Genomics 96 (2010) 352–361

Contents lists available at ScienceDirect

Genomics

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

Genome-wide analysis of the *cis*-regulatory modules of divergent gene pairs in yeast

Chien-Hao Su [a,b], Ching-Hua Shih [a,1], Tien-Hsien Chang [c,d], Huai-Kuang Tsai [a,e,*]

[a] *Institute of Information Science, Academia Sinica, Taipei 115, Taiwan*
[b] *Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan*
[c] *Department of Molecular Genetics, The Ohio State University, Columbus, OH 43210, USA*
[d] *Genomics Research Center, Academia Sinica, Taipei 115, Taiwan*
[e] *Research Center for Information Technology Innovation, Taipei 115, Taiwan*

ARTICLE INFO

Article history:
Received 22 March 2010
Accepted 27 August 2010
Available online 6 September 2010

Keywords:
Yeast
Divergent gene pairs
Regulatory mechanisms
Transcription factor binding sites
cis-regulatory modules

ABSTRACT

In budding yeast, approximately a quarter of adjacent genes are divergently transcribed (divergent gene pairs). Whether genes in a divergent pair share the same regulatory system is still unknown. By examining transcription factor (TF) knockout experiments, we found that most TF knockout only altered the expression of one gene in a divergent pair. This prompted us to conduct a comprehensive analysis *in silico* to estimate how many divergent pairs are regulated by common sets of TFs (*cis*-regulatory modules, CRMs) using TF binding sites and expression data. Analyses of ten expression datasets show that only a limited number of divergent gene pairs share CRMs in any single dataset. However, around half of divergent pairs do share a regulatory system in at least one dataset. Our analysis suggests that genes in a divergent pair tend to be co-regulated in at least one condition; however, in most conditions, they may not be co-regulated.

1. Introduction

A divergent (head-to-head or bidirectional) gene pair comprises two adjacent genes, whose transcription start sites (TSSs) are located on opposite strands of DNA with adjacent 5′-ends. It is often assumed that the intergenic regions flanked by such pairs are capable of initiating transcription in both directions. The *GAL1–GAL10* gene pair in budding yeast is a classic example because its intergenic region drives the expression of both genes upon galactose induction [1]. In humans, the number of divergent gene pairs with TSS separated by less than 1000 base pairs is substantial, constituting more than 10% of the human genome [2]. Computational analysis of various biological systems reveals that some divergent gene pairs exhibit highly correlated expression patterns [2–5]; and, in some cases, they appear to share regulatory elements [6–11], suggesting that they could be co-regulated. A study by Collins et al. [12] provided further experimental support for the model. They showed that GA-binding protein (GABP) binds to more than 80% of the intergenic regions in at least one human cell type, and that the binding is correlated with bidirectional transcription activity. Using a reporter assay, they further demonstrated that, in four out of six cases, introduction of a consensus GABP site into promoters that have little or no expression in the reverse direction results in a significant increase in transcription activity in the reverse direction. Thus, there is a high probability that genes in a divergent pair share the regulatory system.

If genes in a divergent pair do share the regulatory system with functional consequences, one would expect their intergenic region to be under higher evolutionary constraint. In addition, the divergent gene pairs would be more likely to co-express than other types of adjacent gene pairs (*i.e.*, convergent and tandem pairs) [13]. However, the proportions of adjacent genes that share similar expression patterns do not differ significantly, regardless of their adjacent types [6,14,15]. For example, our previous analysis [16] of five yeast species demonstrated that the probability of conservation of each of the three adjacent types is low and the difference in the co-expression level is not statistically significant. Similarly, Yanai and Hunter [17] found that the co-expression of gene neighbors in related nematodes is highly divergent and probably evolves under neutral processes. These findings seem to contradict the notion that the intergenic regions of divergent gene pairs need to be highly conserved during evolution. To identify the regulatory mechanism involved in divergent gene pairs, we examined 263 transcription factor (TF) knockout experiments. Our findings indicated that, in most instances, TF knockout only altered the expression of one gene in a divergent pair (for details, see Results and discussion). Thus, we were motivated to conduct a more comprehensive analysis of budding yeast to estimate how many divergent gene pairs share the regulatory system.

We integrated the annotations of TF binding sites (TFBSs) and microarray expression data to infer the *cis*-regulatory modules (CRMs, *i.e.*, common sets of TFs) of genes in divergent pairs in *Saccharomyces*

[*] Corresponding author. Institute of Information Science, Academia Sinica, 128 Sec. 2, Academia Rd, Nankang, Taipei, Taiwan. Fax: +886 2 27824814.
E-mail addresses: d95922033@ntu.edu.tw (C.-H. Su), Ching-Hua.Shih@rice.edu (C.-H. Shih), chang108@gate.sinica.edu.tw (T.-H. Chang), hktsai@iis.sinica.edu.tw (H.-K. Tsai).
[1] Present address: Department of Ecology and Evolutionary Biology, Rice University, Houston, TX 77005, USA.

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.08.008

cerevisiae. Our analysis indicates that only a limited number of divergent gene pairs may share CRMs in a given dataset. However, we also found that approximately half of the divergent pairs share a regulatory system in at least one dataset when all the data collected under different conditions are considered together.

2. Materials and methods

2.1. Identification of transcription factor binding sites (TFBSs)

First, we collected the TFBS annotations of 117 TFs from MacIsaac et al. [18] (http://fraenkel.mit.edu/improved_map/orfs_by_factor.tar.gz), where a site is bound at $p<0.001$ in the location analysis of Harbison et al. [19] and conserved in three out of the four considered yeast species. For TFs without TFBS information, we collected TFBS annotations from the Mining Yeast Binding Sites (MYBS) database [20]. MYBS allows users to identify TFBSs by using DNA-binding affinity data and phylogenetic footprinting data from eight related yeast species. In this study, we chose TFBSs that are conserved in at least two of the other seven yeast species. By combining these two resources, we were able to consider the TFBS annotations of 144 TFs.

2.2. Identification of divergent gene pairs

We adopted the definition of divergent gene pairs in our previous study [16]. That is, two adjacent genes are considered a divergent pair if they are transcribed in different directions from opposite strands of DNA with adjacent 5′-ends, and the length of the intergenic region between two translation start sites is within 700 bp. According to Dobi and Winston [21], the majority of yeast promoters range from approximately 150 to 400 bases. Therefore, in our analysis, we did not consider divergent pairs whose intergenic regions were longer than 700 bases. We downloaded the sequence and annotations from SGD (http://www.yeastgenome.org) [22]. There are 5702 annotated ORFs in *S. cerevisiae*. After removing dubious, silent, and overlapping ORFs, we identified 961 divergent gene pairs.

2.3. Assigning possible cis-regulatory modules (CRMs) to genes

For two genes, G_α and G_β, in a divergent pair, we investigated whether the TFBSs occurred in the intergenic region to enumerate all possible TFBS compositions. Assuming that there were *n* TFBSs in the intergenic region, we generated a total of 2^n types of TFBS compositions, each of which was treated as a group. We then divided the non-divergent genes into different groups according to the TFBS compositions (only for these *n* TFBSs) in their promoters. As a result, each group contained a set of non-divergent genes. Note that a group with insufficient genes (*e.g.*, less than five) was discarded. This is a limitation imposed by the Kolmogorov–Smirnov (KS) statistical test used in our method, which we discuss in the following paragraph.

For each group, we used the expression coherence (EC) score [23] to quantify whether genes in the same group had similar expression profiles. The score is defined as the fraction of gene pairs in the group with a correlation higher than a threshold. For a given expression dataset, the threshold is set as the 95th percentile correlation coefficient value of all the pairwise correlation coefficients among 100 randomly selected genes. To estimate the significance of the EC scores, we used the method proposed by Lapidot and Pilpel [24]. Specifically, for each microarray dataset, and for each set size (varying from 5 to the maximal number of genes in a group), we randomly generated 1000 gene sets (with the same size and microarray dataset) and computed the EC score of each set. The *p* value of a given score was estimated as the number of random sets with scores higher than the given score divided by 1000. For groups with a *p* value ≤ 0.01, their corresponding TFBS compositions were deemed potential CRMs in the dataset. Correction for multiple testing was not applied at this stage. Instead, we applied it in a later stage when examining the degree of co-expression between G_α (or G_β) and those groups with similar expression profiles.

Next, we determined which potential CRM would be the most likely module to regulate G_α (or G_β) in the given dataset. For this, we selected groups whose gene expression profiles were similar to that of G_α (or G_β). Specifically, we let M_i be the members of group *i*, and examined whether the distribution of the *Pearson* correlation coefficients between G_α and M_i was significantly different from that between G_α and M_{bg}. (M_{bg} denotes genes in the *background set*; that is, none of the *n* TFBSs are present in their promoters). We applied the one-sided Kolmogorov–Smirnov (KS) test to examine the above statistical criteria. For the gene G_α, let $E^\alpha(M_i)$ be the set of *Pearson* correlation coefficients between G_α and M_i, and let $E^\alpha(M_{bg})$ be the set of coefficients between G_α and M_{bg}. We tested $H_0: F_{E^\alpha(Mi)} = F_{E^\alpha(Mbg)}$ against $H_1: F_{E^\alpha(Mi)} <_{st} F_{E^\alpha(Mbg)}$ using the one-sided KS test, where *F* denotes the cumulative distribution function of the correlation coefficients in a set. If H_0 is rejected, $F_{E^\alpha(Mi)} <_{st} F_{E^\alpha(Mbg)}$, which means the correlation coefficients in $E^\alpha(M_i)$ are "stochastically greater" than those in $E^\alpha(M_{bg})$.

Finally, we determined the false discovery rates (FDR) [25] by computing the *q* value to correct for possible false positives from multiple tests. The group with the smallest *q* value (≤ 0.01) was selected and its corresponding CRM was assigned to regulate G_α.

2.4. Microarray dataset

We downloaded the expression data from the Stanford Microarray Database (SMD, http://genome-www5.stanford.edu/) [26]. To avoid bias while calculating the co-expression level (the *Pearson* correlation coefficient) for two genes, we only used microarray data derived from more than nine experiments. As a result, we selected the following ten *S. cerevisiae* microarray datasets for the analysis: *crz1p* [27], *alpha* [28], *damage* [29], *glucose* [30], *oxi* [31], *HP* [32], *Hs* [33], *Os* [33], *MD* [33] and *nitrogen* [33]. The datasets contained the gene expression profiles for experiments ranging from the natural processes of the cell (*e.g.*, the cell cycle) and gene response to environmental perturbation (*e.g.*, heat shock). The statistics of the ten datasets are detailed in Table S1. We performed MA lowess [34] and quantile [35] normalization to reduce systematic biases within each microarray, as well as the intensity-dependent effects and biases between microarrays.

3. Results and discussion

3.1. Pilot study of divergent gene pairs that share a regulatory system using TF knockout experiments

To estimate how many intergenic regions of divergent gene pairs share CRMs, the most intuitive way is to check whether genes in a divergent pair are regulated by the same TF(s). Therefore, to investigate this issue, we used the TF knockout data from Hu et al.'s study [36]. Our underlying hypothesis is that, if a TF knockout affects the expression of both genes in a divergent pair, then the genes are probably regulated by the same TF(s). However, if only one of the genes undergoes a significant expression change, the genes are probably not regulated by the same TF(s). Hu et al. profiled the transcriptional responses in relation to the deletion of individual genes that correspond to 263 TFs. Then, they integrated the resulting data, assigned *p* values and identified target genes that exhibited significant differential expressions. We found that 609 divergent gene pairs exhibited significant changes in gene expression (*p* value $\leq$ 0.001) after knocking out at least one of the 263 TFs (see Additional material 1). For 477 divergent gene pairs, alteration of the gene expression was observed in just one of the two genes, but only in a fraction of TF knockouts. The remainder of the TF knockouts yielded no change for either gene. Our analysis clearly demonstrates that at least 78% (477/609) of divergent pairs do not uniformly respond to TF knockouts. This result suggests that divergent gene pairs in budding yeast may not be driven primarily by a common set of TFs in their promoters under the investigated condition. This possibility motivated us to identify the regulators of each gene in divergent pairs *in silico*. Subsequently, we estimated the proportion of divergent gene pairs that were co-regulated under different conditions.

Fig. 1 shows the model used to assign *cis*-regulatory modules (CRMs) for divergent pairs. For two genes, G_α and G_β, in a divergent pair, we took the union of their TFBSs and assumed that combinations of these *n* TFBSs constituted potential CRMs for regulating G_α and G_β. We enumerated all 2^n combinations (assuming there were *n* TFBSs) and matched each combination with non-divergent genes (genes not in divergent pairs) that had the same TFBS composition. This resulted in 2^n groups of non-divergent genes. The non-divergent genes in each group were then evaluated to determine the significance of their co-expression. The procedure yielded a collection of significantly co-expressed groups. Finally, we compared the gene expression profile of G_α (or G_β) in the investigated dataset of a particular condition with that of each group in the collection, and we deduced the most likely CRM that regulated G_α (or G_β).

3.2. Many divergent gene pairs do not share CRMs under ten investigated conditions

We applied the proposed method to the *S. cerevisiae* genome to identify the CRMs for the genes in all divergent pairs. Table 1 summarizes the CRMs identified by using ten microarray datasets (detailed results are presented in Table 2 and Additional material 2). An identified CRM has a set of significantly co-expressed non-divergent target genes and is assigned to regulate at least one gene in a divergent pair. The number of identified CRMs ranged from 8 (in the *damage* dataset) to 34 (in the *MD* dataset). Out of 1922 genes in divergent pairs, we were able to assign CRMs to 28–297 (1.46%–15.45%) genes in different expression datasets. We excluded divergent pairs in which neither gene was assigned a CRM; such pairs were considered to be *without annotation*. For each divergent pair in a given dataset, we used the following labels: *same* to indicate that both genes had the same assigned CRM; *overlapped*, if the genes' CRMs were subsets of each other; *different*, if the genes had different assigned CRMs; and *one-assigned* if only one gene had an assigned CRM. Both *different* and *one-assigned* were considered *not-shared*, while *overlapped* and *same* were considered *shared*. Interestingly, 56%–100% of the divergent pairs that were assigned CRMs belonged to the *not-shared* category (Fig. 2) in the ten investigated datasets. For example, in the *glucose* dataset, 56% and 7% of divergent pairs were labeled as *one-assigned* and *different* respectively. Thus, in this case, 63% of the divergent pairs were *not-shared*. The proportions of pairs that did not share CRMs increased to 92% and 100% respectively in the *HP* and *nitrogen* datasets. This was probably because there were only a few genes (less than 50) with assigned CRMs in the two datasets. However, even without considering these two datasets, the proportion of divergent gene pairs that shared CRMs was still low, ranging from 19% to 44% in the remaining eight datasets.

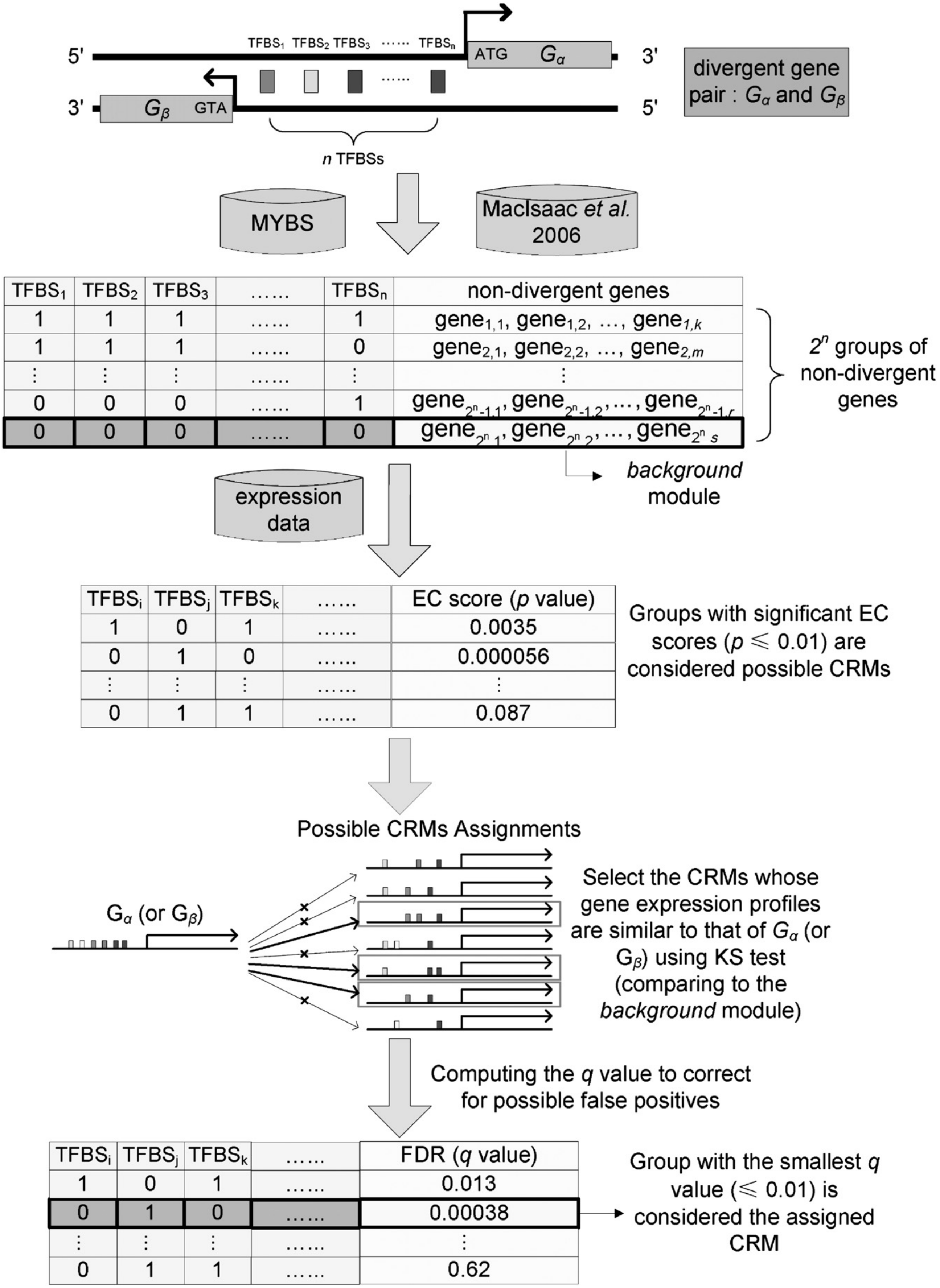
When assigning a CRM to a gene, we selected the most statistically significant CRM. However, as other significant CRMs could also be potential candidates, we might have underestimated the proportion of divergent pairs that share a regulatory system. We examined this possibility using the following stringent criterion: for one gene, all CRMs that satisfied the threshold (q value$\leq$0.01) were deemed potential CRMs. If the respective CRMs of two genes in a divergent pair had at least a single CRM in common, then we labeled the genes as the *same*. Note that the classification of pairs in the *same* and *one-assigned* categories is not affected using this new criterion. After applying the criterion, we found that all pairs in the *overlapped* category were reclassified as the *same* category (but they were already considered as *shared*). In addition, approximately, 47% (31/66) of pairs belonging to the *different* category were reclassified to the *same* category. Overall, the percentage of shared CRMs was still lower than 46% in the ten investigated datasets (Fig. S1), indicating that the proportion of divergent pairs was not underestimated.

To assess whether the results were affected by stochastic noise inherent in genome-scale data analysis, we performed two control experiments. First, since it is known that neighboring genes tend to be co-expressed [15], to avoid a foregone conclusion, we try to preserve the neighbor effects in gene expression and promoter sequence. To this end, we permute the gene expression from the promoters by offsetting all of the gene expression data by one gene, along the chromosome, and retain everything else intact. This preserves the neighbor effects in gene expression and promoter sequence (note that the impact on gene expression of the distance between two adjacent genes were not fully addressed), but it permutes the gene expression of the promoters. The results of this control experiment showed that no gene in the divergent pairs was assigned by regulatory modules in the ten datasets; thus, our preceding results are not merely inherent noise. Second, we randomly scrambled the TFBS annotations of the divergent gene pairs, but left the annotations of non-divergent gene pairs unchanged. The re-assignments were repeated 100 times. In each trial, we found that, on average, less than ten genes (0.52% = 10/1922) in divergent pairs were assigned by regulatory modules, 1.57 (0.082%) divergent gene pairs belonged to the *same* category, and 7.66 (0.4%) pairs were *one-assigned*; however, there were no *overlapped* or *different* pairs. The results of these two control experiments suggest that the CRMs identified by our approach are unlikely to be random expectation.

To determine whether a divergent pair always belonged to the *shared* or *not-shared* categories in all ten datasets, we examined the variations in CRM usage by all divergent pairs (Fig. 3; also see Additional material 2). We grouped the divergent pairs according to their CRM usage into the following three types: (A) divergent pairs that belong to the *shared* category in some datasets and to the *without annotation* category (neither gene in a divergent pair has an assigned CRM) in other datasets, such as the YIL104C_YIL103W pair; (B) divergent pairs that belong to the *shared* category in some datasets and to the *not-shared* category in other datasets, such as the YLR029C_YLR030W pair; and (C) divergent pairs that belong to the *not-shared* category in some datasets, but do not belong to the *shared* category in any dataset. Approximately 49% of the pairs belong to Type C, which indicates that around half of the divergent gene pairs tend to be co-regulated in at least one dataset, but they are not co-regulated in most datasets. In addition, for most divergent pairs in Type C, in all datasets, the CRMs assigned to one gene in a pair were not assigned to the other one gene. Moreover, the CRMs assigned to each gene also varied in different datasets, *e.g.*, the YPR190C_YPR191W pair. This implies that, in a divergent pair, a gene's CRM usage normally varies under different conditions and tends to be different from that of the other gene.

In the ten examined datasets, the proportion of any two genes that were co-expressed (the *Pearson* correlation coefficient >0.6) was determined to be 3–16% by a whole-genome pairwise comparison

Fig. 1. Flowchart for assigning CRMs to genes in divergent gene pairs. First, for each gene, we collected the possible TFBSs within its promoter. Then, we identified genes in divergent pairs and considered the rest as non-divergent. For genes in a divergent pair, we enumerated all possible types of CRMs, assigned non-divergent genes to different CRM types, and checked the degree of co-expression using the expression coherence (EC) score. Then, we selected the groups whose gene expression profiles were similar to the identified genes by comparing with the *background set* (in which genes do not have any of the *related* TFBSs in their promoters). We also used the false discovery rates (FDR) for the correction for multiple testing and obtained the corrected set of regulatory modules. Finally, we determined which CRM was the most likely module by assigning the corresponding CRM with the smallest *q* value.



TFBS$_1$ TFBS$_2$ TFBS$_3$ …… TFBS$_n$
5'
3'
ATG
G_α
3'
5'
G_β
GTA
……
n TFBSs
divergent gene pair : G_α and G_β
MYBS
MacIsaac *et al.* 2006

TFBS$_1$	TFBS$_2$	TFBS$_3$	……	TFBS$_n$	non-divergent genes
1	1	1	……	1	gene$_{1,1}$, gene$_{1,2}$, …, gene$_{1,k}$
1	1	1	……	0	gene$_{2,1}$, gene$_{2,2}$, …, gene$_{2,m}$
⋮	⋮	⋮	……	⋮	⋮
0	0	0	……	1	gene$_{2^n-1,1}$, gene$_{2^n-1,2}$, …, gene$_{2^n-1,r}$
0	0	0	……	0	gene$_{2^n,1}$, gene$_{2^n,2}$, …, gene$_{2^n,s}$

2^n groups of non-divergent genes
background module
expression data

TFBS$_i$	TFBS$_j$	TFBS$_k$	……	EC score (p value)
1	0	1	……	0.0035
0	1	0	……	0.000056
⋮	⋮	⋮	……	⋮
0	1	1	……	0.087

Groups with significant EC scores ($p \leqslant 0.01$) are considered possible CRMs
Possible CRMs Assignments
G_α (or G_β)
×
×
×
×
Select the CRMs whose gene expression profiles are similar to that of G_α (or G_β) using KS test (comparing to the *background* module)
Computing the q value to correct for possible false positives

TFBS$_i$	TFBS$_j$	TFBS$_k$	……	FDR (q value)
1	0	1	……	0.013
0	1	0	……	0.00038
⋮	⋮	⋮	……	⋮
0	1	1	……	0.62

Group with the smallest q value ($\leqslant 0.01$) is considered the assigned CRM

Table 1
Information derived from ten microarray datasets about identified CRMs in divergent gene pairs.

	crz1p	alpha	Damage	Glucose	oxi	HP	Hs	Os	MD	Nitrogen
# of CRMs[a]	25	16	8	30	17	9	32	27	34	11
# of genes with assigned CRMs	287	231	132	297	226	42	266	226	221	28
# of *same*[b] pairs	85	54	20	74	59	3	49	36	33	0
# of *overlapped*[c] pairs	1	5	0	3	3	0	2	1	2	0
# of *different*[d] pairs	7	4	8	14	6	0	11	9	7	0
# of *one-assigned*[e] pairs	101	105	76	115	90	36	142	134	137	28
Proportion of co-expressed[f] genes based on pairwise comparisons (%)	4.5	4.4	3.0	5.2	6.6	9.0	11.0	13.1	9.4	16.1
Co-expressed divergent pairs (%)	10.0	12.9	10.5	8.1	9.1	12.6	17.8	23.6	15.2	17.6
Co-expressed divergent pairs with *shared*[g] CRMs (%)	17.4	35.6	25	28.6	24.2	33.3	44	56.3	42.9	0
Co-expressed divergent pairs with *not-shared*[h] CRMs (%)	2.0	3.3	8.7	1.0	1.2	6.7	5.6	14.7	7.4	50
Co-expressed divergent pairs *without annotated*[i] CRMs (%)	10.3	12.2	10.2	6.6	8.7	12.8	17	22.9	15.2	15.6

[a] # of different CRMs assigned to divergent genes.
[b] *same*: both genes in a divergent pair have the same assigned CRM.
[c] *overlapped*: two genes have overlapping assigned CRMs.
[d] *different*: two genes have different assigned CRMs.
[e] *one-assigned*: only one gene in a pair has an assigned CRM.
[f] Co-expressed: *Pearson* correlation coefficient is greater than 0.6.
[g] *shared*: including *same* and *overlapped.*
[h] *not-shared*: including *different* and *one-assigned.*
[i] *without annotated*: neither gene in a divergent pair has an assigned CRM.

approach. In contrast, for all the divergent pairs, the proportion was 8–24% (for convergent and tandem pairs, the proportions were 8–20% and 8–23% respectively). In the datasets (except the *nitrogen* dataset because no gene pair is shared in the dataset), the proportions of co-expressed divergent pairs in the *shared*, *not-shared*, and *without annotation* categories were 17–56%, 1–15% and 7–23%, respectively (Table 1). Although co-expressed divergent pairs are more likely to fall into the *shared* category under our method, some (1–15%) co-expressed divergent pairs belong to the *not-shared* category. According to our previous study, the proportions of adjacent genes that share similar expression patterns do not differ significantly, regardless of whether they are divergent, convergent, or tandem [16]. In that study, we also found that the divergent relationship was not appreciably favored by selection, and the co-expression of divergent gene pairs could not be attributed simply to a shared regulatory system. Given all the above factors, it is reasonable to infer that only a limited proportion (less than 50%) of divergent gene pairs share a regulatory system under any one investigated condition. However, it should be noted that the ten conditions investigated in our study do not cover the entire transcription program in yeast. Hence, there could be several gene pairs that share a regulatory system under a given condition, but they are not included in our study.

3.3. Identifying condition-dependent CRMs

To ascertain the biological relevance of the identified CRMs with respect to the datasets used in the microarray experiments, we conducted an extensive literature survey. As shown in Table 2, different CRMs were identified in different datasets. For example, 16 CRMs were identified in the *alpha* dataset, in which 11 of the covered 16 TFs are known, or predicted, to be involved in controlling the yeast cell cycle [22,37]. In three cases, where CRMs contain multiple TFBSs (FKH2–MCM1–NDD1 [38], MBP1–SWI4 [39] and MBP1–SWI6 [40]), the synergistic interactions among the TFs during the cell cycle have been documented. Moreover, it is known, or predicted, that some TFs in CRMs form protein complexes or compete with each other for binding, *e.g.* Hap family [41], Met31–Met32–Met4 [42], Swi4–Swi6 [39], Fkh1–Mcm1 [43], Fkh2–Mcm1 [43] and Ino2–Ino4 [44].

Closer inspection of the CRMs activated under different conditions (Table 2) reveals the following pattern: although several CRMs are constitutively active among the ten analyzed datasets, more are transiently activated in terms of one dataset or subset of datasets. For example, two CRMs (2.44%) were active in all datasets, but about 49% of CRMs were only active in one dataset. Therefore, we reasoned that genes whose CRMs are assigned under a particular experimental condition should have a higher probability of participating in functionally related biological processes. We explored this hypothesis by examining the functional enrichment of genes in all divergent pairs with assigned CRMs by using the MIPS Functional Catalogue Database (FunCatDB, http://mips.helmholtz-muenchen.de/proj/funcatDB/) [45], which computes the functional enrichment in terms of hypergeometric p values. Here, we defined a set of genes to be enriched in a particular function if the hypergeometric p value was less than 0.01.

To obtain a more quantitative picture of the relationship between the enrichment of functional categories and genes with assigned CRMs, we counted and compared the total number of enriched functions for genes in divergent pairs with and without assigned CRMs. For example, in the *alpha* dataset, genes in divergent pairs with assigned CRMs were significantly enriched in 10 functional categories (p value<0.01); whereas genes without annotated CRMs were enriched in only one functional category (Fig. 4A). We found that in the ten investigated datasets, genes with assigned CRMs were enriched in a much higher number of functional categories (84 functional categories; p value<0.01) compared to genes without CRMs (only eight functional categories). Moreover, we checked the biological relevance of the enriched functional categories. We found that genes with CRMs assigned under a particular condition were more enriched in the functional categories than genes with CRMs assigned under other conditions, as shown in Fig. 4A. For example, genes in divergent pairs with assigned CRMs in the *alpha* dataset were significantly enriched by genes annotated with "*CELL CYCLE AND DNA PROCESSING*" in the database ($P=8.7\times10^{-6}$ by hypergeometric distribution); such enrichment was not found under other conditions. For genes in divergent pairs without annotated CRMs, no statistically significant enrichment could be found (Fig. 4B).

In addition, to determine whether genes regulated by a CRM are enriched in a specific functional category, we examined the functional enrichment of the target genes by using FunCatDB for each assigned CRM. In the analysis, we only included target genes whose number was larger than or equal to five. We found that, for each assigned CRM, the target genes under a particular condition were significantly enriched in the functional categories (see Additional material 3). To confirm the biological relevance of our findings, we searched the literature for some of the enriched CRMs. We found that RPN4 was enriched in the main category of "*PROTEIN FATE*" and in sub categories "*protein modification*" and "*protein/peptide degradation*" along with

Table 2
Identified CRMs of divergent gene pairs based on the analysis of ten microarray datasets.

CRM[a]	Dataset									
	crz1p	alpha	Damage	Glucose	oxi	HP	Hs	Os	MD	Nitrogen
ABF1							8		13	1
ABF1 RPN4									1	
ABF1 RRPE							1			
AFT1 CIN5			1							
AFT2									1	
BAS1								2	3	
BAS1 GCN4							1	1		
CAD1									2	
CAD1 YAP1 YAP7			1						1	
CBF1									4	
CRZ1	17[b]									
DIG1 MCM1 STE12	2									
ESR1 PAC		2								
FHL1	3		2	4	4	4	3	3	1	2
FHL1 RAP1	27	7	10	13	20	11	15	13	4	9
FHL1 RAP1 SFP1	3	3	6	6	8		7	6	6	1
FKH1								1	6	
FKH1 FKH2	10				1					
FKH1 FKH2 MCM1	1									
FKH1 MCM1			1							
FKH2	1									
FKH2 MCM1	1							1		
FKH2 MCM1 NDD1		1								
FKH2 NDD1	1									
GCN4							4		2	
GCN4 YAP1				14			3	2	7	
GIS1				2						
HAP1						1	4			
HAP2				1			2			
HAP2 HAP3 HAP4 HAP5	1			3				2		
HAP2 HAP4				1						
HAP2 HAP4 HAP5				2						
HAP3				1					1	
HAP3 HAP4 HAP5							1			
HAP3 HAP5				4			1	1		
HAP4				2						
HAP4 HAP5	1			1						
HAP5							1	3		
HCM1		6						2		
HSF1				1	13		6	2	2	
INO2 INO4				1						
MAL63							2			
MBP1		64					3	16	28	1
MBP1 MCM1 SWI6		1								
MBP1 SWI4	5	2			1			2	2	
MBP1 SWI6		1					1	1	1	
MCM1									1	
MCM1 STE12	2									
MCM1 XBP1									1	
MET31						1	1		1	
MET31 MET32 MET4	1									
MET31 MET4				1						
MET4	1				4		1			
MET4 RPN4				1						
MIG1				24			9			
MSN2				6		2	4	3		
MSN2 SKN7				1						
MSN4				1						
NDD1								1		
NDT80					1					
PAC	72	56	51	59	46	9	58	58	39	
PAC RRPE		1								
PHD1										1
RAP1	24			20	11	5	16	13	5	3
RAP1 SFP1	6			4	6		4	4	2	2
REB1 YDR026C							1			
ROX1									1	
RPN4	27			36	44	3	30	7	32	
RPN4 MET4	1									
RRPE	75	41	60	75	59	6	68	66	13	5
SFP1	3			2	3		3	2	3	2
SKN7							2		2	
STB1		10							2	1

(continued on next page)

Table 2 (*continued*)

CRM[a]	Dataset									
	crz1p	alpha	Damage	Glucose	oxi	HP	Hs	Os	MD	Nitrogen
STB1 SWI4	1	6						1	6	
STB1 SWI4 SWI6	1			2						
SWI4		29					4	12	19	
SWI4 SWI6								1		
TEA1				2			1			
UME6				7						
YAP1					2		1		8	
YAP1 YAP7		1			1				1	
YAP7					2					

[a] All CRMs identified in ten datasets.

[b] Each entry represents the numbers of genes assigned to a CRM in the dataset. For example, 17 genes were assigned to CRZ1 in the *crz1p* dataset.

"*structural protein binding*". From studies, profiling the regulatory network of damaged cells in *S. cerevisiae*, DNA excision repair genes and protein degradation genes are known to be modulated by the proteasome-associated protein RPN4 via its binding to the MAG1 upstream repressor [46]. Furthermore, Matsumoto et al. [47] found that RPN4 is involved in the regulation of genes exposed to heat shock conditions (the *Hs* dataset) by comparing their expression levels with *S. cerevisiae* wild-type genes. As another example, we found that MBP1 was enriched in the process of "*CELL CYCLE AND DNA PROCESSING*" in the cell cycle related condition (the *alpha* dataset). Using the crystal structure of the DNA-binding domain of MBP1, Xu et al. [48] demonstrated the importance of this TF in cell cycle control and DNA synthesis. These results consolidate our findings on conditional specificity of CRMs regulating divergent gene pairs and being enriched in biological processes. However, due to the limited data pertaining to CRMs in all experimental conditions in yeast, these results can only be regarded as a guideline, as there could be several other cases not detected by our analysis.

3.4. Identified CRMs tend to occur close to the TSS of their regulated genes

To provide quantitative support for our findings, we also analyzed the physical distances (base pairs) between genes and their assigned CRMs. According to Chen et al. [49], in a divergent gene pair, the TFs whose TFBSs are proximal to a gene tend to regulate it. Therefore, for the *one-assigned* divergent gene pairs, we performed a one-sided KS test to determine whether the distances of the assigned genes to the corresponding CRMs were statistically shorter than those without assigned genes. The distance from a CRM to a gene was calculated as the number of base pairs from the proximal TFBS (belonging to the CRM) to the TSS of the gene. The results showed that the CRMs were indeed closer to their target genes than to the non-target genes in eight datasets (p value$<$0.05). The *alpha*, *damage* and *crz1p* datasets showed highly significant results (p value$<10^{-5}$) (p values were 1.33×10^{-6}, 2.37×10^{-9} and 1.61×10^{-7}, respectively). Based on the evidence from the literature, functional enrichment and distance analysis, we believe that our assigned CRMs play a role in the biological functions elucidated in our work.

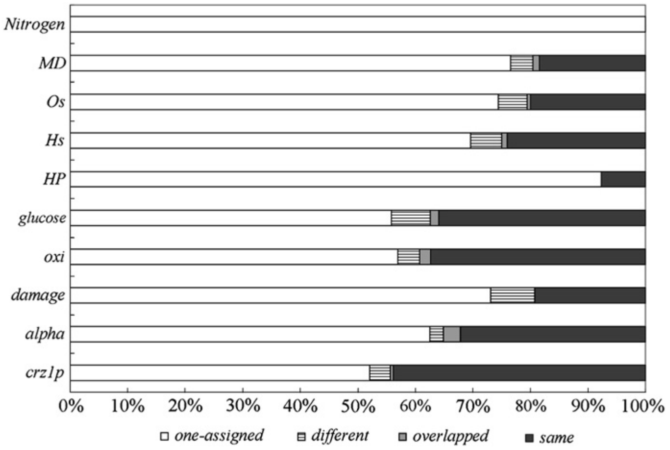


Fig. 2. Distribution of the fraction of *shared*/*not-shared* CRMs of divergent gene pairs in various datasets. Fractions corresponding to various categories, represented in different forms as shown in the inset, are displayed.

3.5. The impact of missing TFBS annotations

The proposed method probably cannot exhaustively identify all CRMs associated with a certain condition due to incomplete TFBS annotations and noisy expression data. Although the TFBS annotations include the binding motif consensus sequences of 144 TFs, some TFBSs are definitely missing. Therefore, we simulated the potential impact of missing TFBSs by removing the occurrences of some TFBSs, and then re-running the whole analysis to estimate the proportion of shared CRMs in divergent gene pairs. We selected five CRMs (FHL1–RAP1, MBP1, PAC, RPN4 and RRPE) that occurred frequently in our results. One at a time, we simulated the changes in the proportion of shared CRMs in divergent gene pairs with and without the TFBS. The impacts of missing CRMs are shown in Table S2. For gene pairs that were assigned to the *shared*, *not-shared* and *without annotation* categories without considering these CRMs, approximately 43%, 89% and 67% of these pairs respectively were *not-shared* in the ten datasets after taking these CRMs into account. These results indicate that when one or more TFBS annotations are available, a substantial portion of gene pairs are likely to be categorized as *not-shared*. This proportion is also consistent with our estimation.

Admittedly, the absence of TFs might have a significant impact on the results and the above simulation might not fully reflect the real situation. Therefore, we also collected four different TF binding datasets to analyze the trend of the proportion of *shared*/*not-shared* regulatory modules of the divergent gene pairs. The four datasets were YEASTRACT [50], MYBS [20], MacIsaac et al. [18] and Badis et al. protein binding microarray (PBM) experiments [51]. We used the datasets to assess the potential impact of incomplete TFBS annotations, since they contained a sufficient number of TFs and were obtained from the literature or experiments. The TFBS annotations of 118 TFs from MacIsaac et al. were gathered as described in Materials and methods. We also collected 104 TFs and the TFBS annotations from MYBS alone. For YEASTRACT, the regulatory associations between TFs and their target genes in *S. cerevisiae* were collected from the literature. After downloading from YEASTRACT, we collected 108 TFs and their target genes. Badis et al. performed *in vitro* PBM experiments to generate the motifs for 112 yeast TFs. Then, they used PWMs and the GOMER program [52] to estimate the probability of a TF binding to somewhere within a promoter. We followed their

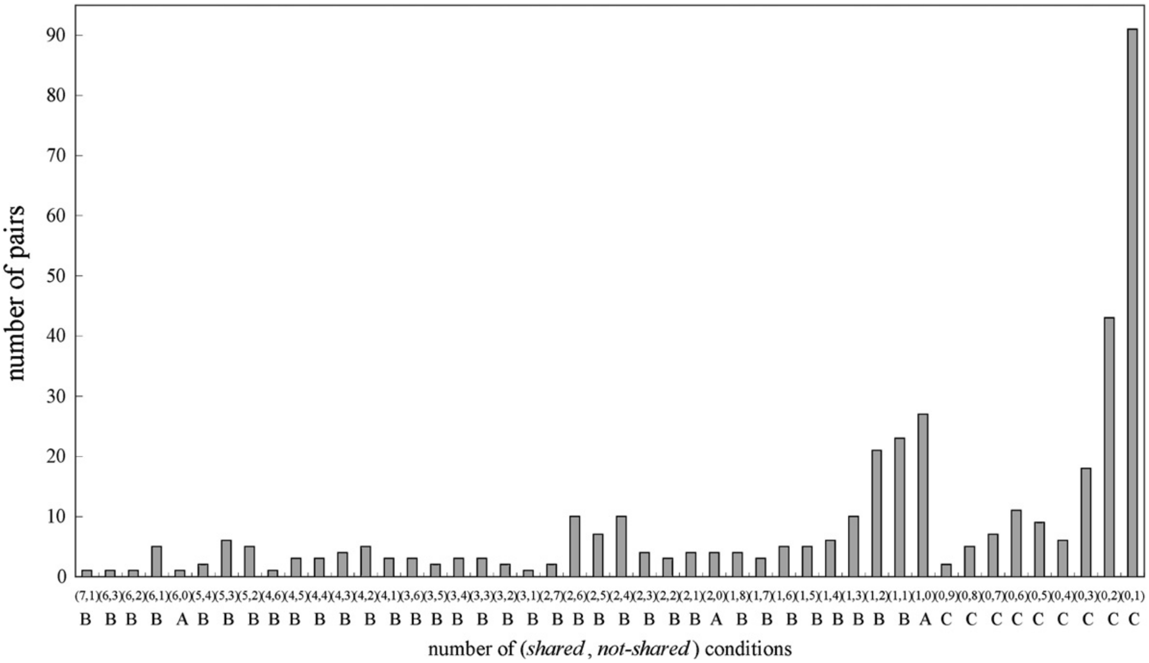


Fig. 3. CRM usage by divergent gene pairs. The figure shows CRM usage of 961 divergent pairs in the ten analyzed datasets. The x-axis represents the number of (*shared*, *not-shared*) datasets. For example, (1,4) represents the group in which the genes in each pair share the same assigned CRM in only one dataset and have different CRMs in the other four datasets. The y-axis represents the number of divergent gene pairs. In addition, the three types of CRM usage are: (A) divergent pairs that belong to the *shared* category in some datasets and to the *without annotation* category in other datasets; (B) divergent pairs that belong to the *shared* category in some datasets and to the *not-shared* category in other datasets; and (C) divergent pairs that belong to the *not-shared* category in some datasets, but do not belong to the *shared* category in any dataset.

criteria and chose the top 200 hits from GOMER as the target genes for each TF. The above four TF binding datasets contained different numbers of TFs. Table S3 provides an overview of the TFs in the four binding datasets. In total, 194 TFs were collected in these four datasets; 36 TFs existed in all four datasets, while 70 TFs only appeared in one dataset. Applying our method to the four datasets, we found that most of the divergent pairs did not share CRMs (Fig. 5). Specifically, the proportions of regulatory modules assigned to the *same*, *different*, *overlapped* and *one-assigned* categories were similar. Therefore, we believe that the true ratio of divergent gene pairs that share CRMs should be similar to our estimation.

3.6. The existence of TFBSs in the intergenic regions of divergent gene pairs

Since different promoters have different numbers of TFBSs and the probability of sharing a CRM might be dependent on the number of binding sites, we examined the relationship between the number of TFBSs and the length of the promoters. As shown in Fig. S1, overall, longer promoters seem to contain more TFBSs. We also checked whether the probability of sharing a CRM is dependent on the number of binding sites. Specifically, we counted the average lengths of promoters and the average numbers of TFBSs of *shared* and *not-shared*

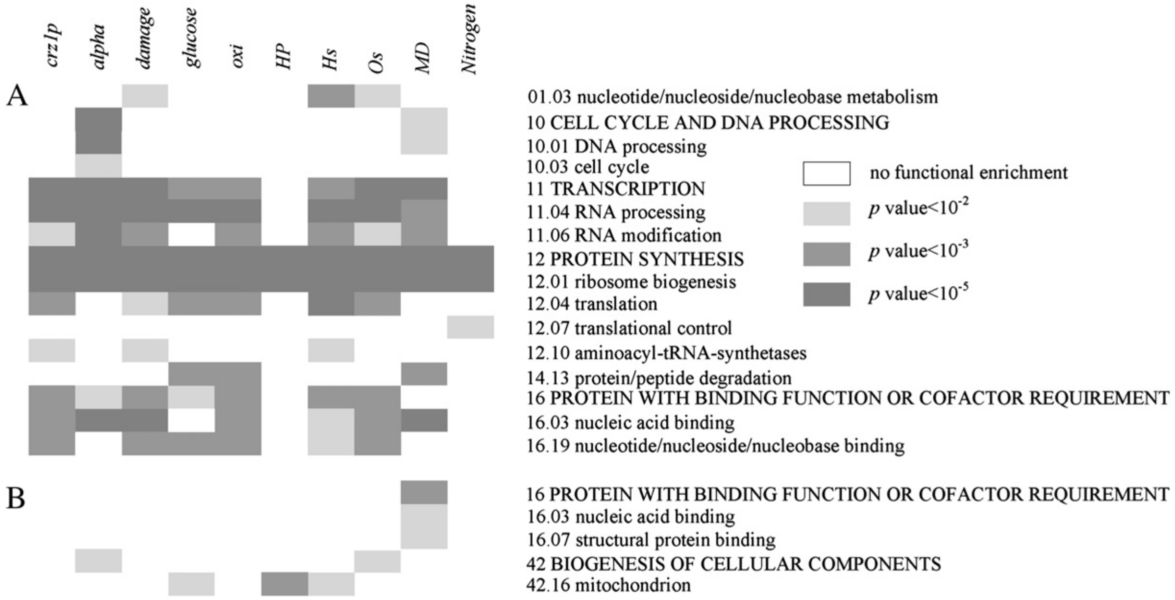


Fig. 4. Functional enrichment of genes in divergent pairs under different expression conditions. (A) With assigned CRMs, and (B) without annotated CRMs.

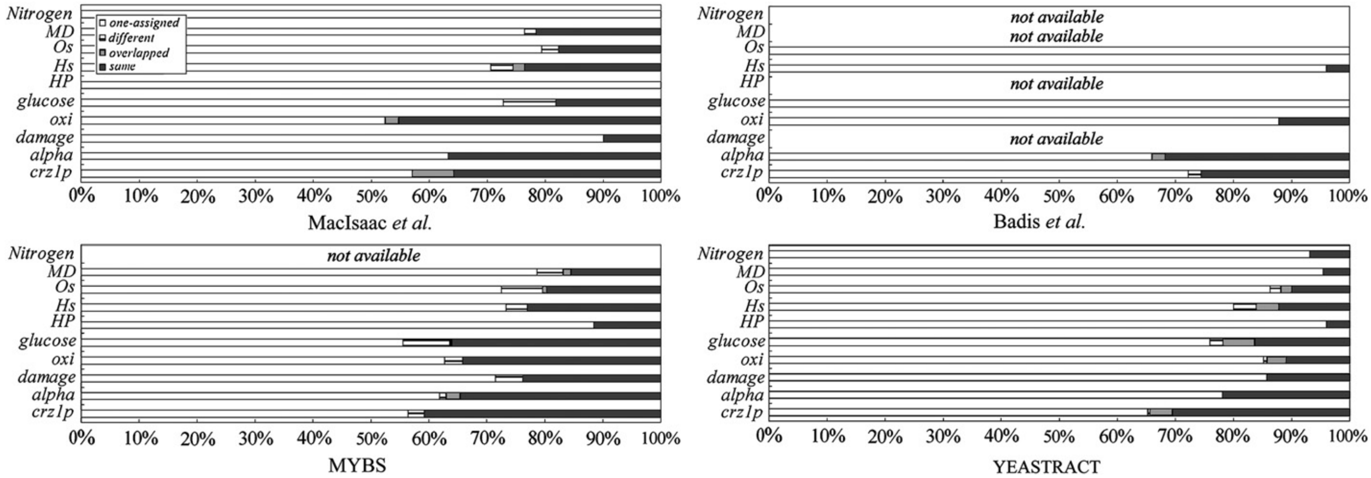


Fig. 5. Distribution of the fraction of *shared*/*not-shared* CRMs of divergent gene pairs in various datasets using four different TF binding datasets. Fractions corresponding to various categories, represented in different forms as shown in the inset, are displayed.

groups in the ten analyzed datasets (Table S4). Although the divergently transcribed pairs in "*shared*" groups tend to have shorter promoters and fewer TFBSs than those in "*not-shared*" groups, the difference (based on the two-sided KS test) is not significant. Only the promoter length in the *glucose* dataset and the number of TFBSs in the *Hs* dataset have *p* values smaller than 0.01.

In terms of the distribution of the number of TFBSs per intergenic region (normalized to the promoter length), we found that the average number of TFBSs in the intergenic regions of divergent gene pairs and uni-directional transcribed promoters was 6.26 (TFBSs/Kbp) and 5.2 (TFBSs/Kbp) respectively. Note that the upper bound on the lengths of uni-directional transcribed promoters was limited to 1000 bp. The average number of TFBSs in a divergent gene pair was larger than that of uni-directional promoters, which was consistent with the observation of Erb and Nimwegen [53]. We also examined whether some TFBSs prefer to locate in the intergenic regions of divergent gene pairs. For each TF, we used a one-sided two sample proportion test to check whether the proportion of divergent pairs with binding sites in the promoters was significantly greater than in non-divergent genes. Among 144 TFs, the following 6 TFs were found to be significantly over-represented in divergent pairs (p value$<10^{-5}$): Abf1, PAC, Rap1, Rpn4, RRPE, and Xbp1. Most of those TFs were also found in our identified CRMs (Table 2). The descriptions provided by the Saccharomyces Genome Database (SGD) [22] also show that some of the above TFs are functionally related. For example, Abf1p and Rap1p are responsible for regulating the chromatin structure [54]. Interestingly, Lin et al. [9] also found that five motifs are over-represented in the intergenic regions of divergent gene pairs in humans. However, the consensus and functions of these five motifs are dissimilar to the six motifs we identified in yeast. Why some TFBSs prefer to reside in the intergenic regions of divergent gene pairs remains an open question.

4. Conclusion

In this paper, we have investigated whether divergent gene pairs share *cis*-regulatory modules. This is a challenging problem because, when two divergent genes share the same intergenic region, it is difficult to assign the appropriate CRM to each gene without ambiguity. Taking advantage of the analyses of the TF knockout experiments, we have shown that, in most instances, TF knockout only alters the expression of one gene in a divergent pair. In an attempt to resolve this problem, we have proposed a novel method for estimating the ratio of divergent gene pairs that share CRMs. The reliability of this assessment was enhanced by an extensive literature survey and functional enrichment analysis. We found that only a limited number of divergent gene pairs appear to share CRMs in one condition, although approximately half of the divergent pairs shared a regulatory system in at least one dataset. Whether this characteristic is common to other systems has yet to be determined. Thus, it would be of great interest to extend this analysis to other systems by using a similar approach.

Supplementary materials related to this article can be found online at doi:10.1016/j.ygeno.2010.08.008.

Funding

This work was supported by the Institute of Information Science, Academia Sinica and the National Science Council of Taiwan (Grant number NSC98-2221-E-001-015) to H.-K.T. In addition, T.-H.C. was supported by the Genomics Research Center, Academia Sinica and the National Science Council (Grants 97-2311-B-001-014-MY3; 98-2311-B-001-015-MY3).

Acknowledgments

We wish to thank the two anonymous reviewers, as well as Krishna B. S. Swamy, Jen-Hao Cheng, Grace Tzu-Wei Huang, Shuwei Li, and Daryi Wang for their valuable suggestions and comments.

References

[1] M. Johnston, R.W. Davis, Sequences that regulate the divergent GAL1–GAL10 promoter in *Saccharomyces cerevisiae*, Mol. Cell. Biol. 4 (1984) 1440–1448.
[2] N.D. Trinklein, S.F. Aldred, S.J. Hartman, D.I. Schroeder, R.P. Otillar, R.M. Myers, An abundance of bidirectional promoters in the human genome, Genome Res. 14 (2004) 62–66.
[3] N. Chen, L.D. Stein, Conservation and functional significance of gene topology in the genome of *Caenorhabditis elegans*, Genome Res. 16 (2006) 606–617.
[4] P.G. Engstrom, H. Suzuki, N. Ninomiya, A. Akalin, L. Sessa, G. Lavorgna, A. Brozzi, L. Luzi, S.L. Tan, L. Yang, G. Kunarso, E.L. Ng, S. Batalov, C. Wahlestedt, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, C. Wells, V.B. Bajic, V. Orlando, J.F. Reid, B. Lenhard, L. Lipovich, Complex loci in human and mouse genomes, PLoS Genet. 2 (2006) e47.
[5] Y.Y. Li, H. Yu, Z.M. Guo, T.Q. Guo, K. Tu, Y.X. Li, Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance, PLoS Comput. Biol. 2 (2006) e74.
[6] S. Kruglyak, H. Tang, Regulation of adjacent yeast genes, Trends Genet. 16 (2000) 109–111.
[7] B. Thomas, Gene clusters and polycistronic transcription in eukaryotes, BioEssays 20 (1998) 480–487.
[8] M.T. Travers, M. Cambot, H.T. Kennedy, G.M. Lenoir, M.C. Barber, V. Joulin, Asymmetric expression of transcripts derived from the shared promoter between the divergently oriented ACACA and TADA2L genes, Genomics 85 (2005) 71–84.
[9] J.M. Lin, P.J. Collins, N.D. Trinklein, Y. Fu, H. Xi, R.M. Myers, Z. Weng, Transcription factor binding and modified histones in human bidirectional promoters, Genome Res. 17 (2007) 818–827.

[10] E. Zanotto, Z.H. Shah, H.T. Jacobs, The bidirectional promoter of two genes for the mitochondrial translational apparatus in mouse is regulated by an array of CCAAT boxes interacting with the transcription factor NF-Y, Nucleic Acids Res. 35 (2007) 664–677.

[11] S.J. Cooper, N.D. Trinklein, E.D. Anton, L. Nguyen, R.M. Myers, Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome, Genome Res. 16 (2006) 1–10.

[12] P.J. Collins, Y. Kobayashi, L. Nguyen, N.D. Trinklein, R.M. Myers, The ets-related transcription factor GABP directs bidirectional transcription, PLoS Genet. 3 (2007) e208.

[13] P.R. Kensche, M. Oti, B.E. Dutilh, M.A. Huynen, Conservation of divergent transcription in fungi, Trends Genet. 24 (2008) 207–211.

[14] L.D. Hurst, E.J. Williams, C. Pal, Natural selection promotes the conservation of linkage of co-expressed genes, Trends Genet. 18 (2002) 604–606.

[15] B.A. Cohen, R.D. Mitra, J.D. Hughes, G.M. Church, A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression, Nat. Genet. 26 (2000) 183–186.

[16] H.K. Tsai, C.P. Su, M.Y. Lu, C.H. Shih, D. Wang, Co-expression of adjacent genes in yeast cannot be simply attributed to shared regulatory system, BMC Genomics 8 (2007) 352.

[17] I. Yanai, C.P. Hunter, Comparison of diverse developmental transcriptomes reveals that coexpression of gene neighbors is not evolutionarily conserved, Genome Res. 19 (2009) 2214–2220.

[18] K.D. MacIsaac, T. Wang, D.B. Gordon, D.K. Gifford, G.D. Stormo, E. Fraenkel, An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*, BMC Bioinform. 7 (2006) 113.

[19] 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.

[20] H.K. Tsai, M.Y. Chou, C.H. Shih, G.T. Huang, T.H. Chang, W.H. Li, MYBS: a comprehensive web server for mining transcription factor binding sites in yeast, Nucleic Acids Res. 35 (2007) W221–W226.

[21] K.C. Dobi, F. Winston, Analysis of transcriptional activation at a distance in *Saccharomyces cerevisiae*, Mol. Cell. Biol. 27 (2007) 5575–5586.

[22] E.L. Hong, R. Balakrishnan, Q. Dong, K.R. Christie, J. Park, G. Binkley, M.C. Costanzo, S.S. Dwight, S.R. Engel, D.G. Fisk, J.E. Hirschman, B.C. Hitz, C.J. Krieger, M.S. Livstone, S.R. Miyasato, R.S. Nash, R. Oughtred, M.S. Skrzypek, S. Weng, E.D. Wong, K.K. Zhu, K. Dolinski, D. Botstein, J.M. Cherry, Gene ontology annotations at SGD: new data sources and annotation methods, Nucleic Acids Res. 36 (2008) D577–D581.

[23] Y. Pilpel, P. Sudarsanam, G.M. Church, Identifying regulatory networks by combinatorial analysis of promoter elements, Nat. Genet. 29 (2001) 153–159.

[24] M. Lapidot, Y. Pilpel, Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription, Nucleic Acids Res. 31 (2003) 3824–3828.

[25] J.D. Storey, R. Tibshirani, Statistical significance for genomewide studies, Proc. Natl. Acad. Sci. USA 100 (2003) 9440–9445.

[26] J. Demeter, C. Beauheim, J. Gollub, T. Hernandez-Boussard, H. Jin, D. Maier, J.C. Matese, M. Nitzberg, F. Wymore, Z.K. Zachariah, P.O. Brown, G. Sherlock, C.A. Ball, The Stanford Microarray Database: implementation of new analysis tools and open source release of software, Nucleic Acids Res. 35 (2007) D766–D770.

[27] H. Yoshimoto, K. Saltsman, A.P. Gasch, H.X. Li, N. Ogawa, D. Botstein, P.O. Brown, M.S. Cyert, Genome-wide analysis of gene expression regulated by the calcineurin/Crz1p signaling pathway in *Saccharomyces cerevisiae*, J. Biol. Chem. 277 (2002) 31079–31088.

[28] 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.

[29] A.P. Gasch, M. Huang, S. Metzner, D. Botstein, S.J. Elledge, P.O. Brown, Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p, Mol. Biol. Cell 12 (2001) 2987–3003.

[30] M. Ronen, D. Botstein, Transcriptional response of steady-state yeast cultures to transient perturbations in carbon source, Proc. Natl. Acad. Sci. 103 (2006) 389–394.

[31] O. Carmel-Harel, R. Stearman, A.P. Gasch, D. Botstein, P.O. Brown, G. Storz, Role of thioredoxin reductase in the Yap1p-dependent response to oxidative stress in *Saccharomyces cerevisiae*, Mol. Microbiol. 39 (2001) 595–605.

[32] M. Shapira, E. Segal, D. Botstein, Disruption of yeast forkhead-associated cell cycle transcription by oxidative stress, Mol. Biol. Cell 15 (2004) 5659–5669.

[33] A.P. Gasch, P.T. Spellman, C.M. Kao, O. Carmel-Harel, M.B. Eisen, G. Storz, D. Botstein, P.O. Brown, Genomic expression programs in the response of yeast cells to environmental changes, Mol. Biol. Cell 11 (2000) 4241–4257.

[34] J. Quackenbush, Microarray data normalization and transformation, Nat. Genet. 32 (Suppl) (2002) 496–501.

[35] B.M. Bolstad, R.A. Irizarry, M. Astrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics 19 (2003) 185–193.

[36] Z. Hu, P.J. Killion, V.R. Iyer, Genetic reconstruction of a functional transcriptional regulatory network, Nat. Genet. 39 (2007) 683–687.

[37] H.K. Tsai, H.H. Lu, W.H. Li, Statistical methods for identifying yeast cell cycle transcription factors, Proc. Natl. Acad. Sci. USA 102 (2005) 13532–13537.

[38] 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.

[39] C. Koch, T. Moll, M. Neuberg, H. Ahorn, K. Nasmyth, A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase, Science 261 (1993) 1551–1557.

[40] V.R. Iyer, C.E. Horak, C.S. Scafe, D. Botstein, M. Snyder, P.O. Brown, Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF, Nature 409 (2001) 533–538.

[41] H.W. Mewes, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, D. Frishman, MIPS: a database for genomes and protein sequences, Nucleic Acids Res. 27 (1999) 44–48.

[42] D.Y. Chiang, D.A. Nix, R.K. Shultzaberger, A.P. Gasch, M.B. Eisen, Flexible promoter architecture requirements for coactivator recruitment, BMC Mol. Biol. 7 (2006) 16.

[43] P.C. Hollenhorst, G. Pietz, C.A. Fox, Mechanisms controlling differential promoter-occupancy by the yeast forkhead proteins Fkh1p and Fkh2p: implications for regulating the cell cycle and differentiation, Genes Dev. 15 (2001) 2445–2456.

[44] J. Hoppen, A. Repenning, A. Albrecht, S. Geburtig, H.J. Schuller, Comparative analysis of promoter regions containing binding sites of the heterodimeric transcription factor Ino2/Ino4 involved in yeast phospholipid biosynthesis, Yeast 22 (2005) 601–613.

[45] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhaupt, M. Munsterkotter, H.W. Mewes, The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, Nucleic Acids Res. 32 (2004) 5539–5545.

[46] S.A. Jelinsky, P. Estep, G.M. Church, L.D. Samson, Regulatory networks revealed by transcriptional profiling of damaged *Saccharomyces cerevisiae* cells: Rpn4 links base excision repair with proteasomes, Mol. Cell. Biol. 20 (2000) 8157–8167.

[47] R. Matsumoto, K. Akama, R. Rakwal, H. Iwahashi, The stress response against denatured proteins in the deletion of cytosolic chaperones SSA1/2 is different from heat-shock response in *Saccharomyces cerevisiae*, BMC Genomics 6 (2005) 141.

[48] R.M. Xu, C. Koch, Y. Liu, J.R. Horton, D. Knapp, K. Nasmyth, X. Cheng, Crystal structure of the DNA-binding domain of Mbp1, a transcription factor important in cell-cycle control of DNA synthesis, Structure 5 (1997) 349–358.

[49] L. Chen, L. Cai, G. Skogerbo, Y. Zhao, R. Chen, Assessing TF regulatory relationships of divergently transcribed genes, Genomics 92 (2008) 316–321.

[50] 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*, Nucleic Acids Res. 34 (2006) D446–D451.

[51] G. Badis, E.T. Chan, H. van Bakel, L. Pena-Castillo, D. Tillo, K. Tsui, C.D. Carlson, A.J. Gossett, M.J. Hasinoff, C.L. Warren, M. Gebbia, S. Talukder, A. Yang, S. Mnaimneh, D. Terterov, D. Coburn, A. Li Yeo, Z.X. Yeo, N.D. Clarke, J.D. Lieb, A.Z. Ansari, C. Nislow, T.R. Hughes, A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters, Mol. Cell 32 (2008) 878–887.

[52] J.A. Granek, N.D. Clarke, Explicit equilibrium modeling of transcription-factor binding and gene regulation, Genome Biol. 6 (2005) R87.

[53] I. Erb, E. van Nimwegen, Statistical features of yeast's transcriptional regulatory code, IEEE Proceedings of the first International Conference on Computational Systems Biology (ICCSB), 2006, pp. 111–118.

[54] A. Yarragudi, T. Miyake, R. Li, R.H. Morse, Comparison of ABF1 and RAP1 in chromatin opening and transactivator potentiation in the budding yeast *Saccharomyces cerevisiae*, Mol. Cell. Biol. 24 (2004) 9152–9164.