Diversification at transcription factor binding sites within a species and the implications for environmental adaptation

Ryan M Ames,* and Simon C Lovell*

*Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester, M13 9PL, UK

Running head: Evolution of transcription factor binding sites

Keywords: gene duplication, evolutionary adaptation, gene regulation, transcription factor binding sites

Corresponding author: simon.lovell@manchester.ac.uk

Abstract

Evolution of new cellular functions can be achieved both by changes in protein coding sequences and by alteration of expression patterns. Variation of expression may lead to changes in cellular function with relatively little change in genomic sequence. We therefore hypothesize that one of the first signals of functional divergence should be evolution of transcription factor binding sites. This adaptation should be detectable as substantial variation

in the transcription factor binding sites of alleles.

New data sets allow the first analyses of intra-species variation from large number of whole-genome sequences. Using data from the *Saccharomyces* Genome Resequencing Project we have analyzed variation in transcription factor binding sites. We find a large degree of variation both between these closely-related strains and between pairs of duplicated genes. There is a correlation between changes in promoter regions and changes in coding sequences, indicating a coupling of changes in expression and function. We show that (i) the types genes with diverged promoters vary between strains from different environments and (ii) that patterns of divergence in promoters consistent with positive selection are detectable in alleles between strains and on duplicate promoters. This variation is likely to reflect adaptation to each strain's natural environment.

We conclude that, even within a species we detect signs of selection acting on promoter regions which may act to alter expression patterns. These changes may indicate functional innovation in multiple genes and across the whole genome. Change in function could represent adaptation to the environment and be a precursor to speciation.

Introduction

Biological function is dependent not only on the function of individual molecules but also the location and timing of their expression. Not all observed phys-

iological differences can be explained by the differences in protein coding regions (Levine and Tjian, 2003); rather, differences in gene expression may lead to large differences in phenotype. Thus evolution of cellular function is dependent not only on change of specificity in molecules coded for by diverged genes, but also on evolution of the expression of these molecules. Changes in expression will be a reflection of alteration of regulatory sequences. Identification of selection acting on regulatory sequences would therefore provide evidence that evolution of gene regulation is a key agent in adaptation to the environment (Carroll, 2005).

Evolution of cellular function within a species can be associated with divergence of alleles. Such divergence can give rise to different strains with differing phenotypes, potentially reflecting adaptation to specific environments. If indeed adaptation to the environment can be observed through evolution of gene regulation (Carroll, 2005) we expect to identify selection acting on the promoter regions of alleles within a single species. Such evidence would provide a role for gene regulation evolution as a precursor to speciation.

Divergence may also be observed between genes related by gene duplication. When divergence is between paralogs pairs, there are a range of possibilities which are describable by a number of models (Ohno, 1970; Innan and Kondrashov, 2010). Differential loss of recently-duplicated genes is common (Kellis, Birren and Lander, 2004; Ames et al., 2010). Of those duplicates which are retained some may remain unchanged when an increased dosage of a gene product gives a selective advantage (Spofford, 1969; Whit-

ton, 2000). Other alternatives are possible: a retained duplicate gene may acquire a novel function (neofunctionalization), the ancestral gene function may be partitioned between the paralogs (subfunctionalization) (Force et al., 1999; Lynch et al., 2001), or the new copy may degenerate (nonfunctionalization). Thus evolution of regulatory sequences may represent a sign of functional innovation both between alleles and between duplicate gene pairs.

Evolution of regulatory sequences has been observed in individual cases. Hox1b duplicates in zebrafish show that degenerative complimentary mutations of *cis*-regulatory elements may lead to differences in expression profiles that in turn cause subfunctionalization (Prince and Pickett, 2002). When humans and chimps are compared, sequences that regulate duplicated genes are found to be evolving rapidly, potentially leading to functional innovation (Kostka, Hahn and Pollard, 2010).

In yeast, several studies have shown that shared number of transcription factor binding sites (TFBSs) or expression correlation between duplicate genes decreases with the age of duplicates (Gu et al., 2002; Papp, Pál and Hurst, 2003). Interestingly, Papp, Pál and Hurst (2003) also show that while the number of shared TFBSs decreases with duplicate age, the total number of sites in each remains constant. The authors conclude that, in yeast, subfunctionalization alone is not the sole mechanism behind divergence of expression. Yeast species have been shown to frequently gain and lose TFBSs (Doniger and Fay, 2007), indicating that the loss of shared binding sites may be followed by gain of novel sites. Indeed there is evidence of

positive selection acting on the promoter regions of a single duplicate pair. Neofunctionalization may therefore also play a significant role in expression divergence of duplicate genes (Papp, Pál and Hurst, 2003).

Variation between individuals within one species is much less well characterized. Recent analyses in yeast show that within-species differences at the genome level are very common (Liti et al., 2009), including differences in duplicate gene content (Ames et al., 2010). Since both subfunctionalization and neofunctionalization of duplicate gene pairs (Papp, Pál and Hurst, 2003) may arise from differences in expression, we hypothesize that there may be substantial differences between regulatory regions not only between species, but also within a single species. Further we hypothesize that much of this difference will be associated with duplicate genes. Since expression differences can give rise to phenotypic differences, we predict there will be evidence of selection of promoter regions in homologous genes within species. Furthermore if neofunctionalization plays a substantial role in the divergence of expression patterns between duplicate genes we would expect to find within promoter regions of duplicate genes evidence of positive selection.

The sequence data from the *Saccharomyces* Genome Resequencing Project (Liti et al., 2009) gives us the first complete genome sequences for many members of a species where the environment for each of the strains is known. Using these genomic sequences and previously annotated duplicates we have analyzed the differences within strain duplicates and between strain's alleles in TFBSs. We find that the association of a transcription factor to a gene is

highly conserved within these closely related strains but we detect substantial variation between the number of sites for each factor between the strains. We find large variation between duplicate genes within strains with older duplicates showing fewer shared motifs across all strains. Change in TFBSs correlates with change in coding sequence between duplicates indicating a possible coupling of change in expression and function. Most strikingly, the types of genes with diverged or conserved promoter regions varies between strains from different environments. Patterns of divergence in promoter regions consistent with positive selection are detectable between strains and on duplicate promoters. We conclude that widespread genomic variation detectable in promoter regions of the same species shows signs of evolution that is shaped by the environment.

Methods

Genomic sequences

The promoter regions of 38 strains of *S. cerevisiae* were analyzed. The genomes were sequenced by the *Saccharomyces* Genome Resequencing Project (SGRP) (Liti et al., 2009). The parallel-alignment assembling (PALAS) assembled and annotated data were used in this study. These data contain imputed values, meaning that genomic regions with no or little coverage from the sequencing runs are inferred from the reference strain data. The imputed values will mean that some of the strains inherent variation will

be lost, leading us to underestimate the amount of variability between the strains. Open reading frames (ORFs) for each gene were extracted from the genomic sequences using the SGRP identified gene positions. Any ORFS labeled as dubious by the *Saccharomyces* genome database (SGD), containing N values, were less than 48 bases or did not have an initial ATG sequence were removed from the analysis.

Promoter regions were defined as the region 500bp upstream from the transcriptional start site (TSS) of that gene (Lawless et al., 2009), taking into account the gene's orientation. If the previous gene's ORF overlapped with this 500bp, the promoter region was defined as the region of genomic sequence between the TSS of the gene and the end of the previous gene. As with the ORF sequences any promoter regions containing 'N' values were removed from the analysis.

Duplicate pairs were annotated within the data sets using previously determined duplicate genes. Whole genome duplicates were annotated using data from Kellis, Birren and Lander (2004). Small scale duplicates were annotated using data from Hakes et al. (2007). For successful annotation each strain must contain both genes of a duplicate pair.

Transcription factor binding sites

TFBSs were identified using a solely computational method and by using experimentally determined consensus sequences and factor associations. For the computational method, position frequency matrices for 177 transcription

factors were downloaded from the JASPAR database (Sandelin et al., 2004). The matrices were converted to binding motifs using the rules set out by D'haeseleer (2006). The binding motifs were scanned against the promoter regions of each gene using a bespoke Java program. All successful hits were recorded as the set of computationally identified binding sites.

Experimentally determined binding sites were derived from Harbison et al. (2004), which were identified using genome-wide location analysis. We chose a subset of 60 transcription factors which had been assigned a single high confidence binding motif and had been shown to interact with specific promoter regions at high confidence ($P \leq 0.001$). These significant motifs were scanned along the promoter regions of their associated genes in each strain using a bespoke Java program. All successful hits were recorded as the set of experimentally determined binding sites.

Binding site turnover

In order to investigate the amount of variation in binding sites between the strains we analyzed the 5056 genes common to all strains. For each of these genes the transcription factors that bound the associated promoter region were identified and the proportion of strains with sites for the factor was calculated. This analysis only counts whether a transcription factor is associated with a gene and does not take into account the number of sites for that factor. Factor conservation was represented as the proportion of strains that share a particular transcription factor for a given gene, averaged over

all the genes.

In addition to determining a transcription factor's association to a specific gene we also analyzed the variation in the number of sites for each transcription factor. This gives us an idea of the amount of binding site gain and loss in the strains. Both of these analyses were repeated with the experimentally and computationally determined sites.

Promoter region divergence

In order to test whether duplicate promoter regions diverge with time since duplication, we looked at the proportion of shared binding sites between duplicates and the synonymous mutation rate (K_s) between the pair. The K_s values are used as a proxy for time since duplication. K_s values were calculated using yn00 of the PAML package (Yang, 2007). For this analysis only duplicate pairs with $K_s < 1.5$ and an effective number of codons > 30 were used as high synonymous substitution rates are unreliable owing to multiple substitutions and a strong codon bias makes K_s a poor proxy of divergence time (Gu et al., 2002). The effective number of codons was determined using CodonW (ftp://molbiol.ox.ac.uk/cu/codonW.tar.Z). This analysis was only carried out for computationally identified sites as the stringent K_s cutoff leaves too few data to analyze for experimental sites.

Functions of duplicates with conserved or diverged promoter regions

To test whether duplicate pairs with highly diverged promoter regions have different functions than those with more conserved promoter regions we used the Gene Ontology (GO) (Ashburner et al., 2000). Genes were annotated with GO terms using pre-computed annotations downloaded from the SGD. Lists of over-represented GO terms were determined for duplicate pairs sharing <30% of binding sites between duplicate genes as a diverged set and sharing >30% as a conserved promoter region set. Duplicates were considered as a sample from all completely sequenced genes for each strain to account for the varying number of genes between strains. Fisher's exact test was used to calculate raw p-values, which were corrected for multiple testing using the false discovery rate correction of Benjamini and Hochberg (1995). This analysis was repeated for diverged sets with <20% and <40% shared binding sites. Because of the small number of experimentally identified TFBSs in duplicate pairs we were only able to apply this analysis to computationally identified binding sites.

Positive selection within and between strains

To identify promoter regions of alleles and duplicate genes that show patterns of change consistent with positive selection we used three separate methods. For the allele analysis we aligned all identified site regions for computationally

identified (5056 genes) and experimentally identified (490 genes) sites for each promoter region using MUSCLE (Edgar, 2004). We defined this data set as the 'binding site' regions. The process was repeated for these sequences with no computational or experimental identified sites and defined this set as the 'non-binding site' regions. For duplicate pairs we also aligned the site and non site regions of the promoter of duplicate genes between strains. The analysis was carried out on duplicate pairs with computationally predicted sites (524 duplicate pairs) and experimentally determined sites (9 duplicate pairs).

Our first method of detecting signs of positive selection aimed to identify those promoter regions which have a much higher genetic distance between site regions than non site regions, indicating that the site regions are changing at an accelerated rate compared to the non binding site regions. The distances between alleles and duplicates were determined using the Kimura "2-parameter" model (Kimura, 1980) as implemented in the PHYLIP package (Felsenstein, 1989). The distance for one promoter region was defined as the average distance between all strains or duplicates site or non site regions. This method used two cutoffs the first required that the genetic distance between binding site regions of alleles must be greater than that of non site regions. The second cutoff required that the genetic distance of the binding site regions must be greater than three standard deviations from the mean of all alleles site distances. This ensured we only selected those alleles showing high rates of substitution.

Secondly, we looked for evidence of selective sweep in non binding site regions. Selective sweep is thought to occur when a beneficial mutation, under positive selection, spreads through a population and the surrounding neutral polymorphisms hitchhike to fixation, resulting in reduced variation at a locus within a population (Smith and Haigh, 1974). Using the genetic distances between alleles and duplicates we identified those promoters showing reduced genetic distance in non binding regions, which were lower than the mean of all alleles or duplicates. We also ensured the the genetic distance of site regions was higher than the mean distance of all alleles or duplicates. This definition identifies promoters with increased change in site regions but also reduced change in non site regions which may be evidence for selective sweep.

Finally, we used maximum likelihood to determine the rates of change in site and non site regions first assuming the sites were evolving at the same rate as non site regions then assuming the regions were evolving at different rates. We performed this analysis using PAML (Yang, 2007) and the HKY model of substitution. The tree used for the allele analysis was taken from Liti et al. (2009) and for duplicate analysis trees were generated using RAxML (Stamatakis, Ludwig and Meier, 2005) from the duplicate gene coding regions using a general time reversible model with gamma distribution (GTRGAMMA). We defined promoter regions that are potentially experiencing positive selection as those showing high rate of change in the site regions and showing a significantly better likelihood scores for inference where we

test whether the regions evolve at the same rate or at different rates. Significance was determined by comparing the support for two rates, defined as:

$$2 \times \{lnL(a) - lnL(b)\} \quad (1)$$

where *lnL(a)* and *lnL(b)* are the likelihood scores assuming the site and non site regions are evolving at different rates and the same rate, respectively. P values were calculated using a χ^2 distribution with one degree of freedom and all p values were corrected for multiple testing using the false discovery rate of Benjamini and Hochberg (1995). We used a corrected p value cutoff of <0.05 to identify promoter regions potentially under positive selection.

For alleles or duplicates which show patterns of change consistent with positive selection, we labelled those promoters which are identified by only one of the above methods as low confidence positively selected promoters. Those promoters identified by two or all three methods were termed medium and high confidence positively selected promoters respectively.

Results

Identifying binding sites

Genomic sequences of *S. cerevisiae* strains assembled by the PALAS method (Liti et al., 2009) are almost complete (Table 1). However, it should be

noted that this data contains imputed sequence, i.e. sequence derived from the reference strain, leading to an under-estimate of the variation between strains. Our conclusions from the data are, therefore, conservative.

Transcription factor binding sites were identified using two methods. The first is a solely computational approach. Consensus motifs for 177 transcription factors from the JASPAR database (Sandelin et al., 2004) were used. The promoter regions of each gene were scanned for each consensus motif and all hits were recorded as binding sites. This method is likely to give an overestimate of the number of binding sites since not all transcription factors will function for every gene.

The second method uses experimentally determined transcription factor consensus sequences and the experimentally determined targets for these factors deemed significant by Harbison et al. (2004). This data set includes 60 consensus sequences of transcription factors that interact with a total of 1974 yeast genes. Here the consensus motifs are applied to the promoter regions of the target genes in order to identify the binding sites. If the consensus motif is not found the binding site is assumed to have been lost. This data set is only expected to find one third of real binding sites and has a false positive rate of 6-10% (Harbison et al., 2004) meaning that this data set is likely to be an underestimate of the number of binding sites.

There are many more genes annotated with computationally predicted sites than the experimentally determined sites (Figure 1). While the promoter regions of the majority of genes have no experimentally identified sites

we find that the vast majority of promoter regions are assigned TFBSs using the computational approach, in a few cases more than 25 transcription factors assigned to a promoter region. There is substantial overlap between the sites identified by both methods and see that 34% of experimentally sites are also detected by the computational method.

Transcription factor binding sites are highly conserved between strains

We determined the degree of conservation of transcription factors interacting with a specific gene (regardless of the number of sites for that factor) across the strains. In general, transcription factor association is highly conserved between strains with the vast majority of genes having the same set of transcription factors in all strains (Figure 2).

In order to identify variation in the number of TFBSs across all strains we compared the number of identified sites for each factor across all strains. For computationally identified TFBSs there is an average of 363±9.4 sites per factor across all the genomes. This analysis was repeated for the experimentally determined sites (Harbison et al., 2004), where there is an average of 15.0±0.7 sites per factor across all the genomes. Here we see a larger proportional variation in experimentally identified sites when compared to the computationally identified sites, which seems surprising given that the computationally identified sites are expected to be an over-estimate. These

results show that while a transcription factors association to a gene is well conserved across all strains the number of sites for these factors has changed, indicating some divergence of TFBSs between the strains.

Selection acts on alleles between strains

In order to determine whether any of these alleles are experiencing selection we examined the genetic distances and evolutionary rates of aligned site regions and non site regions for all alleles. We find that the genetic distances between promoter regions of strains is different in regions containing computationally identified binding sites and regions containing no sites. Overall the mean distance is lower in site regions than non site regions, 0.0046 and 0.0062, respectively (Wilcoxon test, $P<2.2x10^{-16}$). However, there is significantly greater variation of distances in site regions than non site regions as measured by the standard deviation, 0.012 and 0.008, respectively (Levene test, $P=6x10^{-8}$). We find the majority of alleles show no difference in site regions between strains and almost 5% of alleles show show greater average distance in the site regions than the non site regions (Figure 3). The results are the same when we analyze experimentally identified binding sites. Here the mean distance between site regions and non site regions is 0.0012 and 0.006, respectively (Wilcoxon paired test, $P<2.2x10^{-16}$). Additionally, site regions have significantly more variation than non site regions with standard deviations of 0.0056 and 0.0035, respectively (Levene test, $P=7x10^{-7}$).

If the regulation of an allele is under strong selective constraint to be

maintained we might expect stabilizing selection to be indicated by site regions showing less divergence than non site regions. We defined the promoter region of an allele to be under stabilizing selection if the genetic distance between the site regions from all the strains was lower than that of non site regions. Here we find 3934 (77%) alleles with computationally identified sites whose promoter regions are under stabilizing selection. We also see evidence for stabilizing selection in the promoter regions of 467 (95%) alleles with experimentally identified sites.

We also examined those alleles with greater divergence in binding site regions than non-binding site regions. These alleles show patterns of change consistent with positive selection. We used three separate methods to identify those alleles that appear to be experiencing positive selection and categorized them as low, medium and high confidence (Figure 4). Using these definitions we identify 348 alleles that show signs of positive selection with low confidence, 84 with medium confidence and 13 with high confidence (Additional Table 1). When we examine promoter regions of alleles with experimentally identified sites we find 99 regions potentially under positive selection with low confidence, 4 with medium confidence and 2 with high confidence (Additional Table 2).

Divergence of promoter regions in duplicate pairs

In addition to allelic divergence between strains we would also expect to see signs of evolution in promoter regions within strains. After gene duplica-

tion we assume that the expression pattern and thus, promoter regions, are identical and therefore duplicate pairs offer a method of detecting adaptive evolution within a strain. If duplicate genes are acquiring new functions through either neo- or sub-functionalization, we expect them to show some type of divergence over time. If this functional divergence is due to differences in expression there will be a decrease in the proportion of shared binding sites between duplicate genes as they diverge. Using the synonymous mutation rate (K_s) as a proxy for duplicate age we find that older duplicates share fewer computationally predicted binding sites than younger duplicates (Figure 5) and that this relationship is significant (DF=2079, R=0.363, P<2.2 x10^{-16}). This analysis could not be applied to experimentally determined sites as the sparse annotation of experimentally identified sites leaves too few data to analyze.

Divergence of promoter regions is correlated with divergence in coding sequence

We analyzed synonymous (K_s) and non synonymous (K_a) change in the coding regions of the 524 common duplicate pairs with computationally identified sites (Figure 6). The same cutoffs were used as before to select only those pairs with reliable estimates of K_a and K_s. We again used the Kimura "2-parameter" model (Kimura, 1980) to determine the genetic distance between aligned site and non site regions of duplicate pairs. Both the distance between

site and non site regions is significantly positively correlated with K_s (binding sites: R=0.32, P<2.2x10^{-16}, non-binding site: R=0.33, P<2.2x10^{-16}).

Synonymous mutations accumulate in the coding sequence; these are assumed to correlate with time since duplication. Similarly, mutations accumulate in non-binding site regions of promoters, which we expect to be evolving neutrally. In addition mutations also accumulate in the binding site regions of duplicate promoters with increased K_s; these changes may be linked to some change in function or may be mutations of the synonymous sites found in transcription factor binding sites.

Interestingly, there is a different trend in binding site and non-binding site regions when the relationship between K_a and the genetic distance between duplicate promoters is analyzed (Figure 6B). Here there is no correlation between K_a and the distance between non site regions of duplicate promoters (R=0.01, P=0.28) but a significant positive correlation with the site regions of these promoters (R=0.15, P=6.2x10^{-9}). This result suggests that those synonymous changes that might lead to functional change in the coding sequence are correlated with changes in the binding site regions of the corresponding promoter. These changes may lead to an altered expression pattern. This analysis could not be applied to experimentally identified sites as there are too few data points.

Duplicates with diverged promoter regions have different functions from those with conserved promoter regions

We have shown that there is a correlation between K_s and the proportion of shared TFBSs between duplicates and that change in the site regions of promoters correlates with K_a. Next we ask whether certain types of genes are more likely to have diverged promoter regions. We find that duplicate genes with diverged promoter regions (sharing <30% of computationally identified binding sites) have different functions compared with those with more conserved promoter regions (Table 2 and Additional Tables 3 & 4). Those duplicates with conserved promoters have functions involved in the growth of the organism, lipid metabolism and signal transduction. Those duplicates with diverged promoters show a more varied range of functions, including response to varied stimuli and transport and metabolism of sugars.

Interestingly, there is a large amount of variation in over-represented Gene Ontology (GO) (Ashburner et al., 2000) terms between the individual strains. In the conserved set there are a total of 228 unique terms over-represented across all strains, 32 of these are over-represented in a single strain, 154 in more than one strain but not all and 42 in all strains. In the diverged set we see 254 unique over-represented terms across all strains, 54 in a single strain, 146 in multiple strains and 54 in all strains. Of the 42 and 54 terms over-represented in all strains for the conserved and diverged sets respectively,

9 are shared between the two sets indicating that the same types of duplicates may experience divergence of promoter regions in one environment but may be conserved in another environment. Changing the "diverged" and "conserved" cutoffs to 20% or 40% makes little difference to these results (data not shown). The over-represented terms for each strain can be found in Additional File 1. This analysis was only carried out for computationally identified sites because the sparse annotation of experimental sites in duplicate promoters gives a maximum of only 9 duplicate pairs to analyze in the reference strain.

This analysis has revealed that specific types of duplicate genes are more likely to have divergent promoter regions and so might be more likely to diverge in expression pattern. Coupled with the large number of duplicates that share few binding sites this suggests a role for both sub- and neo-functionalisation in the evolution of promoter regions. Interestingly, the data also suggests that the gain and loss of TFBSs is proceeding differently between these strains since there are apparent differences in enrichment between strains. From this observation we hypothesize that differential gain and loss of TFBSs may be an indicator of environmental adaptation, which could be detected as positive selection acting on binding site regions.

Duplicate genes with highly conserved promoter regions maintain similar functions

By identifying duplicates with conserved promoter regions we can identify those duplicates with promoter regions under stabilizing selection. This analysis was only performed on duplicate pairs with computationally predicted binding sites as the experimentally identified site data set only includes 9 duplicate pairs common to all strains. A duplicate pair is defined as being under stabilizing selection if the site regions of the promoters have an genetic distance, measured using the Kimura "2-parameter" method (Kimura, 1980), of less than 0.1 (mean genetic distance is 0.59) and less than the distance for the non site regions of the promoters (Figure 7). These duplicates with conserved promoter regions may still contain sites for different transcription factors but those sites that are shared are highly conserved.

It seems likely that the conservation of common binding sites will act to conserve duplicate gene expression patterns. In the case of subfunctionalization of duplicate pairs the expression pattern of the ancestral gene may be partitioned between the paralogs. Subfunctionalization can therefore be implied if one member of these pairs contains unique sites not found in its paralog but there are also the shared sites that are conserved between the pair.

Interestingly, the duplicate pairs with promoters predicted to be under stabilizing selection show extremely similar functions between genes (Addi-

tional Table 5). Indicating that these promoter regions may be conserved to maintain the function of the genes. The duplicate pair UBX6 (YJL048C) and UBX7 (YBR273C) produce ubiquitin domain containing proteins which interact with each other and Cdc48p in the perinuclear membrane (Decottignies, Evain and Ghislain, 2004). In this case the maintenance of expression patterns is essential for function. We see a similar example with duplicate pair TRE1 (YPL176C) and TRE2 (YOR256C) whose products function together in the degradation of SMF1 (Stimpson, Lewis and Pelham, 2006).

Duplicate promoters show divergence consistent with positive selection

In addition to identifying duplicates with promoters under stabilizing selection we also aimed to identify positive selection acting on the promoter regions of duplicate genes. Here we compared the binding site regions and non-binding site regions between duplicate genes using three different methods of identifying patterns of changes consistent with positive selection. Using data with computationally identified binding sites we find that on average the non-binding site regions have a significantly higher genetic distance, identified using kimura's "2-parameter" model, than the binding site regions, 0.66 and 0.59, respectively (Wilcoxon paired test P=$8.13\text{x}10^{-11}$). This indicates that there is more substitution taking place in the non-binding site regions of duplicate promoters suggesting that most binding site regions are conserved

and non-binding site regions are evolving neutrally.

Despite this it is still possible to detect duplicate promoter regions with high levels of substitution, as the largest genetic distance found for a binding site region is 1.95 and only 0.98 for non-binding site regions (Figure 7). Indeed, the site regions of duplicates have greater variation in distances than that of the non site regions (Levene test, $P<2.2x10^{-16}$) indicating that selection might be acting on these site regions. Using the same definitions as with alleles to identify promoters potentially under positive selection we found 46 duplicate promoters show divergence consistent with positive selection with low confidence and 4 with medium confidence (Additional Table 6). We detect no duplicate promoter regions that have been classed as potentially experiencing positive selection with high confidence evidence. Additionally, we were unable to detect any positive selected duplicate promoters with experimentally determined sites from such a small sample.

Using $K_a/K_s > 1$ as evidence for positive selection in coding regions we attempted to identify any duplicate pairs where both promoter regions and coding regions are experiencing positive selection. Of the 4 duplicate pairs whose coding regions demonstrate signs of positive selection in multiple strains by this criterion, none of these pairs were identified as experiencing positive selection in promoter regions.

The promoter regions which are potentially experiencing positive selection may provide evidence for neofunctionalization, in that there is selection pressure for change in sites that may result in novel binding site formation.

The duplicate pair SSA3 (YBL075C) and SSA4 (YER103W) are members of an essential heat shock family of proteins, *hsp70* (Werner-Washburne, Stone and Craig, 1987). They are expressed at different times and show divergence of binding sites. SSA4 has 2 identified sites for the transcriptional factors MSN2 and MSN4, which do not appear in the SSA3 promoter. Divergence of binding sites as well as unique sites indicates a potential role for neofunctionalization as well as subfunctionalization in duplicate promoter divergence.

Discussion

In this study the promoter regions of 38 strains of *S. cerevisiae* were compared within and between strains with regard to whole genomic content and duplicate genes. We find a large amount of variation between the strains and between duplicate genes in the number of sites for each factor. We further demonstrate that changes in promoter regions are correlated with changes in coding sequences indicating a coupling of changes in expression and function. Perhaps most strikingly we detect divergence between the promoters of alleles and duplicate genes that is consistent with positive selection. These results suggest a role for neofunctionalization in the divergence of gene regulation within and between strains.

We began by identifying ORFs and promoter regions from all strains showing that these strains contained an almost full complement of genes (Table 1). We note that these strains include imputed values (Liti et al., 2009)

and hence under-estimate the variability between the strains. We identified a set of computationally predicted binding sites using transcription factor consensus sequences from JASPAR (Sandelin et al., 2004). The number of computationally predicted sites are likely an over-estimate due to the inclusive nature of the identification method. Alternatively, our prediction of experimentally validated sites is likely an under-estimate (Harbison et al., 2004). The experimental consensus sequences have been assigned to around 550 genes in each strain, whereas in the original study they were assigned to almost 2000 genes (Harbison et al., 2004). There are several reasons for this discrepancy. Firstly, there is an underlying false positive rate in the experimental study that may not be repeated in this study when we find the consensus sequence in the promoter region. Secondly, the promoter region definition in this study may be too strict, and the experimental sites may include sites identified further than 500bp upstream, downstream of the gene or in intron regions. Finally, some of the difference may be present to true variation between the strains. Nevertheless, all of conclusions are consistent regardless of the data set used, giving confidence that our conclusion, albeit conservative, are robust.

A gene's association to a transcription factor is highly conserved between strains of the same species. Meaning that if a specific transcription factor's binding site is present in a particular gene in one strain, all other strains are likely to have at least one site for the factor in their homologous genes. These trends hold for both the computationally and experimentally deter-

mined binding sites (Figure 2). This result is in contrast to that of Doniger and Fay (2007), where more than half of experimentally identified *S. cerevisiae* binding sites were not conserved across closely related species. The difference in these findings could be due to the fact that the strains in this study are much more closely related than the species used by Doniger and Fay (2007). Indeed, when we analyzed the number of sites for each transcription factor across all strains we found more variation. The computationally identified sites showed a 2.6% variation from the mean between strains while the experimentally determined sites showed a 4.6% variation. From this we conclude that while transcription factor association to a gene is conserved across the strains, the actual number of sites for these factors varies between them. If the variation between the number of sites is adaptive we would expect to see evidence of positive selection acting on the promoter regions between alleles.

Since we have identified sites that vary between alleles of these strains we can identify the transcription factors associated with these sites. YAP5 (YIR018W) is a basic leucine zipper (bZIP) transcription factor and its binding sites vary between multiple genes when they are both identified by the computational method and the experimental data. Other bZIP members of the YAP family of bZIP proteins have been shown to be involved in drug resistance (Wu et al., 1993). Indeed YAP5 has been shown to respond to aminotriazole and so part of the organisms response to stress (Fernandes, Rodrigues-Pousada and Struhl, 1997). The changing sites between the strains

for YAP5 might indicate the need for a diverged expression pattern of YAP5 activated genes in some strains.

Interestingly, another bZIP transcription factor is also associated with varying sites between alleles with experimentally identified sites. CIN5 (YOR028C) is a member of the YAP family of bZIP transcription factors (Fernandes, Rodrigues-Pousada and Struhl, 1997) and has been associated with conferring resistance to several drugs. Over-expression of CIN5 leads to increased resistance to cisplatin and two DNA-alkylating agents, methylmethanesulfonate and mitomycin C (Furuchi et al., 2001). We have detected changes in sites between alleles for transcription factors associated with responses to stress. Since the coding regions of these factors appear not to be under positive selection (K_a/K_s of 0.514 and 0.596 for YAP5 and CIN5 respectively) we speculate that the changes in sites might reflect the need for a changed expression pattern of the regulated genes, which may be indicative of adaptation to a specific environment.

We find another transcription factor that shows variation in sites between multiple alleles is also associated with a stress response. GLN3 (YER040W) is a transcription factor that binds to to many genes involved in nitrogen utilization via a zinc finger binding domain (Blinder and Magasanik, 1995). Under nitrogen limiting conditions GLN3 has been shown to increase the expression of nitrogen catabolite repression sensitive genes (Beck and Hall, 1999). The TFBSs that often vary between the strains seem to be regularly associated with transcription factors that mediate a response to stress. This

finding reinforces our hypothesis that that the variation in sites represents the ongoing adaptation of these strains to their environments by altering the expression patterns of specific genes.

The TFBS regions of some promoters are experiencing divergence which is consistent with positive selection. We have identified the promoter regions of a variety of genes which seem to be under positive selection with three methods using both computationally and experimentally identified sites (Additional Tables 1 & 2). The genes associated with these promoter regions show a wide range of functions with genes responsible for transport and uptake of nutrients prominently represented. Notably, we detected several transporter related GO terms that are overrepresented in duplicates showing highly diverged promoter regions (Table 2). It seems likely that adaptation to new environments, as we expect to see in these strains, might be evidenced by adaptations to more efficiently extract nutrients from the environment.

One potentially positively selected promoter region is associated with LYP1 (YNL268W), a lysine specific permease that is responsible for the uptake of lysine and some of its analogues (Grenson, 1966). It has been shown experimentally that overexpression of LYP1 results in increased maximum velocity of lysine uptake (Sychrova and Chevallier, 1993). It seems reasonable that the expression of LYP1 might be altered for increased expression in a lysine limiting environment or in the presence of competition for lysine.

A further example of potential adaptation to new environments through selection on gene regulation of transporters is the transporter SAM3 (YPL274W).

SAM3 encodes a high affinity permease to transport *S*-adenosylmethioine across the plasma membrane of yeast cells, which is required for the utilization of *S*-adenosylmethioine as a sulphur source (Rouillon, Surdin-Kerjan and Thomas, 1999). Again we hypothesize that changes in gene expression in SAM3 would be beneficial to any strain using *S*-adenosylmethioine as a sulphur source.

The promoters of some genes involved in stress response or response to toxins may also be experiencing positive selection. One such gene is FAP1 (YNL023C), which if overproduced, confers rapamycin resistance by competing for binding to Fpr1p (Kunz et al., 2000). Here we have shown how changes to expression of a gene could potentially confer a resistance to drugs.

These results indicate that those genes which show divergence consistent with positive selection in their promoter regions have functions that could aid in adaptation to new environments. Indeed, positive selection has been shown to act on the promoters of neural- and nutrition-related genes (Haygood et al., 2007) and on specific genes important for health in humans (Rockman et al., 2003).

Although we have shown that many promoter regions of alleles show patterns of divergence that are consistent with positive selection it should be noted that positive selection may be difficult to accurately detect for several reasons. Homopolymer runs in binding sites may increase the mutation rate in these reasons due to slippage and so give the site regions a higher rate of change than the surrounding non binding site regions. We find however, that

only 3% and 7.5% of sites having homopolymers of length 4 or more in computationally and experimentally identified sites respectively. These values drop to 0.4% and 1% for homopolymers of lengths 5 or more in computationally and experimentally identified sites respectively. There are alternative sequence features that may have affected our positive selection analysis. For instance, DNA that is in a promoter region but not in a TFBS may be essential for maintaining the structure of DNA. This constraint would cause our methods to annotate a selectively constrained sequence as neutral and thus may make our binding site substitution rate erroneously high, meaning we could incorrectly detect evidence of positive selection. For these reasons we have classed the promoter regions identified is this study as potentially positively selected.

Duplicate genes offer an excellent opportunity to investigate divergence of promoter regions within strains. If duplicate genes are not retained for dosage effects they are expected to diverge at the level of expression or coding sequence. In particular the subfunctionalization hypothesis states that duplicates accumulate complementary degenerative mutations in either regulatory regions or coding sequence in order to partition the ancestral function between the duplicates (Force et al., 1999; Lynch et al., 2001). We have shown a significant negative correlation ($P<2.2x10^{-16}$) between the proportion of shared binding sites and the synonymous mutation rate, which serves as a proxy for age (Figure 5). Previous studies have found the same trend when correlating expression divergence or shared regulatory motifs with some

measure of evolutionary time at the species level (Gu et al., 2002; Papp, Pál and Hurst, 2003; Zhang, Gu and Gu, 2004). Duplicate genes have also been shown to increase expression diversity within several species of *Drosophila* and within strains of yeast (Gu et al., 2004), although this is the first time the sites likely to be responsible for these changes have been identified at the population level.

There is a significant positive correlation between K_s and both binding site and non-binding site regions of promoters (Figure 6A). This result might be expected as we would expect both synonymous changes in protein coding regions and non-binding site regions to evolve neutrally. We might also expect some neutral evolution in binding site regions, as some substitutions within binding motif may not affect binding of the transcription factor.

Interestingly, there is a significant positive correlation between K_a of duplicate genes and the genetic distance of the binding site regions of the corresponding promoters but not the non-binding site regions (Figure 6B). This result suggests that non-synonymous changes that might lead to functional change in the coding sequence are correlated with changes in the site regions of the corresponding promoter that may lead to changes in expression pattern. Indeed, it has been previously shown that there is a significant negative relationship between the expression correlation of duplicate genes and K_a (Gu et al., 2002). Similar results have been obtained when analyzing duplicate genes in *Arabidopsis thaliana*, *Drosophila melanogaster* and humans (Ganko, Meyers and Vision, 2007; Kohn, 2008; Park and Makova, 2009). To-

gether these results indicate that functional change of diverging duplicate genes at the coding sequence level is coupled with change in transcription factor binding sites and expression.

The overrepresentation of different GO terms in duplicates with conserved and diverged promoter regions demonstrates that certain types of genes are more likely to diverge in expression pattern than others (Table 2). Interestingly, previous studies have shown that transporter genes and other membrane proteins show expression divergence after duplication (Gu et al., 2002), which tallies well with this study. Additionally, Gu et al. (2002) have also shown that ribosomal proteins show conserved patterns of expression, yet we detect ribosomal GO terms overrepresented in both our conserved and diverged sets.

We also detect variation in overrepresented GO terms between strains (Additional File 1) and this points to the different utilization of duplicate genes between these strains. It has already been demonstrated that these strains differentially retain duplicates depending on their natural environment (Ames et al., 2010). This result suggests that the same duplicates present across the strains may evolve differently in each strain. If this were the case we would expect to see the evidence of positive selection acting on promoter regions between the strains.

Using the genetic distance between the binding site and non-binding site regions of duplicate promoters we have been able to identify those duplicates with highly conserved binding sites that we believe to be experiencing sta-

bilizing selection (Additional Table 5). We note that under our definition of conserved promoters we may detect duplicates with different compliments of binding sites where any shared sites are highly conserved. These cases may represent duplicate promoters that have undergone subfunctionalization either in expression pattern or gene function. In this case the highly conserved sites might act to ensure amount or timing of gene expression is conserved.

The duplicate pair TRE1 (YPL176C) and TRE2 (YOR256C) show a high amount of conservation in their associated TFBSs. The genes function in the degradation of SMF1 (Stimpson, Lewis and Pelham, 2006), a manganese transporter (Supek et al., 1996), which is vital for the survival of yeast in the presence of heavy metals. As both genes are required for the degradation of SMF1 (Stimpson, Lewis and Pelham, 2006) it seems likely that any strains that encounter heavy metals in their environment will be under selective pressure to maintain both genes and maintain their expression pattern and that this is reflected in the pairs conserved promoter regions across at least some of the strains. In this case the coding regions of the genes may have subfunctionalized so that both are required for SMF1 degradation.

Signs of positive selection acting on the promoter regions of duplicate genes are also detectable. We detected 50 duplicate pairs whose transcription factor binding sites show greater divergence than the non binding site regions(Additional Table 6). Additionally, we can also detect sites unique to one paralog of a duplicate pair which also suggests divergence expression patterns.

The duplicate pair SSA3 (YBL075C) and SSA4 (YER103W) are members of an essential heat shock family of proteins, *hsp70* (Werner-Washburne, Stone and Craig, 1987). SSA3 is expressed after the diauxic shift or in response to heat shock Werner-Washburne, Stone and Craig (1987); Werner-Washburne et al. (1989), while SSA4 is expressed during the diauxic shift and in response to heat, cold or ethanol stress (Werner-Washburne et al., 1989; Boorstein and Craig, 1990; Kandror et al., 2004; Quan et al., 2004). These genes have different expression patterns but a highly conserved amino acid sequence (Boorstein, Ziegelhoffer and Craig, 1994). We see that SSA4 has 2 identified sites for the transcriptional factors MSN2 and MSN4, whereas SSA3 shows non of these sites in the reference strain, which may explain the different expression patterns. These factors have been shown to activate transcription of genes under ethanol stress (Martinez-Pastor et al., 1996). These proteins have an overlap in function, both are involved in the response to heat stress, but otherwise appear to have diverged in order to respond to a wider variety of stressful environments. The divergence of the promoter regions for these duplicates are consistent with positive selection. Diverging regulatory regions that change the timing and conditions of gene expression may be indicative of environmental adaptation.

There is evidence that these promoters may have diverged by both sub- and neo-functionalization. The presence of unique sites in SSA4 might well signify subfunctionalization with these sites being lost from the SSA3 promoter. However, the higher rate of substitution in the binding sites when

compared to the neutral rate of non site regions suggests a selection pressure for changes in TFBSs. This indicates a role for neofunctionalization in the divergence of these duplicate promoters.

Our results indicate that divergence of TFBSs in a variety of duplicate genes show patterns consistent with positive selection. We should note that the same limitations of identifying positive selection in alleles, discussed above, applies to these duplicates. Higher rates of substitution than the neutral rate suggest a selection pressure for changes in TFBSs, which in turn suggests a prominent role for neofunctionalization as well as subfunctionalization in duplicate promoter divergence. This finding adds more evidence to the role of neofunctionalization in duplicate divergence (Papp, Pál and Hurst, 2003).

Conclusion

Variation within a population is the raw material that evolution acts upon. Selection acting on this variation leads to functional adaptation, and so shapes the genome. The *Saccharomyces* Genome Resequencing Project (Liti et al., 2009) is the first large-scale resequencing project that provides multiple genome sequences for a single species, and the only such project where information about the environment from which the organisms were isolated is available. Therefore it remains the best resource to study genomic variation within a species, and the earliest evolutionary events that fix genomic

variation before speciation.

We have previously shown (Ames et al., 2010) that the environment can radically alter the gene content of different strains of *S. cerevisiae* by selecting for retention of a subset of duplicated genes. Here we study the promoter regions both within and between strains. Through evolution of TFBSs an organism can adapt to its environment by the alteration of gene expression patterns. We find that, even within a species, TFBSs vary substantially. Moreover, some binding sites show patterns of divergence consistent with positive selection, indicating functional innovation through neofunctionalization. Many changes in promoters can be rationalized by examining the GO classification of the associated gene, suggesting that, as with duplicate retention, the environment is selecting for the observed differences.

Interestingly, in many cases there are also non-synonymous substitutions in the protein-coding regions. This observation hints at functional adaptation of the protein sequences themselves, concomitant with changes in regulatory regions. Thus, a picture continues emerges of widespread genomic variation within yeast, both in gene content, gene regulation and protein sequence that is shaped by the environment. This view highlights the earliest stages of functional adaptation at the population level, and prior to speciation.

Supplementary material

The following data are available with the online version of this paper. Additional Table 1 lists the 80 alleles whose promoters with computational sites were identified as undergoing positive selection. Additional Table 2 lists the 16 alleles whose promoters with experimentally determined sites were identified as experiencing positive selection. Additional Table 3 shows the over-represented 'Molecular Function' GO terms for duplicates with conserved and diverged promoters common to all strains. Additional Table 4 shows the over-represented 'Cellular Component' GO terms for duplicates with conserved and diverged promoters common to all strains. Additional Table 5 shows the duplicate pairs whose promoters are under stabilizing selection. Additional Table 6 shows the duplicate pairs whose promoters are under positive selection. Finally, Additional File 1 lists all over-represented GO terms for duplicates with conserved and diverged promoters for all strains.

Acknowledgments

We thank Craig Lawless for providing the transcriptional start site data and for useful discussion. We also thank Catherine Walton, Simon Whelan and Daniel Money for useful discussion of the manuscript. This work was supported by the Biotechnology and Biological Sciences Research Council.

References

Ames RM, Rash BM, Hentges KE, Robertson DL, Delneri D, Lovell SC. 2010. Gene duplication and environmental adaptation within yeast populations. Genome biology and evolution. 2:591–601.

Ashburner M, Ball C, Blake J, et al. (11 co-authors). 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–9.

Beck T, Hall M. 1999. The tor signalling pathway controls nuclear localization of nutrient-regulated transcription factors. Nature. 402:689–692.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological). 57:289–300.

Blinder D, Magasanik B. 1995. Recognition of nitrogen-responsive upstream activation sequences of saccharomyces cerevisiae by the product of the gln3 gene. Journal of bacteriology. 177:4190–4193.

Boorstein W, Craig E. 1990. Structure and regulation of the SSA4 HSP70 gene of Saccharomyces cerevisiae. Journal of Biological Chemistry. 265:18912–18921.

Boorstein W, Ziegelhoffer T, Craig E. 1994. Molecular evolution of the HSP70 multigene family. Journal of molecular evolution. 38:1–17.

Carroll SB. 2005. Evolution at two levels: on genes and form. PLoS Biology. 3:1159–1166.

Decottignies A, Evain A, Ghislain M. 2004. Binding of Cdc48p to a ubiquitin-related UBX domain from novel yeast proteins involved in intracellular proteolysis and sporulation. Yeast. 21:127–139.

D'haeseleer P. 2006. What are DNA sequence motifs? Nature biotechnology. 24:423–426.

Doniger S, Fay J. 2007. Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol. 3:932–942.

Edgar R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 32:1792–1797.

Felsenstein J. 1989. PHYLIP-phylogeny inference package (version 3.2). Cladistics. 5:164–166.

Fernandes L, Rodrigues-Pousada C, Struhl K. 1997. Yap, a novel family of eight bzip proteins in saccharomyces cerevisiae with distinct biological functions. Molecular and cellular biology. 17:6982–6993.

Force A, Lynch M, Pickett F, Amores A, Yan Y, Postlethwait J. 1999. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics. 151:1531–1545.

Furuchi T, Ishikawa H, Miura N, Ishizuka M, Kajiya K, Kuge S, Naganuma A. 2001. Two nuclear proteins, cin5 and ydr259c, confer resistance to cisplatin in saccharomyces cerevisiae. Molecular Pharmacology. 59:470–474.

Ganko E, Meyers B, Vision T. 2007. Divergence in expression between duplicated genes in Arabidopsis. Molecular biology and evolution. 24:2298–2309.

Grenson M. 1966. Multiplicity of the amino acid permeases in saccharomyces cerevisiae: Ii. evidence for a specific lysine-transporting system. Biochimica et biophysica acta. 127:339–346.

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

Gu Z, Rifkin S, White K, Li W. 2004. Duplicate genes increase gene expression diversity within and between species. Nature Genetics. 36:577–579.

Hakes L, Pinney J, Lovell S, Oliver S, Robertson D. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biology. 8:R209.

Harbison C, Gordon D, Lee T, et al. (11 co-authors). 2004. Transcriptional regulatory code of a eukaryotic genome. Nature. 431:99–104.

Haygood R, Fedrigo O, Hanson B, Yokoyama K, Wray G. 2007. Promoter re-

gions of many neural-and nutrition-related genes have experienced positive selection during human evolution. Nature genetics. 39:1140–1144.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics. 11:97–108.

Kandror O, Bretschneider N, Kreydin E, Cavalieri D, Goldberg A. 2004. Yeast adapt to near-freezing temperatures by STRE/Msn2, 4-dependent induction of trehalose synthesis and certain molecular chaperones. Molecular cell. 13:771–781.

Kellis M, Birren B, Lander E. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature. 428:617–624.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution. 16:111–120.

Kohn M. 2008. Rapid sequence divergence rates in the 5 prime regulatory regions of young Drosophila melanogaster duplicate gene pairs. Genetics and Molecular Biology. 31:575–584.

Kostka D, Hahn MW, Pollard KS. 2010. Noncoding sequences near duplicated genes evolve rapidly. Genome biology and evolution. 2:518–33.

Kunz J, Loeschmann A, Deuter-Reinhard M, Hall M. 2000. FAP1, a homo-

logue of human transcription factor NF-X1, competes with rapamycin for binding to FKBP12 in yeast. Molecular Microbiology. 37:1480–1493.

Lawless C, Pearson R, Selley J, Smirnova J, Grant C, Ashe M, Pavitt G, Hubbard S. 2009. Upstream sequence elements direct post-transcriptional regulation of gene expression under stress conditions in yeast. BMC genomics. 10:7–27.

Levine M, Tjian R. 2003. Transcription regulation and animal diversity. Nature. 424:147–51.

Liti G, Carter D, Moses A, et al. (11 co-authors). 2009. Population genomics of domestic and wild yeasts. Nature. 458:337–341.

Lynch M, O'Hely M, Walsh B, Force A. 2001. The probability of preservation of a newly arisen gene duplicate. Genetics. 159:1789–1804.

Martinez-Pastor M, Marchler G, Sch
"uller C, Marchler-Bauer A, Ruis H, Estruch F. 1996. The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). The EMBO journal. 15:2227–2235.

Ohno S. 1970. Evolution by Gene Duplication. Springer, New York.

Papp B, Pál C, Hurst L. 2003. Evolution of cis-regulatory elements in duplicated genes of yeast. TRENDS in Genetics. 19:417–422.

Park C, Makova K. 2009. Coding region structural heterogeneity and turnover of transcription start sites contribute to divergence in expression between duplicate genes. Genome Biology. 10:R10.

Prince V, Pickett F. 2002. Splitting pairs: the diverging fates of duplicated genes. Nature Reviews Genetics. 3:827–837.

Quan X, Rassadi R, Rabie B, Matusiewicz N, Stochaj U. 2004. Regulated nuclear accumulation of the yeast hsp70 Ssa4p in ethanol-stressed cells is mediated by the N-terminal domain, requires the nuclear carrier Nmd5p and protein kinase C. The FASEB Journal. NA:309471.

Rockman M, Hahn M, Soranzo N, Goldstein D, Wray G. 2003. Positive selection on a human-specific transcription factor binding site regulating IL4 expression. Current Biology. 13:2118–2123.

Rouillon A, Surdin-Kerjan Y, Thomas D. 1999. Transport of sulfonium compounds. Characterization of the s-adenosylmethionine and s-methylmethionine permeases from the yeast Saccharomyces cerevisiae. Journal of Biological Chemistry. 274:28096–28105.

Sandelin A, Alkema W, Engstrom P, Wasserman W, Lenhard B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research. 32:D91–D94.

Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genetical Research. 23:23–35.

Spofford J. 1969. Heterosis and the Evolution of Duplications. The American Naturalist. 103:407–432.

Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees.

Stimpson H, Lewis M, Pelham H. 2006. Transferrin receptor-like proteins control the degradation of a yeast metal transporter. The EMBO journal. 25:662–672.

Supek F, Supekova L, Nelson H, Nelson N. 1996. A yeast manganese transporter related to the macrophage protein involved in conferring resistance to mycobacteria. Proceedings of the National Academy of Sciences of the United States of America. 93:5105–5110.

Sychrova H, Chevallier M. 1993. Cloning and sequencing of the Saccharomyces cerevisiae gene LYP1 coding for a lysine-specific permease. Yeast. 9:771–782.

Werner-Washburne M, Becker J, Kosic-Smithers J, Craig E. 1989. Yeast Hsp70 RNA levels vary in response to the physiological status of the cell. Journal of bacteriology. 171:2680–2688.

Werner-Washburne M, Stone D, Craig E. 1987. Complex interactions among members of an essential subfamily of hsp70 genes in Saccharomyces cerevisiae. Molecular and cellular biology. 7:2568–2577.

Whitton O. 2000. Polyploid incidence and evolution. Annual Review of Genetics. 34:401–437.

Wu A, Wemmie J, Edgington N, Goebl M, Guevara J, Moye-Rowley W. 1993. Yeast bzip proteins mediate pleiotropic drug and metal resistance. Journal of Biological Chemistry. 268:18850–18858.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution. 24:1586–1591.

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

Table 1: Identified genes, duplicates and sites for all strains

Strain	Genes	Duplicate Genes	Genes with comp sites	Genes with exp sites
273614N	5332	1174	5287	548
322134S	5529	1264	5477	565
378604X	5360	1188	5312	546
BC187	5540	1268	5490	565
DBVPG1106	5507	1250	5458	549
DBVPG1373	5558	1274	5509	568
DBVPG1788	5533	1272	5483	552
DBVPG1853	5500	1254	5449	550
DBVPG6040	5528	1276	5479	563
DBVPG6044	5345	1174	5300	541
DBVPG6765	5573	1284	5523	569
K11	5337	1166	5290	546
L_1374	5558	1276	5507	565
L_1528	5553	1276	5503	566
NCYC110	5345	1182	5299	540
NCYC361	5335	1168	5288	546
REF	5794	1440	5741	586
RM11_1A	5596	1298	5545	572
SK1	5395	1202	5347	544
UWOPS03_461_4	5459	1234	5411	542
UWOPS05_217_3	5505	1256	5456	565
UWOPS05_227_2	5460	1236	5412	544
UWOPS83_787_3	5532	1270	5485	564
UWOPS87_2421	5529	1278	5478	569
W303	5659	1344	5607	582
Y12	5322	1156	5277	541
Y55	5389	1202	5343	545
Y9	5508	1250	5460	561
YIIc17_E5	5338	1172	5291	541
YJM789	5617	1312	5567	576
YJM975	5529	1266	5480	565
YJM978	5370	1192	5322	543
YJM981	5334	1160	5288	546
YPS128	5328	1162	5284	540
YPS606	5322	1158	5278	540
YS2	5532	1264	5483	565
YS4	5343	1172	5297	543
YS9	5341	1172	5293	543

Table 2: Overrepresented 'Biological Process' terms for duplicates with diverged and conserved promoter regions

Conserved		Diverged		Shared	
GO term	Description	GO term	Description	GO term	Description
GO:0006468	Protein amino acid phosphorylation	GO:0000947	Amino acid catabolic process to alcohol via Ehrlich pathway	GO:0007264	Small GTPase mediated signal transduction
GO:0006643	Membrane lipid metabolic process	GO:0005996	Monosaccharide metabolic process		
GO:0006665	Sphingolipid metabolic process	GO:0006006	Glucose metabolic process		
GO:0006672	Ceramide metabolic process	GO:0006007	Glucose catabolic process		
GO:0006793	Phosphorus metabolic process	GO:0006412	Translation		
GO:0006796	Phosphate metabolic process	GO:0006417	Regulation of translation		
GO:0007124	Pseudohyphal growth	GO:0007039	Vacuolar protein catabolic process		
GO:0007165	Signal transduction	GO:0008643	Carbohydrate transport		
GO:0007568	Aging	GO:0008645	Hexose transport		
GO:0007569	Cell aging	GO:0009063	Cellular amino acid catabolic process		
GO:0008360	Regulation of cell shape	GO:0009083	Branched chain family amino acid catabolic process		
GO:0015696	Ammonium transport	GO:0009743	Response to carbohydrate stimulus		
GO:0015892	Siderophore-iron transport	GO:0009850	Auxin metabolic process		
GO:0016049	Cell growth	GO:0009851	Auxin biosynthetic process		
GO:0016301	Kinase activity	GO:0015749	Monosaccharide transport		
GO:0016310	Phosphorylation	GO:0015758	Glucose transport		
GO:0023046	Signaling process	GO:0019318	Hexose metabolic process		
GO:0023060	Signal transmission	GO:0030163	Protein catabolic process		
GO:0032502	Developmental process	GO:0031505	Fungal-type cell wall organization		
GO:0040007	Growth	GO:0035556	Intracellular signal transduction		
GO:0042026	Protein refolding	GO:0042221	Response to chemical stimulus		
GO:0044237	Cellular metabolic process	GO:0042445	Hormone metabolic process		
GO:0044262	Cellular carbohydrate metabolic process	GO:0044257	Cellular protein catabolic process		
GO:0046513	Ceramide biosynthetic process	GO:0046165	Alcohol biosynthetic process		
GO:0046519	Sphingoid metabolic process	GO:0055085	Transmembrane transport		

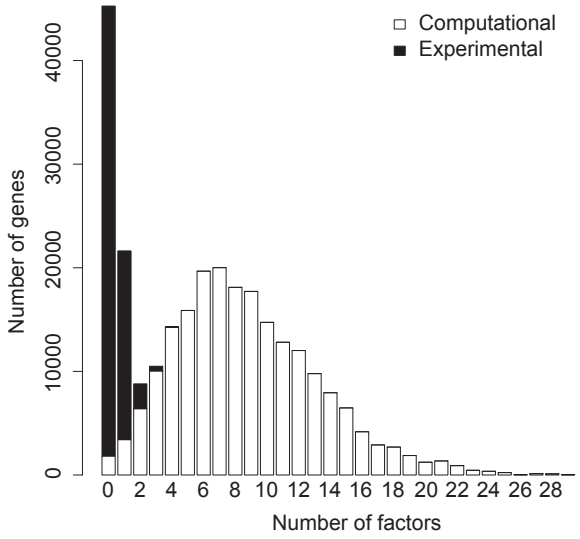

Figure 1: The distribution of transcription factor binding sites across all strains. The number of transcription factors with sites per gene identified computationally (white bars) and experimentally (black bars Harbison et al. (2004)) for all genes across all strains.

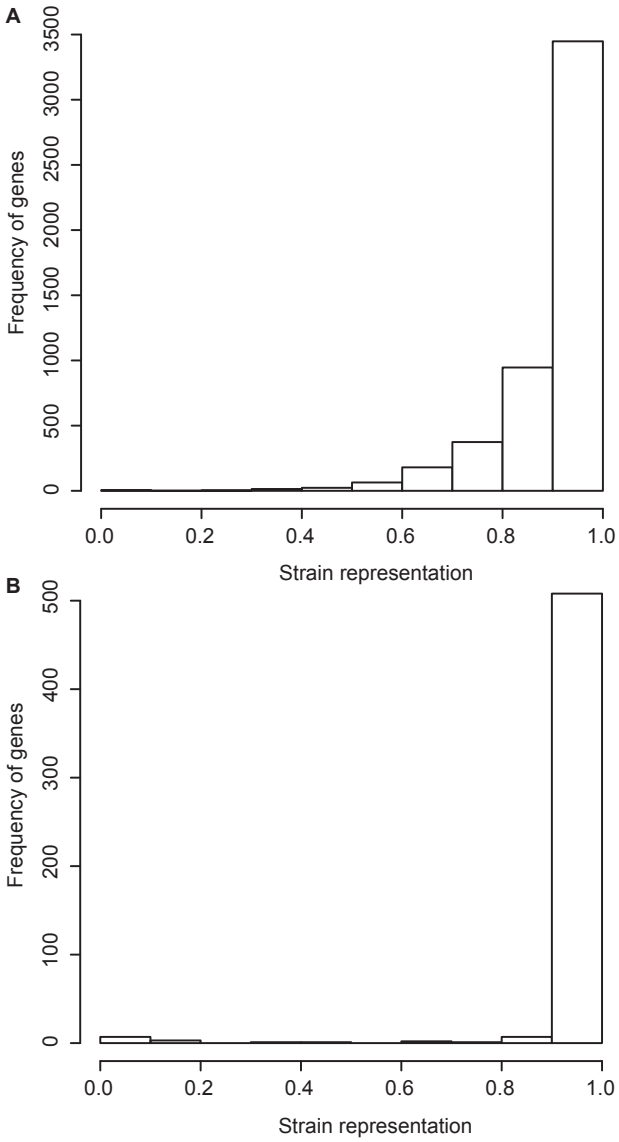

Figure 2: Conservation of transcription factor binding sites across all strains in **(A)** computationally and **(B)** experimentally predicted sites. The strain representation shows the average proportion of strains which contain any number of sites for a given genes associated transcription factors.

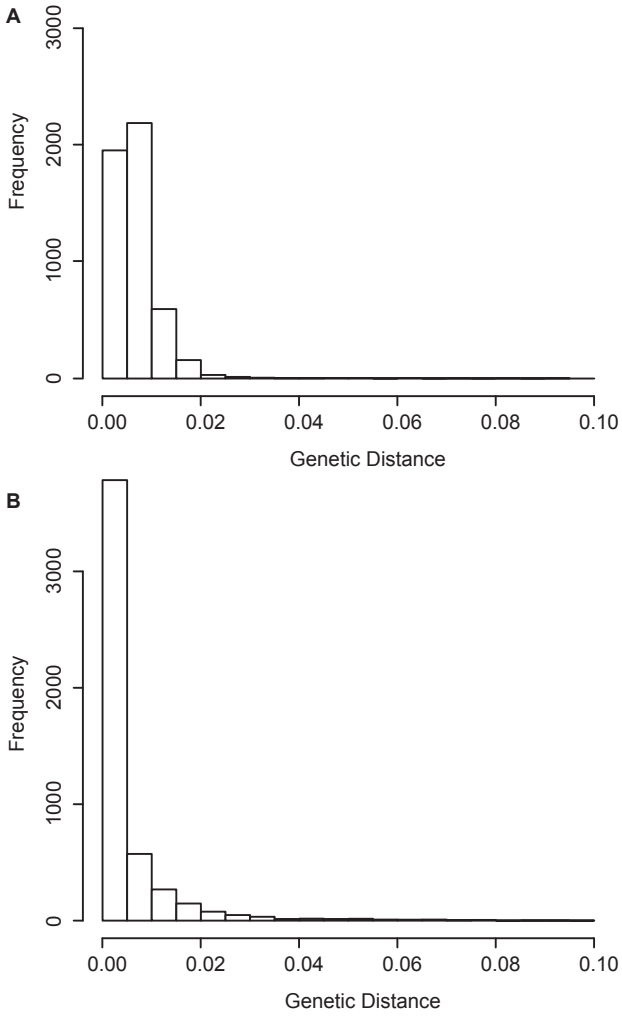

Figure 3: Distribution of distances between alleles in **(A)** non site regions and **(B)** site regions of promoters. Distance is measured using the Kimura "2-parameter" model (Kimura, 1980).

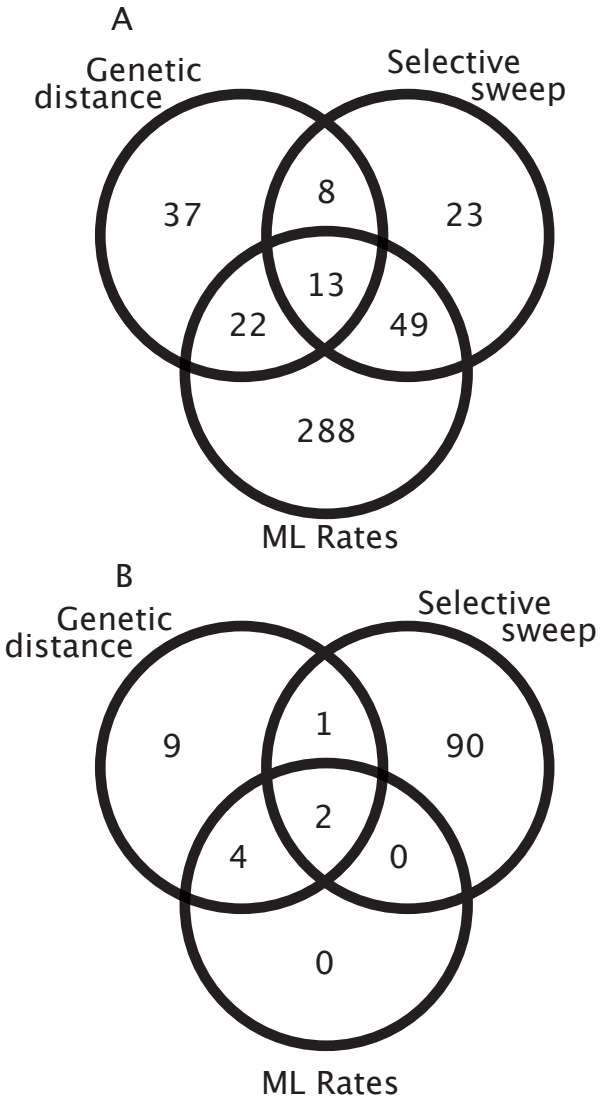

Figure 4: Number of potentially positively selected alleles identified by three methods using **(A)** computationally identified sites and **(B)** experimentally identified sites. We used methods based on genetic distance between binding site and non binding site regions, selective sweep and maximum likelihood inferred rates of change in binding site and non binding site regions. Those alleles identified by only one methods were classified as low confidence. Alleles identified by two methods were classed medium confidence and high confidence alleles were identified by all three methods.

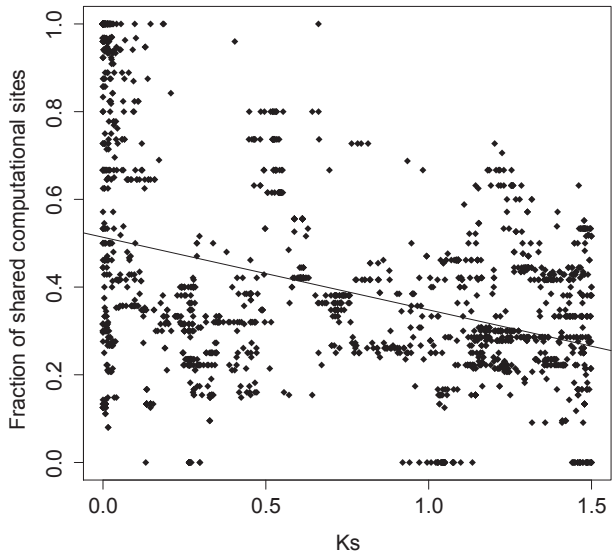

Figure 5: Promoter region binding site divergence between duplicate pairs. The duplicate pairs are from all 38 strains of *S. cerevisiae* and only include duplicates with $K_s < 1.5$ and an effective number of codons > 30 to ensure reliability of the K_s estimates.

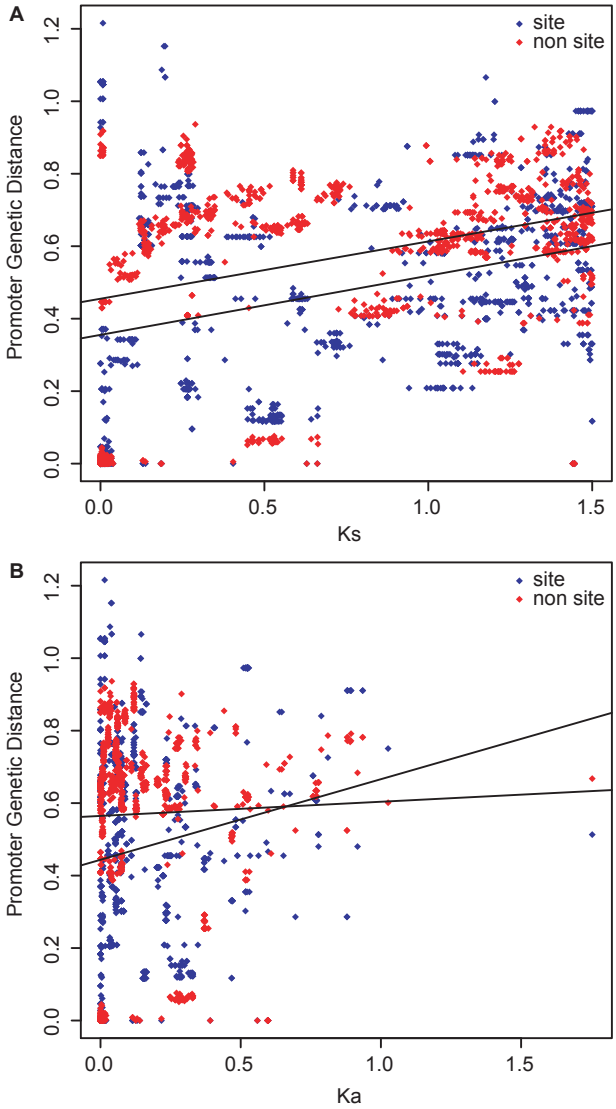

Figure 6: Correlation between divergence of promoter regions and divergence of coding sequences in duplicate genes. Relationship between **(A)** synonymous mutation rate (K_s) and **(B)** non synonymous mutation rate (K_a) in coding regions of duplicate pairs and the genetic distance between site and non site regions of promoters measured using the Kimura "2-parameter" model (Kimura, 1980). Site regions are represented by blue diamonds and non site regions by red diamonds.

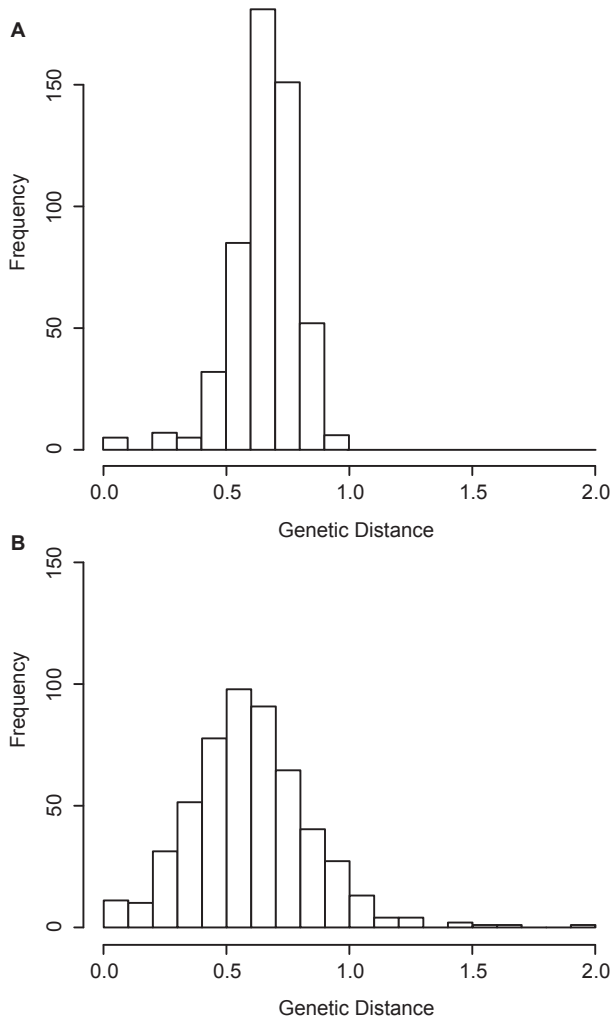

Figure 7: Distribution of genetic distances between duplicate pairs in **(A)** non site regions and **(B)** site regions of promoters. Genetic distance is measured using the Kimura "2-parameter" model (Kimura, 1980).