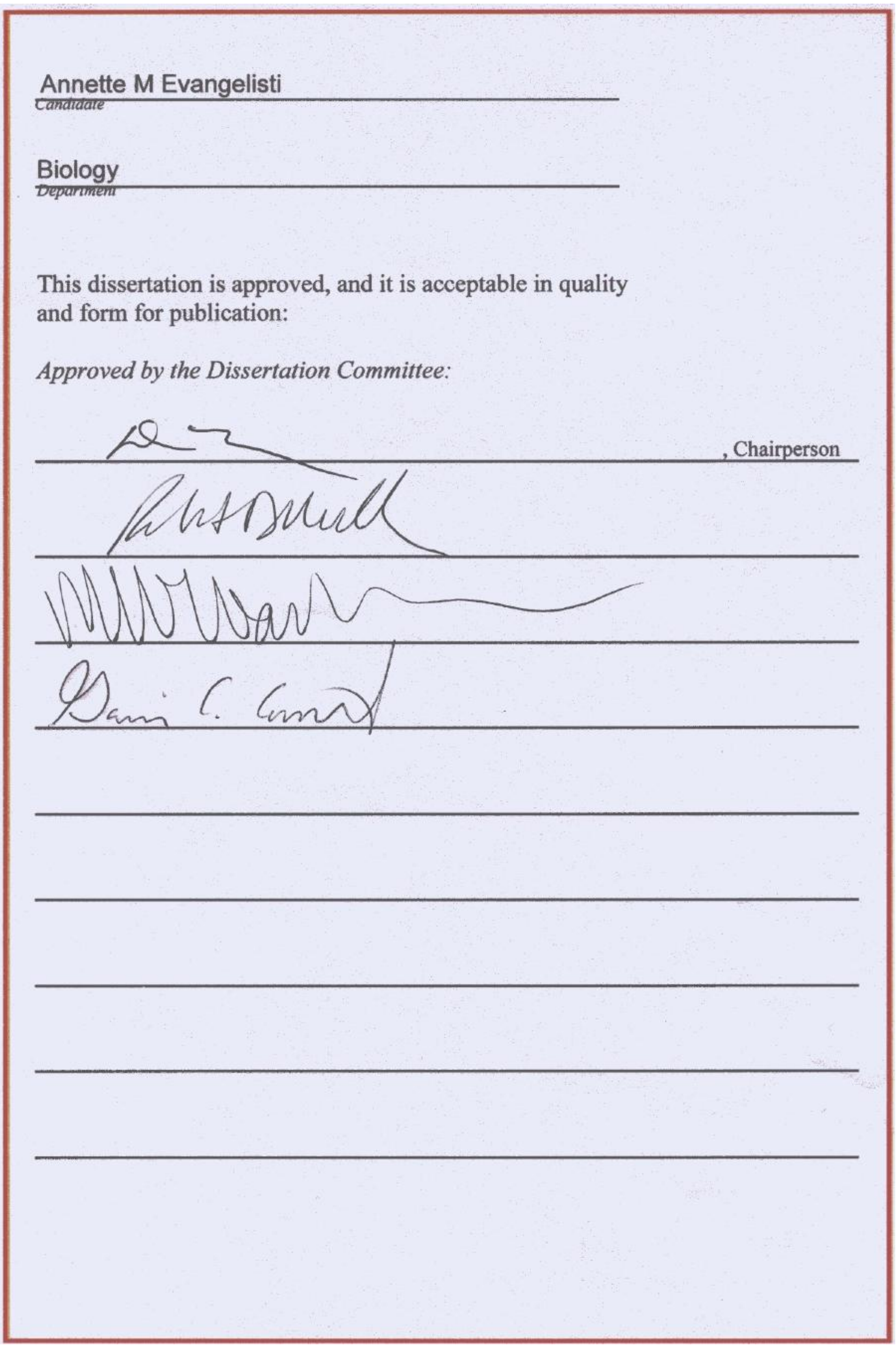
7-1-2010

Always read the introduction : integrating regulatory and coding sequence evolution in yeast

Annette Evangelisti



Annette M Evangelisti
Candidate

Biology
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

, Chairperson

Always read the introduction: Integrating regulatory and coding sequence evolution in yeast

BY

Annette M Evangelisti

B.S., Mathematics, University of Maryland, 1991
M.A. Applied Mathematics, University of Maryland, 1998

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy

Biology

The University of New Mexico
Albuquerque, New Mexico

July, 2010

DEDICATION

To my mother, Edna Evangelisti, you showed me the beauty of knowledge, I hope you would be proud.

ACKNOWLEDGMENTS

I am very grateful to my committee, Don Natvig as chair, Maggie Werner-Washburne, Rob Miller and with extra special thanks to Gavin Conant. Andreas Wagner also played a large role as an early advisor and coauthor on Chapter 4.

This work would not be possible without the Computational Science Fellowship from the Department of Energy, the Research Assistantship from Los Alamos National Lab and the Fellowship from the National Institutes of Health Complexity Grant with P.I. Jim Brown. Also, the data collection and discussion for Chapter 2 was performed in the labs of Vaishali Katju and Ulfar Bergthorsson.

Thanks to the professors that greatly enhanced my experience as a graduate student while I was a Teaching Assistant in their class. Bruce Hofkin, Sam Loker, Eric Toolson and Kelly Howe.

To the people who have shared their academic space with me: your camaraderie, courtesy and expertise was appreciated. Michael Fuller, Michael Gilchrist, Osorio Meirelles, Jeremiah Wright and Cindy Matheisen.

There are many people who have supported me in other ways. They know which forms to fill out, they have helped me navigate the bureaucracy or maybe they are always available with a smile or interesting chat. Richard Cripps, Catherine St Clair, Cheryl Martin, Toby Lucero, Phil Martin.

Then there are all the other people who have offered their friendship, diversion fun and conversation. Kelly Fitzpatrick, Julie McIntyre. Harriet Platero, Elisa LaBeau, Bonnie Lun, Michelle Forys, Sandra Melman, Jana

Gauntt, Patrick Hanington, Jim Cooke, Alexandra Kirk, Bill McKinney, Cris Pedregal Martin and Sina Sepheri.

To my family here in New Mexico, Drew and Sabrina, I offer my sincerest gratitude for bringing love, peace and a center into my life. To my clan on the East Coast, Anita, Llewellyn, Michael, Gloria, Jan and Dave; thanks for keeping me securely fastened to the ground.

Always read the introduction: Integrating regulatory and coding sequence evolution in yeast

BY

ANNETTE M EVANGELISTI

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy

Biology

The University of New Mexico
Albuquerque, New Mexico

July, 2010

Always read the introduction: Integrating regulatory and coding sequence evolution in yeast

ANNETTE M EVANGELISTI

B.S., Mathematics, University of Maryland, 1991
M.A., Applied Mathematics, University of Maryland, 1998
Ph.D., Biology, The University of New Mexico, 2010

ABSTRACT

We analyze duplicate genes in a yeast, *Saccharomyces cerevisiae* with the aim of determining a gene's history and to observe that gene in its genomic context. In Chapter 2 we show that the fate of a duplicate gene pair is in part determined by its genome location. Moreover, we show that for two classes of duplicate genes, resulting from either small-scale duplication or whole-genome duplication, this fate can often be assessed by measuring the patterns of asymmetry in the sequence divergence of the genes in question.
In Chapter 3 we study duplicate genes in the context of their local environments by comparing the patterns of evolution in the coding sequences of duplicate genes for ribosomal proteins with their upstream non-coding sequences. We found that while the coding sequences show strong evidence of recent gene conversion events, similar patterns are not seen in the non-coding regulatory elements. These duplicated ribosomal proteins are not functionally redundant despite their very high degree of protein sequence identity. This analysis confirms that the duplicated proteins have diverged considerably in expression despite their similar protein sequences.

In Chapter 4 we analyze the structure of the transcriptional regulation network and characterize the molecular evolution of both its transcriptional regulators and their regulated genes. We found that both subfunctionalization and neofunctionalization of transcription factor binding play a role in divergence.

Table of Contents

List of Figures

List of Tables

Chapter 1

Always read the introduction: Integrating regulatory and coding sequence evolution in yeast

Annette M. Evangelisti

Introduction

Interpreting the genetic code is like deciphering any other language. Knowing the alphabet or words does not always give rise to understanding. We study a language from different aspects, for each language has a history, a grammar, and contextual meaning. While we know the genetic alphabet and can "read" genes, we cannot consistently predict the function of the protein produced. So in the same manner in which one studies a language we apply this to our investigation of genes. Specifically, we analyze duplicate genes in a yeast, *Saccharomyces cerevisiae* with the aim of determining a gene's history and observe that gene in context to uncover the function of the resulting gene product.

What is our current understanding of how genomes change? Haldane (1933) first predicted the existence and evolutionary importance of gene duplication in 1933, well before DNA sequencing techniques were developed. In 1970, Ohno (1970) expanded Haldane's ideas by suggesting that duplicated genes were candidates to acquire novel function. With the considerable number of published genomes today, it is now known that gene duplication is a common occurrence, but that the rate of duplication varies both among species and among genes. For the majority of duplicate genes, both duplicates experience a short period of relaxed selection, resulting in one member of the pair quickly losing its function (Lynch and Conery 2000). The probability of loss of duplicated genes through genetic drift depends on the species, the mode of duplication, and the expression level of the gene (Taylor and Raes 2004).

Gene duplication occurs at different scales, ranging from whole genome duplication (WGD) to small scale duplications (SSD). Many researchers believe that WGD (or polyploidization events) are a precursor to evolutionary innovation (Comai 2005; Otto 2007; Soltis et al. 2009; Wittbrodt et al. 1998). Most WGD events do not survive the rigors of evolutionary selection but the WGD that have endured have given rise to very diverse and successful descendents (Van de Peer et al. 2009). Various lineages including flowering plants (Blanc et al. 2000; The Arabidopsis Genome Initiative 2000; Tuskan et al. 2006), amoeba (Aury et al. 2006), and vertebrates (Meyer and Van de Peer 2005). The first such event to be detected from a whole-genome sequencing effort was discovered in *S. cerevisiae* (Wolfe and Shields 1997). By mapping the relative genome orders of *S. cerevisiae* and seven related species, Byrne and Wolfe (2005; Wolfe 2000) have provided an essentially complete list of *S. cerevisiae* genes that remain. The nature of the duplicate genes that are not lost are of interest as they may include gene groupings that retain specific functional classes. Genes such as transcription factors, kinases, and ribosomal proteins commonly remain duplicated after WGD but, surprisingly, are not generally duplicated in smaller events (Aury et al. 2006; Blanc and Wolfe 2004; Freeling and Thomas 2006; Maere et al. 2005; Seoighe and Wolfe 1999).

Under what conditions are duplicated genes retained in a genome? There are several different processes that can preserve a duplicate gene in the genome. First is the classic pathway of neofunctionalization whereby one of the duplicate genes acquires a new function (Hughes and Hughes 1993; Ohno 1970) while the other gene retains the ancestral function. The second model (Force et al. 1999a; Hughes and Hughes 1993) involves partitioning the function of the ancestral gene between the two duplicates (subfunctionalization). Lastly, there is the case whereby the duplicate gene retains the ancestral function but the increase in protein expression bestows a selective advantage to the organism that may buffer against mutations (Conant and Wagner 2004; Gonzalez-Gaitan et al. 1994; Gu et al. 2003; Kondrashov and Kondrashov 2006; Nowak et al. 1997; Wagner 1999; Wang et al. 1996).

Methods to examine coding sequence. Transcriptional regulators and the genes whose expression they regulate form large gene regulation networks (Guelzim et al. 2002; Lee et al. 2002; Perez-Rueda and Collado-Vides 2000; Salgado et al. 2004). Analyzing 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. Understanding how the network evolves can shine light on gene evolution. On 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 (Chung et al. 2003; Sole et al. 2002; van Noort et al. 2004; Wagner 2001; Wagner 2003).

Measures for analysis of coding sequence. To understand the forces responsible for preserving a pair of duplicate genes one must study the history of that duplication. One the most useful tools for analyzing that history is the DNA sequence divergence between the two genes. Two measures of divergence that are of particular importance are the nonsynonymous substitution rate (Ka) and the synonymous substitution rate (Ks). Among the many things these measures can assess is whether the two duplicates have diverged at equal rates or if, on the contrary, one evolves more rapidly than the other. The degree to which asymmetry in evolutionary rate occurs after duplication is still somewhat contentious. In *S. cerevisiae*, differing estimates for the frequency of asymmetric divergence have been offered: Kellis et al., suggest 17% of duplicate genes produced by a genome duplication show asymmetry (Kellis et al. 2004) while Conant and Wagner estimated a frequency of 30% (Conant and Wagner 2003b) asymmetry in a more heterogeneous sample of duplicates. Thus, depending on the organisms studied, the genes selected for inclusion and the methods used to identify the divergence, asymmetry may or may not appear to play an important role in duplicate gene divergence.

Methods to examine noncoding effects. The most obvious aspect of the genomic context of a duplicate gene pair is the relative position of two genes in the genome, which is, in turn, determined by the duplication mechanism. For example, it has been shown in mammals that a duplicate gene inserted into the genome by retrotransposition is very likely to evolve faster than its counterpart in the ancestral location (Cusack and Wolfe 2007). Since *S. cerevisiae* underwent a WGD (Dietrich et al. 2004; Dujon et al. 2004; Kellis et al. 2004; Wolfe and Shields 1997), the resulting duplicate pairs were at least initially created with identical genomic contexts. Nevertheless, even among these duplicated genes, strong patterns of asymmetry, some dating to soon after the WGD, have been identified (Byrne and Wolfe 2007; Scannell and Wolfe 2008).

In duplicates produced by SSD, the association between loss of synteny and accelerated evolution is well known (Cusack and Wolfe 2007; Katju and Lynch 2003) and is likely related to duplication mechanism: i.e., a new duplicate lands in an alien genomic context and experiences relaxed selection as a result.

Measure for analysis of noncoding sequence. To calculate the pairwise divergence in the non-coding regions we first extracted the sequence between a gene in question and its 5' neighbor. We then computed pairwise local alignments using the local alignment algorithm of Smith and Waterman (1981). Non-coding DNA tends to evolve rapidly (Lavoie et al. 2010), so to be sure that these alignments represent evolutionarily conserved regions and not simply statistical noise, we compared their local alignment scores against an expected distribution drawn from the genome at large. Scores in the upper 5% of this randomized distribution were inferred to show evolutionary conservation.

How to we apply the aforementioned methods and measures? In Chapter 2 we examine genes in context by showing that the fate of a duplicate gene pair is in part determined by its genome location. Moreover, we show that for two classes of duplicate genes, resulting from either SSD or WGD, this fate can often be assessed by measuring the patterns of asymmetry in the sequence divergence of the genes in question.

In Chapter 3 we study duplicate genes in the context of their local environments by comparing the patterns of evolution in the coding sequences of duplicate genes for ribosomal proteins with their upstream non-coding sequences. We found that while the coding sequences show strong evidence of recent gene conversion events, similar patterns are not seen in the non-coding regulatory elements. These duplicated ribosomal proteins are not functionally redundant despite their very high degree of protein sequence identity. This analysis confirms that the duplicated proteins have diverged considerably in expression despite their similar protein sequences.

In Chapter 4 we analyze the structure of the transcriptional regulation network and characterize the molecular evolution of both its transcriptional regulators and their regulated genes. We found that both subfunctionalization and neofunctionalization of transcription factor binding play a role in divergence.

Chapter 2

Why neighborhoods matter: Accelerated evolution of relocated, duplicated genes in *Saccharomyces cerevisiae*

Annette M. Evangelisti

Abstract

Gene duplication is an important engine of evolutionary innovation. The fate of a newly formed duplicate gene pair is in part determined by the genome locations of the two duplicated genes. Moreover, this fate can often be assessed by analysis of the patterns of asymmetry in the sequence divergence of the genes in question. For two classes of duplicate genes, resulting from either smaller scale duplications (SSD) or whole-genome duplication (WGD), I computed the relative rate of sequence divergence between the gene pairs and compared the gene order in a neighborhood around each of the genes. Duplicates of both types (WGD and SSD) show asymmetric divergence and a pattern of gene loss surrounding one of the genes in the pair. The gene that experiences gene loss in its neighborhood is also the gene with accelerated divergence. While duplicate pairs from SSD are expected to have one gene that experiences gene loss and accelerated divergence in a local region, it is surprising to find the same pattern in the paralogs resulting from a WGD event given the circumstances of their birth, i.e. WGDs are assumed to be equal at birth. These results illustrate the importance of post-duplication events in determining the fate of duplicate genes.

Introduction

Haldane (1933) first predicted the evolutionary potential of gene duplication in 1933, well before DNA sequencing techniques were developed. In 1970 Ohno (1970) expanded these ideas by predicting the importance of gene duplication and the potential for gene duplicates to gain novel functions. With the considerable number of published genomes, it is now known that gene duplication is a common occurrence, but that the rate of duplication varies both among species and among genes. For the majority of duplicate genes, both duplicates would be expected to experience a short period of relaxed selection, resulting in one member of the pair quickly losing its function (Lynch and Conery 2000). The probability of loss of duplicated genes through genetic drift depends on the mode of duplication, the species and the expression level of the gene (Taylor and Raes 2004).

Because duplicate genes which are not lost through drift have the potential to introduce novelty into the genome, gene duplication has a place of importance in any discussion concerning molecular evolution of genes and genomes (Li 1996). There are in fact several processes that can preserve a duplicate gene in the genome. First is the classic pathway of neofunctionalization whereby one of the duplicate genes acquires a new function (Hughes and Hughes 1993; Ohno 1970) while the other gene retains the ancestral function. The second model (Force et al. 1999a; Hughes and Hughes 1993) involves partitioning the function of the ancestral gene between the two duplicates (subfunctionalization). Lastly, there is the case whereby the duplicate gene retains the ancestral function but the increase in protein expression bestows a selective advantage to the organism (dosage selection) (Kondrashov and Kondrashov 2006).

To understand the forces responsible for preserving a pair of duplicate genes it is necessary to study the history of that duplication. One the most useful tools for studying this history is the DNA sequence divergence between the two genes. In particular, two measures of this divergence are the nonsynonymous substitution rate (Ka) and the synonymous substitution rate (Ks). These measures can help determine whether the two duplicates have diverged at equal rates or if, on the contrary, one has evolved more rapidly than the other. The degree to which asymmetry in evolutionary rates occurs after duplication is still somewhat contentious: estimates range from 5% (Kondrashov et al. 2002) to up to 50% (Dermitzakis and Clark 2001; Van de Peer et al. 2001). In the yeast , *Saccharomyces cerevisiae*, differing estimates for the frequency of asymmetric divergence have been offered: Kellis et al., suggest 17% of duplicate genes produced by a genome duplication show asymmetry (Kellis et al. 2004) while Conant and Wagner, in a more heterogeneous sample of duplicates, estimated a frequency of 30% (Conant and Wagner 2003b). Thus, depending on the organisms studied, the genes selected for inclusion and the methods used to identify the divergence, asymmetry may or may not appear to play an important role in duplicate gene divergence. However, some caution in the negative conclusion is warranted. Seoighe and Scheffler have shown that the null hypothesis of symmetric divergence between paralogs is difficult to reject due to issues of statistical power, both with standard molecular clock methods and with their own novel codon model of evolution (Seoighe and Scheffler 2005). Given this low power, even the detection of a small percentage of asymmetrically evolving duplicates may suggest that this process is an important one in duplicate gene evolution.

In this work, I analyzed asymmetry of duplicate evolution in the context of a somewhat underappreciated feature of duplicated genes. While it is natural to assume that immediately after gene duplication the resulting two gene copies are identical, this is not always the case. A study of duplicate genes in C. elegans found that the median of the duplication span fell short of the average gene length, leading to incomplete duplicates for approximately half of the paralogs (Katju and Lynch 2003).. This observation suggests a more general principle: the genomic context of the two genes can have profound effects on duplicate gene evolution and in particular on the patterns of asymmetry in that evolution (Cusack and Wolfe 2007; Katju and Lynch 2003).

The most obvious aspect of the genomic context of a duplicate gene pair is the relative position of two genes in the genome, which is, in turn, determined by the duplication mechanism. For example, it has been shown in mammals that a duplicate gene inserted into the genome by retrotransposition is very likely to evolve faster than its counterpart in the ancestral location (Cusack and Wolfe 2007). The organism studied here, *S. cerevisiae* underwent a whole-genome duplication (WGD) (Dietrich et al. 2004; Dujon et al. 2004; Kellis et al. 2004; Wolfe and Shields 1997), meaning that the resulting duplicate pairs were at least initially created with identical genomic contexts. Nevertheless, even among these duplicated genes, strong patterns of asymmetry, some dating to soon after the WGD, have been identified (Byrne and Wolfe 2007; Scannell and Wolfe 2008).

Here, I compared the two types of duplication present in *S. cerevisiae*, those produced by WGD and by SSD to see whether, in each case, the local genomic neighborhood is associated with particular patterns of asymmetry. Interestingly, I found that the neighborhood matters in both cases: the gene in the less conserved genomic region is more likely to undergo rapid evolution.

Methods

All sequences were downloaded from the *Saccharomyces* Genome Database (http://www.yeastgenome.org/). Table 2lists the source address and the total number of genes downloaded for each genome. *S. castellii* and *S. kluyveri* did not have gene names assigned so a given gene was labeled with the contig number followed by a number reflecting the relative position of the gene in the contig. For example, the 31st coding sequence on contig 795 for *S. kudriavzevii* was named Skud795.31.

Ka is the number of nonsynonymous (amino acid-changing) nucleotide substitutions per nonsynonymous site and Ks is the number of synonymous (amino acid-preserving) nucleotide substitutions per nucleotide site. I used GenomeHistory (Conant and Wagner 2002) to calculate Ka and Ks. GenomeHistory performs a three-step analysis. A gapped BLASTP (Altschul et al. 1997) identifies candidate pairs of duplicate genes, which then undergo pairwise global sequence alignment (Needleman and Wunsch 1970) of the amino acid sequences in question. Finally, Ks and Ka are estimated by maximum likelihood (Yang and Nielsen 2000).

All seven genomes were analyzed by GenomeHistory performing an all-gene to all-gene comparison. Pairs of genes qualified as paralogous in *S. cerevisiae* if Ks |ScerA, ScerB| < 1.0 and if ScerA and ScerB comprised a gene family of exactly two. The outgroup for a duplicate pair of genes was required to be a single copy ortholog to both genes in *S. cerevisiae* and in the case where there was more than one candidate for the outgroup – I chose the genome that is more closely related to *S. cerevisiae*. I identified 53 triplets that satisfied this criterion. The synonymous divergence between at least one of the *S. cerevisiae* genes and the outgroup gene was always greater than the divergence between the two *S. cerevisiae* paralogs.

Table 1 - Genomes and their source.

Genome	Download From	Number of Genes
S. cerevisiae	ftp://genome-ftp.stanford.edu/pub/yeast/sequence/genomic_sequence/orf_dna	6718
S. paradoxus	ftp://genome-ftp.stanford.edu/pub/yeast/sequence/fungal_genomes/S_paradoxus/MIT	8955
S. mikatae	ftp://genome-ftp.stanford.edu/pub/yeast/sequence/fungal_genomes/S_mikatae/MIT	9057
S. kudriavzevii	ftp://genome-ftp.stanford.edu/pub/yeast/sequence/fungal_genomes/S_kudriavzevii/WashU	3768
S. bayanus	ftp://genome-ftp.stanford.edu/pub/yeast/sequence/fungal_genomes/S_bayanus/MIT	9423
S. castellii	ftp://genome-ftp.stanford.edu/pub/yeast/sequence/fungal_genomes/S_castellii/WashU	4677
S. kluyveri	ftp://genome-ftp.stanford.edu/pub/yeast/sequence/fungal_genomes/S_kluyveri/WashU	2968

The first column is the name of the organism, second column is the ftp site and the third column lists the number of genes downloaded for each organism.

To identify the syntenic contexts of a pair of duplicated genes in *S. cerevisiae*, I started with a triplet of genes, ScerA, ScerB and OutAB as described above. ScerA is a gene located on chromosome A, ScerB on chromosome B and OutAB on chromosome O (Figure 1). I aligned segments of chromosome A and B that contained the genes ScerA and ScerB along with the segment that contained the outgroup gene, OutAB. Starting from each gene of interest (ScerA, ScerB, OutAB), I examined the 12 flanking genes on each side. For reference, I label the original gene as position 0 and then number the upstream genes as +1, +2, …, +12 and the downstream genes -1, -2, …, -12. To infer synteny conservation, the 24 genes surrounding OutAB were compared with the genes neighboring the two duplicate genes in the *S. cerevisiae* genome. I thus counted the number of upstream and downstream neighbors of ScerA that had orthologs in the outgroup that are in the 24 gene neighborhood of OutAB (MA). Similarly, MB was the number of orthologs in the neighborhood of ScerB that were also in the neighborhood of OutAB. For each duplicate pair I compared the number of genes found upstream/downstream that had orthologs on either chromosome A or B. If I found a majority of matches on chromosome B, for example, I concluded that there was evidence that the ancestral copy of the gene was on the B chromosome. I used the criteria in Table 2 to determine the Consynteny.

I used Mega4.0 (Tamura et al. 2007) to compute Tajima's Relative Rates test. For all 53 data points I performed a pairwise distance analysis for the two *S. cerevisiae* duplicate pair against the outgroup.

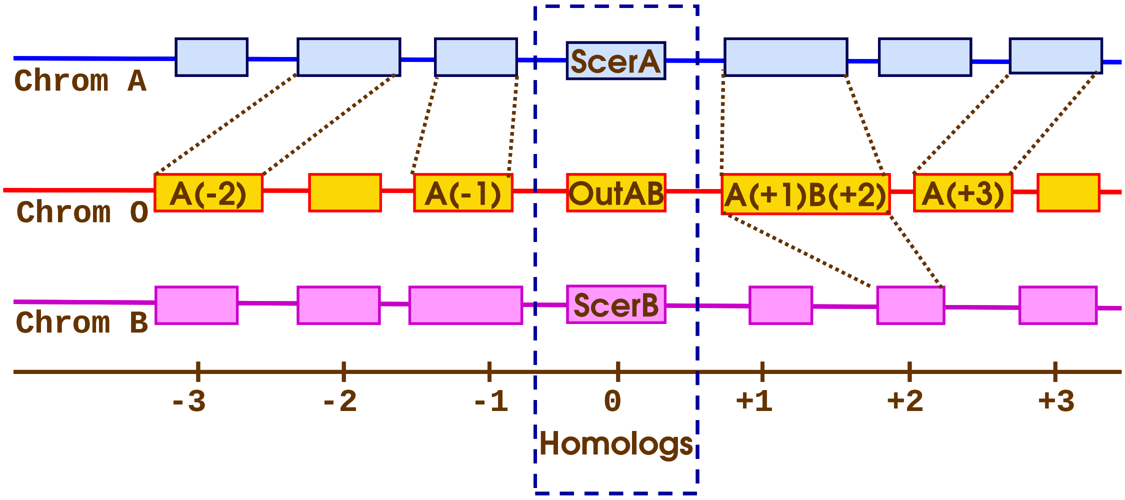


Figure 1 - Illustration of the algorithm used to distinguish the ancestral gene from the Nonconsynteny.

The duplicate pair of *S. cerevisiae* genes, ScerA, ScerB and its single copy ortholog, OutAB are lined up in the center of the diagram on their respective chromosomes; A, B and O. The position of the three paralogous genes is labeled 0, the genes upstream are labeled +1, +2, etc. and the genes downstream are labeled -1, -2, etc. A(-2) indicates that the gene on chromosome O in position -3 is orthologous to the gene on chromosome A in position -2. A(+1)B(+2) indicates that the gene on chromosome O is orthologous to both the gene on chromosome A in position +1 and the gene on chromosome B in position +2.

Table 2 - Criteria for Consynteny call.

$\|MA^a – MB^b\|$	Consynteny location
≤ 2	No call
> 2	ScerA if MA > MB ScerB if MA < MB

a: MA is the number of homologous genes that the 24 genes surrounding OutAB have in common with the 24 neighbors of ScerA
b: MB is the number of homologous genes that the 24 genes surrounding OutAB have in common with the 24 neighbors of ScerB

Results

Comparing duplicate genes pairs from *S. cerevisiae* with six outgroup *Saccharomyces* genomes (Figure 2) of known phylogeny (Kurtzman and Robnett 2003) allowed me to measure sequence divergence and conservation of gene order between the paralogs. The duplicate pairs retained for this study met the following three criteria: 1) the two genes had no other close relatives in the *S. cerevisiae* genome (i.e. they were not members of a larger gene family), 2) there existed an orthologous single copy gene from one of the six outgroups and 3) the rate of synonymous substitutions per nucleotide site (Ks) was less than 1.0 (see Methods). The *S. cerevisiae* paralogs are denoted ScerA, ScerB and their single copy ortholog as OutAB.

The first question explored was whether there was evidence of significant asymmetry in rates of evolution for the duplicate gene pairs. Using Tajima's relative rate test (Tajima 1993),, 47% (25/53) of the pairs showed significant asymmetry using the nucleotide sequences (nt) and 11% (6/53) showed significant asymmetry using the amino acid (aa) sequences ($P < 0.05$ for both tests). Note that all of the paralogous pairs showing significant asymmetry in their amino acid sequences also showed significant asymmetry in their nucleotide sequences.

Next, I sought to place the asymmetry into the context of the location of the genes. Genomic context has a different meaning for genes produced by whole-genome duplication (WGD) and for genes produced by small-scale duplication (SSD). For all the genes, regardless of duplication mechanism, I searched for conserved synteny between paralogs and their

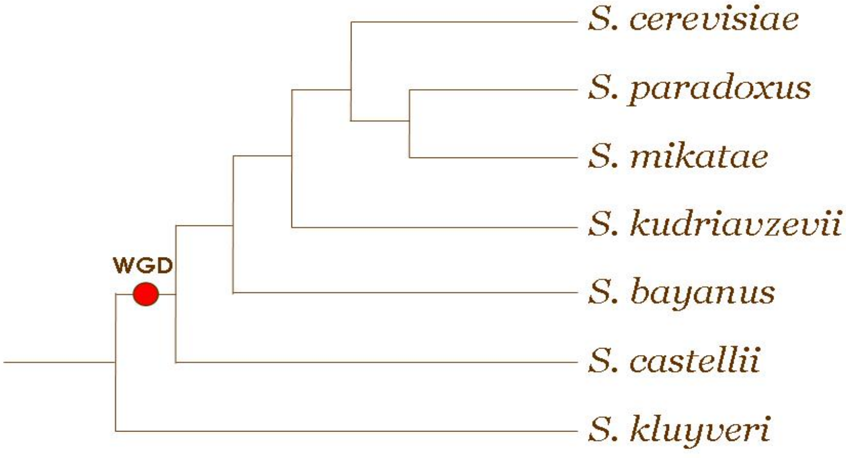


Figure (tree)

Figure 2 - The phylogenetic relationship among *S. cerevisiae* and the six outgroups studied (Kurtzman and Robnett 2003).

The position of the whole genome duplication (WGD) is represented by a red dot. (Dietrich et al. 2004; Dujon et al. 2004; Kellis et al. 2004; Wolfe and Shields 1997).

respective ortholog in the outgroup genome (OutAB; see Methods). For the SSD genes, it is assumed that the derived copy has been inserted into a new genomic location and it will not show conserved synteny. In the case of WGD genes the notion of ancestor/derived is meaningless, but the process of gene loss after WGD (Scannell et al. 2006) can still yield duplicate pairs with differing levels of local synteny. Due to the differing origins of these variations in syntenic context depending on duplication mechanism, I introduce the generic terminology of consynteny, shorthand for conserved synteny, and nonconsynteny, denoting a lack of such conservation.

For 85% (45/53) of the duplicate gene pairs, the differences in synteny between the two genes was sufficient to assign consynteny/nonconsynteny status (see Methods). For these pairs I calculated Ka of |Consynteny, OutAB| and Ka |Nonconsynteny, OutAB|, where Ka is the number of nonsynonymous substitutions per nucleotide site (Li 1997). A Wilcoxon's matched paired test showed a significantly higher proportion of the fast evolving duplicate copies were in the nonconsynteny class ($P = 1 \times 10^{-3}$; N = 45; see Table 3).

Since the previous analysis included gene pairs produced both by SSD and by WGD, the next step was to divide the data into ohnologs (duplicate genes derived from the WGD) and nonohnologs (duplicate genes of any other origin) and repeat the analysis. The Wilcoxon's matched paired test for Ka of |Consynteny, OutAB| and Ka |Nonconsynteny, OutAB| for the ohnolog and nonohnolog groups was significant for both groups again showing that a significantly higher proportion of the fast evolving duplicate are in the non conserved synteny group ($P = 1 \times 10^{-2}$ for N = 34 (ohnologs); $P = 4 \times 10^{-2}$ for N = 11 (nonohnologs) (Table 3).

Table 3 - Wilcoxon's matched pairs test for asymmetric divergence of duplicate pairs.

Measure of divergence	**Genes analyzed**	**N**	**P value**[a]
Ka[b]	All	45	1×10^{-3}
	Symm. AA seqs[c]	40	0.02
	Symm. NT seqs[d]	21	0.4
Ks[e]	All	45	9×10^{-8}
	Symm. AA seqsc	40	9×10^{-7}
	Symm. NT seqsd	21	9×10^{-3}

a: P values are bold if they are significant at α = 0.05 level.
b: Values compared were Ka|Consynteny, OutAB| vs Ka|Nonconsynteny, OutAB|
c: Only includes sequences with non-significant asymmetry in amino acid sequence by Tajima test.
d: Only includes sequences with non-significant asymmetry in nucleotide sequence by Tajima test.
e: Values compared were Ks|Consynteny, OutAB| vs Ks|Nonconsynteny, OutAB|

In an additional test for divergence, I divided the 45 paralogous pairs between Ks |Consynteny, OutAB| and Ks |Nonconsynteny, OutAB| , where Ks is the number of synonymous substitutions per nucleotide site between two sequences. The faster evolving member of the duplicate pair in terms of Ks was again the gene with the less conserved synteny (Wilcoxon's test; P = 9 x 10^{-8}; N = 45). This pattern is also observed for the ohnolog and nonohnolog groups individually (Wilcoxon's matched pairs test; P = 1 x 10^{-6}; N = 34 and P = 3 x 10^{-2}; N = 11, respectively).

Discussion

I found that both duplicate genes produced by SSD and by WGD show an association between increased rates of sequence evolution and loss of local synteny. Thus, the loss of upstream or downstream duplicated genes in the region surrounding one member of a duplicate pair seems to coincide with faster sequence evolution in that gene. For the WGD-produced duplicates, this pattern is particularly interesting because such duplicates are "identical at birth". Asymmetry in such WGD duplicates has already observed (Byrne and Wolfe 2007; Scannell and Wolfe 2008), but its association with differences in genomic context appears to be novel.

There are caveats to consider when evaluating the results of this study. Because the depth of coverage differs in the genome sequences considered, it is possible that some duplicated genes paralogous to *S. cerevisiae* genes may have been missed in the outgroup genome. Another potential problem is the inherent difficulty in detecting asymmetric divergence with this data set by observing the nucleotide sequence alone.

Gene conversion, the process by which a portion of the nucleotide sequence of one gene in a duplicate pair replaces the nucleotide sequence in its corresponding gene, has been documented in *S. cerevisiae* (Li 1996; Sharp and Cowe 1991). The effect of gene conversion is that a molecular-clock-like measurement of time since divergence will be reset and the duplicate pair will appear as if newly duplicated. Such events could have subtle but important effects on this analysis. I selected the duplicate gene pairs for analysis on the basis that they were more closely related to each other than to the outgroup gene. Gene conversion could give rise to such pairs in a manner where, although the two sequences are closely related, the genomic neighborhoods are in fact more distantly related. However, I note that at each phase of this study I chose the most conservative of approaches to minimize these potential errors.

In duplicates produced by SSD, the association between loss of synteny and accelerated evolution is well known (Cusack and Wolfe 2007; Katju and Lynch 2003) and likely related to the duplication mechanism: i.e., a new duplicate lands in an alien genomic context and experiences relaxed selection as a result. The source of this effect in the WGD-produced duplicates is less clear. However, one hypothesis is that the loss of genes, particularly upstream genes, could have the same disruptive effects on promoters that has been hypothesized to underlie the asymmetry observed in mammals (Cusack and Wolfe 2007). Here one can see that regardless of the mechanism that produced the duplicate pair the neighborhood in which a gene finds itself plays an important role in determining its ultimate fate.

Acknowledgments

I am extremely grateful to G. Conant for his insight, many fruitful discussions and his patience. Also, I wish to thank V. Katju, and U. Bergthorsson for helpful discussions on this data.

Chapter 3

Gene conversion among duplicated yeast ribosomal proteins does not extend to upstream regulatory regions

Annette M. Evangelisti

Coauthor:
Gavin C. Conant
Division of Animal Sciences and Informatics Institute
University of Missouri
Columbia MO, U.S.A.

Abstract

By comparing the patterns of evolution in the coding sequences and upstream non-coding sequences of yeast ribosomal proteins duplicated in a genome duplication, we find that while the coding sequences show strong evidence of recent gene conversion events, similar patterns are not seen in the non-coding regulatory elements. This result suggests a potential explanation of the somewhat puzzling fact that duplicated ribosomal proteins are not functionally redundant despite their very high degree of protein sequence identity. Analysis of the patterns of regulatory network evolution after genome duplication confirms that the duplicated proteins have diverged considerably in expression despite their similar protein sequences.

Introduction

With the completion of the genomic sequencing of numerous organisms, it has become evident that polyploidization (or whole-genome duplication, WGD) events have occurred in diverse lineages including flowering plants (Blanc et al. 2000; The Arabidopsis Genome Initiative 2000; Tuskan et al. 2006) amoeba (Aury et al. 2006) and vertebrates (Meyer and Van de Peer 2005). The first such event to be detected in a whole-genome sequence was that in *Saccharomyces cerevisiae* (Wolfe and Shields 1997): striking confirmation of this event was found with the two-to-one mapping of chromosomal regions in *S. cerevisiae* to the genomes of other yeasts lacking the WGD (Dietrich et al. 2004; Dujon et al. 2004; Kellis et al. 2004). Polyploidization events are often followed by substantial losses of duplicated genes (Semon and Wolfe 2007). Which of the two duplicate copies is lost is generally thought to be selectively neutral: if two populations lose alternative copies such reciprocal gene loss can contribute to reproductive isolation and hence speciation (Scannell et al. 2006). The nature of the duplicate genes that are not lost is also of interest: functional classes of genes such as transcription factors, kinases and ribosomal proteins commonly remain duplicated after WGD but, surprisingly, are not generally duplicated in smaller events (Aury et al. 2006; Blanc and Wolfe 2004; Freeling and Thomas 2006; Maere et al. 2005; Seoighe and Wolfe 1999). Wolfe (2000) has proposed the name ohnologs (in honor of Susumu Ohno) for these duplicate genes surviving from WGD. By mapping the relative genome orders of *S. cerevisiae* and seven related species, Byrne and Wolfe (2005) have provided an essentially complete list of *S. cerevisiae* ohnologs.

In S. cerevisiae, approximately 10% of surviving ohnologs are in fact ribosomal proteins (RPs; Byrne and Wolfe 2005; Kim et al. 2009; Planta and Mager 1998). It has been suggested that selection to maintain (high) relative gene dosage among RP genes is at least in part responsible for this over-retention (Birchler and Veitia 2007; Freeling and Thomas 2006; Koszul et al. 2004; Papp et al. 2003). Given this hypothesis, it is suggestive that many of the RP ohnologs are very similar in sequence; in fact, it is thought that these genes have undergone one or more gene conversion events post-WGD (Kellis et al. 2004). Gene conversion occurs when nonhomologous recombination causes the overwriting of substitutions in one gene with the corresponding bases from a paralog. The net effect of such events is to erase the historical sequence divergence between paralogs, and one can plausibly argue that any functional differences between the two genes would be erased simultaneously. Curiously however, there are examples of paralogous RPs in yeast with high sequence identity (>97%) that nonetheless differ in their functional roles (Enyenihi and Saunders 2003; Kaeberlein et al. 2005; Kim et al. 2009; Komili et al. 2007; Ni and Snyder 2001).

Here, we examined the patterns of gene conversion in the yeast ribosomal protein (RP) ohnologs, finding strong evidence for gene conversion in the coding regions of these genes but little evidence of such conversion events in the upstream non-coding regions. An analysis of the RP ohnolog expression network also showed dissimilar expression patterns, consistent with regulatory divergence between the copies being responsible for the observed functional divergence.

Methods

Data sources and orthology inference. A total of 55 previously described WGD-produced duplicate ribosomal proteins (RP; Conant and Wolfe 2006; Planta and Mager 1998) were analyzed. To this set, we added 84 pairs of enzyme genes duplicated at the WGD, identified by cross-referencing to the list of metabolic genes of Kuepfer, Sauer and Blank (2005) to the set of *Saccharomyces cerevisiae* ohnologs (Byrne and Wolfe 2005).

For these two lists (totaling 139 duplicate pairs), we next identified the corresponding orthologous genes in the genome of *S. bayanus* . Orthology inference in post-WGD species is challenging due to reciprocal gene loss, which can give rise to paired homologous genes that are paralogous rather than orthologous {Figure 3`; \Scannell, 2006 #66;, 2007 #88}. We have previously developed a maximum likelihood method that addresses this problem (Conant and Wolfe 2008). Briefly, the analysis begins with an inferred pre-WGD gene order (similar to that of Gordon, Byrne, and Wolfe; 2009). A model of duplicate gene loss after WGD allows us to estimate the relative speciation times of the taxa analyzed and the probability of all possible orthology assignments. Thus, in Figure 3, we estimate with greater than 99.99% confidence that *S. bayanus* gene number 34.11 is the ortholog of *S. cerevisiae* gene RPL26B as opposed to the alternative possible assignment that makes gene 34.11 the ortholog of RPL26A. Importantly, these inferences rest only on the relative gene orders: gene sequences are not considered.

From our list of 55 ribosomal protein gene (RP) duplicates and 77 enzyme duplicates (MP), we thus selected the 29 RP pairs and 76 MP pairs for which the probability of our orthology assignment between *S. cerevisiae* and *S. bayanus* was > 0.98. Thus, these genes represent a set for which we have high confidence orthology information independent of the sequences themselves.

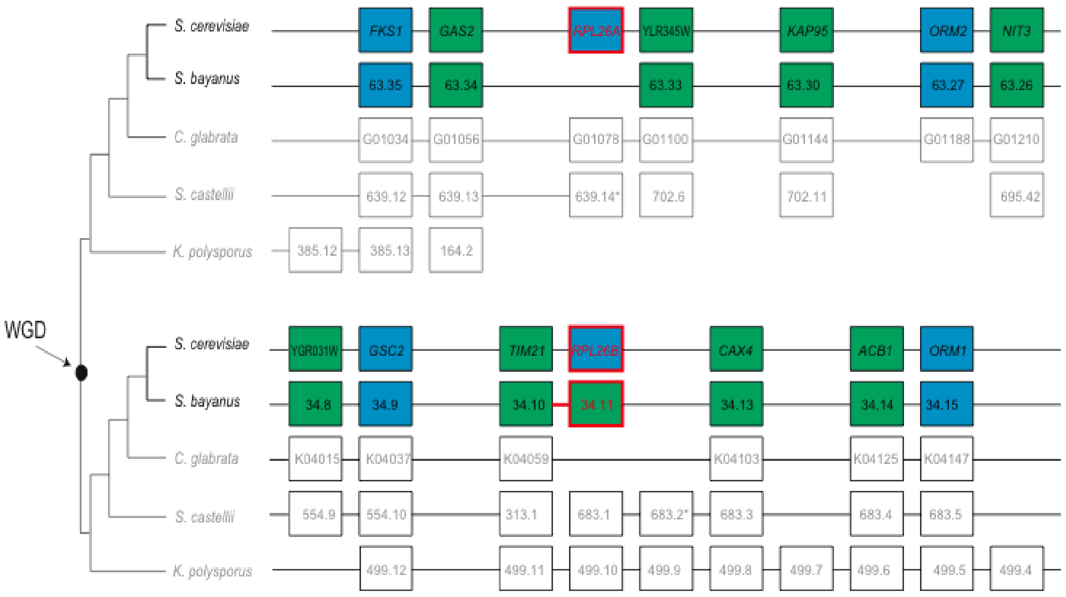


Figure 3 - Illustration of the pattern of genome evolution after WGD in five yeast species in a region surrounding a pair of duplicated ribosomal protein genes (RPL26A and RPL26B).

The upper five tracks and the lower five tracks are inferred to be two orthologous groups. Lines connect genes that are adjacent on their respective contigs or chromosomes. Duplicate genes surviving from WGD are colored blue, green genes are cases where one member of the duplicate pair has been lost post-WGD. The orthology assignments between the paired *S. cerevisiae* and *S. bayanus* genes on the upper and lower tracks are all inferred with greater than 99.99% confidence.

Sequence analyses. We next analyzed the sequence divergence in the coding regions of *S. cerevisiae* ohnolog pairs (Figure 4). To do so, we aligned sequence triplets consisting of two ohnologs from *S. cerevisiae* (Scer1 and Scer2 below) and the *S. bayanus* gene orthologous to Scer1 (Sbay below) using T-Coffee (Notredame et al. 2000). Using these alignments, estimates of the number of nonsynonymous substitutions per nonsynonymous site (Ka) for each of the three branches in Figure 4 were estimated by maximum likelihood as previously described (Conant and Wagner 2003b). Similar calculations were made for the synonymous sites (Ks, data not shown). Note that for most *S. cerevisiae* ohnolog pairs, there are actually two possible triplets, because the corresponding *S. bayanus* genes are also duplicates surviving from

WGD. In such cases, we performed both comparisons (meaning that the identity of Scer1 and Scer2 was switched in the second case).

To test the statistical support for an inference of gene conversion between genes Scer1 and Scer2, we employed a likelihood ratio test (Sokal and Rohlf 1995). First, we identified cases where KaB > Ka1, Ka2, (i.e., the signature of gene conversion; Figure 4) and calculated the likelihood of the sequence alignment under this model (lnLH0). We then constrained the model such that Ka1 = KaB and calculated the likelihood under this alternative model (lnLHA). We compared 2·(lnLH0- lnLHA) to a chi-square distribution with 1 degree of freedom.

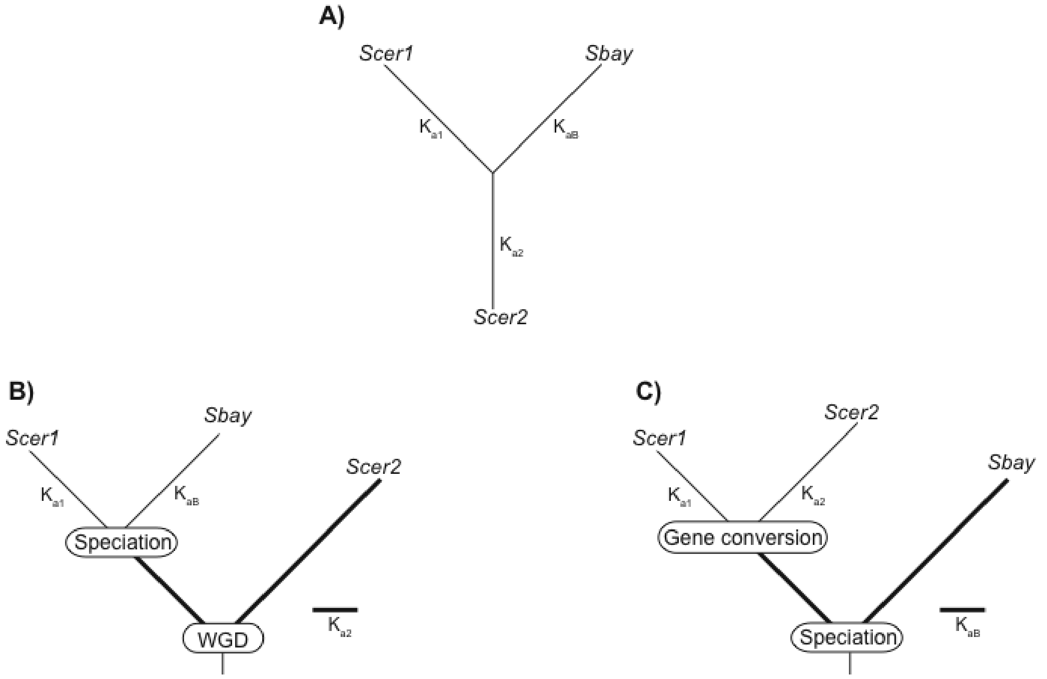


Figure 4 - Analysis of duplicated *S. cerevisiae* genes and a *S. bayanus* ortholog.

A) The format of our triplet-based sequence analysis. Because the models used are time-reversible, only a single, three taxa tree is required. Independent estimates of Ka are made for each branch. B) The expected pattern of branch lengths for the tree in A if the genes follow the known species tree. Note that we expect Ka2 to be large as it represents both the divergence of the gene Scer2 as well as the shared divergence of Sbay and Scer1 post-WGD. C) The expected gene tree if Scer1 and Scer2 have undergone recent gene conversion events. Here we expect KaB to be the largest of the three Ka values, under the same reasoning as in B.

Ribosomal gene expression network divergence after WGD. We have previously described an algorithm for detecting network partitioning among WGD-produced duplicate genes (Conant and Wolfe 2006). As is illustrated in Figure 5, paralogs are divided into two columns with ohnologs opposite each other in a network. Gene expression data for 51 pairs of RP ohnologs were obtained from the expression compendia of Hughes et al. (2000) and overlaid as graph edges. We divide these edges into internal edges, connecting nodes in the same column (arcs or vertical lines in Figure 5), and crossing edges, joining nodes in opposite columns (diagonal lines in Figure 5). Note that the initial assignment of a particular paralog to the first or second column is arbitrary, meaning that there are 2^{n-1} possible unique partitionings of the duplicates into columns. Using the previously described heuristic partitioning algorithm (Conant and Wolfe 2006), we search for the partition among these 2^{n-1} that gives the fewest crossing edges.

To determine if the RP gene expression data showed fewer crossing edges than would be expected by chance, we randomized the networks and recalculated the optimal partitioning. Randomization was performed by selecting every possible quartet of two pairs of duplicates. These four node subgraphs were replaced at random by another four-node subgraph with the same number of edges (Conant and Wolfe 2006). The probability of each such subgraph was calculated based on the inherent asymmetry in interaction degree between paralogs. Thus, we calculated the average fraction p of the total number of interactions for a paralog pair that belonged to the interaction-rich paralog. The probability of an interaction joining two interaction-rich genes

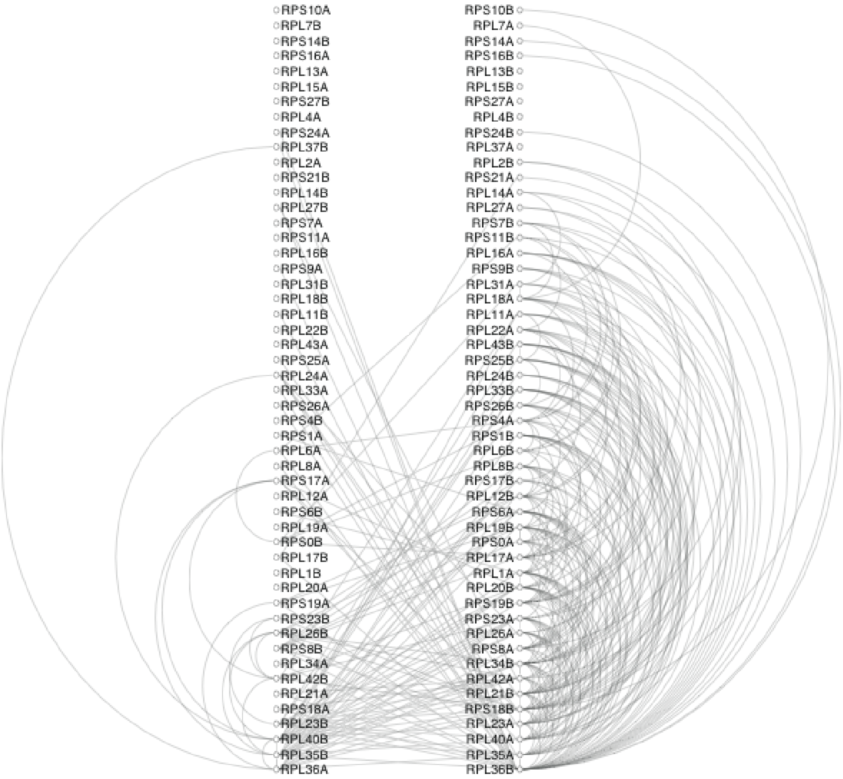


Figure 5 - Gene expressions networks of duplicated ribosomal proteins have diverged since WGD.

Pairs of duplicated ribosomal proteins are arranged opposite each other. Edges connect pairs of genes with correlation in gene expression > 0.8. We searched among the 2^{n-1} permutations of the column arrangements to find this arrangement, which has the minimal number of edges (102) crossing between the two partitions. The minimal number of crossing edges seen in randomized networks was 107, mean was 116. Note also the high degree of asymmetry in the number of interactions seen between duplicated ribosomal proteins.

is thus p^2, while the probability of an interaction joining an interaction-rich and interaction-poor gene is then 2p(1–p). Subgraph probabilities are calculated accordingly. The number of crossing edges in the original network was then compared to the distribution of number of crossing edges seen in 1000 randomized networks.

Results

Strong evidence for numerous gene conversion events among the duplicated ribosomal proteins. Our previous analysis of patterns of gene loss after WGD (Conant and Wolfe 2008) allows us to infer with high confidence that all of the duplicate gene loci discussed here evolved according to the species tree in Figure 4 (Methods). Despite this fact, it is not necessarily the case that the sequences themselves will have evolved under this set of relationships. In particular, a gene conversion event between Scer1 and Scer2 that occurred after the speciation of *S. cerevisiae* and *S. bayanus* would overwrite the historical signal in the sequences of the two genes and give rise to a gene tree of the form of Figure 4.

Using estimates of Ka for triplets of ribosomal protein genes (RP; see Methods), we asked whether the pattern of nonsynonymous divergence in each triplet was most compatible with divergence after WGD (i.e., Ka2 > Ka1, KaB, Figure 4) or with a recent gene conversion event (KaB > Ka1, Ka2, Figure 4). Of the 29 pairs of duplicated RPs in *Saccharomyces cerevisiae*, two follow the pattern expected under WGD, two ohnolog pairs present conflicting patterns of Ka values depending on the *S. bayanus* ortholog used, and the remaining 25 pairs have nonsynonymous divergences consistent with gene conversion. We next assessed whether the signature of gene conversion in the sequence data was strong enough to statistically reject the possibility that phylogenetic relationships within the triplet were simply ambiguous. Thus, we compared a model allowing gene conversion (KaB > Ka1, Ka2) to an alterative model where KaB was constrained to be equal to Ka1. Of the 25 RP duplicate pairs with signatures of gene conversion, 17 showed statistically significant improvement when a model allowing gene conversion was used ($P < 0.05$, likelihood ratio test).

Metabolic genes duplicated at WGD do not show similar patterns of gene conversion. We applied the above approach to a similar set of WGD-duplicated metabolic genes. Among the 76 pairs considered only three show any signs of gene conversion and only 2 of those have significant improvement when the gene conversion model is used ($P < 10^{-7}$, likelihood ratio test). This difference in the proportion of observed gene conversion events between the two groups is highly significant ($P < 10^{-10}$; Fisher's exact test).

Ribosomal protein non-coding regions do not show evidence of gene conversion. For each of the RP gene pairs considered above, we measured the sequence identity in upstream non-coding regions. Among the pairs considered, 15 RP ohnolog pairs had local alignment scores significantly larger than would be expected for unrelated regions. For these pairs, we compared the alignment score S1,2 of the ohnolog pair (Scer1, Scer2) to the scores from the comparison of each paralog to its respective ortholog in *S. bayanus* (i.e., S1,B for Scer1, Sbay and S2,B for Scer2, Sbay). Cases where S1,2 > S1,B , S2,B were interpreted as evidence of upstream gene conversion. We found that only 1/15 (6%) of the pairwise non-coding alignments showed evidence of gene conservation compared to the 12/15 (80%) in the coding regions (from the analysis above;

Table 4). This difference in the prevalence of conversion events between the two groups is highly significant (Table 4 - Prevalence of gene conversion in coding region and non-coding regions.). Interestingly, when this same approach is applied to 39 MP genes, we find very few instances of gene conversion in either region (<10%) and no significant difference in the proportion of conversion events between the non-coding and coding regions (Table 4).

Table 4 - Prevalence of gene conversion in coding region and non-coding regions.

Gene class	Coding regions		Upstream regions		P[a]
	Gene conversion[b]	WGD[c]	Gene conversion[b]	WGD[c]	
RP[d]	12	3	1	14	< .001
MP[e]	1	38	5	34	.2

a: P-value for the test of equal proportions of gene conversion events in the coding and upstream regions (Fisher's exact test).
b: Cases where the two *S. cerevisiae* paralogs share higher sequence identity to each other than either does to its respective ortholog (see text).
c: Cases where at least one *S. cerevisiae* paralog shows higher sequence identity to its ortholog than to the other *S. cerevisiae* paralog.
d: Ribosomal protein gene duplicates.
e: Metabolic gene duplicates.

Analysis of duplicated ribosomal protein gene expression networks. We calculated the network partitioning that resulted in the fewest number of crossing edges for the ribosomal proteins (Figure 5). For these purposes, we defined an edge between any two genes if they shared a correlation (Pearson's r) in gene expression of 0.8 or greater across the set of more than 300 experiments (a threshold of 0.75 produced similar results; data not shown). The network in this analysis showed 102 crossing edges (Figure 5), which, although it appears to be a large number, is significantly smaller than the number of crossing edges seen in any of the randomized networks (P < 0.001). It is also relevant to note the extreme degree of asymmetry evident in this figure: the paralogous ribosomal proteins, despite their sequence similarity, do not have identical expression patterns.

Discussion

We have found that while duplicated RPs created by WGD show strong evidence of gene conversion in their coding regions, the same is not true of the upstream non-coding regions. Such conservation in the coding sequences of RPs is not unexpected as these proteins are highly conserved across a wide range of taxa (Bergmann et al. 2004; McCarroll et al. 2004; Stuart et al. 2003). RPs are also somewhat unusual in their response to genome duplication: they have survived in excess after other WGDs in addition to the yeast WGD (Aury et al. 2006; Blanc and Wolfe 2004; Maere et al. 2005; Seoighe and Wolfe 1999).

One obvious explanation for the similarity in RP coding sequences is selection for high dosages of these proteins. Indeed, there is some evidence for dosage benefits from RP gene duplication (Koszul et al. 2004). However, this explanation is not wholly convincing, particularly as we did not observe these same patterns of gene conversion in yeast metabolic genes, despite the fact that they also likely survived in duplicate partly due to dosage selection (Kuepfer et al. 2005).

Moreover, a number of recent analyses have demonstrated that the duplicated yeast RPs are not, in fact, functionally interchangeable. Thus, several RPs, but not their paralogs, have been shown to be essential for determining bud location in *S. cerevisiae* (Ni and Snyder 2001) and for localizing proteins to that bud (Komili et al. 2007). An equally intriguing case is the difference in protein localization between the RP paralogs Rpl7a and

Rpl7b. Rpl7a is much more highly expressed than is Rpl7b (Ghaemmaghami et al. 2003) but while Rpl7a is only found in the cytoplasm, Rpl7b, despite its lower abundance, is found both in the cytoplasm and in the nucleolus (Kim et al. 2009). This difference does not appear to be caused by differences in the coding sequences of the two genes: replacing the RPL7B sequence with that from RPL7A does not alter localization (Kim et al. 2009). These authors propose that the localization difference is instead driven by preferential incorporation of Rpl7a into ribosomal subunits, meaning that the free protein is rarely present at the site of ribosome subunit assembly in the nucleolus. However, the origins of this difference in incorporation rate remain unclear given the apparent functional equivalence of the two protein sequences.

We are still in the early stages of integrating these diverse observations regarding RP biology. One obvious explanation for the preservation of duplicate genes with (nearly) identical coding sequences is the subfunctionalization model of Force and coauthors (1999a). Under this model, the fact that RPs have what is essentially a generic function (protein synthesis), could imply that the selectively relevant variable is total protein abundance, with the relative contribution of the two paralogs to that abundance being effectively neutral. Subfunctionalization could in these circumstances be either quantitative (only expression of both paralogs gives sufficient protein product) or qualitative (the expression of the two paralogs varies with respect to each other temporally) or a mixture of the two.

As an aside, we note that our inference of subfunctionalization in the function of these ohnologs does not necessarily imply that the selective processes at work were purely neutral (as originally proposed by Force et al.,(1999a)).

Instead, while the large population sizes of yeasts may make such neutral partitioning relatively rare, functional partitioning through other mechanisms remains possible (Innan and Kondrashov 2010). Our network analysis supports a general process of subfunctionalization, showing as it does groups of co-functional paralogs (i.e., network subfunctionalization). In the future, it will be useful to study the temporal and spatial patterns of RP gene expression to discover whether the relative dosages of the paralogs varies across conditions.

Acknowledgements:

We would like to thank K. Byrne for assistance with the Yeast Gene Order Browser (http://wolfe.gen.tcd.ie/YGOB/). This work was partly supported by the Reproductive Biology Group of the Food for the 21st Century program at the University of Missouri.

Chapter 4

Molecular evolution in the yeast transcriptional regulation network

Annette M. Evangelisti

Coauthor:
Andreas Wagner
Department of Biology, The University of New Mexico

This originally appeared in *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution*, 2004, **302B**: 392-411.

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 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 expected if loss of binding played the only role in the divergence of duplicate genes.

Introduction

Transcriptional regulators and the genes whose expression they regulate – their target genes – form large gene regulation networks (Guelzim et al. 2002; Lee et al. 2002; Perez-Rueda and Collado-Vides 2000; 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 (Ito et al. 2001; Lee et al. 2002; Salgado et al. 2004; Uetz et al. 2000; von Mering et al. 2002). 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 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 (Chung et al. 2003; Sole et al. 2002; van Noort et al. 2004; Wagner 2001; Wagner 2003).

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 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 example, the question how gene functions diverge after gene duplication.

Gene duplications play dual roles in evolution. On one hand, gene duplicates that retain similar functions can be a source of gene redundancy, which may buffer organisms against mutations (Conant and Wagner 2004; Gonzalez-

Gaitan et al. 1994; Gu et al. 2003; Nowak et al. 1997; Wagner 1999; Wang et al. 1996). On the other hand, gene duplicates that diverge in function contribute to evolutionary innovation on the biochemical level (Briscoe 2001; Hughes 1994; Zhang et al. 1998). 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. Especially the last mode of divergence has generated considerable recent attention, because it has been argued that it can account for the maintenance of many gene duplicates in eukaryotic genomes (Force et al. 1999a; 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 (Ashburner et al. 2000). They include the biological process a gene acts in, its product’s sub-cellular 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 1999). 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. 1988; Faik et al. 1988); thymidine phosphorylase, which catalyzes the dephosphorylation of thymidine and deoxyuridine, and is the same as an endothelial growth factor (Furukawa et al. 1992; Haraguchi et al. 1994); aconitase, an enzyme in the tricarboxylic acid cycle, which also serves as a translational regulator of ferritin expression (Kennedy et al. 1992); 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 1999). 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 determine their rate of divergence (Gu et al. 2002; Wagner 2000). 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. While primarily descriptive, our analysis provides preliminary answers to the questions raised above, as well as 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 (Bhan et al. 2002; Guelzim et al. 2002; Lee et al. 2002; Perez-Rueda and Collado-Vides 2000; 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 *Saccharomyces 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 experiment (Knop et al. 1999) 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 than 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, 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 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 *Saccharomyces 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 non-synonymous 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. 1997) 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 1988). 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 only five duplicate transcriptional regulators. We thus modified GenomeHistory to carry out a local alignment, using the Smith Waterman algorithm (Smith and Waterman

1981), 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 (Ks) and the non-synonymous divergence (Ka) of every pair of duplicate genes, using a method established by Yang and Nielsen (Yang and Nielsen 2000). Because of the well known multiple substitution problem (Li 1997), both synonymous and non-synonymous divergence estimates show limited reliability for Ka(s)>1.0 respectively. Therefore, we retained only gene pairs with Ka<1 for further analysis.

Orthologous genes. A recent study by Kellis and collaborators reported the genomic DNA sequences of three yeasts, *Saccharomyces mikatae*, *Saccharomyces paradoxus*, and *Saccharomyces bayanus*, closely related to *Saccharomyces cerevisiae* (Kellis et al. 2003). From this study, we used data on synonymous divergence Ks and nonsynonymous divergence Ka between *S. cerevisiae* genes and their unambiguous orthologues from the yeast *Saccharomyces mikatae* (file 'b.KaKs_details-5.xls' at http://www.broad.mit.edu/annotation/fungi/comp_yeasts/). We also used the ratio of non-synonymous to synonymous divergence Ka/Ks averaged for orthologs in the 3 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 non-significant 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 *Saccharomyces 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 6 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. Both the distribution of indegrees and outdegrees have previously been characterized for transcriptional regulation networks, and we will not belabor them here (Bhan et al. 2002; Featherstone and Broadie 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 7 shows an undirected graph whose nodes correspond to only duplicated target genes,

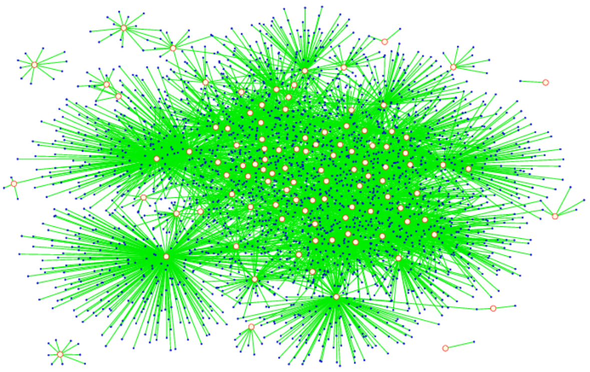


Figure 6 - A graph representation of the transcriptional regulation network.

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.

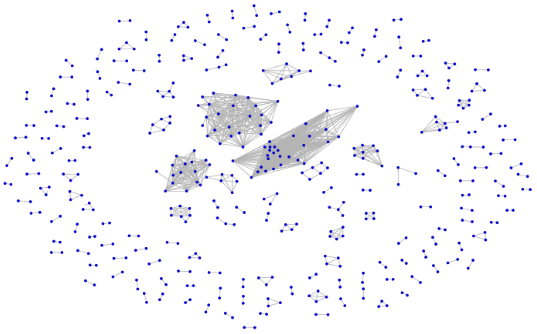


Figure 7 - Gene duplications among target genes of transcriptional regulators.

Blue nodes represent target genes. A gray edge connects two nodes if these two nodes are gene duplicates with amino acid divergence Ka < 1.0.

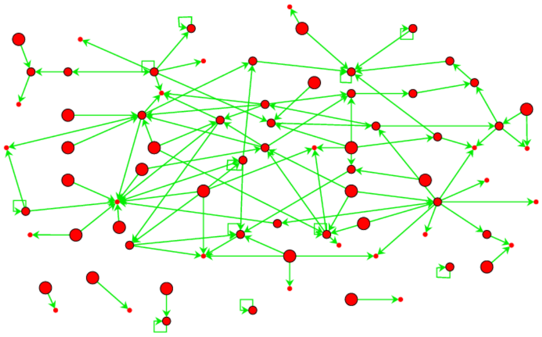


Figure 8 - Regulatory interactions among transcriptional regulators.

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.

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 Figure 6 andFigure 7 shows 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 6. 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 (Lee et al. 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 Ks>1. This and the small number of duplicate regulators render a meaningful statistical analysis of functional divergence after regulator duplication difficult.

Figure 8 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 that have large numbers of target genes are underrepresented in this network Table 5). 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 trend: The higher a regulator’s number of target genes, the smaller the fraction of regulators whose expression it regulates (Figure 9). Again, this statistical association is highly significant (Kendall’s $\tau = -0.50$; $P = 1.17 \times 10^{-7}$; n = 54).

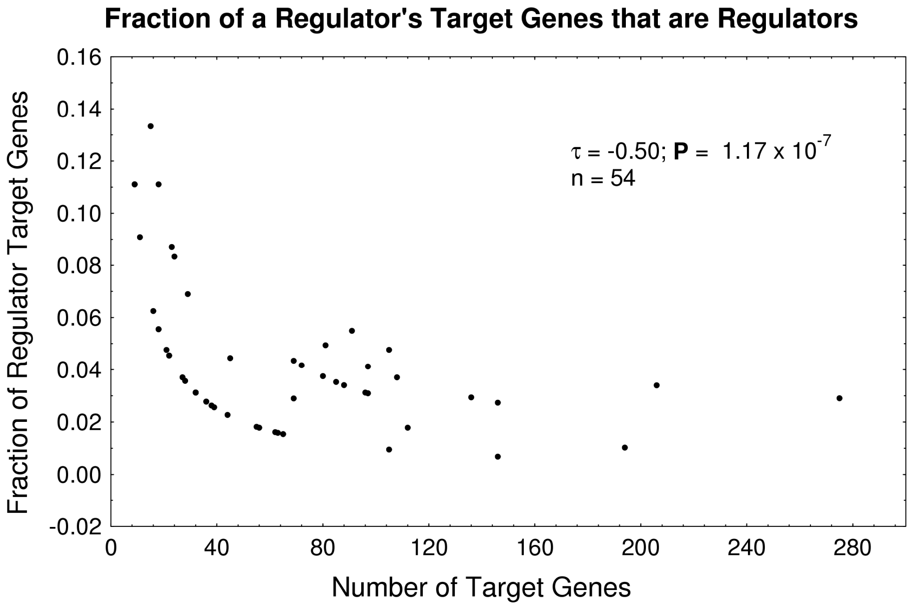


Figure 9 - Number of target genes of transcriptional regulators (horizontal axis) plotted against the fraction of a regulators target genes that are regulators (vertical axis).

Kendall's τ = -0.50; P = 1.17 x 10^{-7}; n = 54.

Table 5 - 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 Pr(x ≤ 2) = 5.08 x 10^{-13}

The null hypothesis is that the columns derive from the same (binomial) distribution. We calculated the parameters defining a binomial distribution from the left column (n = 54, p = 26/54) and used the value of p to find the probability that x, the number of regulators that do not regulate other regulators and that have more than 50 target genes, is smaller than or equal to 2, Pr(x ≤ 2) = 5.08 x 10^{-13}.

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 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? Figure 10 and Figure 11 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 10 shows a weak negative association between a regulator's number of target genes and growth rate on rich medium. 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 rate among five different media when the regulator is eliminated (Figure 11).

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 12 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

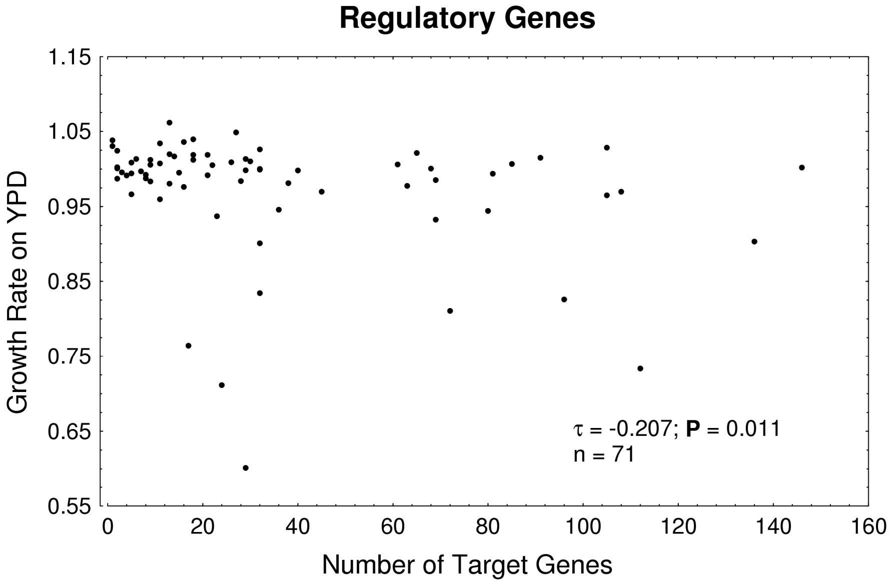


Figure 10 - Growth rate and highly connected regulators.

The horizontal axis shows a regulator's number of target genes. 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 τ = -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.

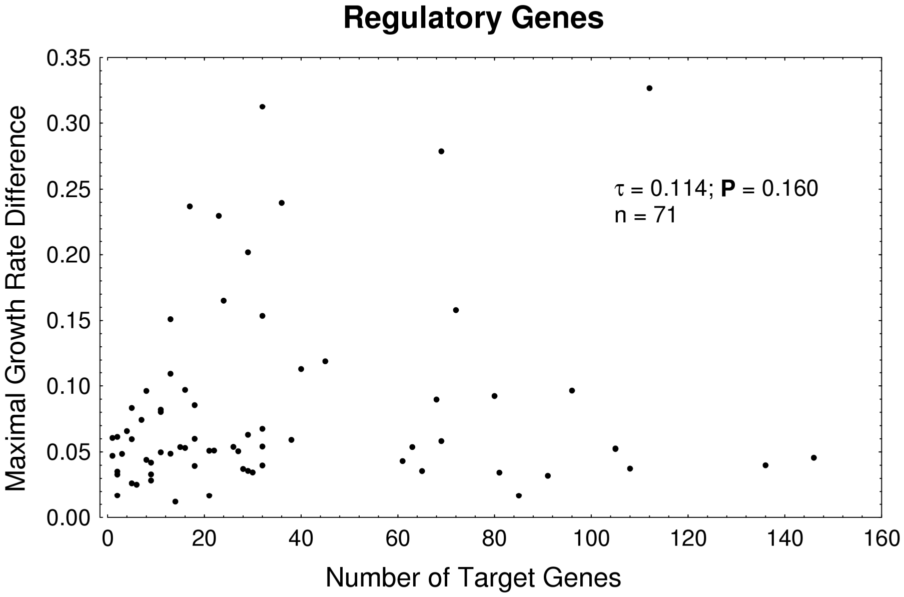


Figure 11 - Growth rate and highly connected regulators.

The horizontal axis shows a regulator's number of target genes. 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 = 0.114; 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.

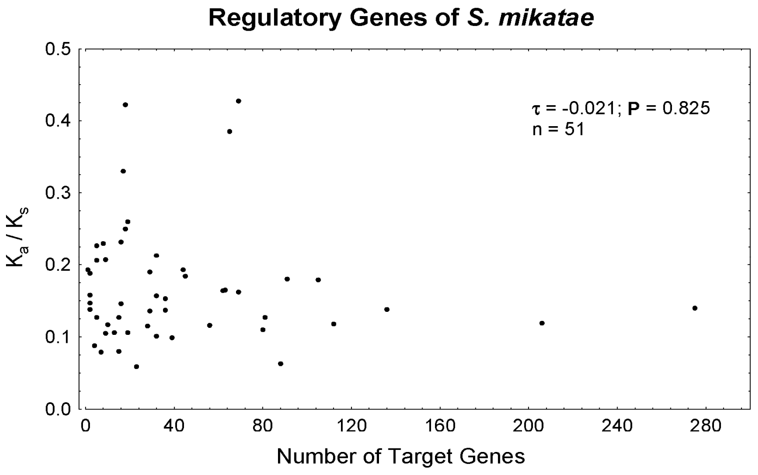


Figure 12 - Do highly connected genes evolve at different rates?.

The horizontal axis shows a regulator's outdegree, that is, its number of target genes. The vertical axis shows the ratio Ka/Ks of non-synonymous 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 τ = -0.021; P = 0.825; n = 51

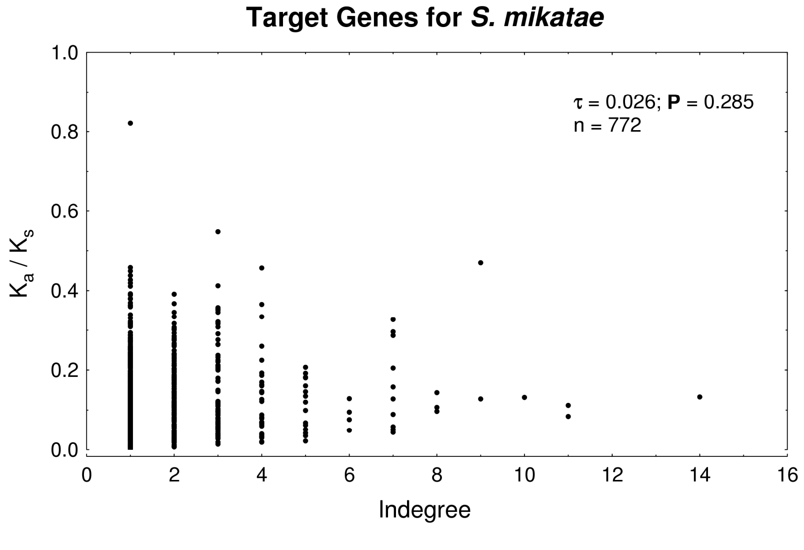


Figure 13 - Do highly connected genes evolve at different rates?

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 Ka/Ks of non-synonymous 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).

S. cerevisiae. We plotted the ratio of non-synonymous to synonymous divergence Ka/Ks as an indicator of evolutionary constraints (Li 1997). It shows that highly connected regulators do not evolve at rates different from other regulators (Kendall's τ = -0.021; P = 0.825; n = 51). Identical results (not shown) hold for non-synonymous divergence Ka instead of Ka/Ks for *S. mikatae*, and also for the average ratio Ka/Ks among orthologs 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 6, 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 14 shows that the answer is no (Kendall's τ = -0.104; P = 0.114; n = 105).

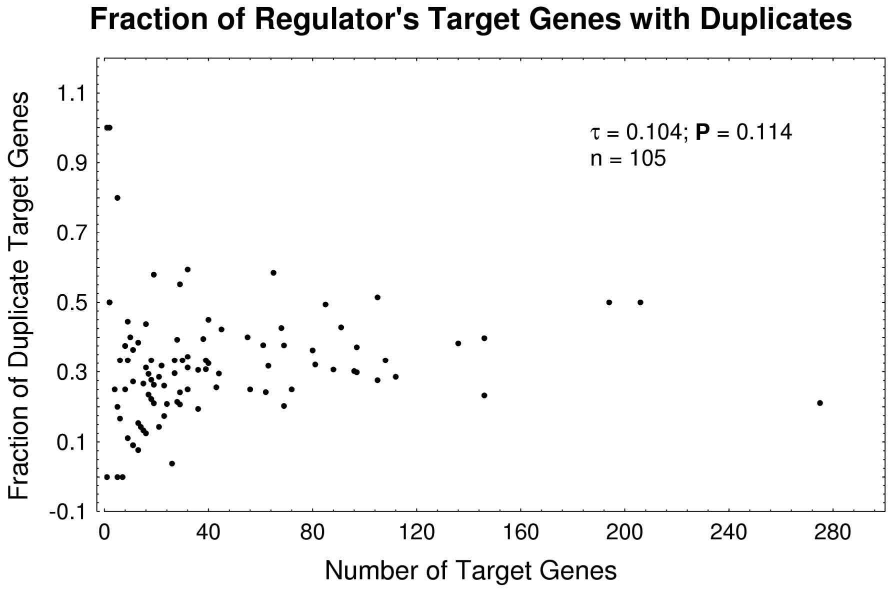


Figure 14 - No significant association exists between a regulator's number of target genes (horizontal axis), and the fraction of target genes that have undergone duplication (vertical axis).

Table 6 - 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

The null hypothesis is that the columns derive from the same (binomial) distribution. We calculated the parameters defining a binomial distribution from the left column (n = 23, p = 7/23) and used the value of p to find the probability that x, the number of single copy regulators that have more than 80 target genes, is greater than or equal to 10, Pr(x ≥ 10) = 0.06

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? Figure 15 andFigure 16 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 15 shows that there is no statistically significant association between indegree and growth rate on rich medium. The same holds for Figure 16, 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 15), as well as a negative association between the maximal difference in growth rate among five media and indegree (Figure 16). 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 13 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 non-synonymous to synonymous divergence Ka/Ks as an indicator of evolutionary constraints (Li 1997). It shows that highly connected target genes do not evolve at different rates than

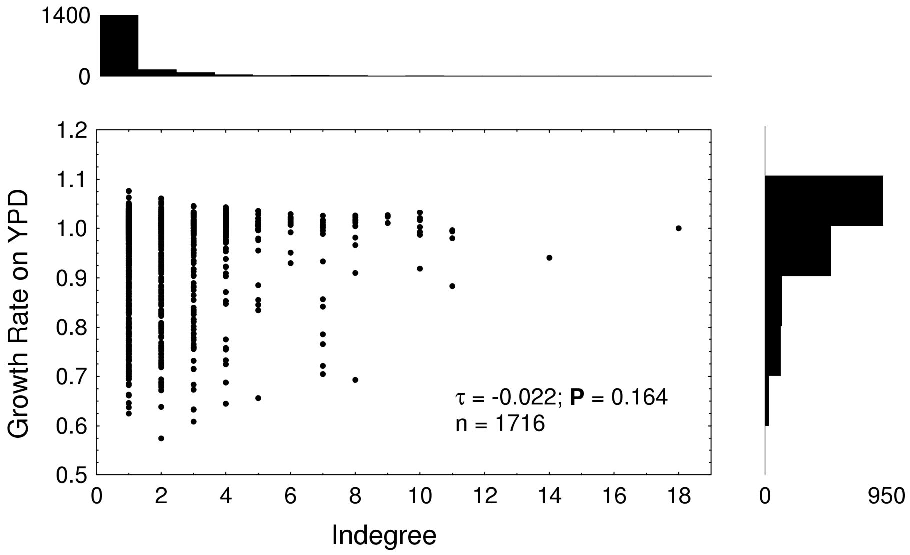


Figure 15 - Growth rate and highly connected target genes.

The horizontal axis shows the number of regulators that bind to the regulatory region of a target gene. 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 τ = -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

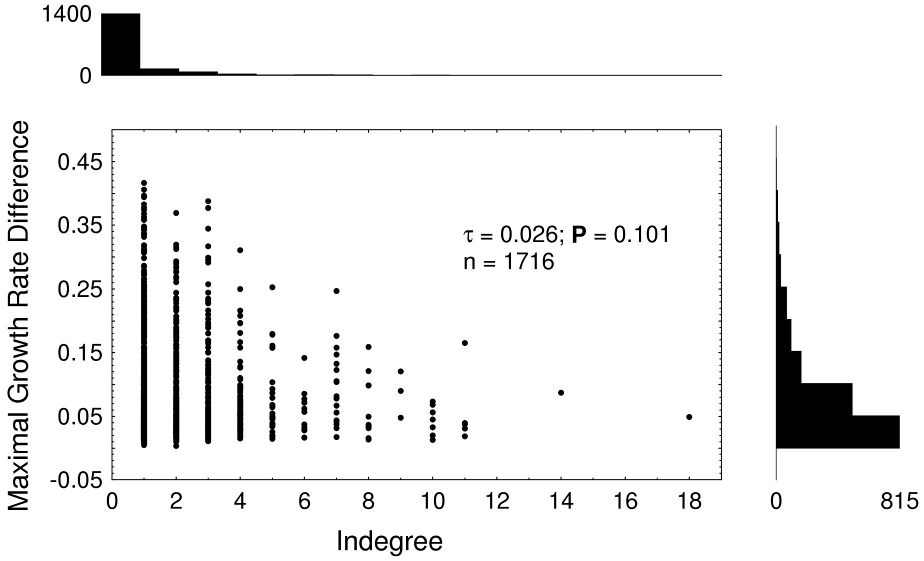


Figure 16 - Growth rate and highly connected target genes.

The horizontal axis shows the number of regulators that bind to the regulatory region of a target gene. 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.

other target genes (Kendall's τ = 0.026; P = 0.285; n = 772). Identical results (not shown) hold for non-synonymous divergence Ka instead of Ka/Ks for *S. mikatae*, as well as for average ratio Ka/Ks for orthologs in the 3 species pairs (*S. bayanus* , *S. paradoxus* and *S. mikatae*), where we averaged the ratio Ka/Ks of the 3. All results show that highly connected target genes do not evolve at different rates than 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 7, 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 ($P = 1.9 \times 10^{-8}$; n = 492). 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 17 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 $\tau = 0.149$) but highly significant ($P = 2.1 \times 10^{-27}$; 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 duplication. We could not address this question for the transcriptional regulators, because of their

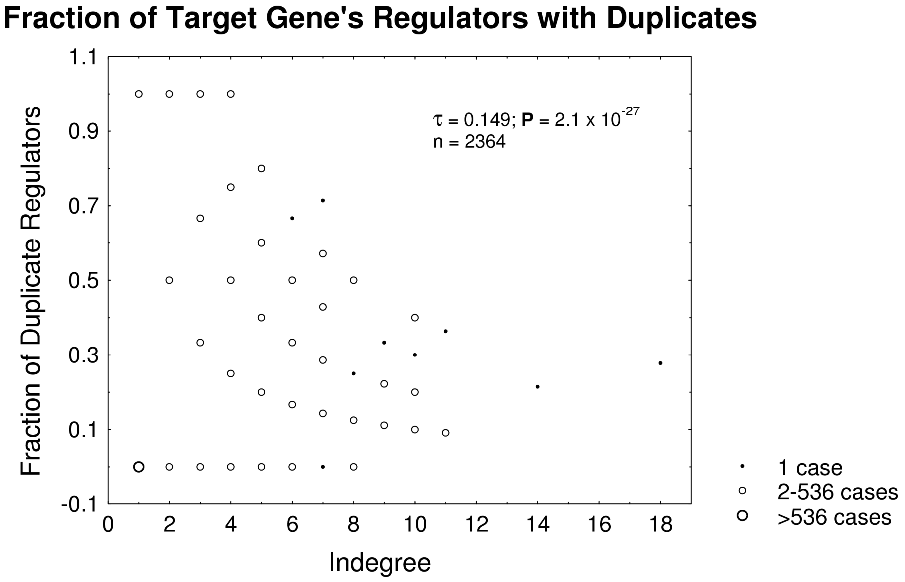


Figure 17 - Regulators of highly connected target genes are more likely to undergo gene duplication

Kendall's τ = 0.149; P = 0.210 x 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.

Table 7 - An exact binomial test to compare the binomial distribution of each column in the table.

InDegree	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 ≤ 5) = 1.90 x 10^{-8}

The null hypothesis is that the columns derive from the same (binomial) distribution. We calculated the parameters defining a binomial distribution from the right column (n = 1871, p = 30/1871) and used the value of p to find the probability that x, the number of single copy target genes that have more than 7 regulators that influence their expression, is less than or equal to 5, Pr(x ≤ 5) = 1.90 x 10^{-8}.

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 T1 and T2, the number d1 of regulators binding to the regulatory region of T1, the number d2 of regulators binding to the regulatory region of T2, as well as the number d12 of regulators binding both target regulatory regions. The fraction of shared regulators is then properly defined as d12/(d1+d2-d12). Figure 18 and Figure 19 shows this fraction of shared regulators as a function of the non-synonymous divergence (Ka) and synonymous or silent divergence (Ks), 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

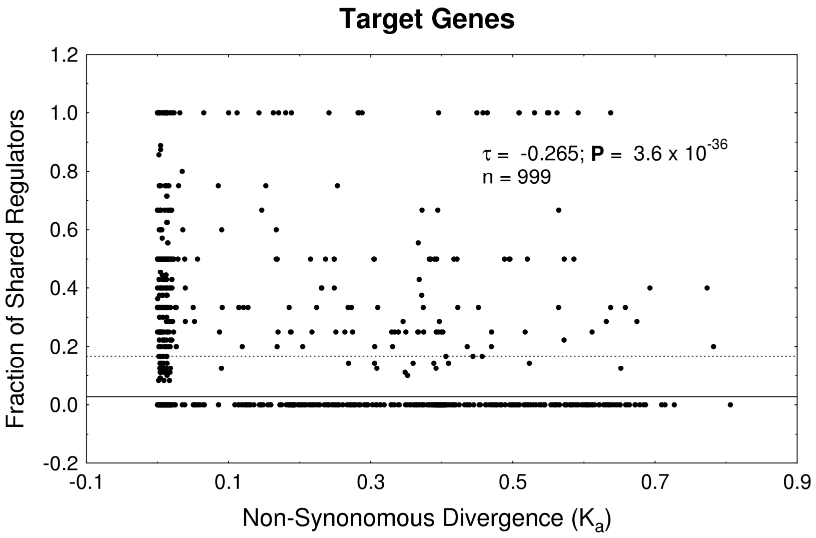


Figure 18 - Negative association between sequence divergence and regulators shared by duplicate target genes.

The vertical axes show the fraction of transcriptional regulators bound to both regulatory regions of a duplicate 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. Sequence divergence as measured by the non-synonymous divergence Ka. (Kendall's τ = -0.265; P = 3.60 x 10^{-36}; n = 999

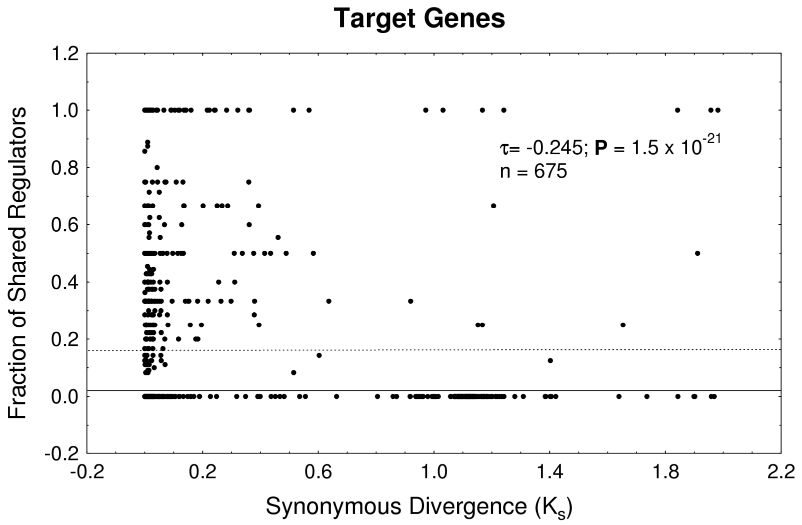


Figure 19 - Negative association between sequence divergence and regulators shared by duplicate target genes.

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). Sequence divergence as measured by the synonymous divergence Ks. (Kendall's τ = -0.245; P = 1.50 x 10^{-21}; n = 675).

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.

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. 1999a). 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 d1+d2 = 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 d1+d2 will decrease after duplication and approach d1+d2 = d, the number of binding interactions before duplication. Conversely, if divergence involved largely addition of new interactions, then d1+d2 should increase after duplication. Figure 20 clearly shows that the second scenario is not the case: d1+d2 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, d12=0 so gene pairs share no transcriptional regulators. If loss of transcription factor binding sites is exclusively responsible for the divergence of duplicates, then the combined degrees d1+d2 of completely diverged duplicates should be identical to the degree d typically found in single-copy genes. Figure 21 shows that this is not the case, regardless of whether one examines very young (Ks<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.

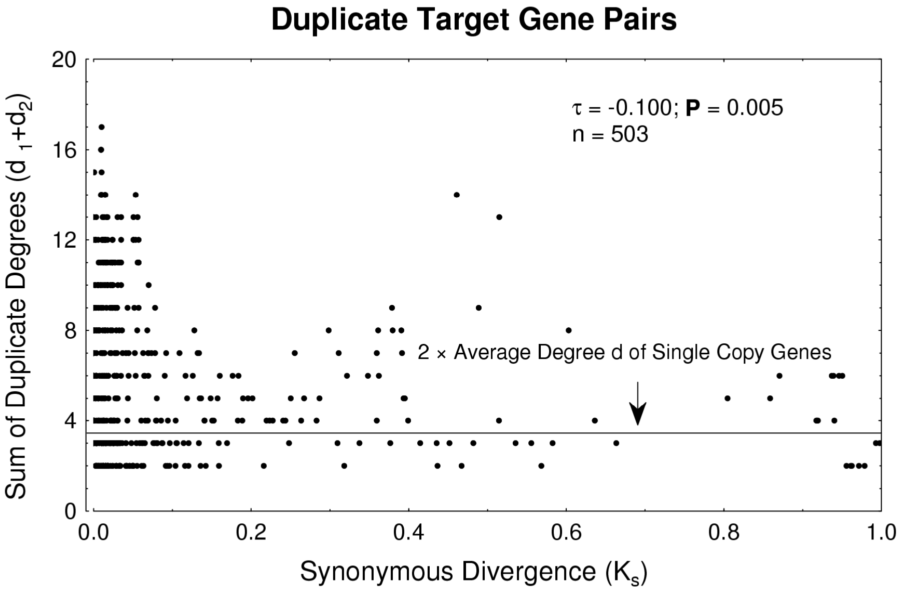


Figure 20 - Sequence divergence and divergence of the number of regulators affecting duplicate target genes.

The horizontal axis indicates synonymous sequence divergence Ks between duplicate target genes. a) The vertical axis indicates the sum d1+d2 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 = -0.100; P = 0.005; n = 503

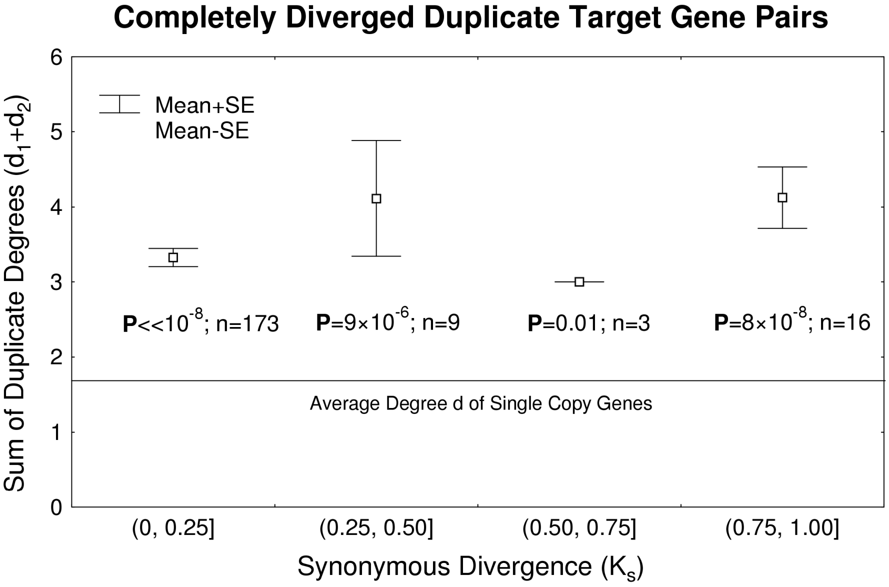


Figure 21 - Sequence divergence and divergence of the number of regulators affecting duplicate target genes Includes only duplicate target gene pairs that have completely diverged since their duplication, i.e., gene pairs where d12 = 0.

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 1995). 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.

Discussion

Our primary focus here was a descriptive analysis of the largest available genome-scale 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 multi-protein complexes, not all of whose members contact DNA (Ptashne 1988). 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. 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 (Bhan et al. 2002; Guelzim et al. 2002; Lee et al. 2002; Perez-Rueda and Collado-Vides 2000; 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 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 (Fraser et al. 2002; Fraser et al. 2003; Hahn et al. 2004; Jeong et al. 2001; Jordan et al. 2003a; Jordan et al. 2003b). The disadvantage of protein interaction data is that such data contain an especially large amount of experimental noise (Gilchrist et al. 2004; von Mering et al. 2002), 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 *Saccharomyces 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 (Gu et al. 2002; Wagner 2000) 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. 1999b; 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. 1999b; 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.

Conclusion

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 (Conant and Wagner 2003a; Milo et al. 2002; Shen-Orr et al. 2002). Even basic regulatory patterns, such as those in the preceding two paragraphs, do currently 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.

List of References

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

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

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

Aury, J.M., O. Jaillon, L. Duret, B. Noel, C. Jubin, B.M. Porcel, B. Segurens, V. Daubin, V. Anthouard, N. Aiach, O. Arnaiz, A. Billaut, J. Beisson, I. Blanc, K. Bouhouche, F. Camara, S. Duharcourt, R. Guigo, D. Gogendeau, M. Katinka, A.M. Keller, R. Kissmehl, C. Klotz, F. Koll, A. Le Mouel, G. Lepere, S. Malinsky, M. Nowacki, J.K. Nowak, H. Plattner, J. Poulain, F. Ruiz, V. Serrano, M. Zagulski, P. Dessen, M. Betermier, J. Weissenbach, C. Scarpelli, V. Schachter, L. Sperling, E. Meyer, J. Cohen, and P. Wincker. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444:** 171-178.

Bergmann, S., J. Ihmels, and N. Barkai. 2004. Similarities and differences in genome-wide expression data of six organisms. *PloS Biology* **2:** 85-93.

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

Birchler, J.A. and R.A. Veitia. 2007. The gene balance hypothesis: From classical genetics to modern genomics. *Plant Cell* **19:** 395-402.

Blanc, G., A. Barakat, R. Guyot, R. Cooke, and I. Delseny. 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* **12:** 1093-1101.

Blanc, G. and K.H. Wolfe. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16:** 1679-1691.

Briscoe, A. 2001. Functional diversification of lepidopteran opsins following gene duplication. *Molecular and Cellular Biology* **18:** 2270-2279.

Bussemaker, H., H. Li, and E. Siggia. 2001. Regulatory element detection using correlation with expression. *Nature Genetics* **27:** 167-171.

Byrne, K.P. and K.H. Wolfe. 2005. The Yeast Gene Order Browser: Combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research* **15:** 1456-1461.

Byrne, K.P. and K.H. Wolfe. 2007. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics* **175:** 1341-1350.

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

Chung, F., L. Lu, T. Dewey, and D. Galas. 2003. Duplication models for biological networks. *Journal of Computational Biology* **10:** 677-687.

Comai, L. 2005. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* **6:** 836-846.

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

Conant, G. and A. Wagner. 2003a. Convergent evolution in gene circuits. *Nature Genetics* **34:** 264-266.

Conant, G. and A. Wagner. 2004. Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. *PROCEEDINGS OF THE ROYAL SOCIETY OF LONDON SERIES B-BIOLOGICAL SCIENCES* **271:** 89-96.

Conant, G.C. and A. Wagner. 2003b. Asymmetric sequence divergence of duplicate genes. *Genome Research* **13:** 2052-2058.

Conant, G.C. and K.H. Wolfe. 2006. Functional partitioning of yeast co-expression networks after genome duplication. *PLoS Biology* **4:** e109.

Conant, G.C. and K.H. Wolfe. 2008. Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. *Genetics* **179:** 1681-1692.

Costanzo, M., J. Hogan, M. Cusick, B. Davis, A. Fancher, P. Hodges, P. Kondu, C. Lengieza, J. Lew-Smith, C. Lingner, K. Roberg-Perez, M. Tillberg, J. Brooks, and J. Garrels. 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 Research* **28:** 73-76.

Cusack, B.P. and K.H. Wolfe. 2007. Not born equal: Increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. *Molecular Biology and Evolution* **24:** 679-686.

Dermitzakis, E.T. and A.G. Clark. 2001. Differential selection after duplication in mammalian developmental genes. *Molecular Biology and Evolution* **18:** 557-562.

Dietrich, F.S., S. Voegeli, S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, R. Pohlmann, P. Luedi, S.W. Choi, R. A., A. Flavier, T.D. Gaffney, and P. Philippsen. 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304:** 304-307.

Dujon, B., D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuveglise, E. Talla, N. Goffard, F. L., A. M., V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J.M. Beckerich, E. Beyne, C. Bleykasten, A. Boisrame, J. Boyer, L. Cattolico, F. Confanioleri, A. De Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J.M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G.F. Richard, M.L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wesolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker, and J.L. Souciet. 2004. Genome evolution in yeasts. *Nature* **430:** 35-44.

Enyenihi, A.H. and W.S. Saunders. 2003. Large-scale functional genomic analysis of sporulation and meiosis in *Saccharomyces cerevisiae*. *Genetics* **163:** 47-54.

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

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

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

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

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

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

Freeling, M. and B.C. Thomas. 2006. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Research* **16:** 805-814.

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

Ghaemmaghami, S., W. Huh, K. Bower, R.W. Howson, A. Belle, N. Dephoure, E.K. O'Shea, and J.S. Weissman. 2003. Global analysis of protein expression in yeast. *Nature* **425:** 737-741.

Gilchrist, M.A., L.A. Salter, and A. Wagner. 2004. A statistical framework for combining and interpreting proteomic datasets. *Bioinformatics* **20:** 689-U290.

Gonzalez-Gaitan, M., M. Rothe, E. Wimmer, H. Taubert, and H. Jackle. 1994. Redundant functions of the genes knirps and knirps-related for the establishment of anterior *Drosophila* head structures. *Proceedings of the National Academy of Sciences of the United States of America* **91:** 8567-8571.

Gordon, J.L., K.P. Byrne, and K.H. Wolfe. 2009. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *Plos Genetics* **5**.

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

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

Guelzim, N., S. Bottani, P. Bourgine, and F. Kepes. 2002. Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genetics* **31:** 60-63.

Hahn, M., G. Conant, and A. Wagner. 2004. Molecular evolution in large genetic networks: Does connectivity equal constraint? *Journal of Molecular Biology* **58:** 203-211.

Haldane, J.B.S. 1933. The part played by recurrent mutation in evolution. *American Naturalist* **67:** 679-682.

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

Hughes, A. 1994. The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London Series B-Biological Sciences* **256:** 119-124.

Hughes, M. and A. Hughes. 1993. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. *Molecular Biology and Evolution* **10:** 1360-1369.

Hughes, T.R., M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H.Y. Dai, Y.D.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S.H. Friend. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102:** 109-126.

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

Ito, T., T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America* **98:** 4569-4574.

Jeffery, C. 1999. Moonlighting proteins. *Trends in Biochemical Sciences* **24:** 8-11.

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

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

Jordan, I., Y. Wolf, and E. Koonin. 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 Evolutionary Biology* **3:** 5.

Kaeberlein, M., R.W. Powers, K.K. Steffen, E.A. Westman, D. Hu, N. Dang, E.O. Kerr, K.T. Kirkland, S. Fields, and B.K. Kennedy. 2005. Regulation of yeast replicative life span by TOR and Sch9 in response to nutrients. *Science* **310:** 1193-1196.

Katju, V. and M. Lynch. 2003. The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. *Genetics* **165:** 1793-1803.

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

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

Kennedy, M., L. Mendemueller, G. Blondin, and H. Beinert. 1992. Purification and characterization of cytosolic aconitase from beef-liver and its relationship to the iron-responsive element binding-protein. *Proceedings of the National Academy of Sciences of the United States of America* **89:** 11730-11734.

Kim, T.Y., C. Ha, and W.K. Huh. 2009. Differential subcellular localization of ribosomal protein L7 paralogs in *Saccharomyces cerevisiae*. *Molecules and Cells* **27:** 539-546.

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

Komili, S., N.G. Farny, F.P. Roth, and P.A. Silver. 2007. Functional specificity among ribosomal proteins regulates gene expression. *Cell* **131:** 557-571.

Kondrashov, F., I. Rogozin, Y. Wolf, and E. Koonin. 2002. Selection in the evolution of gene duplications. *Genome Biology* **3**.

Kondrashov, F.A. and A.S. Kondrashov. 2006. Role of selection in fixation of gene duplications. *Journal of Theoretical Biology* **239:** 141-151.

Koszul, R., S. Caburet, B. Dujon, and G. Fischer. 2004. Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *Embo Journal* **23:** 234-243.

Kuepfer, L., U. Sauer, and L.M. Blank. 2005. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Research* **15:** 1421-1430.

Kurtzman, C.P. and C.J. Robnett. 2003. Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. *FEMS Yeast Research* **3:** 417-432.

Lavoie, H., H. Jogues, J. Mallick, A. Sellam, A. Natel, and M. Whiteway. 2010. Evolutionary tinkering with conserved components of a transcriptional regulatory network. *PLoS Biology* **8:** e1000329.

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

Li, W.-H. 1996. Rates of nucleotide substitution in primates and rodents and the generation time effect hypothesis. *Molecular Phylogenetics and Evolution* **5:** 182-187.

Li, W.-H. 1997. *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Lynch, M. and J. Conery. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151-1155.

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

Maere, S., S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu, M. Kuiper, and Y. Van de Peer. 2005. Modeling gene and genome duplications in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 5454-5459.

Maslov, S., K. Sneppen, K. Eriksen, and K.K. Yan. 2004. Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evolutionary Biology* **4**.

McCarroll, S.A., C.T. Murphy, S.G. Zou, S.D. Pletcher, C.S. Chin, Y.N. Jan, C. Kenyon, C.I. Bargmann, and H. Li. 2004. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nature Genetics* **36:** 197-204.

Meyer, A. and Y. Van de Peer. 2005. From 2R to 3R: Evidence for a fish-specific genome duplication (FSGD). *Bioessays* **27:** 937-945.

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

Needleman, S.B. and C.D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48:** 443-453.

Ni, L. and M. Snyder. 2001. A genomic study of the bipolar bud site selection pattern in *Saccharomyces cerevisiae*. *Molecular Biology of the Cell* **12:** 2147-2170.

Notredame, C., D.G. Higgins, and J. Heringa. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302:** 205-217.

Nowak, M.A., M.C. Boerlijst, J. Cooke, and J. Maynard-Smith. 1997. Evolution of genetic redundancy. *Nature* **388:** 167-171.

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Berlin.

Otto, S. 2007. The evolutionary consequences of polypoidy. *Cell* **131:** 452-462.

Papp, B., C. Pal, and L.D. Hurst. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424:** 194-197.

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

Planta, R.J. and W.H. Mager. 1998. The list of cytoplasmic ribosomal proteins of *Saccharomyces cerevisiae*. *Yeast* **14:** 471-477.

Prince, V. and F. Pickett. 2002. Splitting pairs: The diverging fates of duplicated genes. *Nature Reviews Genetics* **3:** 827-837.

Ptashne, M. 1988. How eukaryotic transcriptional activators work. *Nature* **335:** 683-689.

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

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

Scannell, D.R., K.P. Byrne, J.L. Gordon, S. Wong, and K.H. Wolfe. 2006. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440:** 341-345.

Scannell, D.R. and K.H. Wolfe. 2008. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Research* **18:** 137-147.

Semon, M. and K.H. Wolfe. 2007. Consequences of genome duplication. *Current Opinion in Genetics & Development* **17:** 505-512.

Seoighe, C. and K. Scheffler. 2005. Very low power to detect asymmetric divergence of duplicated genes. In *Comparative Genomics*, pp. 142-152.

Seoighe, C. and K.H. Wolfe. 1999. Yeast genome evolution in the post-genome era. *Current Opinion in Microbiology* **2:** 548-554.

Sharp, P.M. and E. Cowe. 1991. Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* **7:** 657-678.

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

Smith, T.F. and M.S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology* **147:** 195-197.

Sokal, R.R. and F.J. Rohlf. 1995. *Biometry: the principles and practice of statistics in biological research*. W. H. Freeman and Co., New York.

Sole, R., R. Pastor-Satorras, E.D. Smith, and T. Kepler. 2002. A model of large-scale proteome evolution. *Advances in Complex Systems* **5:** 43-54.

Soltis, D.E., V.A. Albert, J. Leebens-Mack, C.D. Bell, A.H. Paterson, C.F. Zheng, D. Sankoff, C.W. dePamphilis, P.K. Wall, and P.S. Soltis. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* **96:** 336-348.

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

Stone, J. and G. Wray. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. *Molecular Biology and Evolution* **18:** 1764-1770.

Stuart, J.M., E. Segal, D. Koller, and S.K. Kim. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302:** 249-255.

Tajima, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics Society of America* **135:** 599-607.

Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24:** 1596-1599.

Taylor, J.S. and J. Raes. 2004. Duplication and divergence: The evolution of new genes and old ideas. *Annual Review of Genetics* **38:** 615-643.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796-815.

Tuskan, G.A., S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R.R. Bhalerao, R.P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G.L. Chen, D. Cooper, P.M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Dejardin, C. Depamphilis, J.

Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjarvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J.C. Leple, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D.R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouze, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C.J. Tsai, E. Uberbacher, and P. Unneberg, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313:** 1596-1604.

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

Van de Peer, Y., S. Maere, and A. Meyer. 2009. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* **10:** 725-732.

Van de Peer, Y., J. Taylor, I. Braasch, and A. Meyer. 2001. The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. *Journal of Molecular Evolution* **53:** 436-446.

van Noort, V., B. Snel, and M.A. Huynen. 2004. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *Embo Reports* **5:** 280-284.

von Mering, C., R. Krause, B. Snel, M. Cornell, S. Oliver, S. Fields, and P. Bork. 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. *Journal of Evolutionary Biology* **12:** 1-16.

Wagner, A. 2000. Decoupled evolution of coding region and mRNA expression patterns after gene duplication: Implications for the neutralist-selectionist debate. *Proceedings of the National Academy of Sciences of the United States of America* **97:** 6579-6584.

Wagner, A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular Biology and Evolution* **18:** 1283-1292.

Wagner, A. 2003. How the global structure of protein interaction networks evolves. *Proceedings of the Royal Society of London Series B-Biological Sciences* **270:** 457-466.

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

Wittbrodt, J., A. Meyer, and M. Schartl. 1998. More genes in fish? *Bioessays* **20:** 511-515.

Wolfe, K. 2000. Robustness - it's not where you think it is. *Nature Genetics* **25:** 3-4.

Wolfe, K.H. and D.C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708-713.

Yang, Z. and R. Nielsen. 2000. Estimating synonymous and non-synonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17:** 32-43.

Zhang, J., H. Rosenberg, and M. Nei. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences of the United States of America* **95:** 3708-3713.