Published online 19 December 2011

Nucleic Acids Research, 2012, Vol. 40, No. 5 ***e38***
doi:10.1093/nar/gkr1252

Identification of co-occurring transcription factor binding sites from DNA sequence using clustered position weight matrices

Young Min Oh[1], Jong Kyoung Kim[2], Seungjin Choi[2,*] and Joo-Yeon Yoo[1,3,*]

[1]Department of Life Sciences, Pohang University of Science and Technology, [2]Department of Computer Science and Division of IT Convergence Engineering, Pohang University of Science and Technology, and [3]Biotechnology Research Center, Pohang University of Science and Technology, Pohang, Republic of Korea

Received July 11, 2011; Revised November 30, 2011; Accepted December 1, 2011

ABSTRACT

Accurate prediction of transcription factor binding sites (TFBSs) is a prerequisite for identifying *cis*-regulatory modules that underlie transcriptional regulatory circuits encoded in the genome. Here, we present a computational framework for detecting TFBSs, when multiple position weight matrices (PWMs) for a transcription factor are available. Grouping multiple PWMs of a transcription factor (TF) based on their sequence similarity improves the specificity of TFBS prediction, which was evaluated using multiple genome-wide ChIP-Seq data sets from 26 TFs. The Z-scores of the area under a receiver operating characteristic curve (AUC) values of 368 TFs were calculated and used to statistically identify co-occurring regulatory motifs in the TF bound ChIP loci. Motifs that are co-occurring along with the empirical bindings of E2F, JUN or MYC have been evaluated, in the basal or stimulated condition. Results prove our method can be useful to systematically identify the co-occurring motifs of the TF for the given conditions.

INTRODUCTION

The ability of every living cell to properly respond to diverse stimuli depends on the genetic information encoded in its genome and signaling cascades that activate appropriate transcription factors (TFs) for gene regulation (1–3). To understand the global network of transcription for controlling diverse cellular responses, it is important to identify the regulatory modules that are responsible for spatial or temporal gene regulation. For this purpose, diverse integrative tools for genomic analysis of DNA sequences, accompanied by information on the transcriptome and interactome, have been actively developed (4).

High-throughput technologies, such as ChIP-chip, ChIP-PET and ChIP-Seq, allow genome-scale mapping of epigenetic modification and protein–DNA interactions in particular genomes (5,6). Integration of accumulated genome-wide experimental data with DNA sequence information allows the construction of a map of the transcriptional regulatory circuits encoded in a genome that can eventually lead to the identification of the regulatory modules for gene regulation. However, annotating the functional transcription factor binding sites (TFBS) in the regulatory modules remains a challenging task (7). The problem derives mainly from the nature of the DNA sequences that are recognized by transcription factors; they are relatively short and degenerate. Furthermore, transcription factors are known to recognize more than one consensus sequence (8), and similar DNA sequences can be recognized by different groups of transcription factors (9).

Because accurate prediction of the putative binding sites of transcription factors is a valuable tool for understanding transcriptional regulatory networks and mechanisms of transcriptional control, numerous computational tools have been generated. The most common method is the pattern matching approach that uses a position weight matrix (PWM) (10–13) or Hidden Markov

*To whom correspondence should be addressed. Tel: +82-54-279-2346; Fax: +82-54-279-2199; Email: jyoo@postech.ac.kr
Correspondence may also be addressed to Seungjin Choi. Tel: 82-54-279-2259, Fax: +82-54-279-2299; Email: seungjin@postech.ac.kr
Present address:
Jong Kyoung Kim, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Models (HMMs) (14). However, prediction of the putative TFBS using the predefined PWM suffers from a high rate of false positive discovery (15). To alleviate this problem, integration of the heterogeneous information (16), such as the DNA sequence conservation score (17,18) DNase-I hypersensitive score (19,20), or nucleosome occupancy (21) and modification information (22,23), has been successfully applied with enhanced prediction performance.

In parallel, approaches using PWM clustering based on the sequence similarity were proposed. In this method, a familial binding profile (FBP) is constructed from the multiple PWMs for each family of transcription factors, improving the sensitivity of *de novo* motif discovery algorithms (24,25). However, a FBP ignores the flanking positions of PWMs that are not aligned but which may be important for discriminating false positives; hence, this method can have low specificity in predicting functional binding sites. An alternative approach is to combine overlapping TFBSs predicted by the original PWMs belonging to the same cluster (15). This program can increase specificity by removing redundant TFBSs, but because it is based on the heuristic scoring system, it is not suitable for comparing scores of overlapping TFBSs. To overcome these problems, we have recently developed a motif-based scanning program (26). It searches STAT TFBS of high affinity scores, using the combined predicted TFBSs from PWMs that show similar binding specificity to STAT family members.

In an attempt to construct an efficient computational tool for predicting TFBSs, we applied the motif-based scanning program to other transcription factors with multiple PWMs. A total of 368 transcription factors and 565 PWMs were considered in this study. Finally, TFBS-scanner was applied to identify the co-occurring *cis*-motifs that might function coordinately. The source code of TFBS-Scanner program is freely available from the supporting webpage, https://sourceforge.net/p/tfbsscanner.

MATERIALS AND METHODS

Analysis of the TF-PWM network

To construct linkages between PWM and its cognate TF, we manually extracted 368 vertebrate TFs and 565 vertebrate PWMs information from the 'Matrix' and 'Factor' database of TRANSFAC professional version 9.4 (Supplementary Table S1A) (11). To map mouse or rat gene to the human gene identifier, Entrez Gene ID (27) (from FTP: Gene), 'HomoloGene' information (28) (from FTP: HomoloGene, build63) of the NCBI were used. Cytoscape 2.7.0 (29) was then used to visualize the TF–PWM interactions (Supplementary Table S1A); each node indicates TF or PWM, and edges for pairwise interaction. A total of 474 PWMs and 368 TFs made up the final graph, which contains multiple connected components (CCs) (Supplementary Table S1B). See the 'TF-PWM network.cys' or PDF files for each CC in the supporting webpage for detail view. For protein interactions between TFs within each CC, a total of 51 837 annotated PPI information has been extracted from the 'interactions' information (27) (from FTP: Gene) of the NCBI. For detail view, see the 'PPI in TF-PWM network.cys' or PDF files for each CC in the supporting webpage. The number of possible PPI is given by $N \times (N-1)/2$, where N is the number of TFs in each CC. To compute the distance between two PWMs, we used the distance measure proposed by Harbison *et al.* (30). Given two PWMs Θ_1,Θ_2 that are already aligned, the distance measure is defined by:

$$d(\Theta_1,\Theta_2) = \frac{1}{W}\sum_{w=1}^{W}\frac{1}{\sqrt{2}}\sum_{l=1}^{4}\left(\Theta_{1,wl}-\Theta_{2,wl}\right)^2,$$

where W is the length of aligned positions. For the optimal alignment between the two PWMs, we used the procedure described by Narlikar *et al.* (21).

Affinity score of a subsequence

To improve the generalization performance of the observed position count matrices, we transformed each position count matrix into a position frequency matrix (PFM) by adding position-dependent pseudo-counts. We used the statistical method of Rahmann *et al.* (31) for position-dependent regularization. For highly conserved positions, we add very small pseudo-counts in order to maintain the strong signal. In contrast, we add relatively large pseudo-counts at poorly conserved positions, preserving the overall composition of nucleotides of the position count matrix. Given a regularized PFM $\Theta_1 \in \mathrm{R}^{W\times 4}$, we represent a sequence s_i as a set of overlapping subsequences $s_{ij}^W = (s_{ij},\cdots,s_{i(j+W-1)})$ of length W to scan TFBSs by sliding a window of length W. We assume that a subsequence can be generated from either the PFM or a background model $\Theta_0 = [\theta_0^T,\cdots,\theta_0^T]^T \in \mathrm{R}^{W\times 4}$, where θ_0 is defined by a zero-order Markov chain. We used six different background models from our previous study (26) to capture the compositional bias of GC content in genomic sequences. Then, we converted the regularized PFM into a PWM Υ by computing the log-odds scores between the PFM and a background model, which is given by:

$$\Upsilon_{wl} = \log\left(\frac{\Theta_{1,wl}}{\Theta_{0,wl}}\right).$$

We decide whether a given subsequence is generated from the PFM or the background model based on the sum of log-odds score:

$$\lambda(s_{ij}^W) = \sum_{w=1}^{W}\sum_{l=1}^{4}\Upsilon_{wl}^{\delta(l,s_{i(j+w-1)})}.$$

Following the statistical method of Rahmann *et al.* (31), we computed the exact distribution of the sum of log-odds scores to measure the statistical significance of the log-odds scores of the subsequences. The exact distribution can be efficiently computed using the positional independence of PWMs and then applying convolution. From the exact distribution P_{Θ_0} given the assumption that the subsequence is generated from the background model, we

computed the type I sequence error probability $\alpha_n(s_{ij}^W)$ of the log-odds score $\lambda(s_{ij}^W)$, which is given by:

$$\alpha_n(s_{ij}^W) \approx 1 - \exp(-nP_{\Theta_0}(x \geq \lambda(s_{ij}^W))).$$

This error quantifies the probability that at least one TFBS occurs within a background sequence of length n [we set n = 500 as proposed by Rahmann *et al.* (31)]. Finally, the affinity score $\gamma(s_{ij}^W)$ of the subsequence is defined as $\gamma(s_{ij}^W) = 1 - \alpha_n(s_{ij}^W)$. Note that the affinity score is not dependent on the length of the PWM. Therefore, we can directly compare the affinity scores of different PWMs.

Clustering PWMs

We constructed three different kinds of PWM clusters for each TF. PWM cluster type 1 is constructed from all PWMs within the CC to which a query TF belongs. PWM cluster type 2 consists of all PWMs that are *directly* linked to the TF of interest. PWM cluster type 3 consists of high-quality PWMs, which were selected based on the following procedures. First, each PWM was assigned a quality score that quantifies how well a given PWM detects true binding sites over noisy sequences (see 'Quality score of PWMs' below). PWMs with low quality scores ($\leq$0.7) were discarded as described previously (26). Second, the PWM with the highest quality score among PWMs of PWM cluster type 2 was selected as the 'representative PWM'. Finally, the inter-motif distance between the representative PWM and the PWMs within the CC was calculated. PWMs with a distance above the cut-off value ($\geq$0.1653) were removed; the value of the distance cutoff was chosen as described previously (26). For the comparison of distance measurements between STAMP (32) and our used method (26), we tested the significance of overlap between two sets of top 9 ranked PWMs excluding the top ranked self PWMs in all 565 PWMs, from the default setting of STAMP and our PWM distance measure. The statistical significance was evaluated by calculating the enrichment *P*-value based on the hypergeometric distribution.

Quality score of PWMs

The quality score $Q(\Theta_1|\Theta_0)$ of a PWM Θ_1 quantifies how well the PWM is separated from a background model Θ_0 (31). We first define the type II error probability of the log-odds score at the given threshold t, given by:

$$\beta(t) = P_{\Theta_1}(x < t).$$

The quality score is then defined by:

$$Q(\Theta_1|\Theta_0) = 1 - \alpha_n(t^*),$$

where the threshold t^* is such that $\alpha_n(t) = \beta(t)$. The quality scores of all the PWMs used in this study are available in Supplementary Table S1A.

Construction of TFBS-Scanner

We reconstructed PWMs from TRANSFAC 9.4 (11) to evaluate the statistical significance of the predicted TFBS. Given a PWM cluster for a query TF, the TFBS-Scanner takes a DNA sequence as input to search for putative TFBSs of the TF. It then chooses a PWM cluster of the TF from the pre-compiled library of PWM clusters and searches all TFBSs of the chosen PWMs at the specified cutoff value of the affinity scores. Finally, all overlapping TFBSs of the PWMs are combined into one with the maximum affinity score.

Z-score of the AUC values

To account for the compositional bias of GC content within the regulatory regions, we defined the Z-score of the AUC value as:

$$z_{ij} = \frac{\mathrm{AUC}_{ij} - \mu_i}{\sigma_i}$$

where AUC_{ij} is the AUC value of the ith TF at the jth ChIP-Seq data set. Here, μ_i and σ_i are the mean and standard deviation of AUC values of the ith TF among the 51 ChIP-Seq data sets.

ChIP-Seq data sets used in this study

We compiled 51 ChIP-Seq data sets from the 'Table Browser' of the UCSC genome browser (33) or from other studies (34–36) (Supplementary Table S2D). To evaluate the classification performance of the TFBS-Scanner, we constructed 51 test sets consisting of positive and negative sets. For a positive set, we used the binding regions for each ChIP-Seq data set. We used, as suggested by Whitington *et al.* (23), the binding regions defined by the authors if the length of the binding region is larger than 500 bp. Otherwise, we expanded the binding regions to 500 bp. A negative set was also constructed by randomly sampling nine sequences for each binding region, following the method of Ernst *et al.* (16). The nine negative sequences were sampled from the non-gapped regions of the same chromosome as the binding region, excluding any overlapping sequences in the positive set.

A binary classifier for ChIP-Seq data sets

Given an input sequence s_i and a PWM cluster $\{\Theta_k\}$, the binary classifier for plotting the ROC curve is given by:

$$f(s_i|\{\Theta_k\}) = \mathrm{mean}_k \max_j \gamma(s_{ij}^{W_k}),$$

where W_k is the length of a PWM Θ_k and $\gamma(s_{ij}^{W_k})$ is the affinity score of subsequences of the input sequence.

Finding overrepresented motifs by MEME

To find the overrepresented motifs from ChIP-Seq data, we used the program MEME (37), which is one of the most popular motif discovery algorithms. We selected input sequences (500 bp) from the top 100 binding regions for each ChIP-Seq data set. Among a total of 51 ChIP-Seq data sets, we only considered 36 data sets released by ENCODE Yale TFBS because they provide *P*-values for each binding region. We used MEME with the following parameters: the distribution of motif occurrences: ZOOPS; the number of different motifs: 10;

the minimum motif width: 10; and the maximum motif width: 20. From the 10 learned motifs, we chose the one that was most similar to the representative PWM of TRANSFAC by visual inspection.

RESULTS

TFBS-Scanner: predicting TFBSs by clustered PWMs

To apply a motif-based scanning program to every TF with multiple position weight matrices (PWMs), we first collected the available information of PWMs and their interactions with cognate TFs. A total of 368 TFs and 565 vertebrate PWMs from the TRANSFAC database were considered; each PWM was linked to its associated TFs when the interaction was supported (see 'Materials and Methods' section). The resulting TF-PWM network is a bipartite graph whose nodes (368 TFs and 474 PWMs) are linked to each other (Supplementary Table S1A). The average number of TFs connected to a single PWM was 1.53, and a total of 61 (10.8%) PWMs had only one TF linkage (PWM:TF = 1:1). The V$EBOX_Q6_01 had the largest connection, with 30 different TFs (Figure 1A). Multiple PWMs were also found to be connected to single TFs; the average number of linked PWMs per TF was 2.35 (Figure 1B).

The TF-PWM network consists of 134 multiple connected components (CCs), defined by the TF-PWM subgraphs in which any two nodes within the same CC are connected to each other through one or more edges (Supplementary Table S1B, 'TF-PWM network.cys' in the supporting webpage). The largest CC, CC-3, is constructed with 69 nodes (36 TFs and 33 PWMs) and 98 edges, whereas the smallest one contains 2 nodes (1 TF and 1 PWM) and 1 edge (Figure 1C). Each CC elicits distinct connectivity between PWM and TF; the average number of PWMs linked to TFs and the average number of TFs connected to PWMs within each CC varies (Figure 1D).

Using 51 837 annotated protein–protein interactions (PPI) from the 'Interaction' information of the NCBI, physical interactions between TFs (Supplementary Table S1C) within each CC have been investigated (See 'Materials and Methods' section). We also checked that 80% of these TF–TF interactions overlapped with the human protein–protein interactions (Supplementary Table S1D), which were physically identified by the

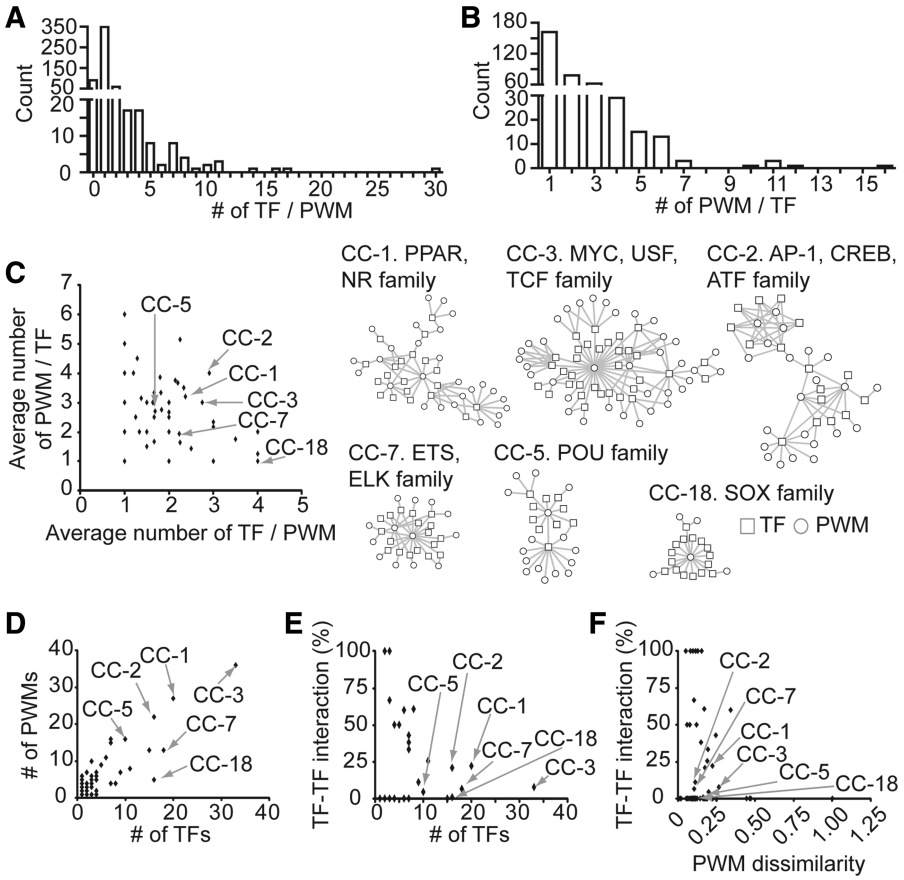

Figure 1. Properties of the TF-PWM network. Histograms of the number of TFs connected to a PWM ($N_{TF\text{-}PWM}$) (**A**) and the number of PWMs for each TF ($N_{PWM\text{-}TF}$) (**B**) in the TF-PWM network derived from TRANSFAC (11). (**C**) Average number of the $N_{TF\text{-}PWM}$ and $N_{PWM\text{-}TF}$ for each connected component (CC) in the TF-PWM network. Subgraphs of the representative CC (denoted as CC-#) are visualized by Cytoscape (29). (**D**) The total number of PWMs and TFs that form each CC are shown. (**E**) Degree of physical interaction among TFs belonging to each CC. TF–TF interaction (%) was calculated by dividing the number of the annotated interactions by the total number of possible interactions for each CC. (**F**) PWM dissimilarity, the mean value of the pairwise PWM–PWM dissimilarity for each CC.

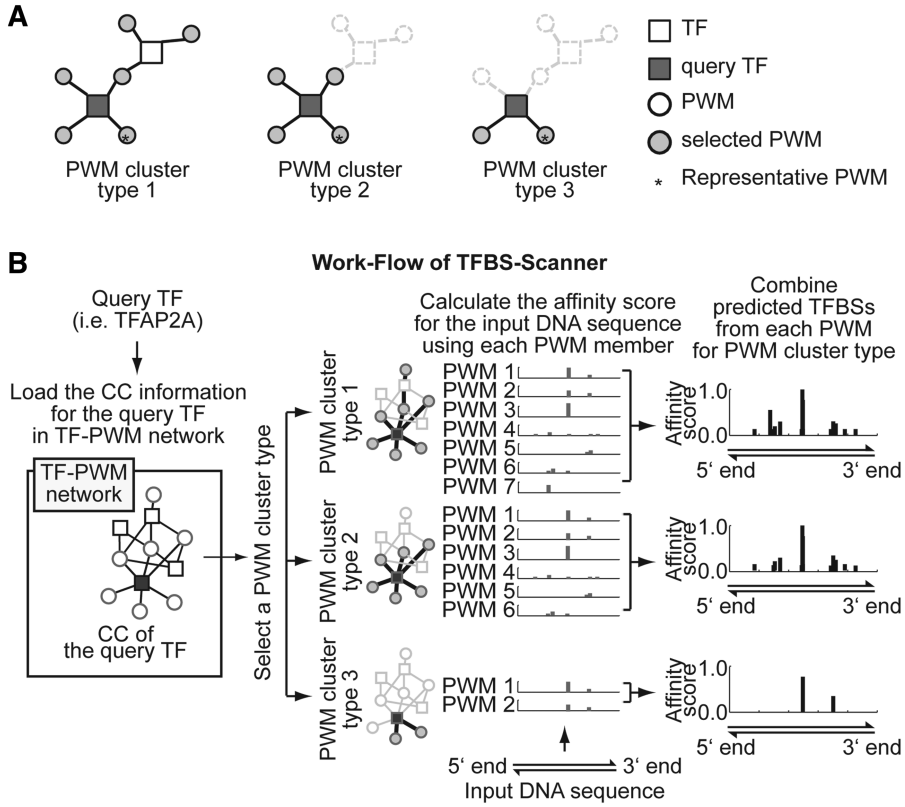

Figure 2. An overview of TFBS-Scanner. (**A**) Diagram of the PWM cluster types used in this analysis. For each query TF, PWM cluster type 1 (left) uses all PWMs that are connected in CC, PWM cluster type 2 (middle) uses PWMs that are directly linked to the query TF, and PWM cluster type 3 (right) uses PWMs that have a quality score and a similarity value higher than a threshold level. Selected PWMs for each PWM cluster type are marked by a gray filled circle. (**B**) Work-flow of TFBS-Scanner. It searches for the putative TFBSs of the query TF (i.e. TFAP2A) in a given DNA sequence. The input DNA sequence is searched to mark putative TFBSs using the selected PWMs, which depends on the chosen PWM cluster type. The predicted TFBSs are combined and accumulated affinity scores are calculated.

Mammalian Two-Hybrid Assay (38). Multiple TFs found in the same CC are likely to interact with each other when the total number of TFs in the CC is relatively small (Figure 1E). In contrast, as the number of TFs in the CC increases, their physical interactions become weaker. In the case of CC-1, which consists of 20 TFs, including the PPAR and NR family members, only 42 protein–protein interactions have been annotated out of 190 possible interactions (Supplementary Figure S1, 'PPI in TF-PWM.cys' in the supporting webpage). The sequence similarity among multiple PWMs in the same CC was then measured. Interestingly, DNA sequences of the PWM within each CC were quite similar, which was not significantly affected by physical interactions between TFs in the same CC (Figure 1F). These data indicate that the TFs in the same CC might be able to recognize similar DNA sequences.

Analysis of the TF-PWM network clearly shows that most of the TFs are connected to multiple PWMs. Therefore, to examine the effect of PWM clustering on the efficiency of TFBS prediction, we compared three different types of PWM clusters (Figure 2A, Supplementary Table S2A-C; see 'Materials and Methods' section). The TFBS-Scanner takes a DNA sequence as input to search putative TFBSs of any given TFs (Figure 2B). For each query TF, TFBS-Scanner searches a pre-compiled library of the TF-PWM network and loads multiple PWMs for each cluster type. Using the selected set of PWMs, TFBS-Scanner screens the query DNA sequence for TFBSs with affinity scores higher than a pre-defined cut-off value. In the end, overlapping sites among predicted TFBSs are combined to calculate the maximum affinity score.

Validation

To examine the efficacy of TFBS-Scanner, we used 51 genome-wide ChIP-Seq data sets analyzing 26 vertebrate TFs (Supplementary Table S2D). For each ChIP-Seq data set, the positive (derived from the peak regions) and the negative (background) set of DNA sequences were prepared (see 'Materials and Methods' section). The performance of TFBS-Scanner was evaluated based on the standard receiver operating characteristic (ROC) analysis, which plots a ROC curve by varying the threshold of the output of a binary classifier. We designed a binary classifier to score each input sequence by averaging the maximum affinity scores of subsequences among the given PWM cluster (see 'Materials and Methods' section). To quantify the performance, the area under the ROC curve (AUC) was computed to range from 0 to 1 (for

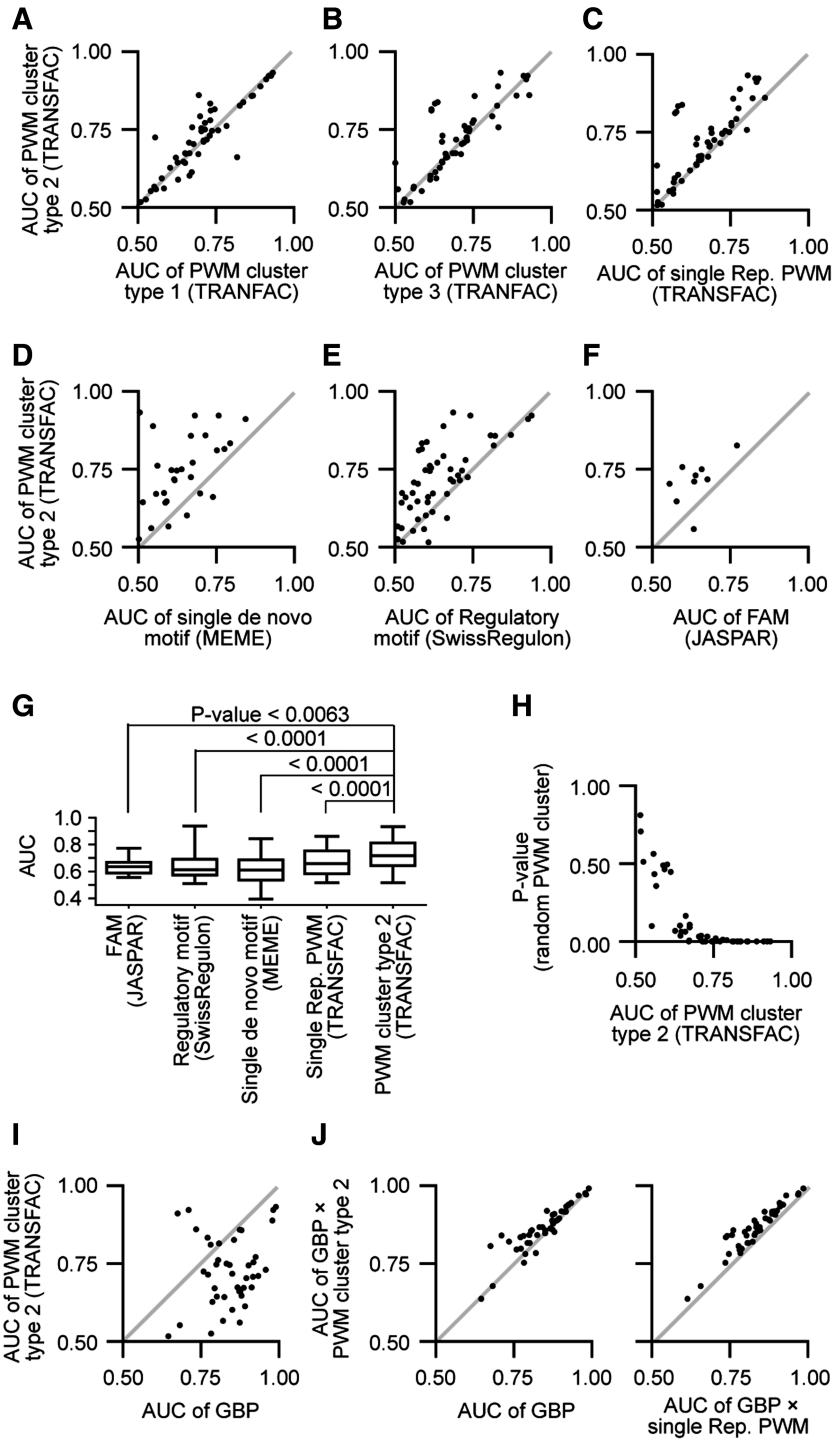

Figure 3. Comparison of the TFBS prediction tools. (**A–F**) Performance comparison between PWM cluster type 2 and PWM cluster type 1 (**A**), PWM cluster type 3 (**B**), single representative PWM (**C**), single *de novo* motif (**D**), regulatory motif from SwissRegulon (**E**) and FAM from JASPAR (**F**). To directly compare the prediction performance between methods, the AUC (Area under curve of an ROC curve) was calculated using the 51 ChIP-Seq data sets (Supplementary Table S2D). (**G**) Performance comparison of TFBS-Scanner to other conventional prediction tools. A box plot shows the AUC values from each tool for the tested ChIP-Seq data sets. *P*-values were calculated using paired *t*-test. (**H**) AUC scores of PWM cluster type 2 were compared to randomly generated PWM cluster, and the empirical *P*-values were shown as a scatter plot. (**I**) Performance comparison between GBP and PWM cluster type 2. (**J**) Performance of the integrated analysis (PWM cluster type 2 plus GBP) was compared to GBP alone or representative PWM plus GBP.

perfect classification). Because a classifier randomly guessing the class label of an input has an AUC value of 0.5, AUC values less than 0.5 have no practical meaning.

Of the 26 TFs tested, PWM cluster type 2 performed better than types 1 and 3 (Figure 3A and B). In addition, prediction efficiency of PWM cluster type 2 was higher than that of single representative PWM (Supplementary Table S2A–C, Figure 3C). Since PWMs were originally built using a limited number of experimentally verified TFBSs, the number of the real TFBSs used to create each PWM is usually biased. It raises the concern that not every PWM might be able to represent real binding *in vivo*. Therefore, we built the single PWMs from the ChIP-Seq data set using the *de novo* motif discovery algorithm MEME (37), and compared the AUC values of the derived PWMs with that of PWM cluster type 2 (Figure 3D). It is noteworthy to mention that the dissimilarity value of the clustered PWMs of TFs used in our analysis was significantly lower than all PWM clusters, which might contribute to enhanced TFBS prediction (Supplementary Figure S2A). To evaluate the distance measure which we used for PWM similarity calculation, we compared results from STAMP (32) and our used method by calculating the enrichment *P*-value based on the hypergeometric distribution (Supplementary Figure S2B, See 'Materials and Methods' section). We observed little differences between the two distance measures.

To evaluate the performance of our PWM clustering methods, we then compared the efficiency of PWM cluster type 2 to other PWM clusters provided by SwissRegulon (39) (Figure 3E), or by the JASPAR FAM database (40) (Figure 3F). The PWM cluster type 2 showed significantly improved performance compared to the SwissRegulon ($P < 0.0001$) or to the JASPAR FAM database, which provides 11 FBPs based on the structural classification of TFs ($P < 0.0063$; Figure 3G). To test the significance of clustering PWMs, we then generated random clusters for each ChIP-Seq data set by selecting the same number of PWMs at random (simulated by 20 000 times), and computed empirical *P*-values of AUC scores. The PWM cluster type 2 showed significant AUC scores for most of the tested ChIP-Seq data sets except for data sets with low AUC scores (Figure 3H). These results indicate that the PWM clustering is useful to improve the efficiency of TFBS prediction.

Both experimental and computational analysis (16) recently reported that TFBSs are preferentially located in the genomic regions scored by general binding preference (GBP). It uses integrated heterogeneous data of epigenetic (22,23), DNase-I hypersensitivity (19,20), and sequence conservation among species. Peak loci for most of the tested ChIP-Seq data sets were largely located in the GBP scored regions with higher AUC score (Figure 3I). By combining the affinity score of TFBS-Scanner with the GBP score, prediction performance was significantly improved (Figure 3J), indicating that integration of the PWM clustering method along with heterogeneous motif-independent data set will be more powerful for TFBS prediction.

Searching co-occurring *cis*-motifs using TFBS-Scanner

Because it is of great interest to understand the TF complexes that cooperatively act in the promoter or enhancer regions (5,41), numerous computational tools have been generated to model representative multiple motifs or regulatory modules that present in the promoter sequences that elicit correlation with their gene expression (42–47). Therefore, we asked whether TFBS-Scanner could be used to determine co-occurring motifs of TFs using ChIP-Seq information, without integration of the gene expression data. For this purpose, we developed a simple but effective computational tool based on clustered PWMs.

If TF-Y binding to DNA is sequence specific and it cooperatively acts with TF-X, then a statistically significant portion of the binding sites for TF-Y will appear in the binding regions of the TF-X. Based on this assumption, we searched for binding sites that frequently co-occurred in the ChIP-Seq data set of TF-X. For this purpose, the AUC values for 368 TFs in the TF-PWM library were first computed using 51 ChIP-Seq data sets. However, due to the compositional bias of GC contents in the regulatory regions (48), the higher mean percentage of GC in each PWM cluster arbitrarily shows higher AUC values (Figure 4A). As a result, PWMs with high GC content were found to be highly enriched for most of the ChIP-Seq data sets (Figure 4B). To solve this problem, the AUC values were converted to a Z-score, where the sample mean and standard deviation of the AUC values are computed for each TF among a total of 51 ChIP-Seq data sets. The Z-score was not affected by the GC contents of the PWM clusters (Figure 4C and D).

Modeling co-occurring *cis*-motifs that are conserved in the diverse cell type: a case of E2F family

Family members of the E2 factor (E2F) family play diverse roles in the transcriptional regulation of cellular proliferation and differentiation. Because E2F family members share a high degree of structural and biochemical similarity and recognize similar E2F DNA sequences (49), it was of interest to identify the co-occurring motifs in the E2F occupied genomic loci. For this purpose, four ChIP-Seq data sets were analyzed in the diverse cellular system using either E2F1 or E2F4 (Supplementary Table S2D). In every assay, binding sequences specific for E2F were highly enriched at the peak loci, as determined by TFBS-Scanner using PWM cluster type 2 (Figure 5A). To identify the co-occurring DNA motifs in the E2F binding regions, Z-scores of the 368 TFs were then calculated (Figure 5B and Supplementary Table S3). As expected, TFBSs of the E2F1, E2F2, E2F3, E2F4, E2F5 and E2F7 exhibited the highest Z-score and the distribution of the normalized cumulative affinity score of these sites in the binding loci overlapped (Supplementary Figure S3). Along with E2F, PWMs of distinct TFs, such as AHR, ARNT, ETF, HIC1, EGR, NRF1, SP1 or TEAD, were calculated to have higher Z-scores than the cut-off value. Although binding sequences for these TFs were quite dissimilar to that of E2Fs (Supplementary Figure S4), distribution patterns of the normalized cumulative affinity

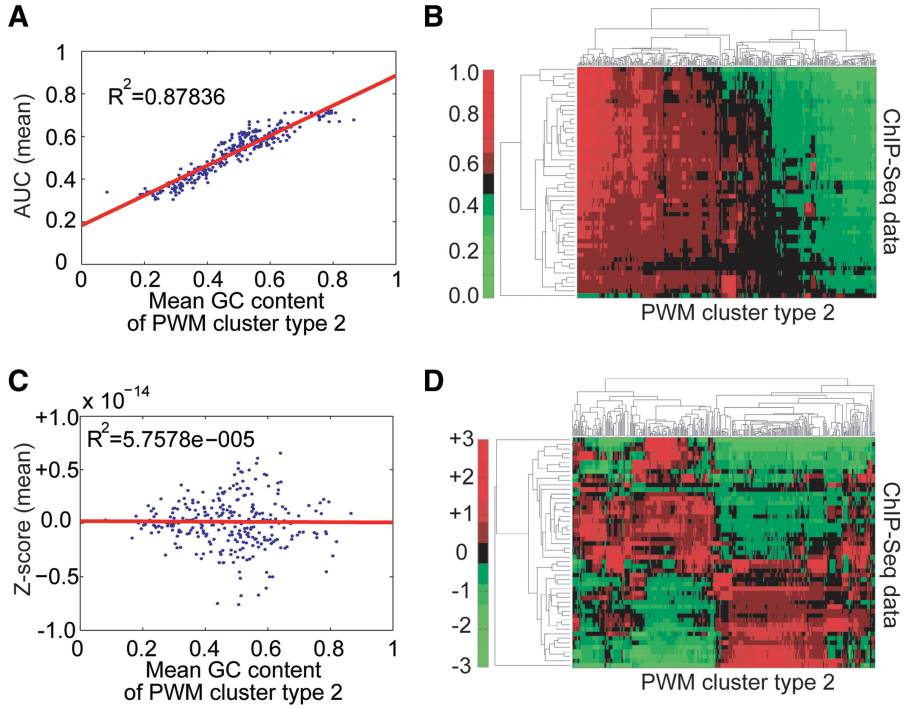

Figure 4. Conversion of AUC to Z-scores. (**A**) Scatter plot showing the positive correlation between the mean AUC values and the mean GC content of PWM clusters (PWM cluster type 2). For each PWM cluster, the mean value of the AUC is taken over the 51 ChIP-Seq data sets. The mean GC content of a PWM cluster is computed by averaging the sum of the frequency of 'G' and 'C' over all the PWMs belonging to the PWM cluster. Red is the best-fit linear regression line with a positive slope. (**B**) Cluster analysis of 368 PWM clusters (PWM cluster type 2) from 51 ChIP-Seq data sets using AUC values. (**C**) Scatter plot of the mean Z-score and the mean GC content of PWM clusters. (**D**) Cluster analysis of 368 PWM clusters (PWM cluster type 2) from 51 ChIP-Seq data sets using Z-scores.

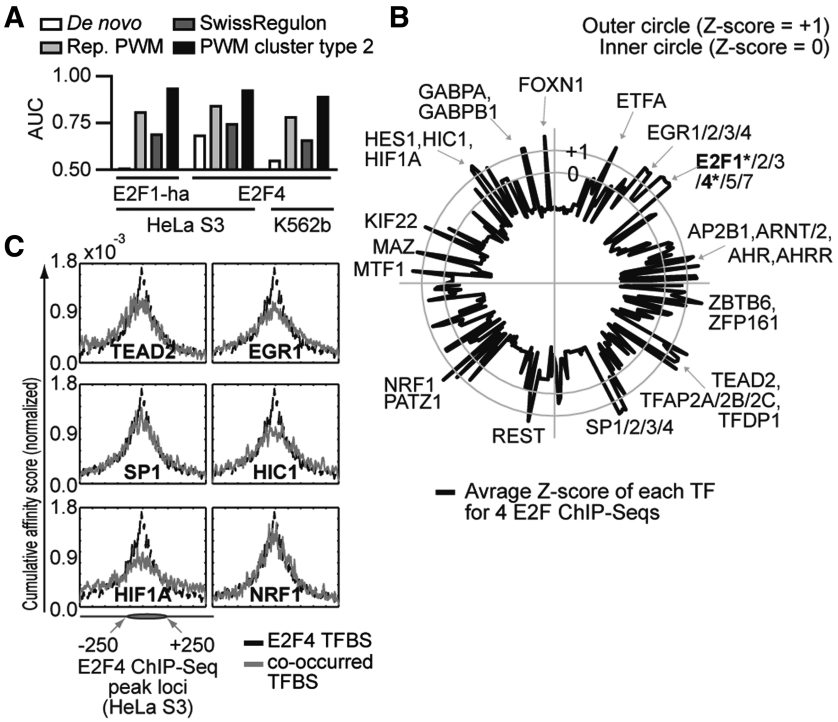

Figure 5. Co-occurring motifs of E2F1 and E2F4. (**A**) The AUC (Area under curve of an ROC curve) for the E2F1 and E2F4 ChIP-Seq data sets using different PWMs or PWM clusters are shown. (**B**) StarGlyph Z-score distribution of the binding sites for 368 TFs in the E2F ChIP peak loci, where the TFs are listed in alphabetical order. +1 Z-score is indicated in the outer circle, and zero Z-score in the inner circle. The query TF is marked by bold characters and an asterisks. (**C**) Using the E2F4 ChIP-Seq data set, the occurrence of the predicted TFBSs with high Z-scores are shown along with E2F4 TFBS in the E2F4 binding region. The affinity scores of the predicted TFBSs were normalized by dividing the sum of the affinity scores along the defined region (–1000 bp to approximately +1000 bp).

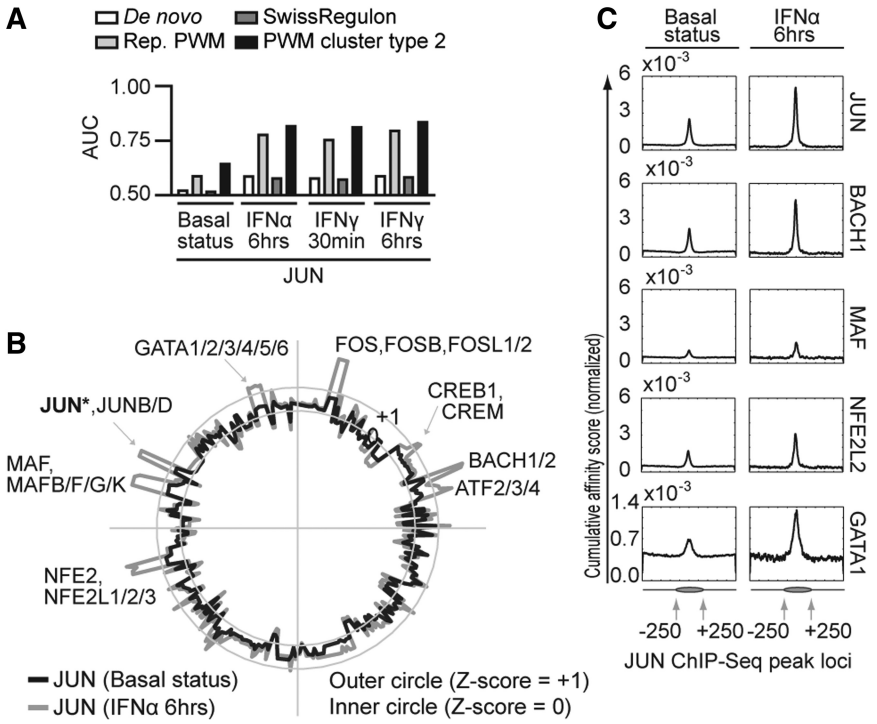

Figure 6. Condition specific co-occurring motifs of JUN. (**A**) The AUC (Area under curve of an ROC curve) for the JUN ChIP-Seq data sets using different PWMs or PWM clusters are shown. (**B**) StarGlyph distribution of Z-scores for 368 TFs from the JUN ChIP-Seq data sets for basal status or IFNα stimulation. (**C**) The motifs for JUN and other TFs with high Z-scores are enriched at the center of the JUN ChIP peak loci, which became evident upon stimulation.

score of these TFs in the ChIP-Seq peak loci were quite similar to that of E2F (Figure 5C), indicating that these TFs might act in conjunction with E2F for target gene regulation. In support of this prediction, functional interactions between E2F and AHR, ARNT, HIC1, EGR, SP1 and SP3 for transcriptional regulation have been previously reported (50–56). In addition, TFBSs for NRF1, TEAD2, and ETF are known to co-occur with that of the E2F family in their target genes (57,58).

Modeling signal-specific co-occurring *cis*-motifs: a case of JUN family

We next asked whether our method could be used to identify signal-specific co-occurring motifs. For this purpose, we chose four sets of ChIP-Seq data assayed with JUN in K562 cells after IFNα or IFNγ stimulation (Supplementary Table S2D). JUN is a basic region-leucine zipper protein and belongs to the family of AP-1, which consists of JUN, FOS, MAF and ATF subfamilies (59). Activated AP-1 recognizes and binds to TGA(C/G)TCA with high affinity or TGACGTCA with low affinity, and acts as a regulatory factor of diverse cellular processes, such as proliferation, transformation and apotosis (59). Cooperative interactions between AP-1 and other TFs in the promoter have been suggested as the regulatory mechanism of target gene specificity governed by AP-1 (60).

For un-stimulated K562 cells, JUN binding loci did not possess prominent JUN binding sequences, as judged by conventional SwissRegulon or by TFBS-Scanner using clustered PWMs (Figure 6A). However, upon stimulation with IFNα, binding sequences for JUN became dominant in the peak loci of JUN ChIP. Along with JUN, motifs for BACH1, FOS, GATA, MAF or NFE2 also exhibited significantly higher Z-scores than the cut-off value, which became prominent after stimulation (Figure 6B andSupplementary Table S3). The distribution of the normalized affinity scores of these TFs overlapped with that of JUN, and their co-occurrence became stronger after stimulation with IFNα (Figure 6C). In the JUN ChIP-Seq data sets for IFNγ stimulation, major co-occurring motifs were very similar to that of IFNα stimulation (Supplementary Table S3). These data suggest that JUN might coordinate with BACH1, FOS, GATA, MAF or NFE2L2 for gene regulation specific for IFNγ or IFNα· In support of this prediction, inter-relation between the AP-1 motif and *cis*-element of GATA or NFE have been previously reported (61–63). Similar to JUN, binding sites for MYC were significantly enriched at the peak loci of the MYC ChIP-Seq after IFNα or IFNγ stimulation, along with co-occurring motifs of NHMHB2/3, HAND1/2, HMX, MAX, MYB, USF1/2 or XBP1 (Supplementary Figure S5). To evaluate the performance of clustered PWM over single PWM, we finally compared the Z-score of PWM cluster type 2 with that of the representative PWM (Supplementary Figure S6). In the both E2F1 and JUN ChIP-Seq data set, PWM clustering usually showed better performance than the representative PWM. These results suggest that our statistical framework using clustered PWMs is useful at identifying co-occurring motifs of TFs.

DISCUSSION

The code for gene regulation is widely accepted to be encoded in the four-letter alphabet of the genome, but decoding the rules for the transcription regulatory circuits that govern diverse cellular responses remains challenging. To reconstruct the transcriptional regulatory network, an important intermediate step is to efficiently predict the binding sites of TFs in the regulatory regions. In this article, we developed a computational framework for predicting the putative TFBSs for a group of TFs for a DNA sequence of choice. Our approach uses modified PWMs based on a previous statistical framework (26) and clustered PWMs based on their sequence similarity and quality. The TFBS-Scanner utilizes 368 TFs and 565 PWMs data sets, which are available in an up-to-date public database.

The performance of most conventional TFBS prediction tools heavily depends on the selection of a PWM that represents a genuine motif for a given TF. However, without a reference set of binding for TFs, it is impossible to identify the best-performing PWMs from pre-existing pools. Furthermore, accumulating evidence suggests that a number of TFs can be associated with more than one binding motif, and a single PWM can be recognized by functionally distinct TFs (5,8). Therefore, conventional motif scanning programs, which assume a single PWM for a single TF, have a fundamental problem. To maximize prediction efficiency, we strategically developed a quality scoring and clustering tool for PWMs. Using the genome-wide ChIP binding peaks, we were able to identify the set of the reference PWM for a given TF and a PWM clustering type that maximizes prediction efficiency. Findings of this article support the idea that approaches combining multiple PWMs associated with a given TF outperform algorithms relying on a single PWM. In addition, it is noteworthy to mention that integration of the GBP score (16) further improved the prediction efficiency of the TFBS-Scanner.

In general, there is growing evidence to support the view that a single motif cannot explain all *in vivo* binding regions (5). In addition, *in vitro* binding of the purified TF protein demonstrates that it can recognize more than one species of the binding motifs (8). Therefore, using a set of qualified PWMs rather than a single PWM to identify *in vivo* binding regions should be more effective and practical. However, it is worth noting that our clustered PWMs cannot detect binding regions of TFs that are indirectly bound or that have bindings dependent on other factors that act in neighboring regions. Although a TF has the ability to recognize DNA sequences directly, not every TF solely depends on direct interaction with DNA sequences *in vivo*. Instead, protein-aided recruitment by the TF adjacent to other TFs may facilitate the formation of stable transcription regulatory complexes (41,64,65). In recent reports, overlapping localization on the genome loci and combinatorial interactions among TFs were strongly suggested as one of the features that explains complex transcriptional regulation of tissue or condition specificity (66,67).

To identify co-occurring motifs of gene regulation, several approaches have been proposed. Identification tools of the enriched motifs from ChIP-chip or ChIP-Seq data sets can be grouped into two classes based on their underlying assumptions. The methods belonging to the first class are based on the assumption that a true motif is located at the centers of the binding regions, whereas insignificant motifs are uniformly distributed (68). However, these approaches require a cutoff to scan the TFBSs of a predefined motif. The second class, which is based on the assumption that a true motif should be overrepresented in the binding regions compared to background sequences, overcomes the drawback of the cutoff for scanning TFBSs by using standard ROC analysis (69,70). Although these approaches can reduce the percentages of false positives by taking into account the GC content of background sequences, it could actually filter out a true GC-rich motif. In contrast to these approaches, our method randomly selects the background sequences without considering the GC content, and then eliminates false positive GC-rich motifs by normalizing the AUC scores. Because our method utilizes Z-scores to select statistically significant motifs, its efficiency depends on the size of the ChIP-Seq data sets to approximate the exact distributions of the AUC scores of each TF.

The genome encodes the blueprints of every possible transcriptional regulatory network. A condition-specific regulatory network of transcription emerges from multiple protein–protein and protein–DNA interactions, activated by specific signaling pathways. Our computational approach provides the groundwork to build a map of these potential networks. Accompanied by genome-wide information of all protein–protein and protein–DNA interactions, our program will serve as a helpful tool to reconstruct the functional network that governs specific cellular responses.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–3, Supplementary Figures 1–6.

FUNDING

Funding for open access charge: National Research Foundation of Korea (NRF) grants funded by the Korean Ministry of Education, Science and Technology (KRF-2008-313-C00604, 2010-0028453, 2011-0018012) and Regional Core Research Program/Anti-aging and Well-being Research Center; WCU Program (Project No. R31-10100); National Core Research Center for Systems Bio-Dynamics (2010-0028447). TJ Park Postdoctoral Fellowship (to J.K.K.).

Conflict of interest statement. None declared.

REFERENCES

1. Brivanlou,A.H. and Darnell,J.E. Jr (2002) Signal transduction and the control of gene expression. *Science*, **295**, 813–818.

2. Emerson,B.M. (2002) Specificity of gene regulation. *Cell*, **109**, 267–270.
3. Spiegelman,B.M. and Heinrich,R. (2004) Biological control through regulated transcriptional coactivators. *Cell*, **119**, 157–167.
4. Hawkins,R.D., Hon,G.C. and Ren,B. (2010) Next-generation genomics: an integrative approach. *Nat. Rev. Genet.*, **11**, 476–486.
5. Farnham,P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, **10**, 605–616.
6. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
7. Hannenhalli,S. (2008) Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, **24**, 1325–1331.
8. 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.
9. Ehret,G.B., Reichenbach,P., Schindler,U., Horvath,C.M., Fritz,S., Nabholz,M. and Bucher,P. (2001) DNA binding specificity of different STAT proteins. Comparison of in vitro specificity with natural target sites. *J. Biol. Chem.*, **276**, 6675–6688.
10. 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.
11. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
12. Pachkov,M., Erb,I., Molina,N. and van Nimwegen,E. (2007) SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res.*, **35**, D127–D131.
13. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
14. Frith,M.C., Spouge,J.L., Hansen,U. and Weng,Z. (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.*, **30**, 3214–3224.
15. Cartharius,K., Frech,K., Grote,K., Klocke,B., Haltmeier,M., Klingenhoff,A., Frisch,M., Bayerlein,M. and Werner,T. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**, 2933–2942.
16. Ernst,J., Plasterer,H.L., Simon,I. and Bar-Joseph,Z. (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Gen. Res.*, **20**, 526–536.
17. Loots,G. and Ovcharenko,I. (2007) ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. *Bioinformatics*, **23**, 122–124.
18. Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.*, **26**, 225–228.
19. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
20. John,S., Sabo,P.J., Thurman,R.E., Sung,M.H., Biddie,S.C., Johnson,T.A., Hager,G.L. and Stamatoyannopoulos,J.A. (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.*, **43**, 264–268.
21. Narlikar,L., Gordan,R. and Hartemink,A.J. (2007) A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.*, **3**, e215.
22. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.
23. Whitington,T., Perkins,A.C. and Bailey,T.L. (2009) High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res.*, **37**, 14–25.
24. Kielbasa,S.M., Gonze,D. and Herzel,H. (2005) Measuring similarities between transcription factor binding sites. *BMC Bioinformatics*, **6**, 237.
25. Schones,D.E., Sumazin,P. and Zhang,M.Q. (2005) Similarity of position frequency matrices for transcription factor binding sites. *Bioinformatics*, **21**, 307–313.
26. Oh,Y.M., Kim,J.K., Choi,Y., Choi,S. and Yoo,J.Y. (2009) Prediction and experimental validation of novel STAT3 target genes in human cancer cells. *PLoS One*, **4**, e6911.
27. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
28. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Federhen,S. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.
29. Killcoyne,S., Carter,G.W., Smith,J. and Boyle,J. (2009) Cytoscape: a community-based framework for network modeling. *Methods Mol. Biol.*, **563**, 219–239.
30. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B., Yoo,J. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
31. Rahmann,S., Muller,T. and Vingron,M. (2003) On the power of profiles for transcription factor binding site detection. *Stat. Appl. Genet. Mol. Biol.*, **2**, Article7.
32. Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.*, **35**, W253–W258.
33. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
34. Wederell,E.D., Bilenky,M., Cullum,R., Thiessen,N., Dagpinar,M., Delaney,A., Varhol,R., Zhao,Y., Zeng,T., Bernier,B. *et al.* (2008) Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res.*, **36**, 4549–4564.
35. 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.
36. Marson,A., Levine,S.S., Cole,M.F., Frampton,G.M., Brambrink,T., Johnstone,S., Guenther,M.G., Johnston,W.K., Wernig,M., Newman,J. *et al.* (2008) Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, **134**, 521–533.
37. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
38. 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.
39. Pachkov,M., Erb,I., Molina,N. and van Nimwegen,E. (2007) SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res.*, **35**, D127–D131.
40. Sandelin,A. and Wasserman,W.W. (2004) Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J. Mol. Biol.*, **338**, 207–215.
41. Panne,D., Maniatis,T. and Harrison,S.C. (2007) An atomic model of the interferon-beta enhanceosome. *Cell*, **129**, 1111–1123.
42. Yu,X., Lin,J., Zack,D.J. and Qian,J. (2006) Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.*, **34**, 4925–4936.
43. Aerts,S., Van Loo,P., Thijs,G., Moreau,Y. and De Moor,B. (2003) Computational detection of cis -regulatory modules. *Bioinformatics*, **19(Suppl. 2)**, ii5–ii14.
44. Makeev,V.J., Lifanov,A.P., Nazina,A.G. and Papatsenko,D.A. (2003) Distance preferences in the arrangement of binding motifs

and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res.*, **31**, 6016–6026.
45. Yu,X., Lin,J., Masuda,T., Esumi,N., Zack,D.J. and Qian,J. (2006) Genome-wide prediction and characterization of interactions between transcription factors in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **34**, 917–927.
46. Hu,J., Hu,H. and Li,X. (2008) MOPAT: a graph-based method to predict recurrent cis-regulatory modules from known motifs. *Nucleic Acids Res.*, **36**, 4488–4497.
47. Ho Sui,S.J., Mortimer,J.R., Arenillas,D.J., Brumm,J., Walsh,C.J., Kennedy,B.P. and Wasserman,W.W. (2005) oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res.*, **33**, 3154–3164.
48. Ho,J.W., Bishop,E., Karchenko,P.V., Negre,N., White,K.P. and Park,P.J. (2011) ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis. *BMC Genomics*, **12**, 134.
49. Dimova,D.K. and Dyson,N.J. (2005) The E2F transcriptional network: old acquaintances with new faces. *Oncogene*, **24**, 2810–2826.
50. Azkargorta,M., Fullaondo,A., Laresgoiti,U., Aloria,K., Infante,A., Arizmendi,J.M. and Zubiaga,A.M. (2010) Differential proteomics analysis reveals a role for E2F2 in the regulation of the Ahr pathway in T lymphocytes. *Mol. Cell. Proteomics*, **9**, 2184–2194.
51. Elena,C. and Banchio,C. (2010) Specific interaction between E2F1 and Sp1 regulates the expression of murine CTP:phosphocholine cytidylyltransferase alpha during the S phase. *Biochim. Biophys. Acta*, **1801**, 537–546.
52. Rotheneder,H., Geymayer,S. and Haidweger,E. (1999) Transcription factors of the Sp1 family: interaction with E2F and regulation of the murine thymidine kinase promoter. *J. Mol. Biol.*, **293**, 1005–1015.
53. Usskilat,C., Skerka,C., Saluz,H.P. and Hanel,F. (2006) The transcription factor Egr-1 is a regulator of the human TopBP1 gene. *Gene*, **380**, 144–150.
54. Zhang,B., Chambers,K.J., Leprince,D., Faller,D.V. and Wang,S. (2009) Requirement for chromatin-remodeling complex in novel tumor suppressor HIC1-mediated transcriptional repression and growth control. *Oncogene*, **28**, 651–661.
55. Zhang,H.J., Li,W.J., Yang,S.Y., Li,S.Y., Ni,J.H. and Jia,H.T. (2009) 8-Chloro-adenosine-induced E2F1 promotes p14ARF gene activation in H1299 cells through displacing Sp1 from multiple overlapping E2F1/Sp1 sites. *J. Cell. Biochem.*, **106**, 464–472.
56. Zhao,L. and Chang,L.S. (1997) The human POLD1 gene. Identification of an upstream activator sequence, activation by Sp1 and Sp3, and cell cycle regulation. *J. Biol. Chem.*, **272**, 4869–4882.
57. Cam,H., Balciunaite,E., Blais,A., Spektor,A., Scarpulla,R.C., Young,R., Kluger,Y. and Dynlacht,B.D. (2004) A common set of gene regulatory networks links metabolism and growth inhibition. *Mol. Cell*, **16**, 399–411.
58. Zellmer,S., Schmidt-Heck,W., Godoy,P., Weng,H., Meyer,C., Lehmann,T., Sparna,T., Schormann,W., Hammad,S., Kreutz,C. *et al.* (2010) Transcription factors ETF, E2F, and SP-1 are involved in cytokine-independent proliferation of murine hepatocytes. *Hepatology*, **52**, 2127–2136.
59. Shaulian,E. and Karin,M. (2002) AP-1 as a regulator of cell life and death. *Nat. Cell Biol.*, **4**, E131–E136.
60. Chinenov,Y. and Kerppola,T.K. (2001) Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity. *Oncogene*, **20**, 2438–2452.
61. Walters,M. and Martin,D.I. (1992) Functional erythroid promoters created by interaction of the transcription factor GATA-1 with CACCC and AP-1/NFE-2 elements. *Proc. Natl Acad. Sci. USA*, **89**, 10444–10448.
62. Yeligar,S.M., Machida,K. and Kalra,V.K. (2010) Ethanol-induced HO-1 and NQO1 are differentially regulated by HIF-1alpha and Nrf2 to attenuate inflammatory cytokine expression. *J. Biol. Chem.*, **285**, 35359–35373.
63. Kim,H., Jung,Y., Shin,B.S., Song,H., Bae,S.H., Rhee,S.G. and Jeong,W. (2010) Redox regulation of lipopolysaccharide-mediated sulfiredoxin induction, which depends on both AP-1 and Nrf2. *J. Biol. Chem.*, **285**, 34419–34428.
64. Kiuchi,N., Nakajima,K., Ichiba,M., Fukada,T., Narimatsu,M., Mizuno,K., Hibi,M. and Hirano,T. (1999) STAT3 is required for the gp130-mediated full activation of the c-myc gene. *J. Exp. Med.*, **189**, 63–73.
65. Dooley,K.A., Millinder,S. and Osborne,T.F. (1998) Sterol regulation of 3-hydroxy-3-methylglutaryl-coenzyme A synthase gene through a direct interaction between sterol regulatory element binding protein and the trimeric CCAAT-binding factor/nuclear factor Y. *J. Biol. Chem.*, **273**, 1349–1356.
66. Jin,F., Li,Y., Ren,B. and Natarajan,R. (2011) PU.1 and C/EBP{alpha} synergistically program distinct response to NF-{kappa}B activation through establishing monocyte specific enhancers. *Proc. Natl Acad. Sci. USA*, **108**, 5290–5295.
67. Kuo,D., Licon,K., Bandyopadhyay,S., Chuang,R., Luo,C., Catalana,J., Ravasi,T., Tan,K. and Ideker,T. (2010) Coevolution within a transcriptional network by compensatory trans and cis mutations. *Genome Res.*, **20**, 1672–1678.
68. Lupien,M., Eeckhoute,J., Meyer,C.A., Wang,Q., Zhang,Y., Li,W., Carroll,J.S., Liu,X.S. and Brown,M. (2008) FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, **132**, 958–970.
69. Tuteja,G., Jensen,S.T., White,P. and Kaestner,K.H. (2008) Cis-regulatory modules in the mammalian liver: composition depends on strength of Foxa2 consensus site. *Nucleic Acids Res.*, **36**, 4149–4157.
70. Gordan,R., Hartemink,A.J. and Bulyk,M.L. (2009) Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res.*, **19**, 2090–2100.