Genomics 92 (2008) 316–321

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

# Assessing TF regulatory relationships of divergently transcribed genes

Lan Chen [a,c,1], Lun Cai [a,c,1], Geir Skogerbø [b], Yi Zhao [a,*], Runsheng Chen [a,b,*]

[a] Bioinformatics Research Group, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[b] Bioinformatics Laboratory, Institute of Biophysics, Chinese Academy of Sciences, Datun Road 15, Beijing 100101, China
[c] Graduate School of the Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

*Article history:*
Received 17 December 2007
Accepted 24 July 2008
Available online 16 September 2008

*Keywords:*
Divergently transcribed genes
Gene regulation
ChIP-chip data
Microarray data

ABSTRACT

Ambiguously located transcription factor (TF) binding sites may introduce a large number of potentially erroneous regulatory associations into models of transcriptional regulatory networks. We have used a two-step expression similarity strategy to distinguish between likely and unlikely regulatory associations for TFs located between divergently transcribed genes in the yeast genome. Most regulatory associations of divergently transcribed genes could be assigned to either high-confidence (HC) or low-confidence (LC) groups. In support of our result, we found that most of the previously characterized regulatory associations reported in the literature fell into the HC group rather than the LC group. Moreover, genomic distance analysis showed that TF binding sites tend to be located in relative proximity to the gene that is most likely to be regulated by this TF. Finally, removal of low-confidence (i.e., most probably erroneous) regulatory associations from the transcriptional regulatory network barely affected its basic architecture.



## Introduction

To understand the mechanisms of regulating transcription, it is necessary to identify transcription factors (TFs) and their binding sites (TFBSs) within the genome. Various genetic and biochemical techniques, including ChIP-on-chip technology [1], have been widely used in the identification of TF binding sites of gene promoter regions across the genome. In 2001, Simon et al. [2] identified the binding sites of nine cell-cycle TFs in yeast, and Lee et al. [3] later systemically analyzed the genome-wide locations of binding sites for 106 yeast TFs. In 2004, the binding sites of 203 TFs were determined under various environmental conditions [4]. The information related to these binding sites has been used to analyze the mechanisms of transcriptional regulation [5,6], and has been further collected into transcription factor databases, such as TRANSFAC [7], YEASTRACT [8], and SCPD [9].

Since intergenic distances in the yeast genome are not very large, the region between two divergently transcribed genes inevitably contains the promoters of both genes. We extracted all the genomic binding sites of the two transcription factors Gal4 and Step12 from the experiments of Young and co-workers [1]. In these data, we found that 11 out of 45 promoters bound by Gal4, and 6 out of 39 promoters bound by Step12, are located between two divergently transcribed genes. For large-scale genome-wide ChIP-on-chip data [3,4], we found that 35% of the sites bound by one or more TF were located upstream of two divergently transcribed genes, indicating that a large fraction of the TF binding sites fall within this type of regions. However, once a TF binds to such a region, the ChIP-on-chip data do not enable us to distinguish which of the two divergently transcribed genes is really regulated by the mentioned TF, or whether both of them are. Therefore, a considerable number of the annotated regulatory associations between TFs and regulated genes may actually be incorrect. If these genome-wide location data are assembled into transcriptional regulatory networks [5,6] and incorporated into transcription factors databases [8], they could bias the analysis, leading to flawed conclusions regarding their biological functions.

A number of studies have utilized expressional correlations between TFs and their target regulatory genes in order to assess the regulatory annotations from ChIP-on-chip data [1,10]. For example, in the supporting web sites of two subsequent studies reporting results from ChIP-on-chip technology [3,4], the authors recommend that the gene expression data could be used to "discipline" the data of divergently transcribed genes, which was also done in their previous study on the two TFs [1]. Other studies focused on evaluating ChIP-on-chip genomic location data also based their analysis on the correlation between a TF and its candidate regulated genes [10]. Boulesteix and Strimmer [10] used a statistical approach to infer true transcription factor activities from a combination of mRNA expression and DNA-protein binding measurements. However, the noise in the expression data and the fact that a gene may be regulated by several TFs affect the accuracy of such predictions. As a consequence, these procedures are only able to assess a part of such regulatory associations.

The present report describes a new strategy designed to filter out erroneous regulatory associations from divergently transcribed genes. The strategy is based on the expressional correlations among the

* Corresponding authors. R. Chen, Bioinformatics Research Group, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China. Y. Zhao, Bioinformatics Laboratory, Institute of Biophysics, Chinese Academy of Sciences, Datun Road 15, Beijing 100101, China.
*E-mail addresses:* biozy@ict.ac.cn (Y. Zhao), crs@sun5.ibp.ac.cn (R. Chen).
[1] These authors contributed equally to this work.

0888-7543/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2008.07.007

target genes regulated by the same TF, and involves two steps (see Supplementary Figure 1): In the first step we identified reliable target genes for each TF and calculated their average co-expression levels under different experimental conditions. These co-expression levels can be considered as a co-expression profile (Profile 1) of all the target genes for a specific TF, and to some extent reflect the regulatory effects of this TF on its target genes. In the second step, for each of the divergently transcribed genes bound by a specific TF, we calculated the co-expression levels between this candidate gene and the reliable target genes of the same TF (Profile 2) under the same set of experimental conditions. Profile 2 reveals the regulatory effects of this TF on the candidate gene. If the TF actually regulates the expression of the candidate gene, a strong correlation between the two profiles would be expected. Thus, the correlation between the two profiles reflects the probability with which the TF may be regulating the candidate gene. A similar idea has employed for the detection of co-operational relations among multiple TFs [11].

Our approach has two advantages compared to previous studies. First, our method calculates the average expressional correlations among multiple genes regulated by the same TF. It does not just depend on expressional correlations between single TF-gene pairs, and therefore should be more robust. Second, taking into account the variation in the expression of genes under different conditions will collect more information concerning the relationship between TFs and the regulated genes, and hence will lead to being more able to distinguish between correct and erroneous regulatory annotations.

This study reports on the assessment of the reliability of transcriptional annotations of the divergently transcribed genes through the above described method. The results were validated through comparison with the available literature. The correlation between the TF binding sites and the reliability of the regulatory associations were also assessed. Furthermore, the effects of low-confidence (i.e., likely incorrect) regulatory association on the motifs of the transcriptional regulatory network were examined.

## Results

### *Disciplining divergently transcribed genes with microarray data*

In total we identified 1274 divergently transcribed gene pairs whose common promoter regions were bound by 1 to 27 different TFs. If a TF binds to the common promoter region of a divergently transcribed gene pair, this TF and each gene of this gene pair constitute a potential regulatory association, which need to be assessed by our method. Altogether we found 2901 candidate regulatory association pairs, corresponding to 5802 candidate regulatory associations. Our method requires a set of established regulatory associations for each TF. However, at present only a limited number of established regulatory associations between yeast TFs and their target genes are available in the literature. Therefore, we extracted regulatory associations from nondivergently transcribed genes of the ChIP-on-chip data from Young's group [3,4] (see *Materials and Methods*) and named these "reliable associations" (RA). A two-step correlation value was calculated for each candidate regulatory association (see *Materials and Methods*).

Depending on whether this two-step correlation value was above or below the given threshold (see *Materials and Methods*), a regulatory association was defined as being either high-confidence (HC) or low-confidence (LC). A total of 2572 (44.4%) of the 5802 candidate regulatory associations fell within the HC group and 2172 (37.4%) within the LC group, while the remaining 1058 (18.2%) were referred as *nonassigned* due to the lack of expression profiles or of corresponding RA data. The 2901 candidate regulatory association pairs could be categorized into three groups. Group 1 contains 696 pairs (24%) for which both regulatory associations were assigned to the HC group, suggesting that the corresponding TF may be regulating both of the divergently transcribed genes. Group 2 contains 914 (31.5%) pairs, within which only one regulatory association of each pair was assigned to the HC group and the other to the LC group. Hence around half of the annotated regulatory associations in this group are probably incorrect. The remaining 1291 (44.5%) pairs fell into Group 3, in which both regulatory associations of a pair were either assigned as LC or nonassigned.

### *Experimental evidence for high confident regulatory associations*

To evaluate our method, we collected well-studied and experimentally verified regulatory associations of six cell cycle TFs and their target genes (see *Materials and Methods*). Of the 27 regulatory associations also found in our data, 23 (85%) fell in the HC group, while only 4 (15%) fell in the LC group, thus, most of verified regulatory associations were assigned as "high confidence" regulatory associations, indicating that our method is able to identify true regulatory relationships. Furthermore, we downloaded a large set of curated regulatory associations in *Saccharomyces cerevisiae* from the database YEASTRACT (see *Materials and Methods*) [8]. Of the regulatory associations downloaded from this database, 683 have been assigned as HC or LC associations in our analysis. Of the 2572 HC and 2172 LC regulatory associations in our results, a significantly ($p$-value$<9\times10^{-30}$ by Fisher's Exact Test) higher number of HC (19.6%) than LC (8.2%) regulatory associations were supported by the experimental evidence. The above data thus all strongly suggest that our evaluation of candidate regulatory associations from divergently transcriptional gene pairs accords with results from traditional experimental work.

### *TF binding site location bias of divergently transcribed genes*

Transcription factor binding sites (TFBS) commonly show significant positional preferences relative to the transcriptional start sites of human genes, and TFBSs of bidirectional promoters exhibit less positional specificity than that of unidirectional promoters [12]. We therefore investigated the positional preferences of TFBSs relative to the transcriptional start sites of divergently transcribed genes in yeast. In order to do so, we downloaded TFBS data with "moderate conservation" cutoff and $p$-value of 0.005 from Harbison et al. [4]. The distance from the TFBS to a gene (TFBS distance) was calculated as the number of base pairs from the center of the TFBS to the transcriptional start site of the gene. If multiple TFBSs corresponded to one TF, the positional center of these TFBSs was taken to represent the binding site of the TF. Only TFBS distances smaller than 2 kb were considered for analysis.

TFBS distance information was established for 1389 candidate regulatory association pairs. For the 498 pairs with one regulatory association in the HC group and the other in the LC group, the average TFBS distance for the regulatory associations in the HC group (363 bp, SD 244 bp) was much smaller than that in the LC group (466 bp, SD 315 bp). These results suggest that in divergently transcriptional gene pairs, the TF binds closer to the gene it actually regulates.

Moreover, in order to obtain detailed information regarding the positional preferences of TF binding sites, the difference in distance between the TFBS and the two corresponding divergently transcribed genes was calculated for each regulatory association pair. For pairs with both regulatory associations falling in the HC group (i.e., HC/HC), the cumulative distribution of distance differences (absolute values) was significantly different ($p$-value=0.05, Kolmogorov-Smirnov test) from pairs with one regulatory association in the HC group and the other in the LC group (i.e., HC/LC), the absolute TFBS distances differences consistently being larger in HC/LC than in HC/HC pairs (Fig. 1A). This suggests that when both of the divergently transcribed genes are regulated by a common upstream TF, the TF binding site

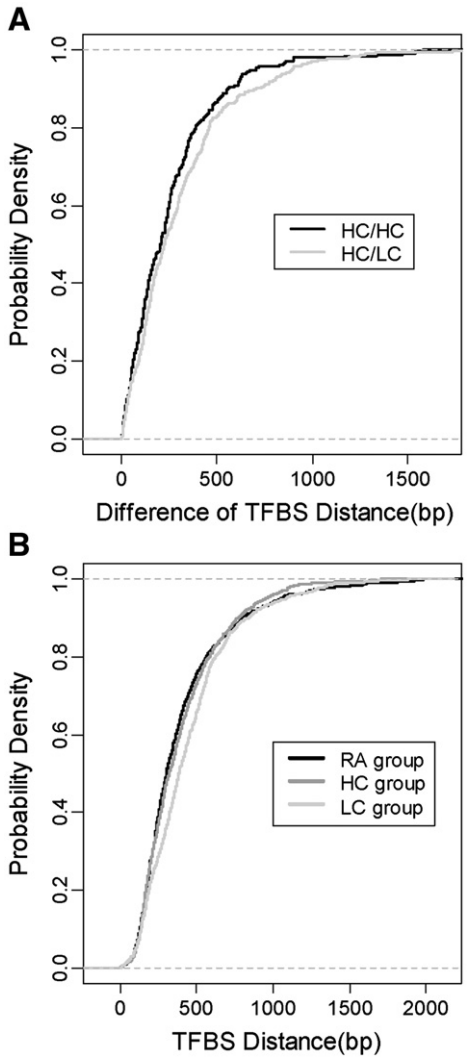


**Fig. 1.** Distribution of the distances between TFBS and gene transcriptional start sites. (A) Cumulative distributions of differences in distance between TFBS and two transcriptional start sites of divergently transcribed genes, for HC/HC pairs and HC/LC pairs. The cumulative distribution of distance difference of HC/LC pairs being generally lower than that of HC/HC pairs indicates that a higher proportion of the distance differences of the HC/LC pairs have larger values. (B) Cumulative distributions of distances between TFBS and the transcriptional start sites of genes in the RA, HC and LC groups. The distribution of the LC group is generally lower than that of the RA and HC groups.

tends to be positioned around the center of the intervening region, while when only one of divergently transcribed genes is regulated by the TF, its binding site tends to be closer to the actually regulated gene.

To further investigate the effects of the confidence level on the TFBS distance, we compared the cumulative TFBS distance distributions for regulatory associations in the RA (nondivergent genes), HC, and LC groups (Fig. 1B). The TFBS distance distribution of the HC group was not significantly different ($p$-value = 0.40, Kolmogorov-Smirnov test) from that of the RA group, whereas the TFBS distance distribution of the LC group was significantly different from the two others ($p$-value = $2.16 \times 10^{-07}$ and $p$-value = $1.47 \times 10^{-11}$ for the HC and the RA comparisons, respectively, Kolmogorov-Smirnov test). The TFBS distances in the LC group were generally longer than those in the two other groups. Our results suggest that, even in the compact yeast genome, the distance between the *cis*-regulatory element and core promoter may influence the strength and choice of transcriptional regulatory associations.

### *LC regulatory associations and transcriptional regulatory network motifs*

Incorrect regulatory associations may affect the topological structure of transcriptional regulatory network constructed from ChIP-on-chip data. Our analysis suggests that divergently transcribed genes can be divided into regulatory associations with different confidence levels. To investigate the potential impact of purportedly incorrect regulatory associations, we progressively removed LC regulatory associations from the original transcriptional regulatory network, and examined the numerical changes in six network motifs in the remaining transcriptional regulatory network. As a control, the same number of HC regulatory associations were also progressively removed from the same network (see *Materials and Methods*). The changes in number of network motifs resulting from this procedure were then examined.

The six network motifs examined included autoregulations (a regulator that binds to its own promoter), multicomponent loops (a regulatory circuit involves a chain of two or more regulators in which the last regulator binds the first), regulator chains (three or more regulators acting in sequence), feedforward loops (in its simplest form a feedforward loop contains a regulator controlling a second regulator, both regulating a common target gene), single-input motifs (a single regulator controlling a set of target genes), and multi-input motifs (a set of three or more regulators acting on a common set of three or more target genes) [3]. A common characteristic of the first four types of motifs is that they reflect the complexity of the relationships among the regulators, and to some extent represent the framework of the transcriptional regulatory network. Removing LC regulatory associations resulted in a much slower decrease in the numbers of these motifs than when the same number of HC regulatory associations were removed (Figs. 2A–D), indicating that HC and LC regulatory associations represent very different types of network edges. The fact that the removal of LC regulatory associations has only a limited effect on the motifs of the transcriptional regulatory network suggests that these edges largely represent false regulatory annotations. Because such edges are randomly inserted into the network, their impact on the overall structure (or framework) of the network is of minor importance.

The fifth type of network motif, the single-input motif, offers the potential for a single TF to regulate a subset of genes in a metabolic pathway [3]. The last type of motif, the multi-input motif, may potentially be coordinating gene expression under various conditions [3]. Unlike the first four motifs mentioned above, these latter two types of motifs represent the relationship between regulators and target genes, and can be seen as corresponding to the detailed structures of the network. Removal of HC and LC regulatory associations from the transcriptional regulatory network had similar impacts on both single-input motifs and multi-input motifs (Figs. 2E and 2F), showing that introducing low-confidence (or incorrect) regulatory associations would have an impact on the detailed structure of the transcriptional regulatory network.

The results listed above indicate that incorporation of low-confidence regulatory associations barely affects the overall architecture of the transcriptional regulatory network, while it significantly affects the detailed structure of the network (i.e., the relationships between TF and its target genes). To some extent, these results suggest that the transcriptional regulatory network can tolerate a number of false positive edges without much change to the basic architecture of the network.

## Discussion

To obtain genome-wide gene regulatory data, transcription factor binding site locations have been examined with ChIP-on-chip technology [1–4]. When a TF binds to a specific genomic site located

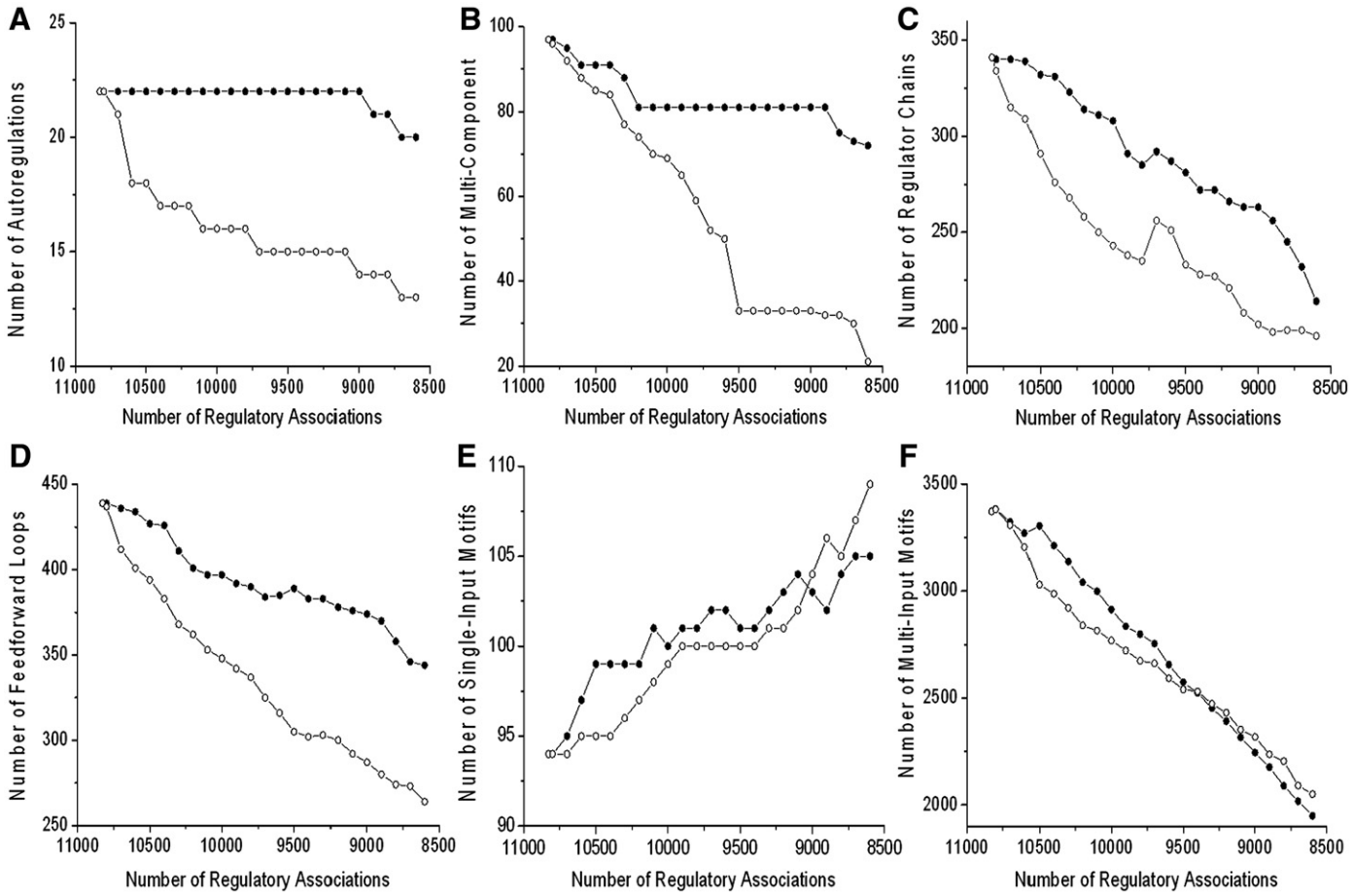


**Fig. 2.** Effects caused by the removing of regulatory associations in either LC group or the HC group on six kinds of network motifs. After progressive removal of the same number of regulatory associations (*X*-axis indicates the number of remaining regulatory associations after removing) in the LC group (LC out, line with filled circle) or the HC group (HC out, line with empty circle) from the transcriptional regulatory network, the numbers of autoregulations (A), multicomponent loops (B), regulator chains (C), feedforward loops (D), single-input motifs (E), and multi-input motifs (F) were examined.

upstream of two divergently transcribed genes, it is not straightforward to determine which of the two genes is regulated by this TF. On the other hand, assigning the same TF as a regulatory factor of both the divergently transcribed genes may introduce a number of incorrect regulatory associations into the transcriptional regulatory networks constructed from these data. Our study used large-scale microarray data to evaluate the confidence levels of regulatory annotations of the divergently transcribed genes. The results were supported by experimentally validated regulatory relationships reported in the literature. Moreover, analysis of the TFBSs-gene distance relations as well as network motifs further supported our evaluations.

The apparent correlation between TFBS proximity and the confidence level of regulatory associations of divergently transcribed genes is an interesting observation. Although some have shown that there is an effect of the distance between a gene and a TFBS, particularly at very short distances (a few tens of base pairs [13–15]), effects of longer distances appear not to have been investigated in any great detail [16]. It is also evident from mammalian studies not only that *cis*-regulatory elements may be active over very large distances (1 Mb or more [17,18]), but also that when cloned in close proximity to a reporter gene, such distant *cis*-regulatory elements are able to drive reporter expression in a way that resembles the endogenous regulatory structure. Thus, strong effects of TFBS-gene distances are perhaps not to be expected, and one might assume that the relatively compact nature of the yeast genome would further preclude any significant role for TFBS-gene distances. Nevertheless, the data presented here could suggest that intermediate distances (i.e., a few hundreds of base pairs) might be playing a role in determining the strength of regulatory associations.

Biological networks are systems constructed from a variety of large-scale data, such as protein-protein, protein—DNA, and protein-metabolite interactions, most of which are found to be robust (reviewed in [19]). Our results show that the overall structure of the transcriptional regulatory network constructed from the large-scale ChIP-on-chip data in yeast has the ability to tolerate incorrect regulatory associations. However, incorporating the less reliable regulatory associations can significantly affect the detailed structures of the network, and removal of such likely false edges from the network may be important for the network's ability to guide future experimental work.

## Materials and methods

### *ChIP-on-chip data*

Two sets of genome-wide ChIP-on-chip location data were obtained from Young's group [3,4] with a threshold of 0.001 [1]. The first dataset contains 107 TFs that bind the promoter regions of 2364 genes, constituting 4359 potential regulatory associations [3]. The second dataset contains 180 TFs that bind the promoter regions of 3621 genes, forming 11,297 potential regulatory associations [4]. Altogether there were 188 TFs that bound the promoter regions of 4029 genes, constituting a transcriptional regulation network involving 12,637 potential regulatory associations, of which 5802

regulatory associations were related to the divergently transcribed genes and assessed in the above analyses.

*Microarray data*

We integrated yeast microarray datasets from the Stanford Microarray Database [20], the Rosetta Compendium data [21], and the NCBI Gene Expression Omnibus [22]. These microarray datasets were divided into several expression files according to the similarities of their experimental conditions. As indicated by a previous study, even after normalization for microarray datasets, the average correlations of random gene pairs are not always zero [23]. Expression files with average correlations deviating significantly from zero may lead to unwanted noise, and we therefore deleted expression files whose absolute mean correlation values were larger than 0.05. Moreover, genes with missing data more than 50% in each file were discarded. At last, there were 36 expression files left, with 7 to 56 experimental data points in each file (in total 452 experiment data points). Each expression file contains 5397 genes (see Supplementary Material).

*Cell cycle regulatory associations*

Regulation of the yeast cell cycle has been widely studied in the past decades. We therefore downloaded experimentally verified regulatory associations of six cell cycle TFs collected by Spellman et al. [20] from the Web site http://genome-www.stanford.edu/cellcycle/data/rawdata/. Totally, there were 70 cell cycle regulatory associations that overlapped with the regulatory associations derived from the ChIP-on-chip data studied in this work. Among these, 27 regulatory associations fell within the HC or LC groups (see supplementary material), and the remainder belong to the RA or *nonassigned* groups.

*Regulatory associations from YEASTRACT*

In total, 26,484 curated regulatory associations were downloaded from YEASTRACT [8] (http://www.yeastract.com/) with the parameter "Documented." To ensure that the validation data set was independent of our input data, we deleted from the YEASTRACT data all regulatory associations derived from the ChIP-on-chip data [3,4] and the microarray data (above) utilized in this work. This left us with 683 curated regulatory associations overlapping regulatory associations in the HC or LC groups.

*Two-step correlation calculation*

For each TF we first obtained the corresponding RA group from the ChIP-on-chip data. If a TF binding site is located in a promoter region that only corresponds to a single gene (i.e., nondivergently transcribed genes), we consider the corresponding regulatory association as being less ambiguous than a regulatory association of a divergently transcribed gene. Based on this idea, we extracted regulatory associations of the nondivergently transcribed genes for all TFs from the ChIP-on-chip data. We then used the expressional profiles from the microarray data (above) to calculate the expressional correlations among the nondivergently transcribed genes for each TF. Individual genes showing a negative average expressional correlation to all other gene were removed. The regulatory associations corresponding to the remaining genes were assigned as the "reliable associations" (RA) of this TF. To reduce the number of false positives, only TFs with more than 10 RA genes were included in the analysis.

If the binding sites of a TF were located in the common promoter region of two divergently transcribed genes, this TF and each of the divergently transcribed genes constitute potential regulatory associations. A two-step correlation was calculated for each potential regulatory association. In the first step, we extracted all genes in the RA group for a given TF and calculated the average Pearson co-expression correlation of these genes in each of the 36 expression files. For each TF, the average co-expression correlations formed a vector representing the variation in the level of co-expression of its target genes. We then extracted each gene from the divergently transcribed pairs bound by the same TF and calculated the average co-expression correlation between this gene and all the genes in the RA group in the same set of expression files. These average co-expression correlations in the 36 expression files formed another vector, which represented the co-expression level between this gene and the genes in the RA group. In the second stage of the analysis, a two-step correlation was obtained by calculating the Pearson correlation between the two vectors (see Supplementary Figure 1). Candidate regulatory associations with two-step correlations higher than our pre-set threshold were assigned to the "high-confidence (HC)" group, whereas those with correlations lower than the threshold were assigned to the "low-confidence (LC)" group.

*Determination of the two-step correlation threshold*

To determine the threshold value for the two-step correlation, we compared the two-step correlation distributions between regulatory associations in the RA group and randomly selected TF-gene pairs. The procedure to calculate the two-step correlation of RA group is as follows: For a given TF, we took one regulatory association out of its RA group. The remaining regulatory associations in this RA group were regarded as a "new" RA group of this TF. Then the two-step correlation was calculated between this "out" regulatory association and the "new" RA group. We repeated this process for all the regulatory associations in this RA group, and then for all TFs; at last we obtained the two-step correlation distribution of the RA groups for all the TFs. To calculate the two-step correlation distribution of randomly selected TF-gene pairs, we created 5000 random TF-gene pairs, by randomly selecting a TF from all TFs with an RA group, and then randomly selecting target genes to assign to this TF from the 5397 yeast genes. Two-step correlations were calculated for all the randomly selected TF-gene pairs. The two-step correlations distributions of the RA group and randomly selected TF-gene pairs were plotted (Fig. 3) As shown in the figure, the two-step correlations distribution of the randomly selected TF-gene pairs was approximately symmetric, while for the RA group the distribution was asymmetric with a peak much higher than that of the randomly selected TF-gene pairs. As a compromise between false positives and

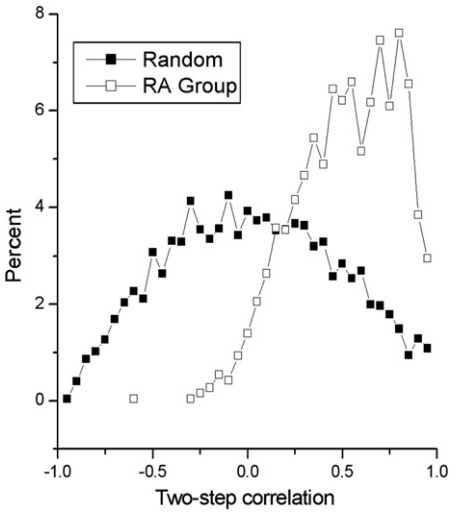


**Fig. 3.** Two-step correlation distributions. Distributions of two-step correlations of RA group (RA Group) and randomly selected TF-gene pairs (Random).

false negatives, the cross point (0.2) of these two distributions was selected as the threshold value. Candidate regulatory associations with two-step correlation values above this threshold were assigned as high-confidence (HC), whereas those below it were assigned as low-confidence (LC).

*Removing the regulatory associations from the transcriptional regulatory network*

Each HC or LC regulatory association has a two-step correlation value, which can be seen as a measure of the reliability of this regulatory association. The LC regulatory associations were sorted by their two-step correlation values from lowest to the highest. LC regulatory associations were removed from the transcriptional regulatory network according to this order; that is, the least reliable (or most probably incorrect) regulatory associations were removed first. When the HC regulatory associations were removed from the transcriptional regulatory network, the most reliable regulatory associations were removed first (that is, the edges with the highest two-step correlation values).

*Software tools and available results*

The statistical analysis was performed with the R software [24]. The evaluation results of all the regulatory associations of divergently transcribed gene pairs are available at http://www.ebiomed.org/pub/diverg.html.

## Acknowledgments

We are grateful to Lisa Caviglia for her elaborate correction of this manuscript, and to Liu Changning and He Shunmin for helpful suggestions on this manuscript. This work was supported by the National Key Basic Research and Development Program (973) under Grants 2002CB713805 and 2003CB715907, the National Sciences Foundation of China under Grants 30630040, 30570393 and 30600729, and the Data Sharing Network of China Essential Medicine Science under Grant 2005DKA32402.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2008.07.007.

## References

[1] B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T.L. Volkert, C.J. Wilson, S.P. Bell, R.A. Young, Genome-Wide Location and Function of DNA Binding Proteins, Science 290 (2000) 2306–2309.
[2] I. Simon, J. Barnett, N. Hannett, C.T. Harbison, N.J. Rinaldi, T.L. Volkert, J.J. Wyrick, J. Zeitlinger, D.K. Gifford, T.S. Jaakkola, R.A. Young, Serial Regulation of Transcriptional Regulators in the Yeast Cell Cycle, Cell 106 (2001) 697–708.
[3] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J.-B. Tagne, T.L. Volkert, E. Fraenkel, D.K. Gifford, R.A. Young, Transcriptional Regulatory Networks in Saccharomyces cerevisiae, Science 298 (2002) 799–804.
[4] C.T. Harbison, D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, J.-B. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K. Pokholok, M. Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, R.A. Young, Transcriptional regulatory code of a eukaryotic genome, Nature 431 (2004) 99–104.
[5] A.J.M. Walhout, Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping, Genome Res. 16 (2006) 1445–1454.
[6] N.M. Luscombe, M. Madan Babu, H. Yu, M. Snyder, S.A. Teichmann, M. Gerstein, Genomic analysis of regulatory network dynamics reveals large topological changes, Nature 431 (2004) 308–312.
[7] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, M. Pru{beta}, F. Schacherer, S. Thiele, S. Urbach, The TRANSFAC system on gene expression regulation, Nucl. Acids Res. 29 (2001) 281–283.
[8] M.C. Teixeira, P. Monteiro, P. Jain, S. Tenreiro, A.R. Fernandes, N.P. Mira, M. Alenquer, A.T. Freitas, A.L. Oliveira, I. Sa-Correia, The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae, Nucl. Acids Res. 34 (2006) D446–451.
[9] J. Zhu, M. Zhang, SCPD: a promoter database of the yeast Saccharomyces cerevisiae, Bioinformatics 15 (1999) 607–611.
[10] A.-L. Boulesteix, K. Strimmer, Predicting Transcription Factor Activities from Combined Analysis of Microarray and ChIP Data: A Partial Least Squares Approach, Theoret. Biol. Med. Modelling 2 (2005) 23.
[11] X.J. Zhou, M.-C.J. Kao, H. Huang, A. Wong, J. Nunez-Iglesias, M. Primig, O.M. Aparicio, C.E. Finch, T.E. Morgan, W.H. Wong, Functional annotation and network reconstruction through cross-platform integration of microarray data, Nat. Biotechol. 23 (2005) 238–243.
[12] J.M. Lin, P.J. Collins, N.D. Trinklein, H.X. Yutao Fu3, R.M. Myers, Z. Weng, Transcription factor binding and modified histones in human bidirectional promoters, Genome Res. 17 (2007) 818–827.
[13] C.A. Spek, R.M. Bertina, P.H. Reitsma, Unique distance- and DNA-turn-dependent interactions in the human protein C gene promoter confer submaximal transcriptional activity, Biochem. J. 340 (Pt 2) (1999) 513–518.
[14] W.P. Tansey, F. Schaufele, M. Heslewood, C. Handford, T.L. Reudelhuber, D.F. Catanzaro, Distance-dependent interactions between basal, cyclic AMP, and thyroid hormone response elements in the rat growth hormone promoter, J. Biol. Chem. 268 (1993) 14906–14911.
[15] R. Schule, M. Muller, H. Otsuka-Murakami, R. Renkawitz, Cooperativity of the glucocorticoid receptor and the CACCC-box binding factor, Nature 332 (1988) 87–90.
[16] G.A. Wray, M.W. Hahn, E. Abouheif, J.P. Balhoff, M. Pizer, M.V. Rockman, L.A. Romano, The Evolution of Transcriptional Regulation in Eukaryotes, Mol. Biol. Evol. 20 (2003) 1377–1419.
[17] L.A. Lettice, T. Horikoshi, S.J.H. Heaney, M.J. van Baren, H.C. van der Linde, G.J. Breedveld, M. Joosse, N. Akarsu, B.A. Oostra, N. Endo, M. Shibata, M. Suzuki, E. Takahashi, T. Shinka, Y. Nakahori, D. Ayusawa, K. Nakabayashi, S.W. Scherer, P. Heutink, R.E. Hill, S. Noji, Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly, Proc. Nat. Acad. Sci. U.S.A. 99 (2002) 7548–7553.
[18] T. Vavouri, G.K. McEwen, A. Woolfe, W.R. Gilks, G. Elgar, Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key, Trends Genet. 22 (2006) 5–10.
[19] A.-L. Barabasi, Z.N. Oltvai, NETWORK BIOLOGY: UNDERSTANDING THE CELL'S FUNCTIONAL ORGANIZATION, Nat. Rev. Genet. 5 (2004) 101–113.
[20] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization, Mol. Biol. Cell 9 (1998) 3273–3297.
[21] T. Babak, B.J. Blencowe, T.R. Hughes, A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription, BMC Genom. 6 (2005) 104.
[22] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, R. Edgar, NCBI GEO: mining tens of millions of expression profiles-database and tools update, Nucl. Acids Res. 35 (2007) D760–D765.
[23] A. Ploner, L. Miller, P. Hall, J. Bergh, Y. Pawitan, Correlation test to assess low-level processing of high-density oligonucleotide microarray data, BMC Bioinformat. 6 (2005) 80.
[24] R Development Core Team, R: a language and environment for statistical computing, R Foundation for StatisticalComputing, Vienna, Austria, 2005.