

# Gene Complexity and Gene Duplicability

Xionglei He and Jianzhi Zhang\*

Department of Ecology and Evolutionary Biology  
University of Michigan  
Ann Arbor, Michigan 48109

## Summary

Eukaryotic genes are on average more complex than prokaryotic genes in terms of expression regulation, protein length, and protein-domain structure [1–5]. Eukaryotes are also known to have a higher rate of gene duplication than prokaryotes do [6, 7]. Because gene duplication is the primary source of new genes [8, 9], the average gene complexity in a genome may have been increased by gene duplication if complex genes are preferentially duplicated. Here, we test this “gene complexity and gene duplicability” hypothesis with yeast genomic data. We show that, on average, duplicate genes from either whole-genome or individual-gene duplication have longer protein sequences, more functional domains, and more *cis*-regulatory motifs than singleton genes. This phenomenon is not a by-product of previously known mechanisms, such as protein function [10–13], evolutionary rate [14, 15], dosage [11], and dosage balance [16], that influence gene duplicability. Rather, it appears to have resulted from the sub-neofunctionalization process in duplicate-gene evolution [17]. Under this process, complex genes are more likely to be retained after duplication because they are prone to subfunctionalization, and gene complexity is regained via subsequent neofunctionalization. Thus, gene duplication increases both gene number and gene complexity, two important factors in the origin of genomic and organismal complexity.

## Results and Discussion

Gene duplicability is determined by the product of the rate of mutations generating duplication and the probability that a duplicate is fixed and retained in the genome of a species [9]. Separating the two components is important for understanding gene duplicability but is difficult except in two types of duplications. First, for retroduplicates that are dead on arrival, retention is entirely by chance and gene duplicability is governed by the mutation rate only. However, because these duplicates are nonfunctional, they are eventually lost. Second, in whole-genome duplication, all genes within a genome are duplicated at the same time and gene duplicability equals the retention rate. Thus, genome duplication is an ideal situation for examining factors that influence gene retention after duplication.

### Genes Retained from the Yeast Genome Duplication

We took advantage of the well-characterized genome duplication that occurred ~100 million years (MY) ago

in an ancestor of the yeast *Saccharomyces cerevisiae* [18, 19] and of the rich functional genomic data of the yeast. *S. cerevisiae* has a total of 5773 protein-coding nuclear genes. Syntenic comparison of the genome sequence of *S. cerevisiae* and that of the yeast *Kluyveromyces waltii*, which diverged from *S. cerevisiae* before the occurrence of the genome duplication, identified 900 genes (i.e., 450 pairs) that were produced by the genome duplication and are still retained in *S. cerevisiae* [19]. We compared these 900 genes with the rest of the *S. cerevisiae* genes, whose duplicates from the genome duplication have been lost. The retained 900 genes have a mean protein length of  $549.6 \pm 12.3$  (standard error of the mean [SEM]) amino acids, 16% greater than the mean length of the other 4873 genes in the genome ( $475.6 \pm 5.3$ ), and their difference is significant ( $p < 10^{-7}$ , two-tailed Z test;  $p < 0.0001$ , two-tailed Mann-Whitney U test). This difference is not due to a small number of outliers, and this is evident from the protein-length distribution, which shows that the retained duplicates are less frequent than the other genes in the bins of fewer than 400 amino acids but are more frequent in the bins of more than 400 amino acids (Figure 1A). A similar result was obtained when only 450 duplicate genes, one from each pair, were used in the statistical comparison with the rest of the *S. cerevisiae* genes. To reduce the likelihood of including erroneously annotated genes, we followed the procedure in [20] and examined a subset of 4270 *S. cerevisiae* genes that had been examined previously (i.e., they have gene names in addition to open reading frame [ORF] names). Again, the mean length ( $560.2 \pm 14.4$ ) of the 716 retained duplicates (with gene names) is significantly greater than that ( $516.5 \pm 6.4$ ) of 3554 other genes ( $p < 0.003$ , one-tailed Z test;  $p = 0.0001$ , one-tailed U test). A simple explanation of these observations is that longer proteins have a higher probability than shorter ones of being retained after the genome duplication. Alternatively, protein length might have increased in retained duplicates. These two hypotheses can be distinguished by examining *K. waltii*. We found that the mean protein length for the *K. waltii* orthologs of the 450 pairs of *S. cerevisiae* duplicates ( $529.1 \pm 17.2$  amino acids) is 13% greater than that for the other 4332 genes ( $470.1 \pm 5.2$ ) in the *K. waltii* genome ( $p < 0.001$ , one-tailed Z test;  $p < 0.0001$ , one-tailed U test), in strong support of the first hypothesis. A similar result was obtained when *Ashbya gossypii*, another yeast species that diverged from *S. cerevisiae* before the genome duplication [21], was examined. Specifically, *A. gossypii* orthologs (mean =  $535.5 \pm 17.0$  amino acids) of the *S. cerevisiae* duplicates derived from the genome duplication are on average 10% longer than all other genes (mean =  $484.9 \pm 5.3$ ) in the *A. gossypii* genome ( $p < 0.003$ , one-tailed Z test;  $p < 0.001$ , one-tailed U test).

Next, we measured protein complexity for the retained genes from the genome duplication in *S. cerevisiae*. For this, we used the predicted functional domains of *S. cerevisiae* proteins compiled in the Munich Information Center for Protein Sequences (MIPS). The mean

\*Correspondence: jianzhi@umich.edu

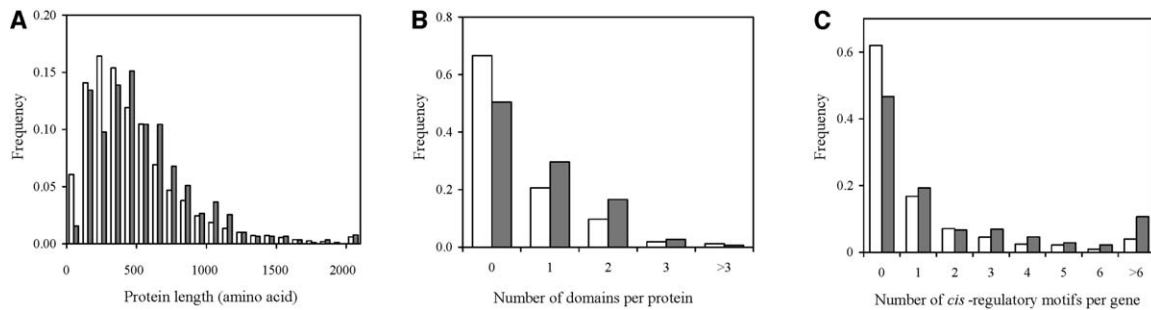


Figure 1. Genes Retained after the Genome Duplication Are More Complex than Other Genes in the Genome of the Yeast *S. cerevisiae*

Shown are the distributions of (A) protein length, (B) number of domains per protein, and (C) number of *cis*-regulatory motifs per gene among genes retained after the genome duplication (filled bars) and all other genes (open bars) in *S. cerevisiae*. The distributional difference between the open and filled bars is significant in each of the three panels ( $p < 10^{-16}$ ,  $p < 10^{-18}$ , and  $p < 10^{-10}$ , respectively;  $\chi^2$  test).

number of domains per protein for the 450 pairs of retained duplicates ( $0.74 \pm 0.03$ ) is 42% greater than the corresponding number ( $0.52 \pm 0.01$ ) for other *S. cerevisiae* genes ( $p < 10^{-11}$ , one-tailed Z test;  $p < 0.0001$ , one-tailed U test). When only genes with names are considered, the above two numbers become  $0.80 \pm 0.03$  and  $0.60 \pm 0.02$ , respectively ( $p < 10^{-7}$ , one-tailed Z test;  $p < 0.0001$ , one-tailed U test). The greater number of domains in retained duplicates is also obvious in the domain-frequency distribution (Figure 1B). For example, the proportion of multidomain ( $\geq 2$  domains) proteins is 19.9% among retained duplicates, in comparison to 12.8% among the rest of the genome ( $p < 10^{-7}$ ,  $\chi^2$  test).

We further examine the regulatory complexity of genes by counting *cis*-regulatory motifs in the intergenic region upstream of each protein-coding gene. These motifs were recently identified by the ChIP-chip experiment (see [Experimental Procedures](#)) in conjunction with computational analysis and are expected to be highly reliable [22]. We excluded from our analysis genes with divergent promoters (i.e., the promoters of two adjacent genes are located in the intergenic region between the two genes) [23]. The mean number of regulatory motifs per gene in the retained duplicates ( $2.34 \pm 0.20$ ) is over twice that ( $1.16 \pm 0.06$ ) in other *S. cerevisiae* genes ( $p < 10^{-8}$ , one-tailed Z test;  $p < 0.0001$ , one-tailed U test). When only genes with names are considered, the number of regulatory motifs per gene becomes  $2.54 \pm 0.24$  for the duplicates and  $1.31 \pm 0.07$  for the other genes ( $p < 10^{-6}$ , one-tailed Z test;  $p < 0.0001$ , one-tailed U test). The distribution of the number of regulatory motifs per gene confirms that this difference reflects a general genomic pattern (Figure 1C). The regulatory-motif dataset also included information on the transcriptional factors that bind the motifs [22]. We found that the mean number of transcriptional factors regulating each retained duplicate ( $1.39 \pm 0.10$ ) is significantly greater than that regulating other *S. cerevisiae* genes ( $0.77 \pm 0.03$ ) ( $p < 10^{-8}$ , one-tailed Z test;  $p < 0.0001$ , one-tailed U test).

#### Genes Retained after Individual Gene Duplications

To test whether higher duplicability for complex genes is also true for individually duplicated genes, we sepa-

rated all protein-coding genes in the *S. cerevisiae* genome into duplicate genes and singleton genes. Duplicate genes are those with at least one duplicate in the genome and are detected by all-against-all BLASTP searches [24]. In contrast, singletons do not have detectable duplicates in the genome. Because the assignment of a gene to either duplicate or singleton genes depends on the specified BLASTP E-value cutoff, multiple cutoffs were used. Here, we only present those results obtained with E-value =  $10^{-5}$  because our conclusion was supported at different cutoffs (see [Why Do Complex Genes Have Higher Duplicability?](#)). We identified 3012 duplicate genes, including 889 genes from the genome duplication and 2123 genes from individual-gene duplications. Eleven genes defined by synteny to be from the genome duplication [19] did not pass the above BLASTP cutoff and were treated as singletons here. This only made our comparison between duplicate and singleton genes more conservative. We found that the average protein length for duplicate genes ( $556.3 \pm 7.0$  amino acids) is 35% greater than that ( $411.6 \pm 6.4$ ) for singletons ( $p < 10^{-46}$ , one-tailed Z test;  $p < 0.0001$ , one-tailed U test) (Figure 2A). A 26% difference is observed between duplicate genes ( $577.6 \pm 8.1$ ) and singletons ( $457.2 \pm 8.1$ ) when only genes with names are considered ( $p < 10^{-25}$ , one-tailed Z test;  $p < 0.0001$ , one-tailed U test). After the separation of duplicate genes into those generated from the genome duplication and those from individual-gene duplications, both groups exhibit significantly greater protein length than singletons do ( $p < 10^{-21}$ , one-tailed Z test;  $p < 0.0001$ , one-tailed U test). We similarly analyzed the number of domains per protein and number of *cis*-regulatory motifs per gene. In both cases, duplicate genes, regardless of whether they are from the genome duplication or individual-gene duplications, are significantly more complex than singletons (Figures 2B and 2C). It is interesting to note that genes resulting from the genome duplication tend to have more *cis*-regulatory motifs than those from individual duplications (Figure 2C), although the cause of this phenomenon is unknown.

#### Gene Complexity and the Number of Paralogs

The observed greater complexity of duplicate genes than of singleton genes led to the prediction of a posi-

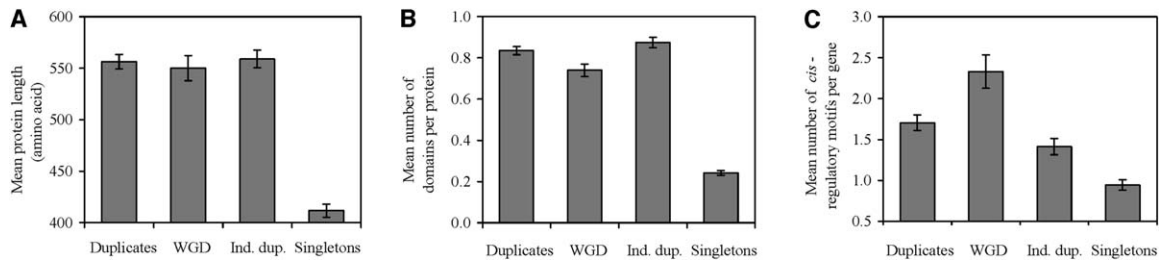


Figure 2. Duplicate Genes Are More Complex than Singleton Genes in the Genome of the Yeast *S. cerevisiae*

Shown here are the average (A) protein length, (B) number of domains per protein, and (C) number of *cis*-regulatory motifs per gene among duplicate and singleton genes in the genome of the yeast *S. cerevisiae*. Error bar shows the standard error of the mean. Duplicates include those resulting from the whole-genome duplication (WGD) and individual-gene duplications (Ind. dup.). For all panels, singletons show significantly smaller values than duplicates, either from the whole-genome duplication or individual-gene duplications.

tive correlation between the complexity of a gene and the number of paralogs ( $N$ ) that it has in the genome. Such correlation was indeed observed when gene complexity was measured by protein length (Figure 3A) or the number of protein domains per gene (Figure 3B). However, the number of *cis*-regulatory motifs per gene increases only slightly when  $N$  is between 1 and 4 and stops increasing or even decreases when  $N > 4$  (Figure 3C). This observation is generally consistent with an earlier study [23] based on a dataset of computationally determined regulatory motifs, and it suggests that the mechanism enhancing the duplicability of genes under sophisticated regulation becomes insignificant when the gene family gets bigger (see Caveats).

**Why Do Complex Genes Have Higher Duplicability?**

There are several factors known to influence gene duplicability, and it is worth examining whether our observation that complex genes have higher duplicability than simple genes is due to any of these factors. First, proteins belonging to protein-protein complexes tend to have reduced rates of gene retention after duplication because duplication generates imbalance in the concentration of the subcomponents of the complex [11]. This imbalance problem, however, does not exist in genome duplications because all subcomponents

are duplicated simultaneously without causing imbalance. Because our observation was made in both whole-genome and individual-gene duplications, it cannot be explained by the imbalance hypothesis.

Second, it has been proposed that haploinsufficient genes have a higher duplicability than haplosufficient genes [16]. Haploinsufficient genes are those that show reduced fitness when one of the two alleles in a diploid becomes nonfunctional, whereas haplosufficient genes show no such fitness reduction. The rationale behind this dosage hypothesis is that the duplication of haploinsufficient genes would confer an immediate advantage because additional products of these genes supposedly lead to increased fitness [16]. Because the dosage effect influences the retention rate after gene duplication, we examined the genome-duplication data for haploinsufficient and haplosufficient genes separately. There are only 28 retained duplicates that are haploinsufficient, 22 of which encode ribosomal proteins. Thus, this sample is not large enough for a meaningful analysis. For haplosufficient genes, the phenomenon of higher duplicability for complex genes still holds (see Table S1 in the Supplemental Data available with this article online), indicating that this phenomenon is independent of the dosage effect.

Third, genes from certain functional categories are

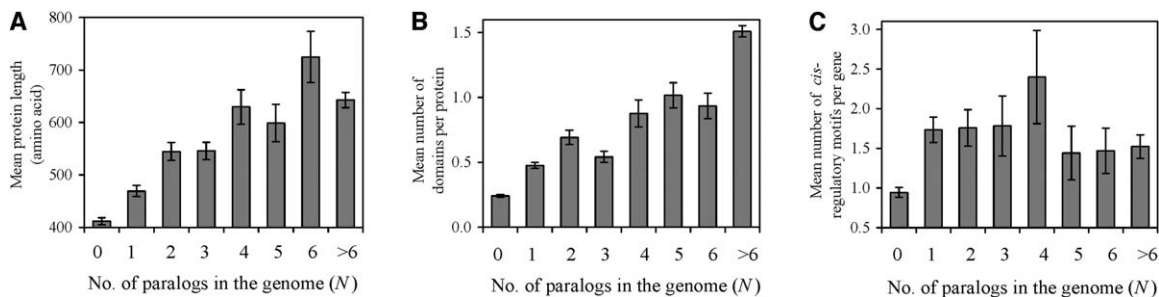


Figure 3. Relationship between Gene Complexity and Number of Gene Paralogs

Shown here are the average (A) protein length, (B) number of domains per protein, and (C) number of *cis*-regulatory motifs per gene among genes with different numbers of paralogs ( $N$ ) in the yeast *S. cerevisiae*. Error bar shows the standard error of the mean. Singletons have  $N = 0$ . Protein length and  $N$  are positively correlated ( $n = 5773$ , linear correlation coefficient  $r = 0.154$ ,  $p < 10^{-31}$ ; Spearman's rank correlation coefficient  $\rho = 0.28$ ,  $p < 10^{-103}$ ). The same is true for the correlation between the number of domains per protein and  $N$  ( $n = 5773$ ,  $r = 0.350$ ,  $p < 10^{-165}$ ;  $\rho = 0.44$ ,  $p < 10^{-271}$ ). The positive correlation between the number of *cis*-regulatory motifs per gene and  $N$  is significant ( $r = 0.032$ ,  $n = 2546$ ,  $p = 0.05$ ;  $\rho = 0.12$ ,  $p < 10^{-8}$ ), but it is mainly due to the difference between genes with  $N = 0$  and those with  $N > 0$ .

known to have higher duplicability than genes from other categories [10–13]. To examine whether our observation could be a result of this effect, we compared gene complexity between the duplicates retained from the genome duplication and all other genes in the genome for each functional category. Note that a gene may be classified into more than one functional category. Our analysis showed that in 21 of 24 comparisons, the retained duplicates are significantly more complex than other genes in the genome, whereas the remaining three comparisons (all on protein size) show no significant differences (Table S2). These results strongly suggest that our observation is not a by-product of variable duplicability of genes belonging to different functional categories.

Fourth, it has been shown that conserved proteins have enhanced duplicability [14, 15]. This could inflate the estimates of the number of domains in duplicates because conserved sequences are more likely to be detected as domains. However, we found that the average number of domains in the duplicates retained from the genome duplication is still greater than that of other genes in *S. cerevisiae*, even after we controlled for the rate of protein evolution, which we estimated by comparing orthologous proteins in *K. waltii* and *A. gossypii* (Table S3).

Because higher duplicability of complex genes than simple genes is observed for both the whole-genome duplication and individual-gene duplications in *S. cerevisiae*, the underlying cause is likely a greater probability of retention for complex genes after duplication rather than a difference in the rate of mutations that generate duplication. It has been demonstrated both theoretically and empirically that partition of ancestral functions (i.e., subfunctionalization) occurs frequently between young duplicates, presumably via complementary degenerate mutations [25–29]. For instance, rapid subfunctionalization after gene duplication has been observed in the yeast for protein interactions [17, 30]. We observed from the *S. cerevisiae* genome duplication data that the numbers of protein domains and regulatory motifs per gene are on average 40%–100% higher for the retained duplicates than for the rest of the genome. This level of difference is probably large enough to cause a difference in the rate of subfunctionalization after duplication, consequently generating a difference in gene retention [26]. Although gene complexity inevitably decreases by subfunctionalization after duplication, previous genomic analysis also revealed subsequent gradual but substantial neofunctionalization [17, 23], explaining why retained duplicates are still more complex than singletons in spite of initial subfunctionalization. It is quite possible that the protein domains or *cis*-regulatory motifs experiencing degenerate mutations shortly after duplication do not deteriorate completely. Instead, they may evolve into domains or motifs with altered specificity or function. Thus, our observation that duplicate genes are more complex than singleton genes is explainable by the sub-neo-functionalization process following gene duplication [17].

Hughes [25] proposed a duplicate-gene evolution model that is sometimes referred to as the “adaptive-conflict” model [31]. In this model, the progenitor gene can conduct multiple pleiotropically constrained func-

tions. Gene duplication enables both copies to become specialized in distinct subsets of the ancestral functions with improved performances, likely by fixations of advantageous mutations. Although this model could explain high retention of complex genes by subfunctionalization, it cannot explain why gene complexity is regained after subfunctionalization; it is difficult to imagine that specialization would generally increase gene complexity.

It should be noted that gene duplicability is also influenced by the rate of mutations that generate duplication. This factor becomes more important in large gene families because the probability of unequal crossover that produces duplication should increase with gene-family size. Consequently, the role of gene retention in determining gene duplicability becomes less prominent. This could explain why gene complexity does not increase with gene-family size in large families (Figure 3). Other factors, including gene function, may also play critical roles in determining the size of large gene families.

#### Caveats

Our analyses may have several caveats. First, the protein-domain dataset used here was based on computational predictions that may contain false negatives and/or false positives. It is possible that the protein-domain annotation is more complete for duplicates than for singletons because duplicates tend to have more homologous sequences in GenBank. However, the ascertainment bias should be minimal because of the availability of many homologous sequences in GenBank for even singleton genes. This is particularly true for the yeast because over a dozen yeast genomes have been sequenced. The regulatory motifs we analyzed were identified by highly accurate experimental methods in conjunction with computational confirmation. Although there may be some false negatives, we see no obvious reason that they would bias our analyses. Furthermore, we conducted an independent analysis with a different regulatory-motif dataset that was based purely on computational predictions [32]. Although the number of motifs per gene is much higher in this dataset, we observe the same trend that genes with more motifs have higher duplicability (Figures S1–S3). Second, although the recognition of retained duplicates from the genome duplication was based on synteny [19] and should not bias our analysis, the separation of duplicate and singleton genes of the *S. cerevisiae* genome was based on BLASTP searches, which have potential biases. Specifically, it is possible that longer proteins are more easily hit in a BLASTP search than shorter ones are, which would result in an upward bias in protein-length estimates for duplicate genes. However, a previous study found BLASTP (E-value cutoffs between  $10^{-3}$  and  $10^{-9}$ ) to be insensitive to protein length [33]. In the present work, we observed only small differences (2.7%) in mean protein length for duplicate genes when a variety of E-value cutoffs (from  $10^{-3}$  to  $10^{-20}$ ) were used, suggesting that such BLASTP biases are minimal in our analysis (Table S4). Furthermore, the number of *cis*-regulatory motifs and protein length are independent from each other (Figure S4); thus, BLASTP searches do not affect our comparison of regulatory motifs between duplicate and singleton genes.

## Implications

In summary, our results show preferential retention of complex genes after duplication. Although gene duplicability is probably determined by multiple factors [9–17, 25–30, 34], our finding may be of special importance in genome evolution. First, the phenomenon we observed at the genomic level supports the role of subfunctionalization in duplicate-gene retention [26]. Because subfunctionalization via complementary degenerate mutations is applicable to all organisms, with a more prominent role in species with smaller populations than with larger populations [7, 28], our finding is expected to be more pronounced for higher organisms, whose populations are generally smaller than those of yeasts. However, subfunctionalization may be unimportant to duplicate retention if the duplicates were already functionally divergent when generated; such duplicates include retroduplicates that differ in expression patterns from their mother genes upon birth or allopolyploidy-generated duplicates that have already been functionally divergent upon polyploidization. Second, because of biased retention of complex genes by subfunctionalization and subsequent events of neofunctionalization [17], gene duplication not only provides raw genetic materials but also materials that are more complex than the genomic average. This feature of gene duplication might be an important force in the evolution of genomic and organismal complexity. Third, the type of differential gene duplicability revealed here may be the basis of the “rich-gets-richer” mechanism that is used to explain the genome-wide power-law distributions of gene-family sizes [35] and numbers of protein-protein interactions [36]. Fourth, the preferential duplication of complex genes is due to the neutral process of subfunctionalization, although later neofunctionalization might be adaptive and involves positive selection. Our finding thus supports and extends the model in which neutral and passive evolution, in addition to the well-recognized role of natural selection and adaptation, plays a prominent role in the origin of genomic complexity [7].

## Experimental Procedures

### Genomic-Sequence Data

The *S. cerevisiae* genome sequence was downloaded from Saccharomyces Genome Database (SGD) (<http://www.yeastgenome.org/>). After Ty transposable elements and mitochondrial genes were excluded, a total of 5773 ORFs was obtained. Among these, 4270 ORFs have gene names. The genomic information of the yeast *Kluyveromyces waltii* was downloaded from <http://www.broad.mit.edu/seq/YeastDuplication>. Five-thousand forty-seven *K. waltii* ORFs have *S. cerevisiae* homologs. After removing Ty elements, mitochondrial genes, ORFs with no DNA or protein sequence information, and 22 ORFs with less than 50 nucleotides (16 of them had zero nucleotide and were apparently incorrect annotations), we obtained 4782 ORFs. Information about the 450 *S. cerevisiae* gene pairs that were products of the genome duplication, as well as about their corresponding single-copy *K. waltii* homologs, was obtained from <http://www.broad.mit.edu/seq/YeastDuplication>. The genomic information of the yeast *Ashbya gossypii* was downloaded from the National Center for Biotechnology Information (NCBI) (<http://www.ncbi.nlm.nih.gov/>). On the basis of the analysis in [21], 528 nonredundant pairs of duplicates resulting from the genome duplication are still retained in *S. cerevisiae*. We compared the single-copy *A. gossypii* homologs of these 528 *S. cerevisiae* gene pairs with all other genes (4190) in the *A. gossypii* genome.

### Determination of Gene Complexity

Predictions of protein domains for *S. cerevisiae* were obtained from MIPS (<ftp://ftpmips.gsf.de/yeast/catalogues/motifs/>). The *cis*-regulatory motifs in the intergenic region upstream of each *S. cerevisiae* ORF were determined by ChIP-chip experiments in conjunction with computational confirmation [22] and were downloaded from [http://jura.wi.mit.edu/fraenkel/regcode/release\\_v24/txtfiles/](http://jura.wi.mit.edu/fraenkel/regcode/release_v24/txtfiles/). ChIP-chip stands for chromatin immunoprecipitation (ChIP), followed by the identification of immunoprecipitated genomic fragments through the use of whole-genome DNA chips. We used the criteria of “binding  $p < 0.005$ , conserved in at least one other yeast,” meaning that the type I error in the experiment was lower than 0.005 and that the motif sequence was conserved in at least another *sensu stricto* *Saccharomyces* species [22]. In accordance with [23], we discarded ORFs with divergent promoters and used the remaining 2546 ORFs whose *cis*-regulatory-motif information was available. We also used another dataset of regulatory motifs, which were predicted with Gibbs sampling algorithm [32] and were compiled in [23]. For this dataset, 3226 ORFs with *cis*-regulatory-motif information were used after the ORFs with divergent promoters were discarded.

### Identification of Duplicate Genes

All-against-all BLASTP [24] searches among *S. cerevisiae* proteins were conducted to separate duplicate and singleton genes. Given the possibility that BLASTP overestimates the mean protein length for duplicate genes, we used five different E-value cutoffs ( $10^{-3}$ ,  $10^{-5}$ ,  $10^{-9}$ ,  $10^{-13}$ , and  $10^{-20}$ ). The results show that the bias was minimal (Table S4). We therefore present only those results based on the cutoff of E-value =  $10^{-5}$ . Similar to [23], we defined the number of paralogs that a gene has in the genome as the number of nonself BLASTP hits that this gene has (E-value =  $10^{-5}$ ).

### Haploinsufficient and Haplosufficient Genes

The single-gene-deletion fitness data for heterozygotes and homozygotes of *S. cerevisiae* were downloaded from [http://www-deletion.stanford.edu/YDPM/YDPM\\_index.html](http://www-deletion.stanford.edu/YDPM/YDPM_index.html). We used the time course 2 datasets derived from growth in the YPD medium and considered those genes whose homozygous deletion strains have fitness values lower than 0.95. A gene is then regarded as haplosufficient when its heterozygous deletion strain has fitness higher than 0.99 or haploinsufficient when its heterozygous deletion strain has fitness lower than 0.95.

### Gene Function Categories

We downloaded functional-classification information for *S. cerevisiae* proteins from <ftp://ftpmips.gsf.de/yeast/catalogues/funecat> (the funecat-2.0 data 28102004 version). There are 19 functional categories in the file, and we restricted our analysis to the largest eight categories because the other categories contain too few genes for meaningful statistical analysis.

### Control for the Rate of Protein Evolution

We conducted all-against-all BLASTP searches between *K. waltii* and *A. gossypii* proteins (E-value =  $10^{-10}$ ) and identified 4096 reciprocal best hits, which were regarded as orthologous pairs. We aligned orthologous genes according to the protein-sequence alignment made by ClustalW [37] and estimated the number of nonsynonymous substitutions per nonsynonymous site ( $d_N$ ) between the orthologs via PAML with default parameters [38]. Information of gene orthology between *S. cerevisiae* and *K. waltii* was obtained from <http://www.broad.mit.edu/seq/YeastDuplication> with manual revisions. We obtained the  $d_N$  values between *K. waltii* and *A. gossypii* orthologs for 822 *S. cerevisiae* duplicate genes (i.e., 411 pairs) that have been retained after the genome duplication as well as for 3522 other *S. cerevisiae* genes.

### Supplemental Data

Four figures and four tables are available with this article online at <http://www.current-biology.com/cgi/content/full/15/11/1016/DC1/>.

## Acknowledgments

We thank L. Hurst and B. Papp for providing the predicted yeast regulatory motifs they compiled and P. Philippsen for providing the whole-genome duplication dataset derived from *Ashbya gossypii*. W. Grus, F. Kondrashov, W.-H. Li, D. Webb, and four anonymous reviewers provided constructive comments. This work was supported in part by National Institutes of Health grant GM67030 to J.Z.

Received: January 19, 2005

Revised: April 13, 2005

Accepted: April 19, 2005

Published: June 7, 2005

## References

- Huang, L., Guan, R.J., and Pardee, A.B. (1999). Evolution of transcriptional control from prokaryotic beginnings to eukaryotic complexities. *Crit. Rev. Eukaryot. Gene Expr.* 9, 175–182.
- Zhang, J. (2000). Protein-length distributions for the three domains of life. *Trends Genet.* 16, 107–109.
- Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., et al. (2000). Comparative genomics of the eukaryotes. *Science* 287, 2204–2215.
- Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Apic, G., Gough, J., and Teichmann, S.A. (2001). An insight into domain combinations. *Bioinformatics* 17 (Suppl 1), S83–S89.
- Yang, J., Lusk, R., and Li, W.H. (2003). Organismal complexity, protein complexity, and gene duplicability. *Proc. Natl. Acad. Sci. USA* 100, 15661–15665.
- Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. *Science* 302, 1401–1404.
- Ohno, S. (1970). *Evolution by Gene Duplication* (New York: Springer-Verlag).
- Zhang, J. (2003). Evolution by gene duplication—an update. *Trends Ecol. Evol.* 18, 292–298.
- Conant, G.C., and Wagner, A. (2002). GenomeHistory: A software tool and its application to fully sequenced genomes. *Nucleic Acids Res.* 30, 3378–3386.
- Papp, B., Pal, C., and Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194–197.
- Marland, E., Prachumwat, A., Maltsev, N., Gu, Z., and Li, W.H. (2004). Higher gene duplicabilities for metabolic proteins than for nonmetabolic proteins in yeast and *E. coli*. *J. Mol. Evol.* 59, 806–814.
- Kondrashov, F.A., Rogozin, B., Wolf, Y.I., and Koonin, E.V. (2002). Selection in the evolution of gene duplications. *Genome Biol.* 3 RESEARCH0008.10.1186/gb-2002-3-2-research0008.
- Davis, J.C., and Petrov, D.A. (2004). Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol.* 2, 318–326.
- Jordan, I.K., Wolf, Y.I., and Koonin, E.V. (2004). Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol. Biol.* 4, 22.
- Kondrashov, F.A., and Koonin, E.V. (2004). A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet.* 20, 287–290.
- He, X., and Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169, 1157–1164.
- Wolfe, K.H., and Shields, D.C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713.
- Kellis, M., Birren, B.W., and Lander, E.S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428, 617–624.
- Gu, Z., Steinmetz, L.M., Gu, X., Scharfe, C., Davis, R.W., and Li, W.H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–66.
- 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.
- Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., Macisaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.B., Reynolds, D.B., Yoo, J., et al. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* 431, 99–104.
- Papp, B., Pal, C., and Hurst, L.D. (2003). Evolution of *cis*-regulatory elements in duplicated genes of yeast. *Trends Genet.* 19, 417–422.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Hughes, A.L. (1994). The evolution of functionally novel proteins after gene duplication. *Proc. R. Soc. Lond. B. Biol. Sci.* 256, 119–124.
- Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545.
- Stoltzfus, A. (1999). On the possibility of constructive neutral evolution. *J. Mol. Evol.* 49, 169–181.
- Lynch, M., and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459–473.
- Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155.
- Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* 18, 1283–1292.
- Lynch, M., and Katju, V. (2004). The altered evolutionary trajectories of gene duplicates. *Trends Genet.* 20, 544–549.
- Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. (2000). Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296, 1205–1214.
- Lipman, D.J., Souvorov, A., Koonin, E.V., Panchenko, A.R., and Tatusova, T.A. (2002). The relationship of protein conservation and sequence length. *BMC Evol. Biol.* 2, 1–10.
- Gibson, T.J., and Spring, J. (1998). Genetic redundancy in vertebrates: Polyploidy and persistence of genes encoding multidomain proteins. *Trends Genet.* 14, 46–49.
- Huynen, M.A., and van Nimwegen, E. (1998). The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* 15, 583–589.
- Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* 5, 101–113.
- Thompson, J.D., Higgins, D.G., and 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.
- Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556.