5666–5677 Nucleic Acids Research, 2017, Vol. 45, No. 10 Published online 3 May 2017
doi: 10.1093/nar/gkx358

Predicting transcription factor binding motifs from DNA-binding domains, chromatin accessibility and gene expression data

Mahdi Zamanighomi[1], Zhixiang Lin[1], Yong Wang[2], Rui Jiang[3] and Wing Hung Wong[1,4,*]

[1]Department of Statistics, Stanford University, Stanford, CA 94305, USA, [2]Academy of Mathematics and Systems Science, National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, China, [3]MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST, Department of Automation, Tsinghua University, Beijing 100084, China and [4]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA

Received August 25, 2016; Revised April 19, 2017; Editorial Decision April 20, 2017; Accepted April 20, 2017

ABSTRACT

Transcription factors (TFs) play crucial roles in regulating gene expression through interactions with specific DNA sequences. Recently, the sequence motif of almost 400 human TFs have been identified using high-throughput SELEX sequencing. However, there remain a large number of TFs (∼800) with no high-throughput-derived binding motifs. Computational methods capable of associating known motifs to such TFs will avoid tremendous experimental efforts and enable deeper understanding of transcriptional regulatory functions. We present a method to associate known motifs to TFs (MATLAB code is available in Supplementary Materials). Our method is based on a probabilistic framework that not only exploits DNA-binding domains and specificities, but also integrates open chromatin, gene expression and genomic data to accurately infer monomeric and homodimeric binding motifs. Our analysis resulted in the assignment of motifs to 200 TFs with no SELEX-derived motifs, roughly a 50% increase compared to the existing coverage.

INTRODUCTION

A central challenge in current biology is to elucidate transcriptional regulatory mechanisms that influence animal growth and development. Experimental techniques determining target genes of transcription factors (TFs) have led to well characterized transcriptional networks in both low complexity organisms (1–3) and mammals (4–7). Although such approaches continue to provide valuable knowledge, they often demand time-consuming and costly strategies that are limited to a very modest subset of TFs and narrowly focused on particular cell types.

The chromatin immunoprecipitation (ChIP) coupled with DNA sequencing has recently become a powerful method for identifying TF–DNA interactions in mammalian genomes (8,9). However, given the diversity of cell types, environmental conditions, and TFs, it is not feasible for ChIP-seq assays to cover all cellular contexts.

Since TFs typically bind to DNA at sites matching specific sequence motifs, knowledge of the motif for a TF will be useful in determining the potential binding sites of the TF. Of course, the accurate inference of the binding sites in a particular cellular context will also require context dependent experimental data such as chromatin accessible regions (10,11). In any case, knowledge of the TF motif is essential.

In a recent study by the Taipale lab (12), called Taipale hereafter, high-throughput SELEX and ChIP sequencing was employed to analyze sequence preferences of human/mouse TFs. They acquired a total of 843 high-resolution motifs expressed as position weight matrices (PWMs). Taipale analysis identified PWMs that are 13 bp long on average and also recovered numerous homodimers for different structural TF families. These results significantly improved knowledge of human TF motifs compared to existing studies (13–15).

On the other hand, there are still many TFs with unknown PWMs. In fact, Universal Protein Resource (UniProt) has annotated more than 1100 DNA-binding TFs (16). Excluding TFs that possess Taipale PWMs, we arrive at approximately 800 human TFs without experimentally determined motifs. The lack of motif information presents a substantial obstacle in the understanding of the regulatory roles of these TFs.

Existing methods to predict TF motifs in the absence of TF–DNA binding data are mostly based on protein sequences (17–19). They concentrate on amino acid sequences with annotated DNA-binding domains (DBDs) and introduce various features originated from DBDs. Dataset com-

[*]To whom correspondence should be addressed. Tel: +1 650 725 2915; Fax: +1 650 725 8977; Email: whwong@stanford.edu

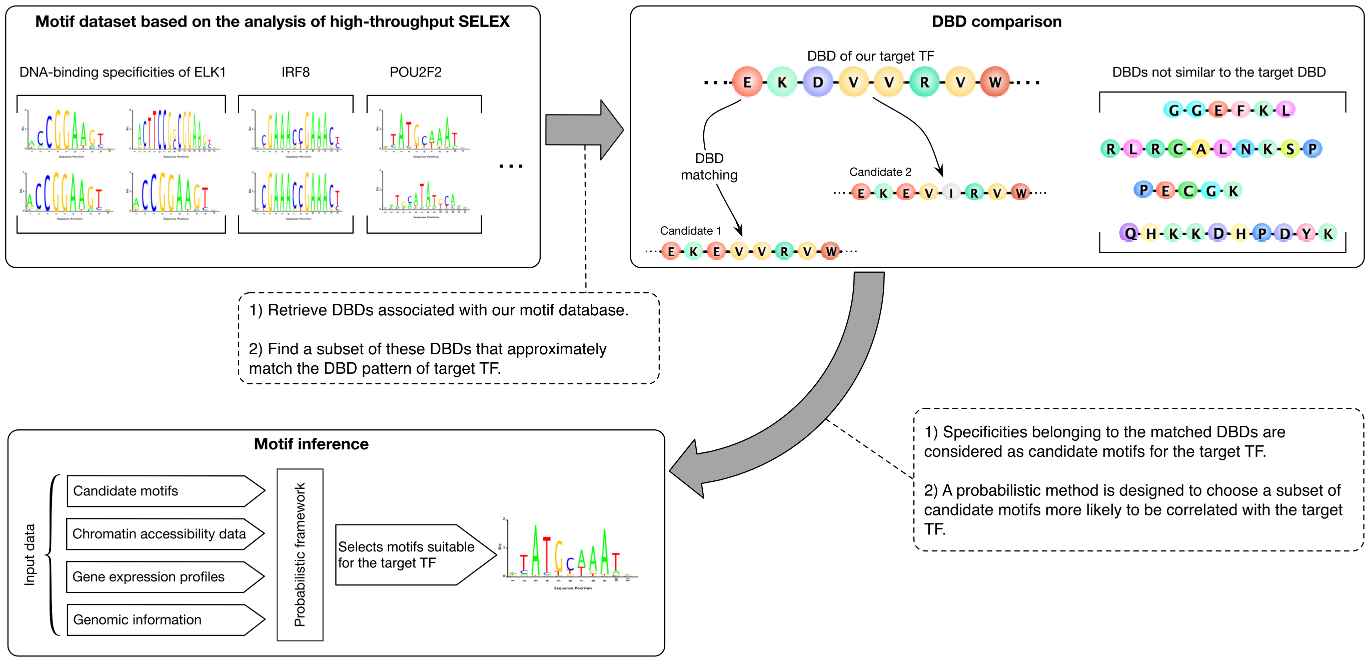
Motif dataset based on the analysis of high-throughput SELEX
DNA-binding specificities of ELK1
IRF8
POU2F2
...
DBD comparison
DBD of our target TF
E K D V V R V W
DBD matching
Candidate 2
E K E V I R V W
Candidate 1
E K E V V R V W
DBDs not similar to the target DBD
G G E F K L
R L R C A L N K S P
P E C G K
Q H K K D H P D Y K
1) Retrieve DBDs associated with our motif database.
2) Find a subset of these DBDs that approximately match the DBD pattern of target TF.
1) Specificities belonging to the matched DBDs are considered as candidate motifs for the target TF.
2) A probabilistic method is designed to choose a subset of candidate motifs more likely to be correlated with the target TF.
Motif inference
Input data
Candidate motifs
Chromatin accessibility data
Gene expression profiles
Genomic information
Probabilistic framework
Selects motifs suitable for the target TF

Figure 1. Overview of DNA-binding specificities inference, beginning with collecting human DBDs and their position weight matrices. The next stage involves mapping TTF to a set of TFs that follow similar DBD patterns as TTF. Finally, our algorithm takes experimental data into account and select candidates describing the best specificities for TTF.

prising TFs/DBDs coupled to PWMs are then used to train features and predict DNA-binding specificities of target TFs. In this work, we show that DBD-based algorithms do not always predict an accurate motif, which suggests a need to improve motif inference by leveraging new experimental data. We develop a pipeline consisting of two steps: (i) based on DBD similarity, we map a target TF (TTF) to a set of Taipale motifs; and (ii) we construct a probabilistic procedure that combines RNA-seq and DNase-seq platforms to select suitable motifs from candidates obtained in the previous step. The proposed approach incorporates high-throughput data across diverse tissue types, takes advantage of genomic information, and reduces our inference algorithm into an optimization problem that can be quickly solved. Our method is named MPAE, which stands for Motif Prediction based on Accessibility and Expression data.

OVERVIEW OF METHODS

A graphical overview of our method is shown in Figure 1. First, we consider a set of DBDs whose DNA-binding specificities are experimentally determined. Next, for any TTF without known motif, we use a DBD-based approach to map our TF of interest to a set of candidate motifs. Finally we use a statistical method, based on gene expression and chromatin accessibility data across a diverse set of cellular contexts, to select a small number of the candidate motifs ($\leq$3) for association to the TTF. In this section, we present an overview for the proposed methods and illustrate their strengths and weaknesses. A more systematic assessment of methods will be described in the 'Results' section.

Motif inference based on DBD similarity

To predict sequence-specific binding motifs for a TF, we implement a DBD similarity-based approach (DBDSim), which applies an agglomerative hierarchical cluster tree to determine a set of TFs that have DBDs analogous to TTF. In particular, we first construct a library consisting of Taipale TFs (12) and retrieve their DBD sequences from Uniprot (21) (Supplementary Table S1). DBDSim calculates pairwise distances among the DBD sequences to create a hierarchical tree. We next build TF clusters based on the tree and measure DBD similarity between each cluster and TTF, named DBDSim score (see 'Materials and Methods' section). We consider motifs associated with clusters that have the highest DBDSim scores as candidate PWMs for our TF of interest.

DBDSim typically yields a modest number of candidate motifs (on average 14) and if the DBD similarity is significant enough, the candidate set is highly likely to contain a good motif for the TTF. However, not all motifs in the candidate sets are good motifs. For example when targeting TF PAX5, the left column in Figure 2A represents motifs identified by DBDSim, each having a very high DBDSim score. However, the second cluster illustrates a significant dissimilarity between PAX3 and PAX7 motifs and the ground truth. Moreover, TF–DNA binding specificities in one cluster can show different structures, leading to the question whether all of them are appropriate predictions? For instance in Figure 2B, MEOX2 is associated with a monomer and two dimeric motifs representing three distinct binding candidates for MEOX1. Despite the significant DBD similarity between MEOX1 and MEOX2, only the monomer is correct motif for MEOX1. Here, it might be suggested

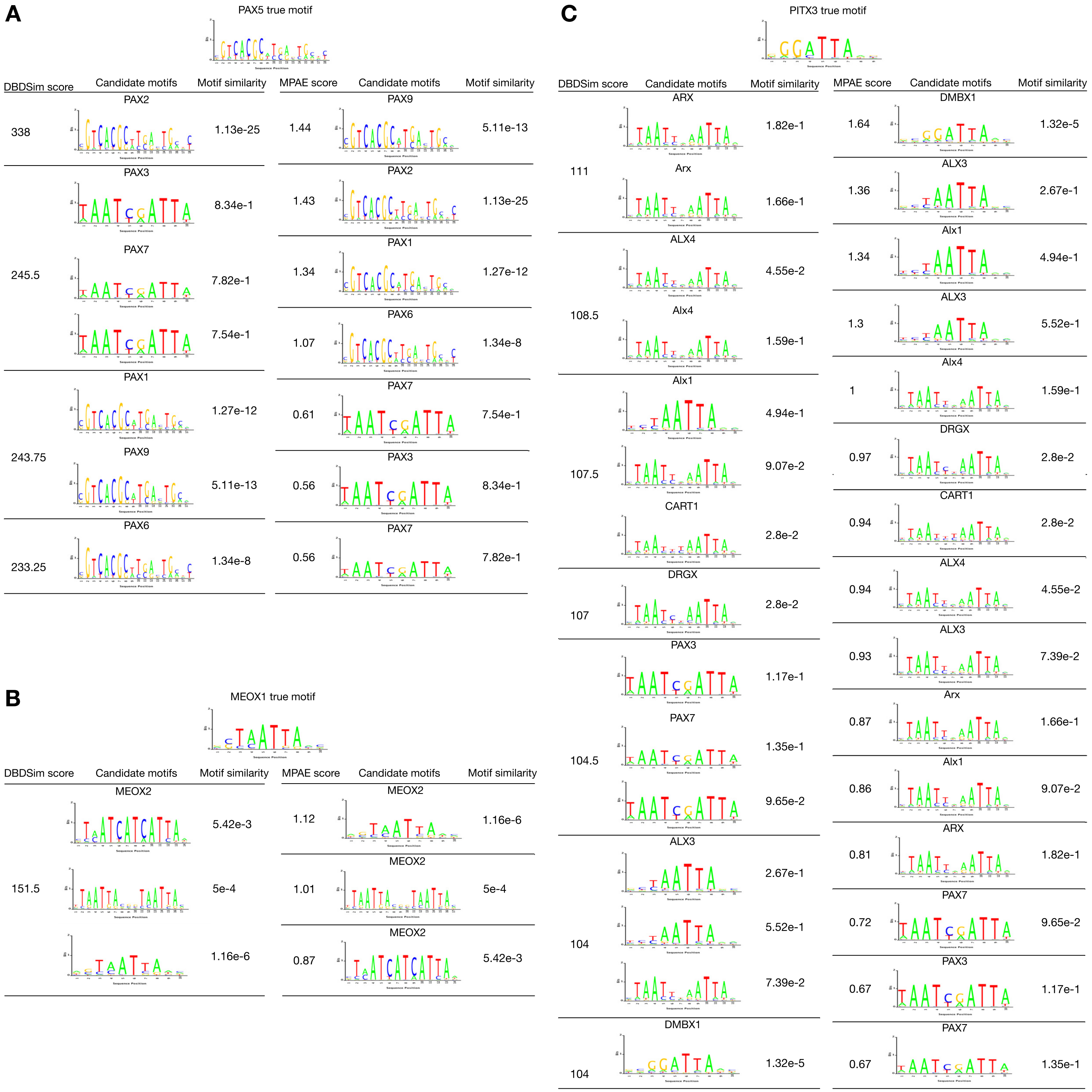

Figure 2. Motif prediction performance. Left panels in **A**–**C** illustrate DBDSim performance and right panels correspond to MPAE method. PAX5, PITX3 and MEOX1 specificities, labeled as true motif, are removed from Taipale dataset. In the first step, we apply DBDSim to recover TFs with significant DBD similarities to the TTFs and ranked them based on DBD scores. The highest score in the left columns indicates the best DBD match to the TF of interest. We compare the PWMs with the true motifs using TOMTOM tools (20), an ungapped alignment scheme that statistically measures similarities between pairs of motifs. TOMTOM motif comparison function is set to the Pearson correlation coefficient and outcomes are reported as *P*-values in Motif similarity columns. Clearly in A, B and C, only part of the candidates are appropriate for the TTFs. We next use our probabilistic approach, MPAE method and reorder candidates in which motifs best suited for the TTFs are assigned to larger α values. We call α the MPAE score. Here, the sum of MPAE scores is equal to the number of motifs suggested by DBDSim. The final candidate rankings, denoted by the right columns, successfully preserve correct specificities while wrong decisions are separated by low MPAE scores.

that MEOX1 and MEOX2 actually possess the same binding profiles since unexpected noises in Taipale experiments did not allow to capture the dimeric motifs of MEOX1. But, we can find many similar examples with notable DBD similarities and different motifs (e.g. ALX3 and ALX4, EN1 and EN2, PRRX1 and PRRX2). It is thus likely that protein structures are responsible for such consistent binding variation. Throughout the paper, we draw our conclusions assuming that experimental TF motifs are complete and accurate.

As another example, the left column in Figure 2C displays a set of candidate TFs that do not have an obvious relation to PITX3 such as the same starting protein names. Additionally, while all selected TFs are almost indistinguishable in terms of DBDScore, according to motif similarity, the single best choice for PITX3 motif is DMBX1.

MPAE: motif prediction from chromatin accessibility and gene expression

The example in Figure 2 and the validation in 'Results' section suggest the need for methods to reduce the false positives in the DBDSim candidate set. Here, we propose a model-based approach (MPAE) for this task. MPAE combines genomic information with matched chromatin accessibility (DNase-seq) and gene expression (RNA-seq) data to re-order TFs in the DBDSim candidate set. We collect data from the ENCODE Project Consortium (22), ROADMAP Epigenomics Project (23) and Pritchard Lab (available at the Coriell Institute for Medical Research) and construct a database that includes matched RNA-seq and DNase-seq profiles on a large variety of cellular conditions (Supplementary Tables S2 and 3). Assuming we have J candidate motifs, for each promoter, say promoter i, we summarize the correlation between TTF expression and candidate binding site openness by $D_i = (D_{i1}, \cdots, D_{ij}, \cdots, D_{iJ})$. Here, D_{ij} represents the summarized data for motif j at promoter i, and it takes a larger value if:

- There are binding sites within the promoter well matched to the PWM of motif j.
- The binding site for motif j is closer to promoter i's transcriptional start site.
- Target TF expression is highly correlated with the binding site openness.

We assume a statistical model for D_i with parameters $\alpha = (\alpha_1, \cdots, \alpha_j, \cdots, \alpha_J)$ for all promoters i, where α_j provides evidence whether motif j is a good candidate (see 'Materials and Methods' section for details). As α_j becomes large, we are more confident that motif j is associated with TTF. The parameter α needs to be estimated from the data. To account for the possibility that TTF may not bind to the promoter, we introduce a binary random variable U_i, which $U_i = 1$ indicates that TTF is bound to promoter i, otherwise $U_i = 0$. We treat U as missing data and implement the expectation-maximization (EM) algorithm to obtain the maximum likelihood estimate (MLE) for α. Finally, we rank candidate motifs according to the estimated $\hat{\alpha}$ and choose the top three candidates as predicted motifs for the target TF.

MPAE tends to select candidate motifs that (i) have stronger correlations between their motif sites openness and TTF expressions across various cell types and (ii) are close to transcription start sites (TSS). Such considerations have been quite effective and generated promising inferences as demonstrated by Figure 2A–C. The right columns in the figure show our probabilistic method correctly rearranges DBDSim candidates and allocates larger MPAE scores, α, to the candidate motifs that are almost identical to PAX5, MEOX1 and PITX3 true motifs.

MATERIALS AND METHODS

Data

We used Taipale motifs that were obtained by high-throughput SELEX and ChIP sequencing (12). All DBDs associated with the PWMs were extracted from www.uniprot.org (Supplementary Table S1). For a given PWM, we scanned the human genome sequence (hg19) and preserved matches above a defined threshold (see 'MPAE' subsection). We restricted motif occurrences to those within 5 kb of TSS. We next collected 217 DNase-seq and RNA-seq pairs, each performed on the same cell type. The experiment accession numbers are listed in Supplementary Table S2 and publicly available at ENCODE Project Consortium (22), ROADMAP Epigenomics Project (23) and Pritchard Lab (the Coriell Institute for Medical Research). To ensure our analysis is not biased by tissue-type, we employed the agglomerative hierarchical clustering algorithm to cluster the 217 matched pairs into 100 groups based on TF expressions. In MPAE procedure, we considered just one matched pair from each cluster (Supplementary Table S2). We then gathered known DNA-binding TFs and retrieved their DBDs from Uniprot (Supplementary Table S3). We finally calculated motif sites openness and DNA-binding TF expressions using the 100 matched samples (Supplementary Material 1).

DBDSim

We employ MATLAB function seqpdist(DBD sequences, 'ScoringMatrix', 'BLOSUM62') and compute pairwise distances between DBD sequences of Taipale TFs. We utilize BLOSUM62 to score amino acid alignments for non-zinc finger proteins. Since zinc fingers often include multiple binding domains and require strict scoring alignment, we perform PAM10 on these proteins. The MATLAB function linkage(DBD distances,'average') is next applied to generate an agglomerative hierarchical cluster tree based on the calculated pairwise distances. Here, linkage is set to use unweighted average distance when comparing clusters. A threshold $c \approx 0.5$ for cutting the tree is defined and clusters are formed when a node and all sub-nodes have inconsistent value below c (MATLAB function cluster(tree,'cutoff',c) is used). The cutoff ensures that DBDs belonging to each cluster are quite similar, i.e. sequences are identical at roughly 85% of their aligned positions or more.

We finally calculate the DBDSim score of cluster C as:

$$\frac{\sum_{m=1}^{|C|} S(C_m, TTF)}{|C|}, \qquad (1)$$

where $|C|$ indicates the total number of TFs within cluster C and $S(C_m, TTF)$ represents the protein alignment score between mth element of C and TTF (nwalign from MATLAB provides score S). Motifs having DBDSim score above a set threshold are predicted as candidate motifs. This threshold is set to be 100 if the DBD of TTF is 60–80 amino acids in length. The threshold is decreased to 50 for shorter protein sequences such as bZIP and bHLH and increased to 200 for zinc finger TFs with long DBDs. These cutoffs, obtained in leave-one-out cross-validation (see Results), are the highest DBDSim scores that capture existing correct motifs for every TTFs. Note that small variations in the thresholds, ±20%, does not notably change the cross-validation results.

MPAE

We propose MPAE to select candidate motifs that are more likely to be a good motif for TTF. Our unsupervised learning algorithm aims to take advantage of DNA accessibility combined with the expression of TTF. The proposed procedure is described in the following.

TF-promoter activity. Our method starts by scanning all promoters for sites with substantial similarity to the candidate PWMs. We use FIMO software to locate motif matches and calculate P-value scores for matches found in the promoters (24). Denote by J the total number of candidate motifs and N the total number of promoters under consideration. The PWM score X_{i,j_l} for candidate motif j at site l in promoter i is defined as:

$$X_{i,j_l} = \begin{cases} -\log_{10}(P\text{-value}), & \text{if } -\log_{10}(P\text{-value}) \geq 5 \\ 0, & \text{otherwise} \end{cases}$$

We perform thresholding on the P-values to exclude sites with poor motif match. The PWM score is not a new concept and simply derived from P-values reported by FIMO. Let Y_{i,j_l} denote the correlation between our TTF expressions and the openness at position j_l. Specifically, we obtain TTF expressions and the j_l openness from 100 matched RNA-seq and DNase-seq samples (Supplementary Material 1) and calculate Y_{i,j_l} as the normalized inner product between the expression and openness vectors. Replacing inner product with Pearson or Spearman correlation increases computational costs and shows negligible changes in Results section. We introduce the TF-promoter activity $D_{i,j}$ as follows:

$$D_{i,j} := \sum_l \left(1 + \delta_i \frac{\text{TSS}_{i,j_l}}{\text{PL}_i}\right)^{-1} X_{i,j_l} Y_{i,j_l}, \qquad (2)$$

where TSS_{i,j_l} shows the distance from position j_l to transcription i start site, PL_i shows promoter i length and δ_i indicates the weight-distance. The promoter regions are assumed to start from the TSS of genes. We fix the promoter length PL_i to $5kbp$ and set $\delta_i = 5$ for every i. If $D_{i,j}$ is zero for all j, we do not take promoter i into consideration. We then scale $D_{i,j}$ with respect to $\sum_{j=1}^{J} D_{i,j}$ to make it sum up to one.

DNA accessibility at the precise locations to which TF is bound can be low because the TF protects its binding sites from DNase I cleavage (25,26). Here, we do not consider chromatin accessibility at the exact binding sites. Instead, we identify open chromatin regions using HOTSPOT software (see Supplementary Material 1 for details). If a motif binding site falls within hotspot region, we allot the hotspot openness to the motif site, otherwise openness is set to zero. In other words, we assume that genomic regions surrounding actual TF binding sites are accessible given the TF is expressed. We hence expect Y_{i,j_l} to play an important role in TF-promoter activity, which has led to promising motif predictions ('Results' section).

To obtain a functional form for TF-promoter activity, we examined multiple linear and non-linear models to achieve a reliable motif inference in leave-one-out cross-validation. We further investigated different promoter lengths ranging from 1 kbps to hundreds of kbps, various thresholding on motif match scores, and several functions to correlate expression and accessibility across diverse cell types. The proposed TF-promoter activity led to not only near-optimal performance but also inexpensive computational algorithms due to the summarization of input data.

The introduced model for TF-promoter activity has parallels with previous studies. For instance, Chen *et al.* (27) developed a statistical framework that integrates ChIP-seq data and expression profiles to identify target genes of TFs. Similar to our model, they assumed that TF binding sites are more probable to occur in ChIP peaks closer to TSS. They also required peak intensity and target expression to follow analogous patterns in different cellular conditions, which roughly corresponds to Y_{i,j_l}. In another work by (10), distance to TSS, motif match score, and experimental data such as histone modifications and chromatin accessibility were statistically combined to infer genome-wide TF binding sites. Note that these studies utilized either expression or accessibility information to elucidate transcriptional regulations while we propose a joint analysis of expression and accessibility suitable for motif prediction.

Ranking candidate motifs by modeling TF-promoter activity. We assume that TTF has regulatory roles on only a subset of promoters, which are called the relevant promoters. Let U_i be the binary indicator of the event that promoter i is a relevant promoter. Since we do not know the relevant promoters, we regard U_i as a latent random variable following a Bernoulli distribution with parameter θ:

$$P(U_i|\theta) = \theta^{U_i}(1-\theta)^{1-U_i}, \qquad (3)$$

where θ is the probability that TTF binds to a random promoter. When $U_i = 1$ (i.e. the promoter is relevant for TTF), the TF-promoter activities $D_i = (D_{i,1}, \ldots, D_{i,j}, \ldots, D_{i,J})$ should be informative to rank the candidate motifs of TTF. In other words, a good motif is likely to give a higher value for $D_{i,j}$. Thus, D_i is assumed to follow a Dirichlet distribution with parameters $\alpha = (\alpha_1, \ldots, \alpha_j, \ldots, \alpha_J)$:

$$f(D_i|U_i = 1, \alpha) = \frac{1}{B(\alpha)} \prod_{j=1}^{J} D_{i,j}^{\alpha_j - 1}$$

$$= \frac{\Gamma\left(\sum_{j=1}^{J} \alpha_j\right)}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \prod_{j=1}^{J} D_{i,j}^{\alpha_j - 1}, \qquad (4)$$

where $B(\cdot)$ is the beta function and α_j represents the belief that motif j is a good candidate for TTF. We assume that $\sum_{j=1}^{J} \alpha_j = J$. Fitting the above model to all relevant promoters, α is inferred and then used to select good motifs for TTF.

When $U_i = 0$ (i.e. the promoter is not relevant for TTF), D_i should not depend on the parameter α. Accordingly, we assume a neutral distribution by setting $\alpha_j = 1$ for all j:

$$f(D_i|U_i = 0, \alpha) = (J - 1)! \tag{5}$$

The expectation-maximization (EM) algorithm. The model parameters (α, θ) are estimated by maximizing the likelihood function $f(D|\alpha, \theta)$ for each TTF separately, as the parameters are assumed to be TTF-specific. Because the latent variable U is unknown, we implement the EM algorithm. We first calculate the complete likelihood $f(D, U|\alpha, \theta)$:

$$\begin{aligned} f(D, U|\alpha, \theta) &= f(D|U, \alpha)P(U|\theta) \\ &= \prod_{i=1}^{N} f(D_i|U_i, \alpha)P(U_i|\theta) \\ &= \prod_{i=1}^{N} f(D_i|U_i = 1, \alpha)^{U_i} f(D_i|U_i = 0, \alpha)^{1-U_i} \theta^{U_i}(1-\theta)^{1-U_i} \\ &= (J-1)! \prod_{i=1}^{N} \left(\frac{\prod_{j=1}^{J} D_{i,j}^{\alpha_j - 1}}{\prod_{j=1}^{J} \Gamma(\alpha_j)} \right)^{U_i} \theta^{U_i}(1-\theta)^{1-U_i} \end{aligned} \tag{6}$$

In the expectation step (E-step), the expectation of the log likelihood function, given the current estimated parameters, (α^t, θ^t), is calculated as:

$$\begin{aligned} Q(\alpha, \theta|\alpha^t, \theta^t) &= \sum_{U} P(U|D, \alpha^t, \theta^t) \log f(D, U|\alpha, \theta) \\ &= \sum_{U} P(U|D, \alpha^t, \theta^t) \log f(D|U, \alpha) + \\ &\quad \sum_{U} P(U|D, \alpha^t, \theta^t) \log P(U|\theta) \end{aligned} \tag{7}$$

To simplify the above equation, we calculate

$$\begin{aligned} P(U_i = 1|D_i, \alpha^t, \theta^t) &= \\ \frac{P(D_i|U_i{=}1,\alpha^t)P(U_i{=}1|\theta^t)}{P(D_i|U_i{=}1,\alpha^t)P(U_i{=}1|\theta^t)+P(D_i|U_i{=}0,\alpha^t)P(U_i{=}0|\theta^t)} \\ &= \frac{\theta^t \frac{1}{B(\alpha^t)} \prod_{j=1}^{J} D_{i,j}^{\alpha_j^t - 1}}{\theta^t \frac{1}{B(\alpha^t)} \prod_{j=1}^{J} D_{i,j}^{\alpha_j^t - 1} + (1-\theta^t)(J-1)!} \\ &= \frac{1}{1 + \frac{1-\theta^t}{\theta^t} \left(\prod_{j=1}^{J} \Gamma(\alpha_j^t)\right) \left(\prod_{j=1}^{J} D_{i,j}^{1-\alpha_j^t}\right)} := q_i^t \end{aligned} \tag{8}$$

Employing q_i^t, the first term in (7) can be rewritten as:

$$-\sum_{j=1}^{J} \log \Gamma(\alpha_j) \sum_{i=1}^{N} q_i^t + \sum_{j=1}^{J} \alpha_j \sum_{i=1}^{N} q_i^t \log D_{i,j} - \sum_{j=1}^{J} \sum_{i=1}^{N} q_i^t \log D_{i,j} + N \log(J-1)! \tag{9}$$

and the second term as:

$$\log \theta \sum_{i=1}^{N} q_i^t + \log(1-\theta) \sum_{i=1}^{N} (1 - q_i^t) \tag{10}$$

See Supplementary Material 1 for details and derivations. Formula (9) is only a function of parameter α while (10) only depends on θ. Therefore in the maximization step (M-step), we maximize (9) and (10) separately. It can be shown that

$$\theta^{t+1} = \frac{\sum_{i=1}^{N} q_i^t}{N} \tag{11}$$

To this end, we estimate parameter α through the following concave optimization

$$\alpha^{t+1} = \arg\max_{\alpha} \sum_{j=1}^{J} \alpha_j \sum_{i=1}^{N} q_i^t \log D_{i,j} - \sum_{j=1}^{J} \log \Gamma(\alpha_j) \sum_{i=1}^{N} q_i^t$$
$$\text{subject to} \sum_{j=1}^{J} \alpha_j = J \text{ and } \alpha_j > 0,\ 1 \le j \le J \tag{12}$$

Although our optimization is concave, the estimate of parameter α is not available at a closed form. Borrowing ideas from (28), we propose a simple and efficient iterative scheme to approximate α. In this scenario, we consider an alternative representation for α:

$$\alpha_j = J w_j \quad \text{and} \quad \sum_{j=1}^{J} w_j = 1 \tag{13}$$

which assures that $\sum_{j=1}^{J} \alpha_j = J$. We emphasize that an initial $\alpha > 0$ always lead to a positive parameter estimation due to the concavity of the cost function when $\alpha > 0$. We next reparameterize our optimization problem with the unconstrained vector z where

$$w_j = \frac{z_j}{\sum_{k=1}^{J} z_k} \tag{14}$$

The gradient of (12) with respect to z_j is

$$\frac{J}{\sum_{k=1}^{J} z_k} \left(-\Psi(J w_j) \sum_{i=1}^{N} q_i^t + \sum_{i=1}^{N} q_i^t \log D_{i,j} + \sum_{k=1}^{J} w_k \Psi(J w_k) \sum_{i=1}^{N} q_i^t - \sum_{k=1}^{J} w_k \sum_{i=1}^{N} q_i^t \log D_{i,k} \right) \tag{15}$$

and $\Psi(x) = d \log \Gamma(x)/dx$. Setting the gradient to zero, we arrive at the update

$$\Psi(\alpha_j) = \frac{1}{\sum_{i=1}^{N} q_i^t} \left(\sum_{i=1}^{N} q_i^t \log D_{i,j} + \sum_{k=1}^{J} w_k^t \Psi(J w_k^t) \sum_{i=1}^{N} q_i^t - \sum_{k=1}^{J} w_k^t \sum_{i=1}^{N} q_i^t \log D_{i,k} \right) \qquad (16)$$

$$w_j^{t+1} = \frac{\alpha_j}{\sum_{k=1}^{J} \alpha_k} \qquad (17)$$

We make use of Newton method to solve equation (16) for α_j where other variables are estimated from the last iteration t (Supplementary Material 1). We then calculate the new w_j^{t+1} as indicated in (17). In a similar fashion, we derive w_j^{t+1} for every j and finally update α^t.

We summarize our MPAE approach in four steps:

- Initialize α with respect to the constraint (12) and also $0 < \theta < 1$.
- Compute q_i^t as shown in (8), E-step.
- Find θ^{t+1} and α^{t+1} using (11), (16), and (17), M-step.
- Perform E-step and M-step until $|\alpha^{t+1} - \alpha^t| < 10^{-6}$ and $|\theta^{t+1} - \theta^t| < 10^{-6}$.

The EM algorithm converges to a local minimum quickly, typically in a few iterations (Supplementary Material 1). We perform the EM multiple times and choose the local optima that results in the largest log marginal likelihood. We finally rank candidate motifs once α is estimated and view candidates with the highest α values as reliable predicted motifs.

RESULTS

Performance analysis

We performed leave-one-out cross-validation to assess the performance of our procedure. Specifically, we inferred the motif of each Taipale human TF that was first removed from the training data. We then used the TOMTOM program (20) to compare the inferred candidate motifs to the omitted Taipale motif by computing the motif-comparison P-value. Any candidate motif with a P-value < 0.0001 is regarded as a 'suitable/correct' motif for the (removed) target TF. Our findings are presented in the following three subsections.

Protein sequences are strongly informative to infer motif models for zinc finger TFs. We first analyze the DBDSim performance for zinc finger TF (ZF-TF) based on cross-validation among the 48 Taipale ZF-TFs. For every target ZF-TF, we define positive outcomes as candidate motifs with DBDSim score above a threshold. Given the positive set, the red curve in Figure 3A shows the average proportion of correct predictions across 48 cases as we change the DBD similarity threshold (see average precision in Supplementary Material 1). The black color at any threshold depicts the proportion of cases that have at least one correct motif in their positive sets, called sensitivity. When DBDSim score between the target ZF-TF and the candidate TF meets 200, we observe that the precision becomes 88% while the sensitivity does not notably drop. However when the DBDSim falls below 200, the precision is likely to be low and we simply refrain from making a prediction. Note that the threshold was obtained for target ZF-TFs that contain three to five ZF domains.

Moreover, DBDSim always allotted the highest DBD score (>200) to the best motifs. In particular, we found 27 cases satisfying the 200 threshold where the percentage of correct motifs within the first cluster, candidates with the highest DBDSim score, was on average 91%. Cross-validation analysis on a subset of these cases are shown in Table 1 (complete results are available in Supplementary Material 2). For the remaining 21 ZF-TFs that did not meet the threshold, we have 12% correct motifs in the first cluster and 4% among the top 10 clusters.

Several methods have been developed to predict ZF motifs from DBD sequences (19,29). They analyze structural models of ZF–DNA interface to estimate new binding specificities. To compare such predictions with DBDSim, we used the recent algorithm in (19), named Persikov method. Recall our 91% average precision derived from the 27 ZF-TFs. Persikov average precision for these cases is just 4%. Even increasing the default P-value to 0.001 and 0.01, we observe low average precisions, 26 and 33%. To enable a fair comparison, we also determined DBDSim performance for the ZF proteins in which Persikov method works best. Setting the P-value to 0.01, Persikov predicts correct motifs for 12 Taipale ZF-TFs. DBDSim, performed on these cases, is able to make predictions for 9 ZF-TFs with 100% average precision. The other 3 cases that do not satisfy the 200 threshold have no similar motif in Taipale dataset and DBDSim properly avoids making a prediction. Decreasing the P-value to 0.0001 when evaluating DBDSim precision among 12 cases, we still achieve 100% precision for the 9 ZF-TFs. On the contrary, Persikov obtains just one prediction having the P-value < 0.0001. To generalize our comparison to all Taipale ZF-TFs, we eliminated the 200 threshold criteria and treated the top cluster as DBDSim predictions. Applying DBDSim and Persikov methods to 48 ZF-TFs, Persikov arrives at 2, 17 and 25% average precision for the P-value 0.0001, 0.001 and 0.01 while DBDSim gives 57, 58 and 60%. Although DBDSim outperforms Persikov, we note that our comparison may be not entirely fair since both methods do not utilize the same training dataset.

Persikov predictions are inaccurate when the total number of ZF domains in a single protein increases. The main challenge involves various unknown cooperations among ZF domains that only activate a subset of domains bound to DNA. DBDSim, on the other hand, is able to handle such complexity provided that a ZF protein with the DBDSim score above the threshold exists in our dataset. Figure 3B demonstrates a comparison of DBDSim and Persikov performance based on the number of ZF domains. For KLF16 and ZNF740, the two methods predict motifs close to the true binding models. As the number of ZFs becomes larger, DBDSim notably outperforms and accomplishes the best motif approximation for GLIS1, HIC2 and ZIC1. We emphasize that many ZF proteins are associated with numerous domains and the architecture of ZF–DNA interactions cannot be properly addressed by current models.

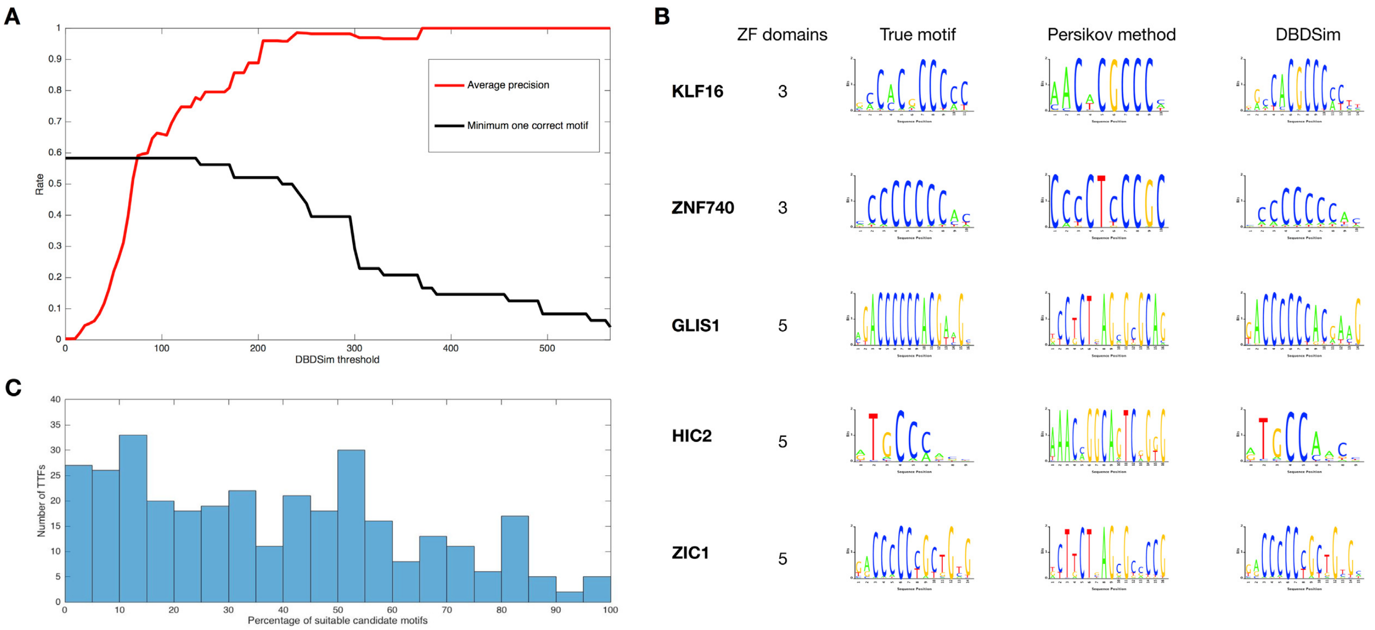

Figure 3. (**A**) DBDSim analysis for ZFs. We impose a cut-off value on DBDSim score to select the candidate set. As this threshold alters (x-axis), the red curve gives the average precision among 48 Taipale ZF-TFs and the black line indicates the fraction of cases whose candidate sets include minimum one correct motif. (**B**) Method comparison of ZF–DNA specificity inference. Second column shows the total number of ZF domains retrieved from Uniprot. The specificities of all five factors, indicated as True motif, have been determined in Taipale data. We denote the proposed algorithm by (19) as Persikov method. Both DBDSim and Persikov method suggest specificities analogous to the grand truth provided that TFs contain three ZF domains. However as ZF domains increases, DBDSim notably outperforms structural model-based predictions. Note that our conclusion was validated across all Taipale ZFs. Here, we only show a few comparisons to avoid repetition. (**C**) DBDSim performance. We performed DBDSim to infer specificities of Taipale human non-ZF TFs assuming their true motifs were missing. The top 10 clusters were used as candidate motifs and the percentage of candidates suitable for each TTF was calculated. Histogram of the percentages shows that DBDSim most likely obtain suitable candidates, however, those are combined with numerous bad candidates that must be disregarded.

Recent study by (30), named Weirauch method, has developed a heuristic DBD-based scheme to expand the assignment of motifs to both ZF and non-ZF TFs. Relying on motif database generated through PBM technique, they measured the levels of DBD similarity between any two proteins that have approximately an identical motif. It was concluded DBD similarity above a rigid threshold for each TF class typically guarantees accurate motif predictions. We observe that Weirauch method when applied to non-ZF proteins gives considerably lower sensitivity compared to DBDSim (see the next subsection), however for ZF, both methods exhibit analogous performance. Applying Weirauch to 48 Taipale ZF-TFs, we are able to infer motifs for 21 ZF proteins representing 98% average precision. DBDSim can also attain the same performance if our threshold increases to 250, but, the sensitivity reduces from 27 cases to 21. Note that Taipale motif database was used for both Weirauch and DBDSim predictions to allow fair comparison.

We emphasize that correct motif predictions for ZF family are always associated with the highest DBDSim score and thus, Weirauch strict thresholding does not significantly reduce the sensitivity of method. But imposing such thresholding on non-ZFs nonetheless sacrifices sensitivity to ensure accurate motif inference. Unlike Weirauch, DBDSim is designed to capture candidate sets that most likely contain suitable motifs for TTF. We then propose MPAE to improve DBDSim's precision while benefiting from its high sensitivity. The detailed analysis is provided in the following subsection.

MPAE accurately predicts motifs for the majority of TF families. For non-ZF proteins, suitable candidates are not necessarily coupled with the highest DBDSim score; e.g. Figure 2B and C. To assess DBDSim performance more systematically, we calculated the percentage of Taipale human TFs (ZF excluded) whose candidate set includes at least one correct motif, defined as the sensitivity of our scheme. The sensitivity was 94% where DBDSim candidate set was constructed using the 10 closest clusters to TTF. We then computed the percentage of suitable motifs among each candidate set. On average, <38% of motifs in the candidate set are correct, Figure 3C depicts histogram of suitable motif percentages. Therefore, DBDSim can offer a candidate set highly likely to contain the correct motif, but a majority of the motifs in the candidate set may be incorrect. This motivated us to develop the MPAE method to select the correct motif(s) among the candidates.

To perform MPAE on non-ZFs, we considered proteins expressed in at least one sample because otherwise the correlation Y_{i,j_l} would become meaningless. In the first step, DBDSim was applied and TTFs not comparable to any available DBDs were dropped, i.e. no prediction will be made for those with DBDSim scores below the aforementioned thresholds. For the remaining TTFs, we used MPAE

Table 1. Leave-one-out cross-validation

Family	Name	Correct consensus	Predicted consensus	*P*-value
znfC2H2	EGR1	TACGCCCACGCATT	ATACGCCCACGCATTT	5.51817×10^{-16}
znfC2H2	GLI2	GACCACCCACGACG	GACCCCCCACGAAG	5.22287×10^{-8}
znfC2H2	KLF14	GGCCACGCCCCCTT	GCCACGCCCCC	1.8514×10^{-10}
znfC2H2	SCRT2	ATGCAACAGGTGG	GAGCAACAGGTGGTT	2.44446×10^{-12}
znfC2H2	SP1	ACCCCGCCCCC	GCCACGCCCCC	8.45718×10^{-10}
znfC2H2	YY2	GTCCGCCATTA	GCCGCCATTAT	7.26281×10^{-6}
znfC2H2	ZBTB7B	GCGACCACCGAA	GCGACCACCGAA	1.58617×10^{-9}
ETS	EHF	AACCCGGAAGTA	AACCCGGAAGTG	2.57183×10^{-8}
ETS	ELK3	ACCGGAAGTA	AACCGGAAATA	2.53826×10^{-6}
E2F	E2F1	ATTGGCGCCAAA	TTTGGCGCCAAA	9.30523×10^{-9}
bHLH	HES5	CGGCACGTGCCA	GACACGTGCC	9.92475×10^{-8}
bHLH	NEUROG2	AACATATGTC	ACCATATGGC	4.12085×10^{-6}
bZIP	BATF3	TGATGACGTCATCA	GATGACGTCATC	3.11216×10^{-8}
bZIP	CREB3	GTGCCACGTCATCA	ATGCCACGTCATCA	6.85621×10^{-10}
homeobox	HOXA10	GGTCGTAAAAAT	GTCGTAAAA	1.19757×10^{-7}
homeobox	MEIS3	TGACAGGTGTCA	TGACAGGTGTCA	1.21951×10^{-12}
homeobox	PRRX1	CCAATTAA	TCTAATTAAA	4.33052×10^{-7}
POU	POU3F3	ATGCATAAATTA	ATGCATAATTTA	7.40732×10^{-14}
RFX	RFX4	CGTTGCCATGGCAACG	CGTTGCCATGGCAACC	3.06322×10^{-15}
AP2	TFAP2B	TGCCCTGAGGGCA	TGCCCTGAGGGCA	1.0674×10^{-11}
NFI	NFIA	TTGGCACGGTGCCAA	TTGGCACGGTGCCAA	9.47109×10^{-11}
HSF	HSF1	TTCTAGAACGTTC	TTCTAGAACGTTC	1.31947×10^{-17}
IRF	IRF4	CCGAAACCGAAACTA	CCGAAACCGAAACT	1.27332×10^{-9}
MADS	MEF2D	ACTATAAATAGA	TCTAAAAATAGA	2.11539×10^{-10}
PAX	PAX2	CGTCACGCTTGACTGCTC	CGTCACGCATGAGTGCTC	1.58414×10^{-21}
HMG	SOX2	GAACAATGGTATTGTTC	AACAATGGTAGTGTT	6.01762×10^{-10}
HMG	SOX8	ATGAATTGCAGTC	ATGAATTGCAGTCAT	1.84138×10^{-10}
forkhead	FOXC2	GTAAATAAACA	TGTAAATAAACAA	1.03226×10^{-9}
forkhead	FOXO1	TTTCCCCACACG	TTTCCCCACACGAC	2.51112×10^{-11}
p53	TP63	AACATGTTGGGACATGTC	AACATGCCCGGGCATGTC	9.91755×10^{-8}
RUNT	RUNX2	TAACCGCAAACCGCAA	TAACCGCAAACCGCAA	7.42584×10^{-18}
T-box	TBX21	TCACACCTTAAAGGTGTGA	TTTCACACCTCAGAGGTGTGAGA	9.11484×10^{-13}
nuclear receptor	ESRRA	TTCAAGGTCAT	GAGGTCATGACCCC	4.25694×10^{-2}
nuclear receptor	RXRG	GAGGTCATGACCCC	GGGGTCATGACCCC	4.18164×10^{-18}

Column 1 and 2 denote TF family and TF name. Column 3 and 4 indicate true and inferred consensus whose motif-comparison *P*-value is depicted in column 5. For predicting ZF motifs, we select candidates with the highest DBD score above 200 and for non-ZF, we consider the top three candidates having $\alpha > 1$. Although multiple motifs were predicted for each TTF, we only show one motif for simplicity. Detailed results are provided in Supplementary Material 2.

method and selected the top three candidates satisfying $\alpha > 1$ (see section 4 in Supplementary Material 1).

The performance was promising where the average fraction of correct motifs within the top three candidates, average precision, was almost 90% (nuclear receptors not included, see the last paragraph). For comparison, we selected the smallest number of DBDSim clusters that (i) contain three candidates and (ii) have the highest DBDSim scores. Given this new candidate set, on average, 71% of motifs were appropriate for TTFs. Since Taipale data contains many TFs with both human and mouse motifs, the above precision for DBDSim can be misleading. Specifically, leave-one-out cross validation is in favor of DBDSim due to the strong similarity between human and mouse DBDs/motifs. Removing mouse TFs from our analysis, the average precision dropped to 66 for DBDSim, but, MPAE performance did not change. We also emphasize that Taipale includes many paralog proteins (e.g. ELF3, ELF4 and ELF5 are paralogs of ELF1), which causes an overestimation of DBDSim performance. For instance, we limited our training data to a subset of Taipale TFs that has less paralog proteins, in total 192 TFs, and obtained 46 and 83 as the average precision for DBDSim and MPAE, respectively.

As another comparison, Weirauch method was applied to non-ZFs (nuclear receptors removed) assuming predicted motifs were taken from Taipale dataset. The average precision was 85% which is 5% below MPAE. Weirauch sensitivity, however, was 69% significantly lower than our 94%. Note that we achieve 94% whether nuclear receptors are considered or not. The sensitivity difference becomes even larger when Taipale training data is limited to the192 TFs with less paralog proteins. In particular, DBDSim combined with MPAE exhibited 86% sensitivity and 83% average precision while Weirauch showed 50 and 79%.

Supplementary Material 2 shows the MPAE evaluation on a subset of TTFs that reflects the overall performance. Table 1 also illustrates part of the evaluation. As indicated in Table 1, the motif inference for the majority of TF families were accurate, however for nuclear receptor family, we observed that some motifs were mispredicted. This could be due to the multifunctional mechanism of receptors, such as ligand binding and heterodimerization (31), which cannot be correctly addressed in the proposed models.

MPAE takes homodimers into account. The SELEX analysis followed by massively parallel sequencing in (32) showed that many TFs can bind to DNA either as monomers or homedimers. To obtain homodimeric binding, Jolma et al. (12) employed the SELEX technique to determine whether TF is bound to two similar sites that are close in a DNA fragment. Given a clear spacing and orientation between the two sites, the homodimer profile were attained.

Dimerization introduces various spacings and orientations among the monomers/half sites. MPAE cross-validation for different families such as SOX, E2F, FOX,

Table 2. Distribution of novel predictions across TF families

Family	Taipale	Predicted	Name	Taipale	Predicted
homeobox	112	37	E2F	6	2
bHLH	35	34	IRF	6	2
znfC2H2	48	32	RFX	4	2
bZIP	20	29	T-box	12	2
forkhead	16	19	p53	1	2
HMG	13	8	MADS	4	1
NFAT	4	6	NFI	3	1
MAD	1	5	PAX	8	1
GATA	3	4	POU	14	1
ETS	24	3	RUNT	2	1
IRX	2	3	SAND	1	1
MYB	2	3	TEA	3	1

Here, 'Family' identifies TF class, 'Taipale' indicates the number of Taipale human TFs and 'Predicted' shows the number of TFs for which motif inference was achieved.

RFX and RUNT revealed that homodimeric orientation and spacing preferences can be anticipated (Supplementary Material 2). For instance, MPAE detects distinct spacing/orientation preferences in SOX family, proteins that mainly display dimeric motifs in Taipale. Furthermore, our method inference is not biased toward monomer nor homodimer and selects candidates that have larger MPAE score across varied cell types. As an example, RUNX2 possesses one Taipale monomer and two homodimers where MPAE prefers homodimeric motif. On the other hand, for ELK1 containing two monomers and one dimer, MPAE nominates monomeric binding.

Identification of DNA-binding specificities for human transcription factors not covered by Taipale

We employed a list of 1988 human proteins from (33) and collected sequence-specific DNA binding TFs (16). The TFs were checked against Uniprot and DBDs are recovered accordingly (Supplementary Table S3). This gives us 1137 TFs covering different structural families. Excluding the Taipale TFs, we arrive at 756 TFs divided into 443 ZFs and 313 non-ZFs. We analyzed these two groups separately, i.e. ZF motif inference were achieved based on DBDSim, whereas MPAE was applied on non-ZF factors.

We were able to elucidate DNA-binding specificities of 32 ZF and 168 non-ZF proteins (Supplementary Table S4). Table 2 exhibits the number of predictions attained for each TF family. The 168 non-ZF predictions substantially expand human motif database, almost 50% growth compared to Taipale. However, the ZF coverage is low since ZF proteins represent the largest class of eukaryotic TFs while Taipale includes only a small set of them. In particular, ZFs with the starting Uniprot names ZB, ZE, ZF, ZI and ZN are numerous and often have ten to thirty domains in a protein. Unfortunately, to the best of our knowledge, few motifs have been experimentally determined for this type of ZFs. Note that even though ZF sensitivity is low, our cross-validation suggests predicted motifs are most likely precise.

To assess the reliability of these new predictions, non-Taiplae TFs, we looked for additional experimental data that may be useful for their validation. We found HOMER ChIP-seq experiments (34) for 60 of the 200 TFs, 32 ZFs and 168 non-ZFs, but across different organisms mostly non-human, *http://homer.salk.edu*. The 60 predicted motifs and the corresponding ChIP-seq derived motifs from HOMER are displayed side by side in part 3 of Supplementary Material 2. It is seen that the majority of the predicted motifs are in good agreement with HOMER motifs. Specifically, 87, 77 and 60% of the predicted motifs have P-values < 0.01, 0.001 and 0.0001 respectively. This unbiased validation by external experimental data implies that the remaining 140 new predictions should also be similarly reliable.

We stress that the expected accuracy of the 200 novel predictions can be higher than 60%, for the P-value 0.0001, due to biological and technical differences between HOMER and our training data. On the other hand, leave-one-out cross-validation might overestimate the expected performance, 90% precision. We therefore presented both leave-one-out and ChIP analysis to enable a fair evaluation on our new predicted motifs.

DISCUSSION

We have shown that the DBD-based approach is not always capable of predicting accurate binding profiles and the joint analysis of open chromatin and gene expression data can further improve inferences. The proposed method is best suited for TFs recognizing short DNA motifs or homodimers. In recent studies (35), many TFs are shown to cooperate with each other and bind to DNA as heterodimers. Such interactions play a crucial role in activating/repressing gene expressions and migrating cells into specific tissues. We expect that our approach can be generalized to identify heterodimers. This will be highly useful as the set of possible TF–TF–DNA interactions is enormous and difficult to ascertain experimentally.

The modular structure of our pipeline allows for the modification of different steps to best predict specificities of particular TFs. Alternative procedures may be implemented in DBDSim to better reflect resemblance between DBDs, for instance taking advantage of 3D-structure of proteins (36,37) as well as utilizing the four key residues of ZF domains (29). Moreover, the performance of MPAE method can further be improved with new motif scanning programs. Current scanning tools (24,34) do not exploit characteristics of dimeric spacing and orientation preferences and can not clearly distinguish dimeric binding models from similar monomers. Furthermore, sequencing technologies are being continually improved to precisely capture genome-wide measurement of chromatin accessibility. Our inference can readily discriminate closely located DNA-binding specificities as a result of reliable open chromatin in short genomic windows.

Motif findings have broad application in discovery of gene regulatory network, functional elements in human genome, and new gene interactions implicated in medical treatment. We have been able to identify binding specificities of 200 human TFs that possess no experimental motifs. Excluding ZF family, the predicted motifs combined with Taipale data covers the majority of human TFs. To reveal unknown ZF specificities, different procedures may be considered. Strategies based on amino acid-nucleotide contact energies allow to predict ZF specificities, but, existing approaches need to address the mechanism of ZF–DNA interactions when large number of ZF domains are presented in a protein. Noticeably, many ZFs include 5 up to 30 binding domains. Experimental determinations of specificities for additional ZF proteins with novel binding domains should be given the highest priority as they will provide not only direct knowledge on those proteins but also new training data to extend DBDSim predictions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Angel Rubio for his helpful comments.

FUNDING

National Institutes of Health (NIH) [R01HG007834, P50HG007735]. Funding for open access charge: NIH [R01HG007834].

Conflict of interest statement. None declared.

REFERENCES

1. Albert,R. and Othmer,H.G. (2003) The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in Drosophila melanogaster. *J. Theor. Biol.*, **223**, 1–18.
2. Karlebach,G. and Shamir,R. (2008) Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.*, **9**, 770–780.
3. Davidson,E.H., Rast,J.P., Oliveri,P., Ransick,A., Calestani,C., Yuh,C.-H., Minokawa,T., Amore,G., Hinman,V., Arenas-Mena,C. *et al.* (2002) A genomic regulatory network for development. *Science*, **295**, 1669–1678.
4. Margolin,A.A., Nemenman,I., Basso,K., Wiggins,C., Stolovitzky,G., Favera,R.D. and Califano,A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**(Suppl. 1), S7.
5. Chen,X., Xu,H., Yuan,P., Fang,F., Huss,M., Vega,V.B., Wong,E., Orlov,Y.L., Zhang,W., Jiang,J. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.
6. Amit,I., Garber,M., Chevrier,N., Leite,A.P., Donner,Y., Eisenhaure,T., Guttman,M., Grenier,J.K., Li,W., Zuk,O. *et al.* (2009) Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science*, **326**, 257–263.
7. Boyer,L.A., Lee,T.I., Cole,M.F., Johnstone,S.E., Levine,S.S., Zucker,J.P., Guenther,M.G., Kumar,R.M., Murray,H.L., Jenner,R.G. *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.
8. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
9. Jothi,R., Cuddapah,S., Barski,A., Cui,K. and Zhao,K. (2008) Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, **36**, 5221–5231.
10. Pique-Regi,R., Degner,J.F., Pai,A.A., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.
11. Cuellar-Partida,G., Buske,F.A., McLeay,R.C., Whitington,T., Noble,W.S. and Bailey,T.L. (2012) Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, **28**, 56–62.
12. 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.
13. Portales-Casamar,E., Thongjuea,S., Kwon,A.T., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W.W. and Sandelin,A. (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **38**, D105–D110.
14. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
15. 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.
16. Bahrami,S., Ehsani,R. and Drabløs,F. (2015) A property-based analysis of human transcription factors. *BMC Res. Notes*, **8**, 82.
17. Schröder,A., Eichner,J., Supper,J., Eichner,J., Wanke,D., Henneges,C. and Zell,A. (2010) Predicting DNA-binding specificities of eukaryotic transcription factors. *PLoS One*, **5**, e13876.
18. Christensen,R.G., Enuameh,M.S., Noyes,M.B., Brodsky,M.H., Wolfe,S.A. and Stormo,G.D. (2012) Recognition models to predict DNA-binding specificities of homeodomain proteins. *Bioinformatics*, **28**, i84–i89.
19. Persikov,A.V. and Singh,M. (2014) De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res.*, **42**, 97–108.
20. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
21. The UniProt Consortium (2014) Activities at the universal protein resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.
22. The ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia of DNA elements) project. *Science*, **306**, 636–640.
23. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nature Biotechnol.*, **28**, 1045–1048.
24. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
25. Hesselberth,J.R., Chen,X., Zhang,Z., Sabo,P.J., Sandstrom,R., Reynolds,A.P., Thurman,R.E., Neph,S., Kuehn,M.S., Noble,W.S. *et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.
26. Fu,Y., Sinha,M., Peterson,C.L. and Weng,Z. (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet*, **4**, e1000138.
27. Chen,J., Hu,Z., Phatak,M., Reichard,J., Freudenberg,J.M., Sivaganesan,S. and Medvedovic,M. (2013) Genome-wide signatures of transcription factor activity: connecting transcription factors, disease, and small molecules. *PLoS Comput. Biol.*, **9**, e1003198.
28. Minka,T.P. (2000) Estimating a Dirichlet distribution. Technical report, Microsoft Research.

29. Kaplan,T., Friedman,N. and Margalit,H. (2005) Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput. Biol.*, **1**, e1.
30. 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.
31. Mangelsdorf,D.J., Thummel,C., Beato,M., Herrlich,P., Schütz,G., Umesono,K., Blumberg,B., Kastner,P., Mark,M., Chambon,P. *et al.* (1995) The nuclear receptor superfamily: the second decade. *Cell*, **83**, 835–839.
32. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpää,M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
33. Ravasi,T., Suzuki,H., Cannistraci,C.V., Katayama,S., Bajic,V.B., Tan,K., Akalin,A., Schmeier,S., Kanamori-Katayama,M., Bertin,N. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.
34. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
35. Jolma,A., Yin,Y., Nitta,K.R., Dave,K., Popov,A., Taipale,M., Enge,M., Kivioja,T., Morgunova,E. and Taipale,J. (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, **527**, 384–388.
36. Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein–DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, **33**, 5781–5798.
37. Morozov,A.V. and Siggia,E.D. (2007) Connecting protein structure with predictions of regulatory sites. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 7068–7073.