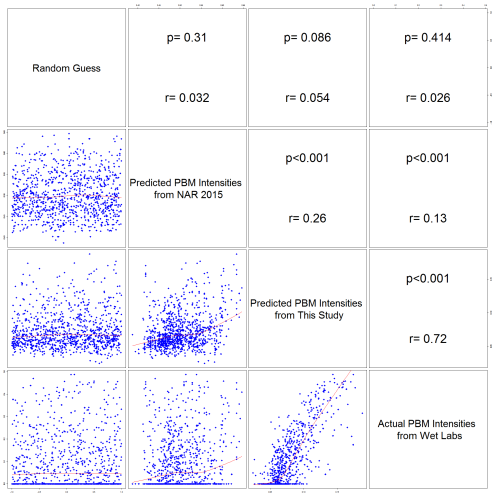
Each dot represents one k-mer. The evaluations are based on Spearman rank correlations (r). Random initialization shows the initial guess before the iterations. The figures are drawn using R where the red curves are local polynomial regression fittings with $\alpha = 2/3$ and the p-values are computed using algorithm AS 89 (Best and Roberts, 1975).



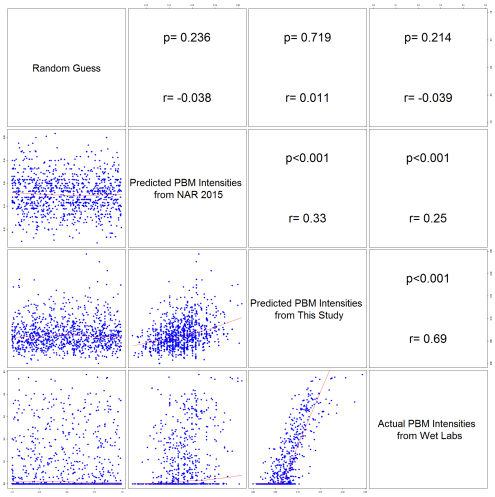
(a) HLH-3 from bHLH Family



(b) CST6 from bZIP Family



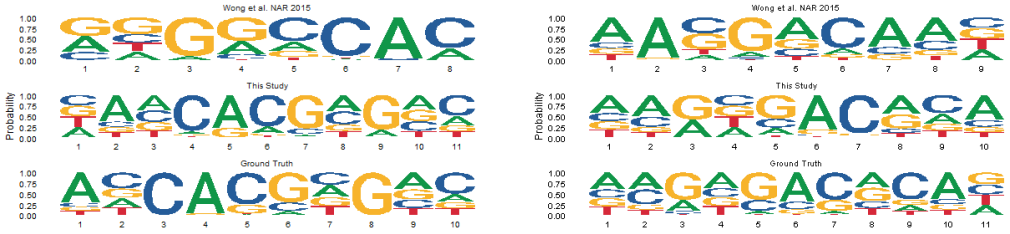
(c) FOXJ1 from Forkhead Family



(d) MSX2 from Homeodomain Family

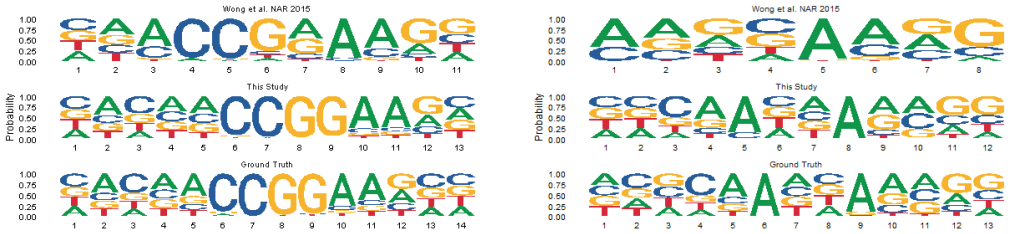
Figure S8 Examples of left-one-out cross-validation predictions on 1000 top k-mers which neighborhood is defined based on the Levenshtein distance threshold of one, related to Figure 5

Each dot represents one k-mer. The evaluations are based on Spearman rank correlations (r). Random initialization shows the initial guess before the iterations. The figures are drawn using R where the red curves are local polynomial regression fittings with $\alpha = 2/3$ and the p-values are computed using algorithm AS 89 (Best and Roberts, 1975).



(a) TYE7 from bHLH Family

(b) XBP1 from bZIP Family

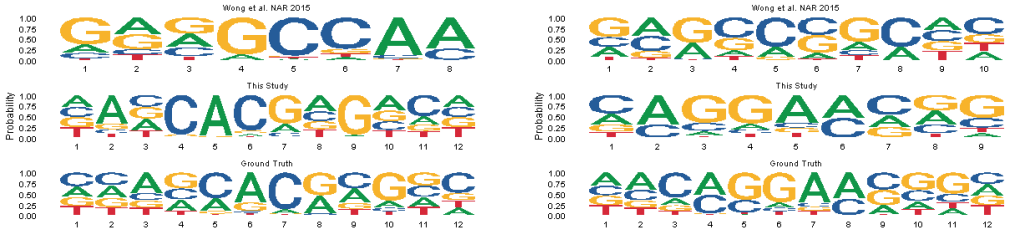


(c) FLI1 from ETS Family

(d) MSX2 from Homeodomain Family

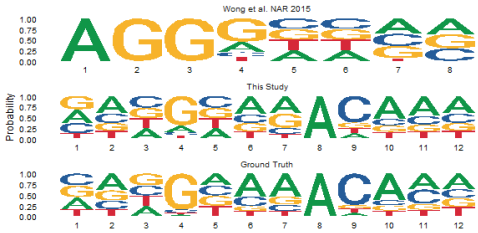
Figure S9 Examples of DNA motif matrices generated by the previous approach (Wong et al., 2015), our approach, and the actual DNA motif matrices as measured using PBM (Robasky and Bulyk, 2011), related to Figure 7

The first two motif matrices are based on the left-one-out cross-validation predictions on 1000 top k-mers which neighborhood is defined based on the LCS distance threshold of one. All settings strictly follow the protocols established by Zhao and Stormo (Zhao and Stormo, 2011).

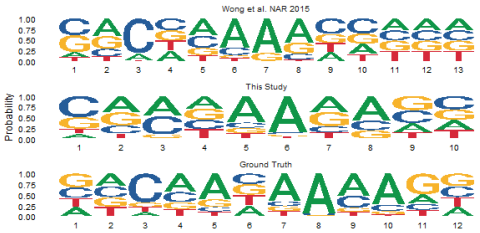


(a) HLH-29 from bHLH Family

(b) EHF from ETS Family



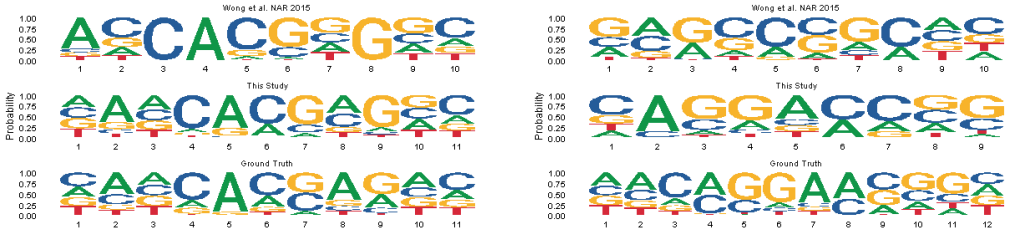
(c) FOXK1 from Forkhead Family



(d) MSX2 from Homeodomain Family

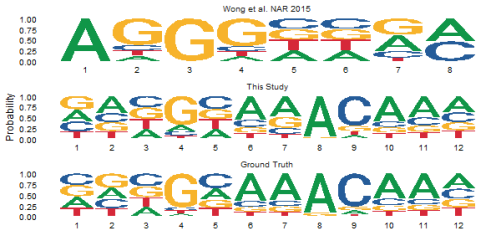
Figure S10 Examples of DNA motif matrices generated by the previous approach (Wong et al., 2015), our approach, and the actual DNA motif matrices as measured using PBM (Robasky and Bulyk, 2011), related to Figure 7

The first two motif matrices are based on the left-one-out cross-validation predictions on 1000 top k-mers which neighborhood is defined based on the Levenshtein distance threshold of one. All settings strictly follow the protocols established by Zhao and Stormo (Zhao and Stormo, 2011).

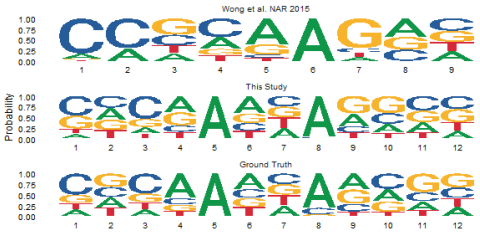


(a) RTG3 from bHLH Family

(b) EHF from ETS Family



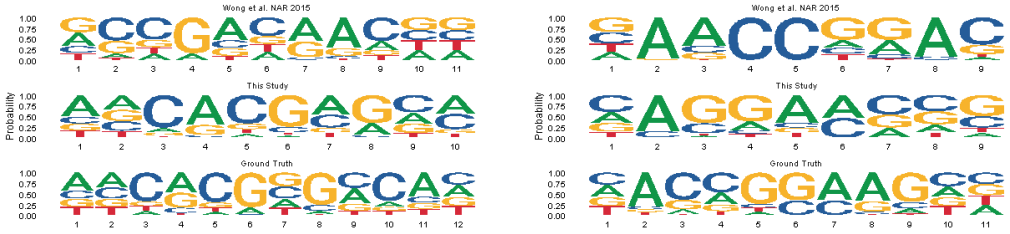
(c) FOXN4 from Forkhead Family



(d) PROP1 from Homeodomain Family

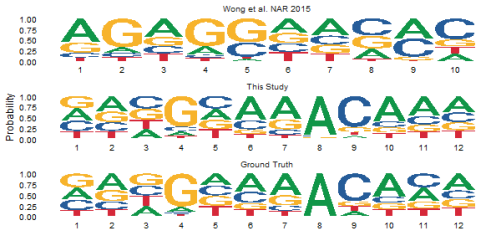
Figure S11 Examples of DNA motif matrices generated by the previous approach (Wong et al., 2015), our approach, and the actual DNA motif matrices as measured using PBM (Robasky and Bulyk, 2011), related to Figure 7

The first two motif matrices are based on the left-one-out cross-validation predictions on 1000 top k-mers which neighborhood is defined based on the hamming distance threshold of two. All settings strictly follow the protocols established by Zhao and Stormo (Zhao and Stormo, 2011).

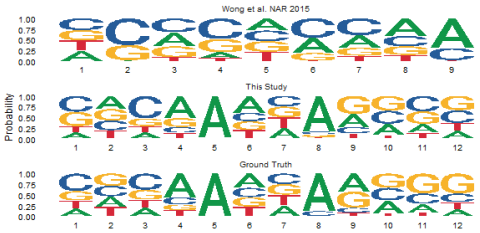


(a) RTG3 from bHLH Family

(b) SPIC from ETS Family



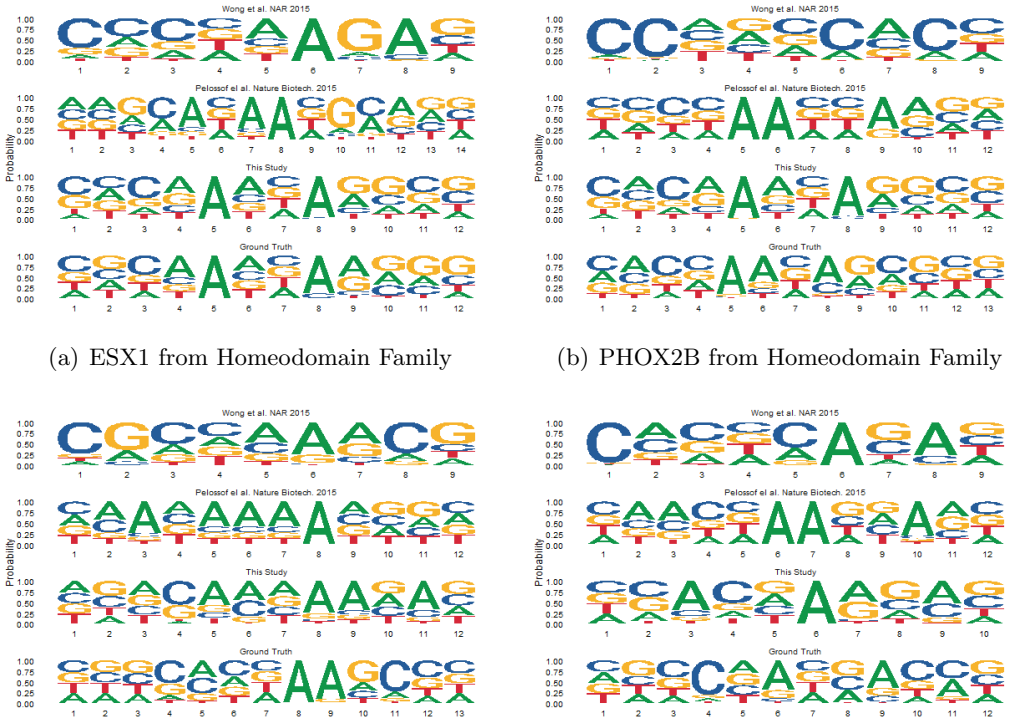
(c) FKH2 from Forkhead Family



(d) ARX from Homeodomain Family

Figure S12 Examples of DNA motif matrices generated by the previous approach (Wong et al., 2015), our approach, and the actual DNA motif matrices as measured using PBM (Robasky and Bulyk, 2011), related to Figure 7

The first two motif matrices are based on the left-one-out cross-validation predictions on 1000 top k-mers which neighborhood is defined based on the hamming distance threshold of three. All settings strictly follow the protocols established by Zhao and Stormo (Zhao and Stormo, 2011).



(a) ESX1 from Homeodomain Family

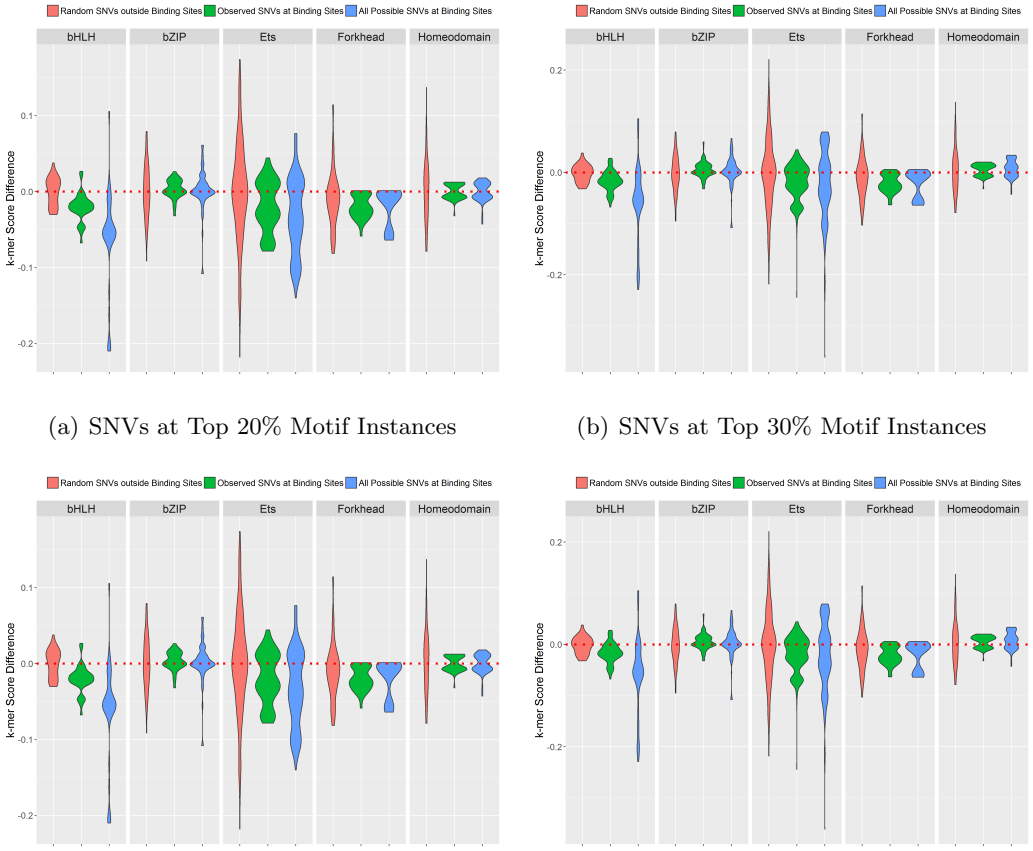
(b) PHOX2B from Homeodomain Family

(c) PITX2 from Homeodomain Family

(d) SIX6 from Homeodomain Family

Figure S13 Examples of Homeodomain DNA motif matrices generated by the previous approach (Wong et al., 2015), Pelossof method (Pelossof et al., 2015), our approach, and the actual DNA motif matrices as measured using PBM (Robasky and Bulyk, 2011), related to Figure 7

The first and third motif matrices are based on the 1000 top k-mers which neighborhood is defined based on the Hamming distance threshold of two. The second motif matrix is generated based on (Pelossof et al., 2015). All settings strictly follow the protocols established by Zhao and Stormo (Zhao and Stormo, 2011).



(a) SNVs at Top 20% Motif Instances

(b) SNVs at Top 30% Motif Instances

(c) SNVs at Top 50% Motif Instances

(d) SNVs at All Motif Instances

Figure S14 Violin plots on the k-mer score difference distributions of the SNVs at top-ranked motif instances via family-specific recognition modeling, related to Figure 9

The observed SNVs are retrieved from the clinically verified dataset, ClinVar (version 20171029), while the DNA motif instances are ranked by TFBStools (Tan and Lenhard, 2016).