JOURNAL OF EXPERIMENTAL ZOOLOGY (MOL DEV EVOL) 000:1–20 (2004)

Molecular Evolution in the Yeast Transcriptional Regulation Network

ANNETTE M. EVANGELISTI AND ANDREAS WAGNER*
Department of Biology, University of New Mexico, Albuquerque, New Mexico, 87131

ABSTRACT We analyze the structure of the yeast transcriptional regulation network, as revealed by chromatin immunoprecipitation experiments, and characterize the molecular evolution of both its transcriptional regulators and their target (regulated) genes. We test the hypothesis that highly connected genes are more important to the function of gene networks. Three lines of evidence—the rate of molecular evolution of network genes, the rate at which network genes undergo gene duplication, and the effects of synthetic null mutation in network genes—provide no strong support for this hypothesis. In addition, we ask how network genes diverge in their transcriptional regulation after duplication. Both loss (subfunctionalization) and gain (neofunctionalization) of transcription factor binding play a role in this divergence, which is often rapid. On the one hand, gene duplicates experience a net loss in the number of transcription factors binding to them, indicating the importance of losing transcription factor binding sites after gene duplication. On the other hand, the number of transcription factors that bind to highly diverged duplicates is significantly greater than would be expected if loss of binding played the only role in the divergence of duplicate genes. *J. Exp. Zool. (Mol. Dev. Evol.) 302B:000–000, 2004.*

INTRODUCTION

Transcriptional regulators and the genes whose expression they regulate—their target genes—form large gene regulation networks (Perez-Rueda and Collado-Vides, 2000; Guelzim et al., 2002; Lee et al., 2002; Salgado et al., 2004). These and other molecular networks, such as protein interaction networks and metabolic networks, are intensely studied, because their characterization has been greatly facilitated by new techniques in genomics and bioinformatics (Uetz et al., 2000; Ito et al., 2001; Lee et al., 2002; von Mering et al., 2002; Salgado et al., 2004). Information about the structure of molecular networks opens a new dimension to studies of molecular evolution, because it allows inquiries that go beyond the evolution of individual genes. Network evolution and gene evolution are of course not independent. On the one hand, we know that mutations at the level of individual genes—including gene duplications—influence the structure of these networks. On the other hand, natural selection acting on the global structure of a network may influence what kind of mutations can be tolerated on the gene level (Wagner, 2001; Sole et al., 2002; Wagner, 2003; Chung et al., 2003; van Noort et al., 2004). Put differently, the structure of the network may influence the evolution of genes and vice versa. This interplay is part of the reason why network evolution is an intriguing and increasingly popular subject of study.

We currently know very little empirically about the evolution of large genetic networks. The first step towards acquiring more knowledge consists of a basic characterization of network structure and of how a gene's connectivity may affect the gene's evolution and the network's function. We here present such a basic analysis for the yeast transcriptional regulation network. Such an analysis may be interesting in its own right, but it also sheds light on questions that biologists have been asking for decades. We illustrate this with one question, how gene functions diverge after gene duplication.

Gene duplications play dual roles in evolution. On the one hand, gene duplicates that retain similar functions can be a source of gene redundancy,

*Correspondence to Andreas Wagner, Department of Biology, 167 Castetter Hall, University of New Mexico, Albuquerque, NM 87131. E-mail: wagnera@unm.edu
Received 9 February 2004; Accepted 5 April 2004
Published online 00 Month 2004 in Wiley Interscience (www.interscience.wiley.com). DOI: 10.1002/jez.b.20027

which may buffer organisms against mutations (Wagner, '99; Gonzalez-Gaitan et al., '94; Wang et al., '96; Nowak et al., '97; Gu et al., 2003; Conant and Wagner, 2004). On the other hand, gene duplicates that diverge in function contribute to evolutionary innovation on the biochemical level (Hughes, '94; Zhang et al., '98; Briscoe, 2001). Which of these roles is predominant? That is, do most gene duplicates retain similar functions long after duplication, or do they diverge rapidly? Furthermore, when two genes diverge in their functions, how does this divergence take place? The two principal possibilities are the acquisition of new functions (neofunctionalization) and the partitioning of existing functions between two duplicates. The last mode of divergence has generated considerable recent attention, because it has been argued that such divergence can account for the maintenance of many gene duplicates in eukaryotic genomes (Force et al., '99a; Lynch and Force, 2000; Prince and Pickett, 2002). However, most evidence regarding the tempo and mode of divergence comes from studies of individual genes and is thus anecdotal.

To answer the above questions one must define, quantify, and compare gene functions. However, to do so raises enormous difficulties, which are encapsulated in the multiple complementary ways to categorize gene functions (Ashbuner et al., 2000). They include the biological process a gene acts in, its product's subcellular localization, and its biochemical activity. These difficulties are also illustrated by the discovery that many genes long thought to have one mundane and well-characterized function—such as enzymatic activity—also have entirely, often completely unanticipated roles (Jeffery, '99). Examples include the glycolytic enzyme phosphoglucose isomerase, which also serves as the cell-signaling molecule neuroleukin, a cytokine causing immune cell maturation, and survival of some embryonic spinal nerve cells (Chaput et al., '88; Faik et al., '88); thymidine phosphorylase, which catalyzes the dephosphorylation of thymidine and deoxyuridine, and is the same as an endothelial growth factor (Furukawa et al., '92; Haraguchi et al., '94); aconitase, an enzyme in the tricarboxylic acid cycle, which also serves as a translational regulator of ferritin expression (Kennedy et al., '92); and carbinolamine dehydratase, which serves in phenylalanine metabolism but also regulates the DNA binding activity of the homeodomain transcription factor hepatic nuclear factor 1α (Jeffery, '99).

With such examples in mind, it may seem utterly hopeless to exhaustively quantify gene function to gain insight into the questions raised above. However, not all is lost. A possible alternative approach consists in studying only one aspect of gene function—however minute—and assay this aspect of gene function for many (duplicate) genes. Take the example of gene expression. When and where a gene is expressed may provide an indication of its function: there are several known cases of gene duplicates in developmental genes, duplicates whose biochemical activity is identical, but whose biological function is different because they are expressed in different tissues or cell populations. With the advent of microarray technology, large-scale measurements of gene expression have become feasible. They can be used to compare this indicator of gene function among many duplicate genes and to determine their rate of divergence (Wagner, 2000; Gu et al., 2002). Other gene function indicators include the molecular interaction partners of a gene product; a gene's synthetic lethal interactions with other genes; the spectrum of transcription factors regulating the expression of a gene (because it may indicate similarity in gene expression); and—specific to genes encoding transcription factors—the regulatory targets of a transcription factor. In this paper, we use the last two indicators of gene function.

The subject of this paper is the transcriptional regulation network of the yeast *Saccharomyces cerevisiae* and the evolution of its genes. Although primarily descriptive, our analysis provides preliminary answers to the questions raised above, as well as to several others. Do gene duplicates diverge in function, or do they retain similar functions, and thus partial redundancy, for a long time? Which is the dominant mode of functional divergence, partitioning of existing transcriptional regulation interactions, or the acquisition of new interactions? Does a gene's connectivity influence its chances to undergo gene duplication, its rate of molecular evolution, or the ability to tolerate mutations? The answers we obtain are preliminary, because information on the network's structure is still limited. Each among several data sets on transcriptional regulation networks (Perez-Rueda and Collado-Vides, 2000; Bhan et al., 2002; Guelzim et al., 2002; Lee et al., 2002; Salgado et al., 2004) has its own weaknesses, which include ascertainment biases and sometimes only indirect evidence for transcriptional regulation. We here chose to use the most recent

and most exhaustive data available, based on a genome-wide chromatin immunoprecipitation experiment (Lee et al., 2002). This analysis involved 106 transcriptional regulators and thousands of likely transcriptional regulation interactions, indicated by the binding of transcriptional regulators to a gene's regulatory regions.

METHODS

Transcriptional regulation data

To identify transcriptional regulators and their target genes—the genes whose expression they regulate—we used results of an immunoprecipitation experiment conducted by Lee and collaborators (Lee et al., 2002). This experiment determined the binding affinity of well-documented transcriptional regulators to regulatory regions of all *S. cerevisiae* genes. The authors started with the 141 best-characterized transcription regulators in the Yeast Proteome Database (Costanzo et al., 2000), and constructed yeast strains in which each of these regulators was tagged with an epitope. Thirty-five of the regulators were eliminated from the study because they were not expressed under the experimental conditions (growth in the rich medium YPD, which contains yeast extract, peptone, and dextrose) or because their tagging was unsuccessful. This left 106 regulators for analysis (Lee et al., 2002).

For each of these 106 regulators, the epitope tag was used in three replicate chromatin immunoprecipitation experiments (Knop et al., '99) to identify genomic DNA to which these regulators bound (Ren et al., 2000). The immunoprecipitated DNA was hybridized to DNA microarrays containing the regulatory regions upstream of known yeast genes. The fluorescence intensity of a spot (regulatory region) on the array indicates the binding strength of a transcriptional regulator to the regulatory region. This indication of binding is quantitative, but for many analyses, a qualitative (all-none) indication of binding and transcriptional regulation is more useful. The authors thus developed an error model of binding that allowed them to assign a probability or P-value of binding for each transcriptional regulator to a gene's regulatory region (Lee et al., 2002). This P-value indicates the confidence one has in a factor's binding to a specific DNA region. We here generally follow the authors' suggestion of equating *bona fide* binding of a transcriptional regulator to a target gene if this P-value is smaller than 10^{-3}. This P-value minimizes the number of false-positive binding interactions, while maximizing the number of true positive regulator-target binding interactions (Lee et al., 2002). Doing so results in 4358 interactions with $P < 10^{-3}$. We also repeated our analysis for drastically less stringent ($P < 10^{-2}$) and more stringent ($P < 10^{-5}$) binding thresholds (results not shown), with no qualitative change to the results we report in detail here.

Connectivity

It is important to be aware that the number of regulatory regions bound by a transcription factors depends on the factor's affinity to its binding sites, as well as on the factor's concentration in the cell. Thus, connectivity is best thought of as a composite variable rather than as a simple number. This does not hold only for our data but for all analyses of molecular interaction networks to date. With this caveat in mind, a natural representation of the transcriptional regulation data generated by Lee (Lee et al., 2002) is a directed graph. A node represents a gene and a directed edge from gene x to gene y indicates that x is a transcription factor that has bound to the regulatory region of gene y at $P < 10^{-3}$. In this case, we will refer to gene x as a transcription factor and to gene y as its target gene. The connectivity of a transcriptional regulator is then the number of edges that emanate from it, its outdegree, which is interpreted as the number of target genes it may regulate. The connectivity of a target gene is its indegree, and reflects the number of transcriptional regulators that bind to the regulatory region of that gene. Because of considerable noise in the data, and because of the influence of binding affinities and protein concentrations we mention above, a gene's connectivity is also best interpreted as a relative measure rather than as an absolute number. In other words, when we call a gene highly connected, we mean highly connected relative to other genes.

Duplicate genes

We identified pairs of duplicate genes in the yeast *S. cerevisiae* using a modified version of a previously published genome analysis tool called GenomeHistory (Conant and Wagner, 2002) (http://www.unm.edu/~compbio/software/GenomeHistory). This tool determines the extent of synonymous and nonsynonymous nucleotide divergence between any two sufficiently similar genes in a whole genome.

Briefly, we used GenomeHistory to carry out a three-step analysis. The first step uses gapped BLASTP (Altschul et al., '97) at an E-value threshold of 10^{-7} to identify candidates for duplicate genes in a whole genome. The second step consists of an amino acid sequence alignment for candidate genes identified in step one to determine pairs of duplicate genes. For our purpose, a global sequence alignment in this step is less than ideal to identify duplicates of transcriptional regulators. The reason is that only parts of transcriptional regulators, especially their DNA binding domains, evolve slowly and are reasonably well conserved in evolution (Ptashne, '88). Other parts, most notably transcriptional activation domains, can evolve very rapidly. The presence of rapidly evolving domains may hinder the identification of gene duplicates through global sequence alignments. This is, for example indicated by the observation that the yeast genome harbors fewer duplicates of transcriptional regulators than of other classes of genes (Conant and Wagner, 2002). For our data set, global alignment yields less than five duplicate transcriptional regulators. We thus modified GenomeHistory to carry out a local alignment, using the Smith Waterman algorithm (Smith and Waterman, '81), of candidate genes identified in the first step. Only gene pairs whose local alignment extended over at least 100 amino acids, and whose amino acid sequence was identical in more than 40% of its residues were included as gene duplicates in the final, third step of the analysis. This third step consists of a maximum likelihood estimate of the synonymous divergence (K_s) and the nonsynonymous divergence (K_a) of every pair of duplicate genes, using a method established by Yang and Nielsen (2000). Because of the well-known multiple substitution problem (Li, '97), both synonymous and nonsynonymous divergence estimates show limited reliability for $K_{a(s)} > 1.0$ respectively. Therefore, we retained only gene pairs with $K_a < 1$ for further analysis.

Orthologous genes

A recent study by Kellis and colleagues reported the genomic DNA sequences of three yeasts, *Saccharomyces mikatae, Saccharomyces paradoxus,* and *Saccharomyces bayanus,* closely related to *S. cerevisiae* (Kellis et al., 2003). From this study, we used data on synonymous divergence K_s and nonsynonymous divergence K_a between *S. cerevisiae* genes and their unambiguous orthologues from the yeast *S. mikatae* (file 'b.KaKs_details-5.xls' at http://www.broad.mit.edu/annotation/fungi/comp_yeasts/). We also used the ratio of nonsynonymous to synonymous divergence K_a/K_s averaged for orthologues in the three species pairs, *S. cerevisiae–S. bayanus, S. cerevisiae–S. paradoxus* and *S. cerevisiae–S. mikatae* (file 'b.KaKs_average.xls').

Growth rates of mutant yeast strains

We utilized results from a genome-scale experiment conducted by Steinmetz and collaborators, which assayed the growth rates of 4,706 homozygous diploid yeast deletion strains (Steinmetz et al., 2002). Briefly, the authors generated a pool containing cells from each deletion strain, and allowed cells in this pool to grow in a variety of media. These included the rich medium YPD mentioned earlier, YPDGE (0.1% glucose, 3% glycerol and 2% ethanol), YPE (2% ethanol), YPG (3% glycerol), and YPL (2% lactate). The investigators assayed the growth rate of individual strains by hybridizing DNA tags that identified each strain to an oligonucleotide microarray. The growth rate thus measured is a growth rate relative to the pool's average growth rate. We here discuss our analysis of publicly available data from one of two replicate experiments (file 'Regression_Tc1_hom.txt' at http://www-deletion.stanford.edu/YDPM/YDPM_index.html) that reported the growth of homozygous mutant strains grown in the five different media listed above. The other replicate experiment yielded qualitatively identical results (not shown). We were able to analyze 1716 genes for which both gene deletion data and transcriptional regulation data was available. We discuss results in detail for only one of the five media, YPD, because the other four media yielded qualitatively identical results (not shown). However, we also report results for a mutant's maximum growth rate difference among the five media to the pool's average growth rate (Steinmetz et al., 2002). This last measure of a gene deletion's effect indicates the greatest growth rate reduction a strain suffers in any of the five media, because most gene deletion strains with a change in growth rate suffer a reduced growth rate. In our statistical analysis of this and other data, we consider the result of any statistical test that rejects a null-hypothesis as highly significant if $P < 0.001$, and as nonsignificant if $P > 0.05$.

RESULTS

Network representation

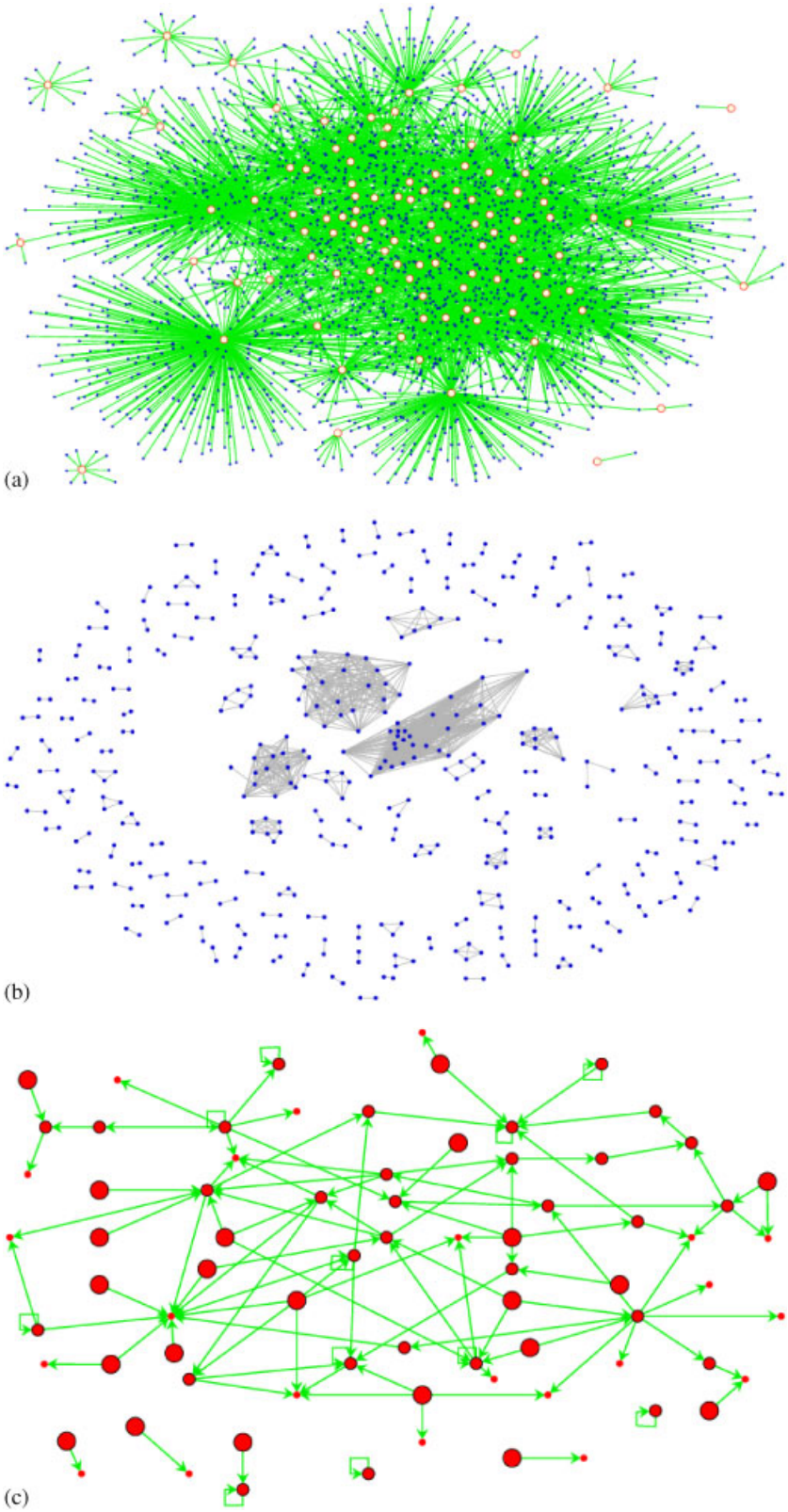
The data we use here (Lee et al., 2002) contains the binding affinity of 106 yeast transcriptional

regulators to regulatory regions of genes in the *S. cerevisiae* genome. This data can be viewed as a directed graph whose nodes are genes. A directed edge from gene x to gene y indicates that x is a transcription factor likely to regulate the expression of gene y. We will refer to the genes whose expression a transcriptional factor regulates as the regulator's target genes. The outdegree of a regulator, that is, the number of directed edges emanating from it, is the number of its target genes. The indegree of a target gene is the number of regulators that potentially influence the target gene's activity by binding to its regulatory region. Figure 1a shows the structure of this network. The majority of the 106 regulators and 2363 target genes are part of one large subgraph or component with 2925 edges. There are four regulators that are at the center of disconnected components, which involve a total of 21 target genes. The distributions of both indegrees and outdegrees have previously been characterized for transcriptional regulation networks, and we will not belabor them here (Featherstone and Broadie, 2002; Bhan et al., 2002; Guelzim et al., 2002; Lee et al., 2002). Information on gene duplications can be superimposed onto this network by introducing undirected edges into the graph: any two nodes are connected by an undirected edge, if they are the products of a gene duplication. Gene duplication is rampant in this transcriptional regulation network. For example, 27% (1688/6267) of target genes have at least one duplicate in the yeast genome. Figure 1b shows an undirected graph whose nodes correspond to duplicated target genes, where an edge between two nodes indicates that they are duplicates of each other. The vast majority of gene families in this graph contain fewer than four genes, with a few larger gene families clustered in the center. The size and complexity of the graphs in Figures 1a and 1b show that little useful information can be extracted from a mere visualization of this data. A more quantitative analysis is called for, an analysis that we will pursue below. We will separately ask similar questions of the two classes of genes—transcriptional regulatory genes and their target genes—constituting the network depicted in Figure 1a. Unfortunately, the distinction between transcriptional regulators and target genes is not clear-cut, because a regulator's expression can itself be transcriptionally regulated. Specifically, of the 106 transcriptional regulators, 50.9% (54/106) are also potentially subject to transcriptional regulation by one of the 106 regulators. We here make the choice to include transcriptional regulators regulated by other regulators in our analysis of target genes. Doing so does not materially affect our results, because transcriptional regulators that are themselves regulated constitute only 2.3% (54/2363) of target genes.

Transcriptional regulators

A majority (83 among 106 or 78%) of regulators are single-copy genes, whereas 23 regulators (22%) have at least one duplicate elsewhere in the genome. Ten transcriptional regulators constitute 5 pairs of duplicates, whereas the duplicates of the remaining 13 duplicate regulators are not among the 106 transcriptional regulators analyzed by Lee and collaborators (2002). However, the duplicates of the 13 regulators whose function has been characterized have been implicated in transcriptional regulation as well, according to information available in the Saccharomyces Genome Database SGD (http://www.yeastgenome.org/). All of the duplications are ancient, as indicated by the fact that all pairs of duplicates involving one regulator have a synonymous divergence of $K_s > 1$. This and the small number of duplicate regulators make it difficult to render a meaningful statistical analysis of functional divergence after regulator duplication.

Figure 1c shows a network representation of all the regulators that have regulatory interactions with other regulators. The majority of the regulators in the network (87% or 66/76) are contained in one large connected component. For this network, we asked whether there are any systematic differences between regulators that affect the expression of other regulators and regulators that do not. We found one such difference. Regulators that do not affect the expression of other regulators and have large numbers of target genes are underrepresented in this network (Table 1). Specifically, about half of all regulators that regulate other regulators have fewer than 50 target genes, and the other half has as many as 250 target genes. In contrast, the vast majority (96%) of other regulators have fewer than 50 target genes, with the remaining 4% having between 50 and 100 target genes. An exact binomial test shows that this difference between the two classes of regulators is highly significant ($P = 5.08 \times 10^{-13}$; $n = 51$). Among those regulators that may affect the expression of other regulators, there is another prominent statistical


(a)
(b)
(c)

trend: The higher a regulator's number of target genes, the smaller the fraction of regulators whose expression it regulates (Fig. 2). Again, this statistical association is highly significant (Kendall's $\tau = -0.50$; $P = 1.17 \times 10^{-7}$ n=54).

Connectivity and importance

The connectivity of a molecule is the result of multiple factors, such as the binding affinity to other molecules—DNA in the case of transcription factors—and a molecule's concentration in the cell. It has been argued that highly connected molecules may be more important to the functioning of a cellular network and to fitness, such that mutations—point mutations, gene deletions, or gene duplications—would have on average a more drastic fitness effect in such molecules (Albert et al., 2000; Jeong et al., 2001). We examined this hypothesis in three complementary ways, by analyzing the effects that mutations in regulators of different outdegree have on the organisms. First, has the removal of a highly connected regulator a more deleterious effect on cell growth? Figures 3a and 3b show the answer to this question, obtained from data on the growth rates of gene deletion strains in yeast transcriptional regulatory genes (Steinmetz et al., 2002). Figure 3a shows a weak negative association between a regulator's number of target genes and growth rate on rich medium.

TABLE 1. An exact binomial test to compare the binomial distribution of each column in the table

Number of Target Genes	Regulators That Regulate Regulators	Regulators That Do Not	Totals
(1, 50]	28	49	77
(50, 275]	26	2	28
Totals	54	51	105

Binomial n=54, p=26/54; Pr ($x \leq 2$)=5.08×10^{-13}.

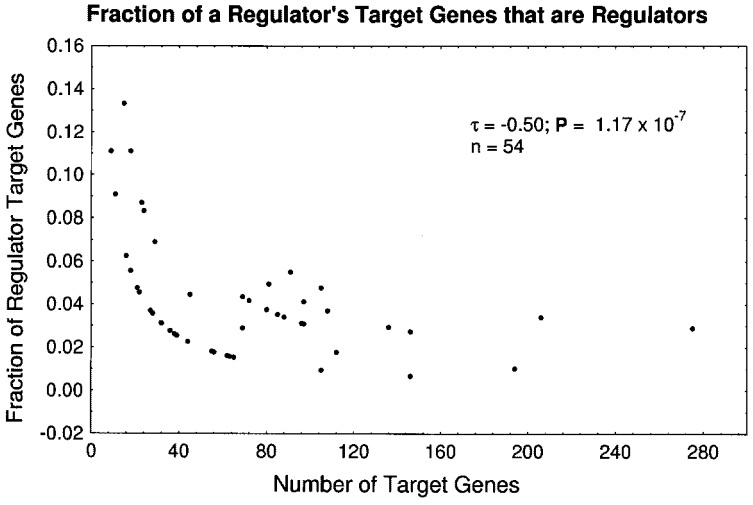


Fig. 2. Number of target genes of transcriptional regulators (horizontal axis) plotted against the fraction of a regulator's target genes that are regulators (vertical axis). (Kendall's $\tau = -0.50$; $P = 1.17 \times 10^{-7}$; n=54).

Fig. 1. a) A graph representation of the transcriptional regulation network. The large red nodes represent transcriptional regulators; the small blue nodes represent target genes; and a green edge between two nodes represents binding of the regulator to a target gene's regulatory region ($P < 0.001$ in the binding model of Lee et al. (2002)). The edges are shown as undirected solely to render the representation less cluttered. Note that all but four regulators are connected in one giant component. b) Gene duplications among target genes of transcriptional regulators. Blue nodes represent target genes. A gray edge connects two nodes if these two nodes are gene duplicates with amino acid divergence $K_a < 1.0$. c) Regulatory interactions among transcriptional regulators. All nodes in this graph represent genes encoding transcriptional regulators. An edge between two nodes represents a potential regulatory relationship between regulator and its target gene, as indicated by the regulator's binding to the target gene's regulatory region ($P < 0.001$ in the error model of Lee et al. (2002)). Three classes of transcriptional regulators are distinguished here, regulators that may influence the expression of other transcriptional regulators but are themselves not transcriptionally regulated (large red circles), regulators that regulate the expression of other regulators and are also transcriptionally regulated (medium red circles), and regulators that do not affect the expression of other transcriptional regulators (small red circles). Squares indicate autoregulation.

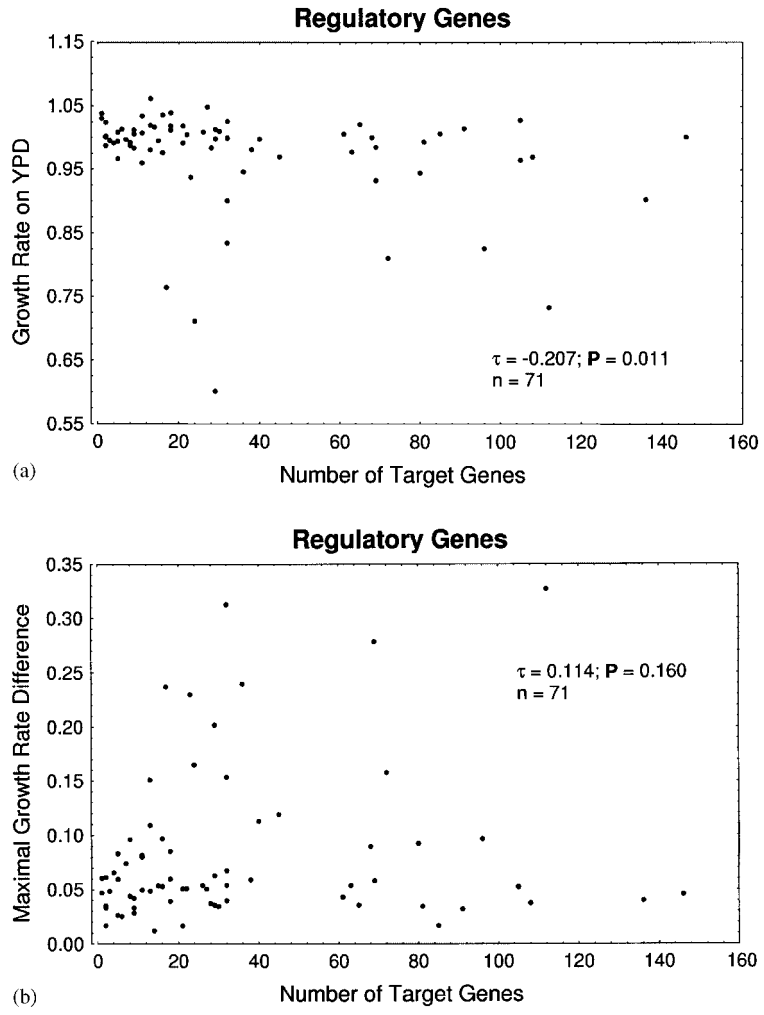


Fig. 3. Growth rate and highly connected regulators. The horizontal axis shows a regulator's number of target genes. a) The vertical axis shows the growth rate of a yeast strain with a homozygous deletion mutant in a transcriptional regulator on the rich medium YPD (Kendall's $\tau = -0.207$; $P = 0.011$; $n = 71$). The growth data is normalized to one. That is, a value of one represents no growth change in the mutant, and a value of less than one indicates slower growth. b) The vertical axis shows the maximum growth rate difference of a mutant to the pool average (Steinmetz et al., 2002) for five different growth media (Kendall's $\tau = 0.114$; $P = 0.160$; $n = 71$). A value of zero indicates that the deletion mutant grows as fast as the wild-type in all five media. The more a value differs from zero, the more the mutant's growth rate is affected in at least one medium. Because most deletions that affect growth cause a reduction in growth rate, this means that large values on the vertical axis indicate a severe growth rate reduction.

That is, deletion of highly connected transcriptional regulators leads to slightly slower growth. However, no significant association exists between a regulator's number of target genes and the maximum difference in growth rates among five different media when the regulator is eliminated (Figure 3b).

In a second attempt to address the above hypothesis, we asked whether highly connected regulators evolve more slowly, that is, whether they are under more severe evolutionary constraints? This would indicate that their encoding genes could tolerate fewer mutations. Figure 4a shows the results of an analysis addressing this question with 51 unambiguous orthologues of the regulators in the genome of the yeast, *S. mikatae,* which is closely related to *S. cerevisiae*. We plotted the ratio of nonsynonymous to synonymous divergence K_a/K_s as an indicator of evolutionary constraints (Li, '97). It shows that highly connected regulators do not evolve at rates different from other regulators (Kendall's $\tau = -0.021$; $P = 0.825$; $n = 51$). Identical results (not shown) hold for nonsynonymous divergence K_a instead of

K_a/K_s for *S. mikatae,* and also for the average ratio K_a/K_s among orthologues in the three species pairs *S. cerevisiae–S. bayanus, S. cerevisiae–S. paradoxus* and *S. cerevisiae*–S. mikatae.

Third, are regulators with many target genes less likely to have undergone a gene duplication sometime in the past? The answer is contained in Table 2, where we categorized regulators by the number of their target genes. Eighty-nine of the 106 regulators (84%) have 80 or fewer target genes, and 17 regulators (16%) have more than 80 target genes. An exact binomial test indicates that single-copy genes are not underrepresented among highly connected regulators (P=0.060; n=83). In other words, high connectivity does not reduce the likelihood that a regulator's duplicate is preserved in the evolutionary record. The converse question is whether a regulator's connectivity may not only influence its own likelihood to undergo duplication, but also the likelihood that any of its target genes undergoes duplication without deleterious effects. We thus asked whether there is a correlation between a regulator's number of target genes and the fraction of these target genes that have undergone duplication. Figure 5 shows that the answer is no (Kendall's τ=0.104; P=0.114; n=105).

Target genes

Just as we did for regulators, we asked, in three complementary ways, whether target genes with high connectivity (indegree) have different propensity to suffer deleterious mutations. First, has the removal of a highly connected target gene a more deleterious effect on cell growth? Figures 6a and 6b show the answer to this question, obtained from data on the growth rates of gene deletion strains in yeast transcriptional regulatory genes (Steinmetz et al., 2002). Figure 6a shows that there is no statistically significant association between indegree and growth rate on rich medium. The same holds for Figure 6b, which uses the difference between indegree and maximum difference in growth rate among five different media as an indicator of deletion effect. However, it is noteworthy that the figure indicates a negative association between the *maximal* reduction in growth rate on rich medium for any gene of a given indegree (Figure 6a), as well as a negative association between the *maximal* difference in growth rate among five media and indegree (Figure 6b). In other words, the maximal effect of a gene deletion decreases with target gene connectivity.

Second, do highly connected target genes, target genes whose expression is influenced by many regulators, evolve more slowly? That is, are they under more severe evolutionary constraints? This would indicate that their encoding genes could tolerate fewer mutations. Figure 4b shows the results of an analysis addressing this question with 772 unambiguous orthologues of the target genes in the genome of the yeast, *S. mikatae,* which is closely related to *S. cerevisiae*. We plotted the ratio of nonsynonymous to synonymous divergence K_a/K_s as an indicator of evolutionary constraints (Li, '97). It shows that highly connected target genes do not evolve at different rates from other target genes (Kendall's τ=0.026; P=0.285; n=772). Identical results (not shown) hold for nonsynonymous divergence K_a instead of K_a/K_s for *S. mikatae*, as for the average ratio K_a/K_s for orthologous in the 3 species pairs (*S. bayanus, S. paradoxus,* and *S. mikatae*). All results show that highly connected target genes do not evolve at different rates from other target genes.

Third and finally, are highly connected target genes less likely to have undergone gene duplications sometime in the past? The answer is contained in Table 3, where we categorized target genes by the number of their regulators. Out of 2363 target genes, 2328 or (98.5%) have seven or fewer regulators, and 35 target genes have more than 7 regulators. An exact binomial test indicates that there are fewer duplicated highly connected target genes than single-copy highly connected target genes (P=1.9 x 10^{-8}; n=1871). In other words, high connectivity may reduce the likelihood that a regulator's duplicate is preserved in the evolutionary record. The converse question is whether the regulators of highly connected target genes show different propensity to undergo gene duplication. Figure 7 shows the indegree of a target gene plotted against the fraction of its regulators that have at least one duplicate in the yeast genome. The association is weak (Kendall's τ=0.149) but highly significant (P=2.1×10^{-27}; n=2364), showing that the regulators of highly connected target genes are slightly more likely to undergo gene duplication.

Divergence after gene duplication

Finally, there is the question about the rate and extent of functional divergence after gene

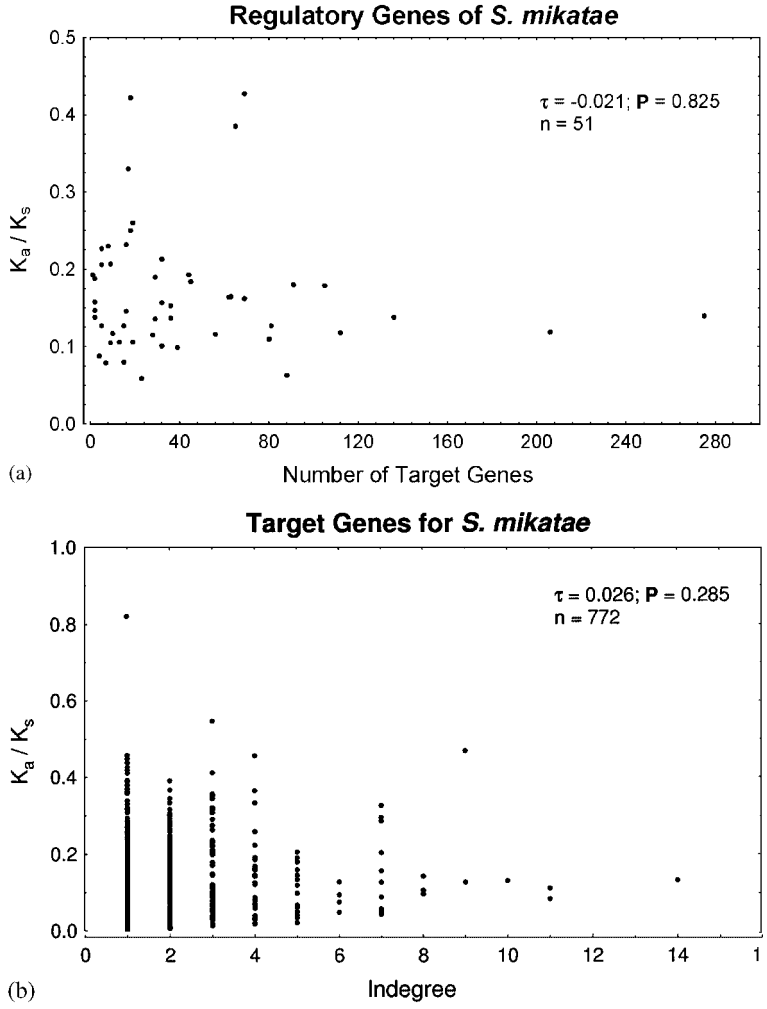


Fig. 4. Do highly connected genes evolve at different rates? a) Regulatory genes. The horizontal axis shows a regulator's outdegree, that is, its number of target genes. The vertical axis shows the ratio K_a/K_s of nonsynonymous to synonymous divergence of the regulatory gene to an unambiguous orthologue in a closely related yeast, *S. mikatae* (Kellis et al., 2003). No significant statistical association is observed (Kendall's $\tau = -0.021$; P=0.825; n=51). b) Target genes. The horizontal axis shows a target gene's indegree, that is, the number of regulators that bind to its regulatory region. The vertical axis shows the average of the ratio K_a/K_s of nonsynonymous to synonymous divergence of the target gene to an unambiguous orthologue in a closely related yeast *S. mikatae*. No significant statistical association is observed (Kendall's τ=0.026; P=0.285; n=772).

duplication. We could not address this question for the transcriptional regulators, because of their small numbers, but we can address it for target genes. The proportion of regulators shared by two target genes can serve as a proxy of their similarity in expression regulation, which is one among several indicators of gene function. We are well aware that two genes with similar expression patterns may have different transcriptional regulators, and vice versa. However, there must be at least some statistical association between two genes' expression similarity and their similarity in the regulators bound to them. Otherwise highly successful approaches to identify regulatory DNA sequences through a combination of DNA sequence and gene expression analysis would not work (Bussemaker et al., 2001).

We determined for every pair of duplicate target genes T_1 and T_2, the number d_1 of regulators binding to the regulatory region of T_1, the number d_2 of regulators binding to the regulatory region of T_2, as well as the number d_{12} of regulators binding both target regulatory regions. The fraction of shared regulators is then properly defined as $d_{12}/(d_1+d_2-d_{12})$. Figures 8a and 8b show this fraction of shared regulators

as a function of the nonsynonymous divergence (K_a) and synonymous or silent divergence (K_s), respectively, between duplicate target genes. The solid line in both panels indicates the average fraction of shared regulators (0.02) between any two randomly chosen target genes in the network. The dotted line indicates the average fraction of shared regulators plus one standard deviation (0.02+0.14=0.16) between any two randomly chosen target genes in the network. Both panels show a highly significant negative association between sequence divergence and the fraction of shared regulators. In addition, it is evident that many duplicate target gene pairs with high sequence similarity have diverged completely in the regulators bound to them. In fact, the statistical association we observe is largely due to an increasing number of duplicates with no shared regulators as duplicates diverge. The statistical association we observe here and the large number of duplicates with no shared regulators is not the result of a conservative binding threshold ($P<0.001$) we used in this analysis. We observe it also for greatly relaxed binding thresholds ($P<0.05$) (results not shown). In sum, divergence after duplication is often rapid.

TABLE 2. An exact binomial test to compare the binomial distribution of each column in the table

Number of Target Genes	Duplicate Regulators	Single Copy Regulators	Totals
(1, 80]	16	73	89
(80, 275]	7	10	17
Totals	23	83	106

Binomial n=23, p=7/23; Pr (x ≥10)=0.06.

A very similar approach allowed us to ask whether duplicate target genes diverge largely through loss of transcriptional regulator binding in one of the genes. This is what recent models of gene divergence emphasizing subfunctionalization of genes suggest (Force et al., '99a). Conversely, it is possible that divergence evolves through the addition of many new transcriptional regulation interactions. Immediately after a gene duplication, if both the coding and the regulatory region are duplicated, the sum of the number of transcription factors binding to both duplicates' regulatory regions is $d_1+d_2=2d$, where d is the number of transcriptional regulators bound to the ancestral gene (before duplication). If divergence occurs only through loss of binding sites, then d_1+d_2 will decrease after duplication and approach $d_1+d_2=d$, the number of binding interactions before duplication. Conversely, if divergence involved largely addition of new interactions, then d_1+d_2 should increase after duplication. Figure 9a clearly shows that the second scenario is not the case: d_1+d_2 decreases after duplication.

Does this mean that only loss of binding sites occurs during divergence? No. It only means that there is a net loss of binding sites during divergence after duplication. To assess whether gain of binding sites is important, we carried out a second analysis, where we focused only on those duplicate gene pairs that have completely diverged, that is, $d_{12}=0$ so gene pairs share no

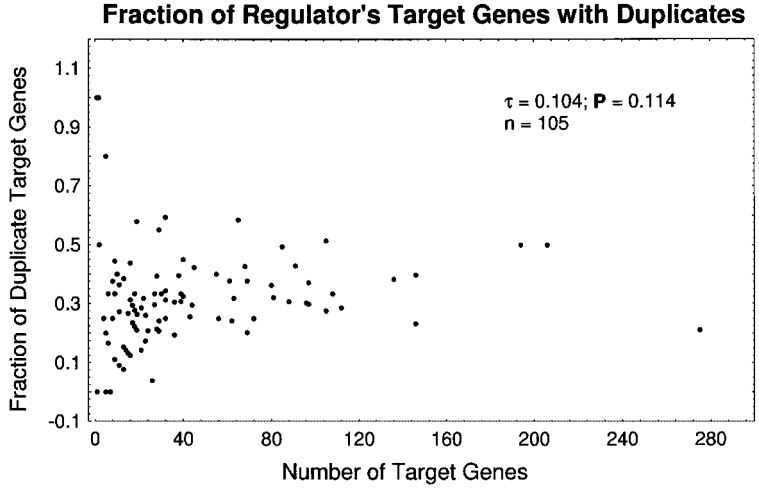


Fig. 5. No significant association exists between a regulator's number of target genes (horizontal axis), and the fraction of target genes that have undergone duplication (vertical axis).

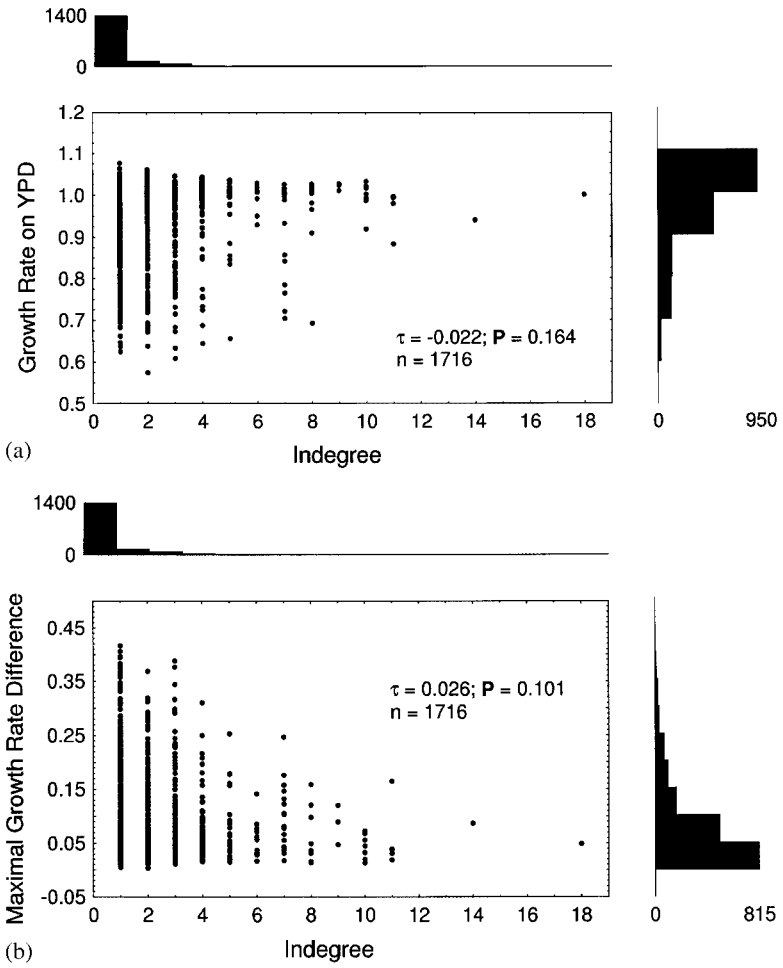


Fig. 6. Growth rate and highly connected target genes. The horizontal axis shows the number of regulators that bind to the regulatory region of a target gene. a) The vertical axis shows the growth rate on YPD medium of a yeast strain with a homozygous deletion mutant in a target gene (Kendall's $\tau = -0.022$; P=0.164; n=1716). The growth data is normalized to one. That is, a value of one represents no growth change in the mutant, and a value of less than one indicates slower growth. b) The vertical axis shows the maximum growth rate difference of a mutant to the pool average (Steinmetz et al., 2002) for five different growth media (Kendall's τ=0.026; P=0.101; n=1716). A value of zero indicates that the deletion mutant grows as fast as the wild-type in all five media. The more a value differs from zero, the more the mutant's growth rate is affected in at least one medium. Because most deletions that affect growth cause a reduction in growth rate, this means that large values on the ordinate axis indicate a severe growth rate reduction.

TABLE 3. An exact binomial test to compare the binomial distribution of each column in the table

In Degree	Duplicate Target Genes	Single Copy Target Genes	Totals
(1, 7]	487	1841	2328
(7, 18]	5	30	35
Totals	492	1871	2363

Binomial n=1871, p= 30/1871; Pr ($x \leq 5$)=1.90×10^{-8}.

transcriptional regulators. If loss of transcription factor binding sites is exclusively responsible for the divergence of duplicates, then the combined degrees d_1+d_2 of completely diverged duplicates should be identical to the degree d typically found in single-copy genes. Figure 9b shows that this is not the case, regardless of whether one examines very young ($K_s<0.25$) or older duplicates. Completely diverged duplicate genes always show a combined degree significantly higher than single-copy genes, which demonstrates that gain of transcription factor binding sites plays a significant role in their divergence.

DISCUSSION

Our primary focus here was a descriptive analysis of the largest available genome-scale

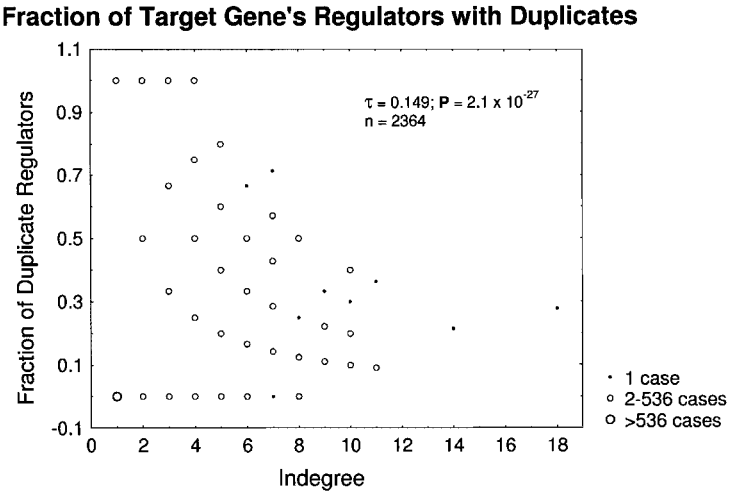


Fig. 7. Regulators of highly connected target genes are more likely to undergo gene duplication (Kendall's τ=0.149; P=0.210×10^{-27};n=2364). The horizontal axis shows the indegree of a target gene, i.e. the number of regulators bound to its regulatory region. The vertical axis shows the fraction of a target gene's regulators that have undergone at least one gene duplication. The vast majority of genes have only one potential regulator. For the majority of the remaining target genes, the fraction of duplicate regulators is smaller than 0.1, in line with the observation that most regulators are encoded by single-copy genes.

experimental data set on the yeast transcriptional regulation network, with an emphasis on how the connectivity of a gene in the network can influence its molecular evolution. In such an analysis, it is expedient to distinguish two classes of genes: regulators and their target genes. Doing so, however, has a disadvantage: there are many fewer regulators than target genes, rendering their statistical analysis more difficult. The problem is aggravated for one important class of mutations that affect a network's structure, gene duplications. All duplications of transcriptional regulators in yeast are ancient, and transcriptional regulators have few gene duplicates when compared with other classes of genes. The latter pattern has been observed previously in an analysis that used global sequence alignment to identify duplicate genes (Conant and Wagner, 2002). Because some domains of transcriptional regulators—especially their DNA binding domains—evolve slowly, whereas other domains evolve rapidly, global sequence alignments may miss duplicate regulators. However, the underabundance of duplicate regulators does not disappear when we use local instead of global sequence alignment to circumvent this problem. For instance, we found here that 27% of target genes have duplicates, whereas only 22% of regulators do. This indicates that duplication of transcriptional regulators has been less prevalent than duplication of target genes in the evolution of the yeast transcriptional regulation network. This paucity of gene duplication in transcriptional regulation gene may be specific to yeasts, because it is not observed in the fruit fly *Drosophila melanogaster* or in the worm *Caenorhabditis elegans* (Conant and Wagner, 2002). It may thus be a peculiarity of the evolutionary history of yeasts rather than a general feature of transcriptional regulation networks. If this is the case, then yeast may not be the best species for this type of study. For the data available and for our purpose it means that we have very limited data to examine the role duplications of regulatory genes have played in this network's evolution.

Caveats

The analysis we carried out here has several caveats. The first of them is that the nature of the experiment limits transcriptional regulators to DNA-binding proteins. However, it is increasingly appreciated that transcriptional regulation in eukaryotes involves large multiprotein complexes, not all of whose members contact DNA (Ptashne and Gann, 2002). Second, the experiments will preferentially identify regulation of genes expressed in rich medium. Thirdly, the data set of 106 transcriptional regulators does not include all transcriptional regulators in yeast. Lastly, the binding of a transcription factor to a target gene's promoter region is indicative but not conclusive of transcriptional regulation.

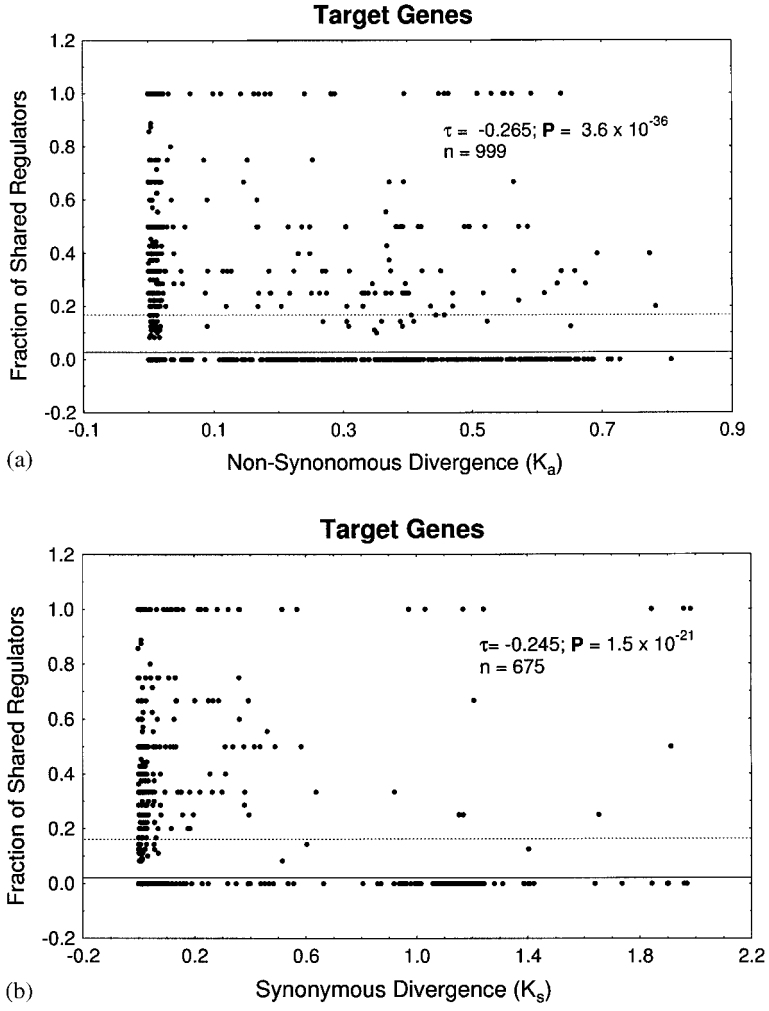


Fig. 8. Negative association between sequence divergence and regulators shared by duplicate target genes. The vertical axes show the fraction of transcriptional regulators bound to both regulatory regions of a duplicate gene pair. The solid lines in both panels indicate the average fraction of shared regulators (0.02) between two randomly chosen target genes in the network. The dotted lines indicate the average fraction of shared regulators plus one standard deviation (0.02 + 0.14=0.16) between any two randomly chosen target genes in the network. These lines are based on one thousand randomly chosen target gene pairs. a) Sequence divergence as measured by the nonsynonymous divergence K_a. (Kendall's $\tau = -0.265$; $P = 3.60 \times 10^{-36}$; n=999). b) Sequence divergence as measured by the synonymous divergence K_s. (Kendall's $\tau = -0.245$; $P = 1.50 \times 10^{-21}$; n=675).

We nevertheless chose to work with this data because it represents by far the largest unbiased body of information on potential transcriptional regulation. Other data sets (Perez-Rueda and Collado-Vides, 2000; Bhan et al., 2002; Guelzim et al., 2002; Lee et al., 2002; Salgado et al., 2004) are not only significantly smaller, they also have other shortcomings, most prominently an ascertainment bias of unknown magnitude that could distort results in unknown ways. We are, however, aware that our results are preliminary and await confirmation through improved experimental data.

Gene connectivity and importance

A prominent hypothesis in the study of biological networks suggests that highly connected molecules are more important to the network, in the sense that the network's global structure—and hence its function—is most severely impaired when such molecules suffer mutations (Albert et al., 2000; Jeong et al., 2001). To begin with, how does one best think of connectivity? Much genome-scale data on molecular networks identifies two molecules as either interacting or not interacting. However, the association of two

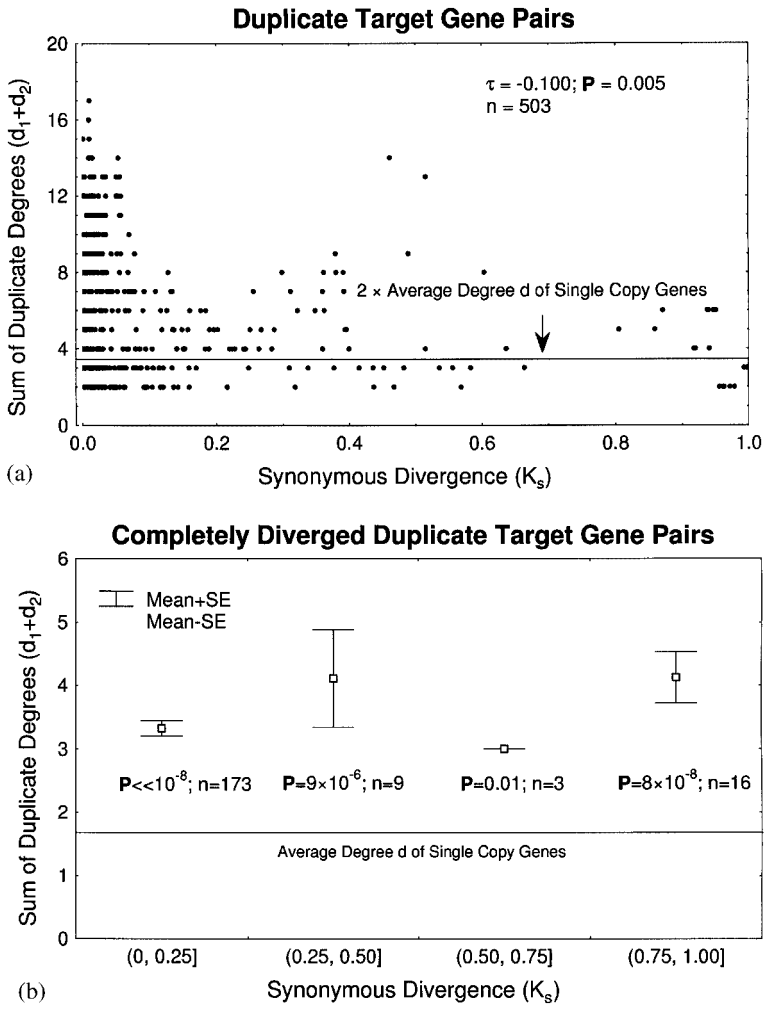


Fig. 9. Sequence divergence and divergence of the number of regulators affecting duplicate target genes. The horizontal axis indicates synonymous sequence divergence K_s between duplicate target genes. a) The vertical axis indicates the sum d_1+d_2 of the number of transcriptional regulators binding to regulatory regions of two duplicate target genes. The solid horizontal line indicates $2d$,where d is the average number of regulators binding to the regulatory region of single-copy genes. Standard errors for d are too close to the mean to be visible in the plot. The number of regulators binding to two duplicate target genes declines with synonymous divergence (Kendall's $\tau = -0.100$; P=0.005; n=503). b) Includes only duplicate target gene pairs that have completely diverged since their duplication, i.e., gene pairs where $d_{12}=0$. Gene pairs are grouped in four bins according to their synonymous divergence. We tested the null hypothesis that the sum of the degrees of completely diverged duplicates is identical to the degree d of single-copy genes using a Mann-Whitney U-test (Sokal and Rohlf, '81). The null-hypothesis is rejected for all four bins examined. This indicates that gain of transcriptional regulation interactions plays a significant role in functional divergence of duplicate target genes.

molecules in a cell is governed by thermodynamic principles. It is influenced by parameters such as dissociation constants and a molecule's concentration in the cell. Proteins have widely varying binding affinities to each other, and widely varying concentrations in the cell. Similarly, transcriptional regulators have widely varying binding affinities to their sites on DNA and widely varying concentrations. Any qualitative data on molecular interactions, such as available genome-scale protein interaction data, captures such variation poorly. The problem is alleviated with the semi-quantitative data that we use here, because this data reflects the confidence one has in the binding of a factor to a regulatory region. However, this data cannot disentangle the effects of concentration and binding affinity. The total connectivity of a transcriptional regulator—its outdegree—should thus be understood as a composite variable influenced

by binding affinities and transcription factor concentrations. It is with this qualification—which holds for all current analyses of molecular interaction networks—that our results should be interpreted.

The hypothesis that connectivity relates to a molecule's importance has been mostly explored with protein interaction networks, with conflicting results (Jeong et al., 2001; Fraser et al., 2002; Fraser et al., 2003; Jordan et al., 2003a; Jordan et al., 2003b; Hahn et al., 2004). The disadvantage of protein interaction data is that such data contain an especially large amount of experimental noise (von Mering et al., 2002; Gilchrist et al., 2004), and that the biological significance of two proteins' interaction is not always clear. In contrast, transcriptional regulation interactions have a clear interpretation: transcription factors regulate genes whose expression is necessary for biological processes. The notion that highly connected regulators are functionally more constrained than other regulators, because they may affect the expression of more target genes, is therefore especially plausible for transcriptional regulation networks.

To address this hypothesis, we first examined whether deletion of highly connected regulators causes more severe growth reduction in yeast. We found a weak statistical association supporting this notion on the rich medium YPD. The problem with interpreting this kind of result is that the growth reduction of a mutant may depend on the growth medium used. So we also asked whether a statistical association exists between a regulator's number of target genes, and the maximal growth defect observed in five different growth media. The statistical association observed in YPD disappeared in this analysis.

A major problem with this type of analysis, in addition to the environmental dependence of mutational effects, is that growth rate reductions much smaller than observable in the laboratory may affect a microbe's fitness, and that a microbe's fitness is not only determined by its growth rate. A complementary analysis thus asks whether highly connected regulators are under more severe evolutionary constraints, in that fewer amino acid changes are preserved in their evolutionary record. To this end, we compared *S. cerevisiae* transcriptional regulators to their orthologues in the closely related yeast *S. mikatae*. We found that regulators with many target genes do not evolve more slowly than other regulators.

Gene duplications are a third class of mutations—aside from gene deletions and point mutations—that may affect network function. A gene duplication can cause an increase in expression of a transcriptional regulator, which may affect the expression of target genes, especially if these target genes are regulated jointly with other regulators. It may be the case that highly connected regulators are less likely to undergo duplications that have been preserved in the evolutionary record. However, we did not observe any such trend. In sum, three independent lines of evidence suggest that the connection between a transcriptional regulator's high connectivity and the network's sensitivity to changes in it is tenuous to nonexistent.

An analogous question can be asked for the target genes of transcriptional regulators instead of the regulators themselves. A highly connected target gene is a target gene to whose regulatory regions many regulators bind. Some such target genes may be combinatorially regulated, whereas others may function in different biological processes, and different regulators may thus regulate their expression at different times. Because of their potential involvement in multiple processes, some highly connected target genes may also be more susceptible to mutations. We find, however, that deletion of highly connected target genes does not generally lead to slower growth. In addition, and contrary to what one might expect, highly connected target genes may evolve slightly faster than other target genes. Only gene duplications show a semblance of the expected pattern: duplicate genes are slightly less abundant among highly connected genes. Taken together, these three lines of evidence show that there is no strong and consistent support for an association between gene connectivity and an organism's ability to tolerate genetic changes in the gene.

Divergence after gene duplication

One question that an analysis of gene networks can address is how gene duplicates diverge in function. This question has two facets, the first of which we already mentioned in the introduction: how rapidly do two genes diverge in their functions? Other studies suggest that indicators of functional similarity among duplicate genes show a highly significant but only weak statistical association with sequence divergence or duplication age. This has been observed for similarity in

gene expression (Wagner, 2000; Gu et al., 2002) and similarity in protein interactions (Wagner, 2001). Our analysis of duplicate target genes of transcriptional regulators confirms this observation. Specifically, the fraction of regulators shared by two duplicate target genes, that is, the fraction of regulators that bind to the regulatory regions of both genes, decreases with the amino acid sequence divergence of the duplicates, as has been observed also by others (Maslov et al., 2004). It also decreases with the divergence of the duplicates at synonymous (silent) sites, an indicator of a gene duplication's age. These statistical associations, although highly significant, are weak. Part of the reason is that even highly similar or recently arisen gene duplicates can have diverged considerably in the regulators bound to them. In other words, divergence in gene regulation after duplication is often rapid.

A second facet of the above question regards the mode of functional divergence after gene duplication. A prominent hypothesis emphasizes the importance of losing some of a gene's functions after duplication, in order for both duplicates to be preserved (Force et al., '99b; Lynch and Force, 2000). Many genes have multiple functions, and when a multifunctional gene becomes duplicated, either duplicate can lose one or more of these functions, as long as they are preserved in the other duplicate. Through selective loss of functions, both duplicates are rendered essential and can no longer be eliminated from the genome. Supporting evidence for this mode of divergence has come from studies of mutational effects in duplicate genes, and from expression studies of duplicate genes in higher organisms, (reviewed in Prince and Pickett, 2002). In gene expression studies, for example, duplicate genes sometimes show a mode of expression restricted to a subset of the expression domains of their ancestral single-copy gene in a related organism. A second mode of divergence that can render one or both duplicates essential is neofunctionalization, the acquisition of new functions. Because degenerative mutations that eliminate transcription factor binding and thus potentially gene expression may be more abundant than mutations that lead to new functions, subfunctionalization might be a much more important mode of divergence than neofunctionalization. However, our analysis here indicates that both modes of divergence play a role. On one hand, gene duplicates experience a net loss in the number of transcription factors binding to them. On the other hand, the number of transcription factors that bind to completely diverged duplicates is significantly greater than expected if loss of binding is solely responsible for the divergence of duplicate genes. With the benefit of hindsight, the importance of neofunctionalization may not be all that surprising. Recent work has shown that new transcriptional regulation interactions can evolve very rapidly in large microbial populations (Stone and Wray, 2001). Part of the reason is that binding sites for transcriptional regulators are short, and that they can often arise by chance alone (Stone and Wray, 2001). In addition, population genetic theory shows that genetic drift, which is necessary for the process of subfunctionalization, is weakest in the large populations of typical microbes, which would render neofunctionalization more prominent in yeast (Force et al., '99b; Lynch and Force, 2000).

Regulators of regulators

Despite the small numbers of transcriptional regulators in this network, we were able to make some intriguing although currently unexplained observations about these regulators. One of them is that regulators which regulate the expression of other regulators tend to have more target genes overall. It would be tempting to call such regulators master regulators. However, the expression of such highly connected regulators is also influenced by other, less highly connected regulators. Thus, when faced with the full complexity of regulatory gene networks, a naive distinction between master regulators and other regulators may be unhelpful in understanding network structure.

A second observation is that regulators with many target genes tend to regulate the expression of a smaller fraction of other regulators than regulators with fewer target genes. There is one obvious candidate explanation for this finding: Mutations in highly connected regulators may have strong pleiotropic effects. A mutation in such regulators may affect the expression of many target genes, and is more likely to be deleterious than a mutation in a less highly connected regulator. If such a mutation affects the expression of another regulator, together with the expression of this regulator's target genes, the likelihood that the mutation is deleterious may be even greater. Highly connected regulators may thus benefit from a reduction in the number of other regulators they regulate. Despite the

plausibility of this argument, our analysis of the relation between connectivity of regulators and their importance to the network does not support it. There is at best a weak link between a regulator's number of target genes and the effects of mutations in the regulator on the organism. In sum, we currently do not have a functional explanation for either of these regulatory patterns.

CONCLUSIONS

Answering questions about the evolutionary forces that affect genetic networks might be helpful in closing the gap between our understanding of biology at the molecular and organismal level of organization. The study we present here shows how much work remains to be done. So far, only the most basic associations between a gene's connectivity and its evolution have been explored. Our study is no exception. The available work does not even allow us to exclude the possibility that the large-scale structure of regulatory networks has little biological significance, and that only small-scale scale network features may be truly of biological importance (Milo et al., 2002; Shen-Orr et al., 2002; Conant and Wagner, 2003). Even basic regulatory patterns, such as those in the preceding two paragraphs, currently do not have a place in a larger understanding of network structure. Not only new data but also new hypotheses will be necessary to assess whether the large-scale structure of biological networks really provides a bridge between molecules and organisms.

ACKNOWLEDGMENTS

We are very grateful to G. Conant for providing us with data on duplicate genes based on local sequence alignments. A.M.E. is supported by The Department of Energy's Computational Science Graduate Fellowship, administered by the Krell Institute. A.W. would like to acknowledge support through NIH grant GM 63882, as well as the continuing support of the Santa Fe Institute.

LITERATURE CITED

Albert R, Jeong H, Barabasi A. 2000. Error and attack tolerance of complex networks. Nature 406:378–382.

Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D. 1997. Gapped Blast and Psi—Blast : A new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin G, Sherlock G. 2000. Gene ontology: Tool for the unification of biology. Nat Genet 25:25–29.

Bhan A, Galas D, Dewey T. 2002. A duplication growth model of gene expression networks. Bioinformatics 18: 1486–1493.

Briscoe A. 2001. Functional diversification of lepidopteran opsins following gene duplication. Mol Cell Biol 18:2270–2279.

Bussemaker H, Li H, Siggia E. 2001. Regulatory element detection using correlation with expression. Nat Genet 27:167–171.

Chaput M, Claes V, Portetelle D, Cludts I, Cravador A, Aburny A, Gras H, Tartar A. 1988. The neurotrophic factor neuroleukin is 90 percent homologous with phosphohexose isomerase. Nature 332:454–455.

Chung F, Lu L, Dewey T, Galas D. 2003. Duplication models for biological networks. J Comput Biol 10:677–687.

Conant G, Wagner A. 2002. GenomeHistory: A software tool and its application to fully sequenced genomes. Nucleic Acids Res 30:3378–3386.

Conant G, Wagner A. 2003. Convergent evolution in gene circuits. Nat Genet 34:264–266.

Conant G, Wagner A. 2004. Duplicate genes and robustness to transient gene knock–downs in Caenorhabditis elegans. P Roy Soc Ser B-Bio 271:89–96.

Costanzo M, Hogan J, Cusick M, Davis B, Fancher A, Hodges P, Kondu P, Lengieza C, Lew-Smith J, Lingner C, Roberg-Perez K, Tillberg M, Brooks J, Garrels J. 2000. The Yeast Proteome Database (YPD) and *Caenorhabditis elegans* Proteome Database (WormPD): Comprehensive resources for the organization and comparison of model organism protein information. Nucleic Acids Res 28:73–76.

Faik P, Walker J, Redmill A, Morgan M. 1988. Mouse glucose-6-phosphate isomerase and neuroleukin have identical 3' sequences. Nature 332:455–456.

Featherstone D, Broadie K. 2002. Wrestling with pleiotropy: Genomic and topological analysis of the yeast gene expression network. BioEssays 24:267–274.

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

Force A, Lynch M, Postlethwait J. 1999b. Preservation of duplicate genes by subfunctionalization. Am Zool 39:78A.

Fraser H, Hirsh A, Steinmetz L, Scharfe C, Feldman M. 2002. Evolutionary rate in the protein interaction network. Science 296:750–752.

Fraser H, Wall D, Hirsh A. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol 3:11.

Furukawa T, Yoshimura A, Sumizawa T, Haraguchi M, Akiyama S, Fukui K, Ishizawa M, Yamada Y. 1992. Angiogenic factor. Nature 356:668.

Gilchrist M, Salter L, Wagner A. 2004. A statistical framework for combining and interpreting proteomic data sets. Bioinformatics (in press).

Gonzalez-Gaitan M, Rothe M, Wimmer E, Taubert H, Jackle H. 1994. Redundant functions of the genes knirps and knirps-related for the establishment of anterior

Drosophila head structures. Proc Natl Acad Sci USA 91:8567–8571.

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

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

Guelzim N, Bottani S, Bourgine P, Kepes F. 2002. Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet 31(1):60–63.

Hahn M, Conant G, Wagner A. 2004. Molecular evolution in large genetic networks: Does connectivity equal constraint? J Mol Biol 58:203–211.

Haraguchi M, Miyadera K, Uemura K, Sumizawa T, Furukawa T, Yamada K, Akiyama S, Yamada Y. 1994. Angiogenic activity of enzymes. Nature 368:198.

Hughes A. 1994. The evolution of functionally novel proteins after gene duplication. P Roy Soc Ser B-Bio 256: 119–124.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 98:4569–4574.

Jeffery C. 1999. Moonlighting proteins. Trends Biochem Sci 24:8–11.

Jeong H, Mason S, Barabasi A, Oltvai Z. 2001. Lethality and centrality in protein networks. Nature 411:41–42.

Jordan I, Wolf Y, Koonin E. 2003a. Correction: no simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors evolve slowly. BMC Evol Biol 3:5.

Jordan I, Wolf Y, Koonin E. 2003b. No simple dependence between protein evolution rate and the number of protein-protein interactions: Only the most prolific interactors tend to evolve slowly. BMC Evol Biol 3:5.

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

Kennedy M, Mendemueller L, Blondin G, Beinert H. 1992. Purification and characterization of cytosolic aconitase from beef-liver and its relationship to the iron-responsive element binding-protein. Proc Natl Acad Sci USA 89: 11730–11734.

Knop M, Siegers K, Pereira G, Zachariae W, Winsor B, Nasmyth K, Schiebel E. 1999. Epitope tagging of yeast genes using a PCR-based strategy: More tags and improved practical routines. Yeast 15:963–972.

Lee T, Rinaldi N, Robert F, Odom D, Bar-Joseph Z, Gerber G, Hannett N, Harbison C, Thompson C, Simon I, Zeitlinger J, Jennings E, Murray H, Gordon D, Ren B, Wyrick J, Tagne J, Volkert T, Fraenkel E, Gifford D, Young R. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298:799–804.

Li W-H. 1997. Molecular Evolution. Massachusetts: Sinauer.

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

Maslov S, Sneppen K, Eriksen K, Yan K-K. 2004. Upstream Plasticity and Downstream Robustness in Evolution of Molecular Networks. BMC Evol Biol 4:9.

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. 2002. Network motifs: Simple building blocks of complex networks. Science 298:824–827.

Nowak MA, Boerlijst MC, Cooke J, Maynard-Smith J. 1997. Evolution of genetic redundancy. Nature 388: 167–171.

Perez-Rueda E, Collado-Vides J. 2000. The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. Nucleic Acids Res 28:1838–1847.

Prince V, Pickett F. 2002. Splitting pairs: The diverging fates of duplicated genes. Nat Rev Genet 3:827–837.

Ptashne M. 1988. How eukaryotic transcriptional activators work. Nature 335:683–689.

Ptashne M, Gann A. 2002. Genes and Signals. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Ren B, Robert F, Wyrick J, Aparicio O, Jennings E, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert T, Wilson C, Bell S, Young R. 2000. Genome-wide location and function of DNA binding proteins. Science 290:2306–2309.

Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V, Santos-Zavaleta A, Bonavides-Martinez C, Collado-Vides J. 2004. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. Nucleic Acids Res 32:D303–D306.

Shen-Orr S, Milo R, Mangan S, Alon U. 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31:64–68.

Smith T, Waterman M. 1981. Identification of common molecular subsequences. J Mol Biol 147:195–197.

Sokal R, Rohlf F. 1981. Biometry. New York: Freeman.

Sole R, Pastor-Satorras R, Smith ED, Kepler T. 2002. A model of large-scale proteome evolution. Adv Complex Syst 5: 43–54.

Steinmetz L, Scharfe C, Deutschbauer A, Mokranjac D, Herman Z, Jones T, Chu A, Giaever G, Prokisch H, Oefner P, Davis R. 2002. Systematic screen for human disease genes in yeast. Nat Genet 31:400–404.

Stone J, Wray G. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol 18: 1764–1770.

Uetz P, Giot L, Cagney G, Mansfield T, Judson R, Knight J, Lockshon D, Narayan V, Srinivasan M, Pochart P, QureshiEmili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg J. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature 403: 623–627.

van Noort V, Snel B, Huynen M. 2004. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. Embo reports (in press).

von Mering C, Krause R, Snel B, Cornell M, Oliver S, Fields S, Bork P. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417:399–403.

Wagner A. 1999. Redundant gene functions and natural selection. J Evol Biol 12:1–16.

Wagner A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: Implications for the neutralist-selectionist debate. Proc Natl Acad Sci USA 97(12):6579–6584.

Wagner A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol Evol 18:1283–1292.

Wagner A. 2003. How the global structure of protein interaction networks evolves. Proc Roy Soc Ser B-Bio 270: 457–466.

Wang Y, Schnegelsberg P, Dausman J, Jaenisch R. 1996. Functional redundancy of the muscle-specific transcription factors Myf5 and myogenin. Nature 379:823–825.

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

Zhang J, Rosenberg H, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA 95:3708–3713.