Nucleic Acids Research Advance Access published November 27, 2013

Nucleic Acids Research, 2013, 1–10
doi:10.1093/nar/gkt1200

Genetic robustness and functional evolution of gene duplicates

Germán Plata[1,2] and Dennis Vitkup[1,3,*]

[1]Department of Systems Biology, Center for Computational Biology and Bioinformatics, Columbia University, New York City, NY 10032, USA, [2]Integrated Program in Cellular, Molecular, Structural, and Genetic Studies, Columbia University, New York City, NY 10032, USA and [3]Department of Biomedical Informatics, Columbia University, New York City, NY 10032, USA

Received September 5, 2013; Revised November 3, 2013; Accepted November 4, 2013

ABSTRACT

Gene duplications are a major source of evolutionary innovations. Understanding the functional divergence of duplicates and their role in genetic robustness is an important challenge in biology. Previously, analyses of genetic robustness were primarily focused on duplicates essentiality and epistasis in several laboratory conditions. In this study, we use several quantitative data sets to understand compensatory interactions between *Saccharomyces cerevisiae* duplicates that are likely to be relevant in natural biological populations. We find that, owing to their high functional load, close duplicates are unlikely to provide substantial backup in the context of large natural populations. Interestingly, as duplicates diverge from each other, their overall functional load is reduced. At intermediate divergence distances the quantitative decrease in fitness due to removal of one duplicate becomes smaller. At these distances, yeast duplicates display more balanced functional loads and their transcriptional control becomes significantly more complex. As yeast duplicates diverge beyond 70% sequence their ability to compensate for each other becomes similar to that of random pairs of singletons.

INTRODUCTION

Survival of biological systems crucially depends on robustness to harmful genetic mutations, i.e. genetic robustness, and to changes in environmental conditions (1–3). Two distinct mechanisms of genetic robustness have been previously discussed. First, alternative signaling and metabolic pathways provide an important mechanism for rerouting in many molecular networks (4,5). Second, a major role in genetic robustness is attributed to gene duplicates (1,6). Gene duplications are frequent in evolution and range in size from small-scale (SSD) to whole-genome events (WGD) (7,8). While in ~90% of the cases one duplicate is eventually lost in evolution (6), duplicated genes that remain in the genome can, at least partially, backup each other's functions. Importantly, functional compensation by duplicates plays a significant role in buffering deleterious human mutations (9).

Genetic robustness due to gene duplicates is inherently tied to their functional divergence. Duplicates that acquire distinct molecular functions (MFs) are naturally unable to compensate for one another. In addition, even if MF is conserved, incomplete compensation between duplicates is possible owing to different expression patterns or dosage effects. Gene duplications are the major source of new genes (10) and several conceptual models of duplicates' evolution have been proposed (11,12). In the neofunctionalization model one duplicate gains new functions, i.e. functions not associated with the ancestral gene, while the other duplicate retains the ancestral functions (10,13,14). In contrast, in the subfunctionalization model both duplicates become indispensable and are retained in evolution by partitioning the ancestral gene functions (15,16). Both these models imply an eventual loss of the ability of duplicates to fully substitute for each other. It is also likely that a significant fraction of duplicates are fixed and retained in genomes owing to selective advantages, such as dosage effects or condition-specific expression patterns, present from the moment of duplication (17,18). In cases of fixation due to a selective advantage, full compensation between duplicates is unlikely.

Even though full compensation between duplicates is not expected in the long term, the ability of duplicates to buffer deleterious mutations of their paralogs has been now demonstrated by several independent observations. These include a lower than expected fraction of essential genes with close duplicates (1), a paucity of pairwise epistatic interactions involving duplicated genes (19), and an excess of aggravating genetic interactions between paralogs (20,21). The contribution of duplicates

*To whom correspondence should be addressed. Tel: +1 212 851 5152; Fax: +1 212 851 5149; Email: dv2121@columbia.edu

to robustness has been primarily considered in the context of qualitative or quantitative growth phenotypes either in nutrient rich or in a small number of laboratory conditions (1,22,23). Although popular in experiments, these conditions are unlikely to approximate well a natural 'milieu' of living systems, which are constantly bombarded by a diverse array of environmental stresses and stimuli. Perhaps more importantly, even if there is a strong compensatory interaction between a pair of duplicates, an evolutionary relevant decrease in fitness can still persist—due to an incomplete buffering—after a damaging mutation in one of the duplicates (24). In the context of long-term evolution, there may not be much difference between mutations leading to the lethal phenotype and mutations associated with a fitness decrease substantially larger than the inverse of the effective population size (25,26). Given that typical population sizes of free-living microbial species are large ($>10^6$–10^8) (27), even a small fitness decrease can be effectively lethal for these organisms. Consequently, quantitative analyses of growth phenotypes, preferably in multiple environmental conditions, are necessary to understand the extent to which compensation between duplicates plays an important role in natural biological populations. Here we perform such an analysis and show that in the context of natural populations, genetic buffering mediated by duplicates is likely to be rare and, surprisingly, it is not a monotonic function of duplicates' divergence.

MATERIALS AND METHODS

Gene and protein sequences for *Saccharomyces cerevisiae*, *Saccharomyces paradoxus*, *Saccharomyces bayanus*, *Saccharomyces castelli*, *Saccharomyces mikatae*, *Saccharomyces kudriavzevii* and *Saccharomyces kluyveri* were obtained from the Saccharomyces Genome Database (SGD; http://downloads.yeastgenome.org/) and the study by Kellis *et al.* (28). Pairs of gene duplicates were identified by sequence homology between proteins within each genome using BLASTP (29). Only duplicates that were bidirectional best hits and could be aligned by >80% of each open reading frame's sequence length were considered in our analysis (30). Following previous studies (1), we excluded ribosomal genes from the analysis owing to their high expression, dominant impact on growth and strong codon adaptation bias. Evolutionary distances between duplicated genes were estimated using the method of Yang and Nielsen (31) implemented in the PAML package (32); the use of other methods, such as maximum likelihood, to estimate Ka and Ks did not significantly change the observed patterns (Supplementary Figure S1A).

We used the data obtained by Hillenmeyer *et al.* (33) to measure the fitness contribution of duplicates across multiple environmental conditions and chemical perturbations. Using a *P*-value cutoff of 0.01, we obtained the number of experimental conditions for which a growth defect was observed for every single gene deletion mutant. We also analyzed quantitative growth measurements for double and single deletion yeast strains obtained from DeLuna *et al.* (34) and Costanzo *et al.* (35). Gene essentiality data was obtained from the study of Giaever *et al.* (36).

To functionally characterize duplicated genes, Gene Ontology (GO) (37) annotations were collected from SGD and Enzyme Commission (EC) annotations from the Comprehensive Yeast Genome Database (CYGD) (38). Transcription factor binding motifs used in our work were compiled from Kafri *et al.* (39) and the high-confidence predictions in Kellis *et al.* (28). We used protein localization data from Huh *et al.* (40), Codon Adaptation Index (CAI) calculations based on the data set by Lu *et al.* (41) and the annotation of protein complexes in CYGD.

RESULTS

Hillenmeyer *et al.* (33) quantified growth phenotypes of single-gene yeast deletion strains in a large collection of environmental conditions. The assembled data set contains ~5.5 million phenotypes of heterozygous and homozygous mutants in ~400 conditions. The sampled conditions represent 27 different environmental stresses and hundreds of perturbations with diverse chemical compounds. Environmental stresses comprised different growth media, media lacking specific vitamins or amino acids, as well as different pH and temperature regimes. This comprehensive collection of phenotypes allowed us to investigate in detail the diversification of duplicates' functions and their contribution to genetic robustness in multiple conditions.

We first investigated how the average number of sensitive conditions, i.e. conditions with a significant growth decrease due to deletion of one duplicate, depends on sequence divergence (Ka) between the duplicated genes (Figure 1A and B). We considered the fraction of different conditions with a growth phenotype as a quantitative measure of compensation capacity for duplicates at various divergence distances. For close duplicates the average number of sensitive conditions is not significantly different from that of a random pair of yeast singletons (Figure 1B, horizontal line). Importantly, this result does not imply that random gene pairs and close duplicates are equivalent in terms of the similarity of their MF. As we demonstrate below, the observed pattern is likely due to a higher overall functional load of close duplicates. Here and throughout the article we use the term 'functional load' of a gene to characterize the average fitness decrease—across considered conditions—due to the gene deletion; we note that, based on the definition above, the functional load is not a measure of the total number of MFs a gene has, but it reflects the gene's overall fitness contribution.

Interestingly, the number of sensitive conditions initially drops as duplicates diverge, decreasing about 30% at the distances corresponding to $Ka \approx 0.1$ ($Ks \approx 1$, see Supplementary Figure S2A and B). As duplicates diverge further, the average number of sensitive conditions increases again, reaching the average for a random pair of yeast singletons at $Ka \approx 0.25$. The trend shown in

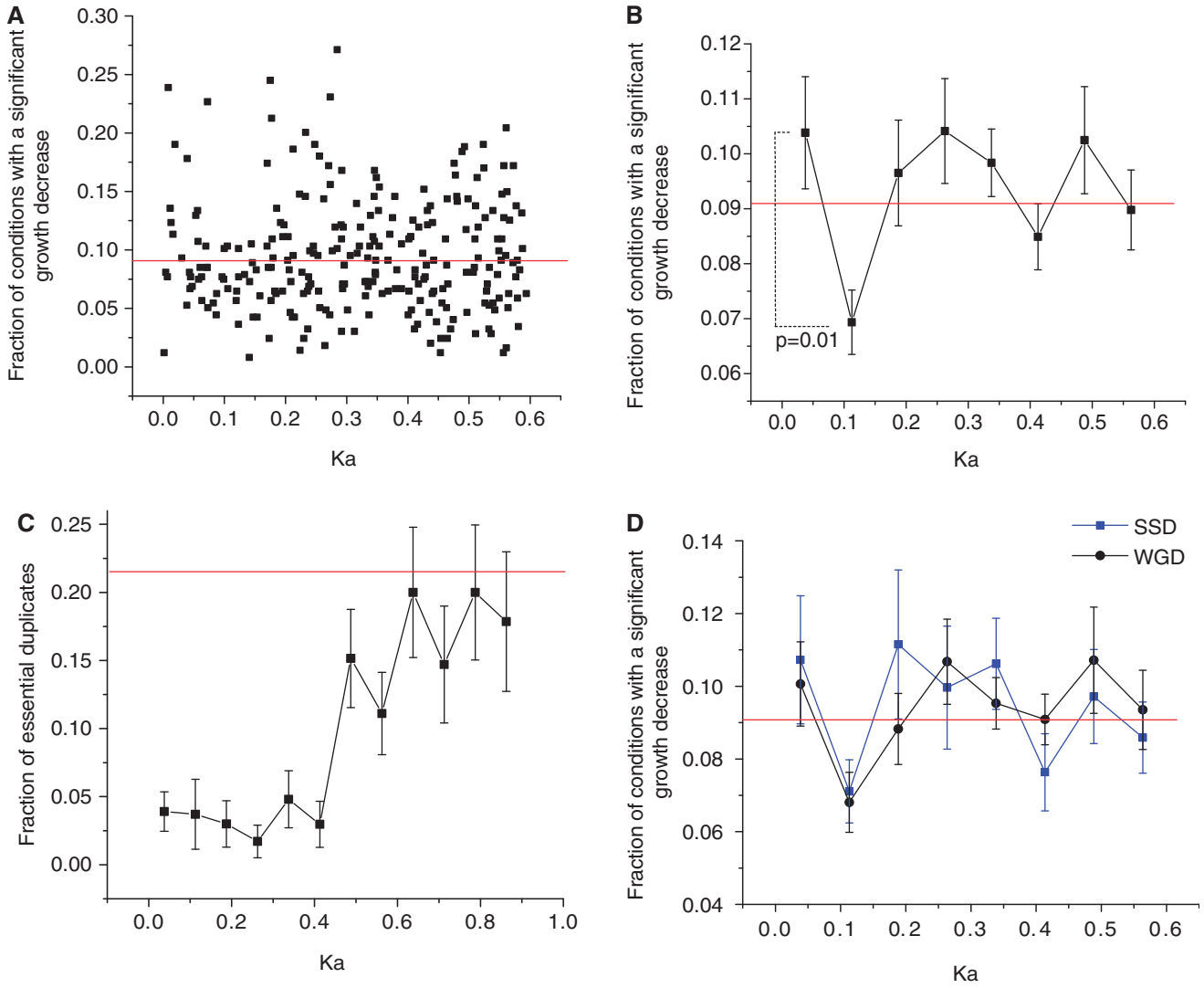

Figure 1. Compensation patterns between yeast duplicates as a function of their evolutionary divergence, Ka, the number of nonsynonymous substitutions per site. (**A**) Scatterplot of the fraction of sensitive conditions, i.e. conditions with detectable growth phenotypes resulting from duplicate gene deletion, versus Ka. Each dot in the figure represents a pair of yeast duplicates. (**B**) The average fraction of sensitive conditions per duplicate pair. The *P*-value was calculated using the Mann–Whitney U test. The horizontal lines in A and B indicate the average fraction of sensitive conditions for a random pair of yeast singletons. (**C**) The average fraction of essential duplicates, i.e. duplicates with a lethal phenotype on deletion, as a function of Ka. The horizontal line indicates the fraction of essential yeast singletons. Gene essentiality data were obtained from the *Saccharomyces* Genome Deletion Project (36). (**D**) Fraction of conditions with a significant growth decrease for deletion of yeast duplicates arising from small-scale (SSD) and whole-genome duplications (WGD). The duplicates were classified as SSD or WGD based on the study by Kellis *et al.* (8) The horizontal line shows the average fraction of sensitive conditions for a random pair of yeast singletons. In the figures, error bars represent the standard error of the mean (SEM).

Figure 1B is not sensitive to the *P*-value cutoff used to determine the significance of the growth decrease observed in mutant strains (Supplementary Figure S3). A similar trend was also observed for the average growth decrease (functional load), measured either by log ratios or Z-scores across all tested conditions (Supplementary Figure S4A and B). Bin-free analyses of the data (Supplementary Figures S1B and C and S2B) also revealed a smaller fitness cost due to the loss of duplicates at intermediate distances (Ka $\approx$ 0.1).

Because most actively growing wild-type yeast populations are diploid (42), we mainly focused our analysis on heterozygous mutant strains. The patterns of functional compensation for heterozygous and homozygous mutants are similar when multiple-drug resistance genes, as defined by Hillenmeyer *et al.* (33), are not considered (Supplementary Figure S4C). The trends also remain similar when only environmental perturbations are analyzed in the homozygous experiments (Supplementary Figure S4D). We also checked that the observed compensation patterns due to closest duplicates are not significantly influenced by additional, i.e. more diverged, paralogs (Supplementary Figure S4E). This lack of significant compensation by diverged duplicates results in an approximately linear relationship between the number of sensitive conditions per yeast protein family and the

family size (Supplementary Figure S5). Finally, the observed compensation patterns were not affected by removal of gene pairs with a high CAI (Supplementary Figure S6A), suggesting that the observed trend cannot be explained by expression-based constraints on the rate of duplicate sequence evolution (Ka) (43) or high expression levels of certain duplicates.

It is interesting to compare the ability of duplicates to buffer mutations leading to any detectable growth decrease beyond a given fitness threshold (Figure 1B) and their role in protecting against the no-growth phenotype, i.e. the likelihood to observe essential genes in duplicate pairs. In Figure 1C, using data from the study by Giaever *et al*. (36), we show the fraction of essential duplicates as a function of their divergence. In agreement with previous studies (1,22,23) we found that the fraction of essential genes remains low and approximately constant for close duplicates, and increases substantially only at divergence distances corresponding to $Ka > 0.4$. Notably, this pattern is qualitatively different from the compensation for quantitative growth phenotypes (Figure 1B), demonstrating the aforementioned impact of using quantitative phenotypes to assess the evolutionarily relevant consequences of mutations. Also in contrast to patterns obtained in studies based on essential genes (22), we observed similar compensation profiles for gene pairs originating from small-scale and genome-wide duplications (Figure 1D). Because all WGD duplicates have the same age, this result suggests that the ability of duplicates to buffer each other's function across multiple conditions depends more strongly on their sequence divergence than on the time since duplication.

It is likely that the observed decrease in the number of sensitive conditions at intermediate divergence distances ($Ka \approx 0.1$) is due to a decrease of the functional load carried at these distances by the union of duplicate genes. To explore this possibility, we considered the quantitative fitness data from DeLuna *et al*. (34) and the synthetic genetic array (SGA) data from Costanzo *et al*. (35). In these studies, the authors performed quantitative growth measurements of yeast strains with individual and simultaneous deletions of duplicates. Using the single deletion phenotypes from the DeLuna *et al*. (Figure 2A) and Costanzo *et al*. studies (Figure 2B), we observed fitness profiles similar to the one obtained based on the data from Hillenmeyer *et al*. (Figure 1B) as a function of Ka, with smaller phenotypic effects at intermediate distances. Interestingly, the overall functional load of duplicate pairs, measured by the phenotype of double deletions, indeed substantially decreases with their divergence (Figure 2C and D). This result suggests that while close duplicates are more likely to have similar functions, their higher functional load makes complete compensation less likely. Because the overall functional load of duplicates remains approximately constant for $Ka > 0.15$, the higher fraction of detectable growth phenotypes at these distances is likely due to a decreased ability for functional compensation as duplicates diverge. Compensation between duplicates quantified by the presence of aggravating interactions between duplicate pairs decreases as a function of sequence divergence (Figure 2E and F) [see (21)].

Besides a smaller overall functional load, it is possible that duplicates at intermediate distances have other properties that favor genetic robustness. To explore this possibility, for each duplicate pair, we looked at the gene with the largest and the gene with the smallest number of sensitive conditions (Figure 3A). Notably, while the duplicate with more conditions (Figure 3A, more sensitive duplicate) follows the average trend for all duplicates (Figure 1B), the duplicate with fewer conditions (Figure 3A, less sensitive duplicate) shows a steady gain in the number of conditions as a function of Ka. Consequently, the functional load of close duplicates, measured by the number of sensitive conditions, is different, and this difference becomes significantly smaller as the genes diverge (Figure 3B. Pearson's $r = -0.64$, $P = 7 \times 10^{-4}$, see also Supplementary Figure S2C). Close duplicates with the larger number of sensitive conditions also show a higher evolutionary constraint, evaluated by the normalized ratio of nonsynonymous to synonymous substitutions per nucleotide site, Ka/Ks (Wilcoxon Signed Rank test $P = 7 \times 10^{-3}$, Figure 2C). This result agrees with previous reports of asymmetric evolution of duplicates in the context of co-expression, genetic interaction and protein–protein interaction networks (19,44,45). The observed asymmetry in the functional load between close duplicates can make buffering difficult. For example, if the less sensitive duplicate is expressed only under specific environmental conditions.

To further explore the mechanism behind the observed backup patterns, we analyzed the functional diversification of yeast duplicates as a function of their sequence divergence (Ka). First, for genes encoding metabolic enzymes we calculated the fraction of gene pairs with conserved EC numbers (Figure 4A); the conservation of EC numbers indicates that corresponding proteins catalyze identical biochemical reactions. Second, we calculated the fraction of shared GO terms describing protein MF for all duplicates (Figure 4B). Both measures showed that the MF of yeast duplicates typically starts to substantially diverge only at about $Ka > 0.4$. The timing of this divergence approximately coincides with a significant increase in the fraction of essential duplicates (Figure 1C). On the other hand, the significant changes in the number of quantitative growth phenotypes are observed when the MF of duplicates is usually still conserved.

A complementary analysis of transcription factor binding sites suggests that gene regulation plays an important role in establishing the observed compensation patterns. It was previously demonstrated that duplicated yeast genes have, on average, a higher number of cis-regulatory motifs than singleton genes (46). Using a comprehensive data set of ~150 known and predicted DNA binding motifs in yeast (28,39), we found that the average number of different motifs regulating a duplicate pair increases significantly at $Ka \approx 0.1$ (Figure 4D, dashed line, Mann–Whitney U test, $P = 0.06$). At this divergence distance, the average number of different motifs per duplicate pair is more than twice the number of motifs for a pair of yeast singletons (Figure 4D, dashed horizontal line). The number of regulatory motifs increases both for the

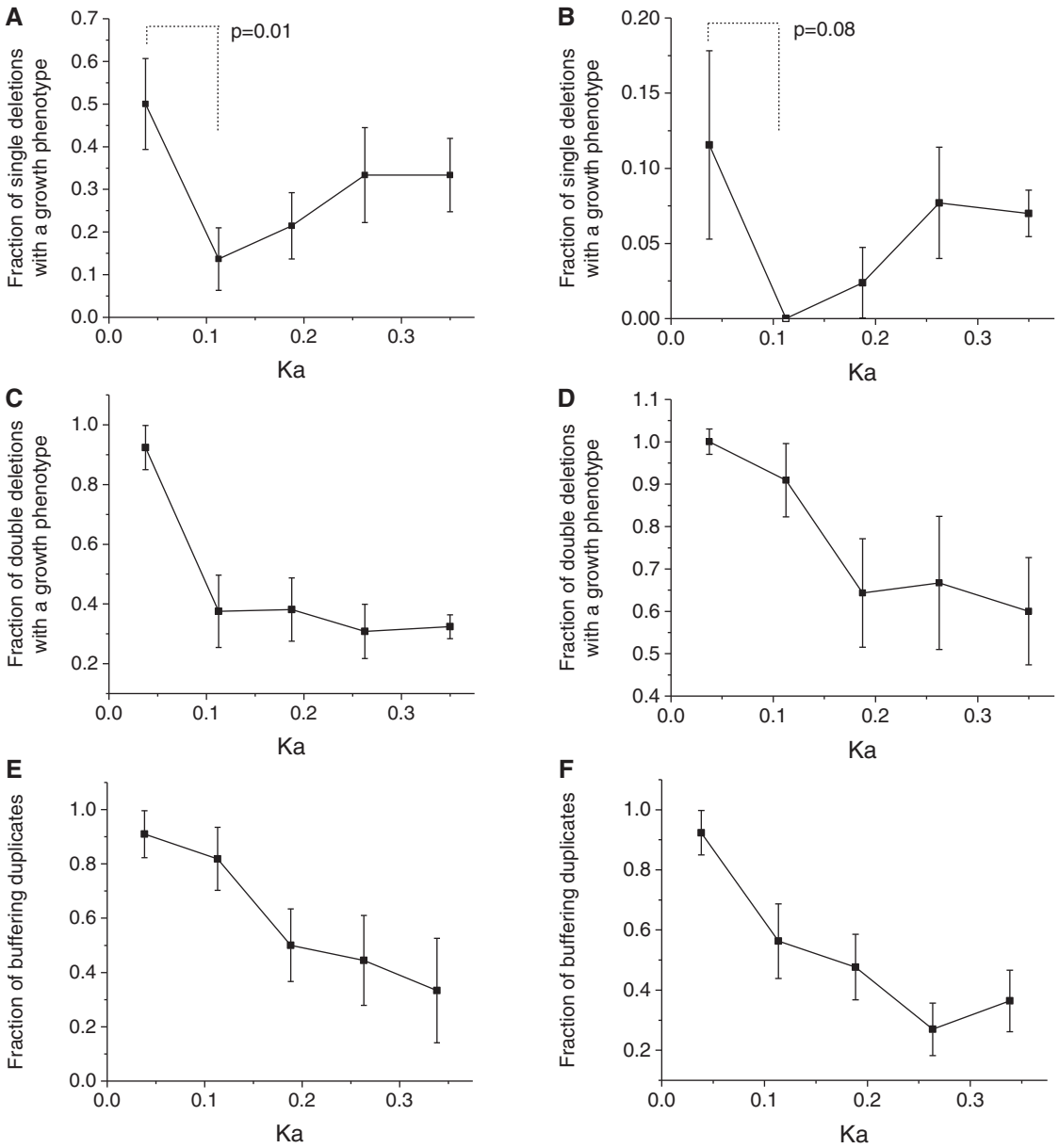

Figure 2. Growth phenotypes for individual and simultaneous deletion of duplicates as a function of their sequence divergence (Ka). The results in the first column (A, C, E) are based on the competition experiments by DeLuna *et al.* (34), and in the second column (B, D, F) on the synthetic genetic arrays (SGA) by Costanzo *et al.* (35). (**A**, **B**) Fractions of single duplicate deletions with a significant growth decrease. (**C**, **D**) Fractions of simultaneous (double) duplicate deletions with a significant growth decrease. Due to different measurement sensitivities of the two studies, different cutoffs were used to determine a significant growth decrease: 1% for DeLuna *et al.* (A, C) and 10% for Costanzo *et al.* (B, D); the presented results are not sensitive to the exact cutoff values (see Supplementary Figure S7). *P*-values were obtained using Fisher's exact test. (**E**, **F**) Fraction of paralogs with a significant negative epistatic interaction from the studies of DeLuna *et al.* and Costanzo *et al.*, respectively. In the figures error bars represent the SEM.

duplicate with the highest and the duplicate with the smallest number of sensitive conditions (Supplementary Figure S8A and B). The increase in complexity of the duplicates regulation at Ka $\approx$ 0.1 is also confirmed by a significant increase (Mann–Whitney U test, $P = 1 \times 10^{-3}$) at these distances of the number of transcription factor mutants (47) affecting duplicate gene expression (Figure 4D, solid line).

While the total number of DNA motifs regulating duplicates initially increases with divergence, the fraction of shared motifs [Supplementary Figure S9A, see also (48)], the overlap in GO terms describing biological processes (Figure 4C) and the overlap in cellular localization observed in fluorescence-tagging experiments (40) decrease (Supplementary Figure S8B). Such a pattern suggests that the increase in regulatory complexity allows duplicates to specialize for different biological processes while mostly preserving common MFs. The ability of duplicates with partially diverged regulatory regions to compensate for each other through expression

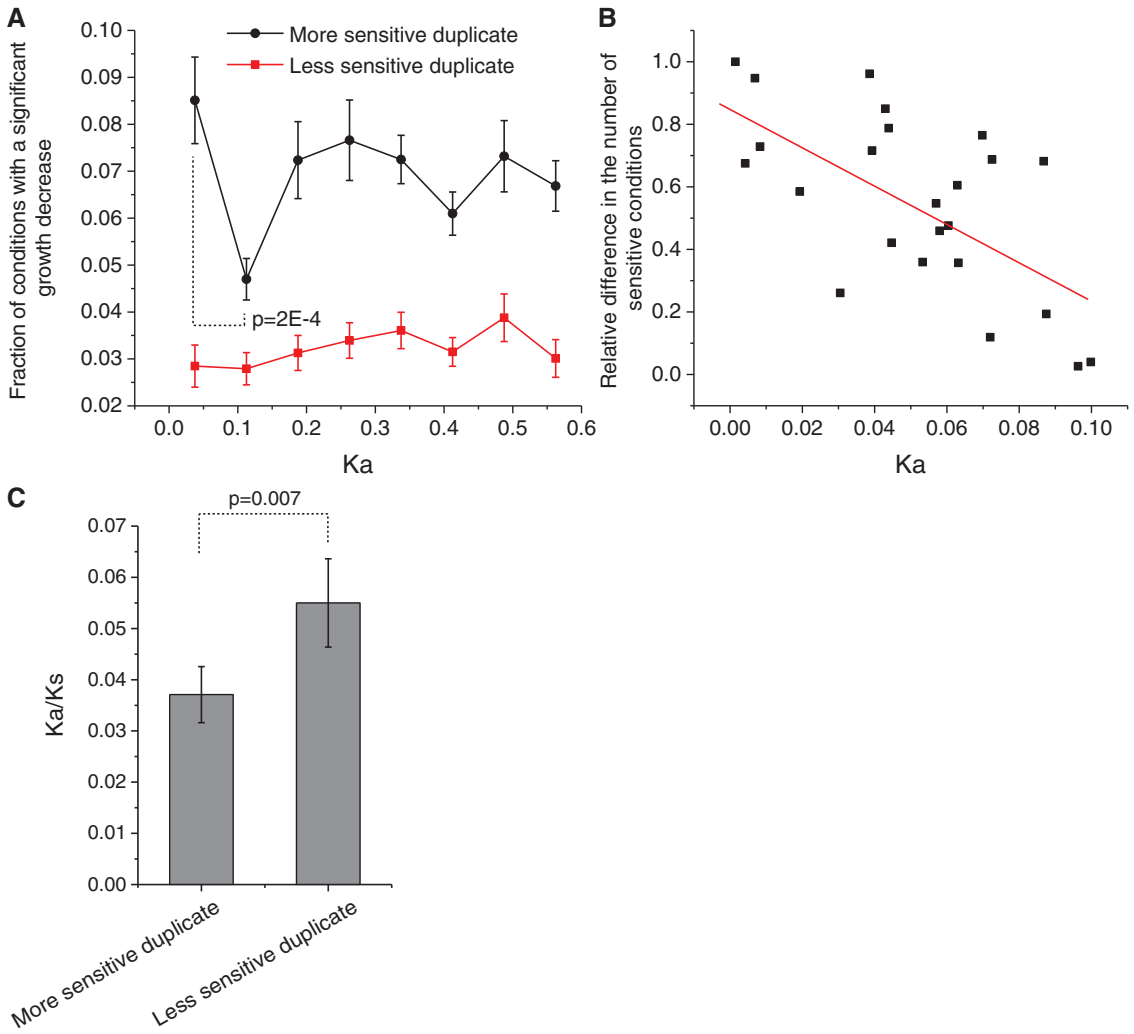

Figure 3. Differences in the number of sensitive conditions between duplicates. (**A**) The average fraction of sensitive conditions for the duplicates with the higher and lower number of sensitive conditions in each pair; Ka values represent sequence divergence between duplicates. The *P*-value is for the Mann–Whitney U test. (**B**) The relative difference in the number of sensitive conditions between duplicates as a function of their initial divergence; Ka values represent sequence divergence between duplicates. The relative difference was calculated as the absolute difference in the number of sensitive conditions between duplicates normalized to the total number of sensitive conditions for the pair (Spearman's $r = -0.60$, $P = 2 \times 10^{-3}$; Pearson's $r = -0.64$, $P = 7 \times 10^{-4}$). (**C**) The average Ka/Ks ratio for the paralogs with the largest (more sensitive) and smallest (less sensitive) number of conditions with a significant growth decrease. Ka/Ks ratios were calculated relative to orthologous sequences in *S. bayanus*. Only duplicates with Ka < 0.15 to each other were considered. The *P*-value is for the Wilcoxon signed rank test.

changes of the intact gene was previously described by Kafri *et al.* (39,49). Also, the recent study by DeLuna *et al.* (50) showed that on deletion of one duplicate, expression changes of the remaining paralog are often need-based, i.e. they happen primarily when the corresponding function is required. Such regulatory backup circuits should, at least in some cases, enable functional compensation between homologs with different expression patterns in wild type. Notably, based on the data from recent study by Springer *et al.* (51), who measured the expression changes of yeast genes when one of two genomic copies was deleted in diploid cells, we observed a significant dosage response only for genes forming recently duplicated pairs (Ka < 0.15, Figure 4E). This suggests that genes with close duplicates are most responsive to dosage effects.

Finally, the patterns of diversification and functional compensation described above should correlate with the process of duplicate loss in evolution. We investigated the retention of yeast duplicates using the complete genomic sequences of seven species: *S. cerevisiae*, *S. paradoxus*, *S. bayanus*, *S. castelli*, *S. mikatae*, *S. kudriavzevii* and *S. kluyveri*. We calculated the number of remaining duplicates as a function of their sequence divergence (Figure 5, see also Supplementary Figure S10A for the corresponding relationships in individual yeast species). This analysis suggests that a relatively brief initial period of high duplicate loss (6) is followed by a long evolutionary period (Ka > 0.1) during which the average loss rate decreases >10-fold (red in Figure 5). Interestingly, the loss rate significantly decreases approximately at the divergence distance when duplicates become more similar in terms

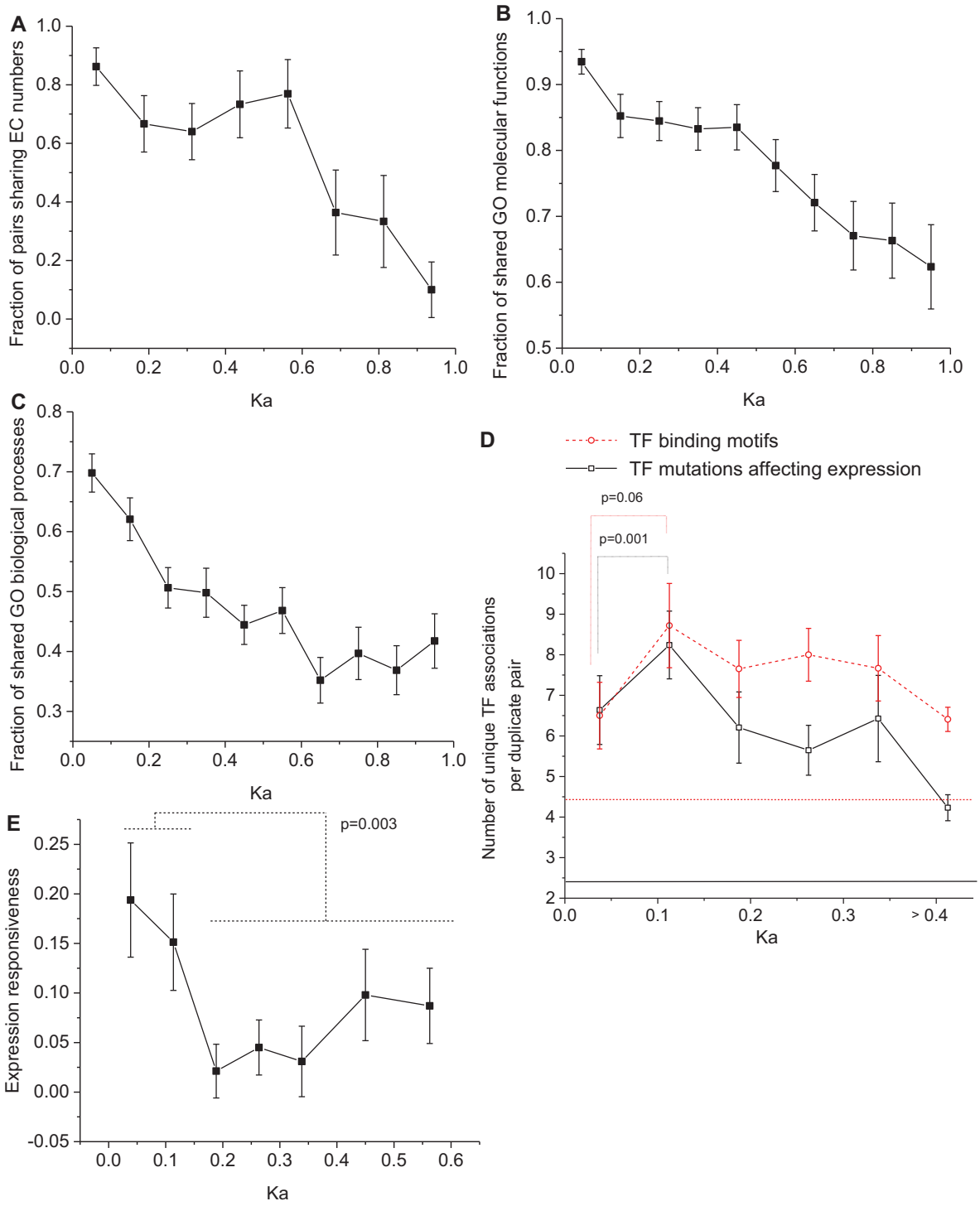

Figure 4. Diversification of duplicates function and regulation. (**A**) Fraction of metabolic duplicates sharing the same EC numbers; conservation of EC numbers indicates catalysis of identical biochemical reactions. (**B**) Fraction of GO MF terms shared between duplicates. (**C**) Fraction of GO Biological Process (BP) terms shared between duplicates. In panels B and C we considered only GO terms with a distance of three or more to the corresponding GO root hierarchy term. (**D**) Dashed line, the average number of different transcription factor binding motifs per duplicate pair. Transcription factor (TF) binding motifs were compiled from the studies of Kafri *et al.* (39) and Kellis *et al.* (28). Solid line, the average number of transcription factor deletions in *S. cerevisiae* that significantly affect the expression of duplicate genes. The data were obtained from the study by Hu *et al.* (47). For comparison we also show the average number of motifs and TF mutants affecting expression for random pairs of yeast singletons (horizontal dashed and solid lines); the *P*-values were calculated using the Mann–Whitney U test. (**E**) The average dosage compensation (responsiveness) of duplicates as a function of sequence divergence (Ka). The data for the average expression responsiveness was obtained from the work of Springer *et al.* (51). In that study, responsiveness was measured in diploid yeast strains as the Log_2 ratio (perturbed versus normal) of expression changes for the remaining gene copy following deletion of the equivalent gene copy on a sister chromosome. The *P*-value was calculated using the Mann–Whitney U test. In all figures error bars represent the SEM.

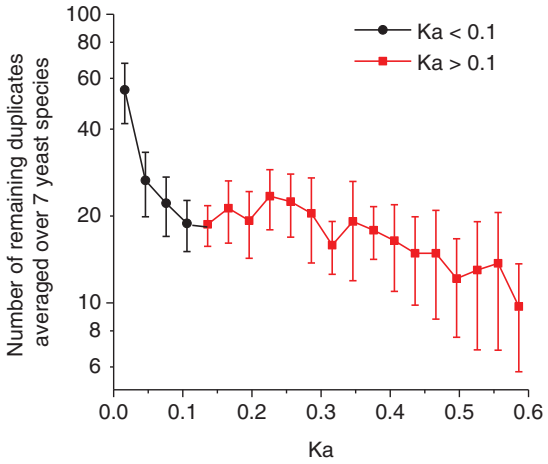

Figure 5. The average number of duplicates retained in the genomes of yeast species as a function of the duplicates divergence Ka, the number of nonsynonymous substitutions per site. The number of remaining duplicates was averaged over the genomes of seven yeast species: *S. cerevisiae*, *S. paradoxus*, *S. bayanus*, *S. castelli*, *S. mikatae*, *S. kudriavzevii* and *S. kluyveri*. See also Supplementary Figure S10 for the number of remaining duplicates in the individual species and for the number of remaining duplicates as a function of Ks. The rate of duplicate loss in evolution is >10 times lower for the distances corresponding to $Ka > 0.1$ compared with the distances at $Ka < 0.1$. In the figure, error bars represent the SEM.

of their functional load (Figure 2B) and when their regulatory complexity significantly increases (Figure 5D). It is likely that the duplicates surviving the initial loss stage develop independent functionalities and are preserved for long times in the genomes of yeast species.

DISCUSSION

In the present study, we analyzed genetic robustness due to duplicates in the context of quantitative growth phenotypes and sensitivities to gene deletions in multiple environmental conditions. Such robustness is important for understanding the buffering of deleterious mutations in large natural biological populations. Our results demonstrate that, contrary to commonly held view, close gene duplicates are unlikely to provide a high level of backup in the context of large natural populations. Consequently, it is unlikely that many duplicates are fixed in natural populations specifically due to selection for robustness.

Our analysis also suggests that duplicate redundancies described in genomics databases, and frequently observed in laboratory experiments, should be considered with caution, at least with respect to their functions in natural biological populations. To investigate this point further, we analyzed a set, compiled by Kafri *et al.* (52), of 112 yeast duplicates reported to be at least partially redundant in research publications. These duplicates have been described as redundant based on their functional overlap and compensatory interactions observed in small-scale experimental studies. Interestingly, based on the number of conditions with quantitative growth phenotypes from the study by Hillenmeyer *et al.* (33), and the quantitative growth measurements by Costanzo *et al.* (35), the duplicates annotated as redundant are not significantly different from all other yeast duplicates (Mann–Whitney U, $P = 0.13$ and 0.35, respectively, Supplementary Figure S11). This demonstrates that, although many yeast duplicates indeed may show functional overlap in some laboratory conditions, their compensation properties will probably be significantly less important in large natural populations due to the ability of purifying selection to efficiently prune mutations causing even a small fitness decrease.

It is likely that several different factors contribute to the relative paucity of functional compensation between paralogs at small divergence distances. A significant fraction of duplications are likely to be fixed owing to dosage effects (17), and functional compensation between such duplicates in the context of natural populations is unlikely. For example, the lack of significant compensation between histone pairs, HTA1-HTA2 and HHT1-HHT2, is likely to be a consequence of their role in maintaining proper histone levels in yeast cells. Gene dosage may explain the inability of some duplicates to backup each other, but it is unlikely to be the only explanation. We showed that even when all duplicate pairs with a high CAI (Supplementary Figure S6A) or pairs forming known protein complexes (Supplementary Figure S6B) are removed from the analysis, the patterns of functional compensation remain similar. Notably, genes with a high CAI have been also associated with higher frequencies of interlocus gene conversion (IGC) (53,54). While IGC can slow down the rate of duplicates sequence divergence (55), analyses based only on WGDs with no evidence of IGC [using data recently reported by Casola *et al.* (56)] revealed essentially identical compensation patterns (Supplementary Figure S12).

Close duplicates are also less likely to compensate for each other probably owing to the aforementioned dichotomy in their functional loads (Figure 3A and B). Many close duplicates can be classified, based on their activity and breadth of expression, into a major and a minor functional isoforms. For example, the glyceraldehyde-3-phosphate dehydrogenase TDH1 is active under various stress conditions, while its isoenzyme TDH2 is used primarily during exponential growth (57). Similarly, the ubiquitin conjugating enzyme UBC4 is expressed during exponential growth, while its duplicate UBC5 is active during stationary phase (58). The difference in functional load for close yeast duplicates is also consistent with the asymmetric partition of functions, interactions and gene expression, observed between close duplicates in other organisms, for example, *Arabidopsis* and Human (45,59,60). This suggests that duplicate-dependent compensation in the context of natural populations may be limited in other species as well.

Our analysis suggests that a typical lifecycle of gene duplicates in yeast consists of several distinct evolutionary stages (11,12). In the first stage (at duplicate distances corresponding to $Ka < 0.05$), duplicates tend to have high overall functional loads and significant asymmetry in the number of sensitive conditions; both of these factors make complete compensation unlikely. The high functional load of close duplicates suggests that adaptive selection plays

an important role in their fixation. In the second stage ($0.05 < Ka < 0.25$), as duplicates diverge further, their overall functional load usually decreases. This may happen, for example, due to relaxation of the environmental conditions, which facilitated the original duplicate fixation. The vast majority of duplicates, likely the paralogs with relatively smaller functional loads (Figure 3C), are lost at this stage (Figure 5). Gene pairs that survive the period of high duplicate loss display more balanced functional loads and complex regulation; these gene pairs are usually retained for long evolutionary times in yeast genomes (Figure 5). Surviving duplicates can provide at least partial compensation at intermediate divergence distances and also serve as an important source of new protein functions. In the third stage ($Ka > 0.3$ or ~70% sequence identity), the lifecycle of duplicates is completed when their functional roles diverge, and their quantitative compensation properties become indistinguishable from those of random pairs of yeast singletons.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Fyodor Kondrashov (Centre for Genomic Regulation) and members of the Vitup laboratory for illuminating discussions regarding the manuscript.

FUNDING

National Institutes of Health (NIH) [GM079759] and the National Centers for Biomedical Computing [U54CA121852]. Funding for open access charge: NIH [GM079759].

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Stelling,J., Sauer,U., Szallasi,Z., Doyle,F.J. III and Doyle,J. (2004) Robustness of cellular functions. *Cell*, **118**, 675–685.
3. Wagner,A. (2005) *Robustness and Evolvability in Living Systems*. Princeton University Press, Princeton.
4. Wagner,A. (2000) Robustness against mutations in genetic networks of yeast. *Nat. Genet.*, **24**, 355–361.
5. Papp,B., Pal,C. and Hurst,L.D. (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, **429**, 661–664.
6. Lynch,M. and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.
7. Wapinski,I., Pfeffer,A., Friedman,N. and Regev,A. (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature*, **449**, 54–61.
8. 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.
9. Hsiao,T.L. and Vitkup,D. (2008) Role of duplicate genes in robustness against deleterious human mutations. *PLoS Genet.*, **4**, e1000014.
10. Ohno,S. (1970) *Evolution by Gene Duplication*. Springer-Verlag, Berlin, New York.
11. Conant,G.C. and Wolfe,K.H. (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.*, **9**, 938–950.
12. Innan,H. and Kondrashov,F. (2010) The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.*, **11**, 97–108.
13. Nadeau,J.H. and Sankoff,D. (1997) Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics*, **147**, 1259–1266.
14. Sidow,A. (1996) Genome duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.*, **6**, 715–722.
15. 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.
16. Lynch,M. and Force,A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**, 459–473.
17. Kondrashov,F.A., Rogozin,I.B., Wolf,Y.I. and Koonin,E.V. (2002) Selection in the evolution of gene duplications. *Genome Biol.*, **3**, RESEARCH0008.
18. Bergthorsson,U., Andersson,D.I. and Roth,J.R. (2007) Ohno's dilemma: evolution of new genes under continuous selection. *Proc. Natl Acad. Sci. USA*, **104**, 17004–17009.
19. VanderSluis,B., Bellay,J., Musso,G., Costanzo,M., Papp,B., Vizeacoumar,F.J., Baryshnikova,A., Andrews,B., Boone,C. and Myers,C.L. (2010) Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Mol. Syst. Biol.*, **6**, 429.
20. Ihmels,J., Collins,S.R., Schuldiner,M., Krogan,N.J. and Weissman,J.S. (2007) Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. *Mol. Syst. Biol.*, **3**, 86.
21. Li,J., Yuan,Z. and Zhang,Z. (2010) The cellular robustness by genetic redundancy in budding yeast. *PLoS Genet.*, **6**, e1001187.
22. Guan,Y., Dunham,M.J. and Troyanskaya,O.G. (2007) Functional analysis of gene duplications in *Saccharomyces cerevisiae*. *Genetics*, **175**, 933–943.
23. Conant,G.C. and Wagner,A. (2004) Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. *Proc. Biol. Sci.*, **271**, 89–96.
24. Thatcher,J.W., Shaw,J.M. and Dickinson,W.J. (1998) Marginal fitness contributions of nonessential genes in yeast. *Proc. Natl Acad. Sci. USA*, **95**, 253–257.
25. Gillespie,J.H. (1998) *Population Genetics, A Concise Guide*. Johns Hopkins Univ. Press, Baltimore.
26. Hartl,D. and Clark,A. (1997) *Principles of Population Genetics*, 3 edn. Sinauer Associates, Sunderland.
27. Lynch,M. (2006) Streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.*, **60**, 327–349.
28. Kellis,M., Patterson,N., Endrizzi,M., Birren,B. and Lander,E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.
29. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
30. Gu,Z., Cavalcanti,A., Chen,F.C., Bouman,P. and Li,W.H. (2002) Extent of gene duplication in the genomes of Drosophila, nematode, and yeast. *Mol. Biol. Evol.*, **19**, 256–262.
31. Yang,Z. and Nielsen,R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, **17**, 32–43.
32. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
33. Hillenmeyer,M.E., Fung,E., Wildenhain,J., Pierce,S.E., Hoon,S., Lee,W., Proctor,M., St Onge,R.P., Tyers,M., Koller,D. *et al.* (2008) The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science*, **320**, 362–365.
34. DeLuna,A., Vetsigian,K., Shoresh,N., Hegreness,M., Colon-Gonzalez,M., Chao,S. and Kishony,R. (2008) Exposing the fitness contribution of duplicated genes. *Nat. Genet.*, **40**, 676–681.

35. Costanzo,M., Baryshnikova,A., Bellay,J., Kim,Y., Spear,E.D., Sevier,C.S., Ding,H., Koh,J.L., Toufighi,K., Mostafavi,S. *et al.* (2010) The genetic landscape of a cell. *Science*, **327**, 425–431.
36. Giaever,G., Chu,A.M., Ni,L., Connelly,C., Riles,L., Veronneau,S., Dow,S., Lucau-Danila,A., Anderson,K., Andre,B. *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature*, **418**, 387–391.
37. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
38. Guldener,U., Munsterkotter,M., Kastenmuller,G., Strack,N., van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., Garcia-Martinez,J., Perez-Ortin,J.E. *et al.* (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.*, **33**, D364–D368.
39. Kafri,R., Bar-Even,A. and Pilpel,Y. (2005) Transcription control reprogramming in genetic backup circuits. *Nat. Genet.*, **37**, 295–299.
40. Huh,W.K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
41. Lu,P., Vogel,C., Wang,R., Yao,X. and Marcotte,E.M. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.*, **25**, 117–124.
42. Gimeno,C.J. and Fink,G.R. (1992) The logic of cell division in the life cycle of yeast. *Science*, **257**, 626.
43. Pal,C., Papp,B. and Hurst,L.D. (2001) Highly expressed genes in yeast evolve slowly. *Genetics*, **158**, 927–931.
44. Conant,G.C. and Wolfe,K.H. (2006) Functional partitioning of yeast co-expression networks after genome duplication. *PLoS Biol.*, **4**, e109.
45. Wagner,A. (2002) Asymmetric functional divergence of duplicate genes in yeast. *Mol. Biol. Evol.*, **19**, 1760–1768.
46. He,X. and Zhang,J. (2005) Gene complexity and gene duplicability. *Curr. Biol.*, **15**, 1016–1021.
47. Hu,Z., Killion,P.J. and Iyer,V.R. (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet.*, **39**, 683–687.
48. Papp,B., Pal,C. and Hurst,L.D. (2003) Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet.*, **19**, 417–422.
49. Kafri,R., Levy,M. and Pilpel,Y. (2006) The regulatory utilization of genetic redundancy through responsive backup circuits. *Proc. Natl Acad. Sci. USA*, **103**, 11653–11658.
50. DeLuna,A., Springer,M., Kirschner,M.W. and Kishony,R. (2010) Need-based up-regulation of protein levels in response to deletion of their duplicate genes. *PLoS Biol.*, **8**, e1000347.
51. Springer,M., Weissman,J.S. and Kirschner,M.W. (2010) A general lack of compensation for gene dosage in yeast. *Mol. Syst. Biol.*, **6**, 368.
52. Kafri,R., Dahan,O., Levy,J. and Pilpel,Y. (2008) Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. *Proc. Natl Acad. Sci. USA*, **105**, 1243–1248.
53. Lin,Y.S., Byrnes,J.K., Hwang,J.K. and Li,W.H. (2006) Codon-usage bias versus gene conversion in the evolution of yeast duplicate genes. *Proc. Natl Acad. Sci. USA*, **103**, 14412–14416.
54. Pyne,S., Skiena,S. and Futcher,B. (2005) Copy correction and concerted evolution in the conservation of yeast genes. *Genetics*, **170**, 1501–1513.
55. Gao,L.Z. and Innan,H. (2004) Very low gene duplication rate in the yeast genome. *Science*, **306**, 1367–1370.
56. Casola,C., Conant,G.C. and Hahn,M.W. (2012) Very low rate of gene conversion in the yeast genome. *Mol. Biol. Evol.*, **29**, 3817–3826.
57. Boucherie,H., Bataille,N., Fitch,I.T., Perrot,M. and Tuite,M.F. (1995) Differential synthesis of glyceraldehyde-3-phosphate dehydrogenase polypeptides in stressed yeast cells. *FEMS Microbiol. Lett.*, **125**, 127–133.
58. Seufert,W. and Jentsch,S. (1990) Ubiquitin-conjugating enzymes UBC4 and UBC5 mediate selective degradation of short-lived and abnormal proteins. *EMBO J.*, **9**, 543–550.
59. Zou,C., Lehti-Shiu,M.D., Thomashow,M. and Shiu,S.H. (2009) Evolution of stress-regulated gene expression in duplicate genes of Arabidopsis thaliana. *PLoS Genet.*, **5**, e1000581.
60. Chung,W.Y., Albert,R., Albert,I., Nekrutenko,A. and Makova,K.D. (2006) Rapid and asymmetric divergence of duplicate genes in the human gene coexpression network. *BMC Bioinformatics*, **7**, 46.