國立交通大學

生物科技系

博士論文

酵母菌基因體的演化分析研究

Evolutionary Analysis of the Yeast Genome

研究生：林勇欣

指導教授：黃鎮剛　教授

中華民國九十五年七月

酵母菌基因體的演化分析研究

Evolutionary Analysis of the Yeast Genome

研究生：林勇欣　　Student：Yeong-Shin Lin

指導教授：黃鎮剛　　Advisor：Jenn-Kang Hwang

國立交通大學

生物科技系

博士論文

A Dissertation
Submitted to Department of Biological Science and Technology
College of Biological Science and Technology
National Chiao Tung University
in partial Fulfillment of the Requirements
for the Degree of
PhD
in

Biological Science and Technology

July 2006

Hsinchu, Taiwan, Republic of China

中華民國九十五年七月

酵母菌基因體的演化分析研究

研 究 生：林勇欣　　　　　　　　　　指導教授：黃鎮剛博士

國立交通大學 生物科技系 博士班

摘　　要

Saccharomyces cerevisiae 經由古老的全基因體複製（whole-genome duplication，WGD）產生的重複基因（duplicate genes）中，有許多基因對之間表現出了較預期低許多的同義分歧（synonymous divergence，K_S），有些基因對之間的序列相似性比該基因和 *S. bayanus* 的同源基因（orthologue）之間的相似性更高，或者是和 *Kluyveromyces waltii*（在 WGD 發生之前分化的物種）的同源基因相比，擁有較慢的演化速度。這樣的減速演化（decelerated evolution）過去被歸因於重複基因之間的基因轉換（gene conversion）。在這篇論文的第一部份，我探討了四個物種中約三百個 WGD 基因對，以及這些基因對在非 WGD 物種中的同源基因，並因此發現了密碼子使用偏移（codon usage bias）以及蛋白質序列的保守性是造成重複基因對的減速演化兩個重要的原因。基因轉換只有在巨大的密碼子使用偏移或是非常保守的蛋白質序列存在的情況下，才能有效地對減速演化造成影響。我更進一步發現，突變型態的改變，或是 tRNA 編碼基因拷貝數目（tDNA copy number）的改變，會造成密碼子使用偏移的改變，也因此導致 *K. waltii* 及 *S. cerevisiae* 之間同義距離（K_S distance）的增加。很有趣地，有些蛋白質在 WGD 物種輻射狀種化之前表現出很快的演化速率，然而，在輻射狀種化之後他們的演化速率卻降得很低，甚至不再有變化。這代表功能上的保守性對於重複基因對的減速演化也有很大的影響。

接下來，我利用功能性基因體及蛋白質結構資料，探討蛋白質複雜性（蛋白質次單元種類的數目，protein complexity）對基因的可移除性（dispensability）及可複製性（duplicability）的影響。結果發現，基因可複製性在異複合體（由兩

種以上的次單元所構成的複合體，hetero-complexes）及同複合體（單體或是單一種次單元所構成的複合體，homo-complexes）之間存在明顯的差異。然而，基因的可移除性則是隨著蛋白質複合體次單元種類數目的增加而逐漸降低。這代表劑量平衡假說（dosage balance hypothesis）雖然能夠解釋蛋白質複合體的基因可複製性，卻無法完美的解釋不同的異複合體之間基因可移除性的差異。可能的情況是當一個異複合體次單元基因被剔除的時候，整個複合體的功能都會受到影響。因此這個基因被剔除造成的適性（fitness）影響會隨著蛋白質複雜性的升高而增加。此外我發現具有多功能區（multi-domain）的多肽基因和只具有單一功能區的多肽基因相比，有較低的可移除性和較高的可複製性。經由WGD產生的重複基因（不含核糖體次單元基因）普遍的比其他的重複基因擁有較高的可移除性。而屬於同一個複合體的次單元通常傾向有類似的表現量和類似的剔除適性影響。最後，我估計重複基因對於基因突變頑抗性（genetic robustness against null mutation）的貢獻大約是9%，比前人估計的要小許多。對酵母菌的基因可移除性來講，蛋白質的複雜性應該比重複基因的影響來的重大許多。

最後我所探討的是蛋白質演化速率。最近的研究指出，酵母菌中蛋白質的演化速率唯一的主要決定因素在於轉譯效率的選擇力（translational selection）上，這可以由mRNA和蛋白質的表現量以及密碼子適應值（codon adaptation index）來表示。本研究則說明蛋白質的結構其實也有舉足輕重的影響。為了要維持蛋白質的結構穩定，包埋在蛋白質內部或是位於蛋白質交互作用面的殘基（residue）通常比暴露在蛋白質表面接觸溶劑的殘基面對更強的演化拘束力。經由淨相關（partial correlation）分析發現，蛋白質中暴露殘基的百分比（P_{exposed}）可以解釋的演化速率變異量，可達到轉譯效率的選擇力所能解釋的一半以上。這個結果和功能性密度（functional density）假說是一致的，也就是說，蛋白質若擁有較多殘基與特定功能相關（如穩定蛋白質結構或是蛋白交互作用），會因此而傾向擁有較慢的演化速率。

Evolutionary Analysis of the Yeast Genome

Student: Yeong-Shin Lin Advisor: Dr. Jenn-Kang Hwang

Department of Biological Science and Technology
National Chiao Tung University

Abstract

Many *Saccharomyces cerevisiae* duplicate genes that were derived from an ancient whole-genome duplication (WGD) unexpectedly show a small synonymous divergence (K_S), a higher sequence similarity to each other than to orthologues in *S. bayanus*, or slow evolution compared to the orthologue in *Kluyveromyces waltii*, a non-WGD species. This decelerated evolution was attributed to gene conversion between duplicates. Using ~300 WGD gene pairs in four species and their orthologues in non-WGD species, the first part of my thesis shows that codon usage bias and protein sequence conservation are two important causes for decelerated evolution of duplicate genes, whereas gene conversion is effective only in the presence of strong codon usage bias or protein sequence conservation. Further, I found that change in mutation pattern or in tDNA copy number changed codon usage bias and increased the K_S distance between *K. waltii* and *S. cerevisiae*. Intriguingly, some proteins showed fast evolution before the radiation of WGD species but little or no sequence divergence between orthologues and paralogues thereafter, indicating that functional conservation after the radiation may also be responsible for decelerated evolution in duplicates.

In the second part, I studied the effects of protein complexity (here defined as the number of subunit types in a protein) on gene dispensability and gene duplicability using functional genomic and protein structural data. I found that the major distinction for gene duplicability in protein complexity is between

hetero-complexes, each of which includes at least two different types of subunits (polypeptides), and homo-complexes, which include monomers and complexes that consist of only subunits of one polypeptide type. However, gene dispensability decreases only gradually as the number of subunit types in a protein complex increases. These observations suggest that the dosage balance hypothesis can explain gene duplicability of complex proteins well, but cannot completely explain the difference in dispensabilities between hetero-complex subunits. It is likely that knocking out a gene coding for a hetero-complex subunit would disrupt the function of the whole complex, so that the deletion effect on fitness would increase with protein complexity. I also found that multi-domain polypeptide genes are less dispensable but more duplicable than single domain polypeptide genes. Duplicate genes derived from the whole genome duplication event in yeast are more dispensable (except for ribosomal protein genes) than other duplicate genes. Further, I found that subunits of the same protein complex tend to have similar expression levels and similar effects of gene deletion on fitness. Finally, I estimated that in yeast the contribution of duplicate genes to genetic robustness against null mutation is ~ 9%, smaller than previously estimated. In yeast, protein complexity may serve as a better indicator of gene dispensability than do duplicate genes.

The last part is a study related to protein evolutionary rate. Recently, translational selection, including mRNA expression, protein abundance, and codon adaptation index, has been suggested as the single dominant determinant of protein evolutionary rate in yeast. This study shows that protein structure is an important determinant as well. Buried residues, which are responsible for maintaining protein structure or located on a stable interaction surface, are under stronger constraints than solvent-exposed residues. Partial correlation analysis shows that the variance of evolutionary rate explained by the proportion of exposed residues (P_{exposed}) can reach more than half of that explained by translational selection. This result suggests that proteins with many residues involved in specific functions (e.g. maintaining structure or protein interaction) may evolve more slowly, which is consistent with the "functional density" hypothesis.

Acknowledgment

感謝黃鎮剛老師這五年來，在學業上的指導及生活上的照顧，讓我有機會進入結構生物學的殿堂，學習這些有趣新穎的題目。在跨領域的研究中，更要感謝黃老師耐心的指導、支持以及包容。不論是對學術研究工作的熱忱，或是待人處事的諄諄教悔，老師都是值得學習的榜樣。在芝加哥一年的生活，則非常感謝李文雄老師的關照，並開拓了我的視野。97年夏天認識李老師之後，陸陸續續受到老師很多的協助，把我當成自己的學生一樣的照顧。李老師對後輩的提攜總是不遺餘力，令人感動。當然，還要感謝曾晴賢老師多年來的關心指導，帶領我走入學術研究的道路。如果說，我現在對研究工作稱得上有一些熱情、堅持與自信，可以說都是在清華的那段歲月所奠定的基礎。我還要感謝楊昀良老師、何信瑩老師、丁照棣老師及黃憲達老師在口試的時候所提供的許多寶貴的建議，以及楊進木老師和彭慧玲老師在論文初期所給予的協助與指導。來自不同領域的意見總是能讓我有機會訓練自己從不同的角度思考，獲益良多。

從傳統生物學跨足到生物資訊，若不是鎮熊、景盛、存操、啟德屢次傾囊相授，並將工作站維護的友善而順暢，憑我這只懂皮毛、勉強堪用的電腦知識，事倍功半可能也不足以形容我的窘狀。另一方面，雖然離開清華到了不同的實驗室，和玉萍的討論總是能激發出許多有趣的想法。研究的路途一路走來，並不覺得孤單。還要感謝雅雯、立青、雲輝、Ay、Joshua、Jake在我待在芝加哥那段期間生活的照應及研究討論。特別是雅雯像家人般的噓寒問暖，協助日常生活大小事務，讓我雖然身處異鄉，仍倍覺溫暖。思民學長帶我入門，也為我立下典範，即使研究的領域不再相同，仍永遠會是我追隨學習的對象。志豪、玉菁、蔚倫，因為你們的努力，才讓我淺薄的想法有實現的機會。未曾謀面的顏聖紘學長指點我許多分類學的概念，這是網路世代才有的機緣。其他黃鎮剛實驗室、李文雄實驗室、曾晴賢實驗室的夥伴，以及我的大學同學們，雖然無法一一將你們的名字列出來，但是當我挫折沮喪時，有你們的加油打氣，遇到困難時，相互扶持，這份情誼，不會忘記。

最後要感謝上琪及家人的支持與體諒。我知道在親友面前很難解釋這十年來，我究竟是在做什麼。因為你們，我才能安心的把自己關進這象牙塔裡。你們永遠是我最好的避風港。

Contents

1896

Abbreviations

ACC	solvent accessible surface area
CAI	codon adaptation index
DSSP	database of secondary structure assignments for all PDB entries
K_A	nonsynonymous distance
K_S	synonymous distance
KS test	Kolmogorov-Smirnov test
MIPS	Munich Information Center for Protein Sequences
ORF	open reading frame
PDB	Protein Data Bank
P_{exposed}	proportion of solvent exposed residues in one protein
PSSM	position-specific scoring matrices
RelACC	relative solvent accessibility
SGD	*Saccharomyces* Genome Database
SVM	supporting vector machine
tDNA	transfer RNA coding gene
WGD	whole genome duplication
WGD genes	duplicated gene pairs derived from WGD
YDPM	Yeast Deletion Project and Proteomics of Mitochondria Database

Chapter 1

General Introduction

For decades, *Saccharomyces cerevisiae*, also called budding yeast or baker's yeast, has been one of the best model organisms for genetics, cellular mechanisms and physiological studies. This unicellular organism, unlike more complex eukaryotes, can be grown on various laboratory conditions, which is important for functional genomics analyses. Moreover, many of the substantial cellular functions are highly conserved from yeast to mammals. In 1996, *S. cerevisiae* became the first completely sequenced eukaryote (Goffeau et al., 1996). Comparative studies with the following sequenced eukaryotic genomes therefore ushered in the "post-genomic era". Here in this dissertation I used yeast genomic data to study a number of interesting evolutionary problems.

One important biological evolutionary mechanism is duplication, which provides extra genetic material that substantially can be remodelled into "novel" gene products. Lynch and Conery (2000) estimated gene duplication rate as one duplication per gene per 100 million years using three completely sequenced eukaryote genomes. Gao and Innan (2004) suggested that this rate should be two orders of magnitude lower because there are many extensive concerted evolution via gene conversion between duplicated genes. Using comparative genomics data from yeast species, Wong, Butler and Wolfe (2002) showed that almost the entire *S. cerevisiae* genome lies in duplicated sister regions, which suggests that the entire genome became duplicated at some point, followed by rearrangement and gene loss. When the genomes of *Kluyveromyces waltii* and *Ashbya gossypii* were completely sequenced (Dietrich et al., 2004; Kellis, Birren and Lander, 2004), the whole-genome duplication (WGD) event was finally confirmed and the ancient gene order was clearly identified. About 10% of the WGD genes have been preserved after massive gene loss. One member of each duplicate pairs often has evolved rapidly into a novel gene with a derived function (Wagner, 2002; Kellis, Birren and Lander, 2004).

While a majority of yeasts cannot grow in the absence of oxygen (aerobic yeasts), a majority species of the *Saccharomyces* complex can survive without any oxygen (Pronk, Steensma and van Dijken, 1996; Moller, Olsson and Piskur, 2001). The *Saccharomyces sensu stricto* yeasts, including *S. bayanus*, *S. cariocanus*, *S. cerevisiae*, *S. kudriavzevii*, *S. mikatae* and *S. paradoxus*, represent an isolated and well-supported monophyletic group with overall phenotypic similarity (Kurtzman and Robnett, 2003). Kwast et al. (2002) showed that some of the segmental duplicated genes have remodelled their expression to become dependent on the presence/absence of oxygen and glucose. The number of shared regulatory motifs in the duplicates decreases with evolutionary times, whereas the total number of regulatory motifs remains unchanged (Papp, Pal and Hurst, 2003a). The ribosomal proteins module can switch from employing one *cis*-element into another through the formation of redundant intermediate promoters harbouring both *cis*-elements in a tightly coupled configuration (Tanay, Regev and Shamir, 2005). The loss of a specific cis-regulatory element from dozens of genes following the apparent WGD event is connected to the change in gene expression for mitochondrial and cytoplasmic ribosomal proteins, and the emergence of the capacity for rapid anaerobic growth for these *Saccharomyces* complex yeasts (Ihmels et al., 2005). These studies suggested that WGD apparently provided new genes or regulatory elements, which were the basis for major remodelling of metabolism, including the development of an efficient glucose repression pathway and oxygen independence, in *Saccharomyces* complex.

Duplicate genes have also been used to explain the genetic robustness against mutations through functional compensation (e.g., Nowak et al., 1997; Gu et al., 2003; Conant and Wagner, 2004; Kafri, Bar-Even and Pilpel, 2005). Most *S. cerevisiae* genes are nonessential under laboratory conditions (Winzeler et al., 1999; Glaever et al., 2002; Steinmetz et al., 2002). Based on the dosage balance hypothesis (Veitia, 2002; Veitia, 2003), i.e., similar dosage among subunits in a protein complex is preferred, Papp, Pal and Hurst (2003b) and Yang, Lusk and Li (2003) have shown that protein complexity is an important determinant of gene duplicability. While based on the dosage theory (Kondrashov and Koonin, 2004), duplication is preferred for highly expressed genes. Deutschbauer et al. (2005) showed that the primary mechanism of haploinsufficiency, which is defined as a dominant phenotype in diploid organisms that are heterozygous for a loss-of-function allele, is due to insufficient protein production. He and Zhang (2005) suggested that duplicate genes have longer protein sequences, more functional domains, and more *cis*-regulatory motifs than singleton genes. They also proposed that non-important (dispensable) genes have higher probability to duplicate (He and Zhang, 2006), although other studies showed that duplicate genes are usually conserved, i.e., with remote orthologues (Davis and Petrov, 2004; Jordan, Wolf and Koonin, 2004).

Dispensable genes (Hirsh and Fraser, 2001; Yang, Gu and Li, 2003; Wall et al., 2005; Zhang and He, 2005) or proteins with more interactions (Fraser et al., 2002) have also been proposed to evolve slowly. However, highly expressed proteins also evolve slowly (Pal, Papp and Hurst, 2001; Rocha and Danchin, 2004; Wall et al., 2005), and usually tend to be indispensable; meanwhile, their interactions may have higher chance to be identified (Bloom and Adami, 2003; Pal, Papp and Hurst, 2003). Recently, Drummond, Raval and Wilke (2006) proposed that translational selection, including mRNA expression, protein abundance, and codon adaptation index, is the single dominant determinant of protein evolutionary rate.

In the following chapters, I first studied how gene conversion and codon usage bias affect the decelerated evolution of WGD genes. Then I collected protein complex data to analyze its relationships with gene dispensability and gene duplicability, and also reveal how protein complexity and protein structure may determine protein evolutionary rate.

E S A 1896

Chapter 2

Codon usage bias versus gene conversion in the evolution of yeast duplicate genes

Introduction

Gene conversion has been extensively studied in yeast (Petes and Hill, 1988; Petes, 2001). Recently, Kellis, Birren and Lander (2004) identified 60 gene pairs in *Saccharomyces cerevisiae* that were derived from an ancient whole-genome duplication (WGD) but showed a small sequence divergence. They suggested that these genes have undergone gene conversion for three reasons. First, in 90% of the cases, both paralogues show decelerated evolution (at least 50% slower than the orthologue in *Kluyveromyces waltii*). Second, nucleotides at fourfold degenerate codon positions for these genes are highly conserved. Third, in about half of the cases, the two paralogues in *S. cerevisiae* are closer in sequence to each other than either is to its syntenic orthologue in *S. bayanus*. Similarly, Gao and Innan (2004) attributed the small synonymous divergence (K_S) between ancient duplicated genes in yeast to gene conversion. However, most WGD gene pairs with decelerated evolution (Kellis, Birren and Lander, 2004) have an extremely strong codon-usage bias (Fig. 1). Codon-usage bias is known to increase with gene expression level (Coghlan and Wolfe, 2000; Akashi, 2001) and can slow down synonymous divergence between duplicate genes (Pal, Papp and Hurst, 2001). Therefore, I am interested to investigate whether codon usage bias rather than gene conversion is more important for the decelerated evolution.

Materials and Methods

Sequence data. I used the whole genome duplication (WGD) gene pairs in *S. cerevisiae* and their orthologues in *K. waltii* (Kellis, Birren and Lander, 2004) and *Ashbya gossypii* (Dietrich et al., 2004), and included their syntenic orthologues from three other species, *S. bayanus*, *S. mikatae* and *S. paradoxus* (Kellis et al., 2003). All sequences were aligned using the amino acid sequences with CLUSTAL W 1.83 (Thompson, Higgins and Gibson, 1994) and their corresponding DNA sequences were therefore used. The synonymous nucleotide divergence (K_S) values were estimated using PAML 3.14 (Yang, 1997). Codon adaptation index (CAI) values (Sharp and Li, 1987), each of which indicates the strength of codon usage bias, were obtained from MIPS (Mewes et al., 2002) for *S. cerevisiae* genes.

Identification of gene conversion events. Numerous methods for gene conversion identification have been developed, but these methods are either not suitable or not powerful enough for the present analysis. For example, S. Sawyer's method uses measures of the distribution of identical synonymous sites between sequence pairs to identify candidate regions of conversion (Sawyer, 1989). This method assumes a neutral evolutionary process for synonymous sites and may therefore not be suitable for yeast genes in which codon usage bias affects synonymous substitution. More importantly, it does not use any outgroup for reference, so it is in general less powerful than phylogeny-based methods. Other methods, such as those of Jakobsen and coworkers (1996; 1997), rely on the examination of site-by-site phylogenies and the phylogeny for each site in a multiple alignment of paralogues and orthologues is tested for its support of conversion. Although these methods are similar to what proposed in this study, they suffer when there are multiple substitutions at individual sites (Drouin et al., 1999). This may again be a problem in my analysis as I am examining the ancient duplicates retained from the whole genome duplication in yeast in which multiple substitutions are common. Therefore, I have developed a related algorithm for conversion identification.

I used WGD orthologues in the 4 genomes, *S. cerevisiae*, *S. bayanus*, *S. mikatae* and *S. paradoxus*. At nucleotide position i, let D_i = the number of nucleotide differences between the two nucleotides in paralogous gene 1 and gene 2 in species 1 (the species under study), and B_{ji} = the number of nucleotide differences in gene j (j = 1, 2) between species 1 and its orthologue in species 2. Let $B_i = (B_{1i} + B_{2i}) / 2$. Sequences with gaps longer than 50% of the alignment were removed. For a gene under study, species with only one (or no) paralogue available are also removed. Gaps are all removed. For *S. cerevisiae*, *S. paradoxus*, or *S. mikatae*, B_i is calculated between the species under study and *S. bayanus*. For *S. bayanus*, B_i is calculated as the average of the differences between *S. bayanus* and the available three species.

Under the null hypothesis of no gene conversion, the distance (number of differences) between the two paralogues in a species should be larger than or equal to the distance between orthologues, i.e., $D_i - B_i \geq 0$, because the duplication event occurred prior to speciation. Dynamic programming is used to select the segment from site m to n that maximizes $\sum_{i=m}^{n}(B_i - D_i)$. This segment has N sites, $N = n - m + 1$. Let $D = \sum_{i=m}^{n} D_i$ and $B = \sum_{i=m}^{n} B_i$. If $N \geq 20$, the binomial probability to observe $D \leq \mathrm{B}$ for a segment of N sites is calculated using the orthologous distance B as the expected distance, i.e., $D = B$. This is a stringent criterion because the WGD event occurred earlier than speciation events. The estimated probability is

$$P(B,D,N) = \sum_{k=0}^{D} \frac{N!}{k!(N-k)!} \left(\frac{B}{N}\right)^{k} \left(1 - \frac{B}{N}\right)^{N-k} \qquad [1]$$

However, this segment always has its first and last sites supporting $B_i > D_i$, which may cause an overestimate of the significance. Therefore, I remove the first or the last site of the segment, and recalculate B and D as $\sum_{i=m+1}^{n} B_i$ and $\sum_{i=m+1}^{n} D_i$, or $\sum_{i=m}^{n-1} B_i$ and $\sum_{i=m}^{n-1} D_i$, and obtain binomial probabilities p_1 and p_2, respectively. The higher value of p_1 and p_2 is used.

The segments thus identified with the paralogous distance significantly smaller than the orthologous distance might potentially be derived from gene conversion. However, many possible segments of N sites can be selected from the entire gene sequence, so it is necessary to take this factor into consideration. Therefore, for each segment with a binomial probability $p < 0.01$ computed from [1], an empirical distribution of B for a segment of length N is constructed using 10,000 bootstrap

1896

samples from $\{B_1, B_2, ..., B_L\}$, where L = alignment length for the gene under consideration. Then, it is possible to determine the significance of D by counting the proportion of samples for which $D < B$. Segments with a binomial probability $p <$ 0.01 and with an empirical probability < 0.01 are considered candidate gene conversions.

Codon usage frequencies and tDNA genes. Relative frequencies of codon usage in orthologues of WGD genes were calculated for the genomes of *K. waltii*, *A. gossypii*, *S. cerevisiae*, *S. bayanus*, *S. mikatae* and *S. paradoxus*. Two sets of gene pairs were obtained. *S. cerevisiae* genes with CAI > 0.5 were classified into the highly expressed set and so were their orthologues in other species, whereas genes with CAI < 0.2 were classified into the lowly expressed set. The Chi-square test was used to examine if a codon is favored in highly expressed genes compared with lowly expressed genes. I obtained tDNA genes of *S. cerevisiae* from MIPS, and used the sequences and genomic BLAST in NCBI to identify orthologues in the other 5 genomes.

Results and Discussion

I first use the hypothetical trees in Fig. 2 to explain that a gene conversion event can distort the branch lengths and the topology of the phylogeny of duplicate genes and their orthologues among species. For example, the distance between paralogues α and a is expected to be longer than that between orthologues α and β (Fig. 2A) but the opposite is true in Fig. 2B because of a gene conversion event. To see how often such a situation has occurred in yeast duplicate genes, I studied ~300 WGD gene pairs in *S. cerevisiae* and their syntenic orthologues from three related species, *S. bayanus*, *S. mikatae* and *S. paradoxus* (Kellis et al., 2003). Because the WGD occurred prior to the radiation of these species, in the absence of gene conversion the synonymous distance (K_S) is expected to be larger between *S. cerevisiae* paralogues than between orthologues in different species. I find that this expectation indeed holds in most cases, with 93.4% of duplicate pairs in *S. cerevisiae* having a paralogous K_S greater than or equal to the K_S between orthologues (Fig. 3). This result indicates that only in a small proportion of these WGD duplicate genes has the tree topology been distorted by gene conversion because only when a point is below the line in Fig. 3 would a distortion in topology have occurred. Interestingly, most *S. cerevisiae* paralogous pairs with a small K_S also show a small K_S between orthologues and many have a high codon adaptation index value (CAI, a large circle in Fig. 3), a measure of codon usage bias (Sharp and Li, 1987). This analysis suggests that decelerated evolution of *S. cerevisiae* paralogues is at least in part due to biased codon usage, which serves as an evolutionary constraint (Pal, Papp and Hurst, 2001; Hirsh, Fraser and Wall, 2005).

Here are two examples illustrating different effects of gene conversion and codon usage bias on the evolution of duplicate genes. The first one is the gene pair YGR138C / YPR156C indicated by the red arrow in Fig. 3. The small circle indicates that these two genes have a weak codon usage bias (CAI 0.310 / 0.261), which is also reflected in the large K_S distance between orthologues. However, contrary to expectation, the K_S distance between the two *S. cerevisiae* paralogues is smaller than those between orthologues (Fig. 3), suggesting that gene conversion has occurred between the two *S. cerevisiae* paralogues. Indeed, the phylogenetic tree in Fig. 4A shows that the paralogues in each of the first three species are clustered, indicating gene conversions in these species after speciation. The second example is the gene pair YML063W / YLR441C indicated by the green arrow in Fig. 3. The large circle indicates a strong codon usage bias (CAI 0.769 / 0.696), which is reflected by small K_S values. The tree topology is as expected (Fig. 4B), so it provides no evidence of gene conversion. Despite this, the tree branches in Fig. 4B are in general much shorter than those in Fig. 4A. Clearly, codon usage bias can slow down sequence evolution in the entire tree, whereas gene conversion can shorten only sequence divergences between paralogues, but not those between syntenic orthologues.

To pursue the analysis further, I reconsidered the 66 duplicate gene pairs identified by Gao and Innan (2004) to have a small K_S between *S. cerevisiae* paralogues. I found that 57 of them were duplicated before the divergence between *S. cerevisiae* and *S. bayanus* and only one of these 57 pairs (YGL147C / YNL067W) is not from WGD (Dietrich et al., 2004; Kellis, Birren and Lander, 2004). In the 57 phylogenies for these 57 pairs, only 8 pairs showed a completely distorted tree topology (suggesting conversion in all lineages) like Fig. 4A, 23 pairs showed a partially distorted topology, while about half of them (26 pairs) showed no topology distortion (Table 1). I note that with the exception of two (YDL131W / YDL182W and YDR312W / YHR066W) all of the 57 pairs have a strong codon usage bias (CAI > 0.5). Therefore, in many of these gene pairs the small K_S values between *S. cerevisiae* paralogues (and between orthologues) might be largely due to strong codon usage bias constraint.

The above phylogenetic analysis, however, is not powerful enough for detecting all gene conversion events because conversion events involving only a small DNA region are unlikely to change the tree topology. For this purpose, a statistical method has been developed to detect gene conversion events and has been applied to ~300 WGD duplicate gene pairs in *S. cerevisiae*, *S. paradoxus, S. mikatae* and *S. bayanus*. The main purpose is to see whether gene conversion occurred primarily in high CAI genes. Indeed, Table 2 shows that about half of the genes with CAI ≥ 0.7 have undergone gene conversion events, while only 2% of the genes with CAI < 0.5 have conversions ($p < 10^{-8}$ for all species). Apparently, codon usage bias increases the rate of gene conversion by reducing the rate of sequence divergence. In the absence of strong codon usage bias, synonymous divergence between duplicate genes increases with time, and the chance of gene conversion is concomitantly reduced.

Another intriguing observation was that for most duplicate gene pairs that show a small protein distance, the divergence between the *K. waltii* - *A. gossypii* and

Saccharomyces sensu stricto species lineages is much longer (e.g. Fig. 5). This observation has been taken as evidence of gene conversion in the *Saccharomyces* species under study (Kellis, Birren and Lander, 2004). However, I notice that in these genes the protein distances are short not only between paralogues in the same species but also between orthologues in different WGD species, indicating that protein sequence conservation rather than gene conversion was the major cause of decelerated evolution. In the period immediately following the WGD event the duplicate proteins had apparently evolved rather rapidly (Fig. 5), likely due to relaxed functional constraints following WGD or the emergence of anaerobic growth, which has been found to be connected with *cis*-regulatory element evolution (Ihmels et al., 2005) . During this period gene conversion might have played a key role in maintaining the sequence similarity between the two paralogues. However, the rate of evolution had evidently become very slow prior to the radiation of the four *Saccharomyces* species (Fig. 5) and this largely explains why the sequence divergence is small between not only paralogues but also orthologues.

As for synonymous substitutions, previous studies showed that overlooking nucleotide composition differences (Tarrio, Rodriguez-Trelles and Ayala, 2001) or codon-usage patterns (Christianson, 2005) among sequences can mislead phylogenetic reconstruction. An examination of the codon usage patterns reveals that genes in *K. waltii* and *A. gossypii* have a stronger preference for G and C at third codon positions than genes in the four *Saccharomyces* species (Table 3). This may be one reason for the large K_S values in highly expressed genes between the *K. waltii* - *A. gossypii* lineage and the *Saccharomyces* lineage.

It was proposed that codon-usage bias is generally correlated with overall genome GC content, which is largely determined by mutational processes (Chen et al., 2004). Moreover, in most prokaryotic genomes, codons that are favored in highly expressed genes are well conserved (Rocha, 2004). In this study, the codon preferences for these yeast species also agree with their genome GC content, i.e., 44% and 52% for *K. waltii* and *A. gossypii*, and 38% ~ 40% for the four *Saccharomyces* species. However, although most favored codons are the same among these species (Table 3), I found a switch of the preferred codon of glutamine (Gln) between CAA and CAG and a switch of the preferred codon of glutamic acid (Glu) between GAA and GAG between *S. cerevisiae* and *A. gossypii*. As shown in Table 4, these switches might be due to changes in tDNA gene copy number. For instance, the numbers of tDNA-Glu genes for anticodons TTC and CTC are 14 and 2 in *S. cerevisiae* but 3 and 8 in *A. gossypii*, and this may explain why the GAA codon is preferred in *S. cerevisiae*, whereas GAG is preferred in *A. gossypii*. Such a difference in codon preference can increase the synonymous distance between species. The tDNA gene phylogeny suggests that the change of gene copy number can be derived from a point mutation at anticodon or from duplication/deletion of tDNA genes in the genome (Fig. 6).

Codon-usage bias is a compromise between compositional constraint (genomic GC content) and natural selection acting at the level of translation (Powell and Moriyama, 1997; Musto et al., 1999; Kliman, Irving and Santiago, 2003). If these two forces act in the same direction, for example, a preferred codon ending in G or C in a GC-rich genome, codon-usage bias could be extremely strong for highly expressed genes. On the other hand, the two forces may counteract each other; for example, a preferred codon ending in G or C in an AT-rich genome may only have its frequency slightly higher than 50% for highly expressed genes. This might explain why the high divergence between the *K. waltii* - *A. gossypii* and the *Saccharomyces sensu stricto* species mostly occurred in highly expressed genes.

Gao and Innan (2004) estimated the expected length of concerted evolution in *S. cerevisiae* as 25 million years based on the theory the same group previously proposed (Teshima and Innan, 2004) (*f* = 9/51; 51 gene pairs shows concerted evolution at the divergence time between *S. cerevisiae* and *S. bayanus*, while 9 gene pairs are still under concerted evolution at the divergence time between *S. cerevisiae* and *S. paradoxus*). I selected 18 gene pairs for which the paralogues and orthologues in *S. cerevisiae*, *S. paradoxus* and *S. bayanus* are all available and with CAI ≥ 0.7. Gene conversion was detected in 11 *S. cerevisiae* gene pairs. When I used *S. paradoxus* to calculate the orthologous distance instead, 6 gene pairs still have gene conversion events detectable. The expected length of concerted evolution for *S. cerevisiae* genes with CAI ≥ 0.7 thus estimated is 70 million years ($f = 6/11$; from *S. cerevisiae* - *S. bayanus* divergence to *S. cerevisiae* - *S. paradoxus* divergence). Note that this value may be underestimated because these genes are highly constrained and have evolved slowly. Informative sites indicating gene conversion may be too few to make the statistics significant. However, a similar estimate was obtained assuming the duration of concerted evolution started at the WGD event and the WGD occurred 100 million years ago ($f = 12/21$; from WGD to *S. cerevisiae* - *S. bayanus* divergence). Using the same method, I can estimate the expected lengths of concerted evolution for *S. cerevisiae* genes with CAI between 0.5 and 0.7, and CAI < 0.5 as 20 million years and 10 million years, respectively ($f = 4/31$ and $4/238$; from WGD to *S. cerevisiae* - *S. bayanus* divergence).

In summary, this study suggests that codon usage bias and protein functional conservation might have been more important than gene conversion for the decelerated evolution of WGD duplicate genes in yeasts. Note that gene conversion occurs only occasionally, whereas codon usage constraint and functional constraint of proteins are constant forces that slow down sequence evolution. Furthermore, the rate of gene conversion decreases as sequence divergence increases. For this reason gene conversion may not be an effective means for long-term maintenance of sequence similarity between duplicate genes in the absence of codon usage constraint or functional constraint. In contrast, both codon usage constraint and protein functional constraint can slow down sequence evolution in the absence of gene conversion. Of course, the three factors can have synergistic effects in maintaining high sequence similarity between paralogues.

Chapter 3

Protein complexity, gene duplicability and gene dispensability in the yeast genome

Introduction

Previous studies have suggested that most genes (~80%) of the budding yeast (*Saccharomyces cerevisiae*) are nonessential under laboratory conditions (Winzeler et al., 1999; Glaever et al., 2002; Steinmetz et al., 2002). Two mechanisms have been proposed for explaining this phenomenon. The first is the existence of duplicate genes (e.g., Nowak et al., 1997; Gu et al., 2003; Conant and Wagner, 2004; Kafri, Bar-Even and Pilpel, 2005); that is, the loss of function in one copy can be compensated by the other copy or copies. The second mechanism stems from alternative metabolic pathways, regulatory networks, and so on (Wagner, 2000). Papp, Pal and Hurst (2004) used an *in silico* metabolic flux model of the yeast metabolic network to address the dispensability issue. They estimated that up to 68% of "dispensable" genes might actually be important, but under conditions yet to be examined in the laboratory, 15-28% of dispensable genes are compensated by a duplicate, while only 4-17% are buffered by flux reorganization of the metabolic network.

In this study, I pursue the gene dispensability issue from the viewpoint of protein complexity. The number of domains in a polypeptide (He and Zhang, 2005) and the number of subunits in a protein complex (Yang, Lusk and Li, 2003) have been used to describe gene complexity and protein complexity, respectively. Here I define “domain complexity” as the number of domains in a polypeptide and “protein complexity” as the number of different subunit types in a protein complex. Although the number of protein interactions has been shown to correlate with protein deletion lethality (Jeong et al., 2001), there are four reasons to investigate protein complexity. First, the protein-protein interaction study was based on high-throughput data, which may have high false positive and false negative rates (von Mering et al., 2002). Second, subunits in a large complex without direct physical interactions to each other may not be detected by yeast two-hybrid analyses. Third, the number of protein interactions may reflect the number of functions or reactions that a polypeptide is involved, while a large complex may have only one specific function. Fourth, I am also interested in comparing monomers and homo-multimers, which is not feasible from protein interaction data.

Utilizing data on the fitness of heterozygotes for knockouts of essential genes in yeast, Papp, Pal and Hurst (2003b) found a greater decrease in heterozygote fitness if the gene is involved in a protein complex than if it is not, supporting the dosage balance hypothesis (Veitia, 2002; Veitia, 2003). However, homozygous gene deletion of a complex subunit may disrupt the protein function, which may be difficult to compensate by duplicated genes or alternative pathways if the function is cooperatively performed by multiple subunits. Further, Phadnis and Fry (2005) showed a negative correlation between homozygous effects and dominance of mutations (the ratio of heterozygous to homozygous effects) for all major categories of genes, which implies heterozygous and homozygous gene deletions may not have the same trend of fitness effect. It is therefore interesting to investigate whether the fitness effect of homozygous gene deletion increases with protein complexity and domain complexity.

The second purpose of this study is to re-examine the issue of the effect of protein complexity on gene duplicability. Although Papp, Pal and Hurst (2003b) and Yang, Lusk and Li (2003) have shown that protein complexity is an important determinant of gene duplicability, the relationship between protein complexity and gene duplicability is still not very clear. This is particularly so with respect to the question of whether homo-complexes tend to have a higher gene duplicability than hetero-complexes; although Yang, Lusk and Li (2003) considered this question, their data was not sufficiently large to draw a clear conclusion.

The third purpose is to investigate whether duplicate genes derived from the whole genome duplication event in the yeast (Wolfe and Shields, 1997; Dietrich et al., 2004; Kellis, Birren and Lander, 2004) and non-WGD duplicate genes show similar relationships among protein complexity, duplicability, and dispensability. He and Zhang (2006) found that a less severe fitness consequence of deleting a duplicate gene than deleting a singleton gene is at least in part due to the reason that duplicate genes are intrinsically less important than singleton genes. I wish to obtain a better estimate of the contribution of duplicate genes to gene dispensability in the yeast genome because the estimate by Gu et al. (2003) did not subdivide duplicate genes into WGD and non-WGD genes and did not consider the possibility of different gene duplicabilities for homo- and hetero-complexes. In this study, for simplicity, I include monomers, which consist of a single polypeptide, in the class of homo-complexes because, as will be seen later, monomers and homo-complexes show small differences in both gene dispensability and gene duplicability.

Materials and Methods

Identification of duplicate genes and singletons. An all-against-all FASTA (Pearson and Lipman, 1988) search was conducted for the whole set of *S. cerevisiae* protein sequences to obtain the list of singleton (single-copy) and duplicate genes as described in Gu et al. (2003). A whole genome duplication dataset was obtained from the genes listed in either Kellis, Birren and Lander (2004) or Dietrich et al. (2004). Although some gene pairs in the WGD dataset are quite diverged and may not satisfy the duplicate gene definition, they are still used in this analysis. The genes that did not satisfy the criteria for being singletons or duplicate genes were classified as twilight zone genes. The proportion of singleton families, *P*, was calculated as the number of singletons divided by the sum of the number of singletons and the number of duplicate gene families; 1 - *P* is used as a measure of gene duplicability.

Data on fitness effect of gene deletion. The growth rates of each yeast single-gene-deletion strain under various conditions were obtained from Steinmetz et al. (2002) (YDPM, http://www-deletion.stanford.edu/YDPM/YDPM_index.html) with five growth media: YPD, YPDGE, YPG, YPE and YPL; and from Glaever et al. (2002) with six extra conditions: YPGal, Minimal, Ph8, NaCl, Sorbitol and Nystatin. Each strain contains the precise homozygous diploid deletion of one ORF in the yeast genome. Genes annotated as essential in MIPS (Mewes et al., 2002) (http://mips.gsf.de/) or in YDPM were removed from this growth rate dataset because there is a possibility that an essential strain could be detected due to cross hybridization of a tag from another non-essential strain. The remaining genes were used and I calculated the fitness values (*f*) as the extent of survival and reproduction of the deletion strain relative to the pool of all strains grown and measured collectively (Gu et al., 2003). Essential genes annotated in both MIPS and YDPM were sequentially included, and their fitness values were assumed to be 0. All genes were subdivided into four groups according to their *f* values: (1) the deletion has a weak or no fitness effect in all conditions studied if $f_{\mathrm{min}} \geq 0.95$, where f_{min} is the smallest *f* value among all 11 growth conditions; (2) the deletion has a moderate effect if $0.8 \leq f_{\mathrm{min}} < 0.95$; (3) the deletion has a strong effect if $0 < f_{\mathrm{min}} < 0.8$; and (4) the deletion is lethal and *f* is set as 0. To avoid including pseudogenes and erroneously predicted genes, only ORFs with gene names in MIPS, YDPM or SGD (http://www.yeastgenome.org/) were kept for further analyses. Dispensable genes are defined as genes with a weak or no deletion fitness effect, i.e., $f_{\mathrm{min}} \geq 0.95$.

Similar to Papp, Pal and Hurst (2003b), I only used the growth rates of heterozygous strains obtained on YPD substrate from Steinmetz et al. (2002) to estimate their haplosufficiency. Only genes with two measurements from repeat experiments were retained, and average growth rates were calculated. Relative heterozygous fitness was calculated as the relative growth rate to the pool of all strains.

Collection of protein complexity data. Domain complexity data is obtained from Deng et al. (2002). Protein complexity is defined here as the number of different polypeptide types in a protein complex, not as the number of polypeptide subunits as defined in Yang, Lusk and Li (2003). The information of protein complexity was assembled from the complex or subunit descriptions in Swiss-Prot/TrEMBL (http://us.expasy.org/sprot/), MIPS, and SGD. A protein was regarded as a complex only when the descriptions of all components agreed with each other. A careful manual survey of published papers was made to verify these annotations. For example, in MIPS category 100, calcineurin B includes three entries; however, they do not form a hetero-trimer, but, instead, a regulatory subunit and two catalytic subunits form two kinds of hetero-dimers. I also used each gene name and several keywords to find literature on PubMed (http://www.ncbi.nlm.nih.gov/) and Google Scholar (http://scholar.google.com/) to increase the dataset. Homo-complexes (each composed of only one polypeptide type) were divided into monomers, homo-dimers, and homo-multimers, while hetero-complexes (each composed of more than one gene type) were classified according to the number of subunit polypeptide types. Polypeptides appearing in more than one complex were classified as multi-complex subunits, and the largest complex that a protein is involved was designated for the polypeptide. Cytoplasmic and mitochondrial ribosomal proteins were treated separately from other proteins.

Fitness values and expression levels among complex subunits. Since a protein complex could be a functional unit, its components should have similar deletion fitness effects. To test this hypothesis, hetero-complex genes, not including ribosomal and multi-complex proteins, were subdivided into dispensable (i.e., with a weak or no gene deletion effect) and indispensable, or lethal and nonlethal to examine if subunits of the same complex tend to have the same effect. I also wish to know, after excluding those dispensable and lethal genes, whether the fitness values of the subunits of a protein complex are still more similar than random gene pairs. For this purpose, I only keep genes with a strong or moderate deletion effect. The mean fitness difference between complex subunits is calculated and compared with the distribution of mean difference between randomly selected gene pairs. This random selection was repeated

10^7 times. For comparison, the fitness difference between duplicate genes (Gu et al., 2003) is also examined using the present method.

Similar procedures were applied to compare protein expression levels among complex subunits. TAP-tagged protein abundance data (Ghaemmaghami et al., 2003) were obtained from Yeast GFP Fusion Localization Database (http://yeastgfp.ucsf.edu/). Codon adaptation index (CAI) values, each of which indicates the strength of codon usage bias, were from MIPS.

Results

Protein complexity and gene dispensability. Previous studies used either only complex/non-complex dataset (Ge et al., 2001; Papp, Pal and Hurst, 2003b; Poyatos and Hurst, 2004; Teichmann and Veitia, 2004; Phadnis and Fry, 2005) or used proteins of no recorded interaction as monomers (Yang, Lusk and Li, 2003). I collected a more extended and reliable protein complex dataset, so that an analysis of different protein complexities is feasible. Table 5 shows the fitness effects of gene deletions for subunits of homo-complexes and subunits of hetero-complexes. The proportions of genes with weak (or lethal) fitness effect of deletion for monomers, homo-dimers, and homo-multimers are not significantly different from one another ($p > 0.1$). Thus, the number of subunits in a homo-complex protein, including monomers, does not seem to affect gene dispensability significantly. In contrast, subunits of a hetero-complex tend to have a lower dispensability than subunits of a homo-complex, especially when the number of subunit types becomes larger than 2. This trend is also observed for the proportion of genes with lethal deletion effect (Table 5).

Protein complexity and gene duplicability. I compared the proportions of singleton, duplicate, and twilight zone genes for homo- and hetero-complex subunits (Table 6). The proportion of duplicate genes (including WGD and non-WGD duplicates) is consistently higher than 40% for all homo-complex proteins; the differences between monomers and homo-dimers or homo-multimers are not significant ($p > 0.1$). In contrast, subunits of hetero-complexes have a much lower proportion of duplicate genes; the proportion decreases from 25% to 16% as the number of complex subunit types increases from 2 to ≥ 9, though this weak trend is not statistically significant ($p > 0.1$). In terms of the proportion of singleton families (P), the P value increases from 75% to 91% as the number of subunit types in a hetero-complex increases from 2 to ≥ 9. Note that the differences in gene duplicability between subunits of homo-complexes (monomers, homo-dimers, or homo-multimers) and subunits of hetero-complexes (subdivided according to their subunit types) are all significant ($p < 0.01$). Yang, Lusk and Li (2003) showed that complex proteins are less duplicable than monomers. This study further indicates that in terms of gene duplicability the major distinction is between homo-complexes and hetero-complexes. It is likely that only duplication of a gene for a subunit in a hetero-complex may cause dosage imbalance.

I then compared the proportion of haploinsufficient genes (heterozygous deletion fitness value obtained on YPD substrate < 0.99) among indispensable genes (homozygous deletion fitness value obtained on YPD substrate < 0.95) for homo-complex subunits. I found that homo-multimers are significantly more haploinsufficient (7/24) than homo-dimers + monomers (6/73, $p < 0.05$), which suggests that maintaining a sufficient protein dosage is more essential for homo-multimers (Kondrashov and Koonin, 2004). This result implies that many duplicates of genes for homo-multimer subunits were possibly retained due to protein dosage requirement. Compared to monomers, most duplicates of genes for homo-multimer subunits were from non-WGD events ($p < 0.05$, Table 6). This result supports the above observation because unlike WGD, which occurs rarely, non-WGD duplication can occur more frequently and duplicate genes can be retained if there is an increased requirement of protein dosage. I also found that the low duplicability of subunits of large hetero-complexes (composed of 9 or more subunit types) is largely due to their small number of WGD duplicates ($p < 0.05$, Table 6).

Ribosomal proteins. Ribosomes are the largest protein complexes in the yeast proteome. The WGD duplicates have been retained for most of the cytoplasmic ribosome proteins, but not for mitochondrial ones (Table 6). This phenomenon might be explained (1) by the dosage theory (Kondrashov and Koonin, 2004), i.e., after the WGD event a larger dosage would be required in the cytoplasm than in the mitochondria, and (2) by the dosage balance hypothesis (Veitia, 2002; Veitia, 2003), i.e., similar concentrations of subunits in the same protein complex are selectively preferred; otherwise, the imbalanced dosage of subunits may significantly reduce the final concentration of the protein complex. When a singleton cytoplasmic ribosomal subunit is deleted, its function cannot be compensated and the whole ribosome is not functional (10 out of 13 singleton genes have a lethal deletion effect), whereas deletion of a subunit with duplicates may only cause dosage deficiency and imbalance, but may not be lethal (91 out of 107 duplicate genes have strong or moderate deletion effects, but only 3 of them are lethal). Interestingly, most mitochondrial ribosome subunits are not essential (only with strong deletion effects), despite the fact that they are singleton genes.

Sequence similarity and gene dispensability. The dataset was subdivided into WGD and non-WGD sets, and also homo-complexes, hetero-complexes, and proteins without complex annotation (excluding ribosomal proteins). These genes were further subdivided according to the K_A of each gene to its most similar paralogue in the genome. Their cumulative distributions of fitness effect of gene deletion were compared (Fig. 7). Surprisingly, the correlation between gene deletion fitness effect and K_A is weak, especially for WGD genes (Kolmogorov-Smirnov test, $p > 0.1$), though this correlation was considered as strong evidence of functional compensation among duplicates (Gu et al., 2003). On the other hand, non-WGD hetero-complex subunits with $K_A < 0.4$ are more dispensable than subunits with $K_A > 0.4$ (Fig. 7B, $p < 0.001$); however, subunits with $K_A > 0.4$ are less dispensable than twilight zone and singleton genes ($p < 0.001$). For non-WGD homo-complex subunits, genes with $K_A > 0.4$ have similar fitness distributions (Fig. 7D), while genes with $K_A < 0.4$ are more dispensable ($p < 0.05$). Similar results are found for non-WGD genes without protein complex annotations (Fig. 7F). In this case, gene dispensability is increased when K_A is < 0.6 ($p < 0.05$).

Because protein complexity is an important determinant of gene duplicability (Papp, Pal and Hurst, 2003b; Yang, Lusk and Li, 2003), one may suspect that the higher dispensability of subunits of a homo-complex protein is mainly due to a higher proportion of duplicate genes for subunits of homo-complex proteins in the genome. Figure 7 indicates that when the distance of each gene to its most similar paralogue is controlled, homo-complex subunits still are much more dispensable than hetero-complex subunits, especially for non-WGD genes. This result suggests that the higher dispensability for homo-complex subunits is not due to their abundance of duplicate genes. I further analyzed gene dispensability with protein complexity for hetero-complex subunits. When I removed duplicate genes to regenerate the relationships between fitness effect of gene deletion and protein complexity (Fig. 8), the observation that gene dispensability decreases as the number of subunit types in a protein complex increases (Table 5) still holds, except that the dispensability of homo-complex subunits is slightly decreased. Therefore, I suggest that the higher dispensability of genes coding for subunits of small hetero-complexes (or homo-complexes) cannot be attributed to functional compensation of duplicated genes. On the other hand, protein complexity may serve as a better indicator of gene dispensability than does gene duplication, as will be discussed later.

Domain complexity and protein complexity. Since a protein domain may be the functional unit, one may expect multi-domain polypeptides to have lower dispensability. Indeed, Figure 9 shows that multi-domain polypeptides (with ≥2 domains) are significantly less dispensable than single-domain polypeptides. This difference is more significant when polypeptides for which no domain information is available are included in single-domain polypeptides. I found that 43% of hetero-complex subunits and 55% of homo-complex subunits are multi-domain polypeptides ($p < 0.001$). This result suggests that the proportion of multi-domain polypeptides cannot explain the low dispensability of hetero-complex subunits. On the other hand, one may suspect that the larger number of domains in the homo-complex avoids the need for a hetero-complex, implying that it might be the total number of domains of all the subunits of a protein complex that is a determinant of gene duplicability. To test this hypothesis, I only consider subunits of homo-complex or hetero-complex for which the summation of domain numbers in a complex is 2~4. Duplicate genes are excluded. The result indicates that hetero-complex subunits are still less dispensable than homo-complex ones (51 genes out of 192 genes with weak deletion fitness effects for hetero-complex subunits; 30 genes out of 70 genes for homo-complex subunits; $p < 0.05$). Among these genes, hetero-complex subunits should have fewer domains than homo-complex subunits. Therefore, I suggest protein complexity should be a more important determinant of gene dispensability than domain complexity.

Previous studies showed that domain complexity (He and Zhang, 2005) and protein complexity (Papp, Pal and Hurst, 2003b; Yang, Lusk and Li, 2003) both are important determinants of gene duplicability. Therefore, it is interesting to investigate whether these two factors correlate with each other since homo-complex subunits have a higher proportion of multi-domain polypeptides. Table 7 reveals that when the domain number in a polypeptide increases from one to > 2, its gene duplicability (1 - P) also increases from 35% to 64% for homo-complex subunits ($p < 10^{-2}$), and from 10% to 45% for hetero-complex ones ($p < 10^{-9}$). Moreover, homo-complex subunits are more duplicable than hetero-complex subunits when the number of domains is controlled ($p < 10^{-7}$ for single-domain polypeptides; $p < 10^{-3}$ for polypeptides with 2 domains; the difference is not significant for polypeptides with > 2 domains due to the small sample size). This result suggests domain complexity and protein complexity are largely independent with respect to gene duplicability.

Similar dispensabilities and expression levels for the subunits of a complex. Complex subunits were subdivided into dispensable (i.e. with a weak or no gene deletion effect) and indispensable, or into lethal and nonlethal genes (Table 8). The proportions of each combination of subunit pairs found in the same complex and those of randomly selected gene pairs were compared. The observed number of subunit pairs with the same fitness effect category was found to be much higher than expected. Therefore, complex subunits tend to display similar fitness effects of gene deletion. It has been reported that proteins in the same interaction module also have similar dispensability (Poyatos and Hurst, 2004). Because most genes are distributed at the two extreme ends of fitness effect of gene deletion, it is interesting to ask whether the above conclusion still holds if only genes with strong or moderate deletion effects are considered. The answer is yes, no matter which growth condition is considered (Table 9). Although duplicated genes may have a chance to compensate each other's function (Gu et al., 2003), I found that under most conditions duplicate gene pairs are not as similar to each other in gene deletion effect on fitness as the subunits of a complex. The reason might be that many duplicated genes have already functionally diverged, whereas the subunits of a complex usually play the same functional role.

Under the dosage balance hypothesis, complex subunits should have similar protein expression levels. Using the same method described above, i.e., comparing with randomly selected gene pairs, I find that similarity indeed exists for protein expression levels of complex subunits. The mean logarithm difference of TAP-tagged protein abundance values between hetero-complex subunits is significantly less than the mean difference between random gene pairs ($p << 10^{-7}$). For proteins that do not have abundance data (~one third of the genes), the codon adaptation index (CAI) was used to infer the expression level. I found that the mean difference in CAI values between subunits of a protein complex is only half of the mean difference between random gene pairs ($p << 10^{-7}$). This result is comparable to Ge et al.'s (2001) finding that genes encode interacting proteins tend to have similar expression profiles.

Discussion

Different trends of dispensability and duplicability for hetero-complexes. It was noted above that although subunits of hetero-complexes composed of 2 subunit types are less dispensable compared with subunits of homo-complexes (Fig. 8), the difference is not significant ($p > 0.1$). In contrast, the dispensability of hetero-complexes composed of 3~4 subunit types is significantly lower ($p < 0.01$). In other words, when the number of subunit types increases, gene dispensability decreases gradually, instead of a sharp difference between homo- and hetero-complexes. On the other hand, although gene duplicability (1 - *P*) correlates with protein complexity, the difference in gene duplicabilities between subunits of hetero-complexes composed of 2 and > 9 subunit types is not significant. Although the insignificance could be due to a small sample size, both duplicabilities are significantly less than the duplicability of homo-complex subunits. The duplicability dramatically decreases from 46% ~ 51% for homo-complex subunits to 9% ~ 25% for hetero-complexes subunits (Table 6). The reason might be only duplication of a gene for a hetero-complex subunit may cause serious dosage imbalance. This observation suggests that the dosage balance hypothesis can explain gene duplicability of complex proteins well (Papp, Pal and Hurst, 2003b; Yang, Lusk and Li, 2003), but cannot completely explain the difference in dispensabilities between hetero-complex subunits.

It is likely that knocking out a gene coding for a hetero-complex subunit would disrupt the function of the whole complex. This viewpoint is supported by the above result that subunits in the same complex tend to have similar deletion fitness effects. If the function of a protein complex is determined by most or all of its subunits, to compensate its lost function may need another complex either from duplicated genes or from alternative pathways. This effect may be more harmful than a complex concentration reduction derived from subunit duplication or heterozygous deletion (dosage imbalance). On the other hand, the formation of a large protein complex may take a long time in evolution. Therefore, losing the function of a large complex may be more severe than losing the function of a small one. This might explain why the dispensability of hetero-complexes decreases with protein complexity.

Functional compensation by duplicate genes. Non-WGD genes were subdivided into hetero-complexes, homo-complexes, and genes without complex annotations (Fig. 7). The contribution of duplicate genes to genetic robustness was estimated using Gu et al.'s method (2003). The result indicates that the dispensability of 1, 10, and 106 genes out of 104, 93, and 1086 dispensable genes might be attributed to gene duplication for these three categories, respectively. The proportion of the contribution of duplicate genes to genetic robustness is thus estimated to be 9% (117/1283) for non-WGD genes. He and Zhang (2006) found that less important genes tend to have a higher gene duplicability than important genes and suggested that this difference can partly account for a less severe fitness effect of deleting a duplicate gene than deleting a singleton gene. In the case studied here, the high dispensability of non-WGD genes with $K_A < 0.4$ may partly be due to recent duplications of less important genes, rather than all from functional compensation by duplicates. Therefore, the contribution of duplicate genes to dispensability may not be as high as previously estimated (23%, Gu et al., 2003).

It is worth noting that, except for ribosomal proteins, most of the ~400 WGD gene pairs that have been retained (Dietrich et al., 2004; Kellis, Birren and Lander, 2004) are dispensable (Fig. 7). For genes with the same protein complexity and with the same range of K_A to their most similar paralogues in the genome, WGD genes are consistently more dispensable than non-WGD duplicate genes. The difference is statistically significant for hetero-complexes, for homo-complexes with $K_A > 0.4$, and for genes without complex annotation and with $K_A > 0.6$ (KS test, $p < 0.05$). This result implies that in the majority of cases the dispensability of WGD genes may not be due to functional compensation from their duplicates, because functional compensation should have similar effects for WGD and non-WGD duplicates. An alternative explanation is that dispensable genes might have a higher chance to be retained than indispensable genes following the WGD event. This result echoes the previous observation that dispensable (less important) genes have a higher duplicability (He and Zhang, 2006). The reason Gu et al. (2003) overestimated the contribution of duplicate genes to dispensability is likely because their singleton dataset includes many hetero-complex subunits, while duplicate gene dataset includes many homo-complex subunits and WGD genes, which are dispensable intrinsically.

Another set of WGD genes are cytoplasmic ribosomal proteins, which, as mentioned earlier, tend to be indispensable. While deletion of a singleton cytoplasmic ribosomal subunit usually has lethal effect, deletion of a ribosomal subunit with duplicates may only cause dosage deficiency and imbalance (strong or moderate effect), but may not be lethal (Table 5). This fact suggests functional compensation exist for these WGD ribosomal proteins. However, although deletion of a ribosomal subunit with duplicates may not be lethal, such deletion is still evolutionarily deleterious. It is likely that their duplicates are retained mainly due to dosage requirement, but not due to functional compensation.

Chapter 4

Protein structure and evolutionary rate

E S A
1896

Introduction

The issue of what factors determine protein evolutionary rate has drawn much attention in recent years. Two major theories have been proposed to explain the variance of protein evolutionary rates. One is that functionally less important proteins evolve faster than more important ones (Ohta, 1973; Kimura and Ohta, 1974; Wilson, Carlson and White, 1977), which was supported by the weak but significant correlation between gene dispensability and protein evolutionary rate (Hirsh and Fraser, 2001; Yang, Gu and Li, 2003; Wall et al., 2005; Zhang and He, 2005). The other is that the rate is primarily determined by the proportion of residues involved in specific functions (functional density, Zuckerkandl, 1976). Fraser et al. (2002) suggested that proteins with more interactions evolve more slowly because they have higher functional density. Other studies reported that a protein evolves slowly if (1) the protein is highly expressed (Pal, Papp and Hurst, 2001; Rocha and Danchin, 2004; Wall et al., 2005), (2) the protein is involved in stable complexes (Teichmann, 2002), (3) the protein is involved in interaction hubs situated in single modules (Fraser, 2005), or (4) the protein occupies a central position in networks (Hahn and Kern, 2005). However, some of the above results were questioned because the levels of gene expression were not controlled (Bloom and Adami, 2003; Pal, Papp and Hurst, 2003). Recently, Drummond, Raval, and Wilke (2006) proposed that translational selection is the single dominant determinant, and provided an explanation why gene or protein expression level governs the evolutionary rate (Drummond et al., 2005). Most of these studies only considered characters of a whole protein, but did not look into differences in evolutionary constraints among residues. It is therefore interesting to investigate whether there are other dominant determinants.

Materials and Methods

Yeast genomic data. I obtained nonsynonymous rates (K_A) from Wall et al. (2005), protein interaction modules from Han et al. (2004), mRNA expression level from

Holstege et al. (1998), protein abundance from Ghaemmaghami et al. (2003), and codon adaptation index (CAI) values from Drummond, Raval and Wilke (2006); CAI indicates the strength of codon usage bias (Sharp and Li, 1987). Protein complexity and gene dispensability data are obtained as described in chapter 2. Genes without gene names were excluded. Principal component regression was performed using *R* with the package ‘pls’ (Ihaka and Gentleman, 1996). Protein abundance and mRNA expression level was log transformed.

Solvent accessibility predicted using homology model. PDB homologues for each open reading frame (ORF) were obtained from the *Saccharomyces* Genome Database (SGD, http://www.yeastgenome.org/). The solvent accessible surface areas (ACC) for each residue of the PDB homologue with the highest *p* value were obtained from DSSP (http://swift.cmbi.ru.nl/gv/dssp/) (Kabsch and Sander, 1983). Relative solvent accessibility (RelACC) was the ACC for each residue subdivided by the maximum value of ACC for the certain amino acid (represented using percentage), which is estimated from a Gly-X-Gly extended tripeptide conformation. Here I define residues with RelACC higher than 25% as exposed residues, and the others as buried. The proportion of exposed residues (P_{exposed}) for each PDB homologue was thus calculated.

Solvent accessibility predicted using supporting vector machine (SVM). SVM prediction was performed as described in Hsu (2005) using 480 proteins from Kim and Park (2004) as training dataset, and position-specific scoring matrices (PSSM), secondary structure profiles, and hydropathy indexes as feature factors. A 7-fold cross validation test yields 78% accuracy.

Results and Discussion

I first studied the relationship between the number of hetero-complex subunit types (excluding ribosomal proteins) and nonsynonymous substitution rates (K_A) for

Saccharomyces cerevisiae genes, because the total number of subunits present in a macromolecular complex or the fraction of a protein's residues directly involved in intermolecular contacts should be more relevant than a protein's total number of interaction partners (Bloom and Adami, 2004). The negative correlation between these two factors ($n = 479$, $p = 1.3\times 10^{-8}$, $R^2 = 6.6\%$) is consistent with Teichmann's (2002) finding that stable complex proteins evolve more slowly. However, the number of hetero-complex subunit types also correlates with mRNA expression level ($n = 760$, $p = 1.3 \times 10^{-15}$, $R^2 = 8.1\%$). Partial correlation analysis shows that the correlation between protein complexity and K_A is reduced when translational selections, especially mRNA expression, are controlled (Table 10). This result suggests that subunits of a large complex tend to be highly expressed, or, alternatively, protein complex data are possibly biased, i.e., highly expressed protein complexes have a higher chance to be identified and studied.

A recent study indicated that residues in the buried core and residues on the solvent-exposed surfaces are under different selection pressures and with different substitution patterns (Tseng and Liang, 2006). While the core residues of a protein are important in maintaining protein structure, only part of the surface residues are involved in ligand binding, enzymatic reactions or protein-protein interactions. If residues on the stable protein interaction surface are also treated as buried ones when the protein complex is considered as a whole, I speculate that exposed residues may evolve at a faster rate than buried residues. I obtained PDB homologue for yeast open reading frames (ORFs) from *Saccharomyces* Genome Database (SGD), and calculated the proportion of exposed residues (P_{exposed}) of the sequences. Partial correlation analysis shows that P_{exposed} indeed correlates with K_A (Table 10). However, the proportion of response's variance explained by the component, R^2, is not high. One possible reason is that some PDB structures only include one or partial subunits, but not necessarily the entire protein complex. In this case, residues on the stable interaction surface, which should be buried *in vivo*, are mistakenly treated as exposed ones in the PDB structure. Therefore, P_{exposed} may be overestimated for large complex subunits.

To overcome this problem, I used supporting vector machine (SVM) to predict P_{exposed} directly for each yeast ORF without the use of three-dimensional structures. The result indicates that although the accuracy of SVM prediction is only 78%, the predicted P_{exposed} does represent substantial information, and significantly correlate with evolutionary rate (Table 10). The variance of evolutionary rates explained by P_{exposed} is more than half of that explained by mRNA expression or protein abundance, and even slightly more than that explained by codon usage bias measured by the codon adaptation index (CAI, Sharp and Li 1987). It should be noticed that, although partial correlation analysis may be unreliable when data are noisy and the correlation is weak (Drummond, Raval and Wilke, 2006), the correlation found here is very strong. On the other hand, principal component regression analysis may also cause some misleading. The reason is that principal component transformation only guarantees that the transformed components are independent to each other and the prior components contain most information, but does not guarantee that the transformed vectors are biologically meaningful. When three translational selection related predictors are used to perform the analysis (Table 11), they compose the majority part of the first component and contribute 26.5% of the variance of K_A, while P_{exposed} contributes only about 5%. However, an opposite result is derived when two protein structure-related predictors and CAI are used (Table 12). To represent translational selection more relevantly, I used the first component in the principal component analysis for the three predictors (mRNA expression, protein abundance and CAI) to perform partial correlation with P_{exposed} and K_A (Table 10). The result described above is still hold. This suggests that P_{exposed} contributes at least 5%~10% to variation in protein evolutionary rate and should be the most important known determinant except for gene expression.

Fraser (2005) showed that party hubs (Han et al., 2004; proteins interact with most of their partners simultaneously) evolve slower than date hubs (proteins interact with different partners at different times). Since most party hubs are protein complexes, while date hubs are not (Han et al., 2004), I compared P_{exposed} between them. Not surprisingly, party hubs have a smaller P_{exposed} (49.7%) compared with date hubs (56.9%, t-test $p = 5.0 \times 10^{-5}$). It is reasonable to speculate P_{exposed} should also explain part of the difference of evolutionary rates between party and date hubs. Similarly, subunits of a large hetero-complex should have more protein interactions, and be less dispensable (as described in chapter 2). P_{exposed} may therefore underlie the correlations (Hirsh and Fraser, 2001; Fraser et al., 2002; Yang, Gu and Li, 2003; Wall et al., 2005; Zhang and He, 2005) between these two factors and evolutionary rate.

It is worth noting that, proteins with high P_{exposed} may evolve slowly or fast, whereas proteins with a low P_{exposed} always have low evolutionary rates (Fig. 10). This result suggests that protein three-dimensional structure only provides a general index, i.e., buried resides can not freely evolve. Some exposed residues may be functionally important and thus conserved, e.g., residues at active sites or ligand binding sites. While translational selection governs the rate of evolution for the whole protein (Drummond, Raval and Wilke, 2006), this study shows that “functional density” (Zuckerkandl, 1976) negatively correlates with protein evolutionary rate, i.e., a protein with more residues involved in specific functions may evolve more slowly. I expect that a better correlation will be found when the proportion of functionally important residues can be appropriately defined rather than a rough estimation using the proportion of exposed residues.

References

Akashi H. (2001). Gene expression and molecular evolution. *Curr. Opin. Genet. Dev.* **11:** 660-666.

Bloom J.D., Adami C. (2003). Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein-protein interactions data sets. *BMC Evol. Biol.* **3:** 21.

Bloom J.D., Adami C. (2004). Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: Response. *BMC Evol. Biol.* **4:** 14.

Chen S.L., Lee W., Hottes A.K., Shapiro L., McAdams H.H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. *Proc. Natl. Acad. Sci. USA* **101:** 3480-3485.

Christianson M.L. (2005). Codon usage patterns distort phylogenies from or of DNA sequences. *Am. J. Bot.* **92:** 1221-1233.

Coghlan A., Wolfe K.H. (2000). Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast* **16:** 1131-1145.

Conant G.C., Wagner A. (2004). Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. *Proc. R. Soc. Lond. B Biol. Sci.* **271:** 89-96.

Davis J.C., Petrov D.A. (2004). Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol.* **2:** 318-326.

Deng M., Mehta S., Sun F., Chen T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Res.* **12:** 1540-1548.

Deutschbauer A.M., Jaramillo D.F., Proctor M., Kumm J., Hillenmeyer M.E., Davis R.W., Nislow C., Giaever G. (2005). Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169:** 1915-1925.

Dietrich F.S., Voegeli S., Brachat S., Lerch A., Gates K., Steiner S., Mohr C., Pohlmann R., Luedi P., Choi S. et al. (2004). The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **304:** 304-307.

Drouin G., Prat F., Ell M., Clarke G.D.P. (1999). Detecting and characterizing gene conversions between multigene family members. *Mol. Biol. Evol.* **16:** 1369-1390.

Drummond D.A., Bloom J.D., Adami C., Wilke C.O., Arnold F.H. (2005). Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. USA* **102:** 14338-14343.

Drummond D.A., Raval A., Wilke C.O. (2006). A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23:** 327-337.

Fraser H.B. (2005). Modularity and evolutionary constraint on proteins. *Nat. Genet.* **37:** 351-352.

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

Gao L.-z., Innan H. (2004). Very low gene duplication rate in the yeast genome. *Science* **306:** 1367-1370.

Ge H., Liu Z., Church G.M., Vidal M. (2001). Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* **29:** 482-486.

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

Glaever G., Chu A.M., Ni L., Connelly C., Riles L., Veronneau S., Dow S., Lucau-Danila A., Anderson L., Andre B. et al. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418:** 387-391.

Goffeau A., Barrell B.G., Bussey H., Davis R.W., Dujon B., Feldmann H., Galibert F., Hoheisel J.D., Jacq C., Johnston M. et al. (1996). Life with 6000 genes. *Science* **274:** 546-567.

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

Hahn M.W., Kern A.D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* **22:** 803-806.

Han J.-D.J., Bertin N., Hao T., Goldberg D.S., Berriz G.F., Zhang L.V., Dupuy D., Walhout A.J.M., Cusick M.E., Roth F.P. et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430:** 88-93.

He X., Zhang J. (2005). Gene complexity and gene duplicability. *Curr. Biol.* **15:** 1016-1021.

He X., Zhang J. (2006). Higher duplicability of less important genes in yeast genomes. *Mol. Biol. Evol.* **23:** 144-151.

Hirsh A.E., Fraser H.B. (2001). Protein dispensability and rate of evolution. *Nature* **411:** 1046-1049.

Hirsh A.E., Fraser H.B., Wall D.P. (2005). Adjusting for selection on synonymous sites in estimates of evolutionary distance. *Mol. Biol. Evol.* **22:** 174-177.

Holstege F.C.P., Jennings E.G., Wyrick J.J., Lee T.I., Hengartner C.J., Green M.R., Golub T.R., Lander E.S., Young R.A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95:** 717-728.

Hsu W.-L. (2005). Prediction of protein relative solvent accessibility from amino acid sequence. Master thesis, National Chiao Tung University, Hsinchu.

Ihaka R., Gentleman R. (1996). R: a language for data analysis and graphics. *J. Comput. Graph. Stat.* **5:** 299-314.

Ihmels J., Bergmann S., Gerami-Nejad M., Yanai I., McClellan M., Berman J., Barkai N. (2005). Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* **309:** 938-940.

Jakobsen I.B., Easteal S. (1996). A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Bioinformatics* **12:** 291-295.

Jakobsen I.B., Wilson S.R., Easteal S. (1997). The partition matrix: exploring variable phylogenetic signals along nucleotide sequence alignments. *Mol. Biol. Evol.* **14:** 474-484.

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

Jordan I.K., Wolf Y.I., Koonin E.V. (2004). Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol. Biol.* **4:** 22.

Kabsch W., Sander C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577-2637.

Kafri R., Bar-Even A., Pilpel Y. (2005). Transcription control reprogramming in genetic backup circuits. *Nat. Genet.* **37:** 295-299.

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

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

Kim H., Park H. (2004). Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins* **54:** 557-562.

Kimura M., Ohta T. (1974). On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. USA* **71:** 2848-2852.

Kliman R.M., Irving N., Santiago M. (2003). Selection conflicts, gene expression, and codon usage trends in yeast. *J. Mol. Evol.* **57:** 98-109.

Kondrashov F.A., Koonin E.V. (2004). A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet.* **20:** 287-291.

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

Kwast K.E., Lai L.-C., Menda N., James D.T., III, Aref S., Burke P.V. (2002). Genomic analyses of anaerobically induced genes in *Saccharomyces cerevisiae*: functional roles of Rox1 and other factors in mediating the anoxic response. *J. Bacteriol.* **184:** 250-265.

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

Mewes H.W., Frishman D., Guldener U., Mannhaupt G., Mayer K., Mokrejs M., Morgenstern B., Munsterkotter M., Rudd S., Weil B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30:** 31-34.

Moller K., Olsson L., Piskur J. (2001). Ability for anaerobic growth is not sufficient for development of the petite phenotype in *Saccharomyces kluyveri*. *J. Bacteriol.* **183:** 2485-2489.

Musto H., Romero H., Zavala A., Jabbari K., Bernardi G. (1999). Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: Compositional constraints and translational selection. *J. Mol. Evol.* **49:** 27-35.

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

Ohta T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature* **246:** 96-98.

Pal C., Papp B., Hurst L.D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics* **158:** 927-931.

Pal C., Papp B., Hurst L.D. (2003). Rate of evolution and gene dispensability. *Nature* **421:** 496-497.

Papp B., Pal C., Hurst L.D. (2003a). Evolution of *cis*-regulatory elements in duplicated genes of yeast. *Trends Genet.* **19:** 417-422.

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

Papp B., Pal C., Hurst L.D. (2004). Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429:** 661-664.

Pearson W.R., Lipman D.J. (1988). Improved tools for biological sequence analysis. *Proc. Natl. Acad. Sci. USA* **85:** 2444-2448.

Petes T.D. (2001). Meiotic recombination hot spots and cold spots. *Nat. Rev. Genet.* **2:** 360-369.

Petes T.D., Hill C.W. (1988). Recombination between repeated genes in microorganisms. *Annu. Rev. Genet.* **22:** 147-168.

Phadnis N., Fry J.D. (2005). Widespread correlations between dominance and homozygous effects of mutations: Implications for theories of dominance. *Genetics* **171:** 385-392.

Powell J.R., Moriyama E.N. (1997). Evolution of codon usage bias in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **94:** 7784-7790.

Poyatos J.F., Hurst L.D. (2004). How biologically relevant are interaction-based modules in protein networks. *Genome Biol.* **5:** R93.

Pronk J.T., Steensma H.Y., van Dijken J.P. (1996). Pyruvate metabolism in *Saccharomyces cerevisiae*. *Yeast* **12:** 1607-1633.

Rocha E.P.C. (2004). Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Res.* **14:** 2279-2286.

Rocha E.P.C., Danchin A. (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol. Biol. Evol.* **21:** 108-116.

Sawyer S. (1989). Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6:** 526-538.

Sharp P.M., Li W.-H. (1987). The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15:** 1281-1295.

Steinmetz L.M., Scharfe C., Deutschbauer A.M., Mokranjac D., Herman Z.S., Jones T., Chu A.M., Giaever G., Prokisch H., Oefner P.J. et al. (2002). Systematic screen for human disease genes in yeast. *Nat. Genet.* **31:** 400-404.

Tanay A., Regev A., Shamir R. (2005). Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast. *Proc. Natl. Acad. Sci. USA* **102:** 7203-7208.

Tarrio R., Rodriguez-Trelles F., Ayala F.J. (2001). Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. *Mol. Biol. Evol.* **18:** 1464-1473.

Teichmann S.A. (2002). The constraints protein–protein interactions place on sequence divergence. *J. Mol. Biol.* **324:** 399-407.

Teichmann S.A., Veitia R.A. (2004). Genes encoding subunits of stable complexes are clustered on the yeast chromosomes: An interpretation from a dosage balance perspective. *Genetics* **167:** 2121-2125.

Teshima K.M., Innan H. (2004). The effect of gene conversion on the divergence between duplicated genes. *Genetics* **166:** 1553-1560.

Thompson J.D., Higgins D.G., Gibson T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673-4680.

Tseng Y.Y., Liang J. (2006). Regions and application in protein function inference: A Bayesian Monte Carlo approach. *Mol. Biol. Evol.* **23:** 421-436.

Veitia R.A. (2002). Exploring the etiology of haploinsufficiency. *Bioessays* **24:** 175-184.

Veitia R.A. (2003). Nonlinear effects in macromolecular assembly and dosage sensitivity. *J. Theor. Biol.* **220:** 19-25.

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

Wagner A. (2000). Robustness against mutations in genetic networks of yeast. *Nat. Genet.* **24:** 355-361.

Wagner A. (2002). Asymmetric functional divergence of duplicate genes in yeast. *Mol. Biol. Evol.* **19:** 1760-1768.

Wall D.P., Hirsh A.E., Fraser H.B., Kumm J., Giaever G., Eisen M.B., Feldman M.W. (2005). Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci. USA* **102:** 5483-5488.

Wilson A.C., Carlson S.S., White T.J. (1977). Biochemical evolution. *Annu. Rev. Biochem.* **46:** 573-639.

Winzeler E.A., Shoemaker D.D., Astromoff A., Liang H., Anderson K., Andre B., Bangham R., Davis R.W. (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285:** 901-906.

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

Wong S., Butler G., Wolfe K.H. (2002). Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc. Natl. Acad. Sci. USA* **99:** 9272-9277.

Yang J., Gu Z., Li W.-H. (2003). Rate of protein evolution versus fitness effect of gene deletion. *Mol. Biol. Evol.* **20:** 772-774.

Yang J., Lusk R., Li W.-H. (2003). Organismal complexity, protein complexity, and gene duplicability. *Proc. Natl. Acad. Sci. USA* **100:** 15661-15665.

Yang Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13:** 555-556.

Zhang J., He X. (2005). Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol. Biol. Evol.* **22:** 1147-1155.

Zuckerkandl E. (1976). Evolutionary processes and evolutionary noise at the molecular level. I. Functional density in proteins. *J. Mol. Evol.* **7:** 167-183.

Table 1. Gene pairs duplicated before the divergence between *S. cerevisiae* and *S. bayanus* that have a small K_S between *S. cerevisiae* paralogues as identified by Gao and Innan (2004).

Gene pairs	CAI values	Gene pairs	CAI values
With completely distorted tree topology		With no topology distortion	
YHL033C / YLL045C	0.842 / 0.849	YER074W / YIL069C	0.816 / 0.756
YGL135W / YPL220W	0.832 / 0.821	YJL190C / YLR367W	0.812 / 0.523
YBR031W / YDR012W	0.803 / 0.812	YMR230W / YOR293W	0.802 / 0.840
YBR009C / YNL030W	0.734 / 0.627	YDR450W / YML026C	0.775 / 0.733
YHR203C / YJR145C	0.709 / 0.695	YDR418W / YEL054C	0.766 / 0.605
YMR186W / YPL240C	0.581 / 0.518	YER056C-A / YIL052C	0.763 / 0.781
YDL131W / YDL182W	0.329 / 0.321	YDL061C / YLR388W	0.760 / 0.653
YDR312W / YHR066W	0.160 / 0.189	YDR447C / YML024W	0.757 / 0.810
With partially distorted tree topology		YNL302C / YOL121C	0.757 / 0.794
YBR181C / YPL090C	0.846 / 0.837	YOR234C / YPL143W	0.730 / 0.747
YEL034W / YJR047C	0.814 / 0.704	YGR118W / YPR132W	0.726 / 0.789
YBR189W / YPL081W	0.810 / 0.507	YER131W / YGL189C	0.711 / 0.781
YCR031C / YJL191W	0.805 / 0.590	YDR500C / YLR185W	0.711 / 0.700
YHR141C / YNL162W	0.795 / 0.769	YLR441C / YML063W	0.696 / 0.769
YLR029C / YMR121C	0.783 / 0.436	YJL136C / YKR057W	0.693 / 0.596
YFR031C-A / YIL018W	0.773 / 0.764	YBR191W / YPL079W	0.691 / 0.733
YGL147C / YNL067W	0.771 / 0.778	YJL177W / YKL180W	0.680 / 0.809
YDL083C / YMR143W	0.764 / 0.677	YGR034W / YLR344W	0.677 / 0.631
YDL136W / YDL191W	0.759 / 0.798	YGR214W / YLR048W	0.668 / 0.733
YGL031C / YGR148C	0.759 / 0.756	YMR242C / YOR312C	0.665 / 0.697
YBL072C / YER102W	0.747 / 0.718	YLR448W / YML073C	0.627 / 0.672
YDL075W / YLR406C	0.737 / 0.630	YMR194W / YPL249C-A	0.620 / 0.800
YBR048W / YDR025W	0.733 / 0.705	YIL133C / YNL069C	0.611 / 0.723
YGR085C / YPR102C	0.727 / 0.781	YNL096C / YOR096W	0.597 / 0.747
YGR027C / YLR333C	0.716 / 0.612	YJR094W-A / YPR043W	0.571 / 0.872
YBL027W / YBR084C-A	0.708 / 0.686	YLR264W / YOR167C	0.561 / 0.528
YNL301C / YOL120C	0.680 / 0.812		
YHL001W / YKL006W	0.680 / 0.684		
YDL082W / YMR142C	0.652 / 0.742		
YLR287C-A / YOR182C	0.642 / 0.748		
YBL087C / YER117W	0.624 / 0.648		
YBL002W / YDR224C	0.563 / 0.658		

Table 2. Number of gene pairs (with detected gene conversion events / total)

	CAI ≥ 0.7	0.7 > CAI ≥ 0.5	CAI < 0.5	*p* value
S. cerevisiae	12 / 21	4 / 31	4 / 238	$< 10^{-8}$
S. paradoxus	9 / 18	3 / 28	3 / 215	$< 10^{-8}$
S. mikatae	8 / 20	4 / 29	3 / 161	$< 10^{-8}$
S. bayanus	15 / 21	11 / 31	6 / 246	$< 10^{-8}$

Only detected conversion events longer than 20 bp were reported.

Table 3. Relative frequencies of codon usage for different anticodons in yeast species

		S. cerevisiae	*S. paradoxus*	*S. mikatae*	*S. bayanus*	*K. waltii*	*A. gossypii*
a.a.	Codon	Relative frequencies of codon usage (highly / lowly expressed genes)					
Ala	GCU	**0.69** / 0.32	**0.67** / 0.32	**0.68** / 0.33	**0.62** / 0.29	**0.45** / 0.31	**0.35** / 0.18
Ala	GCC	**0.24** / 0.20	**0.25** / 0.21	**0.24** / 0.20	**0.31** / 0.25	**0.44** / 0.22	**0.39** / 0.24
Ala	GCA	0.03 / 0.32	0.03 / 0.31	0.03 / 0.31	0.02 / 0.29	0.03 / 0.28	0.05 / 0.22
Ala	GCG	0.00 / 0.13	0.00 / 0.13	0.01 / 0.12	0.00 / 0.14	0.03 / 0.16	0.16 / 0.34
Arg	CGU	0.13 / 0.13	**0.15** / 0.12	0.14 / 0.13	**0.15** / 0.12	**0.18** / 0.15	**0.14** / 0.12
Arg	CGC	0.00 / 0.06	0.00 / 0.06	0.00 / 0.06	0.01 / 0.07	0.03 / 0.15	0.15 / 0.28
Arg	CGA	0.00 / 0.08	0.00 / 0.08	0.00 / 0.08	0.00 / 0.07	0.00 / 0.11	0.00 / 0.07
Arg	CGG	0.00 / 0.04	0.00 / 0.04	0.00 / 0.04	0.00 / 0.05	0.01 / 0.10	0.05 / 0.24
Arg	AGA	**0.85** / 0.46	**0.82** / 0.45	**0.84** / 0.47	**0.84** / 0.46	**0.77** / 0.30	**0.63** / 0.14
Arg	AGG	0.01 / 0.24	0.02 / 0.24	0.02 / 0.23	0.01 / 0.23	0.02 / 0.17	0.03 / 0.15
Asn	AAU	0.13 / 0.62	0.15 / 0.61	0.18 / 0.62	0.11 / 0.55	0.07 / 0.51	0.06 / 0.41
Asn	AAC	**0.87** / 0.38	**0.85** / 0.39	**0.82** / 0.38	**0.89** / 0.45	**0.93** / 0.49	**0.94** / 0.59
Asp	GAU	0.43 / 0.67	0.45 / 0.66	0.46 / 0.68	0.39 / 0.61	0.16 / 0.55	0.14 / 0.44
Asp	GAC	**0.57** / 0.33	**0.55** / 0.34	**0.54** / 0.32	**0.61** / 0.39	**0.84** / 0.45	**0.86** / 0.56
Cys	UGU	**0.89** / 0.61	**0.91** / 0.60	**0.92** / 0.63	**0.92** / 0.58	**0.61** / 0.48	**0.43** / 0.35
Cys	UGC	0.11 / 0.39	0.09 / 0.40	0.08 / 0.37	0.08 / 0.42	0.39 / 0.52	0.57 / 0.65
Gln	CAA	**0.98** / 0.67	**0.97** / 0.66	**0.97** / 0.66	**0.99** / 0.65	0.56 / 0.54	0.21 / 0.30
Gln	CAG	0.02 / 0.33	0.03 / 0.34	0.03 / 0.34	0.01 / 0.35	0.44 / 0.46	**0.79** / 0.70
Glu	GAA	**0.95** / 0.69	**0.95** / 0.68	**0.93** / 0.68	**0.96** / 0.67	0.31 / 0.54	0.13 / 0.38
Glu	GAG	0.05 / 0.31	0.05 / 0.32	0.07 / 0.32	0.04 / 0.33	**0.69** / 0.46	**0.87** / 0.62
Gly	GGU	**0.93** / 0.39	**0.92** / 0.39	**0.93** / 0.41	**0.91** / 0.38	**0.80** / 0.31	**0.52** / 0.20

Gly	GGC	0.05 / 0.21	0.05 / 0.22	0.04 / 0.20	0.07 / 0.24	0.15 / 0.28	0.40 / 0.40
Gly	GGA	0.01 / 0.26	0.02 / 0.25	0.02 / 0.25	0.01 / 0.21	0.04 / 0.25	0.02 / 0.16
Gly	GGG	0.00 / 0.14	0.01 / 0.15	0.01 / 0.14	0.01 / 0.17	0.01 / 0.16	0.06 / 0.25
His	CAU	0.28 / 0.66	0.30 / 0.65	0.33 / 0.67	0.27 / 0.62	0.09 / 0.54	0.08 / 0.44
His	CAC	**0.72** / 0.34	**0.70** / 0.35	**0.67** / 0.33	**0.73** / 0.38	**0.91** / 0.46	**0.92** / 0.56
Ile	AUU	**0.48** / 0.45	**0.47** / 0.44	**0.49** / 0.44	0.38 / 0.40	0.28 / 0.41	0.26 / 0.34
Ile	AUC	**0.51** / 0.24	**0.51** / 0.25	**0.49** / 0.24	**0.60** / 0.30	**0.71** / 0.34	**0.71** / 0.41
Ile	AUA	0.01 / 0.31	0.02 / 0.31	0.02 / 0.31	0.02 / 0.30	0.01 / 0.25	0.02 / 0.25
Leu	UUA	0.16 / 0.28	0.17 / 0.27	0.21 / 0.28	0.16 / 0.24	0.01 / 0.13	0.01 / 0.09
Leu	UUG	**0.75** / 0.26	**0.71** / 0.27	**0.68** / 0.27	**0.73** / 0.28	**0.67** / 0.21	**0.57** / 0.18
Leu	CUU	0.01 / 0.13	0.03 / 0.13	0.02 / 0.13	0.02 / 0.12	0.06 / 0.19	0.05 / 0.13
Leu	CUC	0.00 / 0.06	0.00 / 0.06	0.01 / 0.06	0.01 / 0.07	0.05 / 0.15	0.07/ 0.16
Leu	CUA	0.06 / 0.15	0.06 / 0.14	0.07 / 0.14	0.07 / 0.15	0.12 / 0.14	**0.15** / 0.13
Leu	CUG	0.01 / 0.12	0.02 / 0.12	0.02 / 0.12	0.01 / 0.15	0.08 / 0.18	0.15 / 0.32
Lys	AAA	0.17 / 0.60	0.19 / 0.59	0.21 / 0.60	0.15 / 0.56	0.09 / 0.50	0.04 / 0.33
Lys	AAG	**0.83** / 0.40	**0.81** / 0.41	**0.79** / 0.40	**0.85** / 0.44	**0.91** / 0.50	**0.96** / 0.67
Met	AUG	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
Phe	UUU	0.21 / 0.61	0.23 / 0.60	0.25 / 0.61	0.16 / 0.54	0.15 / 0.51	0.15 / 0.44
Phe	UUC	**0.79** / 0.39	**0.77** / 0.40	**0.75** / 0.39	**0.84** / 0.46	**0.85** / 0.49	**0.85** / 0.56
Pro	CCU	0.12 / 0.32	0.14 / 0.31	0.15 / 0.32	0.13 / 0.28	0.27 / 0.33	0.18 / 0.20
Pro	CCC	0.00 / 0.17	0.01 / 0.18	0.01 / 0.17	0.02 / 0.21	0.04 / 0.20	0.10 / 0.22
Pro	CCA	**0.88** / 0.38	**0.85** / 0.37	**0.84** / 0.38	**0.85** / 0.36	**0.67** / 0.32	**0.61** / 0.24
Pro	CCG	0.00 / 0.14	0.01 / 0.15	0.00 / 0.13	0.00 / 0.15	0.01 / 0.15	0.10 / 0.33
Ser	UCU	**0.53** / 0.24	**0.53** / 0.24	**0.57** / 0.24	**0.49** / 0.22	**0.42** / 0.22	**0.29** / 0.16

Ser	UCC	**0.36** / 0.14	**0.35** / 0.15	**0.32** / 0.14	**0.41** / 0.17	**0.41** / 0.15	**0.44** / 0.17
Ser	UCA	0.04 / 0.22	0.04 / 0.22	0.04 / 0.22	0.04 / 0.20	0.03 / 0.18	0.03 / 0.12
Ser	UCG	0.01 / 0.10	0.01 / 0.11	0.01 / 0.10	0.00 / 0.12	0.06 / 0.14	0.17 / 0.21
Ser	AGU	0.03 / 0.17	0.03 / 0.17	0.03 / 0.18	0.03 / 0.16	0.02 / 0.15	0.01 / 0.11
Ser	AGC	0.03 / 0.12	0.03 / 0.12	0.03 / 0.11	0.03 / 0.12	0.06 / 0.17	0.05 / 0.22
Thr	ACU	**0.52** / 0.32	**0.51** / 0.32	**0.53** / 0.33	**0.42** / 0.29	**0.34** / 0.30	0.22 / 0.20
Thr	ACC	**0.44** / 0.19	**0.44** / 0.20	**0.42** / 0.19	**0.55** / 0.23	**0.59** / 0.23	**0.59** / 0.25
Thr	ACA	0.03 / 0.33	0.04 / 0.32	0.04 / 0.33	0.03 / 0.30	0.04 / 0.28	0.04 / 0.24
Thr	ACG	0.00 / 0.16	0.01 / 0.16	0.01 / 0.15	0.00 / 0.17	0.03 / 0.19	0.15 / 0.31
Trp	UGG	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
Tyr	UAU	0.15 / 0.59	0.16 / 0.59	0.18 / 0.60	0.09 / 0.54	0.07 / 0.45	0.07 / 0.37
Tyr	UAC	**0.85** / 0.41	**0.84** / 0.41	**0.82** / 0.40	**0.91** / 0.46	**0.93** / 0.55	**0.93** / 0.63
Val	GUU	**0.56** / 0.37	**0.55** / 0.36	**0.56** / 0.36	**0.45** / 0.32	0.33 / 0.31	**0.23** / 0.21
Val	GUC	**0.41** / 0.18	**0.41** / 0.19	**0.40** / 0.18	**0.51** / 0.22	**0.50** / 0.22	**0.46** / 0.23
Val	GUA	0.01 / 0.24	0.01 / 0.24	0.02 / 0.25	0.01 / 0.21	0.01 / 0.20	0.01 / 0.14
Val	GUG	0.03 / 0.21	0.03 / 0.22	0.02 / 0.22	0.02 / 0.25	0.15 / 0.27	0.29 / 0.41

a. a.: amino acid. Favored codons in highly expressed genes relative to lowly expressed genes are shown in bold ($p < 0.05$).

Table 4. Numbers of tDNA genes for different anticodons in yeast species

a. a. - Anticodon	*S. cerevisiae*	*S. paradoxus*	*S. mikatae*	*S. bayanus*	*K. waltii*	*A. gossypii*
Asp - ATC	0	0	0	0	0	0
Asp - GTC	16	19	16	16	10	10
Cys - ACA	0	0	0	0	0	0
Cys - GCA	4	3	4	4	4	3
Gln - TTG	9	9	9	9	6	4
Gln - CTG	1	1	1	1	4	4
Glu - TTC	14	15	14	14	7	3
Glu - CTC	2	2	2	2	9	8
His - ATG	0	0	0	0	0	0
His - GTG	7	7	7	7	4	5

Table 5. Relationships between protein complexity and fitness effect of gene deletion

Protein structure [a]	Total # of genes	Proportions (numbers) of genes with lethal, strong, moderate and weak deletion effect on fitness			
		Lethal	Strong	Moderate	Weak
Monomer	109	19% (21)	22% (24)	18% (20)	41% (44)
Homo-dimer	127	20% (26)	17% (22)	20% (26)	42% (53)
Homo-multimer	83	16% (13)	24% (20)	17% (14)	43% (36)
Hetero-complex (2)	166	25% (41)	22% (37)	19% (32)	34% (56)
Hetero-complex (3~4)	219	37% (81)	24% (52)	18% (40)	21% (46)
Hetero-complex (5~8)	190	59% (112)	15% (29)	12% (23)	14% (26)
Hetero-complex (9~)	213	58% (124)	28% (60)	8% (18)	5% (11)
Cytoplasmic ribosome	126	13% (17)	29% (37)	45% (57)	12% (15)
Mitochondrial ribosome	62	3% (2)	84% (52)	5% (3)	8% (5)

[a] The number in the parentheses for hetero-complexes indicates the number of subunit types.

Table 6. Relationships between protein complexity and gene duplicability

Protein structure [a]	Total # of genes	Proportions (numbers) of duplicate, twilight zone, and singleton genes				Proportion of singleton families, *P*
		WGD Duplicate	Non-WGD Duplicate	Twilight	Singleton	
Monomer	109	26% (28)	14% (15)	30% (33)	30% (33)	54%
Homo-dimer	127	20% (26)	22% (28)	23% (29)	35% (44)	54%
Homo-multimer	83	18% (15)	30% (25)	19% (16)	33% (27)	49%
Hetero-complex (2)	166	12% (19)	13% (22)	39% (65)	36% (60)	75%
Hetero-complex (3~4)	219	10% (23)	7% (16)	31% (67)	52% (113)	85%
Hetero-complex (5~8)	190	10% (19)	11% (20)	36% (69)	43% (82)	80%
Hetero-complex (9~)	213	3% (7)	13% (28)	30% (63)	54% (115)	91%
Cytoplasmic ribosome	126	80% (101)	5% (6)	5% (6)	10% (13)	21%
Mitochondrial ribosome	62	0% (0)	0% (0)	16% (10)	84% (52)	100%

[a] The number in the parentheses for hetero-complexes indicates the number of subunit types.

Table 7. Relationships between protein complexity, domain complexity and gene duplicability

Protein structure	Proportions (numbers) of duplicate, twilight zone, and singleton genes			Proportion of singleton families, *P*
	Duplicate	Twilight	Singleton	
Homo-complex subunits with one domain	31% (44)	25% (36)	44% (63)	65%
Homo-complex subunits with 2 domains	49% (48)	23% (23)	28% (27)	47%
Homo-complex subunits with > 2 domains	47% (36)	34% (26)	19% (15)	36%
Hetero-complex subunits with one domain	12% (54)	27% (120)	61% (273)	90%
Hetero-complex subunits with 2 domains	19% (33)	44% (77)	37% (64)	79%
Hetero-complex subunits with > 2 domains	28% (47)	51% (86)	21% (35)	55%

One-domain polypeptides include polypeptides for which no domain information is available.

Table 8. Expected and observed proportions (numbers) of pairwise fitness combinations for hetero-complex subunits

	Expected proportion (all possible pairwise combinations)	Observed proportion (combinations found in the same complex)	*p* value (Fisher's exact test)
Dispensable vs. Dispensable	3.2% (9045)	4.4% (71)	8.7×10^{-3}
Dispensable vs. Indispensable	29.5% (83295)	12.2% (197)	5.1×10^{-62}
Indispensable vs. Indispensable	67.3% (190036)	83.4% (1350)	7.6×10^{-49}
Lethal vs. Lethal	19.8% (55945)	45.4% (735)	2.2×10^{-120}
Lethal vs. Nonlethal	49.5% (139695)	18.6% (301)	1.1×10^{-147}
Nonlethal vs. Nonlethal	30.7% (86736)	36.0% (582)	6.2×10^{-6}

The numbers of dispensable, indispensable, lethal, and nonlethal genes are 135, 617, 335, and 417, respectively.

Table 9. The mean value of fitness difference between randomly selected gene pairs and between complex subunits or duplicate gene pairs

	Complex subunits / random pairs		Duplicate genes / random pairs	
Conditions	Mean of fitness difference	*p* value	Mean of fitness difference	*p* value
YPD	0.116 / 0.142	3.6×10^{-6}	0.120 / 0.147	3.7×10^{-4}
YPDGE	0.105 / 0.141	$< 1 \times 10^{-7}$	0.109 / 0.141	3.3×10^{-6}
YPG	0.136 / 0.232	$< 1 \times 10^{-7}$	0.156 / 0.230	$< 1 \times 10^{-7}$
YPE	0.140 / 0.245	$< 1 \times 10^{-7}$	0.169 / 0.247	$< 1 \times 10^{-7}$
YPL	0.131 / 0.209	$< 1 \times 10^{-7}$	0.130 / 0.206	$< 1 \times 10^{-7}$
YPgal	0.105 / 0.133	3.0×10^{-5}	0.099 / 0.129	3.0×10^{-4}
Minimal	0.138 / 0.158	7.3×10^{-3}	0.123 / 0.173	6.6×10^{-6}
Ph8	0.132 / 0.176	$< 1 \times 10^{-7}$	0.142 / 0.161	2.6×10^{-2}
NaCl	0.111 / 0.135	5.4×10^{-4}	0.112 / 0.140	1.1×10^{-3}
Sorbitol	0.100 / 0.122	7.5×10^{-5}	0.124 / 0.127	3.7×10^{-1}
Nystatin	0.115 / 0.145	4.2×10^{-5}	0.117 / 0.139	8.9×10^{-3}

Only genes with strong or moderate deletion effect on fitness are included.

Table 10. Partial correlation analyses between six factors and K_A

The factor correlated with K_A	n	p	R^2	The factor controlled
Number of hetero-complex subunit types	465	3.1×10^{-2}	1.0%	mRNA expression
Number of hetero-complex subunit types	372	8.5×10^{-5}	4.0%	Protein abundance
Number of hetero-complex subunit types	479	3.1×10^{-9}	6.9%	CAI
P_{exposed} (PDB homologues)	984	1.3×10^{-5}	1.9%	mRNA expression
P_{exposed} (PDB homologues)	739	3.4×10^{-5}	2.3%	Protein abundance
P_{exposed} (PDB homologues)	1026	5.2×10^{-7}	2.4%	CAI
P_{exposed} (predicted using SVM)	2153	$< 10^{-15}$	10.9%	mRNA expression
P_{exposed} (predicted using SVM)	1602	$< 10^{-15}$	12.5%	Protein abundance
P_{exposed} (predicted using SVM)	2267	$< 10^{-15}$	13.2%	CAI
mRNA expression	2153	$< 10^{-15}$	17.4%	P_{exposed} (SVM)
Protein abundance	1602	$< 10^{-15}$	13.9%	P_{exposed} (SVM)
CAI	2267	$< 10^{-15}$	11.8%	P_{exposed} (SVM)
P_{exposed} (predicted using SVM)	1568	$< 10^{-15}$	11.3%	Translational selection[a]
Translational selection[a]	1568	$< 10^{-15}$	18.3%	P_{exposed} (SVM)

[a] The first component in the principal component analysis for mRNA expression, protein abundance, and CAI values.

Table 11. Results of principal component regression analysis on four predictors and K_A for 1568 genes

	Principal components			
	1	2	3	4
Percent variance in predictors	58.3	23.0	10.0	8.7
Percent variance explained (R^2) in K_A	26.5***	4.9***	0.1	0.1
Percent contributions				
mRNA expression	**32.6**	0.4	6.3	**60.7**
Protein abundance	**30.8**	1.4	**65.4**	2.4
CAI	**30.4**	6.1	**28.0**	**35.5**
P_{exposed} (predicted using SVM)	6.2	**92.1**	0.3	1.4

* $p < 0.05$; ** $p < 10^{-6}$; *** $p < 10^{-9}$. Bold indicates that the indicated predictor contributes at least 20% to the indicated component.

Table 12. Results of principal component regression analysis on five predictors and K_A for 1026 genes

	Principal components				
	1	2	3	4	5
Percent variance in predictors	30.1	23.0	21.2	16.1	9.6
Percent variance explained (R^2) in K_A	19.3***	12.1***	0.6*	0.0	1.7*
Percent contributions					
P_{exposed} (predicted using SVM)	**47.1**	0.6	4.4	2.2	**45.7**
P_{exposed} (PDB homologues)	**40.4**	11.8	3.7	1.5	**42.6**
Dispensability	0.3	**28.3**	**37.9**	**33.0**	0.4
Protein length	0.9	**27.1**	**51.2**	9.9	10.9
CAI	11.2	**32.3**	2.8	**53.4**	0.4

* $p < 0.05$; ** $p < 10^{-6}$; *** $p < 10^{-9}$. Bold indicates that the indicated predictor contributes at least 20% to the indicated component.

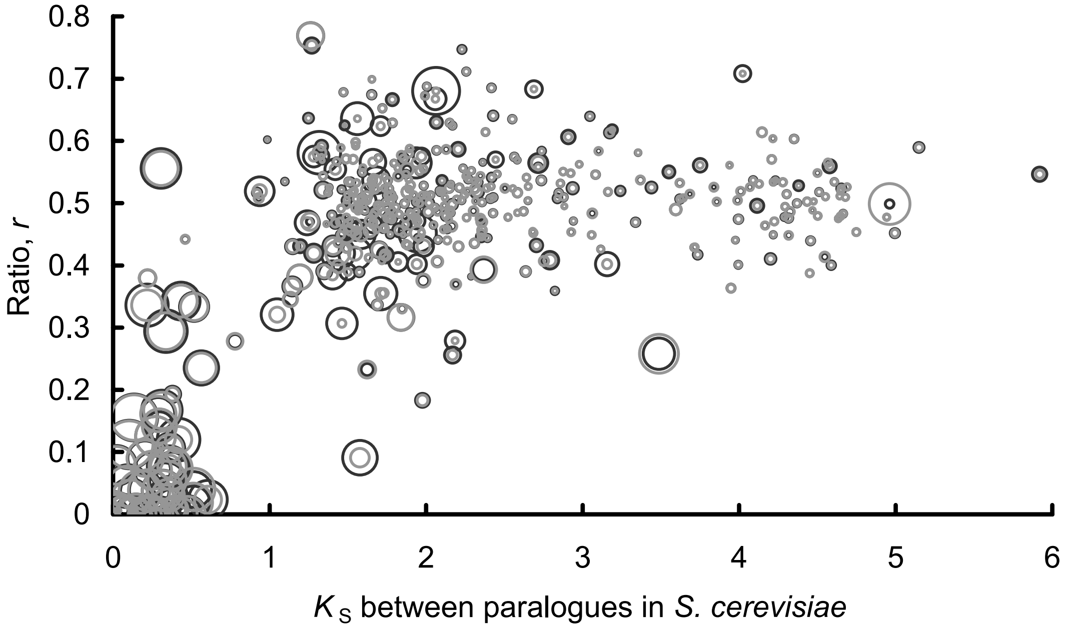

Figure 1. Let *div* $_{(a,\,b)}$ represent protein divergence between sequences *a* and *b*; *SC1*, *SC2*, and *K* represent *S. cerevisiae* copy1, copy2 and their orthologue in *K. waltii* following Kellis, Birren and Lander (2004). The ratio *r* is defined as *div* $_{(SC1,\,SC2)}$ / (*div* $_{(K,\,SC1)}$ + *div* $_{(K,\,SC2)}$). Gene pairs showing decelerated evolution (small ratio *r*) correspond to gene pairs with small K_S values. Copy 1 and 2 are represented as dark gray and light gray circles. The circle diameter indicates its CAI value, which represents its codon usage bias (Sharp and Li, 1987). For gene pairs with K_S less than 0.75, 87% genes (92/106) show CAI values higher than 0.5.

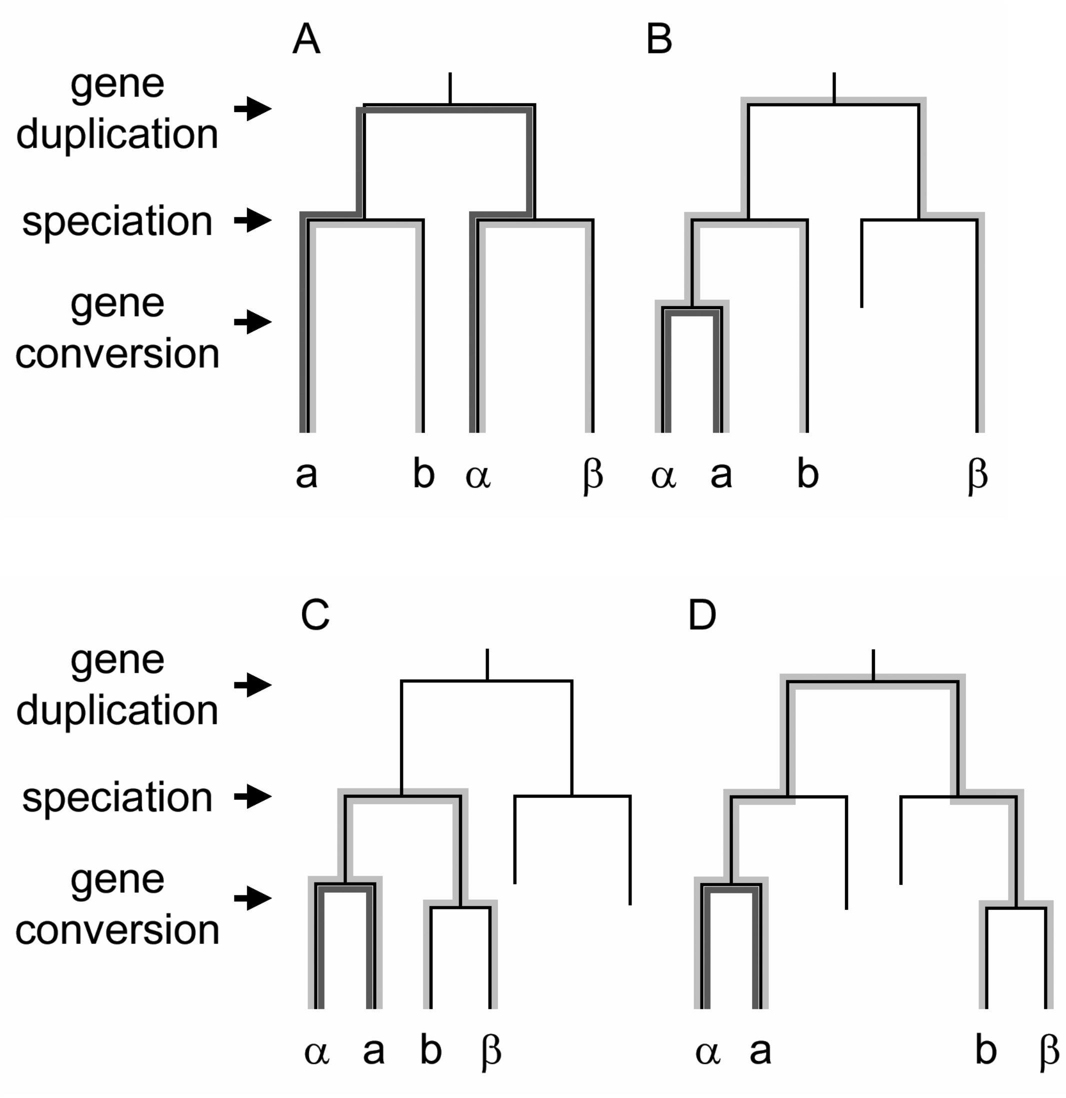

Figure 2. (A) Genes a and α (b and β) are paralogues derived from a gene duplication, and a and b (α and β) are orthologues derived from a speciation event. Dark gray and light gray lines indicate the distances between paralogues (a and α) and orthologues, respectively. (B) α was converted by a. (C) α was converted by a, and β was converted by b. (D) α was converted by a, and b was converted by β. Note that gene conversion can reduce the distance between paralogues but tends to increase the distance between syntenic orthologues.

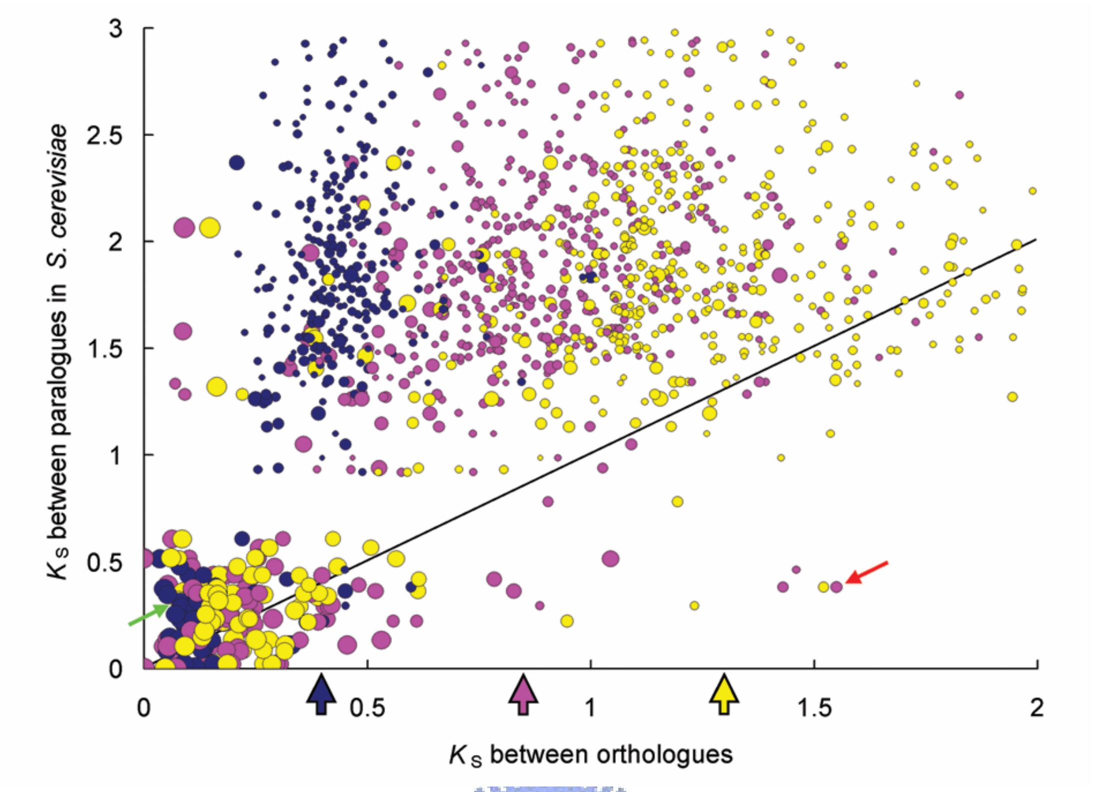

Figure 3. K_S between paralogues in *S. cerevisiae* (distance between a and α in Fig. 2) vs. K_S between orthologues (distances between a and b, or α and β in Fig. 2). Dark blue: *S. cerevisiae* vs. *S. paradoxus*; pink: *S. cerevisiae* vs. *S. mikatae*; yellow: *S. cerevisiae* vs. *S. bayanus*; open arrows indicate the average distances for these species pairs under weak codon usage bias (K_S = 0.4, 0.8, and 1.3). Circle sizes indicate the CAI values of the genes in *S. cerevisiae*. The slope line indicates that the distance between paralogues is equal to that between orthologues. The red and green solid arrows indicate gene pairs YGR138C / YPR156C between *S. cerevisiae* and *S. mikatae*, and YML063W / YLR441C between *S. cerevisiae* and *S. paradoxus*, respectively. Genes with incomplete sequences, paralogous pairs with K_S > 3 and orthologous pairs with K_S > 2 are not included in this figure.

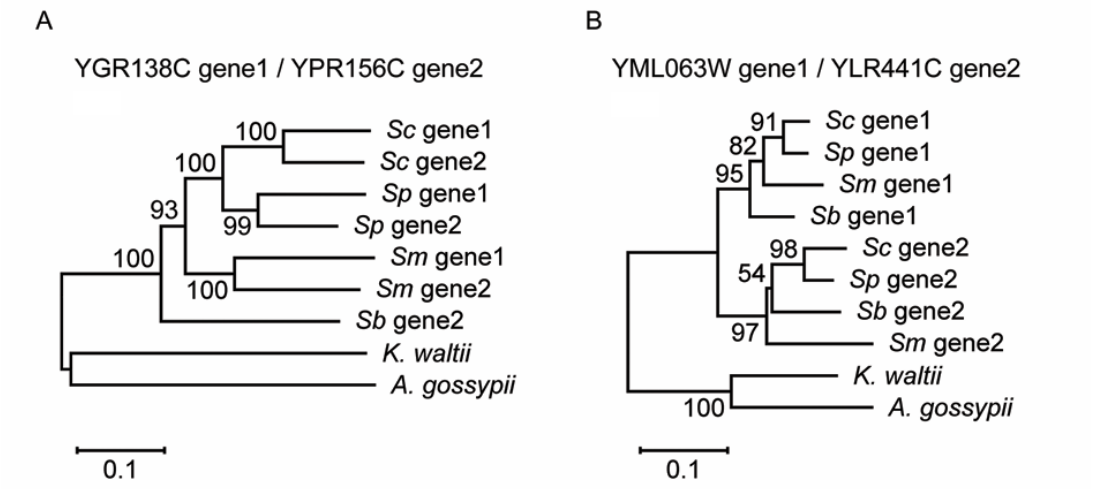

Figure 4. (A) Neighbor-joining tree (K_S distances) of the WGD gene pair YGR138C / YPR156C in *S. cerevisiae* (*Sc*), their orthologues in *S. paradoxus* (*Sp*), *S. mikatae* (*Sm*), and *S. bayanus* (*Sb*), and the outgroups in *K. waltii* (Kellis, Birren and Lander, 2004) and *A. gossypii* (Dietrich et al., 2004) (the red arrow in Fig. 3; CAI = 0.310 / 0.261). The orthologue of YGR138C in *S. bayanus* was not completely sequenced and not included in this figure. (B) YML063W / YLR441C (the green arrow in Fig. 3; CAI = 0.769 / 0.696). The numbers at branch nodes are bootstrap values.

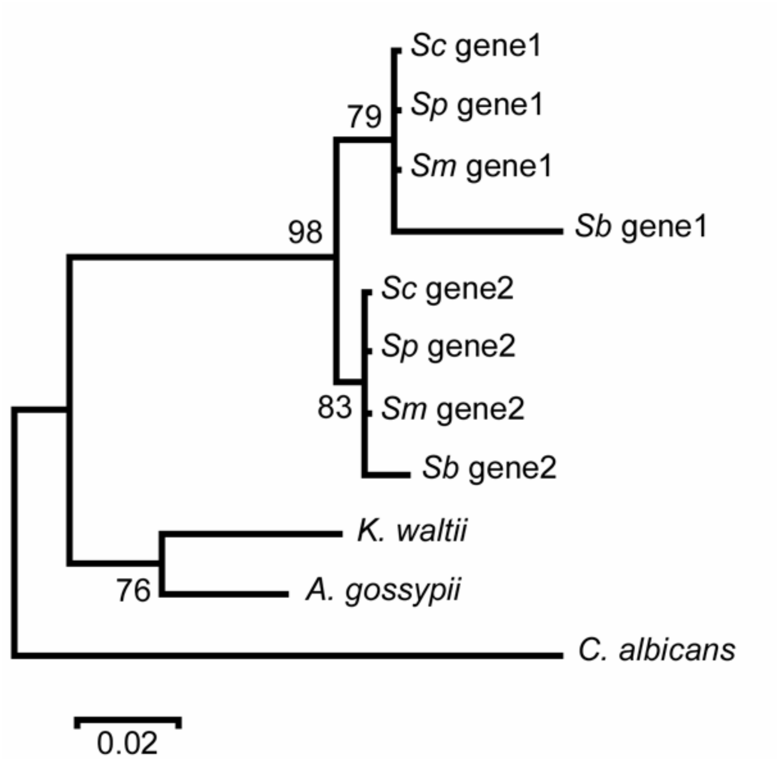

Figure 5. Neighbor-joining tree of the whole genome duplicated ORFs of *S. cerevisiae* (*Sc*), and their orthologues in *S. paradoxus* (*Sp*), *S. mikatae* (*Sm*), and *S. bayanus* (*Sb*), and outgroups *K. waltii*, *A. gossypii* and *C. albicans* for YER131W (gene 1) / YGL189C (gene 2) (cytoplasmic small ribosomal subunits; CAI = 0.711 / 0.781). The tree was constructed using protein Poisson distances. The numbers at branch nodes are bootstrap values.

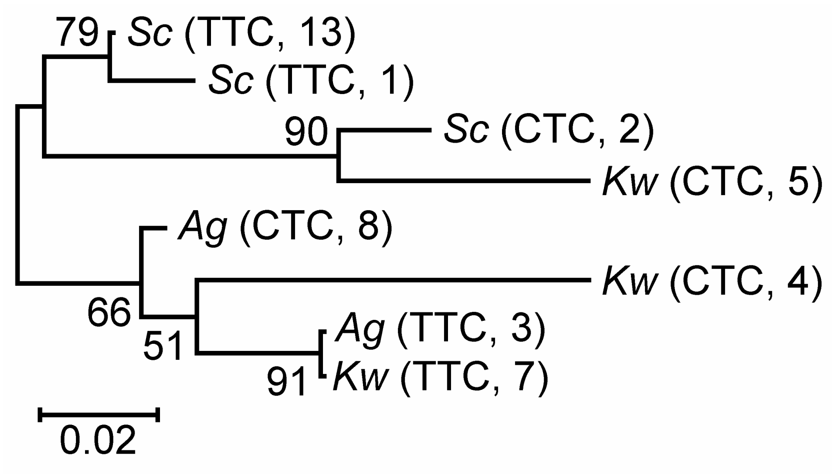

Figure 6. The neighbor-joining tree of tDNA-Glu genes among three yeast species (*S. cerevisiae*, *Sc*; *K. waltii*, *Kw*; *A. gossypii*, *Ag*). The triplet and number in the parenthesis indicate, respectively, the tDNA anticodon and the gene copy number in the corresponding genome. The numbers at branch nodes are bootstrap values. This phylogeny suggests the switch between anticodons occurred at least twice in the evolution of the tDNA-Glu gene in these yeast species.

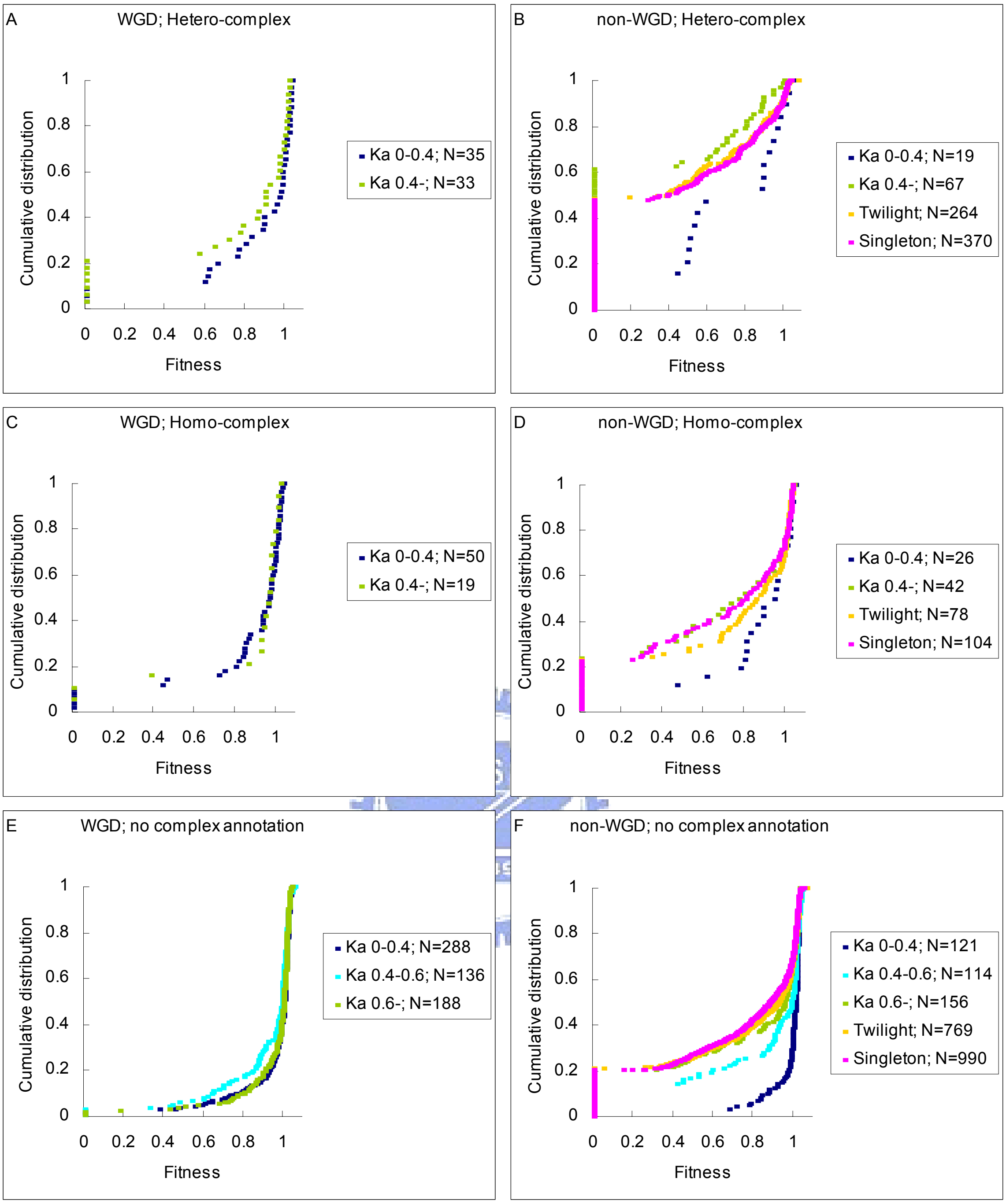

Figure 7. Cumulative fitness distribution of gene deletions of WGD (A, C, E) and non-WGD (B, D, F) duplicate genes for hetero-complexes (A, B), homo-complexes (C, D), and proteins without complex annotation (E, F). Duplicate genes are further subdivided according to the K_A of each gene to its most similar paralogue in the genome. N indicates gene number.

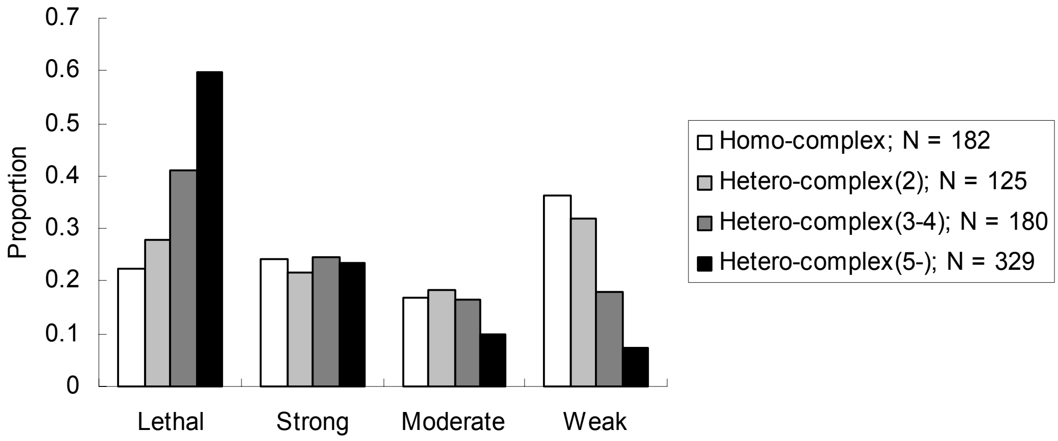

Figure 8. Fitness distribution of gene deletions after exclusion of duplicate genes. Homo-complexes include monomers, homo-dimers, and homo-multimers. The number in the parentheses for hetero-complexes indicates the number of subunit types. N indicates gene number.

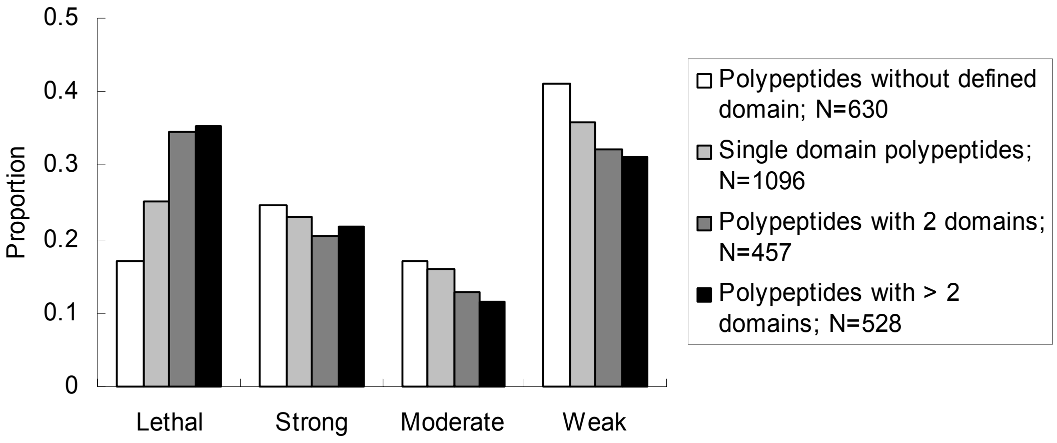

Figure 9. Fitness distribution of gene deletions for polypeptides subdivided according to their domain annotation after exclusion of duplicate genes. Single-domain polypeptides are, on average, more dispensable than multi-domain polypeptides (proportion of weak effect genes, $p < 0.05$; proportion of lethal genes, $p < 10^{-6}$), while polypeptides with 2 or > 2 domains have similar dispensability ($p > 0.1$). N indicates gene number.

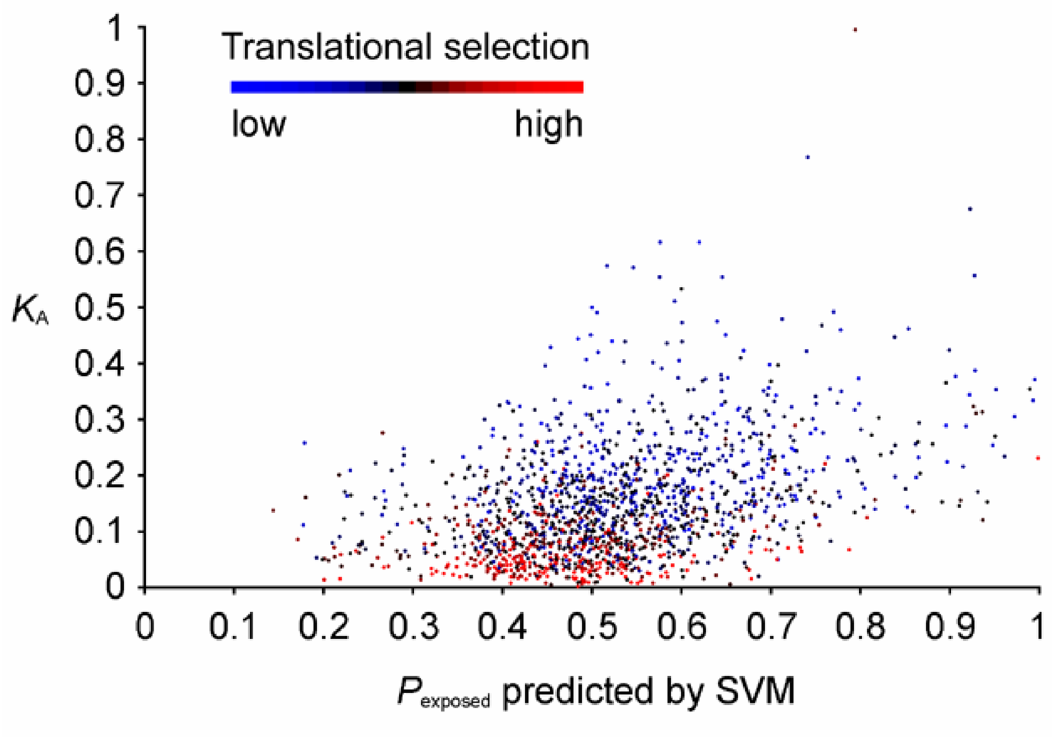

Figure 10. The relationship between nonsynonymous substitution rate (K_A) and $P_{exposed}$ predicted by SVM. Red circles indicate highly expressed genes, while blue circles indicate lowly expressed ones. The strength of translational selection is based on the first component in the principal component analysis for mRNA expression, protein abundance, and CAI values.