Published online 18 September 2006 *Nucleic Acids Research, 2006, Vol. 34, No. 17* ***4925–4936***
doi:10.1093/nar/gkl595

Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues

Xueping Yu[1], Jimmy Lin[1], Donald J. Zack[1,2,3,4] and Jiang Qian[1,*]

[1]Wilmer Institute, [2]Department of Molecular Biology and Genetics, [3]Department of Neuroscience and [4]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Maumenee Building 844, 600 N. Wolfe Street, Baltimore, MD 21287, USA

Received April 25, 2006; Revised July 13, 2006; Accepted August 1, 2006

ABSTRACT

Tissue-specific gene expression is generally regulated by more than a single transcription factor (TF). Multiple TFs work in concert to achieve tissue specificity. In order to explore these complex TF interaction networks, we performed a large-scale analysis of TF interactions for 30 human tissues. We first identified tissue-specific genes for 30 tissues based on gene expression databases. We then evaluated the relationships between TFs using the relative position and co-occurrence of their binding sites in the promoters of tissue-specific genes. The predicted TF–TF interactions were validated by both known protein–protein interactions and co-expression of their target genes. We found that our predictions are enriched in known protein–protein interactions (>80 times that of random expectation). In addition, we found that the target genes show the highest co-expression in the tissue of interest. Our findings demonstrate that non-tissue specific TFs play a large role in regulation of tissue-specific genes. Furthermore, they show that individual TFs can contribute to tissue specificity in different tissues by interacting with distinct TF partners. Lastly, we identified several tissue-specific TF clusters that may play important roles in tissue-specific gene regulation.

INTRODUCTION

One of the fundamental questions in biology is to understand how different tissues achieve specificity. Given the same DNA template, how are different tissue types determined? What are the different genes that are expressed and how are these different genes regulated in different tissues? With current high throughput technology, researchers can now measure gene expressions in various tissues on a large scale (1,2). However, it is still a challenge to understand the intricate and complex control of these genes.

There are more than 25 000 genes in the human genome, and they demonstrate dramatic diversity in terms of expression levels and tissue expression patterns. Despite this tremendous diversity, all genes are controlled by <2000 tranranscription factors (TFs) (3). This limited set of TFs is thought to be able to control the larger set of expression patterns through combinatorial regulation, in which multiple factors work in combination to control individual genes.

To study tissue-specific gene expression, Wasserman and colleagues employed the concept of a regulatory module (cluster of TF binding sites) to predict muscle- and liver-specific regulatory regions (4,5). Using known tissue-specific TFs based on experimental evidence, they were able to not only recover many known tissue-specific regulatory regions, but also predict novel genes that contribute to tissue specificity. The idea of regulatory module has also been applied to study of gene regulation in fly development (6).

Despite the success of these approaches, they cannot be applied on a large scale to many tissues due to the limited state of our current knowledge about TFs. One requirement of these methods is to have a list of TFs that are known to be relevant to the tissue of interest. For example, the analysis of liver specific gene regulation depended upon a priori knowledge about six TFs with experimentally determined roles in liver gene expression (5). Biological knowledge on individual tissues is crucial to the quality of *in silico* prediction of tissue-specific gene regulation. Unfortunately, current knowledge of TFs that contribute to the tissue-specificity is limited, and this in turn has limited the large scale bioinformatic study of tissue-specific gene regulation.

To circumvent this limitation, we have been working to develop computational methods to analyze tissue-specific gene regulation that are less dependent on specific information about individual TFs. Our approach seeks to identify TFs that are important to tissue specificity by focusing on patterns of co-occurrence of pairs of DNA binding sites. Instead

*To whom correspondence should be addressed. Tel: +1 443 287 3882; Fax: 1+ 410 502 5382; Email: jiang.qian@jhmi.edu

of searching for single TFs that have a role in tissue-specific gene expression, we look for interacting TF pairs that may co-regulate tissue-specific genes. Our approach has been tested in the yeast model system (7). The method is based on the hypothesis that TF complex instead of individual TF is the functional unit in tissue-specific gene regulation; one can better identify TFs that contribute to tissue-specificity in the context of TF interactions than single TFs. Such analysis not only yields a list of TFs that may play a role in tissue-specific gene regulation, but also provides information about interactions between specific TFs.

In this paper we describe the application of this approach to human TF interactions. We first derived, from publicly available gene expression databases, a list of genes that are preferentially expressed in 30 tissues. These sets of tissue-specific genes represent 'signatures' of the transcriptomes of the tissues of interest. We then searched the upstream regions of these genes for all known TF binding sites, and predicted TF pairs that may co-regulate their expression. Based on this analysis, we present several conclusions about how non-tissue specific TFs can in combination direct tissue-specific expression, and also develop a network model of some of the interactions involved.

METHODS

Identification of tissue-specific genes

We utilized the NCBI EST database to obtain a set of genes that are preferentially expressed in the tissue which are termed as tissue-specific genes. EST sequences are clustered into UniGene. Let $e_i(g)$ be the number of ESTs corresponding to gene g in tissue i. The total number of ESTs in UniGene for g is $E(g) = \sum_i e_i(g)$. Given the total size of EST libraries in tissue i, s_i, the expected number of ESTs in tissue i for each gene is proportional to $p_i = s_i/\sum_i s_i$. For gene g, if it is expressed equally across all tissues (i.e. not differentially expressed), the expected number of ESTs in tissue i is equal to $f_i = E(g)p_i$.

The expression enrichment (EE) is defined as $\mathrm{EE}_i(g) = e_i(g)/f_i(g)$, which is the ratio between observed to expected number of ESTs for gene g in tissue i. A large value of $\mathrm{EE}_i(g)$ suggests that gene g is preferentially expressed in tissue i. To evaluate the possibility that a large EE is due to chance rather than a reflection of true differential expression, we calculated a P-value for the enrichment score. The P-value for gene g in tissue i was calculated according to the formula

$$P_i(g) = \sum_{x=e_i(g)}^{E(g)} \binom{E(g)}{x} p_i^x (1-p_i)^{E(g)-x}.$$

$P_i(g)$ is in essence the probability of observing $e_i(g)$ or more ESTs for a non-differentially expressed gene g in tissue i given the total number of ESTs for gene g [$E(g)$] and the size distribution of the tissue libraries (p_i).

We defined the genes as tissue-specific genes if they satisfy the two criteria (i.e. $EE_i(g) > 5$ and $P_i(g) < 10^{-3.5}$).

Promoters sequences and conservation scores

For analysis, we used only those UniGenes (used to identify tissue-specific genes) that had corresponding RefSeq entries (hg17). We defined the promoter region of a gene as the 1 kb non-coding sequence upstream of the gene's TSS and 5′-untranslated region (5′-UTR), with the TSS defined as the 5′-most position of the RefSeq sequence. 5′-UTR is trimmed to 1 kb if it is longer than that. The average length of a promoter is ~1480 bp. Although we are aware that some TF binding sites are located far from the transcription start site (TSS), the promoter regions we defined have the highest density of known TF binding sites (8), and thus it is more likely to detect real TF interactions in these regions.

Since some genes share the some promoter sequences (e.g. genes with alternative splicing), we removed these redundant promoters. In total, we identified 17 859 unique promoters for our analysis. In our calculations, we utilized information about sequence conservation and excluded repeat element sequences (e.g. Alu) (9). Conservation was determined from an eight-way alignment of human, chimpanzee, dog, mouse, rat, chicken, zebrafish and fugu sequences. The conservation was quantified by a conservation score between 0 and 1, with 1 as highest conservation (10). The conservation score was obtained from UCSC genome database (11). We utilized a relatively low cutoff (0.3) so that more human-specific binding motifs would not be excluded. The combined removal of repeat elements and the conservation constraint reduces the average promoter length to ~590 bp.

Position weight matrices and genome-wide search

We obtained human position weight matrices (PWMs) from TRANSFAC database and literatures (12). Some matrices are very similar to each other and not distinguishable. We compared the similarity between matrices and removed the redundancy according to one of the two criteria: (i) similarity >0.95; (ii) similarity >0.90 and sharing the same name (The details of the matrix comparison can be found in Supplementary Data). Finally 306 unique matrices are used for our study.

The match scores between a matrix and a promoter sequence were calculated for all possible positions along the promoter. The top 0.015% matches (including non-conservative ones) were utilized for the next stage of analysis. On average, each matrix has 3500 hits in conserved and non-repetitive regions, which corresponds to approximately one site occurring per five promoters. This cutoff is somewhat arbitrary and may cause some false positive hits. The study of TF pairs is expected to increase the prediction specificity.

Identification of tissue-specific TF interactions

We predict two TFs interacting with each other if their binding sites have over-represented co-occurrence in the promoters of tissue-specific genes and the distances (in unit of bp) between the two sites are significantly different from random expectation.

The P-value for two binding sites consists of two contributions. One is from the binding site pair co-occurrence and the other is from the distance constraint. The overall P-value is defined as:

$$P = P_{\mathrm{occ}} P_{\mathrm{d}}$$

where P_{occ} evaluates the over-representation of a binding site pair occurrence (g) in the tissue-specific promoters compared

to its occurrence (G) in all promoters in the human genome, and P_d is used to evaluate the deviation of the observed distance distribution from a random expectation. P_{occ}, is calculated according to

$$P_{occ} = \sum_{k=g}^{G} \binom{G}{x} \left(\frac{n}{N}\right)^x \left(1-\frac{n}{N}\right)^{G-x},$$

where n is the number of tissue-specific genes; N is the total number of human genes. Note that some TF binding sites occur multiple times in one promoter. In such cases, when calculating g and G we counted all combinations between the TF binding sites. An alternative option would have been to count these cases as a single binding site pair occurrence. We compared these two options and found that counting all possible pairs yielded better performance (Supplementary Data).

The contribution of distance constraint between two sites in promoter sequences, P_d, is calculated by comparing the observed distance distribution with a background distribution. For each TF pair, we calculated the distances between their binding sites in the promoters. If one or both sites occur multiple times in one promoter, we included the distances between all combinations. Even though we filtered out the sites in repetitive elements or non-conserved regions, we calculated the true distance between two sites as if these regions were not masked out.

The background distance distribution is considered to be from random site pairs that do not interact with each other. Since these random sites do not have any biological meaning, they could occur in repetitive and non-conserved regions. Given the length of one promoter sequence (L) and binding site pair distance (d), the number of all possible arrangements for the site pair is $L - w_f - w_b - d + 1$, where w_f and w_b are the widths of the two sites. The chance of observing distance d is proportional to the number of arrangements for a given d, and can be normalized as

$$f_d(L) = \frac{L - w_f - w_b - d + 1}{\sum_{i=1}^{L-w_f-w_b+1} (L - w_f - w_b - i + 1)}.$$

Given the length distribution $F(L)$ of promoter sequences in human, the random distribution of the binding site distances is $f_d = \sum_L F(L) f_d(L)$. P_d was calculated by comparing the observed distance distribution and f_d using the Kolmogorov–Smirnov test.

In addition, we also calculated the maximum of the integrated probability differences (A_d) of the observed and expected distributions (the gray area in Figure 1B). From the comparison of positive control and random binding site pairs, we set the thresholds for A_d to 0.24 and $-\log(P)$ to 6.2. Considering the multiple testing correction (i.e. $306 \times 307/2$ TF pairs), the corresponding alpha for the P-value threshold is $\sim$0.03.

Significance evaluation by permutation

We used an alternative method to evaluate the significance of the predicted TF pairs. After we obtained the hits for all TF binding matrices in conserved and non-repetitive regions (see above), we permuted the matrix labels on the hits. With the randomized labels, we calculated the P-values for each TF pair as described above.

Sensitivity and enrichment

We collected a set of known direct protein–protein interactions as positive controls to evaluate our prediction. Total 480 TF–TF interactions were obtained from databases of DIP (13) and TRANSFAC (12). We measured sensitivity and enrichment for our prediction.

The sensitivity is defined as the ratio between recovered positive controls and total positive controls. For example, we predicted total N interactions in 30 tissues. N_p of them are known TF–TF interactions. Since some interactions could occur in multiple tissues, we may have N' and N'_p unique interactions in predicted and known interactions, respectively. The sensitivity is therefore equal to $N'_p/480$.

The enrichment is defined as the ratio between observed (D_o) and expected density (D_e) of positive controls in the predicted interactions. The observed density is calculated according to $D_o = N_p/N$. The expected positive density (D_e) is the ratio between all positive controls and all possible TF–TF interactions. Note that here the same TF–TF interaction is considered as distinct interactions if it occurs in different tissues. Thus, the number of all possible interactions is $M(M+1)/2 * 30$, where M is the total number of TFs in the study (i.e. $M = 306$). For positive controls, we have no information about how many tissues one interaction associates with. We estimated it as N_p/N'_p. The number of all positive controls in 30 tissues is $480 * N_p/N'_p$. Therefore,

$$D_e = \frac{480 * N_p/N'_p}{M(M+1)/2 * 30}$$

Expression coherence

We calculated the correlation coefficients of any gene pairs in the entire genome to get the background distribution of the correlations, and the value at the top highest 5% of the distribution was set to be a threshold. The thresholds are different for different microarray datasets (tissues). The expression coherence (EC) for a group of genes is the fraction of gene pairs with correlation coefficient higher than the threshold. Thus, for a group of randomly picked genes, the EC value is expected to be 0.05.

RESULTS

Identifying tissue-specific genes

We first derived groups of tissue-specific genes that are preferentially expressed in a particular tissue. Using the available 5.3 millions human EST sequences, which map to 54 000 UniGene clusters (NCBI) (14,15), we calculated the gene expression pattern for each UniGene in 30 human tissues. The list of the tissues can be found in Table 1.

To determine whether a gene is preferentially expressed in a tissue, we defined EE as the ratio between observed expression level in a tissue versus averaged expression level across the 30 tissues. For each gene, 30 EE values were calculated, one for each of the 30 tissues. We also calculated the P-value

Table 1. Summary of the predictions

	Number of specific genes	Top three interactions			Total number of interactions	Top three hubs			Total number of TF involved
Bladder	247	ARNT:SREBP1	MAX:SREBP1	SREBP1:MYC/MAX	230	SREBP1	NRF1	USF	108
Blood	508	ELF1:PEA3	ETF:NRF1	PEA3:PU.1	208	PEA3	ELF1	FOXJ2	124
Bone	145	EF-C:MIF1	EF-C:RFX1	GATA-3:MIF1	342	ATF	EF-C	MIF1	169
Bone marrow	343	ETF:NRF1	MAX:SREBP1	NRF1:STAT3	245	SREBP1	NRF1	ELK1	127
Brain	308	AP2alpha:ETF	C/EBP:PBX1	ETF:LBP1	831	C/EBP	CHX10	FREAC7	245
Cervix	279	CREB:NRF1	ELK1:NRF1	ETF:NRF1	94	NRF1	SMAD-3	E2F	57
Colon	260	AFP1:HNF1	CDP:FOXO4	CDP:HNF1	196	HNF1	CDP	AFP1	108
Eye	289	CHX10:CRX	CHX10:GATA6	CRX:PITX2	365	CRX	CHX10	FOXJ2	168
Heart	270	AP1:MEF2	AP1:RSRFC4	FOXJ2:POU3F2	607	MEF2	SRF	RSRFC4	205
Kidney	460	COUP-TF/HNF4:HNF1	CRX:HNF1	GCNF:HNF1	418	HNF1	HNF4alpha	COUP-TF/HNF4	184
Larynx	371	AP1:NFE2	AP1:TCF11/MAFG	BACH1:NFE2	462	ETF	NRF1	SRF	204
Liver	393	CDP:HNF1	C/EBPgamma:HNF1	E4BP4:HNF1	202	HNF1	COUP-TF/HNF4	HNF4alpha	106
Lung	157	ETF:MYC/MAX	ETF:LBP1	ETF:TBP	56	NRF1	ETF	LHX3	59
Lymph node	475	c-Ets-1:c-Ets-2	ELK1:PU.1	ICSBP:c-Ets-2	183	NRF1	LBP1	ICSBP	105
Mammary gland	158	E2F:SRY	ETF:TAX/CREB	ETF:ZID	85	ETF	FAC1	RSRFC4	70
Muscle	326	AP-4:MEF2	MEF2:MYOD	MEF2:RSRFC4	1052	MEF2	RSRFC4	LBP1	227
Ovary	206	AP2alpha:VDR	CREB:MAZ	DBP:VDR	109	MZF1	MAZ	ETF	93
Pancreas	206	ATF:HEB	ATF:NF-muE1	ATF:NRL	225	ATF	MYOD	E12	135
PNS[a]	126	CREB:RSRFC4	HNF1:SRY	MYOD:OCT1	406	POU3F2	HFH-3	CREB	148
Placenta	379	ATF1:CHX10	ATF1:LHX3	ATF:CHX10	135	LHX3	CHX10	ATF	79
Prostate	201	AFP1:LHX3	CART1:LHX3	C/EBPgamma:LHX3	59	LHX3	TBP	SRF	51
Skin	209	AREB6:ALX4	AREB6:ARP-1	AREB6:E47	132	AREB6	VDR	LMO2 complex	102
Small intestine	133	CART1:LHX3	HNF1:LHX3	HNF1:NKX6-2	948	LHX3	HNF1	NKX6-2	239
Soft tissue	182	C/EBPgamma:FOXO4	FOXO1:SF1	FOXO4:TBP	311	LUN-1	FOXO4	FOXO1	146
Spleen	168	E12:c-Ets-1	GCM:NFKB1	LBP1:MYOD	107	LBP1	RSRFC4	NFKB1	82
Stomach	270	AP2gamma:ETF	AREB6:NFE2	ER:HIF1	73	AREB6	ETF	BACH1	77
Testis	860	AP2:NRF1	EGR1:NRF1	EGR3:NRF1	232	NRF1	NF-Y	SP1	120
Thymus	107	ATF:TAX/CREB	c-Ets-1:c-Ets-2	ETF:NRF1	219	TAX/CREB	ATF	c-Ets-1	113
Tongue	480	ATF:CREB	ATF:CREBP1	CREBP1:CREB	465	ATF	NRF1	CREB	158
Uterus	58	E4BP4:POU1F1	E4BP4:POU3F2	NKX6-2:POU3F2	63	POU3F2	OCT1	E4BP4	45

[a]PNS, peripheral nervous system.

for each EE to evaluate its statistical significance. With these two values, we defined a gene as 'tissue specific' if it had an EE in a particular tissue larger than 5 and a P-value $<10^{-3.5}$ (for details see Methods). Clearly, by this definition, 'tissue specific' is a relative term and does not mean that a particular gene is expressed only in a specific tissue. The numbers of tissue-specific genes for each tissue are shown in Table 1. On average, there are ~290 tissue-specific genes for each tissue. In total, we identified 7261 tissue-specific genes for the 30 tissues.

As an evaluation, we also performed an independent gene search according to a gene's annotated functions. Based on the gene ontology (GO) annotation (16), we searched the logical combination of keywords to find the known tissue-related genes. For example, we used 'eye or retina' to search eye-related genes, while we used 'bone but not bone marrow' to search bone-related genes. In most cases we found a large overlap between the two sets of genes that resulted from gene expression and function annotation. This means that many genes with known tissue-related functions are actually also preferentially expressed in the tissue and vice versa. However, for some tissues, the overlap is not significant. These tissues are often not well studied and GO information is lacking. Therefore, using expression data from the EST database provides a more objective and unbiased determinant of tissue-specificity.

From our analysis, we found that most tissue-specific genes (85%) are specific to only one tissue. However, some tissue-specific genes are shared by several tissues. For example, muscle and heart share 171 tissue-specific genes, which is not surprising because skeletal and cardiac muscle have in common a number of structural, biochemical and functional characteristics.

Expression enrichment and occurrence enrichment

Among the identified tissue-specific genes, nearly 10% (605 out of 7261) of them encode TFs. Intuitively, these tissue-specific TFs are likely candidates to play an important role in regulating tissue-specific expression in their respective tissues. If this hypothesis is correct, one might expect the DNA binding sites for these TFs to be over-represented in the promoters of their respective tissue-specific genes (the promoter region is defined as 1 kb non-coding upstream to the TSS). To evaluate this possibility, we studied TF binding site occurrence enrichment (OE). OE measures the enrichment of a TF binding site in the promoters of a group of tissue-specific genes compared to its overall occurrence in all promoters of the genome. It is defined as the ratio between observed and expected TFs' binding site occurrence in the promoters of the tissue-specific genes. To calculate OE, we searched TF binding sites in the promoter sequences. To increase the chance that the matched sequences are biologically functional, we considered only hits in evolutionarily conserved regions. In our calculation, we only considered the binding sites falling in the regions with conservation scores >0.3. In addition, we also excluded the regions annotated as repeat elements (e.g. Alu).

We first examined the relationship between EE and OE for the 306 TF binding sites available in TRANSFAC databases and literature (12). A few TFs have both large EE and large OE in a tissue; these TFs presumably play a significant role in regulating their corresponding tissue-specific genes. For example, CRX has large EE and OE in eye and it is known to be a TF that regulates many retina-related genes such as rhodopsin (17,18). Similarly, MYOD has a large EE and OE in muscle, and it is known to play an important role in muscle gene regulation (19). However, there is no overall correlation between EE and OE; the TFs that are preferentially expressed in a particular tissue do not necessary have more occurrences of their respective binding sites in the promoters of genes with the same tissue specificity (Supplementary Figure 1). Furthermore, this observation is not affected by our choice of the conservation threshold (Supplementary Figure 2).

We hypothesize that many TFs without large EE and OE are also important in regulating tissue-specific genes. For example, MEF2 is not preferentially expressed in muscle (its EE in muscle is 0.4), yet it is known to regulate muscle-specific genes (20). As an additional approach to distinguish which TFs are relevant to tissue-specific gene regulation, we studied the combinatorial occurrence of TFs' binding sites. This effort was based on the assumption that tissue specificity is likely to be achieved by a set of TFs instead of single TFs. In order to look for this type of combinatorial specificity, we attempted to identify TF pairs that are likely to co-regulate tissue-specific genes based on the relationship between their binding sites.

Distance constraint of TF binding sites can help identifying interacting TFs

Previous bioinformatics efforts to predict interacting TFs have often been based on the co-expression of their target genes or co-occurrence of DNA binding sites (21–24). Here, we have added information about distance between two binding sites. Our argument is that if two TFs interact with each other, the locations of their binding sites in the promoters are not independent of each other. The distance between their binding sites in the promoters should differ from a random distribution. More specifically, we suggest that the binding sites of two interacting TFs should tend to be close to each other.

For most of TF pairs, their DNA binding sites do not show any preference for a particular distance. The distance distribution is very close to random expectation (Figure 1A). However, for some known interacting TF pairs, the distances between their respective binding sites are not uniformly distributed (Figure 1B). The chances for the two binding sites at short distances are significantly enriched, while at larger distance (>200 bp) the chances are depleted compared with random expectation. Figure 1C illustrates an example of an interaction between multiple binding sites of the same TF, which we call homotypic interactions. We found that multiple copies of SP1 binding sites often occur in promoters, and that their distance distribution indicates that these binding sites tend to be close to each other.

More significantly, we found that some TF pairs display enhanced preference for short distances in the context of certain tissue-specific genes. For instance, MYOD and MEF2 are both known to be involved in muscle gene regulation and interact with each other (25,26). If we searched for

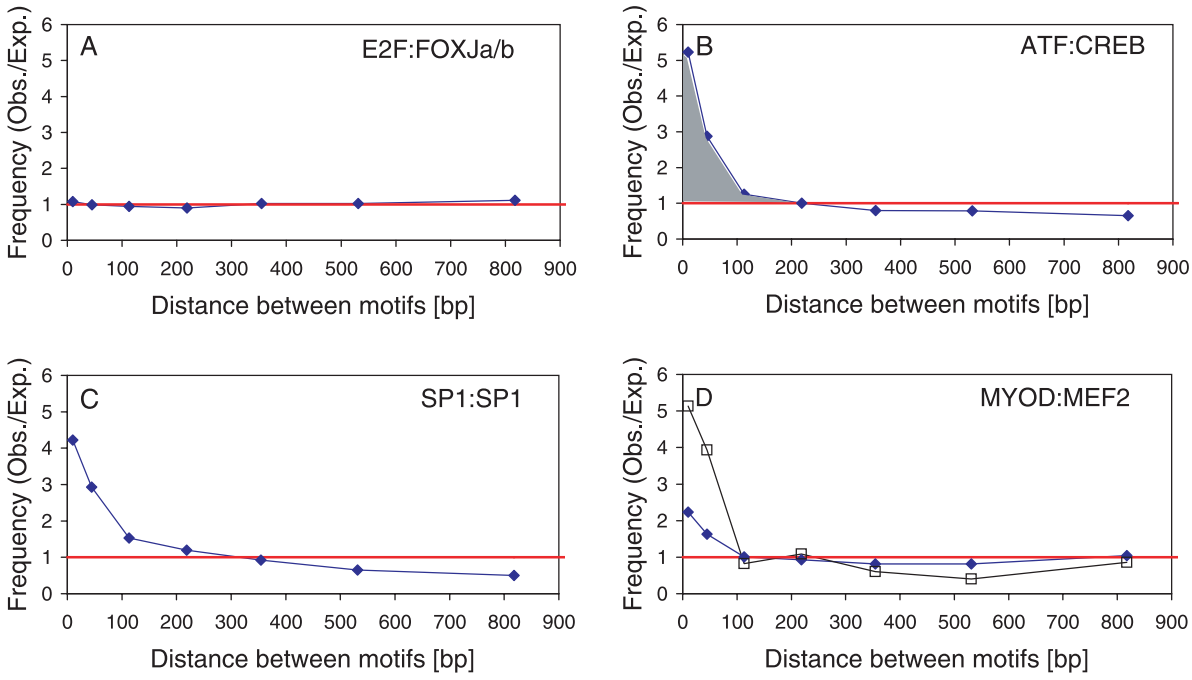

Figure 1. Distance distributions between TF binding sites in promoters. (**A**) shows an example of two TFs without preferential distance between their binding sites in promoters. (It should be noted that here is currently no published data implicating interaction between E2F and FOXJa/b.) The line of $y = 1$ represents the expected distribution. (**B** and **C**) show the distance distributions for two examples with interacting TFs. (**D**) is an example for tissue-specific interaction. The short distance enrichment is enhanced in muscle specific genes (open squares) compared with that in entire genome (filled diamonds). The gray area in (B) was used to quantify the difference between observed and expected distributions.

their binding sites in the promoters of the entire genome, there is a slight enrichment for shorter distances. However, when we restricted analysis to the promoters of the 326 muscle-specific genes (Table 1), we found considerable enrichment for short distances (Figure 1D). This observation suggests that the interaction between MYOD and MEF2 may be muscle-specific.

Inspired by such observations, we examined the distances between all possible 46 971 (306 * 307/2) TF binding site pairs between the 306 TFs in the 30 groups of tissue-specific genes. To quantify the short distance enrichment between two sites, we calculated the area between the observed and expected distance distributions (Figure 1B). Larger areas indicate greater deviation from random expectation, and hence stronger interactions, while area close to 0 indicate little or no interaction. In addition, we calculated the *P*-value (for P_d see Methods) to evaluate the significance of the difference between two distributions. In conjunction with the information of co-occurrence of the binding sites (for P_{occ} see Methods), we predicted tissue-specific interactions between TFs.

Prediction of tissue-specific interactions

Based on the co-occurrence and distance distribution of their binding sites, we calculated *P*-values for all possible pairs of TFs. Each TF pair is associated with 30 *P*-values, one for each of the 30 tissues. Figure 2A shows the *P*-values for the pair of MYOD and MEF2 (both factors are known to regulate muscle-specific genes); in support of the fidelity of our approach, the highest *P*-value occurs in muscle ($P < 10^{-26}$). Besides the known TF interactions, we also made predictions on novel interactions. For example, the interaction between PAX2 and FOXJ2 is predicted to be eye-specific, as the *P*-value for their binding sites is most significant in the eye ($P < 10^{-14}$).

Overall, we predicted 9060 tissue-specific TF interactions (~300 for each tissue). The detailed numbers and the top three most significant predictions are listed in Table 1. Most of the interactions (76%) are specific to only one tissue. We made schematic interaction networks for each individual interactome (Supplementary Figure 2). The difference between these networks is striking. Although the number of participating TFs is limited (306), the connectivity in each networks are clearly different. These unique interactomes might contribute to the unique behavior of the individual tissues.

Evaluation by known protein–protein interactions

We undertook several approaches to assess the validity our TF pair predictions. One was to compare our predictions to known information about protein–protein interactions. Although we realize that our predicted interactions between TFs do not necessary reflect direct physical interaction (for example, interaction could be mediated through another co-factor), we felt that this comparison would provide a useful rough assessment. We collected known protein–protein interactions from TRANSFAC (12) and DIP (13) databases, and a recent large scale yeast two hybrid screen (27). In total, we collected 480 known TF–TF interactions for our analysis.

Sensitivity and specificity are widely used statistics for evaluation of a predictive method. However, in our case,

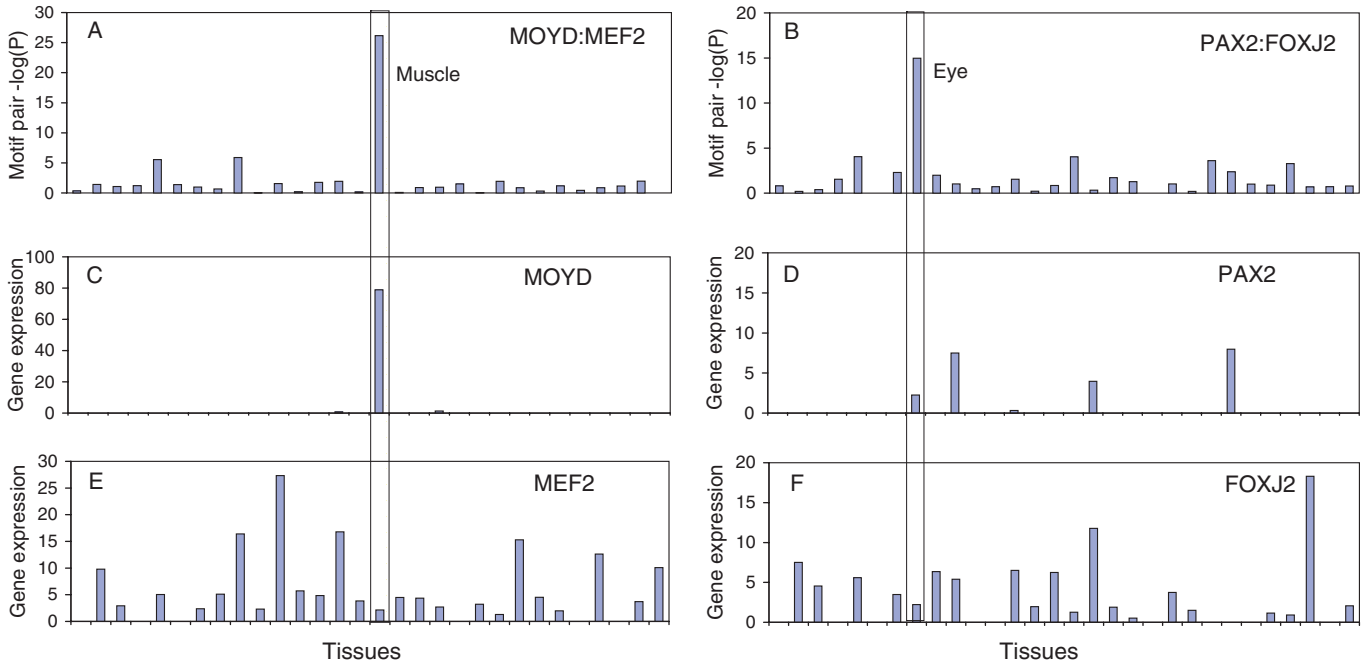

Figure 2. Examples of interactions. The *x*-axis is the tissue types (in the same order as Table 1). Two interactions are found for muscle (**A**) and eye (**B**) based on the relationship between the TF binding motifs. The *y*-axis is the *P*-values from the prediction. (**C**–**F**) show the relative gene expression levels of single TFs in an arbitrary unit.

because the majority of the TF interactions are unknown, we could not derive a reasonable estimate for specificity. Instead, we calculated the enrichment for known interactions in the prediction. Enrichment is defined as the ratio between observed and expected fraction of known protein–protein interactions in the predicted set (Methods). In order to get an overview of the behavior of the enrichment and the sensitivity, we plotted them as a function of *P*-value (Figure 3A). As expected, if more stringent *P*-value thresholds are used, the number of known TF interactions recovered decreases. At the same time, the enrichment of known interactions in the prediction increases with more stringent thresholds. This indicates that the known interactions are not randomly distributed in the prediction set and tend to have more significant *P*-values from our prediction. With the information from the plot, we chose an arbitrary threshold of $P = 10^{-6.2}$ for our prediction of TF interactions. At this threshold, >40% of the known interactions are recovered, with 84-fold enrichment for known interactions. Note that this approach may underestimate the sensitivity of the method because the known protein–protein interactions may also include non-tissue-specific TF interactions, which is not the focus in this study.

We also compared the *P*-value distribution for known TF interactions and unknown pairs (Figure 3B). One can see that on average known TF interaction pairs have more significant *P*-values compared to the other TF pairs. However, it is also clear that the two distributions have a large overlap, indicating that the unknown TF pair set actually includes many interacting TF pairs that have not yet been determined. We performed a random simulation by permutation (Methods). The *P*-values obtained from this simulation are further shifted toward the left (less significant).

Evaluation by co-expression of target genes

As a second independent evaluation of our TF interaction predictions, we analyzed the expression patterns of the potential target genes of the TF pairs. Our working hypothesis was that the expression profiles of a set of genes controlled by common TFs are more likely to be correlated and behave similarly than those of randomly selected genes (28). Therefore, if we observed significant co-expression of their target genes, we inferred that the TF pair is regulating the target genes; otherwise, the TF pair is likely to be inactive. We used EC as a measurement of similarity between the expression profiles of a group of genes (21). Note that the EST data used for identification of tissue-specific genes show the average expression level of a gene, while the expression profiles we used here are the detailed expression fluctuation across various conditions. The gene expression profiles analyzed were based on microarray experiments obtained from SMD database (29). We collected expression profiles in five tissues, including brain (30–32), eye (33), liver (34), lymph node (35) and pancreas (36). The experimental conditions are quite heterogeneous for different tissues. For example, the expression levels in various compartments of eye (e.g. retina, corneal) were measured; while for the liver, different tumor samples were used for expression.

We first examined whether the target genes of detected TF pairs are mainly tissue-specific. For a given interacting TF pair that is specific to a tissue (e.g. eye), we first identified a set of genes whose promoter sequences contain both binding sites. Among these genes, some of them are tissue-specific (e.g. eye-specific), and some are not. We calculated the EC for the two groups. For example, we identified CRX and NRL as an eye-specific TF pair. A whole-genomic scan resulted in 569 genes with both binding sites in their

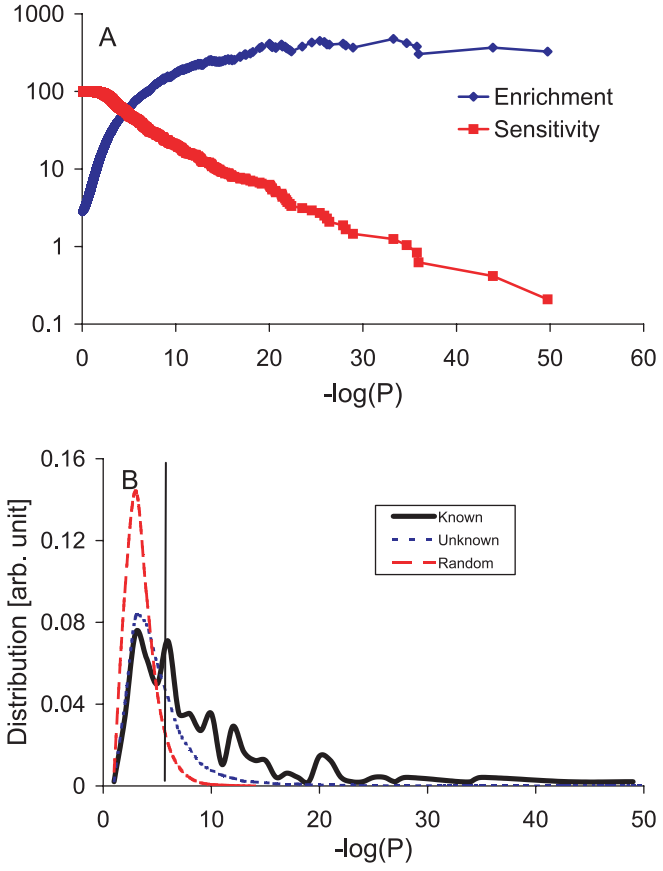

Figure 3. Evaluation of prediction by protein–protein interactions. (**A**) Two curves represent the sensitivity and enrichment in function of *P*-values. (**B**) *P*-value distributions for known protein–protein interaction, TF pairs without known interactions, and pairs from permutation simulation. The vertical line indicates the threhold chosen for this study.

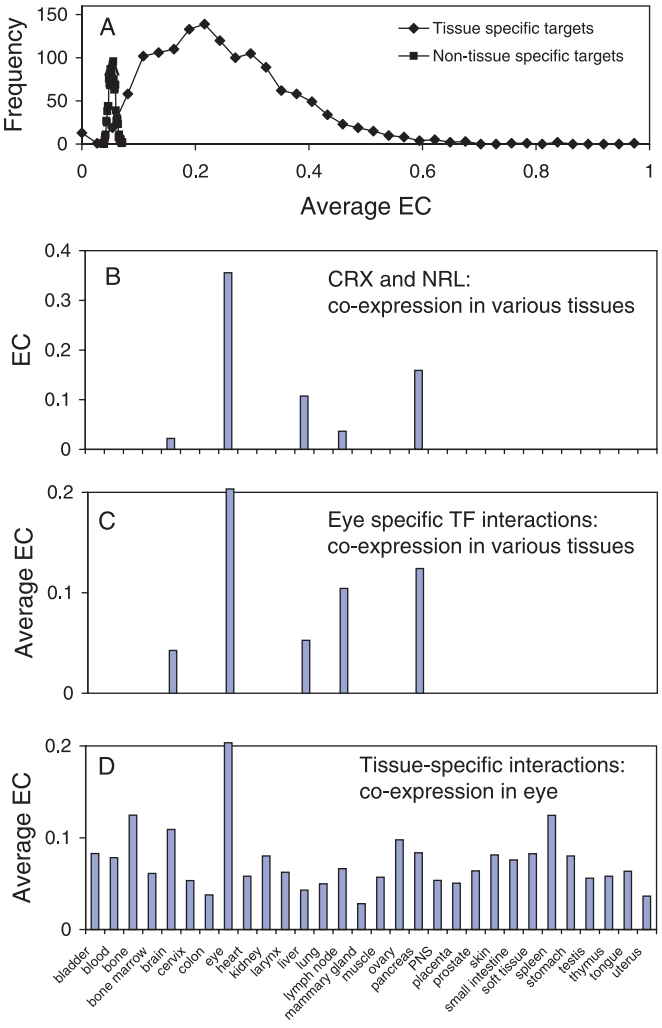

Figure 4. Evaluation of prediction by gene expression. (**A**) The expression coherence (EC) distributions for tissue specific genes and non-tissue specific genes. (**B**) The EC of the target genes of eye-specific pair CRX and NRL in different tissues. (**C**) The average EC of the target genes of all eye specific TF pairs in different tissues. (**D**) For all tissue specific TF pairs, we calculated the co-expression of their target genes using the microarray data set obtained in eye. The eye specific TF pairs have the highest co-expression in eye.

promoters and EST database analysis identified 18 of them as eye-specific. For this set of the 18 eye-specific genes, we calculated the EC to be 0.36 which is significantly higher than the value of the non-eye-specific gene group (0.06) or the general background (0.05). Interestingly, the average distance between the two binding sites in the eye-specific genes is shorter than that in non-eye specific genes (394 versus 553 bp). The results of both the co-expression and position analyses indicate that CRX and NRL may mainly regulate eye-specific genes. This conclusion is not only true for this specific example. In fact, the same observation was made for most of the predicted TF pairs: the ECs for non-tissue specific genes are narrowly distributed ~0.05; while the ECs for tissue-specific genes are much higher (Figure 4A).

We next checked whether the tissue specific genes are also co-expressed in other tissues. For example, we examined if the 18 target genes of CRX and NRL are co-expressed in tissues other than eye. By using microarray data measured in four other tissues, we obtained the EC values for these 18 genes in these tissues (Figure 4B). The co-expression levels in other tissues are lower than that in eye, indicating that the TF pair preferentially exerts its effect on targets in the eye. Similarly, we performed this evaluation with the 365 predicted eye-specific TF pairs and found that the average ECs for the target genes of these TF pairs are higher in the eye than in four other tissues (Figure 4C). Conversely, if we calculated the ECs for all tissue-specific TF pairs using microarray data set measured in eye, we found that eye-specific TF pairs have the highest co-expression (Figure 4D). In summary, we found that (1) the target genes of tissue-specific TF pairs are mostly the genes with the same tissue specificity, and (2) the TF pairs regulate their target genes predominantly in the tissue of interest.

General properties of the TF–TF interaction networks

One notable feature of the predicted TF interaction network is that most TFs involved in tissue-specific regulation are themselves not preferentially expressed in the tissue of interest, i.e. non-tissue-specific TFs can regulate tissue-specific gene expression. As one illustration of this point, of the 98 TFs involved in eye specific interactions, only 11 of them are themselves eye specific. As a specific example of this

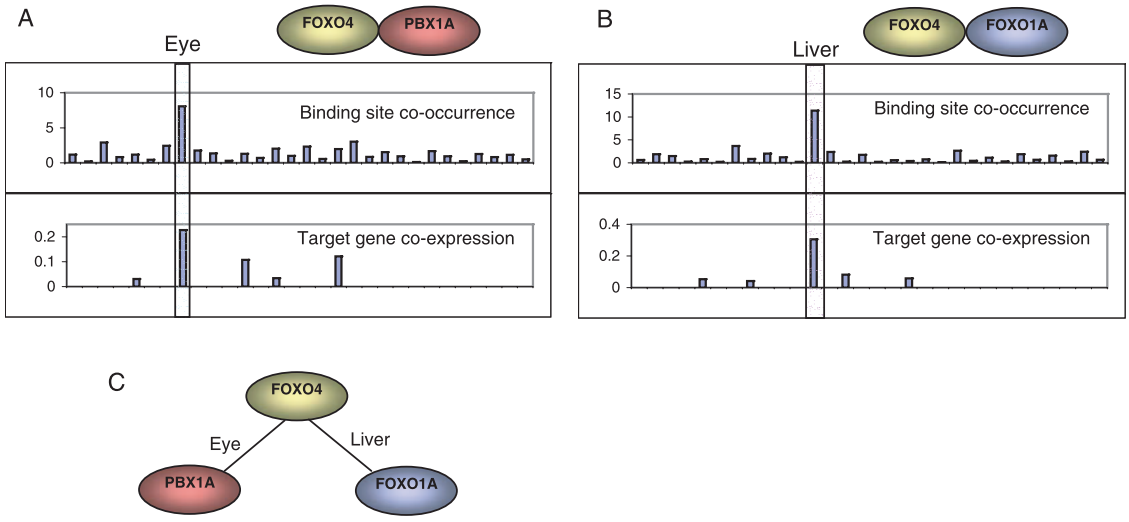

Figure 5. One TF participating in multiple tissue specificity. (**A**) Both the binding site co-occurrence and target gene co-expression indicate that FOXO4 and PBX1A have eye-specific interaction. (**B**) FOXO4 and FOXO1A have liver-specific interaction. (**C**) FOXO4 can participate in regulation of eye- and liver-specific gene expression by interacting with PBX1A or FOXO1A.

dichotomy between TF expression and target specificity, MYOD and MEF2 were predicted to interact specifically in muscle (Figure 2A). While MYOD is preferentially expressed in muscle, MEF2 is not muscle specific at all (Figure 2C and E). As another example, PAX2 and FOXJ2 interact with each other in the eye, yet neither of these factors is highly expressed in the eye (Figure 2D and F). These examples indicate that tissue specificity might be achieved by combination of multiple TFs. By interacting with other TFs, non-tissue specific TFs can co-regulate tissue-specific genes.

Another interesting feature we observed from our TF analysis is that an individual TF can contribute to tissue specificity in different tissues by interacting with different TFs. We demonstrate this characteristic with an example in Figure 5. FOXO4 and PBX1A are predicted to interact in an eye-specific manner based on their binding site distribution and co-expression of their target genes. Similarly, FOXO4 and FOXO1A demonstrate a specific interaction in liver. Therefore, FOXO4 can contribute to tissue specificity in both eye and liver, depending on whether it interacts with PBX1A or FOXO1A, respectively. This example again indicates that the 'function' of a TF is better defined by its interactions with other TFs ('the company it keeps') rather than by the characteristics of individual TFs. In our Supplementary Data, we present other examples of single TFs participating in multiple tissue specificity.

Tissue-specific interaction clusters

A predicted interaction map provides an opportunity to examine the local and detailed interactions specific to various tissues. Figure 6A shows a global view of the interaction network generated from our predictions. In the figure, the different colors of nodes (TFs) and edges (interactions) indicate different tissue types. The thickness of edges reflects the significance of the predicted interaction. From the overall network, we can appreciate the complex inter-relations and see hints of modularity in the network. If we focus upon the most significant interactions (interaction with $P < 10^{-20}$), a clear pattern emerges (Figure 6B). Interestingly, interactions with the same tissue types form tissue-specific clusters. This type of clustering is not expected to occur spontaneously from a random network. To illustrate the properties of the network, we discuss five clusters in detail.

Muscle and heart clusters. Muscle- and heart-specific interaction clusters are interconnected, which is not surprisingly due to the biological similarity of skeletal and heart muscle. The clear centrally controlling TF is MEF2 which is a well known TF in muscle (25). MEF2 is in fact the most connected node (hub) both in muscle- and heart-specific interactions; these hubs presumably represent the most central TFs in these tissues (Table 1 lists the top three hubs for each tissue). The interaction between MEF2 and MYOD is essential for inducing myogenesis in transfected fibroblasts (25). Besides MYOD, most of predicted interaction partners for MEF2 are known to be in involved in muscle and heart gene regulation (e.g. MYOD, SRF, RSRFC4 and AP1) (4,19). MEF2 is also predicted to work with LF-A1, PAX2 and ARP1 to control transcription in muscle. These three TFs' interactions mostly occur in other tissues instead of muscle and heart. This means that even though some TFs may mostly be involved in transcription in other tissues, depending on the partner, they may regulate different tissues.

Kidney and liver clusters. The interaction clusters for kidney and liver are linked through HNF1, which is known to be predominantly expressed in liver and kidney and regulate a variety of liver-specific genes (37). Based on our predictions, it is in fact found to function as a hub in four tissues: kidney, liver, small intestine and colon (Table 1). Our results suggest possible important regulatory roles in these tissues. HNF1 is known to function as a dimer (38), consistent with our prediction of a homotypic interaction for this factor. We also recovered two known liver-specific regulators, HNF4 and HNF3 (5,39). Both are found to interact with HNF1. Our analysis of TF interaction not only provides a collection of TFs

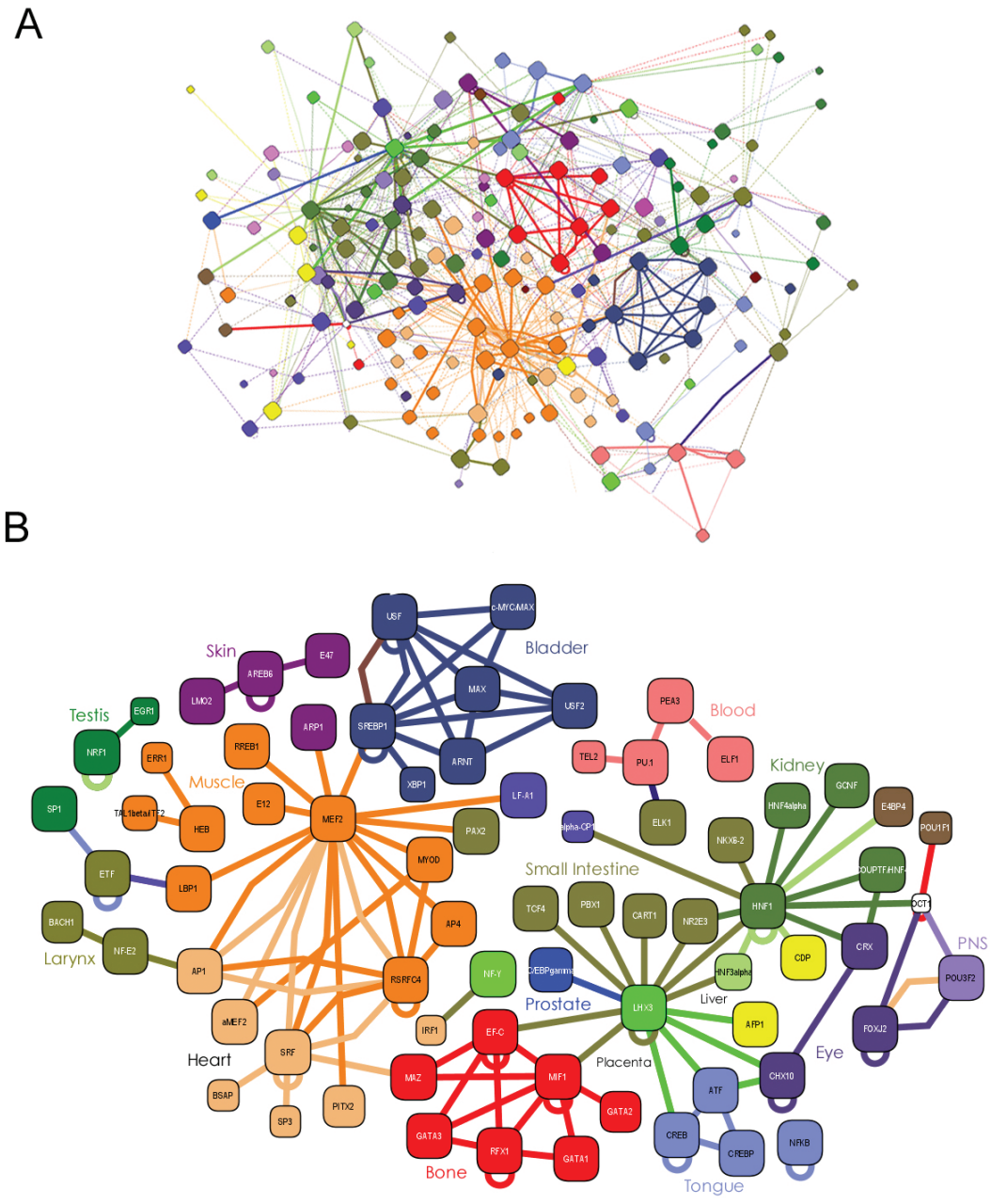

Figure 6. Network of predicted TF interactions. (**A**) The predicted TF interaction network. (**B**) The most significant interactions with $-\log(P) > 20$. The colors of nodes (TFs) and edges (interactions) indicate different tissue types. If the majority of interactions through one TF are from one tissue type, the node is colored with that tissue type. Otherwise, the node is blank. The size of nodes indicates the numbers of interactions through this TF. The thickness of edges indicates the significance of interactions.

involved in tissue-specific gene regulation, but also reveals the relationships between them.

Bone cluster. In bone-specific interaction cluster, there are seven TFs: EF-C, MAZ, MIF1, GATA1, GATA2, GATA3 and RFX1. Among them, EF-C and MIF1 are the bone-specific hubs (Table 1). Interestingly, they also form two homotypic interactions. MAZ1 participates also in a co-regulatory interaction with SRF in heart. Most of these TFs and interactions are previously unknown and our results provide new hypothesis for experimental testing.

DISCUSSION

In this paper, we have demonstrated that one can predict tissue-specific TF interactions based on the observed non-random distribution of their respective DNA binding sites within the upstream regions of genes. The predicted TF interactions are likely to co-regulate tissue-specific genes, and thus contribute to the tissue specificity of gene expression. Our predictions were evaluated by comparison to known protein–protein interactions and by assessing the degree of gene expression correlations between target genes. We found that TFs that themselves are not preferentially

expressed in a tissue can combine with other factors to contribute to tissue specificity. In addition, individual TFs can participate in tissue-specific interactions in multiple tissues by interacting with distinct partners in the different tissues. These findings are consistent with and strengthen the rationale for ongoing experimental transcription studies that are beginning to focus on transcription complexes as the functional unit in gene regulation. Our large scale study demonstrates in detail that biological function can be better defined in terms of TF interactions rather than in terms of single TFs alone.

To identify the interaction between TFs, we examined the distance distribution between their binding sites within the upstream regions of genes. From our previous analysis on yeast promoter regions (7), we discovered that many interacting TF pairs display strong preferences in terms of the distances between their binding sites. We termed the preferred distances as characteristic distances. In this study on a mammalian system, we did not find evidence for characteristic distances between two interacting TFs. The lack of particular characteristic distance in the mammalian system can be attributed to both biological and technical reasons. Biologically, the mammalian system is much more complex. One TF may have many interacting TFs. The sum of many characteristic distances could cancel out single or several peaks in the distance distribution. Technically, the signal to noise ratio of identifying functional TF binding sites is much lower in the mammalian system. Although we used evolutionary conservation to increase the specificity of our prediction of TF binding sites, it is likely that our analysis still included many non-functional matches. This is in contrast to the situation in yeast where the chance of identifying 'true', biologically active, TF binding sites is much higher. One reason for this is that promoter structure is much simpler in yeast. Another reason is that we were able to integrate the large scale chromatin-immunoprecipitation on microarray (ChIP-chip) dataset (40,41) in our analysis, and such large scale ChIP-chip data is not available for the mammalian system.

Based a similar approach, Zhu *et al*. (42) identified interacting TF pairs based on over-represented co-occurrence of binding motifs in the promoters of entire genome. Their approach may miss many tissue-specific TF interactions as identified in our study. In addition, the interactions detected in their work may be not as statistically significant as the tissue-specific interaction described here. For instance, for an eye specific interaction, many non-eye specific genes also contain their binding sites in the promoters, and they will introduce noise in the prediction. Other methods have been proposed to find cooperative TF pairs. One of such program is Co-Bind (24), which utilizes a Gibbs sampling approach to identify the interactions between two TFs. The difference between our method and Co-Bind is that we studied the interactions between known TF binding sites while Co-Bind detects interactions between novel sites. However, Co-Bind has not yet been tested on mammalian systems.

As one of our future directions, we could predict TF interactions for more specific gene groups such as cell-type specific expression patterns. For instance, we detected eye specific interactions in this study. We could further break down genes according to the specific cell type in which they are expressed, such as photoreceptor versus ganglion cell. However, we are aware of some of the technical challenges that would be associated with such an approach. One is data availability. Currently, we do not have sufficient datasets for this type of work. However, the increasing availability of large *in situ* hybridization studies (43), as well as ongoing advances in laser capture microscopy and cell purification methods, coupled with microarray and other gene profiling methodologies suitable for small sample preparations, may soon make such studies possible. A second technical problem that will have to be addressed relates to statistical issues that may arise from the availability of only small groups of genes specific to particular cell types.

Our analysis provides a global picture of TF combinatorial regulation of tissue-specific genes. In addition, we made specific predictions in various tissues for experimental testing. However, it is worthy to note that only 306 TF binding sites are currently available and included in this study. Thus, the predicted interaction network is far from complete, and will need to be further developed and modified as more experimental data accumulates related to the nature and specificity of TF–DNA interactions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The research is supported in part by grants from the National Institute of Health (EY015684 to J.Q. and EY001765 to D.J.Z.), funds supporting the Guerrieri center for genetic engineering and molecular ophthalmology and by a gift from Mr and Mrs Robert and Clarice Smith. Funding to pay the Open Access publication charges for this article was provided by NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Shalon,D., Smith,S.J. and Brown,P.O. (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*., **6**, 639–645.
2. Velculescu,V.E., Zhang,L., Zhou,W., Vogelstein,J., Basrai,M.A., Bassett,D.E., Jr, Hieter,P., Vogelstein,B. and Kinzler,K.W. (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.
3. Messina,D.N., Glasscock,J., Gish,W. and Lovett,M. (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res*., **14**, 2041–2047.
4. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol*., **278**, 167–181.
5. Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res*., **11**, 1559–1566.
6. Berman,B.P., Nibu,Y., Pfeiffer,B.D., Tomancak,P., Celniker,S.E., Levine,M., Rubin,G.M. and Eisen,M.B. (2002) Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.
7. 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.
8. Qian,J., Lin,J., Luscombe,N.M., Yu,H. and Gerstein,M. (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, **19**, 1917–1926.
9. Smit,A.F. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev*., **9**, 657–663.
10. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al*. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*., **15**, 1034–1050.
11. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996–1006.
12. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*., **28**, 316–319.
13. Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*., **30**, 303–305.
14. Schuler,G.D., Boguski,M.S., Stewart,E.A., Stein,L.D., Gyapay,G., Rice,K., White,R.E., Rodriguez-Tome,P., Aggarwal,A., Bajorek,E. *et al*. (1996) A gene map of the human genome. *Science*, **274**, 540–546.
15. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for 'expressed sequence tags'. *Nature Genet*., **4**, 332–333.
16. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet*., **25**, 25–29.
17. Chen,S., Wang,Q.L., Nie,Z., Sun,H., Lennon,G., Copeland,N.G., Gilbert,D.J., Jenkins,N.A. and Zack,D.J. (1997) Crx, a novel Otx-like paired-homeodomain protein, binds to and transactivates photoreceptor cell-specific genes. *Neuron*, **19**, 1017–1030.
18. Furukawa,T., Morrow,E.M. and Cepko,C.L. (1997) Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. *Cell*, **91**, 531–541.
19. Weintraub,H., Dwarki,V.J., Verma,I., Davis,R., Hollenberg,S., Snider,L., Lassar,A. and Tapscott,S.J. (1991) Muscle-specific transcriptional activation by MyoD. *Genes Dev*., **5**, 1377–1386.
20. Yu,Y.T., Breitbart,R.E., Smoot,L.B., Lee,Y., Mahdavi,V. and Nadal-Ginard,B. (1992) Human myocyte-specific enhancer factor 2 comprises a group of tissue-restricted MADS box transcription factors. *Genes Dev*., **6**, 1783–1798.
21. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet*., **29**, 153–159.
22. Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res*., **31**, 7024–7031.
23. Bulyk,M.L., McGuire,A.M., Masuda,N. and Church,G.M. (2004) A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli*. *Genome Res*., **14**, 201–208.
24. GuhaThakurta,D. and Stormo,G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.
25. Molkentin,J.D., Black,B.L., Martin,J.F. and Olson,E.N. (1995) Cooperative activation of muscle gene expression by MEF2 and myogenic bHLH proteins. *Cell*, **83**, 1125–1136.
26. Black,B.L., Molkentin,J.D. and Olson,E.N. (1998) Multiple roles for the MyoD basic region in transmission of transcriptional activation signals and interaction with MEF2. *Mol. Cell. Biol*., **18**, 69–77.
27. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al*. (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, **437**, 1173–1178.
28. Yu,H., Luscombe,N.M., Qian,J. and Gerstein,M. (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet*., **19**, 422–427.
29. Ball,C.A., Awad,I.A., Demeter,J., Gollub,J., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Matese,J.C., Nitzberg,M., Wymore,F. *et al*. (2005) The Stanford Microarray Database accommodates additional microarray platforms and data formats. *Nucleic Acids Res*., **33**, D580–D582.
30. Liang,Y., Diehn,M., Watson,N., Bollen,A.W., Aldape,K.D., Nicholas,M.K., Lamborn,K.R., Berger,M.S., Botstein,D., Brown,P.O. *et al*. (2005) Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. *Proc. Natl Acad. Sci. USA*, **102**, 5814–5819.
31. Bredel,M., Bredel,C., Juric,D., Harsh,G.R., Vogel,H., Recht,L.D. and Sikic,B.I. (2005) High-resolution genome-wide mapping of genetic alterations in human glial brain tumors. *Cancer Res*., **65**, 4088–4096.
32. Bredel,M., Bredel,C., Juric,D., Harsh,G.R., Vogel,H., Recht,L.D. and Sikic,B.I. (2005) Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas. *Cancer Res*., **65**, 8679–8689.
33. Diehn,J.J., Diehn,M., Marmor,M.F. and Brown,P.O. (2005) Differential gene expression in anatomical compartments of the human eye. *Genome Biol*., **6**, R74.
34. Chen,X., Cheung,S.T., So,S., Fan,S.T., Barry,C., Higgins,J., Lai,K.M., Ji,J., Dudoit,S., Ng,I.O. *et al*. (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell*, **13**, 1929–1939.
35. Juric,D., Sale,S., Hromas,R.A., Yu,R., Wang,Y., Duran,G.E., Tibshirani,R., Einhorn,L.H. and Sikic,B.I. (2005) Gene expression profiling differentiates germ cell tumors from other cancers and defines subtype-specific signatures. *Proc. Natl Acad. Sci. USA*, **102**, 17763–17768.
36. Iacobuzio-Donahue,C.A., Maitra,A., Olsen,M., Lowe,A.W., van Heek,N.T., Rosty,C., Walter,K., Sato,N., Parker,A., Ashfaq,R. *et al*. (2003) Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNA microarrays. *Am. J. Pathol*., **162**, 1151–1162.
37. Courtois,G., Morgan,J.G., Campbell,L.A., Fourel,G. and Crabtree,G.R. (1987) Interaction of a liver-specific nuclear factor with the fibrinogen and alpha 1-antitrypsin promoters. *Science*, **238**, 688–692.
38. Mendel,D.B., Khavari,P.A., Conley,P.B., Graves,M.K., Hansen,L.P., Admon,A. and Crabtree,G.R. (1991) Characterization of a cofactor that regulates dimerization of a mammalian homeodomain protein. *Science*, **254**, 1762–1767.
39. Parviz,F., Matullo,C., Garrison,W.D., Savatski,L., Adamson,J.W., Ning,G., Kaestner,K.H., Rossi,J.M., Zaret,K.S. and Duncan,S.A. (2003) Hepatocyte nuclear factor 4alpha controls the development of a hepatic epithelium and liver morphogenesis. *Nature Genet*., **34**, 292–296.
40. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.
41. Horak,C.E., Luscombe,N.M., Qian,J., Bertone,P., Piccirrillo,S., Gerstein,M. and Snyder,M. (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev*., **16**, 3017–3033.
42. Zhu,Z., Pilpel,Y. and Church,G.M. (2002) Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J. Mol. Biol*., **318**, 71–81.
43. Blackshaw,S., Fraioli,R.E., Furukawa,T. and Cepko,C.L. (2001) Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes. *Cell*, **107**, 579–589.