Asymmetric Functional Divergence of Duplicate Genes in Yeast

Andreas Wagner

Department of Biology, University of New Mexico

Most duplicate genes are eliminated from a genome shortly after duplication, but those that remain are an important source of biochemical diversity. Here, I present evidence from genome-scale protein-protein interaction data, microarray expression data, and large-scale gene knockout data that this diversification is often asymmetrical: one duplicate usually shows significantly more molecular or genetic interactions than the other. I propose a model that can explain this divergence pattern if asymmetrically diverging duplicate gene pairs show increased robustness to deleterious mutations.

Introduction

Soon after a gene duplication, degenerative mutations are likely to eliminate duplicate genes from the genome (Li 1997, pp. 284–287; Lynch and Conery 2000). But gene duplications occur continuously and at high rates in eukaryotes, which accounts for the fact that up to 50% of a eukaryotic genome may consist of duplicate genes (Lynch and Conery 2000; Rubin et al. 2000). These persisting duplicate genes are perhaps the most prominent source of biochemical innovation of gene products. But little is known about how this innovation occurs or about how gene duplicates diverge in general.

Studying functional divergence among duplicate genes requires a definition of gene function, but no such universal definition is possible. The reason is that there are several complementary ways of classifying gene functions (Ashburner et al. 2000). For instance, gene products can be characterized biochemically, e.g., as enzymes or transcription factors. Second, they can be characterized through their time and locus of expression, e.g., expression during a cell cycle stage, in the cytoplasm, or during brain development. Third, they can be characterized genetically through mutations and through other genes that these mutations affect. This list is not necessarily complete.

Functional genomics has added much information to each of these categories, especially in model organisms like the yeast *Saccharomyces cerevisiae*. First, monitoring expression through microarrays (Chu et al. 1998; Eisen et al. 1998; Spellman et al. 1998; Gasch et al. 2000) provides spatiotemporal expression information for thousands of genes at once. This information is indicative of the biological process a gene is involved in. Second, genome-wide protein-protein interactions can characterize physical interactions among thousands of gene products (Bartel et al. 1996; Fromont-Racine, Rain, and Legrain 1997; Uetz et al. 2000; Ito et al. 2001). Third, large-scale gene knockout screens in combination with microarray experiments indicate which genes' expression level is affected by a mutated gene (Hughes et al. 2000). Thus, even in the absence of a detectable phenotype—all too frequent in knockout experiments—a putative function can sometimes be assigned using genetic interactions with known genes.

Attempts to identify gene functions according to any of the above criteria, whether they use genomic or pregenomic techniques, yield one key message: most genes have more than one, if not many functions. They are expressed at multiple times and in multiple places, they affect multiple biological processes when mutated, or they interact with proteins with diverse biochemical and biological roles (Bender et al. 1983; Li and Noll 1994; Jack and Delotto 1995; Slusarski, Motzny, and Holmgren 1995; Kirchhamer, Yuh, and Davidson 1996; Schwikowski, Uetz, and Fields 2000; Wagner 2001). This multifunctionality has important implications for the divergence of duplicate genes: duplicate genes often diverge through loss of complementary (sub)functions in each duplicate (Force, Lynch, and Postlethwait 1999; Lynch and Force 2000; Wagner 2000). Examples abound. To name but two, the *ZAG1* and *ZMM2* genes are paralogues in the maize genome. They are orthologues of the Arabidopsis *AGAMOUS* gene, which is involved in carpel and stamen development. Each of them appears to have largely lost one of their ancestral expression domains: *ZAG1* is expressed at high levels in developing carpels and *ZMM2* is expressed in developing stamens. A null mutation in *ZAG1* affects only early carpel development (Coen and Meyerowitz 1991; Schmidt et al. 1993; Mena et al. 1996). Force, Lynch, and Postlethwait (1999) report on the zebrafish engrailed genes *eng1* and *eng1b*, the likely results of a teleost-specific gene duplication of the tetrapod *En1* gene. In mice and chicken, *En1* is expressed in the developing pectoral appendage bud and in specific neurons of the developing hindbrain and spinal cord. In zebrafish, *eng1* retained expression in the pectoral appendage bud, whereas *eng1b* is only expressed in the hindbrain and the spinal cord. Similar patterns of divergence may be quite common in zebrafish (Ekker et al. 1995; Lee, Xu, and Breitbart 1996; Ekker et al. 1997).

Studies focussing on individual gene pairs fall short of identifying general divergence patterns of many duplicate genes. At first sight, analyzing functional divergence of many duplicate genes may seem like a hopeless task. Because it is not even straightforward to classify one gene's function, how would one compare the functions of many divergent duplicates? Functional genomic

Key words: protein interaction networks, microarrays, gene knockout, biochemical innovation.

Address for correspondence and reprints: Andreas Wagner, Department of Biology, University of New Mexico, 167A Castetter Hall, Albuquerque, New Mexico 817131-1091. E-mail: wagnera@unm.edu.

Mol. Biol. Evol. 19(10):1760–1768. 2002
 ISSN: 0737-4038

experiments provide a crude remedy for this problem. Despite their disadvantage of providing largely qualitative information about genetic and molecular interactions of genes, their great advantage is that they do so for thousands of genes at once. They thus yield insight about one aspect—however minute—of gene function, such as the protein interaction partners of a gene, gene expression patterns affected through mutating a gene, or the response of gene expression to environmental challenges. It is this aspect of gene function I will focus on.

Methods

Gene Duplication Data

Data on yeast gene duplicates were kindly provided by John Conery (Department of Computer Science, University of Oregon) and generated as described in Lynch and Conery (2000). Briefly, gapped BLAST (Altschul et al. 1997) was used for pairwise amino acid sequence comparisons of all yeast open reading frames as obtained from GenBank. All protein pairs with a BLAST alignment score greater than 10^{-2} were retained for further analysis. Then, the following conservative approach was followed to retain only unambiguously aligned sequences. Using the protein alignment generated by BLAST as a guide, a sequence pair was scanned to the right of each alignment gap. All sequences from the end of the gap through the first ''anchor'' pair of matched amino acids were discarded. All subsequent sequences (exclusive to the anchor pair of amino acids) were retained if a second pair of matching amino acids was found within less than six amino acids from the first. This procedure was then repeated to the left of each alignment gap (see Lynch and Conery [2000] for more detailed description and justification). The retained portion of each amino acid sequence alignment was then used jointly with DNA sequence information to generate nucleotide sequence alignments of genes. For each gene pair in this data set, the fraction K_s of synonymous (silent) substitutions per silent site as well as the fraction K_a of replacement substitutions per replacement site were estimated using the method of Li (1993).

Protein Interaction Data and Analysis

Data for 899 pairwise interactions among 985 yeast proteins, as reported in Uetz et al. (2000), were obtained from http://depts.washington.edu/sfields/projects/YPLM/Nature-plain.html on February 15, 2000. There are 43 proteins that have been reported to interact among themselves. Before further analysis all such self-interactions were eliminated. (Self-interactions are interactions between two protein products of the same gene, such as interactions that might occur for homodimerizing proteins.) The resulting protein interaction network was then represented as a graph using the Library of Efficient Data types and Algorithms (LEDA) (Mehlhorn and Naher 1999). Within this graph representation, common and different protein interactions among gene family members are easily analyzed (Wagner 2001). To analyze protein interaction data not generated by two-hybrid experiments, I used information on physical interactions among yeast proteins obtained from the Munich Institute for Protein Sequences (MIPS) database (Mewes et al. 1999, http://mips.gsf.de/proj/yeast/CYGD/db/index.html). I eliminated from these data all protein interactions generated only by two-hybrid experiments. The remaining 899 interactions involve 680 proteins. I did not distinguish between genes with only one paralogue and genes that occur in multigene families in the analysis of either data set.

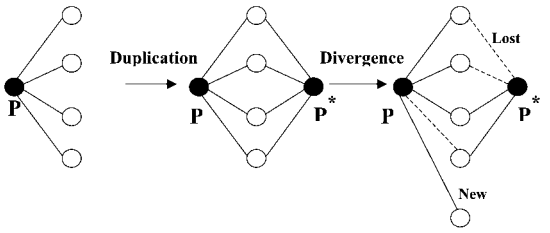

FIG. 1.—Asymmetric divergence in protein-protein interactions. Circles stand for proteins, and lines for interactions among proteins. Shortly after a gene duplication, the products P and P* of a duplicate gene that are part of a protein interaction network interact with the same proteins. Eventually, some or all common interactions may be lost (dashed lines), and new interactions may be gained by either protein.

I used the following numerical approach to test (and reject) the null hypothesis that the number of interactions in products of paralogous genes has diverged symmetrically. (Notice that this hypothesis does not regard the mechanism of divergence, only its pattern.) The approach proceeds by (1) reconstructing the (identical) numbers of interactions of two proteins immediately after duplication of their encoding genes, and (2) emulating the process of symmetric divergence. Consider two proteins P and P* that have d_1 and d_2 protein interactions, respectively, and that share b of these interaction partners (fig. 1). It follows that P and P* have $d_1 - b$ and $d_2 - b$ nonshared interactions, respectively, adding to a total of $d_1 + d_2 - 2b$ nonshared interactions. Each of these interactions might have arisen through the evolutionary loss of an interaction that was shared after duplication or through the evolutionary gain of an interaction since the duplication. To not restrict myself to only one of these possibilities, I assume that after duplication interactions are lost with some probability P_l and gained with probability $(1 - P_l)$. Because interactions are gained or lost probabilistically, one cannot unambiguously reconstruct the ancestral state of interactions, that is, the number of interactions P and P* had immediately after duplication. But it is possible to reconstruct a likely ancestral state simply by noting that the number of lost interactions after duplication follows a binomial distribution $B(d_1 + d_2 - 2b, P_l)$. This ancestral state, the number of interactions of each protein immediately after duplication, is simply given by $b + n_l$, where n_l is a random number distributed as $B(d_1 + d_2 - 2b, P_l)$. (The total number of interactions gained by the two duplicates then immediately follows as $n_g = [d_1 + d_2 - 2b] - n_l$.) Equipped with these two numbers, I then applied the null hypothesis of symmetric divergence to emulate each protein's divergence from this ancestral state. According to the null hypothesis, the number of interactions lost and gained by protein P

since duplication is given by random numbers n_{l1} with distribution B(n_l, 0.5) and n_{g1} with distribution B(n_g, 0.5), respectively. The factor 0.5 in these distributions reflects the assumption of symmetric divergence in the null hypothesis. Thus, according to the null hypothesis, protein P should have $(b + n_l) - n_{l1} + n_{g1}$ interactions. The number of interactions of protein P* immediately follows as $(b + n_l) - (n_l - n_{l1}) + (n_g - n_{g1})$.

I numerically applied this approach, which I have explained for only one protein pair, to all protein pairs considered here. In this way, I generated a distribution of the number of interactions under the null hypothesis of symmetric divergence with a given probability P_l of interaction loss. I then asked whether the statistical association between the number of interactions is the same in the null model as in the empirical data. The answer was unequivocally no, regardless of the value of P_l used. But only three special cases are treated in the main text: (1) divergence through loss of interactions only ($P_l = 1$); (2) divergence through gain of interactions ($P_l = 0$); and (3) divergence through equiprobable loss and gain of interactions ($P_l = 0.5$).

Environmental Stress and Gene Expression

To assay the differential expression response of yeast paralogues to environmental stresses, I used data provided by Gasch et al. (2000) for the following conditions: heat shock (25–37°C, after 30 min), reverse heat shock (37–25°C, 30′), H_2O_2 and Menadione exposure, both of which generate reactive oxygen species (60′ and 80′, respectively), dithiothreitol, a reducing agent interfering with protein folding (90′), diamide, an agent oxidizing sulfhydryl groups, (40′), hyperosmotic shock mediated by 1 M sorbitol (60′), hypo-osmotic shock mediated by transfer of cells from 1 M sorbitol to medium lacking sorbitol (30′), amino acid starvation (2 h), nitrogen depletion (1 day) , and stationary phase (7 days). I considered genes whose expression level was changed at least threefold relative in response to a stressor to be affected significantly. Because the expression response to most environmental stresses is transient, I chose a time point (indicated above in parentheses) approximately halfway through the measured response time series for each environmental stress to assess significant change. I then counted the number of stressors to which each member of a paralogous gene pair responded and did so for all 5,460 duplicate pairs with $K_a < 0.75$. For 40.4% (2,210) of these gene pairs, neither gene in the pair showed a response to any of the stressors applied. Such gene pairs are not suitable for this analysis, and I have thus eliminated them. I also excluded 162 further gene pairs (2.96%), where at least one stress condition induced the expression of one gene but repressed that of the other. Because of cross-hybridization, very closely related duplicates cannot be distinguished through microarray analysis, but the analysis of Gasch et al. (2000, fig. 5) suggests that gene pairs with $K_s > 0.5$ are readily distinguishable. I thus excluded an additional 4.5% (247) of the paralogues with $K_s < 0.5$ from the analysis. The null hypothesis of symmetric divergence was assessed in exactly the same way as that for protein-protein interactions, except that d_1 and d_2 now do not correspond to the number of protein interactions but instead to the number of expression responses that two duplicate genes show when exposed to the environmental stressors considered here (b is the number of environmental stressors to which both duplicate genes respond).

Gene Perturbations and Gene Expression

Data summarizing the effects of 271 gene deletions (and other treatments) on gene expression were made available as supplemental material to Hughes et al. (2000), file data_expts_1-300_ratios.txt. From this data set, which contains $\log_{10}$-transformed expression ratios of 6,312 genes for each mutation, I eliminated all data derived from haploid and aneuploid deletion strains, as well as data on nongenetic treatments. The remaining data contain information on null mutation effects for a total of 21 paralogous gene pairs, the most closely related 11 of which ($K_a < 1$) are discussed here. For each member gene of each paralogue, I determined what other genes were affected in their expression level by a synthetic-null mutation in the gene. I also determined the number of genes that were affected by a null mutation in each paralogue. I considered a gene as affected by a null mutation if its level of mRNA expression had changed by more than threefold in response to the mutation.

The total number of gene pairs to test for symmetric divergence is much smaller than that available for protein interactions and environmental stress response, but the number of affected genes per null mutation is much larger. This means not only that the above test must be modified but also that it is now possible to test each individual gene pair for symmetric divergence. I present an exact test only for the two extreme cases of loss and gain of function after duplication. Let d_1 and d_2 be the number of genes affected by a synthetic null mutation in genes 1 and 2, respectively, of a paralogous pair. Let b be the number of genes affected by both mutations. Under the null hypothesis of symmetric (equiprobable) loss-of-effects on other genes, a null mutation in either duplicate would have affected $d_1 + d_2 - b$ other genes immediately after duplication. $l_1 = d_2 - b$ and $l_2 = d_1 - b$ of these effects were subsequently lost in genes 1 and 2, respectively, adding to a total of $d_1 + d_2 - 2b$ lost effects. A disparity between l_1 and l_2 indicates asymmetry in divergence. The probability P of a disparity as big as or bigger than that actually observed, by chance alone, is calculated by summing over the tails of a binomial distribution B($d_1 + d_2 - 2b$, ½), so

$$P = 2 \sum_{0}^{\min(l_1,l_2)} \binom{d_1 + d_2 - 2b}{1/2} \left(\frac{1}{2}\right)^{d_1+d_2-b} \qquad l_1 \neq l_2,$$

and $P = 1$ for $l_1 = l_2$. The factor 2 in front of the summation sign indicates that this is a two-tailed test. The second null hypothesis, that of symmetric gain-of-effects by the two duplicates, is tested in the same way. The only difference is that $\min(l_1, l_2)$ above is replaced

Table 1
Differential Effects of Null Mutations in Paralogous Genes on Yeast Gene Expression

Gene 1/Gene 2	K_a	Number of Effects		Symmetric Divergence[c]
		1/2[a]	Common[b]	
MBP1/SW14	**0.46**	**5/149**	**1**	**7.69×10^{-39}**
ERP2/ERP4	0.42	6/1	0	0.13
CLB6/CLB2	0.52	17/17	0	1
ISW1/ISW2	0.45	20/13	3	0.1
YER041W/RAD27	**0.85**	**0/14**	**0**	**1.2×10^{-4}**
YHR022C/VPS21	**0.87**	**12/3**	**0**	**0.035**
PPR1/CAT8	**0.78**	**14/0**	**0**	**1.2×10^{-4}**
SIR2/HST3	**0.75**	**16/97**	**2**	**1.9×10^{-16}**
PAU2/YOR009W	**0.75**	**1/10**	**0**	**0.011**
ALD5/YHR039C	0.78	6/4	0	0.75
DIG2/DIG1	**0.65**	**9/23**	**2**	**5.2×10^{-3}**

[a] The number of genes whose expression is affected by a null mutation in genes 1 and 2, respectively.

[b] The number of genes affected by a null mutation in either gene.

[c] *P* values in this column indicate the probability that a difference in the number of affected genes equal or greater than the observed difference is due to equiprobable loss or gain of effects in the duplicates. Rows in bold type have $P < 0.05$. See *Methods* for functional annotations of gene names in the table.

with $\min(g_1, g_2)$, where $g_1 = d_1 - b$ and $g_2 = d_2 - b$ is the number of effects gained. This shows that the *P* values of the two scenarios are identical, suggesting that a mixed model of gain and loss would yield qualitatively identical results.

The following are brief annotations (Mewes et al. 1999) of all genes listed in table 1 (in order of appearance), with the exception of genes with seven-letter names, which correspond to genes of completely uncharacterized functions—MBP1: subunit of the MBF transcription factor; SWI4: transcription factor; ERP2: p24 protein involved in membrane trafficking; ERP4: similarity to human COP-coated vesicle membrane protein; CLB6: B-type cyclin; CLB2: G2/M-specific cyclin; ISW1 and ISW2: strong similarities to Drosophila ISW1 gene; RAD27: ssDNA endonuclease and 5′-3′exonuclease; VPS21: GTP-binding protein; CAT8: transcription factor involved in gluconeogenesis; SIR2: silencing regulatory protein and DNA-repair protein; HST3: silencing protein; PAU2: strong similarity to members of the Srp1p/Tip1p family; ALD5: aldehyde dehydrogenase 2 (NAD+); DIG1 and DIG2: MAP kinase–associated proteins, down-regulator of invasive growth and mating. Further information on the genes affected by a particular perturbation is available at http://www.rosettainpharmatics.com/publications/cell_hughes.htm as well as at the Munich Information Center for Protein Sequences (http://mips.gsf.de/proj/yeast/).

Results

Asymmetric Divergence in Protein-Protein Interactions

Genome-scale screens of protein interactions using the yeast two-hybrid assay have been carried out in several organisms (Bartel et al. 1996; Fromont-Racine, Rain, and Legrain 1997; Ito et al. 2000; Uetz et al. 2000). Their results are comprehensive maps of protein-protein interactions comprising many proteins encoded by a genome. Interpreting these maps is still difficult because they may contain significant numbers of false-positive and false-negative interactions (Ito et al. 2001) and because they collapse the spatial and temporal dimensions of gene expression into a still-life image of protein interactions. But these maps have also demonstrated their usefulness in predicting the spatial expression domain and functional annotation of many proteins from their interaction partners (Schwikowski, Uetz, and Fields 2000). They can also answer questions about global patterns of interactions, questions whose answers do not depend on the veracity of each individual interaction but only on statistical interaction patterns.

More than 30% of yeast genes whose products interact with proteins have one or more gene duplicates in the yeast genome (Wagner 2001). How do gene duplications influence the structure of the protein interaction network? Figure 1 shows a hypothetical protein P that interacts with four other proteins. Immediately after duplication of the gene encoding P, P and its duplicate P* share all four interactions. As the duplicates diverge in sequence, they also diverge in their protein interactions. Each protein may occasionally gain new interactions. But if mutations are more likely to cause loss of an interaction, as suggested by the prevalence of degenerative mutations in general (Li 1997), then most divergences will be due to loss of originally common protein interactions. Here, I use the number of interaction partners a protein has as a crude one-dimensional indicator of protein function. The number of common and different interactions between two duplicates then indicates their functional divergence.

Figure 2*a* shows the number of interaction partners for 1,734 pairs of paralogous genes in the network described by Uetz et al. (2000). These comprise all paralogous pairs with $K_a < 1$ nonsynonymous substitutions per nonsynonymous site, corresponding to genes with less than 60% amino acid divergence. The abscissa and ordinate axes show the number of protein interactions for the first and second protein member of each pair. The number of common interactions in these pairs is small: even among the most recent paralogues (synonymous substitutions per synonymous site $K_s < 0.5$) less than 60% share any interactions at all, and this number dwindles to less than 15% for more distant paralogues ($K_s > 1$) (Wagner 2001).

Figure 2*a* shows a distinct L-shape, indicating that in many protein pairs, where one partner has many interactions, the other one has disproportionately few. This negative correlation in the number of interaction partners between duplicates is statistically highly significant (Spearman $r_s = -0.58$, $P \ll 10^{-3}$; Pearson $r = -0.15$, $P \ll 10^{-3}$, df = 1,732). Could it have occurred by chance alone, that is, through random symmetric (equiprobable) loss or gain of interactions in either member of a pair? To find out, I numerically tested this null hypothesis of symmetric divergence as described in *Methods.* I did so under multiple scenarios distinguished by the relative importance given to evolutionary gains and losses of protein interactions after gene duplications. More specifically, each scenario assumes that the prob-

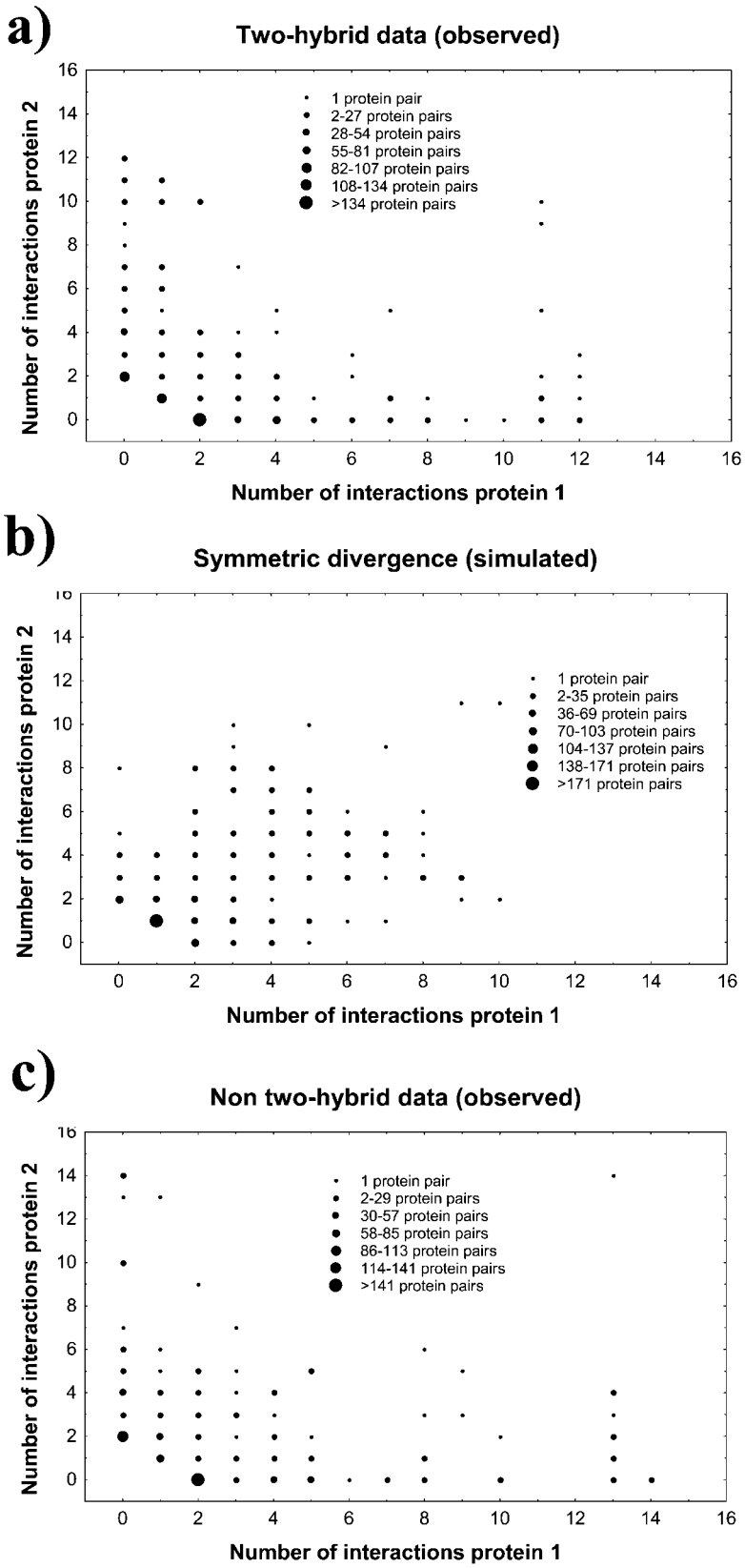

FIG. 2.—Asymmetric divergence in protein-protein interactions: the number of interaction partners of protein 1 versus protein 2 plotted for the two protein products of 1,734 paralogous yeast gene pairs with $K_a < 0.75$. Values of zero on either axis indicate that one member of a paralogous pair is not part of the protein interaction network. It may have lost all protein interactions (but may have retained biological functions not mediated through protein interactions). Pairs where one member has no protein interactions and the other member has only one interaction have been excluded from the plot. The thickness of each dot indicates the number of protein pairs with the number of interactions shown in the inset. (*a*) Two-hybrid data as reported in Uetz et al. (2000). (*b*) Simulated data. The expected distribution of interactions for the same 1,734

abilities of interaction loss and gain are equal to some probability P_l and $1 - P_l$. I report results only for three representative scenarios, although others yielded qualitatively identical results. The first scenario assumes that all divergence of interactions in duplicate genes is due to loss of interactions since the duplication. ($P_l = 1$). For illustration, figure 2*b* shows the distribution of interactions expected under this scenario, as generated from a stochastic simulation of the divergence of 1,734 pairs of paralogous genes. The L-shape of the plot in figure 2*a* disappears in this scenario of symmetric divergence, as does the highly negative statistical association (Spearman $r_s = -0.08$, $P \ll 10^{-3}$; Pearson $r = 0.44$, $P \ll 10^{-3}$, df = 1,732).

The second scenario assumes that all divergence is due to symmetric (equiprobable) gain of interactions ($P_l = 0$) in the two duplicates. It yields identical results (Spearman $r_s = -0.1$, $P \ll 10^{-3}$; Pearson $r = 0.46$, $P \ll 10^{-3}$, df = 1,732). The third scenario assumes that divergence is due to a mix of both loss and gain of interactions ($P_l = 0.5$), where both duplicates lose or gain interactions symmetrically, that is, with equal probability. It also leads to a fundamentally different distribution of interactions compared with that observed in the data. (Spearman $r_s = -0.08$, $P \ll 10^{-3}$; Pearson $r = 0.46$, $P \ll 10^{-3}$, df = 1,732). Similar to the simulated data shown in figure 2*b*, the L-shape observed in the data also disappears under the latter two scenarios.

Independent genome-scale two-hybrid experiments using different experimental designs (Uetz et al. 2000; Ito et al. 2001) show limited overlap in the interactions they detect. It is thus advisable to ensure that the observed patterns of divergence are not artifacts of a particular experimental technique. I have repeated the above analysis with yeast protein interaction data taken from the MIPS database (Mewes et al. 1999), from which I eliminated all protein interaction information generated by two-hybrid experiments. The remaining 899 interactions among 680 yeast proteins have been experimentally confirmed using techniques ranging from Western blotting to coimmunoprecipitation. The global pattern of interactions among paralogues follows closely that of the two-hybrid data, an L-shaped distribution indicating asymmetry (fig. 2*c*) and a highly negative statistical association (Spearman $r_s = -0.52$, $P \ll 10^{-3}$; Pearson $r = -0.15$, $P \ll 10^{-3}$, df = 1,357). This pattern is not explicable through symmetric loss of interactions (Spearman $r_s = 0.12$, $P \ll 10^{-3}$; Pearson $r = 0.54$, $P \ll 10^{-3}$, df = 1,357), symmetric gain of interactions (Spearman $r_s = 0.10$, $P \ll 10^{-3}$; Pearson $r = 0.49$, $P \ll 10^{-3}$, df = 1,357), or symmetric gain and loss of interactions (Spearman $r_s = 0.09$, $P \ll 10^{-3}$; Pearson $r = 0.53$, $P \ll 10^{-3}$, df = 1,357).

In summary, protein interactions among products of duplicate genes diverge asymmetrically, i.e., one paralogue has more protein interactions than the other. This asymmetry is statistically highly significant and is not explicable through independent (equiprobable) loss or gain of function in the duplicates.

Asymmetric Response to Environmental Stresses

Unicellular organisms like yeast have evolved elaborate cellular responses, allowing them to adapt to drastic environmental changes. They can not only withstand fluctuations in temperature, osmolarity, environmental acidity, and types and quantity of nutrients but also survive the influence of radiation and toxic chemicals. During environmental change, many genes alter their transcriptional activity. Such changes in mRNA expression profile provide valuable insights into gene functions (Chu et al. 1998; Eisen et al. 1998; Spellman et al. 1998; Gasch et al. 2000). A recent study examined the genomic mRNA expression response of most yeast genes to a variety of environmental stressors (Gasch et al. 2000). To assess the differential response of duplicate genes to these stressors, I analyzed data from 11 different stress responses, including heat shock, hyperosmotic shock, amino acid, and nitrogen starvation (Gasch et al. 2000). I excluded the most closely related paralogues ($K_s < 0.5$) from the analysis because cross-hybridization does not allow them to be distinguished by microarray analysis. For the remaining 2,841 paralogous gene pairs, with $K_a < 0.75$ and $K_s > 0.5$, I identified the number of stressors to which each member of the pair responds.

There is again a pronounced asymmetry in the response of gene duplicates to these stresses, as indicated by a significantly negative statistical association between the number of stresses the first and second gene respond to (Spearman $r_s = -0.33$, $P \ll 10^{-3}$; Pearson $r = -0.1$, $P \ll 10^{-3}$, df = 2,839). Completely analogous to the tests for symmetric divergence in protein interactions, I analyzed whether this association is consistent with the null hypothesis that the paralogues originally responded identically to these 11 stresses but that divergence occurred symmetrically for the two gene duplicates. This hypothesis must be rejected, regardless of whether divergence occurs through loss of responses (Spearman $r_s = -0.003$, $P > 0.5$; Pearson $r = 0.21$, $P \ll 10^{-3}$, df = 2,839), gain of responses (Spearman $r_s = -0.0004$, $P > 0.5$; Pearson $r = 0.21$, $P \ll 10^{-3}$, df = 2,839), or a mix of loss and gain of responses (Spearman $r_s = -0.002$, $P > 0.5$; Pearson $r = 0.21$, $P \ll 10^{-3}$, df = 2,839). In summary, the distinct asymmetry in divergence observed for protein interactions also holds for another aspect of gene function, the response to environmental stress.

Asymmetric Response to Genetic Perturbations

The results of a large-scale gene perturbation experiment in yeast, involving several hundred gene-

←

paralogues shown in (*a*) if divergence after duplication had occurred through independent equiprobable interaction loss in either duplicate. (*c*) Empirical data obtained from experiments not using the two-hybrid assay (Mewes et al. 1999). Notice the distinct L-shape of the empirical data in (*a*) and (*c*), which disappears in the simulated data (*b*).

knockout mutations in combination with microarray measurements of changes in the expression of 6,312 yeast genes, have been reported (Hughes et al. 2000). Measuring the effect of a null mutation in a gene on the expression of all other genes does not distinguish between direct and indirect effects of the mutation. Its advantage, however, is that it is a very comprehensive means to assay genetic interactions.

For the purpose of this article it is relevant that the available data (Hughes et al. 2000) contain information on the knockout effect of 11 paralogous gene pairs with $K_a < 1$. For these 11 gene pairs, I compared the number of genes whose expression is affected by a null mutation in each member of the pair (table 1). Interpreting differences between paralogues in the number of affected genes is complicated because these differences are not only the result of divergence between the paralogues but also include effects from the divergence of genes interacting with each paralogue. But the advantage of a perturbation approach is that it provides a more comprehensive assessment of functional differences between paralogues than a mere analysis of direct physical protein interactions. It exposes how the effects of a mutation ripple through a transcriptional regulation network.

Similar to the analysis discussed above, one can ask whether the observed differences between paralogues can be attributed to independent and equiprobable loss or gain of genetic interactions. For seven out of 11 gene pairs in table 1, both these null hypotheses must be rejected, that is, these seven gene pairs show statistically significant asymmetries in divergence. Eliminating one of two paralogous genes affects a substantially greater number of other genes than eliminating the other.

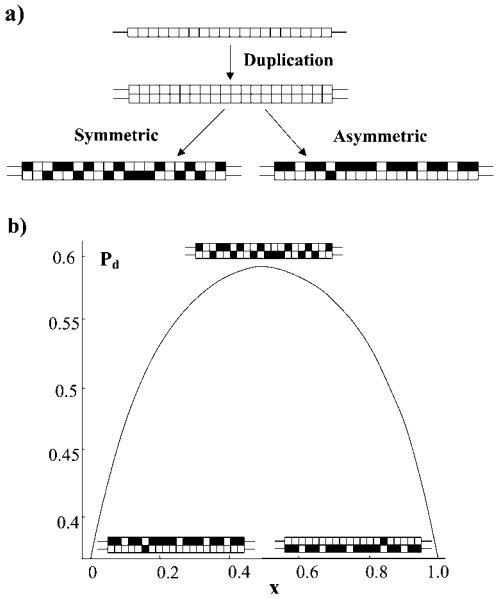

FIG. 3.—Asymmetric divergence and mutational robustness. (*a*) Schematic depiction of symmetric versus asymmetric divergence of two duplicate genes with 20 (sub)functions, represented by white boxes. Black boxes indicate that a gene has suffered a mutational loss of the respective function. In asymmetric divergence, this loss occurs preferentially in one gene. (*b*) The probability P_d that a loss-of-function mutation has a deleterious effect, i.e., it eliminates a function not ''covered'' by the other gene, as a function of x, the degree of asymmetry in divergence. P_d is smallest for maximal asymmetry in divergence, i.e., for $x = 0$ and $x = 1$.

Discussion

In all three data sets, evidence for asymmetric divergence is unequivocal. Gene perturbations affect the expression of a moderate to large number of genes. This makes it possible to derive statistical evidence for asymmetric divergence of paralogous genes from individual gene pairs. Seven out of 11 perturbed gene pairs show such evidence. The number of environmental stresses to which a gene responds is typically smaller and so is the number of protein interactions of gene products. These smaller numbers make it more difficult to derive solid evidence for asymmetric divergence from individual gene pairs. But such evidence emerges when analyzing multiple gene pairs.

What causes asymmetric divergence? Here, I present a simple model of divergence through loss of common functions. Figure 3 explains the basic idea behind this model. It applies to the divergence of genes that have several suitably defined functions (represented by white boxes in fig. 3*a*), as indicated by observed molecular interactions or patterns of gene expression. Immediately after a duplication, two duplicates are identical in all these functions. The model makes only two assumptions about the process of divergence, both of them very simple. First, every function must be exercised by at least one of the two genes. Organisms in which this does not hold will suffer reduced fitness. Second, a loss-of-function mutation (1) affects each of the duplicates with equal probability (1/2), and (2) eliminates one of the affected gene's functions. In this context, what is the probability P_d of suffering a deleterious mutation if the two duplicates have diverged symmetrically versus asymmetrically? Asymmetric divergence means that one duplicate has lost more functions than the other. Assume that since the duplication, duplicates 1 and 2 have lost a fraction l_1 and l_2 of their functions, respectively ($0 < l_1, l_2 \leq 1$). Let $l = l_1 + l_2$ be the total fraction of functions lost ($0 \leq l \leq 1$). If no function is allowed to have been lost in both genes, the probability that a mutational loss of one further function has a deleterious effect is equal to

$$P_d = \frac{1}{2}\left(\frac{l_2}{1 - l_1}\right) + \frac{1}{2}\left(\frac{l_1}{1 - l_2}\right).$$

Upon expressing l_1 and l_2 in terms of the total fraction of functions lost, l, using $x := l_1/l$ ($1 - x = l_2/l$), this expression becomes

$$P_d(x, l) = \frac{1}{2}\left(\frac{1 - x}{(1/l) - x}\right) + \frac{1}{2}\left(\frac{x}{(1/l) - 1 + x}\right).$$

In this expression, a value of $x = 0.5$ indicates symmetric divergence. The pertinent feature of $P_d(x, l)$ is that it is unimodal: regardless of l, it has a maximum at $x = 0.5$ (fig. 3*b*). This means that the probability of a deleterious mutation is greatest if two genes have di-

verged symmetrically. Thus, asymmetric divergence minimizes the risk of deleterious mutations.

If this model is correct, we see asymmetrically diverged gene pairs because organisms harboring them have survived preferentially in the past. Importantly, natural selection would act in an indirect, second-order manner on such gene pairs. In a population polymorphic for gene duplicates at different stages of divergence, different individuals would not necessarily have different fitness levels; rather, the propensity of such individuals to suffer deleterious mutations would be different. Individuals with symmetrically diverged duplicates would thus be preferentially eliminated from the population through deleterious mutations.

One might assume that the selective advantage of having asymmetrically diverged gene duplicates must be minute. After all, differences in fitness do not manifest themselves until new loss-of-function mutations arise. For any organism, the expected waiting time for such a new loss-of-function mutation is proportional to the inverse of the mutation rate μ (Hartl and Clark 1989, p. 98). During this time, symmetrically diverged gene duplicates are free to go to fixation via random drift. Formal population genetic analysis (Wagner 2000) shows that for sufficiently large population sizes ($N > 1/\mu$) the lens of natural selection has sufficient resolving power to perceive differences in mutational robustness and to act on them. For microorganisms like yeast, attainable population sizes may well be in the required range. In addition, this minimally required population size is based on the evolution of only one diverging gene pair (Wagner 2000). It may be much smaller for multiple gene pairs and their cumulative effects on mutational robustness.

The requirement for large effective population sizes suggests a test for the model. In organisms with small effective population sizes, such as many higher vertebrates, we would not expect asymmetric divergence of gene duplicates. (The necessary data are not yet available.) A requirement for persistently large population sizes may also be one of the reasons why the asymmetry observed is not perfect and does not hold for all genes. Depending on a gene and its functions, a loss-of-function mutation may have very subtle fitness effects. In conjunction with fluctuating effective population sizes, the selection pressures for asymmetrical divergence may fluctuate as well. Some genes thus diverge symmetrically, whereas others do not.

The foundation of this speculative model is the assumption that gene duplicates diversify mostly through loss of common functions. The model is thus a neutral model in the sense that adaptive mutations providing fitness benefits play no role in it. Although neutral divergence of gene duplicates has received much attention in recent work (Nowak et al. 1997; Gibson and Spring 1998; Force, Lynch, and Postlethwait 1999; Wagner 1999; Lynch and Force 2000; Wagner 2000) and is probably an important mode of gene evolution, the importance of beneficial mutations must not be neglected (Hughes 1994; Kreitman and Akashi 1995; Walsh 1995; Ludwig, Patel, and Kreitman 1997; Cirera and Aguade 1998; Tsaur, Ting, and Wu 1998). Recent evidence using fully sequenced genomes further underscores the abundance of beneficial mutations and thus the importance of scenarios of sequence divergence that involve such mutations (Fay, Wyckoff, and Wu 2002). Although it is not clear how adaptive mutations might lead to asymmetric functional divergence of gene duplicates, the cause may be as simple as that one adaptive mutation leads to a cascade of further such mutations and consequent functional change. To distinguish between neutral models of asymmetric functional divergence and models involving adaptive mutations will be a major task for future work.

Acknowledgment

Financial support through NIH grant GM63882 is gratefully acknowledged.

LITERATURE CITED

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. H. ZHANG, Z. ZHANG, W. MILLER, and D. J. LIPPMAN. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

ASHBURNER, M., C. A. BALL, J. A. BLAKE et al. (20 co-authors). 2000. Gene ontology: tool for the unification of biology. Nat. Genet. **25**:25–29.

BARTEL, P. L., J. A. ROECKLEIN, D. SENGUPTA, and S. FIELDS. 1996. A protein linkage map of Escherichia coli bacteriophage T7. Nat. Genet. **12**:72–77.

BENDER, W., M. AKAM, F. KARCH, P. A. BEACHY, M. PEIFER, P. SPIERER, E. B. LEWIS, and D. S. HOGNESS. 1983. Molecular genetics of the Bithorax complex in *Drosophila melanogaster*. Science **221**:23–29.

CHU, S., J. DERISI, M. EISEN, J. MULHOLLAND, D. BOTSTEIN, P. O. BROWN, and I. HERSKOWITZ. 1998. The transcriptional program of sporulation in budding yeast. Science **282**:699–705.

CIRERA, S., and M. AGUADE. 1998. Molecular evolution of a duplication: the sex-peptide (Acp70a) gene region of *Drosophila subobscura* and *Drosophila madeirensis*. Mol. Biol. Evol. **15**:988–996.

COEN, E. S., and E. M. MEYEROWITZ. 1991. The war of the whorls: genetic interactions controlling flower development. Nature **353**:31–37.

EISEN, M. B., P. T. SPELLMAN, P. O. BROWN, and D. BOTSTEIN. 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95**:14863–14868.

EKKER, M., M. A. AKIMENKO, M. L. ALLENDE, R. SMITH, G. DROUIN, R. M. LANGILLE, E. S. WEINBERG, and M. WESTERFIELD. 1997. Relationships among msx gene structure and function in zebrafish and other vertebrates. Mol. Biol. Evol. **14**:1008–1022.

EKKER, S. C., A. R. UNGAR, P. GREENSTEIN, D. P. VONKESSLER, J. A. PORTER, R. T. MOON, and P. A. BEACHY. 1995. Patterning activities of vertebrate hedgehog proteins in the developing eye and brain. Curr. Biol. **5**:944–955.

FAY, J. C., G. J. WYCKOFF, and C. I. WU. 2002. Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature **415**:1024–1026.

FORCE, A., M. LYNCH, and J. POSTLETHWAIT. 1999. Preservation of duplicate genes by subfunctionalization. Am. Zool. **39**:460.

FROMONT-RACINE, M., J. C. RAIN, and P. LEGRAIN. 1997. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. Nat. Genet. **16**:277–282.

GASCH, A. P., P. T. SPELLMAN, C. M. KAO, O. CARMEL-HAREL, M. B. EISEN, G. STORZ, D. BOTSTEIN, and P. O. BROWN. 2000. Genomic expression programs in the response of yeast cells to environmental change. Mol. Biol. Cell **11**: 4241–4257.

GIBSON, T. J., and J. SPRING. 1998. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet. **14**:46–49.

HARTL, D. L., and A. G. CLARK. 1989. Principles of population genetics. Sinauer Associates, Sunderland, Mass.

HUGHES, A. L. 1994. The evolution of functionally novel proteins after gene duplication. Proc. R. Soc. Lond. Ser. B Biol. Sci. **256**:119–124.

HUGHES, T. R., M. J. MARTON, A. R. JONES et al. (22 co-authors). 2000. Functional discovery via a compendium of expression profiles. Cell **102**:109–126.

ITO, T., T. CHIBA, R. OZAWA, M. YOSHIDA, M. HATTORI, and Y. SAKAKI. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA **98**:4569–4574.

ITO, T., K. TASHIRO, S. MUTA, R. OZAWA, T. CHIBA, M. NISHIZAWA, K. YAMAMOTO, S. KUHARA, and Y. SAKAKI. 2000. Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc. Natl. Acad. Sci. USA **97**:1143–1147.

JACK, J., and Y. DELOTTO. 1995. Structure and regulation of a complex locus: the cut gene of Drosophila. Genetics **139**: 1689–1700.

KIRCHHAMER, C. V., C. H. YUH, and E. H. DAVIDSON. 1996. Modular cis-regulatory organization of developmentally expressed genes: 2 genes transcribed territorially in the sea-urchin embryo and additional examples. Proc. Natl. Acad. Sci. USA **93**:9322–9328.

KREITMAN, M., and H. AKASHI. 1995. Molecular evidence for natural selection. Annu. Rev. Ecol. Syst. **26**:403–422.

LEE, K. H., Q. H. XU, and R. E. BREITBART. 1996. A new tinman-related gene, nkx2.7, anticipates the expression of nkx2.5 and nkx2.3 in zebrafish heart and pharyngeal endoderm. Dev. Biol. **180**:722–731.

LI, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**: 96–99.

LI, W.-H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

LI, X. L., and M. NOLL. 1994. Evolution of distinct developmental functions of 3 Drosophila genes by acquisition of different cis-regulatory regions. Nature **367**:83–87.

LUDWIG, M. Z., N. H. PATEL, and M. KREITMAN. 1997. Evolution of the even-skipped stripe-2 enhancer of Drosophila. Dev. Biol. **186**:A27.

LYNCH, M., and J. S. CONERY. 2000. The evolutionary fate and consequences of duplicate genes. Science **290**:1151–1155.

LYNCH, M., and A. FORCE. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics **154**: 459–473.

MEHLHORN, K., and S. NAHER. 1999. LEDA: a platform for combinatorial and geometric computing. Cambridge University Press, Cambridge.

MENA, M., B. A. AMBROSE, R. B. MEELEY, S. P. BRIGGS, M. F. YANOFSKY, and R. J. SCHMIDT. 1996. Diversification of C-function activity in maize flower development. Science **274**:1537–1540.

MEWES, H. W., K. HEUMANN, A. KAPS, K. MAYER, F. PFEIFFER, S. STOCKER, and D. FRISHMAN. 1999. MPS: a database for genomes and protein sequences. Nucleic Acids Res. **27**:44–48.

NOWAK, M. A., M. C. BOERLIJST, J. COOKE, and J. MAYNARD-SMITH. 1997. Evolution of genetic redundancy. Nature **388**: 167–171.

RUBIN, G. M., M. D. YANDELL, J. R. WORTMAN et al. (54 co-authors). 2000. Comparative genomics of the eukaryotes. Science **287**:2204–2215.

SCHMIDT, R. J., B. VEIT, M. A. MANDEL, M. MENA, S. HAKE, and M. F. YANOFSKY. 1993. Identification and molecular characterization of zag1, the maize homolog of the Arabidopsis floral homeotic gene agamous. Plant Cell **5**:729–737.

SCHWIKOWSKI, B., P. UETZ, and S. FIELDS. 2000. A network of protein-protein interactions in yeast. Nat. Biotechnol. **18**: 1257–1261.

SLUSARSKI, D. C., C. K. MOTZNY, and R. HOLMGREN. 1995. Mutations that alter the timing and pattern of cubitus interruptus gene expression in *Drosophila melanogaster*. Genetics **139**:229–240.

SPELLMAN, P. T., G. SHERLOCK, B. FUTCHER, P. O. BROWN, and D. BOTSTEIN. 1998. Identification of cell-cycle regulated genes in yeast by DNA microarray hybridization. Mol. Biol. Cell **9**:2155.

TSAUR, S. C., C. T. TING, and C. I. WU. 1998. Positive selection driving the evolution of a gene of male reproduction, Acp26aa, of Drosophila: divergence versus polymorphism. Mol. Biol. Evol. **15**:1040–1046.

UETZ, P., L. GIOT, G. CAGNEY et al. (20 co-authors). 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature **403**:623–627.

WAGNER, A. 1999. Redundant gene functions and natural selection. J. Evol. Biol. **12**:1–16.

———. 2000. The role of pleiotropy, population size fluctuations, and fitness effects of mutations in the evolution of redundant gene functions. Genetics **154**:1389–1401.

———. 2001. The yeast protein interaction network evolves rapidly and contains few duplicate genes. Mol. Biol. Evol. **18**:1283–1292.

WALSH, J. B. 1995. How often do duplicated genes evolve new functions? Genetics **139**:421–428.

HERVE PHILIPPE, reviewing editor

Accepted June 7, 2002