

Higher Duplicability of Less Important Genes in Yeast Genomes

Xionglei He and Jianzhi Zhang

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

Gene duplication plays an important role in evolution because it is the primary source of new genes. Many recent studies showed that gene duplicability varies considerably among genes. Several considerations led us to hypothesize that less important genes have higher rates of successful duplications, where gene importance is measured by the fitness reduction caused by the deletion of the gene. Here, we test this hypothesis by comparing the importance of two groups of singleton genes in the yeast *Saccharomyces cerevisiae* (Sce). Group *S* genes did not duplicate in four other yeast species examined, whereas group *D* experienced duplication in these species. Consistent with our hypothesis, we found group *D* genes to be less important than group *S* genes. Specifically, 17% of group *D* genes are essential in Sce, compared to 28% for group *S*. Furthermore, deleting a group *D* gene in Sce reduces the fitness by 24% on average, compared to 38% for group *S*. Our subsequent analysis showed that less important genes have more *cis*-regulatory motifs, which could lead to a higher chance of subfunctionalization of duplicate genes and result in an enhanced rate of gene retention. Less important genes may also have weaker dosage imbalance effects and cause fewer genetic perturbations when duplicated. Regardless of the cause, our observation indicates that the previous finding of a less severe fitness consequence of deleting a duplicate gene than deleting a singleton gene is at least in part due to the fact that duplicate genes are intrinsically less important than singleton genes and suggests that the contribution of duplicate genes to genetic robustness has been overestimated.

Introduction

Recent progress in genomics has revived the interest of evolutionary biologists in gene duplication, as substantial evidence suggests that gene duplication plays a major role in genomic and organismal evolution by providing raw genetic material from which new genes are derived (Ohno 1970; Zhang 2003). Many analyses have focused on identifying factors that influence the rate of successful gene duplication or gene duplicability (Force et al. 1999; Conant and Wagner 2002; Yang, Lusk, and Li 2003; Papp, Pal, and Hurst 2003a; Yang and Petrov 2004; Kondrashov and Koonin 2004; Marland et al. 2004; Zhang and Kishino 2004; He and Zhang 2005a, 2005b). In a genome, different genes contribute differently to the survival and reproduction of organisms and thus have different importance. In this work, we define the importance of a gene by the fitness reduction caused by the deletion of the gene in standard laboratory conditions. Three considerations lead us to predict that gene importance and gene duplicability are negatively correlated. First, it was found that proteins belonging to large protein complexes tend to have reduced gene duplicability because duplication generates imbalance in the concentration of the subcomponents of the complex and therefore is selected against (Yang, Lusk, and Li 2003; Papp, Pal and Hurst 2003a). Known as the “centrality and lethality” rule (Jeong et al. 2001), proteins involved in more protein-protein interactions (including protein complexes) are on average more important than those with fewer interactions. From the above two observations, it may be predicted that less important genes have higher duplicability than more important ones. Second, maintaining genetic stability, particularly the stability of central cellular and developmental processes, may be essential for the survival of organisms. Because gene duplication could cause genetic perturbation by doubling gene dosage, one expects that important genes tend to have reduced duplicability.

Key words: gene duplicability, yeast, gene dispensability, gene importance, functional compensation, genetic robustness.

E-mail: jianzhi@umich.edu.

Mol. Biol. Evol. 23(1):144–151, 2006

doi:10.1093/molbev/msj015

Advance Access publication September 8, 2005

Consistent with this prediction, it has been shown that genes functioning in early developmental stages (presumably more important) have lower duplicability than those functioning in late developmental stages (less important) in worms (Castillo-Davis and Hartl 2002; Yang and Li 2004). Third, in yeasts, essential genes, which cause lethality when deleted, are on average regulated by fewer transcription factors than are nonessential genes, a phenomenon that likely reflects the fact that essential genes tend to be housekeeping and have simple expression regulation (Yu et al. 2004). This observation suggests that essential genes should also have fewer *cis*-regulatory motifs than nonessential genes have, as will be demonstrated later. We recently showed that genes with more *cis*-regulatory motifs have higher duplicability (He and Zhang 2005a). This is because when these genes are duplicated, the daughter genes are subject to a higher probability of subfunctionalization that enhances the chance of gene retention, and subsequent neofunctionalization restores the high number of motifs (He and Zhang 2005a, 2005b). Thus, it can be predicted that less important genes have a higher rate of successful duplication.

The above prediction, if verified, has a significant biological implication. Previous studies in yeasts and nematodes found that deleting a duplicate gene tends to cause a less severe phenotype than deleting a singleton gene (Gu et al. 2003; Kamath et al. 2003; Conant and Wagner 2004). This phenomenon is generally ascribed to functional compensation among duplicate genes and is often referred to as genetic robustness by gene duplication (Gu et al. 2003). However, our prediction of higher duplicability of less important genes suggests that the above phenomenon may simply be due to a difference in the intrinsic importance between singletons and duplicates and that the contribution of duplicate genes to genetic robustness may have been overestimated. Here, we directly test our prediction using the fitness effects of single-gene deletions in the yeast *Saccharomyces cerevisiae* (Sce) and the genomic sequences of several related yeast species. The fitness effect is used as a measure of gene importance in Sce. However, because of the confounding factor of potential functional compensations among duplicates, we

cannot directly compare the fitness effects of singleton and duplicate genes in *Sce*. Instead, we limit our analyses to *Sce* singleton genes but measure the duplicability of these genes by examining whether their orthologs have duplicated in four related yeast genomes. Our strategy is similar to that of Davis and Petrov (2004), who studied the relationship between gene duplicability and gene sequence conservation by measuring gene duplicability in *Sce* and sequence conservation in two insects.

Materials and Methods

Genomic Data

Yeast genome sequences were downloaded from the following URLs: *Sce* and *Saccharomyces bayanus* (Sba), ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/; *Kluyveromyces waltii* (Kwa), <http://www.broad.mit.edu/seq/YeastDuplication>; *Ashbya gossypii* (Ago), <ftp://ftp.ncbi.nih.gov/genomes/Fungi/>; *Debaryomyces hansenii* (Dha) and *Yarrowia lipolytica* (Yli), <ftp://ftp.ncbi.nih.gov/genbank/genomes/FUNGI>; *Candida albicans* (Cal), http://www.candidagenome.org/download/sequence/genomic_sequence/; and *Schizosaccharomyces pombe* (Spo), http://www.sanger.ac.uk/Projects/S_pombe/.

The *Sce* homozygous single-gene-deletion fitness data (Steinmetz et al. 2002) were downloaded from http://www-deletion.stanford.edu/YDPM/YDPM_index.html. Following Gu et al. (2003), the lowest fitness value across five growth conditions (YPD, YPDGE, YPE, YPG, and YPL) was used for each strain. In addition, a list of essential genes (Giaever et al. 2002) was downloaded from http://www-sequence.stanford.edu/group/yeast_deletion_project/Essential_ORFs.txt. After excluding genes with contradictory fitness information between the two data sets, a total of 5,724 genes with fitness values were retained for further analysis.

To identify haploinsufficient genes, we downloaded the heterozygous single-gene-deletion fitness data of *Sce* from http://www-deletion.stanford.edu/YDPM/YDPM_index.html. We used the time course two data sets 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 if its heterozygous deletion strain has fitness higher than 0.99 or haploinsufficient if its heterozygous deletion strain has fitness lower than 0.95.

The *cis*-regulatory motifs for *Sce* genes were originally predicted using Gibbs sampling algorithm (Hughes et al. 2000) and were compiled by Papp, Pal, and Hurst (2003b). 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) (Papp, Pal, and Hurst 2003b). We also considered whether to use the *cis*-regulatory motif data set generated by the ChIP-chip method (Harbison et al. 2004). Although this method is unlikely to produce false positives, it may have missed some true motifs. There are only 11 group *D* genes (see *Results* for explanation) that have at least one motif in this data set. We thus did not use this data set in further analysis.

We downloaded *Sce* protein functional classification information from <ftp://ftpmips.gsf.de/yeast/catalogues/funcat> (the funcat-2.0 data 28102004 version). There are 19 functional categories in the file, and we restricted our analysis to the largest nine categories because the other categories contain too few genes for meaningful statistical analysis. A gene may belong to more than one functional category. The information of *Sce* protein complexes was from ftp://genome-ftp.stanford.edu/pub/yeast/literature_curation/go_protein_complex_slim.tab.

The nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* genome sequences were downloaded from Ensembl (<ftp://ftp.ensembl.org/pub/>), and the *C. elegans* RNAi phenotype data set was generated by Kamath et al. (2003).

Data Analyses

We conducted genome-wide all-against-all BlastP searches (Altschul et al. 1990) among *Sce* proteins (E value cutoff = 0.1). Those genes with only self-hits are referred to as singletons. After excluding mitochondrial genes and Ty elements, we obtained 1,587 singleton genes. To examine gene duplicability, these *Sce* singleton genes were used as query sequences to BlastP all proteins from the genomes Dha, Cal, Yli, and Spo (E value = 10^{-10}). We manually corrected erroneous BlastP results that were apparently due to annotation errors in the genome sequences, such as the assignment of the same DNA sequence to two genes. DNA sequences of homologous genes in the five yeasts were aligned according to the protein sequence alignment by ClustalW (Thompson, Higgins, and Gibson 1994). The number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) between a pair of homologous sequences were estimated by the likelihood method implemented in PAML with default parameters (Yang 1997). To examine the effect of pericentromeric and subtelomeric gene location on our result, we removed 20 genes that are closest to the centromere and 20 genes closest to the telomere on each chromosome arm. These genes represent ~22% of all the genes in the *Sce* genome.

Results

Identification of Group *S* and Group *D* Genes

We identified 1,587 singleton genes from the yeast *Sce* and determined their homologous genes in four other yeast species Dha, Cal, Yli, and Spo. The phylogeny of the five yeasts is (((*Sce*, (*Dha*, *Cal*)), *Yli*), *Spo*) (Wolfe 2004; see Supplementary Fig. 1, Supplementary Material online). The four non-*Sce* species were chosen from more than a dozen completely sequenced yeast genomes because they are the most divergent from *Sce*. Therefore, many duplication events are expected to have occurred in these species after their separation from *Sce*. Note that the well-characterized genome duplication in *Sce* (Wolfe and Shields 1997) occurred after the separation of *Sce* from these species (Kellis, Birrin, and Lander 2004; Wolfe 2004; Supplementary Fig. 1, Supplementary Material online). We did not examine more divergent organisms such as animals and plants

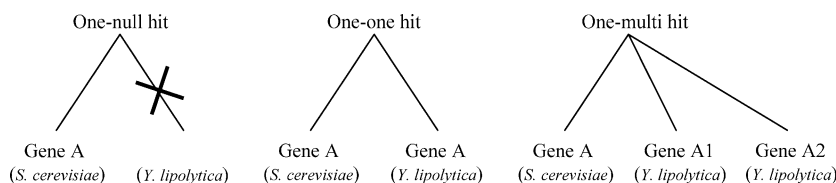


FIG. 1.—Three possible situations when a *Saccharomyces cerevisiae* singleton gene is BlastP searched against another genome such as *Yarrowia lipolytica*.

because homologue identification becomes more difficult in distantly related species. As illustrated in figure 1, a Sce singleton gene may have 0 (one-null), 1 (one-one), or >1 (one-multi) homologous genes in each of the other four yeasts. For example, we identified 613 one-null, 928 one-one, and 46 one-multi cases in Dha (table 1).

We combined the one-one cases from the four species and extracted a nonredundant list of Sce genes. We further removed from this list those genes that belong to one-multi cases in any of the four species. The final list contained 925 Sce genes, referred to as group *S* (standing for singleton) (table 1). These genes are singletons in Sce. Moreover, a group *S* gene has an ortholog in at least one other yeast and has no more than one homolog in any of the yeast species examined. Similarly, we lumped the one-multi cases from the four species and extracted a nonredundant list of 135 Sce genes known as group *M* (standing for multiple) (table 1). These 135 genes are singletons in Sce, but each has at least two homologs in at least one of the other yeasts. Note that in both group *S* and group *M*, it is possible that a Sce gene does not have an ortholog in one, two, or three of the four other yeasts examined.

A one-multi case between Sce and another species is generated by either lineage-specific gene loss in Sce (fig. 2A) or lineage-specific gene duplication in a non-Sce species (fig. 2B). Because we are interested in gene duplication, not gene loss, it is necessary to distinguish between the two scenarios. Let us use Yli-specific gene duplication as an example. The divergence time between the duplicate genes Yli-A1 and Yli-A2 is shorter than the divergence time between Sce-A and Yli-A1 and that between Sce-A and Yli-A2 (fig. 2B). This relationship does not hold in Sce-specific gene losses (Fig. 2A). Ideally, divergence times among the genes can be measured approximately by d_S among the gene sequences (Li 1997) because d_S increases linearly with time and is only minimally affected by changes in selective pressure during evolution, which often occurs after gene duplication (Zhang 2003). However, the gene sequences analyzed here are too divergent for d_S to be accurately es-

timated. We thus used nonsynonymous nucleotide distance (d_N). Because accelerated protein evolution often follows gene duplication, the use of d_N as a proxy for time will tend to overestimate the divergence time between Yli-A1 and Yli-A2 as well as that between Yli-A1 (or Yli-A2) and Sce-A. But the former overestimation is more severe than the latter overestimation because the former estimation is affected by both Yli-A1 and Yli-A2, while the latter is affected by only one of the two genes. This bias will lead to misclassifying lineage-specific gene duplication as gene loss. However, this error is acceptable because it only reduces the size of our sample of lineage-specific gene duplication and reduces statistical power. Following this strategy, we computed d_N among Sce-A, Yli-A1, and Yli-A2 for each one-multi case and identified those cases in which d_N between A1 and A2 is smallest among the three pairwise d_N 's. These cases are most likely due to Yli-specific gene duplication. Only the two best hits are considered if Sce-A has >2 homologs in Yli. Following the same strategy, Dha-, Cal-, and Spo-specific gene duplication cases were identified from the one-multi cases. We then combined these cases and extracted a nonredundant list of 81 Sce genes from the 135 group *M* genes. These 81 genes are referred to as group *D* (standing for duplicate). Obviously, group *D* is a subset of group *M* and has all the properties of group *M*. In addition, group *D* genes experienced lineage-specific duplications (rather than lineage-specific gene losses) in at least one non-Sce yeast genome. The names of group *S*, *M*, and *D* genes are provided in Supplementary data set 1 (Supplementary Material online).

Table 1
BlastP Search Results of 1,587 Sce Singleton Genes Against Four Other Yeast Genomes

Genomes Searched Against	One-Null Cases	One-One Cases	One-Multi Cases
Dha	613	928	46
Cal	591	930	66
Yli	742	802	43
Spo	872	670	44
Nonredundant set		925	135

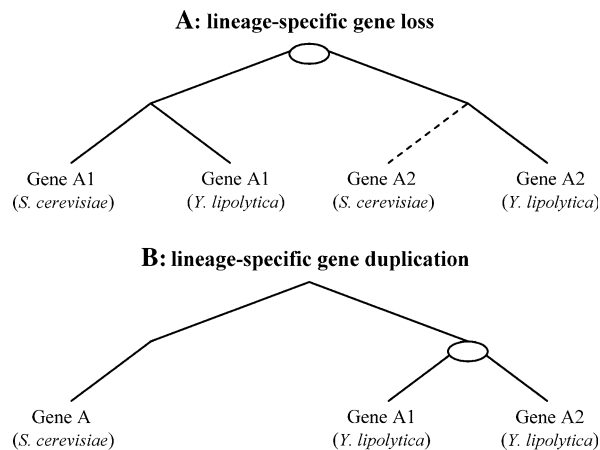


FIG. 2.—Two evolutionary scenarios that can generate one-multi cases shown in figure 1. A circle at an interior node indicates gene duplication and a dotted line indicates gene loss.

Table 2
Comparison Between Group S and Group D genes

	Group S	Group D	P Value
Number of genes	925	81	
Number of genes with fitness information	865	77	
Number of essential genes	246	13	
Percentage of essential genes	28.4	16.9	0.03 ^a
Mean fitness reduction when a gene is deleted	0.38 (0.01 ^b)	0.24 (0.04 ^b)	0.0003 ^c
Mean number of <i>cis</i> -regulatory motifs	15.8 (0.37 ^b)	18.3 (1.42 ^b)	0.05 ^c
Percentage of genes in protein complexes	27.9	17.3	0.04 ^d
Percentage of genes in large protein complexes ^e	20.6	17.3	0.47 ^f

^a $\chi^2 = 4.74$ (df = 1).

^b Standard error of the mean.

^c One-tailed Mann-Whitney *U* test.

^d $\chi^2 = 4.25$ (df = 1).

^e Large complexes have >10 protein components.

^f $\chi^2 = 0.52$ (df = 1).

Comparison Between Group S and Group D Genes

We compared the importance of group *D* and group *S* genes of *Sce* by using the fitness values of single-gene-deletion *Sce* strains. Note that both groups *D* and *S* are singleton genes in *Sce*. Their difference is that group *D* genes duplicated in at least one other yeast examined here, whereas group *S* genes did not duplicate. The fitness values for 77 of the 81 group *D* genes and 865 of the 925 group *S* genes were available (table 2). We found that group *D* genes cause significantly smaller fitness reduction when deleted, compared to group *S* ($P = 0.0003$; table 2). The mean fitness reduction when a group *D* gene is deleted is 0.24 ± 0.04 , compared to 0.38 ± 0.01 when a group *S* gene is deleted. The percentage of essential genes is also significantly lower in group *D* genes (16.9) than in group *S* genes (28.4) ($P = 0.03$; table 2). Furthermore, we classified genes into four categories of importance according to the fitness effect of gene deletion. Compared with group *S*, group *D* is more prevalent in the least important category but less prevalent in the other three categories (fig. 3). All these results suggest that group *D* genes are less important than group *S* genes, supporting our hypothesis of a negative correlation between gene importance and gene duplicability.

To verify the above result in a larger sample, we increased the E value from 10^{-10} to 10^{-5} when we BlastPed *Sce* genes against other yeast genomes. In addition, we measured gene duplicability in six yeast genomes instead of four. The two additional genomes are *Kwa* and *Ago*. The phylogeny of *Sce* and these six yeast species is ((((*Sce*, *Kwa*), *Ago*), (*Dha*, *Cal*)), *Yli*), *Spo*) (Wolfe 2004; Supplementary Fig. 1, Supplementary Material online). The new analysis resulted in 1,152 group *S* genes and 108 group *D* genes. We found that the percentage of essential genes is 24.9 in group *S* and 16.2 in group *D* ($\chi^2 = 3.74$, $P = 0.05$). The mean fitness reduction when a group *S* gene is deleted is 0.34 ± 0.01 , compared to 0.22 ± 0.04 for group *D* ($P = 0.0006$, one-tailed Mann-Whitney *U* test). Thus, this enlarged sample showed a sim-

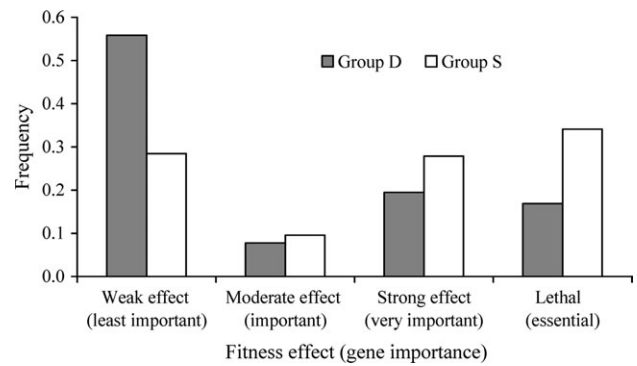


FIG. 3.—Group *D* genes are intrinsically less important than group *S* genes. Group *D* genes duplicated in at least one other yeast, whereas group *S* did not. Both groups are singleton genes in *Saccharomyces cerevisiae* and therefore are not subject to among-duplicate functional compensations. The fitness effect of a gene measures the reduction in fitness when the gene is deleted. Weak effect: fitness of the single-gene-deletion strain > 0.99; moderate effect: $0.95 < \text{fitness} < 0.99$; strong effect: $0 < \text{fitness} < 0.95$; and lethal: fitness = 0. The fitness effects of group *D* genes are significantly lower than those of group *S* genes ($P = 0.0003$, Mann-Whitney *U* test).

ilar pattern that group *D* genes are less important than group *S* genes.

Another way to verify our result is to compare group *S* and group *M*. Although group *M* contains genes that experienced gene loss, all group *M* genes experienced duplication at some time in evolution (fig. 2), in contrast to group *S* genes, which did not duplicate. In other words, gene duplicability is higher for group *M* than group *S*. For both the enlarged data set and the original data set, we found group *M* to be less important than group *S*. For instance, in the enlarged data set, there are 1,152 group *S* genes and 200 group *M* genes. The percentage of essential genes is 24.9 in group *S* and 17.6 in group *M* ($\chi^2 = 4.56$, $P = 0.03$). The mean fitness reduction when a group *S* gene is deleted is 0.34 ± 0.01 , compared to 0.23 ± 0.03 for group *M* ($P < 0.0001$, one-tailed Mann-Whitney *U* test).

Why Do Less Important Genes Have Higher Duplicability?

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. There are no documented factors causing less important genes to duplicate more at the mutational level. The phenomenon of frequent large-scale genetic alterations such as segmental duplication and chromosomal rearrangement in pericentromeric and subtelomeric regions (Eichler and Sankoff 2003) could lead to a high duplication mutation rate of less important genes if less important genes tend to reside in these regions. However, after the removal of group *D* genes that are located in these regions (see *Materials and Methods*), the percentage of essential genes (18.8) and the mean fitness reduction by gene deletion (0.27 ± 0.05) are virtually unchanged (both $P > 0.5$), suggesting that the chromosomal location hypothesis cannot explain our finding. Furthermore, the observed high frequency of duplicates in pericentromeric and subtelomeric regions is probably not due to a high rate of duplication but a high rate of acceptance of duplicates originating from other genomic regions (Lander et al. 2001; Kellis et al. 2003).

We then explore several factors that may affect the fixation and retention rate of duplicate genes. As aforementioned, one possible explanation for the higher retention rate of less important genes is that genes involved in protein complexes tend to be more important (Jeong et al. 2001) but their duplicates tend to be selected against because they cause dosage imbalance (Papp, Pal, and Hurst 2003a). We found that 17% of group *D* genes and 28% of group *S* genes belong to protein complexes ($P = 0.04$; table 2). Thus, it is likely that the dosage imbalance effect has contributed to the observed negative correlation between gene importance and gene duplicability. We also examined the involvement of group *S* and *D* genes in large protein complexes because a previous study showed that genes involved in large complexes (>10 components) have the most severe reduction in gene duplicability (Yang, Lusk, and Li 2003). About 17% of group *D* genes and 21% of group *S* genes belong to large protein complexes. This difference, although not significant ($P = 0.47$, table 2), is in the expected direction.

The second possible explanation is that less important genes have more *cis*-regulatory motifs, which could facilitate subfunctionalization after gene duplication and increase the retention probability of duplicate genes (He and Zhang 2005a). We found that the average number of *cis*-regulatory motifs is indeed higher in less important genes than in more important genes when the entire yeast genome is considered (fig. 4). For group *S* and group *D* genes, we found that the latter have on average 16% more motifs than the former ($P = 0.05$; table 2). Thus, our observation may also be due to *cis*-regulatory motifs.

In addition to the above two possible explanations, duplications of important genes may cause more genetic perturbations than duplications of less important genes and thus may be more likely to be selected against. For example, genes functioning in early developmental stages (presumably more important) have lower duplicability than those functioning in late developmental stages (less important) in worms (Castillo-Davis and Hartl 2002). However, it is difficult to test whether group *S* genes cause more perturbations than group *D* genes when duplicated, as no such information is available in *Sc*.

Haploinsufficient genes have also been suggested to have an elevated duplicability in comparison to haplosufficient genes because doubling the gene dosage of haploinsufficient genes is more likely to be beneficial immediately after gene duplication (Kondrashov and Koonin 2004). Haploinsufficient genes refer to those that show reduced fitness when one of the two alleles becomes nonfunctional, whereas haplosufficient genes show no such fitness reduction. We, however, do not see a significant difference in gene importance between haploinsufficient and haplosufficient genes in yeast. Specifically, we found that $79/146 = 54\%$ of haploinsufficient genes and $653/1180 = 55\%$ of haplosufficient genes are essential ($\chi^2 = 0.08$, $P = 0.78$). The mean fitness reduction caused by deleting a haploinsufficient gene is 0.324, compared to 0.287 for haplosufficient genes ($P = 0.31$, two-tailed Mann-Whitney *U* test). Thus, the influence of haploinsufficiency on gene duplicability cannot generate the correlation between gene importance and duplicability. Based on a different definition of haploinsufficiency, a recent

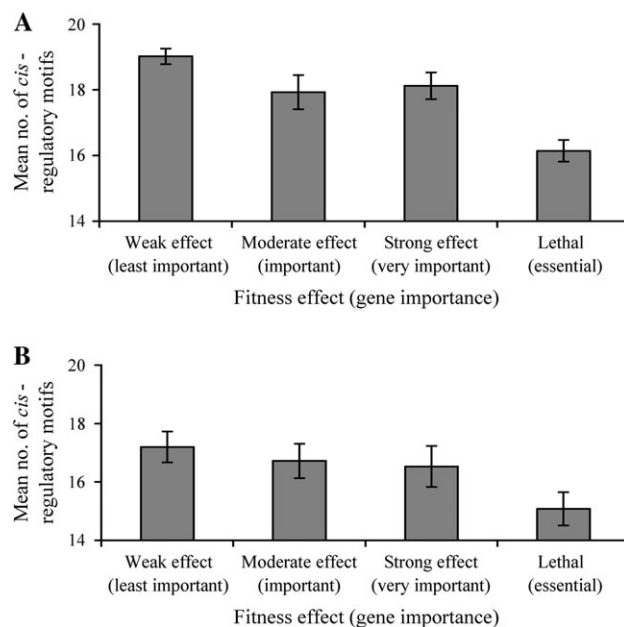


FIG. 4.—More *cis*-regulatory motifs are found in less important genes in the yeast genome. The fitness effect of a gene measures the reduction in fitness when the gene is deleted. Weak effect: fitness of the single-gene-deletion strain > 0.99; moderate effect: $0.95 < \text{fitness} < 0.99$; strong effect: $0 < \text{fitness} < 0.95$; and lethal: fitness = 0. The error bar represents one standard error of the mean. (A) All genes in *Saccharomyces cerevisiae*. The number of genes used in the four fitness categories are 1,520, 299, 621, and 502, respectively. (B) Singleton genes in *S. cerevisiae*. The number of genes used in the four fitness categories are 325, 232, 164, and 146, respectively. In both panels, the fitness effect and number of regulatory motifs are negatively correlated (Spearman's rank correlation $r = -0.12$, $P < 0.0001$ for panel A and $r = -0.09$, $P = 0.013$ for panel B).

study reported that haploinsufficient genes are more likely to be essential than haplosufficient genes (Deutschbauer et al. 2005). This result would predict a positive correlation between gene importance and duplicability, opposite of our observation. Thus, the phenomenon of higher duplicability of less important genes does not appear to be explainable by haploinsufficiency.

It is known that gene duplicability differs among gene functional categories (Conant and Wagner 2002; Papp, Pal, and Hurst 2003a; Marland et al. 2004; He and Zhang 2005a). We examined nine largest functional categories and found that group *D* has a significant higher proportion of metabolic genes ($P = 0.04$) and a significantly lower proportion of genes involved in biogenesis of cellular components ($P = 0.04$), compared with group *S* (table 3). Interestingly, the observation of a higher percentage of metabolic genes in group *D* is in line with a recent study that showed high duplicability of metabolic genes (Marland et al. 2004). However, we found that among the 1,587 *Sc* singleton genes, metabolic genes and nonmetabolic genes are similar in gene importance. For example, the proportion of essential genes among metabolic genes (24.8%) is not significantly different from that among nonmetabolic genes (21.4%) ($\chi^2 = 1.59$, $P = 0.21$). When deleted, the mean fitness reduction is 0.32 ± 0.02 for metabolic genes and 0.30 ± 0.01 for nonmetabolic genes ($P = 0.25$, two-tailed Mann-Whitney *U* test). Thus, although metabolic genes have a higher duplicability, they do not cause the negative

Table 3
Numbers (and Proportions) of Genes Belonging to
Various Functional Categories

Functional Category	Group <i>S</i>	Group <i>D</i>	<i>P</i> Value ^a
Metabolism	243 (0.263)	30 (0.370)	0.037
Cell cycle and DNA processing	123 (0.133)	8 (0.099)	0.380
Transcription	175 (0.189)	9 (0.111)	0.081
Protein synthesis	86 (0.093)	12 (0.148)	0.108
Protein fate (folding, modification, and destination)	156 (0.169)	13 (0.160)	0.851
Protein with binding function or cofactor requirement	139 (0.150)	9 (0.111)	0.340
Cellular transport, transport facilitation, and routes	136 (0.147)	10 (0.123)	0.563
Biogenesis of cellular components	146 (0.158)	6 (0.074)	0.044
Unclassified proteins	203 (0.220)	19 (0.235)	0.753

^a *P* value from the χ^2 test of the hypothesis that there is no difference between groups *S* and *D* in terms of the proportion of genes belonging to the functional category concerned.

correlation between gene importance and duplicability. Compared with other singleton genes, those involved in biogenesis of cellular components have a greater fitness reduction when deleted ($P = 0.001$, two-tailed Mann-Whitney *U* test), although the percentage of essential genes between the two groups is similar ($\chi^2 = 1.73$, $P = 0.19$). This suggests that the differential duplicability among genes of different functional categories could potentially explain our observation.

A positive correlation between gene sequence conservation and gene duplicability was recently reported, although the underlying cause is unknown (Davis and Petrov 2004; see also Jordan, Wolf, and Koonin 2004). This result was obtained from an analysis similar to ours in both approach and sample size. Specifically, the authors estimated gene conservation using d_N between orthologous sequences of two insects and measured the duplicability of the corresponding genes in yeasts (Davis and Petrov 2004). Using the data provided by Zhang and He (2005), we obtained d_N between orthologous genes of *Sce* and *Sba*, a species that is closely related to *Sce*. The average d_N is 0.0753 ± 0.0065 for group *D* and 0.0868 ± 0.0018 for group *S* ($P = 0.01$, one-tailed Mann-Whitney *U* test). Thus, surprisingly, our group *D* genes are both more conserved in sequence and less important in function, when compared to group *S* genes. This appears to be contradictory to the notion that genes conserved at the sequence level tend to be more important (Wall et al. 2005; Zhang and He 2005). However, we note that the genome-wide correlation between sequence conservation and gene importance, although statistically significant, is weak (Zhang and He 2005). This is particularly true for singleton genes (Yang, Gu, and Li 2003; Zhang and He 2005), as is the case of group *S* and *D* genes.

Discussion

In this work, we first predicted from several considerations that less important genes have higher gene duplicability

than more important genes and then tested it by comparing the importance of two groups of singleton genes in the yeast *Sce*, avoiding the confounding factor of potential functional compensations among paralogous genes. Group *S* genes did not duplicate in four other yeast species examined, whereas group *D* experienced duplication in at least one of these species. Consistent with our hypothesis, we found group *D* genes to be less important than group *S* genes. In our analysis, the importance of a *Sce* gene was measured by the amount of fitness reduction caused by the deletion of the gene. It may be argued that this kind of fitness data does not accurately reflect the importance of a gene because the fitness was evaluated in a limited number of laboratory conditions, which can be very different from natural conditions (Papp, Pal, and Hurst 2004). We regard the fitness effect as a first approximation of gene importance. Previous analysis showed that this approximation, while crude for individual genes, is reasonably good at the genomic level because several predictions on gene importance have been confirmed when this approximation was used (Jeong et al. 2001; Gu et al. 2003; Zhang and He 2005). Thus, when no other genome-wide indices of gene importance are available, this fitness effect data set remains the most appropriate for measuring gene importance. One caveat of our analyses is that we examined gene duplicability in non-*Sce* species, while gene importance was measured in *Sce*. However, as long as evolutionary changes in gene duplicability or gene importance are unbiased, our results are valid. The fact that we could detect a discrepancy in gene importance between groups *S* and *D* suggests that the duplicability and/or importance of a gene are relatively conserved among the five yeasts examined. Because gene duplicability or importance can potentially change during evolution (Zhang and He 2005), making it more difficult to detect the discrepancy between groups *S* and *D*, we predict that the true difference between the two groups would be even more pronounced if we could measure gene duplicability and gene importance in the same organisms.

Does the correlation between gene duplicability and gene importance that we observed in yeasts also hold in other organisms? Finding a suitable species to address this question is difficult because the analysis requires the phenotypic information of genome-wide single-gene deletions as well as the genome sequence of at least one related species. One possibility is the nematode *C. elegans*, for which a genome-wide RNA interference (RNAi) experiment has been conducted and a related species (*C. briggsae*) has been completely sequenced. However, RNAi is often ineffective in blocking protein production, and a lack of RNAi phenotype may be due to this reason or high dispensability of the targeted gene. We thus limited our analysis to those genes with RNAi phenotypes. Applying the same approach as used for the yeasts, we identified 300 group *S* and 6 group *D* genes with RNAi phenotypes. Both group *S* and group *D* genes are singletons in *C. elegans*. The difference is that group *S* genes did not duplicate in *C. briggsae*, whereas group *D* genes did. We found that 57% of group *S* genes and 33% of group *D* genes cause various degrees of embryonic lethality in RNAi experiments ($P = 0.05$, one-tailed Mann-Whitney *U* test). Furthermore, 18% of group *S* genes and none of group *D* genes cause various degrees of sterility in RNAi

Table 4
Proportion of Sce Singleton and Duplicate Genes That Are Essential

	E Values					
	10^{-1}	10^{-3}	10^{-5}	10^{-9}	10^{-13}	10^{-20}
Number of duplicate genes identified	3,782	3,001	2,701	2,339	2,160	1,920
Number of essential duplicate genes	683	489	401	320	282	239
Proportion of duplicates that are essential (PE_d)	0.181	0.163	0.148	0.137	0.131	0.124
Proportion of singletons that are essential (PE_s)	0.222	0.222	0.222	0.222	0.222	0.222
PE_d/PE_s	0.813	0.734	0.669	0.616	0.588	0.561

NOTE.—Only genes with fitness-effect information are counted here. Essential genes are those that cause lethality when deleted. Singleton genes have only self-hits at E value = 0.1.

experiments ($P = 0.08$, one-tailed Mann-Whitney U test). Thus, there is some indication that the observation of higher duplicability of less important genes is not limited to yeasts.

Our analysis suggests that higher duplicability of less important genes in yeasts is related to at least three factors. First, genes involved in protein complexes tend to be more important (Jeong et al. 2001) and tend to have lower duplicability because of the dosage imbalance caused by duplication (Papp, Pal, and Hurst 2003a). Second, less important genes have more *cis*-regulatory motifs, which facilitate sub-functionalization after gene duplication and increase the retention probability of duplicate genes (He and Zhang 2005a). Third, genes from some functional categories are both less important and more duplicable than genes from other categories. However, our data set is not large enough for us to separate the contributions of these factors. In the future, it would be interesting to study whether these factors independently contribute to our observation. Because of the small sample size, we did not test the cause of our observation in nematodes.

Regardless of the underlying cause, the finding of higher duplicability of less important genes has significant implications. Gu et al. (2003) observed that deleting a duplicate gene tends to cause a smaller fitness reduction than deleting a singleton gene in *Sce*. A similar observation was made from the genome-wide RNAi experiments in *C. elegans* (Kamath et al. 2003; Conant and Wagner 2004). Under the assumption that duplicate genes and singleton genes have similar intrinsic importance, Gu et al. (2003) concluded that their observation was due to functional compensation among duplicate genes, which, as they estimated, accounts for at least a quarter of no-phenotype gene deletions. Davis and Petrov (2004) argued that the extent of functional compensation could be even larger based on their observation of higher duplicability of more conserved genes and the presumption that conserved genes cause greater fitness reduction when deleted. However, as aforementioned, sequence conservation of a gene only weakly correlates with its fitness effect (Zhang and He 2005) and therefore has no direct relevance to functional compensation. Our present study directly measures gene importance and gene duplicability and suggests that the observation of Gu et al. (2003) was, at least in part, due to the simple fact that less important genes have a higher rate of successful duplication. We noticed that 16.9% of group *D* genes and 28.4% of group *S* genes are essential, with the ratio between the two percentages being $\lambda = 0.595$. In the entire *Sce* genome, the proportion of essential

genes among duplicates varies from $PE_d = 12.4\%$ to 18.1%, as the E value cutoff used in BlastP searches changes from 10^{-20} to 10^{-1} (table 4). The proportion of essential genes among all singletons of *Sce* is $PE_s = 22.2\%$, if we use the definition that singletons have only self-hits at the E value cutoff of 10^{-1} . Thus, PE_d/PE_s varies from 0.561 to 0.813 (table 4). It is interesting that PE_d/PE_s is either similar to or higher than λ . Because λ was derived from *Sce* singletons without the involvement of among-duplicate functional compensation, the above comparison between PE_d/PE_s and λ suggests the possibility that functional compensation is not needed at all for explaining the dispensability discrepancy between duplicates and singletons in *Sce*. The average percentage of essential genes in groups *S* and *D* is 27.5%, while the corresponding number in group *D* is 16.9%. The average percentage of essential genes in the entire *Sce* genome is 19.2%. If our observations in groups *S* and *D* can be applied to the entire yeast genome, one would predict that a gene that will have successful duplication in *Sce* has a probability of $(16.9\%/27.5\%) \times 19.2\% = 11.8\%$ to be essential. This said, we caution that the proportion of essential genes in group *D* was estimated from a relatively small number of genes and may contain a large sampling error. Furthermore, the observation that the fitness effect of deleting a duplicate gene is positively correlated with the sequence dissimilarity of this gene to its closest paralog (Gu et al. 2003; Conant and Wagner 2004) strongly suggests functional compensation among duplicates. On the other hand, rapid divergence in sequence, expression, and function of duplicates following duplication argues against functional compensation between old duplicates (Wagner 2005). We conclude that it is necessary to reevaluate the prevalence of among-duplicate functional compensation and requantify the contribution of duplicate genes to genetic robustness.

Supplementary Material

Supplementary Fig. 1 and data set 1 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

Acknowledgments

We thank Soochin Cho, Wendy Grus, Zhenglong Gu, Michael Lynch, Ondrej Podlaha, Ken Wolfe, and an anonymous reviewer for valuable comments. This work was

supported by research grants from National Institutes of Health and University of Michigan to J.Z.

Literature Cited

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410.
- Castillo-Davis, C. I., and D. L. Hartl. 2002. Genome evolution and developmental constraint in *Caenorhabditis elegans*. *Mol. Biol. Evol.* **19**:728–735.
- Conant, G. C., and A. Wagner. 2002. GenomeHistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res.* **30**:3378–3386.
- . 2004. Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. *Proc. Biol. Sci.* **271**:89–96.
- Davis, J. C., and D. A. Petrov. 2004. Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol.* **2**:318–326.
- Deuschbauer, A. M., D. F. Jaramillo, M. Proctor, J. Kumm, M. E. Hillenmeyer, R. W. Davis, C. Nislow, and G. Giaever. 2005. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169**:1915–1925.
- Eichler, E. E., and D. Sankoff. 2003. Structural dynamics of eukaryotic chromosome evolution. *Science* **301**:793–797.
- Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**:1531–1545.
- Giaever, G., A. M. Chu, L. Ni et al. (72 co-authors). 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**:387–391.
- Gu, Z., L. M. Steinmetz, X. Gu, C. Scharfe, R. W. Davis, and W. H. Li. 2003. Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**:63–66.
- Harbison, C. T., D. B. Gordon, T. I. Lee et al. (20-authors). 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**:99–104.
- He, X., and J. Zhang. 2005a. Gene complexity and gene duplicability. *Curr. Biol.* **15**:1016–1021.
- . 2005b. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**:1157–1164.
- Hughes, J. D., P. W. Estep, S. Tavazoie, and G. M. Church. 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**:1205–1214.
- Jeong, H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. 2001. Lethality and centrality in protein networks. *Nature* **411**:41–42.
- Jordan, I. K., Y. I. Wolf, and E. V. Koonin. 2004. Duplicated genes evolve slower than singletons despite the initial rate increase. *BMC Evol. Biol.* **4**:22.
- Kamath, R. S., A. G. Fraser, Y. Dong et al. (13 co-authors). 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**:231–237.
- Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**:241–254.
- Kellis, M., B. W. Birren, and E. S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**:617–624.
- Kondrashov, F. A., and E. V. Koonin. 2004. A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet.* **20**:287–290.
- Lander, E. S., L. M. Linton, B. Birren et al. (255 co-authors). 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.
- Li, W. H. 1997. *Molecular evolution*. Sinauer, Sunderland, Mass.
- Marland, E., A. Prachumwat, N. Maltsev, Z. Gu, and W. H. Li. 2004. Higher gene duplicabilities for metabolic proteins than for nonmetabolic proteins in yeast and *E. coli*. *J. Mol. Evol.* **59**:806–814.
- Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, New York.
- Papp, B., C. Pal, and L. D. Hurst. 2003a. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**:194–197.
- . 2003b. Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet.* **19**:417–422.
- . 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**:661–664.
- Steinmetz, L. M., C. Scharfe, and A. M. Deuschbauer et al. (11 co-authors). 2002. Systematic screen for human disease genes in yeast. *Nat. Genet.* **31**:400–404.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 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.
- Wagner, A. 2005. Distributed robustness versus redundancy as causes of mutational robustness. *Bioessays* **27**:176–88.
- Wall, D. P., A. E. Hirsh, H. B. Fraser, J. Kumm, G. Giaever, M. B. Eisen, and M. W. Feldman. 2005. Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci. USA* **102**:5483–5488.
- Wolfe, K. H. 2004. Evolutionary genomics: yeasts accelerate beyond BLAST. *Curr. Biol.* **14**:R392–R394.
- Wolfe, K. H., and D. C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**:708–713.
- Yang, J., and W. H. Li. 2004. Developmental constraint on gene duplicability in fruit flies and nematodes. *Gene* **340**:237–240.
- Yang, J., Z. Gu, and W. H. Li. 2003. Rate of protein evolution versus fitness effect of gene deletion. *Mol. Biol. Evol.* **20**:772–774.
- Yang, J., R. Lusk, and W. H. Li. 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. *Comput. Appl. Biosci.* **13**:555–556.
- Yu, H., D. Greenbaum, H. X. Lu, X. Zhu, and M. Gerstein. 2004. Genomic analysis of essentiality within protein networks. *Trends Genet.* **20**:227–31.
- Zhang, J. 2003. Evolution by gene duplication—an update. *Trends Ecol. Evol.* **18**:292–298.
- Zhang, J., and X. He. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol. Biol. Evol.* **22**:1147–1155.
- Zhang, Z., and H. Kishino. 2004. Genomic background predicts the fate of duplicated genes: evidence from the yeast genome. *Genetics* **166**:1995–1999.

Kenneth Wolfe, Associate Editor

Accepted September 1, 2005