Predominant Gain of Promoter TATA Box after Gene Duplication Associated with Stress Responses

Yangyun Zou,[1,2] Wei Huang,[1,2] Zhenglong Gu,[3] and Xun Gu*,[1,2,4]

[1]MOE Key Laboratory of Contemporary Anthropology and Center for Evolutionary Biology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China

[2]Center for Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai, China

[3]Division of Nutritional Sciences, Cornell University

[4]Department of Genetics, Development, and Cell Biology, Iowa State University

***Corresponding author:** E-mail: xgu@iastate.edu.

Associate editor: Naruya Saitou

Research article

Abstract

TATA box, the core promoter element, exists in a broad range of eukaryotes, and the expression of TATA-containing genes usually responds to various environmental stresses. Hence, the evolution of TATA-box in duplicate genes may provide some clues for the interrelationship among environmental stress, expression differentiation, and duplicate gene preservation. In the present study, we observed that the TATA box is significantly overrepresented in duplicate genes compared with singletons in human, worm, Arabidopsis, and yeast genomes. We then conducted an extensive functional genomic analysis to investigate the evolution of TATA box along over 700 yeast gene family phylogenies. After reconstructing the ancestral TATA-box states (presence or absence), we found that significantly higher numbers of TATA box gain events than loss events had occurred after yeast gene duplications—the overall gain–loss ratio is about 3–4 to 1. Interestingly, these TATA-gain duplicate genes on average have experienced greater expression divergence from the ancestral expression states than their most closely related TATA-less duplicate partners, but only under environmental stress conditions (asymmetric evolution); indeed, under normal physiological conditions, they have similar expression divergence (symmetric evolution). Moreover, we showed that TATA-gain duplicates are enriched in stress-associated functional categories but that is not the case for TATA-ancestral duplicates (those inherited from their ancestors prior to duplication). Together, we conclude that after the gene duplication, gain of the TATA box in duplicate promoters may have played an important role in yeast duplicate preservation by accelerating expression divergence that may facilitate the adaptive evolution of the organism in response to environmental changes.

Key words: gene duplication, TATA box, environmental stress, expression divergence, gene family, maximum parsimony.

Introduction

TATA box is the core promoter element for TATA-binding protein (TBP) binding, and exists in a broad range of eukaryotes (Orphanides et al. 1996). For instance, in *Saccharomyces* yeasts, approximately 20% of genes bear the TATA box (Basehoar et al. 2004). Genes with the TATA box, called TATA-containing genes (TATA(+)), and genes without any TATA box, called TATA-less genes (TATA(−)), may represent two distinct pathways of transcription regulation: TATA-containing genes depend more strongly on the SAGA complex (Spt-Ada-Gcn5-Acetyltransferase)-mediated TBP binding, whereas TATA-less genes more on TFIID complex (transcription Factor II D)-dominated TBP binding (Basehoar et al. 2004; Huisinga and Pugh 2004). Substantial evidence suggested that TATA-containing genes are usually associated with environmental stress responses and variably expressed, while most TATA-less genes tend to be housekeeping genes and are constitutively expressed (Basehoar et al. 2004; Tirosh et al. 2006; Landry et al. 2007).

Recently, Kim et al. (2010) observed that yeast duplicate genes were enriched with TATA-containing genes, raising an interesting issue about the relationship among TATA box, environmental stress, and duplicate gene preservation. Many hypotheses have been proposed to explain why so many duplicate genes have been retained in a genome, under the process of subfunctionalization (Force et al. 1999; Lynch and Force 2000; Qian et al. 2010) or neofunctionalization (Ohno 1970), as well as the gene dosage effect (Kondrashov and Koonin 2004; Kondrashov and Kondrashov 2006; Innan and Kondrashov 2010). Since expression divergence after the duplication was considered as the first step toward duplicate preservation in these models, substantial analyses have been carried out on the pattern of expression divergence between duplicates (e.g., Wagner 2000; Gu et al. 2002, 2004, 2005; Makova and Li 2003; Huminiecki and Wolfe 2004; Li et al. 2005, 2009; Chain et al. 2008; Ha et al. 2009), the *cis*-regulatory motif divergence (Papp et al. 2003; Zhang et al. 2004), the *trans*-regulatory divergence (Gu et al. 2005; Leach et al. 2007; Zou et al. 2009b), and the epigenetic divergence (Rodin et al. 2005; Li et al. 2008; Zheng 2008). However, except for a few studies (Ha et al. 2007; Hanada et al. 2008; Zou et al. 2009a), the effects of environmental stresses have not been sufficiently taken into account in this subject.

Because these TATA-containing genes are usually associated with various stress responses, understanding the evolution of TATA-box in duplicate genes would provide new insights into the mechanism how the expression divergence between duplicates could be affected by environmental factors. To explain the enrichment of the TATA box in duplicates, Kim et al. (2010) suggested that the TATA box may contribute to gene duplicability, that is, TATA-containing genes are more likely to be duplicated. However, as we shall discuss in this study, duplicate genes may actually gain new TATA boxes after the gene duplication that have facilitated the stress-response dependent expression differentiation.

In our current study, we first examine whether the enrichment of the TATA box in duplicate promoters is a general pattern across different organisms. To this end, we conduct the genome-wide analysis in the human, Arabidopsis, worm, and yeast. Second, we test the TATA-box gain hypothesis, taking advantage of the well-developed molecular phylogenic analysis and substantial yeast functional genomic data. Briefly speaking, we infer the ancestral TATA box state (presence or absence) for over 700 yeast gene family phylogenies to test whether the gain of the TATA box after the gene duplication is statistically more frequent than the loss. We further investigate the associations among the gain of the TATA box in duplicates, expression divergence, and the organismal responses to various environmental stresses. Our goal is to show how the *cis*-regulatory motif TATA box impacts the yeast duplicate gene preservation through stress-dependent expression divergence and environment adaptation.

Materials and Methods

TATA-Containing and TATA-Less Gene Classification

TATA box in human, worm, and Arabidopsis was predicted according to the description of previous study (Tirosh et al. 2006). The classification of TATA-containing and TATA-less genes in *Saccharomyces cerevisiae* was obtained from Basehoar et al. (2004). Only protein-coding genes were used for the analysis.

Identification of Duplicate and Singleton Genes

We downloaded amino acid sequences of all genes in *Homo sapiens*, *Caenorhabditis elegans*, and *S. cerevisiae* from Ensembl database (http://www.ensembl.org/info/data/ftp/index.html). Peptide sequences of *Arabidopsis thaliana* were obtained from Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org). A genome-wide all-against-all BlastP search was performed to determine duplicate and singleton genes in human, worm, Arabidopsis, and yeast. If a gene had more than one isoform, the longest peptide was used to BlastP searches. Singleton was considered to be a protein that did not hit any other proteins in BlastP search with *E*-value = 0.1. A gene was defined as a duplicate gene if nonself hit had an *E*-value less than or equal to 1.0×10^{-20}. Duplicate gene and its best nonself BlastP hit with *E*-value cutoff of 1.0×10^{-20} were identified as duplicate pairs. Other criteria with *E*-value thresholds of $1.0 \times 10^{-5}$and 1.0×10^{-10} for duplicate gene determination were also applied to robust the analyses and results. Whole genome duplication (WGD) derived duplicate genes of *H. sapiens*, *A. thaliana*, and *S. cerevisiae* were obtained according to the definition of previous studies (Blanc et al. 2003; Byrne and Wolfe 2005; Makino and McLysaght 2010). Duplicate genes except for WGD genes were classified as duplicates generating from small scale duplication (SSD).

The rate of synonymous substitutions (K_S) between yeast duplicate pairs was estimated using PAML (Yang and Nielsen 2000) with default parameters.

Inference for Ancestral TATA Box State of Yeast Gene Families

The protein sequences of hemiascomycetous yeasts *S. cerevisiae* and other four related yeast species, *Candida glabrata*, *Kluyveromyces lactis*, *Debaryomyces hansenii*, and *Yarrowia lipolytica* were downloaded from Génolevures database (http://www.genolevures.org/; Dujon et al. 2004), where protein families are classified based on all-to-all sequence comparisons and algorithmic consensus clustering (Nikolski and Sherman 2007) and literature curation. We then obtained > 800 protein families with the number of *S. cerevisiae* members larger than 1. Multiple protein sequence alignment for each gene family was generated by ClustalX (Thompson et al. 1997). T-Coffee program for multiple sequence alignment was also used to validate the analyses and results (Notredame et al. 2000). The phylogeny of each protein family was inferred from aligned amino acid sequences using the neighbor-joining (NJ) method (Saitou and Nei 1987) with Poisson correction distance in the software of Molecular Evolutionary Genetics Analysis 4.0 (MEGA4) (Tamura et al. 2007). Bootstrap test with 1,000 resamplings was used to evaluate the reliability of tree topologies (Felsenstein 1985). We applied relatively distant organism *Drosophila melanogaster* as outgroup to determine the root of the phylogeny. Specifically, we searched orthologous genes from *D. melanogaster* for each *S. cerevisiae* paralog of a family by BlastP method on Ensembl database (http://www.ensembl.org/). The best hit of each search with *E*-value less than 10^{-20} was selected as the homolog of that gene, then, the protein sequences were retrieved from Ensembl to reconstruct and root the phylogenetic trees. If *D. melanogaster* as outgroup fails to determine the root of tree, midpoint of the longest span between two taxa was jointly used to root the tree.

We then mapped the TATA box state of each gene into the given phylogeny. Except for well-classified TATA-containing and TATA-less genes in *S. cerevisiae*, we also predicted the TATA box status in other four yeast species using TATA consensus sequence TATA(A/T)A(A/T)(A/G) (supplementary material, Supplementary Material online), and observed that the TATA box is expectedly conserved in these yeast species, but the position in promoters is variable (supplementary figs. S1 and S2, Supplementary Material online), which is consistent with the previous study (Tirosh et al. 2007). So we simplified the inferred phylogeny, just using the topology represented by the single *S. cerevisiae* species, then mapped the TATA box data from Basehoar et al. (2004) into the reduced phylogeny. After removing protein families with TATA box information not available, we obtained 767 families for further analysis, where 542 for only two paralogs of *S. cerevisiae*, and 225 for at least three paralogs.

Maximum parsimony (MP) was applied to infer the TATA box state (present or absent) of ancestral *S. cerevisiae* genes and estimate the minimum possible number of TATA box switches (gains or losses) for each gene family (Fitch 1971). In some cases, the parsimony inference may generate more than one alternative scenarios of evolutionary change that are equally parsimonious (Farris 1970), which may imply different evolutionary trajectory of the TATA box. Therefore, we used parsimony optimization algorithms called the accelerated transformation (ACCTRAN) and the delayed transformation (DELTRAN) to resolve the ambiguity (Farris 1970; Swofford and Maddison 1987). The illustration example was showed in supplementary figure S3 (Supplementary Material online).

Yeast Expression Data and Definition of Expression Divergence between Duplicate Genes

Expression profile data of *S. cerevisiae* induced by environmental stresses were downloaded from Gasch et al. (2000) (http://www-genome.stanford.edu/yeast_stress). Expression pattern data of developmental processes were collectively obtained from Spellman et al. (1998) (cell cycle related) and Chu et al. (1998) (sporulation related). The expression values were normalized according to the recommendation of the original studies.

We used evolutionary distance (*E*) defined by Gu et al. (2005) to measure expression divergence between two duplicate copies. Specifically, for any duplicate gene 1 and 2, let x_{1k} and x_{2k} be its expression level, respectively, in the *k*th microarray experiment, $\bar{x}_1$ and $\bar{x}_2$ be the mean of expression level in *k*th microarray experiments, respectively, where $k = 1, \ldots m$. The formula of expression distance (*E*) between gene 1 and 2 is as follows:

$$\hat{E}_{12} = \sum_{k=1}^{m}[(x_{1k} - \bar{x}_1) - (x_{2k} - \bar{x}_2)]^2/(m - 1)$$

Functional Analysis of Yeast Duplicates

Biological process information based on Gene Ontology Slim (GO Slim) was downloaded from *Saccharomyces* Genome Database *(http://downloads.yeastgenome.org/literature_curation/go_slim_mapping.tab)*. GO Slim terms were used to classify 45 biological process categories. Yeast duplicate genes were then assigned to GO Slim classification according to their annotation. One duplicate may contain more than one GO Slim category. In order to get the confidence level (*P*-values) of functional enrichment of these genes, we also conducted web-based gene enrichment analysis using Fisher's exact test with DAVID bioinformatics resources (Dennis et al. 2003; Huang et al. 2009), which contain multiple other categories of functional features, including annotations like pathway, protein domain, and protein interaction.

Results

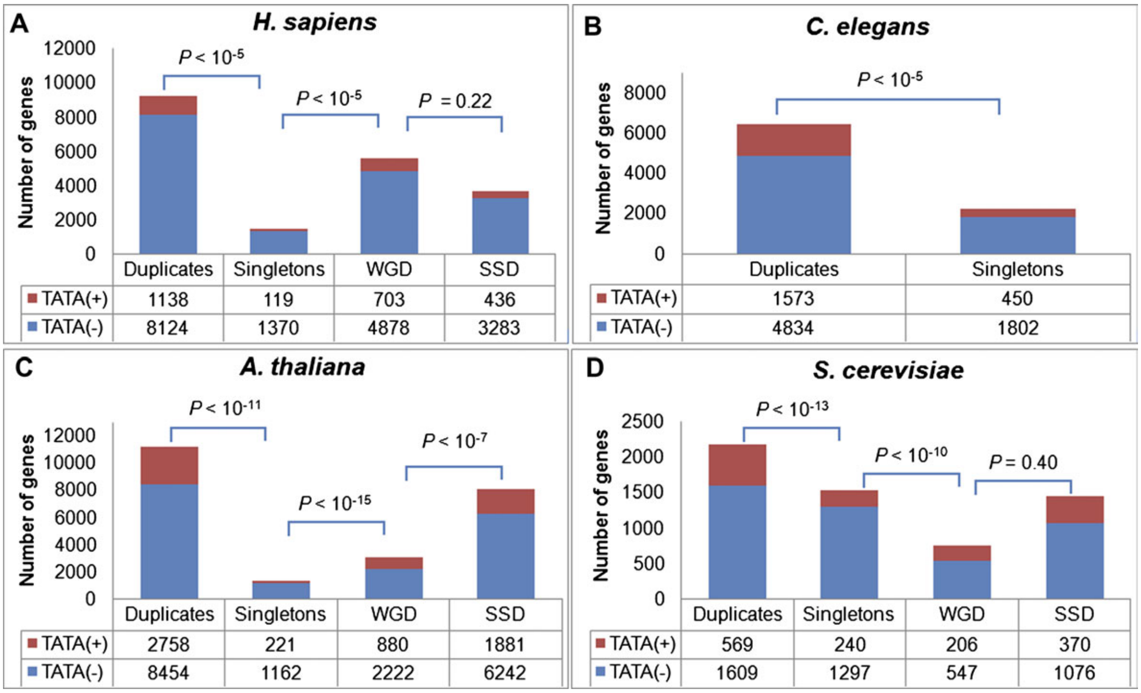
Overrepresentation of the TATA Box in Duplicate Gene Promoters

We surveyed four genomes, human, worm, Arabidopsis, and yeast, to conduct a comprehensive analysis for comparing the proportion of TATA-containing genes. In each genome, TATA-containing genes and TATA-less genes were identified according to the previous studies (Basehoar et al. 2004; Tirosh et al. 2006). Overall, the percentage of TATA-containing genes in worm, Arabidopsis, and yeast remains roughly the same, which is 22.5%, 23.7% and 19.2%, respectively. Impressively, we found that this percentage in the human genome is considerably low (11.2%); the difference from the rest of three genomes is statistically highly significant ($P < 10^{-10}$). In mouse case, we also observed the low percentage, about 10.7% (data not shown). Though the inherent mechanism remains unclear, it appears that fewer TATA-containing genes in mammalian genomes may be the general pattern.

Moreover, we observed that in all four genomes, the TATA box is significantly enriched in duplicate promoters compared with single-copy genes (fig. 1). For instance, in Arabidopsis, the percentage of TATA-containing genes is 24.6% in duplicates versus 16.0% in single-copy genes ($P < 10^{-11}$), so that the enrichment fold (in duplicates) is about 24.6/16.0 = 1.54. This enrichment fold is 1.23 in worm and 1.67 in yeast. Interestingly, in the human genome, the enrichment fold is about 1.54, almost the same as that in Arabidopsis, though the proportion of TATA-containing genes in human is almost 50% less than that in Arabidopsis.

We therefore conclude that, in general, TATA-containing genes have been enriched in duplicates roughly up to about 1.5-fold. It should be noticed that in our study, duplicate genes were defined as those having at least one nonself hit with the cutoff *E*-value = 10^{-20} in the all-against-all BlastP searches. Meanwhile, single-copy genes do not have any discernable duplicate copy with the cutoff *E*-value = 0.1 in the same genome. Our main results hold when other common *E*-value cutoffs for duplicate gene determination, such as 1.0×10^{-5} or 1.0×10^{-10} were used.

The above findings may imply the potential influence of the TATA box on duplicate gene evolution. But, its underlying mechanism is still unclear. Moreover, to explain the overrepresentation of the TATA box in duplicate promoters, Kim et al. (2010) suggested that the TATA box may be associated with gene duplicability, that is, TATA-containing genes have the higher probability to be duplicated or maintained. But, it could also be caused by the gain of new TATA boxes after the gene duplication. Apparently, such simple enrichment analysis also cannot distinguish between these two possibilities. To address these



A *H. sapiens*

	Duplicates	Singletons	WGD	SSD
🟥 TATA(+)	1138	119	703	436
🟦 TATA(-)	8124	1370	4878	3283

B *C. elegans*

	Duplicates	Singletons
🟥 TATA(+)	1573	450
🟦 TATA(-)	4834	1802

C *A. thaliana*

	Duplicates	Singletons	WGD	SSD
🟥 TATA(+)	2758	221	880	1881
🟦 TATA(-)	8454	1162	2222	6242

D *S. cerevisiae*

	Duplicates	Singletons	WGD	SSD
🟥 TATA(+)	569	240	206	370
🟦 TATA(-)	1609	1297	547	1076

FIG. 1. TATA box is overrepresented in duplicate gene promoters. The figure shows the number of all duplicates, singletons, WGD, and SSD genes classified as TATA-containing genes and TATA-less genes in (*A*) human, (*B*) worm, (*C*) Arabidopsis, and (*D*) yeast. The statistical inference was conducted by Chi-square test via R program. This figure is available in color online and in black and white in print.

issues, we implemented a genome-wide phylogenetic analysis in yeast, taking advantage of the availability of substantial yeast functional genomics.

Promoter TATA Box Evolution in Yeast Gene Families

We wish to reconstruct ancestral states of the TATA box and infer gain or loss events of the motif during evolution along the phylogeny of gene family. Treating the TATA box as a discrete molecular character with two states (present or absent), we can use a lot of computational methods (e.g., Fitch 1971; Swofford and Maddison 1987; Maddison WP and Maddison DR 1992; Schluter 1995; Pagel 1999; Huelsenbeck et al. 2000). See Materials and Methods for the detail of the computational pipeline designed to carry out a genome-wide analysis. In short, we reconstructed over 700 phylogenies of yeast protein families, each of which has at least two duplicate genes in *S. cerevisiae*. Due to the computational efficiency of genome-wide analysis, we mainly used the NJ method to infer the gene family tree. Then, the ancestral TATA box state was reconstructed by MP with ACCTRAN and DELTRAN optimization algorithms.

We used the yeast tricalbin family as an example to illustrate our analysis. The tricalbins comprise a novel, synaptotagmin-like protein family which play a potential role in membrane trafficking and sorting (Schulz and Creutz 2004). In *S. cerevisiae*, it contains YOR086C (TCB1), YNL087W (TCB2), YML072C (TCB3), as well as a distantly related protein YOR296W, a putative protein with unknown function. We collected the orthologous protein sequences of tricalbin family from five hemiascomycetous yeast species and *D. melanogaster*. Figure 2A shows the inferred phylogeny by the NJ method with the Poisson distance. High bootstrapping values at most nodes suggest that the inference is statistically reliable. Using other tree-making methods such as parsimony (MEGA, Tamura et al. 2007), maximum likelihood (ML) (Phyml, Guindon and Gascuel 2003), and bayesian inference (BI) (MrBayes, Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003), resulted in virtually the same topology (not shown). When the two fruit fly genes were used as the outgroup, the rooted phylogeny can be simplified to the topology that contains genes only from a single genome *S. cerevisiae* (fig. 2B).

According to Basehoar et al. (2004), for four yeast tricalbin genes, only TCB2 is TATA-containing gene and others are TATA-less genes. Then one can easily map the status of the TATA box onto the phylogeny of the yeast tricalbin gene family. Apparently, the ancestral status of the TATA box can be reconstructed unambiguously by the MP method. Most likely, the common ancestral TATA box state of the tricalbin family was absent, and there was one TATA box gain event occurred in the lineage toward TCB2, after the gene duplication that resulted in TCB1 and TCB2 (fig. 2B). Functional analysis suggested that TCB1 and TCB3, both TATA box absent, are functionally redundant, and involved in the process of calcium-responsive lipid binding, whereas TCB2 (TATA-containing), insensitive to calcium, may have a unique role in mediating the creation of hetero-oligomeric tricalbin complexes (Schulz and Creutz 2004).

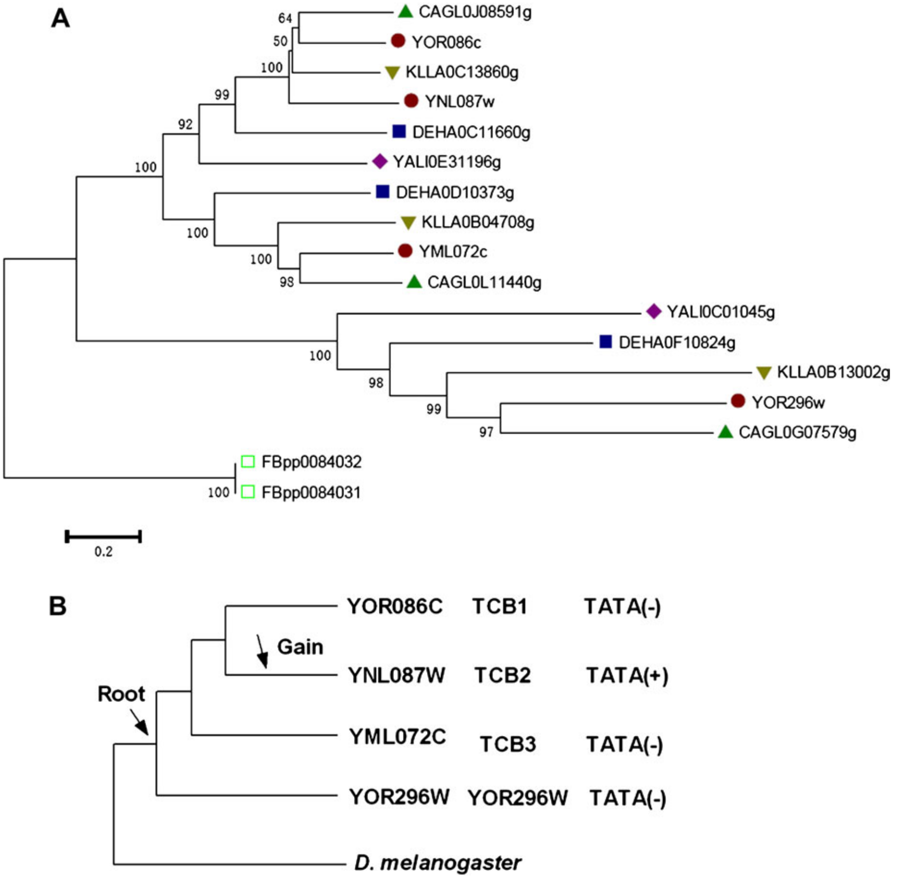


Fig. 2. Illustration example for the yeast tricalbin family. (*A*) represents the NJ tree of the tricalbin protein family, while (*B*) ancestral TATA box state reconstruction process of the tricalbin family by the MP method. Bootstrapping with 1,000 replicates was conducted to evaluate the reliability of evolutionary relationship among member genes. *Drosophila melanogaster* was used as outgroup to determine the root of NJ tree. Different signs in the tips of tree represent homologous tricalbin genes from different species, where solid circle: *Saccharomyces cerevisiae*, Solid triangle: *Candida glabrata*, solid inverted triangle: *Kluyveromyces lactis*, solid square: *Debaryomyces hansenii*, solid diamond: *Yarrowia lipolytica*, and blank square: outgroup *D. melanogaster*. This figure is available in color online and in black and white in print.

More TATA Box Gains than Losses after the Yeast Gene Duplication

Among 542 gene families with only two duplicate genes of *S. cerevisiae*, there are 368 gene families that both duplicates are absence of TATA-box in their promoter regions, 49 families with both duplicates being TATA-containing (table 1). The ancestral state inference of the TATA box for these families is obvious. However, for the rest 125 two-duplicate gene families with only one TATA-containing duplicate, we are not able to infer the ancestral status of the TATA box, though we know at least one TATA box status change occurred.

For gene families with three or more duplicate genes in *S. cerevisiae*, we may infer gain and loss events of the TATA box along the phylogeny (fig. 2), using the parsimony-based algorithms called ACCTRAN and DELTRAN (Farris 1970; Swofford and Maddison 1987). Overall, for 225 yeast gene families with three or more duplicates in *S. cerevisiae*, we identified 101 parsimonious gain or loss events (table 2). Comparing the results between the ACCTRAN and DELTRAN analyses is helpful to evaluate the bias caused by the inference ambiguity of parsimony method. Generally, ACCTRAN optimization tends to infer more TATA box loss than the gain, whereas DELTRAN is on the contrary. We found only three cases show the difference between these two treatments, but the main result is the same, that is, both algorithms show that the number of TATA-box gain events is much larger than that of TATA-box loss events.

Table 1. Ancestral TATA Box State Reconstruction for *S. cerevisiae* Gene Families by Parsimony Principle.

Two Paralogs		At Least Three Paralogs	
Ancestral TATA Box State	Number of Families	Ancestral TATA Box State	Number of Families
TATA (+)	49	TATA (+)	27
TATA (-)	368	TATA (−)	161
Ambiguity	125	Ambiguity	37

Table 2. Events of TATA Box Switches During the Yeast Gene Family Evolution According to Different Parsimony Optimizations.

Parsimony Optimizations	TATA Box Gains	TATA Box Losses	Binomial Test (*P-value*)
ACCTRAN	75	26	$<10^{-5}$
DELTRAN	78	23	$<10^{-7}$

The gain-to-loss ratio is about 2.9~3.4, and the null hypothesis of equal gain and loss was rejected highly significantly (table 2). Moreover, for those gene families with at least three paralogous members, except for 37 gene families that have ambiguous ancestral state of the TATA box, we inferred that the ancestral genes of 161 gene families were likely to be TATA-less, whereas those of 27 gene families were TATA-containing (table 1). Since at the genome level, the TATA box is found in about 20% of yeast genes (Basehoar et al. 2004), the expected proportion of TATA-containing genes to TATA-less genes is approximately 0.25. However, we observed that the ratio of the inferred presence/absence of the promoter TATA box in ancestral genes of these yeasts is even lower (27/161 = 0.168); with a marginal significance, it is less than the expected value 1/4 (Binomial test, $P < 0.001$). With a more conservative test to assign the ambiguous case equally to TATA(+) and TATA(−) groups, we found the ratio (0.25) is the same as the genome expectation. Whereas for yeast gene families with only two duplicate genes, putative presence/absence ratio of the ancestral TATA box status is much lower (49/368 = 0.133), compared with that of gene families with at least three members (table 1). These results suggested that TATA-containing genes are not duplication preferred, and the enrichment of the TATA box in yeast duplicate genes may be the consequence of contiguous gains of new TATA box after gene duplications. Since the promoter TATA box plays an essential role in regulating transcription in response to environmental stress (Basehoar et al. 2004), duplication-related gain of the TATA box in the promoter region may provide the genetic resource for evolutionary innovations and adaptation in response to the environmental stress.

Functional Preference of TATA-Containing Duplicates

If acquiring the promoter TATA box after the gene duplication was to enhance the capability of organismal responses to environmental stresses, we may predict that such acquisition may occur in some gene families with special functional preference. We designed the following analysis to test this hypothesis, with the help of yeast gene families and Gene Ontology. First, we consider three groups of yeast duplicate genes according to the evolutionary pattern of the TATA box.

1) $D_{TATA(+)}$-gain group, i.e. any duplicate gene in this group may have acquired a new TATA box after the gene duplication. In some gene families, a single TATA-box gain event may result in multiple closely related duplicates to be TATA-containing. That is, why we obtained 87 duplicate genes based on 75 gain events (ACCTRAN, table 2). 2) $D_{TATA(+)}$-loss group, i.e. any duplicate gene in the group may have lost its ancestral TATA box after the gene duplication. Similarly, we obtained 31 duplicate genes from 26 loss events (ACCTRAN, table 2). 3) $D_{TATA(+)}$-ancestral group, that is, without any detectable gain or loss, the TATA box in any duplicate gene in this group was inherited from their ancestral TATA-containing genes. Our study collected 84 $D_{TATA(+)}$ ancestral genes from yeast gene families with at least three *S. cerevisiae* paralogous members. List of genes in each group can be found in supplementary tables S3, S4 and S5, (Supplementary Material online), respectively.

We then carried out the GO-based analysis on biological process category. Impressively, 42.5% (37 of 87) $D_{TATA(+)}$-gain duplicates are found in the GO Slim transport, and 14.9% in the GO Slim response to chemical stimulus, and 11.5% in the GO Slim vesicle-mediated transport (fig. 3A). In other words, most $D_{TATA(+)}$-gain duplicates are overdistributed in these three biological processes apparently associated with the stress response. By contrast, duplicate genes in the $D_{TATA(+)}$-loss group do not show any biological process preference (Data not shown). Interestingly, we found that these $D_{TATA(+)}$-ancestral genes were moderately enriched in some GO Slims of biological processes, such as generation of precursor metabolites and energy, cellular carbohydrate metabolic process (fig. 3B).

Using DAVID bioinformatics resources (Dennis et al. 2003; Huang et al. 2009), we further statistically tested the patterns of enrichment in terms of multiple functional categories. Indeed, Fisher's exact test shows that $D_{TATA(+)}$-gain duplicates are significantly enriched in the transport ($P < 10^{-3}$), including drug transporter activity ($P < 10^{-3}$), ATP-binding cassette transporter ($P < 10^{-2}$), cell membrane ($P < 10^{-5}$), and extracellular ($P < 10^{-7}$), etc., all of which are believed to be involved in the direct response to the external signal stimuli (supplementary fig. S4, Supplementary Material online). The significance holds in each case ($P < 0.01$) after the Benjamini–Hochberg correction. Note that genes in the $D_{TATA(-)}$-loss group show some degree of functional enrichment in several biosynthetic and metabolic processes ($P < 0.01$) (supplementary fig. S5, Supplementary Material online). In addition, DAVID functional enrichment analysis reveals that $D_{TATA(+)}$-ancestral duplicates represent significant enrichment in the glycolysis/gluconeogensis and related metabolic processes ($P < 10^{-4}$ with Benjamini–Hochberg correction; supplementary fig. S6, Supplementary Material online). Together, our results suggested that, in the yeast, while $D_{TATA(+)}$-ancestral duplicates are mainly responsible for normal development or growth of the organism, $D_{TATA(+)}$-gain paralogs may have a primary role in adapting to changing extracellular environments by directly responding to external signals.

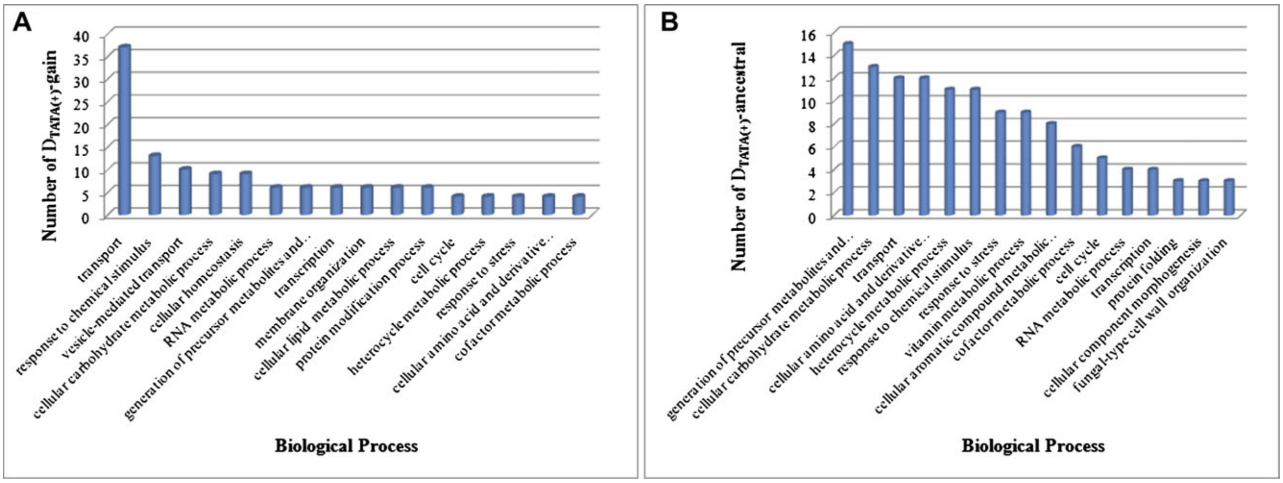


Fig. 3. Distribution of duplicates for (*A*) $D_{TATA(+)}$-gain and (*B*) $D_{TATA(+)}$-ancestral groups by biological process classification (Gene Ontology criterion). This figure is available in color online and in black and white in print.

Evolutionary Time of TATA Box Gain Events

In above analyses, we did not take the effect of young or old gene families into account. This is related to an intriguing question whether the TATA box gain and loss events had occurred constantly or episodically during the course of yeast evolution.

To tackle this issue, we first divided duplicate pairs into two categories, the TATA (+/+(−)) group for duplicate pairs with at least one copy containing a TATA box, and the TATA (−/−) group for the case when both copies are TATA-less genes. Then, we calculated the distance of synonymous substitutions (K_S) between each duplicate pair. As synonymous substitution distance is statistically unreliable when K_S becomes larger, we removed duplicate pairs with $K_S > 2.0$ for the analysis. On average $K_S = 0.93 \pm 0.06$ for TATA(+/+(−)) and $K_S = 0.74 \pm 0.06$ for TATA(−/−), and the K_S-difference between two categories is significant ($P < 0.05$, Wilcoxon rank sum test). Since K_S between duplicate genes was the proxy of evolutionary time since the duplication, our result demonstrates that the TATA box may be enriched in relatively ancient duplicate genes.

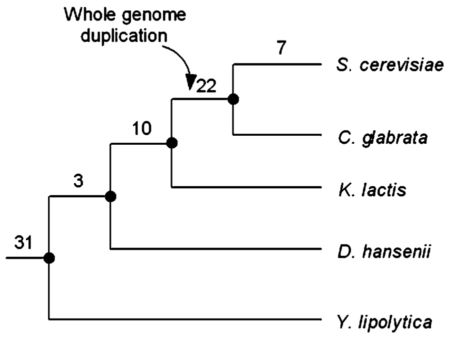


Fig. 4. The age distribution of gene duplications where duplicates experienced TATA box gain events. The relative age of duplication event was estimated from reconciled tree based on gene tree and species tree. Numerals above the branches indicate the number of the gene duplication occurring in the lineage where duplicated copies undergone TATA box switches (gain) thenceforward. Black circles denote the events of speciation.

Furthermore, based on species tree of yeasts (Dujon 2006), we roughly estimated the phylogenetic time interval of the TATA box gain event by the method of gene tree/species tree reconciliation (Page and Charleston 1997). If the orthologous genes of some yeast species were not available, the time interval estimation was dubious. Hence, after excluding these unclear cases, we obtained 73 TATA box gain events in paralogs whose gene duplication relative time can be determined by the reconciled tree. The result shows an apparent two-mode pattern (fig. 4): Gain events were preferentially occurred 1) on relatively recent paralogous genes (model I, 29 events, ~40%) duplicated after the split of *S. cerevisiae* and *K. lactis*, and 2) on very ancient duplicates (mode II, 31 events, ~42%) before the speciation of *S. cerevisiae* and *Y. lipolytica*.

Since the WGD of yeast organisms was happened after the divergence of *Saccharomyces* from *Kluyveromyces* but before the speciation of *S. cerevisiae* and *C. glabrata* (Wolfe and Shields 1997), we wonder whether the TATA box gain events in the recent model I were mainly arose from the WGD. Interestingly, according to the definition of WGD duplicate genes by Byrne and Wolfe (2005), we found that, duplicate genes from 33 gain events (total 75) were actually inferred as paralogous pairs from the WGD.

Gain of the TATA Box after Gene Duplication Accelerates Asymmetric Expression Divergence Related to Environmental Stresses

A number of studies have shown that genes with the TATA box are associated with the response to environmental stresses (Basehoar et al. 2004; Huisinga and Pugh 2004). Indeed, TATA-containing genes tend to have a high rate of gene expression variability between species (Tirosh et al. 2006), as well as increased expressional genetic variations (Landry et al. 2007). Based on these observations, one may hypothesize that the acquisition of the TATA box after

Table 3. Pairwise *E*-distances between Four Paralogs of the Yeast Tricalbin Family.

E_{env}/E_{dev}	TCB1	TCB2	TCB3	YOR296W
TCB1	1	0.8432	0.1820	0.3555
TCB2	0.6504	1	0.5977	0.6302
TCB3	0.2321	0.3358	1	0.2571
YOR296W	0.2141	0.5739	0.2497	1

the duplication may accelerate the asymmetric expression divergence specific to environmental stresses.

We used two *S. cerevisiae* expression data sets to test this hypothesis. One is the expression profile data set induced by various environmental stresses (Gasch et al. 2000), and the other is the normal developmental expression profile data set (Chu et al. 1998; Spellman et al. 1998). For each yeast gene family with three or more duplicate genes of *S. cerevisiae*, we calculated the (pairwise) expression distance (*E*) matrix from the environmental stress data set and the developmental data set separately, denoted by E_{env} and E_{dev}, respectively. For instance, table 3 shows the *E*-distances between each pair of four duplicates of *S. cerevisiae* tricalbin family, upper-diagonal for E_{env}, and lower-diagonal for E_{dev}. Apparently, TCB2 (TATA-containing) has the largest expression distances compared with the rest of three duplicates under the condition of environmental stresses.

Following Gu et al. (2005), we designed a simple test for the stress-specific asymmetric expression divergence driven by the acquisition of the promote TATA box. Consider a gene family or subfamily with three duplicates (or groups), denoted by *A*, *B*, and *C*, respectively (fig. 5A). Suppose that only gene (group) *A* is TATA-containing, and duplicate gene (group) *C* is the outgroup. Let *E(A,C)* and *E(B,C)* be the (mean) expression distances between *A* and *C*, as well as *B* and *C*, respectively. The hypothesis that the TATA box acquisition in gene *A* may initiate higher expression divergence responding to environmental stresses predicts $E_{env}(A,C) > E_{env}(B,C)$, whereas $E_{dev}(A,C) = E_{dev}(B,C)$ in the case of normal development. Given the phylogeny (fig. 2), we assign TCB2 (TATA-containing), TCB1, and TCB3 as gene *A*, *B* and *C*, respectively. We then calculated $E_{env}(A,C) = 0.\ 60$ and $E_{env}(B,C) = 0.18$ (table 3). Bootstrapping test (Gu et al. 2005) shows $E_{env}(A,C) > E_{env}(B,C)$ is statistically significant ($P < 0.01$). Moreover, there is no statistical significance to reject the null hypothesis $E_{dev}(A,C) = E_{dev}(B,C)$ ($P > 0.05$).

To further measure the degree of stress-dependent expression asymmetry in the gene *A* (TCB2), we calculated the expression branch length of the group *A* and the group *B*, denoted by *E(A)* and *E(B)*, respectively. Since $E_{env}(A,B) = 0.84$, we have $E_{env}(A) = [E_{env}(A,B) + E_{env}(A,C) - E_{env}(B,C)]/2 = 0.63$ and $E_{env}(B) = [E_{env}(A,B) + E_{env}(B,C) - E_{env}(A,C)]/2 = 0.21$. Hence, since the duplication, the rate of (stress-dependent) expression divergence in the gene *A* (TCB2) is about as 0.63/0.21 = 3 folds as that in the gene *B* (TCB1) (fig. 5B). As expected, under the normal developmental conditions, the expression divergence between these two gene groups is virtually the same (fig. 5C).

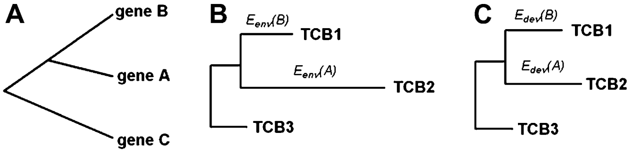


Fig. 5. Relative-rate test for asymmetric evolution of expression divergence after the acquisition of the TATA box in the yeast tricalbin family. (*A*) Schematic tree for three duplicates (groups). Relative (*B*) stress-dependent and (*C*) development-dependent expression rate test between TCB2 with the TATA box gain and TCB1 without.

In short, our analysis suggests that greater expression divergence of TCB2 under environmental stresses may have occurred as the result of the TATA-box acquisition. Tentatively we identified stress conditions in which TCB2 has its own unique expression pattern (fig. 6), such as the conditions of heat shock, sulfhydryl oxidizing agent diamide, etc., where the expression of TCB2 is down-regulated.

An immediate follow-up question is whether our finding in the yeast tricalbin family is a genome-wide pattern. To this end, we extended the analysis to all 75 TATA-box gain events. Paired *t*-test has shown that the stress-dependent expression branch length of group *A* ($E_{env}(A)$) is significantly greater than that of group *B* ($E_{env}(B)$) ($P < 0.01$) (fig. 7A). Furthermore, an opposite trend that may occur when the promoter TATA box has been lost after the gene duplication, as shown by our analysis on 26 TATA-box loss events; yet the statistical significance is marginal ($P < 0.10$) (fig. 7B). As expected, the evolution of the TATA box gain or loss did not cause any asymmetric expression divergence between duplicates under normal development conditions ($P > 0.10$).

Discussion

In this study, we have carried out a genome-wide analysis to reveal the evolutionary pattern of the *cis*-regulatory element TATA box after the gene duplication and its association with expression evolution and environmental stress responses. Clearly, duplicate gene promoters have higher TATA box preference than that of single-copy genes, as shown in the human, worm, Arabidopsis, and yeast genomes. To further explore the underlying mechanism, we developed a computational pipeline for the genome-wide analysis of ancestral TATA box inference along the phylogeny of a gene family. Applying this method to yeast genomes, we found that ancestors of most yeast gene families were TATA box absent. After the gene duplication, a greater number of TATA box gain events than that of loss events had occurred. Our findings suggest that overrepresentation of the TATA box in extant yeast duplicate promoters may not be the consequence of high duplicability of TATA-containing genes (Kim et al. 2010), instead, we argue that more likely, the subsequent TATA box gain events after the gene duplication.

Moreover, our further analysis indicates that, after acquiring the TATA box, these duplicate genes on

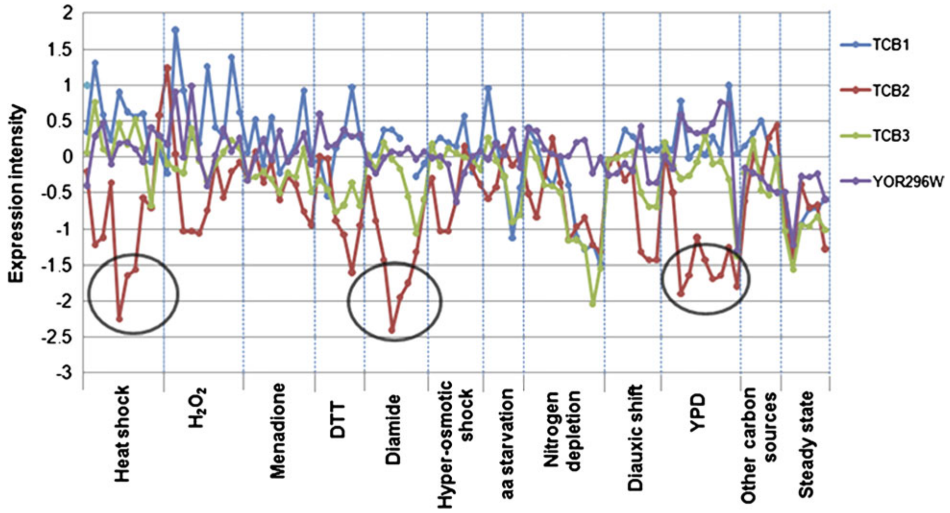


FIG. 6. Expression profile of the yeast tricalbin family in response to externally environmental stresses. Expression intensity was normalized according to what the original study recommended. This figure is available in color online and in black and white in print.

average may have experienced greater expression divergence that distinctly differed from the ancestral expression pattern, but only under the environmental stresses, not under the normal physiological conditions. As expected, these TATA-gain duplicate genes are particularly enriched in those stress-related functional categories, but not the case for TATA-ancestral duplicates. These lines of evidence together support the hypothesis that the gain of the TATA box after the duplication may play a nontrivial role in yeast duplicate preservation by accelerating relatively stress-responsive expression divergence.

Duplicate Genes and Environmental Stress Responses

Our current study provides some clear evidence to illustrate the relationship between the gene duplication and the stress-driven expression divergence, supporting the notion that the gene duplication may contribute to environmental adaptation by providing new stress sensing and regulatory response mechanisms. It has been reported that duplicate genes were overrepresented in several stress-associated functional categories (Maere et al. 2005; Moore and Purugganan 2005; Rizzon et al. 2006; Hanada et al. 2008). For instance, Hanada et al. (2008) observed that lineage-specific tandem duplicates tend to be involved in responses to environmental stimuli in Arabidopsis. Besides, Zou et al. (2009) suggested that duplicate genes could experience asymmetric partitioning of ancestral stress responses, likely due to differential loss of DNA regulatory elements. Since the ability to timely and properly cope with environmental stresses is crucial to the survival of all living organisms, and the gene duplication has been thought as one main genetic resource of evolutionary novelties, gain of the TATA box (a stress-sensitive regulatory motif) after the duplication could be an important mechanism for organisms to adaptively deal with drastically changing environments via more flexible expression regulation program. Yet, we cannot rule out the possibility that organisms may adapt to variable environments by duplicating stress-response genes.

TATA Box and WGD

Our results also reveal that about half of TATA box gain events were occurred in WGD duplicate genes. According

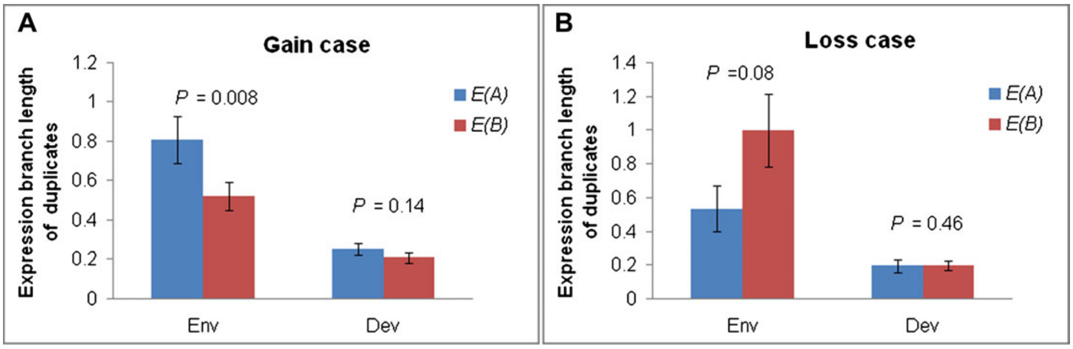


FIG. 7. Comparison of relative rate of expression divergence between the group A and the group B in (*A*) gain case and (*B*) loss case. Here, expression variations disturbed by external environments and internal developmental processes were shortly denoted as Env and Dev, respectively. Paired *t*-test was applied to obtain *P* -values. This figure is available in color online and in black and white in print.

to the expression divergence hypothesis for TATA-gain duplicates, it is expected that WGD genes may have higher expression variability. The notion is supported by the study of Guan et al. (2006), which proposed that WGD paralogs represent more diversity in expression pattern and upstream regulatory regions compared with paralogs from SSD. To further know the association between gene duplication mechanisms and the TATA box, we subdivided duplicate genes into WGD generating paralogs and SSD derived paralogs and observed that WGD paralogs have no differentiation in the TATA box preference with SSD derived duplicates except in Arabidopsis (fig. 1). Such phenomenon may due to the overrepresentation of TATA-ancestral duplicates in SSD. Interestingly, we found that in the land plant Arabidopsis, a higher percentage of the WGD generating paralogs is TATA-containing than that of SSD derived duplicates (fig. 1). We assumed that it might because Arabidopsis genome had undergone multiple-round WGD events (Blanc and Wolfe 2004), therefore having more chance to acquire the TATA box in their WGD duplicate promoters. Further analysis will address these issues.

Technical Issues in Ancestral State Inference

There is a more straightforward approach to address the issue we focused, that is, to compare the gene that are duplicated in one species but is a singleton in another species (Qian et al. 2010), then ask whether the duplicate copy (or copies) are more likely to have a TATA box. For example, duplicate genes in *S. cerevisiae* and their singleton ortholog in a pre-WGD species can be chosen for the analysis. But, this implementation has some limitations. First, if we directly applied the well-acknowledged post-WGD species like *S. cerevisiae* and pre-WGD species like *K. lactis*, the inconsistent of phylogenetic tree with WGD hypothesis may bring the bias to the inference. Otherwise, it still needs phylogenetic analysis to determine whether the duplication event was occurred after the split of two species used. Second, such simple approach just focuses on single lineage-specific gene duplication, such as one round of WGD, and misses much information of duplication events that happened in other species. Therefore, we think that our analysis based on multiple-species phylogenetic analysis (reconstructing the ancestral TATA state) should be more reliable than that based on less species.

Now, inference of ancestral states has been widely used to study phenotypic or genetic evolution for ecological and morphological characters or genetic/genomic features along a species tree or a gene family tree (Jermann et al. 1995; Moran 2004; Leschen and Buckley 2007; Ekman et al. 2008). The MP method (Fitch 1971; Swofford and Maddison 1987; Maddison WP and Maddison DR 1992) is generally robust against random errors, but requires the assumption of slow evolution, otherwise, the inference could be misleading (Felsentein 1973; Frumhoff and Reeve 1994; Schultz et al. 1996; Cunningham et al. 1998). Our study used the MP inference, mainly because the evolution of the TATA box turnover is slow. For 225 gene families with at least three paralogs of *S. cerevisiae*, we have observed 105 families have no experience in the change of the TATA box along the phylogeny, 93 families bear only one possible state change. In these case, different methods such as ML (Schluter 1995; Pagel 1999) or the BI (Huelsenbeck et al. 2000), give virtually the same result. We have checked the ancestral state inference for 27 families that may be involved in more than one TATA box changes by three (MP, ML, and BI) methods under the same phylogeny. The only difference is about some cases when MP assigns the 'uncertain ancestral state', namely, 50% chance for either presence or absence of the TATA box, whereas ML or BI may give slightly weight in favor of one state, depending on the option of parameter setting. In this sense, the MP method may provide more conserved analysis.

In our genome-wide phylogenetic analyses and ancestral TATA inference, the inferred topology should be reliable, particularly for the evolutionary relationship of duplicate genes. To evaluate the error caused by the phylogenetic tree with relatively low statistical confidence, we selected phylogenies whose bootstrap values of all nodes of gene duplications are larger than 60%, 70%, 80% and 90%, respectively and then reanalyzed the TATA box evolution. As shown in supplementary tables S1 and S2 (Supplementary Material online), our main result was not altered by various cutoffs. Secondly, we used several other major tree-making algorithms, including the MP, ML, and bayesian to rule out any tree-making dependent misleading. Since these methods are time-consuming when the gene family is large, we randomly selected 30 gene families to perform the analysis. Though there were minor differences among the topologies constructed by different methods, the evolutionary relationship of different duplicate genes in each gene family was not changed in all cases. Hence, we conclude that the phylogenetic uncertainty is unlikely to affect our main results. We notice that some bayesian inference methods (Ronquist 2004; Vanderpoorten and Goffinet 2006) considered the phylogenetic uncertainty and the uncertainty in character mapping simultaneously, but the reliability of these methods remains to see.

Finally, we would like to address several caveats in our study. One is that the ancestral TATA box states of 37 families (about 16%) are ambiguous according to the parsimony principle. We also reconstructed the TATA box state of ancestors of these families by bayesian MCMC method with BayesTraits software (Pagel et al. 2004). The obtained posterior probability of two states of the TATA box (present or absent) for each family is very close to 50%, which also failed to confidently infer ancestral TATA box of these families. Thereby, more information beyond sequence is needed. The other caveat is the determination of the root of the phylogenetic tree, which is always a tough problem in the phylogenetic analysis. For 120 yeast gene families that may be involved in TATA box turnover, we have to determine the root of the phylogeny. Typically, we used animal such as *D. melanogaster* as an outgroup to ascertain the root of the trees. Although 71 families were successfully rooted, there were 49 families that could not be rooted because duplication events had occurred before the split of animal/fungi. Since we observed that in these unable-rooted families, the branch lengths with respect to duplicate genes differed considerably, suggesting that one can intuitively speculate the root based on the longest branch length. We thus tentatively used the midpoint approach as a second way to help rooting the tree. For comparison, we applied the midpoint approach to those 71 gene families that can be rooted by the outgroup approach. We found only five families gives the different gain–loss inference, which did not alter our main results. In short, in spite of some technical limitations and caveats, we are confident that our main findings are solid.

Supplementary Material

Supplementary materials, tables S1–S5, and figures S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournal.org/).

Acknowledgments

We are grateful to Yanwu Zeng for his suggestions on some technical issues. Great thanks also go to two anonymous reviewers' constructive comments, conducing to refine our study. This work was partially supported by the Fudan University Innovation Fund for Graduate Students.

References

Basehoar AD, Zanton SJ, Pugh BF. 2004. Identification and distinct regulation of yeast TATA box-containing genes. *Cell* 116:699–709.

Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res.* 13:137–144.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell.* 16:1667–1678.

Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* 15:1456–1461.

Chain FJ, Ilieva D, Evans BJ. 2008. Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evol Biol.* 8:43.

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I. 1998. The transcriptional program of sporulation in budding yeast. *Science* 282:699–705.

Cunningham CW, Omland KE, Oakley TH. 1998. Reconstructing ancestral character states: a critical reappraisal. *Trends Ecol Evol.* 13:361–366.

Dennis G Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4:P3.

Dujon B. 2006. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet.* 22:375–387.

Dujon B, Sherman D, Fischer G, et al. 2004. (67 co-authors). Genome evolution in yeasts. *Nature* 430:35–44.

Ekman S, Andersen HL, Wedin M. 2008. The limitations of ancestral state reconstruction and the evolution of the ascus in the Lecanorales (lichenized Ascomycota). *Syst Biol.* 57:141–156.

Farris JS. 1970. Methods for computing wagner trees. *Syst Zool.* 19:83–92.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791.

Felsentein J. 1973. Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from discrete characters. *Syst Zool.* 22:240–249.

Fitch MW. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst Zool.* 20:406–416.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.

Frumhoff PC, Reeve HK. 1994. Using phylogenies to test hypotheses of adaptation: a critique of some current proposals. *Evolution* 48:172–180.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell.* 11:4241–4257.

Gu X, Zhang Z, Huang W. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc Natl Acad Sci U S A.* 102:707–712.

Gu Z, Nicolae D, Lu HHS, Li W-H. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* 18:609–613.

Gu Z, Rifkin SA, White KP, Li W-H. 2004. Duplicate genes increase gene expression diversity within and between species. *Nat Genet.* 36:577–579.

Guan Y, Dunham MJ, Troyanskaya OG. 2006. Functional analysis of gene duplications in Saccharomyces cerevisiae. *Genetics* 175: 933–943.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.

Ha M, Kim E-D, Chen ZJ. 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proc Natl Acad Sci U S A.* 106:2295–2300.

Ha M, Li W-H, Chen ZJ. 2007. External factors accelerate expression divergence between duplicate genes. *Trends Genet.* 23:162–166.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.* 148:993–1003.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 4:44–57.

Huelsenbeck JP, Rannala B, Masly JP. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. *Science* 288: 2349–2350.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.

Huisinga KL, Pugh BF. 2004. A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. *Mol Cell.* 13:573–585.

Huminiecki L, Wolfe KH. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res.* 14:1870–1879.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet.* 11:97–108.

Jermann TM, Opitz JG, Stackhouse J, Benner SA. 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374:57–59.

Kim Y, Lee JH, Babbitt GA. 2010. The enrichment of TATA box and the scarcity of depleted proximal nucleosome in the promoters of duplicated yeast genes. *J Mol Evol.* 70:69–73.

Kondrashov FA, Kondrashov AS. 2006. Role of selection in fixation of gene duplications. *J Theor Biol.* 239:141–151.

Kondrashov FA, Koonin EV. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet.* 20:287–290.

Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL. 2007. Genetic properties influencing the evolvability of gene expression. *Science* 317:118–121.

Leach LJ, Zhang Z, Lu C, Kearsey MJ, Luo Z. 2007. The role of cis-regulatory motifs and genetical control of expression in the divergence of yeast duplicate genes. *Mol Biol Evol.* 24:2556–2565.

Leschen RA, Buckley TR. 2007. Multistate characters and diet shifts: evolution of Erotylidae (Coleoptera). *Syst Biol.* 56:97–112.

Li J, Musso G, Zhang Z. 2008. Preferential regulation of duplicated genes by microRNAs in mammals. *Genome Biol.* 9:R132.

Li W-H, Yang J, Gu X. 2005. Expression divergence between duplicate genes. *Trends Genet.* 21:602–607.

Li Z, Zhang H, Ge S, Gu X, Gao G, Luo J. 2009. Expression pattern divergence of duplicated genes in rice. *BMC Bioinformatics.* 10(Suppl 6):S8.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–473.

Maddison WP, Maddison DR. 1992. MaClade: analysis of phylogeny and character evolution. Sunderland (MA): Sinauer Associates.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A.* 102:5454–5459.

Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A.* 107:9270–9274.

Makova KD, Li WH. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.* 13:1638–1645.

Moore RC, Purugganan MD. 2005. The evolutionary dynamics of plant duplicate genes. *Curr Opin Plant Biol.* 8:122–128.

Moran AL. 2004. Egg size evolution in tropical American arcid bivalves: the comparative method and the fossil record. *Evolution* 58:2718–2733.

Nikolski M, Sherman DJ. 2007. Family relationships: should consensus reign?—consensus clustering for protein families. *Bioinformatics* 23:e71–e76.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol.* 302:205–217.

Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

Orphanides G, Lagrange T, Reinberg D. 1996. The general transcription factors of RNA polymerase II. *Genes Dev.* 10:2657–2683.

Page RDM, Charleston MA. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol.* 7:231–240.

Pagel M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst Biol.* 48:612–622.

Pagel M, Meade A, Barker D. 2004. Bayesian estimation of ancestral character states on phylogenies. *Syst Biol.* 53:673–684.

Papp B, Pal C, Hurst LD. 2003. Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet.* 19:417–422.

Qian W, Liao B-Y, Chang AY-F, Zhang J. 2010. Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends Genet.* 26:425–430.

Rizzon C, Ponger L, Gaut BS. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. *PLoS Comput Biol.* 2:e115.

Rodin SN, Parkhomchuk DV, Riggs AD. 2005. Epigenetic changes and repositioning determine the evolutionary fate of duplicated genes. *Biochemistry(Mosc).* 70:559–567.

Ronquist F. 2004. Bayesian inference of character evolution. *Trends Ecol Evol.* 19:475–481.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.

Schluter D. 1995. Uncertainty in ancient phylogenies. *Nature* 377:108–110.

Schultz TR, Cocroft RB, Churchill GA. 1996. The reconstruction of ancestral character states. *Evolution* 50:504–511.

Schulz TA, Creutz CE. 2004. The tricalbin C2 domains: lipid-binding properties of a novel, synaptotagmin-like yeast protein family. *Biochemistry* 43:3987–3995.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell.* 9:3273–3297.

Swofford DL, Maddison WP. 1987. Reconstructing ancestral character states under Wagner parsimony. *Math Biosci.* 87:199–229.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol.* 24:1596–1599.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25:4876–4882.

Tirosh I, Berman J, Barkai N. 2007. The pattern and evolution of yeast promoter bendability. *Trends Genet.* 23:318–321.

Tirosh I, Weinberger A, Carmi M, Barkai N. 2006. A genetic signature of interspecies variations in gene expression. *Nat Genet.* 38:830–834.

Vanderpoorten A, Goffinet B. 2006. Mapping uncertainty and phylogenetic uncertainty in ancestral character state reconstruction: an example in the moss genus Brachytheciastrum. *Syst Biol.* 55:957–971.

Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proc Natl Acad Sci U S A.* 97:6579–6584.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708–713.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol.* 17:32–43.

Zhang Z, Gu J, Gu X. 2004. How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? *Trends Genet.* 20:403–407.

Zheng D. 2008. Asymmetric histone modifications between the original and derived loci of human segmental duplications. *Genome Biol.* 9:R105.

Zou C, Lehti-Shiu MD, Thomashow M, Shiu SH. 2009a. Evolution of stress-regulated gene expression in duplicate genes of Arabidopsis thaliana. *PLoS Genet.* 5:e1000581.

Zou Y, Su Z, Yang J, Zeng Y, Gu X. 2009b. Uncovering genetic regulatory network divergence between duplicate genes using yeast eQTL landscape. *J Exp Zool B Mol Dev Evol.* 312B:722–733.