doi: 10.1111/j.1420-9101.2009.01851.x

Detecting positive selection in the budding yeast genome

Y.-D. LI*†, H. LIANG†[1], Z. GU‡, Z. LIN†, W. GUAN*, L. ZHOU*, Y.-Q. LI* & W.-H. LI†

*College of Life Sciences, Zhejiang University, Hangzhou, China
†Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA
‡Division of Nutritional Sciences, Cornell University, Ithaca, NY, USA

Keywords:

expression variation;
gene essentiality;
nucleotide substitution;
positive selection;
promoter evolution;
Saccharomyces cerevisiae.

Abstract

Available abundant genomic data allows us to study the evolution of the yeast genome at a fine scale. In this study, we examined the adaptive evolution of coding and promoter regions in three *Saccharomyces cerevisiae* strains. First, using a maximum-likelihood approach, we identified 76 positively selected genes (PSG) whose coding regions likely have undergone positive selection in the recent past. These genes show significant bias in terms of biological function and they show over-representation of charged amino acids at positively selected sites. Next, using recent data on yeast transcription-start sites to define core-promoter regions, we identified 31 positively selected promoters, and their corresponding genes are significantly enriched in transmembrane transporter function. We found PSG show no correlation with promoter adaption or expression variation, suggesting that positive selection on coding regions and positive selection on promoter regions are not coupled. Together, our study provides insights into the evolution of *S. cerevisiae* strains from different environments.

Introduction

The budding yeast *Saccharomyces cerevisiae* is probably the best-studied eukaryotic organism. Since it was first introduced in fermentation, yeast strains have been evolving under different ecological niches (Querol *et al.*, 2003). Directed evolution experiments demonstrate that yeast strains can quickly adapt to specific environmental conditions. For example, after 450 generations of glucose-limited growth, the predominant cell type in the population could utilize glucose more efficiently (Brown *et al.*, 1998).

So far three *S. cerevisiae* strains, each with a distinct living history, have been sequenced at the genome level. The lab strain S288c was isolated from a rotten fig about 70 years ago, and it has long become a common lab strain (Gu *et al.*, 2005). The pathogenic strain YJM789 (YJM) was isolated from an acquired immune deficiency syndrome patient with *S. cerevisiae* pneumonia in 1989, and it is used as a model organism for fungal infections (Wei *et al.*, 2007). The wild strain RM11-1a (RM) was collected from a California vineyard and introduced into the lab in 1996 (Ronald *et al.*, 2006). The sequence divergence between S288C and RM/YJM is ~0.5–1%, similar to that between human and chimpanzee. Thus, these yeast strains are suitable for studying how positive selection contributes to genetic diversity among closely related genomes.

For a gene, adaptive changes may take place either in the coding region or in the regulatory region. These two types of adaptive changes have different functional consequences: the first one modifies the amino-acid sequence encoded in the gene, while the latter may lead to change at the expression level. For the coding regions, through decades of efforts, many computational methods have been developed to detect the signal of positive selection based on polymorphism or divergence data (Li *et al.*, 2008). Conventionally, positive selection can be inferred from the ratio of nonsynonymous to synonymous substitution rates (K_N/K_S), with $K_N/K_S > 1$ indicating the presence of positive selection. However,

Correspondence: W.-H. Li, Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA, and Y.-Q. Li, College of Life Sciences, Zhejiang University, Hangzhou 310035, China.
Tel.: 1 773 702 3104; fax: 1 773 702 9740; e-mail: wli@uchicago.edu
[1]Present address: Department of Bioinformatics and Computational Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA.

this criterion is often too strict to detect positive selection, because positive selection may operate on a small fraction of sites. Recently, the maximum-likelihood methods for detecting individual positively selected sites have been shown to be more powerful (Wong *et al.*, 2004; Zhang *et al.*, 2005). With these approaches, positively selected genes (PSG) have been identified in primates (Clark *et al.*, 2003; Bakewell *et al.*, 2007), *Drosophila* species (Eyre-Walker, 2006) and different strains of *Escherichia coli* (Petersen *et al.*, 2007).

Since the 1970s, evolution of noncoding regulatory regions has attracted wide interests. Recently, several studies have been carried out to detect positive selection on noncoding regions at the genome level (Pollard *et al.*, 2006; Prabhakar *et al.*, 2006; Haygood *et al.*, 2007; Liang *et al.*, 2008). For example, Prabhakar *et al.* (2006) found an excess of human-specific substitutions in conserved noncoding sequences near neuronal genes. In yeast, the contribution of positive selection on noncoding regulatory sequences to genetic diversity remains poorly studied.

In this study, we used computational approaches to identify genes subjected to positive selection on coding or on promoter regions in the above three strains of *S. cerevisiae*. First, using a recently developed branch-site-likelihood approach, we found that yeast genes with positively selected coding regions show significant bias in terms of biological function. Next, with the available transcription-start-site (TSS) data of the *S. cerevisiae* genome (Nagalakshmi *et al.*, 2008), we focused on core promoters and detected cases of adaptive evolution. Finally, integrating the above results, we showed that positive selection at core promoters and positive selection at coding regions are not coupled in the evolution of the three yeast strains.

Materials and methods

Constructing orthologous gene sets

The genome sequences of *S. cerevisiae* (S288c and RM11-1α), *S. mikatae* and *S. paradoxus* were downloaded from the Broad Institute (Kellis *et al.*, 2003), and the genome of YJM789 was from the Stanford Genome Technology Center (Wei *et al.*, 2007). All open-reading frame (ORF) sequences were translated into amino-acid sequences, and the ORF with unknown amino acids or stop codons were discarded. The orthologous gene sets among S288c, YJM, *S. mikatae* and *S. paradoxus* were obtained by the synteny-based orthology designation (Kellis *et al.*, 2003; Wei *et al.*, 2007). The one-to-one orthologous pairs between S288c and RM were identified by the best-reciprocal hit approach, with an *E*-value cutoff of 10^{-6}, a sequence identity ≥ 40% and ≥ 75% alignable residues (Li *et al.*, 2009). Finally, we obtained ~3300 unambiguously defined orthologous gene sets among the five *Saccharomyces* genomes.

Orthologous proteins were aligned by CLUSTALW (Larkin *et al.*, 2007) with the default parameter settings, and the aligned amino-acid sequences were then reversely translated into nucleotide sequences for further analysis.

We reconstructed *S. cerevisiae* and *S. paradoxus* ancestral sequence by codeml with the *S. cerevisiae* sequence to determine the pattern of substitutions along the *S. cerevisiae* lineage (Berglund *et al.*, 2009). Weak base pairs (W) are A or T base pairs, and strong base pairs (S) are G or C base pairs. We defined the W → S bias of a gene as follows: W → S bias = number of W → S/(number of W → S + number of S → W).

Detecting positive selection on coding regions

To detect positive selection on protein-coding genes, we used a branch-site test implemented in the PAML program (Yang, 2007), which has been shown to be more robust than other tests (Zhang *et al.*, 2005). The branch-site models permit the ω ratio to vary both among codon sites and among lineages. In the test, positive selection in the 'foreground' lineage of interest is allowed, and the 'background' lineages are assumed to evolve without positive selection. We assessed positive selection in two types of phylogeny (Fig. 1a,b). The first one included the three *S. cerevisiae* strains and *S. paradoxus*, where each strain lineage was designated as the foreground branch and the others as background branches respectively. The input phylogeny was constructed by the neighbour-joining method based on the concatenated nucleotide sequences of ~3300 orthologous genes (Rokas *et al.*, 2003), which is consistent with the phylogenetic relationships reported in previous studies (Gu *et al.*, 2005; Ronald *et al.*, 2006). In the second analysis, each *S. cerevisiae* strain was independently used as the

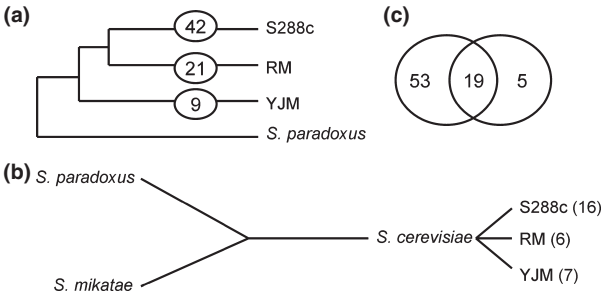

Fig. 1 The positively selected genes (PSG) in three *Saccharomyces cerevisiae* strains by the branch-site likelihood ratio method. (a) The numbers of PSG were identified in the phylogeny including the three *S. cerevisiae* strains and *S. paradoxus* (out group), as shown in circle. (b) The numbers of PSG were identified in the phylogeny where each *S. cerevisiae* strain was independently compared with *S. paradoxus* and *S. mikatae*, as shown in bracket. (c) The overlaps between two PSG sets identified.

foreground branch with *S. paradoxus* and *S. mikatae* as the background branches. The first analysis was sensitive for positive selection in the recent history of *S. cerevisiae* strains, while the second analysis can also detect adaptive evolution since the divergence of *Saccharomyces* species from the other two yeast species. PSG were identified by comparing the alternative model (model = 2, NSsites = 2, fix_omega = 0) with the null model (model = 2, NSsites = 2, fix_omega = 1, omega = 1). Each analysis was carried out several times with different initial values to ensure the global maximum value.

Considering the multiple-testing effect, $P = 0.001$ or 0.0005 was chosen as a cutoff, corresponding to a false discovery rate (FDR) of 20% for the two analyses respectively. For each PSG set, amino-acid residues with a posterior probability of >95% were identified as positively selected sites. The frequencies of positively selected amino acids were compared with the background amino-acid composition (the amino-acid composition of all proteins) in our dataset by the binomial tests.

Constructing the core-promoter alignments

The TSS coordinates in the *S. cerevisiae* (S288c) genome was obtained from Nagalakshmi *et al.* (2008). As transcription factor binding sites are mainly located at −240 and + 60 bp of TSS (Venters & Pugh, 2009), we extracted −240 bp upstream and + 60 bp downstream of the TSS and defined the ∼300 bp as the core-promoter region (the regions overlapping with any known coding regions were excluded). The core-promoter sequences of YJM and RM strains were obtained by the BLAT search (Kent, 2002), and were confirmed with the location relative to their coding genes. The core-promoter sequences of *S. paradoxus* were directly extracted from the UCSC multiple genome alignments (http://hgdownload.cse.ucsc.edu/). Gaps were removed from these alignments. Together, there were 4257 core promoters in our analysis.

Detecting positive selection on core promoters

To choose a good local neutral reference for core-promoter regions, we compared the divergence rates (%) of pseudogenes, introns and four-fold degenerate sites between *S. cerevisiae* (S288c) and *S. paradoxus*; the divergence rate was defined as the percentage of non-conserved nucleotides between the two species. Pseudogene annotations were downloaded from http://www.pseudogene.org (Karro *et al.*, 2007), and introns were extracted from Zhang *et al.* (2007). Four-fold degenerate sites were parsed from the coding-region alignments, and only genes with more than 20 sites were included in our analysis. We found that among the above regions, four-fold degenerate sites evolved at the highest rate (Table 3). So we used them as the (near) neutral control.

To infer positive selection on core promoters in *S. cerevisiae*, core promoters in each strain were compared with the orthologous sequences in *S. paradoxus*. A core promoter was inferred as potentially under positive selection, if (i) its substitution rate (K_p) was higher than the average substitution rate of four-fold degenerate sites (K_4) in the whole dataset and (ii) it evolved significantly faster than the four-fold sites in the same gene, based on Fisher's exact test (one-tailed) on the proportions of conserved sites in the two types of regions. At FDR = 20%, $P < 0.005$ was chosen as a cutoff.

In the above method, as the substitution rate is the average between *S. cerevisiae* and *S. paradoxus*, the detected signals mainly reflect the evolution before the divergence of *S. cerevisiae* strains. To detect positively selected promoters among the strains, we used the phylogeny in Fig. 1a and baseml in PAML to calculate the substitution rates of core promoters (K_p) and four-fold degenerate sites (K_4) with the REV model in the three stains respectively (Yang, 2007). Then we used the substitution rate ratio (K_p/K_4) as a rough index to classify core promoters that potentially underwent positive selection.

Functional analysis of positively selected genes

The gene ontology (GO) annotation of yeast genes was extracted from the SGD database (Hong *et al.*, 2008); and the GO term analysis was carried out by BiNGO (Maere *et al.*, 2005). Overrepresented GO terms for PSG were identified at FDR = 20%.

Fitness data of single-gene deletion in S288c were obtained from Gu *et al.* (2003). The relationship between PSG in S288c and essential genes was evaluated by Fisher's exact test.

The expression data of strains S288c and RM were obtained from Brem *et al.* (2002), and genes with differential expression were defined by a given *P*-value of Wilcoxon–Mann–Whitney test. At various *P*-value cutoffs, whether PSG showed differential expression was tested by the chi-squared test. The expression level of PSG was also compared with non-PSG by the Wilcoxon–Mann–Whitney test, in terms of both average and maximal values. All the statistical analyses were performed in R (http://www.r-project.org).

Results

Positive selection on coding regions in *Saccharomyces cerevisiae*

Using the branch-site-likelihood approach, we identified 72 and 24 genes whose coding regions have potentially undergone adaptive evolution through comparative analysis among strains or among species respectively ($P = 0.001$ or 0.0005, FDR = 20%). As shown in

Fig. 1a,b, there are 42 PSG in S288c, 21 in RM and nine in YJM through the strain-based comparison; and there are 16 PSG in S288c, 6 in RM and 7 in YJM through the species-based comparison (Table S1). For analysis based on species comparison, two genes (YIR390W-A and YPR124) were identified as PSG in all three strains. Remarkably, these two PSG sets overlap extensively: more than 80% of PSG in the species-based comparison are also included in the strain-based set (Fig. 1c), highlighting the robustness of our approach. Therefore, we combined these two PSG sets in the subsequent analyses. Moreover, our results clearly show that there are more PSG in the S288c lineage than in the RM and YJM lineages (Fig. 1).

In the analysis, we used the tree of genome-wide alignments as the underlying phylogeny for positive selection inference. Thus, recombination among the strains may inflate false positives (Anisimova *et al.*, 2003). We obtained the same results for these PSG when individual gene trees were employed, indicating that intragenic recombination did not significantly influence our results. Moreover, the accelerated evolution may be caused by nucleotide substitution bias other than positive selection in human (Berglund *et al.*, 2009). In our yeast analysis, we found that the evolution of PSG is not affected by nucleotide substitution bias. As shown in Table 1, the evolutionary rate of PSG is significantly greater than non-PSG, but the weak (AT) to strong (GC) bias is similar in four gene categories.

Functional bias of genes with positively selected coding regions

To characterize the biological relevance of PSG, we performed the GO term analysis on PSG relative to the whole gene set we used. As shown in Table 2, PSG are significantly enriched in some categories: cell wall and membrane are over-represented in cellular component; metal ion transport in biological process; and transmembrane transporter activity in molecular function.

With fitness data of single-gene deletion in S288c, we found that positive selection tends to operate on nonessential genes. However, the proportion of essential genes among PSG is not significantly lower than that in non-PSG (0.15 vs. 0.21, $P = 0.23$, Fisher's exact test).

In addition, through analysing the amino-acid distribution at the positively selected sites, we found that charged amino acids are highly over-represented, especially for Glu (binomial test, $P = 3 \times 10^{-10}$). Consistently, hydrophobic amino acids, such as Tyr and Leu, are significantly under-represented at positively selected sites (binomial test, $P < 0.001$; Table S2).

To test whether PSG are associated with expression variation, we incorporated the expression data of strains

Table 1 Patterns of nucleotide substitution in genes on the *Saccharomyce cerevisiae* (S288c) lineage.

Category	No. genes	Average length (bp)	No. substitutions (per gene)	S → W	W → S	W → S bias*	K_N/K_S*
PSG	61	1437	113	51	48	0.468	0.119
Non-PSG	2704	1147	73	33	31	0.476	0.087
Non-PSG (fastest†)	100	861	66.5	28	27.5	0.492	0.372
Non-PSG (slowest†)	100	546	22	9	10	0.5	0.0001

*The statistical difference between positively selected gene (PSG) and non-PSG determined by Wilcox rank sum test: $P < 0.0018$ (K_N/K_S), $P = 0.424$ (W → S bias).
†The fastest and slowest in terms of evolutionary ratio (K_N/K_S); 'Weak' (W) designates A or T base pairs, and 'Strong' (S) designates G or C base pairs.

Table 2 Functional bias of positively selected genes identified in the coding regions and promoter regions respectively.

GO term	Categories	No. genes†	No. PSG†	*P*-value*
Coding regions				
Cell wall	Cellular component	37	6	0.0002
Metal ion transport	Biological process	26	4	0.0037
Metal ion transmembrane transporter activity	Molecular function	13	3	0.004
Cellular bud membrane	Cellular component	4	2	0.0042
Promoter regions				
Transmembrane transporter activity	Biological process	7	5	0.0071

The gene ontology (GO) term categories are separated by coding and promoter regions and ranked according to the *P*-values.
*The raw *P*-values are shown, calculated by the binomial test in BiNGO. Significantly overrepresented GO terms were identified at a false discovery rate = 20%.
†For coding region, the total number of genes is 3222 and the total number of PSG is 75; and for promoter region, the total number of genes is 122 and the total number of PSG is 31.

S288c and RM. At various cutoffs of defining genes with differential expression, we found no correlation between PSG and gene expression differentiation (37 of 76 PSG are differentially expressed in S288c/RM). Moreover, PSG do not show higher expression levels than non-PSG.

Positive selection on core promoters in *Saccharomyces cerevisiae*

Unlike coding regions, where all the sites can be classified into synonymous and nonsynonymous for comparison, a key factor for accurately inferring positive selection in noncoding regions is to employ good local neutral regions as a reference. For this purpose, we compared the divergence rates of pseudogenes, introns and four-fold degenerate sites. As shown in Table 3, the divergence rate of four-fold degenerate sites is substantially higher than that of pseudogenes or introns. Thus, we chose four-fold sites as the neutral control.

To identify positively selected promoters, we first required K_p to be higher than the average K_4 in the whole dataset, thus reducing false positives because of functional constraints at four-fold degenerate sites (Fig. 2a). We then required that the promoter should evolve significantly faster than the four-fold sites in the same gene (Table 3). As shown in Fig. 2b, we observed many more genes with a fast evolving promoter (with very low *P*-values) than random expectation. At FDR = 20% and $P < 0.005$, we identified 31 promoters as positively selected. For these promoters, the GO term analysis showed that their corresponding genes are significantly enriched in transmembrane transporter function, which is similar to the functional bias of coding-region PSG. We also found that these genes show no statistically significant difference from the other genes in the genome in terms of both fitness effect and gene expression.

Relationships between the evolution of promoter and coding regions

Integrating genes with positively selected coding regions and core promoters, we found no significant overlap between the two gene sets and only one gene YJR030C is shared by both gene sets. Because the positively selected core promoters we identified mainly capture the adaptive signal since the divergence of the *S. cerevisiae* species, as a complementary test, we also calculated K_p in each *S. cerevisiae* strain after diverging from the common ancestor of the three strains. Using the four-fold degenerate sites in the corresponding coding regions as a nearly neutral control, we classified genes into two groups ($K_p > K_4$ and $K_p \leq K_4$) and test if genes in the first group are more likely to undergo adaptive evolution at their coding regions. Again, we found no difference in the two groups in terms of the proportion of coding-region PSG.

Meanwhile, we found that the proportion of $K_p > K_4$ in the lab strain S288c is substantially higher than those in YJM and RM (Table 4), suggesting that positive selection at the regulatory sequences has been more prevailing in the S288c lineage. This finding parallels the trend observed for adaption at the coding regions.

Table 3 Summary statistic of evolutionary rate analysis.

DNA type	No. genes	Median *D**	Mean *D* (SD)
Pseudogenes	133	0.267	0.273 (0.0124)
Introns	262	0.231	0.224 (0.0041)
Four-fold degenerate sites	3262	0.286	0.289 (0.0011)
Core promoters	4452	0.163	0.17 (0.0063)

**D*, divergence rate between *Saccharomyce cerevisiae* (S288c) and *S. paradoxus*.

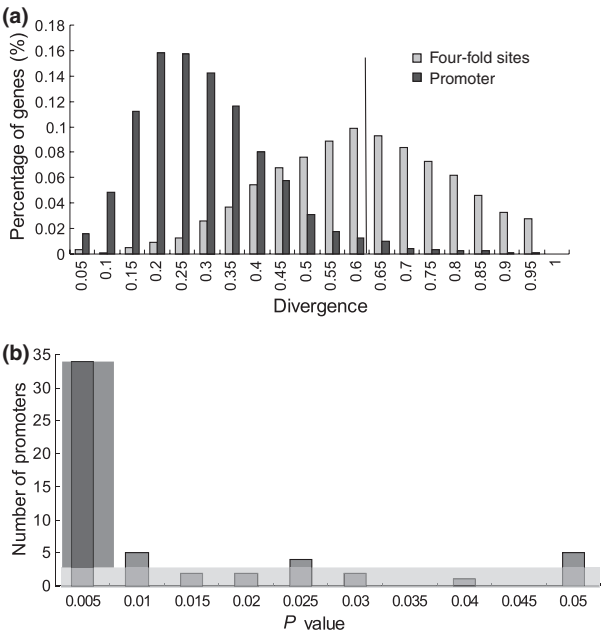

Fig. 2 The positively selected core promoters inferred based on the comparison of *Saccharomyces cerevisiae* and *S. paradoxus*. (a) The distributions of K_p and K_4. The mean K_4 is indicated as a vertical line. Only 122 core promoters with K_p greater than mean K_4 are included in the next step. (b) The observed distribution of the core promoters that evolved significantly faster than neutral controls. The light grey square represents the uniform distribution expected from chance alone. The grey square represents positively selected promoters identified in our study [$P < 0.005$, false discovery rate (FDR) = 20%].

Discussion

In this study, we focused on detecting positive selection in three *S. cerevisiae* strains from distinct ecological niches. Using a branch-site likelihood approach, we first identified yeast genes whose coding regions may contain

Table 4 Evolution of core promoters in three *Saccharomyces cerevisiae* strains.

Strains	Mean K_p*	Mean K_4	No. genes with $K_p > K_4$†	Prop. of genes with $K_p > K_4$ (%)
S288c	0.00182 (0.00166–0.00201)	0.00406 (0.00354–0.00467)	686	26.8
YJM	0.00172 (0.00153–0.00193)	0.0021 (0.0019–0.00231)	456	17.8
RM	0.00127 (0.00113–0.00142)	0.00247 (0.00217–0.00283)	523	20.4

*The 95% confidence intervals were calculated from 1000 bootstrap sampling. K_p is significantly different between S288c and RM (Wilcoxon rank sum test, $P < 0.005$).
†K_p, the substitution rate of core promoter; K_4, the substitution rate of four-fold degenerate sites. Both are calculated on each of the three strain lineages since the divergence of their common ancestor with PAML.

adaptive changes as the divergence from the *S. paradoxus* lineage. These PSG show significant functional bias. In particular, genes related to cell wall, membrane and transmembrane transport activity are greatly enriched. We speculate that because the function of these genes is more directly affected by the external environment, they have a better chance to evolve under positive selection.

As for positively selected sites, we found that they are different from the background amino-acid composition: negatively charged amino acids (e.g. Glu) are greatly over-represented; while hydrophobic amino acids (e.g. Leu and Tyr) are under-represented. This may be explained by the functional role of different amino acids. Negatively charged residues tend to be functional resides on the surface of a protein and directly facilitate the interactions among partners (Verma *et al.*, 2006). Hydrophobic amino acids are often placed within a protein structure as supporting residues, so there are relatively fewer chances for changes at these sites to improve the protein function.

Utilizing the recent TSS data and a relatively strict method, we identified 31 genes with positively selected core promoters that are enriched in transmembrane transporter function. Interestingly, the functional bias of PSG in coding region and promoter region both is related to environmental interaction (although the number of genes involved is small). This result is consistent with the notion that positive selection mainly occurs in response to changes in environment. For the set of promoter PSG, there is no significant overlap with the PSG of coding regions. The gene YJR030C is detected both in promoter and coding region, and its function remains unknown. However, since it is expressed in carbon limited or depleted chemostat culture (http://www.yeastgenome.org), this gene may be important for yeast cells to adapt to the carbon stress. While the number of positively selected promoters is too small to draw a definite conclusion, our results suggest that adaption at promoter regions and coding regions are largely independent in these yeast strains, which is consistent with a previous study (Fay & Benavides, 2005). Moreover, genes with a positively selected promoter and coding region show no bias in terms of expression.

Finally, compared with the RM wild strain the evolutionary rate is elevated in both promoter and four-fold degenerate sites in the lab strain. This can be mainly explained by relaxed selective constraint because of the reduced effective population size in lab strain S288c (Gu *et al.*, 2005). Meanwhile, we found more adaptive signals in coding regions as well as a higher proportion of genes with $K_p > K_4$ in S288c strain, which also implies more adaptive changes in the lab strain.

Through the effort of decades, significant progress has been made about detecting the signal of adaptive evolution at the molecular level. However, the statistical power of different methods varies greatly, and some of the inferred candidate genes are prone to false positives (Wong *et al.*, 2004; Li *et al.*, 2008). Doniger *et al.* (2008) found scant evidence of positive selection in yeast by the McDonald-Kreitman test, which may be a result of less sensitivity for individual positively selected sites in the method (Li *et al.*, 2008). Furthermore, the performance of detecting methods is affected by the sequence quality and the number of sequences used in comparative analyses. With recent advance of next-generation sequencing techniques, many more genomes in the same species are expected to be available soon (Doniger *et al.*, 2008), which would provide unprecedented opportunities to understand how adaptive molecular changes makes an organism more fit in its own ecological context. More importantly, experimental studies are greatly needed to examine the functional consequence of inferred adaptive changes.

Acknowledgments

The authors thank Dr Rachel Brem for providing yeast gene expression data. This study was supported by NIH grants (GM30998 and GM081724) to W.-H. Li, and grants from the RPST of Zhejiang Province, China (2005C23027) to Y.Q. Li. Y.D. Li was also supported by a fellowship from the government of China.

References

Anisimova, M., Nielsen, R. & Yang, Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164**: 1229–1236.

Bakewell, M.A., Shi, P. & Zhang, J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc. Natl Acad. Sci. USA* **104**: 7489–7494.

Berglund, J., Pollard, K.S. & Webster, M.T. 2009. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol.* **7**: e26.

Brem, R.B., Yvert, G., Clinton, R. & Kruglyak, L. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752–755.

Brown, C.J., Todd, K.M. & Rosenzweig, R.F. 1998. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Mol. Biol. Evol.* **15**: 931–942.

Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D., Kejariwal, A., Todd, M.A., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B., Ferriera, S., Wang, G., Zheng, X., White, T.J., Sninsky, J.J., Adams, M.D. & Cargill, M. 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**: 1960–1963.

Doniger, S.W., Kim, H.S., Swain, D., Corcuera, D., Williams, M., Yang, S.P. & Fay, J.C. 2008. A catalog of neutral and deleterious polymorphism in yeast. *PloS Genet.* **4**: e1000183.

Eyre-Walker, A. 2006. The genomic rate of adaptive evolution. *Trends Ecol. Evol.* **21**: 569–575.

Fay, J.C. & Benavides, J.A. 2005. Hypervariable noncoding sequences in *Saccharomyces cerevisiae*. *Genetics* **170**: 1575–1587.

Gu, Z., Steinmetz, L.M., Gu, X., Scharfe, C., Davis, R.W. & Li, W.H. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**: 63–66.

Gu, Z., David, L., Petrov, D., Jones, T., Davis, R.W. & Steinmetz, L.M. 2005. Elevated evolutionary rates in the laboratory strain of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **102**: 1092–1097.

Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K.D. & Wray, G.A. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat. Genet.* **39**: 1140–1144.

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

Karro, J.E., Yan, Y., Zheng, D., Zhang, Z., Carriero, N., Cayting, P., Harrrison, P. & Gerstein, M. 2007. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.* **35**: D55–D60.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241–254.

Kent, W.J. 2002. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**: 656–664.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. & Higgins, D.G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**: 2947–2948.

Li, Y.F., Costello, J.C., Holloway, A.K. & Hahn, M.W. 2008. ''Reverse ecology'' and the power of populaton genomics. *Evolution* **62**: 2984–2994.

Li, Y.D., Xie, Z.Y., Du, Y.L., Zhou, Z., Mao, X.M., Lv, L.X. & Li, Y.Q. 2009. The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. *Gene* **436**: 8–11.

Liang, H., Lin, Y.S. & Li, W.H. 2008. Fast evolution of core promoters in primate genomes. *Mol. Biol. Evol.* **25**: 1239–1244.

Maere, S., Heymans, K. & Kuiper, M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**: 3448–3449.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. & Snyder, M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–1349.

Petersen, L., Bollback, J.P., Dimmic, M., Hubisz, M. & Nielsen, R. 2007. Genes under positive selection in *Escherichia coli*. *Genome Res.* **17**: 1336–1343.

Pollard, K.S., Salama, S.R., King, B., Kern, A.D., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J.S., Bejerano, G., Baertsch, R., Rosenbloom, K.R., Kent, J. & Haussler, D. 2006. Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* **2**: e168.

Prabhakar, S., Noonan, J.P., Paabo, S. & Rubin, E.M. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**: 786.

Querol, A., Fernandez-Espinar, M.T., del Olmo, M. & Barrio, E. 2003. Adaptive evolution of wine yeast. *Int. J. Food Microbiol.* **86**: 3–10.

Rokas, A., Williams, B.L., King, N. & Carroll, S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**: 798–804.

Ronald, J., Tang, H. & Brem, R.B. 2006. Genomewide evolutionary rates in laboratory and wild yeast. *Genetics* **174**: 541–544.

Venters, B.J. & Pugh, B.F. 2009. A canonical promoter organization of the transcription machinery and its regulators in the *Saccharomyces* genome. *Genome Res.* **19**: 360–371.

Verma, S., Bednar, V., Blount, A. & Hogue, B.G. 2006. Identification of functionally important negatively charged residues in the carboxy end of mouse hepatitis coronavirus A59 nucleocapsid protein. *J. Virol.* **80**: 4344–4355.

Wei, W., McCusker, J.H., Hyman, R.W., Jones, T., Ning, Y., Cao, Z., Gu, Z., Bruno, D., Miranda, M., Nguyen, M., Wilhelmy, J., Komp, C., Tamse, R., Wang, X., Jia, P., Luedi, P., Oefner, P.J., David, L., Dietrich, F.S., Li, Y., Davis, R.W. & Steinmetz, L.M. 2007. Genome sequencing and comparative analysis of *Saccharomyces cerevisiae* strain YJM789. *Proc. Natl Acad. Sci. USA* **104**: 12825–12830.

Wong, W.S., Yang, Z., Goldman, N. & Nielsen, R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168**: 1041–1051.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**: 1586–1591.

Zhang, J., Nielsen, R. & Yang, Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**: 2472–2479.

Zhang, Z., Hesselberth, J.R. & Fields, S. 2007. Genome-wide identification of spliced introns using a tiling microarray. *Genome Res.* **17**: 503–509.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Table S1 List of genes with positive selection on coding region.

Table S2 The amino acid bias at positively selected sites.

Table S3 List of genes with a positively selected core promoter.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Received 22 June 2009; revised 4 September 2009; accepted 14 September 2009