

On the source of mutational robustness
in genetic networks of yeast

Andreas Wagner

University of New Mexico

and

The Santa Fe Institute

University of New Mexico

Department of Biology

167A Castetter Hall

Albuquerque, NM 87131-1091

Phone: +505-277-2021

FAX: +505-277-0304

wagnera@unm.edu

Abstract

Resilience of an organism's physiological and developmental processes against mutations can have two principal causes, the first being overlapping gene functions. In this case, loss-of-function mutations in one gene will have little phenotypic effects if one or more other genes with similar functions serve as a "back-up". The second cause stems from interactions among genes with unrelated functions, and has been documented in metabolic and regulatory gene networks. The subject of this paper is to analyze, on a genome-wide scale, which of these causes of robustness is more important. To this end, I use functional genomics data from the yeast *Saccharomyces cerevisiae* to test a series of hypotheses related to the following: If gene duplications are mostly responsible for robustness, then a correlation is expected between the similarity of two duplicated genes and the effect of mutations in one of these genes. The results demonstrate that interactions among unrelated genes are the major cause of robustness against mutations. This type of robustness is unlikely to be an intrinsic property, but is probably an evolved response of genetic networks to stabilizing selection.

Introduction

A large and growing body of evidence shows that both physiological and developmental processes of eukaryotes display considerable robustness against mutations. For instance, many loss-of function mutations of important developmental genes in higher organisms show weak or no phenotypic effects¹⁻⁶. Several morphological traits of *Drosophila melanogaster* show a conspicuous lack of morphological variation, despite a considerable amount of (hidden) genetic variation⁷⁻¹⁰. A recent study linking the heat-shock protein Hsp90 to the buffering of development against genetic variation provides one potential molecular mechanism for this phenomenon¹¹. On the cellular level, the mere fact that most loss-of-function mutations at enzymatic loci are recessive demonstrates the resilience of biochemical pathways against changes in gene dosage¹². Recent results from systematic knock-out mutations of all genes on chromosome V of the yeast *Saccharomyces cerevisiae* shows that almost 40 percent of all yeast genes have little or no detectable effect on growth rate in five different environments¹³.

Two main causes of robustness. There are two principal mechanistic causes for such resilience against mutations. The first is redundant or overlapping gene functions, where mutations in one gene have little effect if there are one or more genes with similar functions that can serve as a back-up. A plethora of evidence on knock-out mutations in developmental genes supports this scenario, stemming from considerable similarity between mutated gene and “back-up” genes, as well as from many functional studies of such genes^{1-6,14}. Because the majority of eukaryotic genes (e.g., 60 percent in the case of yeast) have duplicates in the same genome, and because many of these duplicates have redundant functions, it seems plausible that redundancy among duplicates is the main cause of resilience against mutations. However, there is a second possible cause stemming from interactions among genes with unrelated biochemical functions. It is best illustrated for the case of recessive mutations in biochemical pathways. Here, gene

products completely unrelated in function, i.e., enzymes that catalyze different chemical reactions, contribute to a reaction chain or network whose goal it is to sustain an optimal flux of metabolites¹⁵. Metabolic control theory, the mathematical description of flux in metabolic networks, demonstrates that the recessiveness of mutations can only be understood as a property of the interactions of all enzymes in the chain, interactions that compensate for mutational changes in gene dosage at one locus^{12,16-18}. In addition, large metabolic networks can compensate even for complete loss-of-function of one or more enzymatic reactions by exploiting alternative metabolic routes¹⁹. Further potential examples include the buffering of development against genetic variation in *Drosophila* via the heat-shock protein Hsp90¹¹, as well as recent modeling work suggesting that regulatory such networks can evolve great resilience against mutations without duplicate genes of similar function²⁰.

Which cause is more important on a genome-wide scale? Both causes of robustness, gene duplications and buffering via interactions of evolutionarily unrelated genes are well documented. How can we decide which of them is more important? The answer lies in the only aspect that distinguishes the two scenarios: the relation of buffering to similarity among genes. Immediately after the duplication of a gene, loss-of-function of one of the duplicates is likely to have no phenotypic effect. This is supported by genetic data on ancient genome duplications, after which many duplicated genes appear to have been lost^{21,22}. If both original and duplicate are retained, they will then diverge in their functions, and hence in their sequences and/or expression patterns. As long as their functions overlap to some extent, phenotypic effects of loss-of-function will be weak, which appears to be the case for the multiple partially redundant genes in vertebrate development¹⁻⁶. Finally, original and duplicate gene will have diverged completely, and mutational effects will be severe. Sequence similarity is one indicator of functional similarity among duplicated genes. Another such indicator is similarity in two genes' spatiotemporal expression pattern. While expression patterns can be similar for reasons unrelated to evolutionary

history, there are several cases of overlapping gene functions in vertebrates where expression pattern is a better indicator of functional divergence^{1,23} than is sequence. In sum, if past gene duplications are the prime cause of mutational robustness within an organism, one would expect a strong correlation between similarity among genes and fitness effects, if many duplicated genes are studied.

In the yeast *Saccharomyces cerevisiae*, data has recently become available that allows a test of this prediction. First, both the complete genome sequence and micro array expression data are available for several important cellular processes²⁴⁻²⁶. Second, yeast underwent a genome duplication approximately 100Mya, the remnants of which are 55 syntenic pairs of blocks of duplicated genes²⁷. Although these are only a small fraction of all duplicated yeast genes, their identical duplication time distinguishes them from other duplicates. Third, Smith *et al.*¹³ generated Ty1-transposon induced loss-of-function mutations in each of 255 genes on yeast chromosome V, and determined the growth rate of the resulting 255 strains with then unprecedented accuracy in six different environments, as well as their mating efficiency. On chromosome V are also 45 genes within 6 blocks that were duplicated in the genome duplication event²⁷.

Are dispensable genes really dispensable? Before embarking on a test of the above prediction, a brief discussion of the knock-out data used here is necessary. The data stems from a liquid-culture competition assay that determined growth rate to approximately 5% resolution in 5 different environments and, in addition, mating efficiency¹³. For any of the genes whose knock-out has no detectable phenotype, the question arises whether this is merely due to the limited number of environments and fitness components assayed. While one can not exclude this with certainty, it is worth noting that almost 90% of the mutants with a detectable phenotype displayed this phenotype in all of the assayed conditions, and only about 10 percent had growth defects specific to one environment¹³. This suggests that for the majority of genes a mutant phenotype

may manifest itself under most conditions, although a rigorous assessment of this hypothesis must await the availability of more data. More problematic is perhaps the limited resolution of growth rate differences (5%) between knock-out mutant and wild-type. While a more recent, smaller study increased the experimental resolution of growth rate differences by one order of magnitude²⁸, and was still unable to detect growth differences for approximately 20% of knock-out mutations, even smaller differences might affect the persistence of a genotype in a population. Whether they do depends on the (unknown) effective population size²⁹ of yeast.

While there may well be no truly dispensable genes, it is worth keeping in mind that experimenter-imposed knock-out mutations are much more drastic genetic perturbations than the more subtle point mutations, insertions, and deletions which genes are usually exposed to. Results from knock-out studies are thus best used as indicators for the effects that these more subtle mutations have on fitness. They show that there is an enormous breadth in the distribution of fitness effects, with many mutations having very subtle effects. The results reported below show that unrelated genes and not redundant duplicate genes are the prime source of these subtle effects.

Results

Phenotypic effect of knock-out and sequence similarity of duplicate genes. I took three complementary approaches to assess whether there is a relationship between the severity of the fitness effect of knock-out mutations on one hand, and similarity in sequence and expression of mutated genes to other yeast genes, on the other hand. First, I assessed the relation between fitness effects and sequence similarity for 45 genes in six blocks on chromosome V that are remnants of a past genome duplication event in yeast²⁷. Each of the 45 genes has a paralogue in a

syntenic block of genes on some other yeast chromosome²⁷. Ref. 13 provided information on the effect of knock-out mutations in the 45 block-duplicated genes. The authors of this study determined the growth rate effects of knock-out mutations in each of 255 genes on chromosome V, and placed them into 4 categories. For the purpose of this study, I encoded fitness effects numerically, ranging from 0 (indistinguishable from wild-type) to 3 (most severe fitness effect, greater than 25% reduction in fitness). I then calculated various measures of statistical associations between fitness effects and sequence similarity among duplicate gene pairs, using six different measures of sequence similarity. Fig. 1 and Table 1 illustrate the result which holds regardless of the similarity measure used: genes whose knock-out has little or no fitness effect do not have paralogues with more similar sequence than do genes with severe fitness effects.

Phenotypic effect of knock-out and similarity in expression pattern of duplicate genes. To assess whether differences among expression patterns may be better indicators for the severity of fitness effects, I pooled publicly available expression data from three separate studies that had determined the expression of all yeast genes during the cell-cycle, during sporulation, and during the diauxic shift (the switch from anaerobic to aerobic metabolism upon depletion of glucose)²⁴⁻²⁶. For each of the 45 duplicated gene pairs, I calculated several measures of similarity in mRNA expression pattern from this data set. I then determined whether there was a correlation between similarity in gene expression pattern and the fitness effects of mutations for the 45 block-duplicated genes. The results, summarized in Fig. 2 and Table 1, show that there is no significant association between severity of fitness effect and similarity in expression pattern.

Must perhaps both similarity in sequence and expression pattern be high in order for loss-of-function mutations to have weak effects? In the data set of 45 genes used here, there were 9 genes with more than 80% amino acid similarity, and a significant Pearson correlation coefficient greater than $r=0.8$ in their expression pattern. While one can not draw statistically sound conclusions from such a small number of genes, the number of genes (7 of 9 [77%]) that lead to a

detectable fitness effect is in fact greater than that for the remainder of the 45 genes (19 of 36 [53%]). This is partly because Table 2 contains several ribosomal genes, for which duplicates are often both highly conserved in sequence and expression pattern, yet show detectable fitness defects when mutated.

Duplicate genes outside duplicate gene blocks. The second of three approaches to determine the relationship between gene similarity and mutational effects involved all genes on chromosome V outside the six duplicated blocks. I used these genes in a BLAST search for the ORFs most similar to them anywhere in the yeast genome (see “Methods”). A correlation analysis with the resulting 17 gene pairs, analogous to that discussed above, yielded exactly the same qualitative results (not shown): there is no statistical association between measures of similarity among the most similar paralogous genes and fitness effect of mutations.

Do genes in duplicated blocks have weaker phenotypic effects than other genes? Shortly after a genome duplication, many duplicated genes may be lost. After sufficient time has elapsed to allow for functional diversification among the retained genes, one might expect that some “equilibrium” distribution of the effects of loss-of-function mutations is attained, where some genes cause severe fitness defects, while others might be dispensable (see Fig. 3 for illustration). Gene redundancy resulting from ancient genome duplications in vertebrates suggests that the amount of time necessary to arrive at such an equilibrium distribution, whatever its nature, may be much longer than the 100 Myrs elapsed since the yeast genome duplication²⁷. This raises the question whether the many weak fitness effects observed in yeast are remnants of the yeast genome duplication event.

To address this question, one has to compare the distribution of mutational effects between (i) a sample of gene pairs retained from the genome duplication, and (ii) a reference set of genes not duplicated in the genome duplication. The more than 200 genes outside duplicated blocks on chromosome V are a reasonable choice for this reference set, because only eight of

them are expected to have paralogous partners that have been retained after the genome duplication²². Fig. 4 compares the distribution of fitness effects for mutations in all duplicated blocks on chromosome V, and the same distribution for all genes outside the duplicated blocks. They are statistically indistinguishable. This suggests that the large number of genes with small knock-out effects in the yeast genome is not primarily a result of the genome duplication.

Many genes with weak or no knock-out effect have no paralogues. In the last of three approaches, I assigned all genes on chromosome V to one of two groups, those whose knock-out had no detectable fitness effect (94 genes), and those whose knock-out had the most severe fitness effect (46 genes). I then asked whether there are any differences between the two sets of genes with regard to (i) the number of similar genes in the genome, (ii) the sequence similarity of the most similar genes, and (iii) the similarity in expression pattern of the most similar genes.

Question (i) was specifically motivated by the possibility that not only one similar gene, but a whole gene family might be jointly responsible for mutational buffering. Fig. 5 shows the answer to this question, the distribution of the number of genes with detectable similarity to chromosome V genes in the two categories. The most striking observation is that 41 (43.6%) of chromosome V genes with no detectable fitness effect also have no similar genes in the yeast genome. This result is not an artefact of too stringent a criterion to determine similarity among genes (see “Methods”). Overall, the two distributions shown in Fig. 5 are significantly different ($\chi^2=40.9$; $df=4$; $P<0.001$), but the differences are not in the direction one would predict if fitness effects of a mutation were correlated with gene family size.

Next, I asked whether paralogues of chromosome V genes with weak fitness effects display greater sequence similarity (Fig. 6a) or greater similarity in expression pattern than paralogues of genes with strong fitness effects (Fig. 6b). Where there are differences in the distribution of similarity, the differences are not consistent with the notion that greater similarity is associated with weaker fitness effects (Fig. 6). Qualitatively identical results (not shown) are

obtained if similarity data from the five or ten most closely related genes is averaged and analyzed.

Discussion

When compared to genes whose loss-of-function results in severe fitness defects, genes with weak or no fitness effects (i) are not more similar to their closest paralogues, both in sequence and temporal expression pattern, (ii) are not part of larger gene families, whose members are, on average, more similar in sequence or expression to the mutated gene, and (iii) have no detectably related yeast genes at all in about half of the cases studied. These cases include 45 paralogous gene pairs which are the remnants of a past genome duplication event. The distribution of fitness effects for these 45 genes is the same as that observed for the rest of a large genomic sample studied. Thus, while gene duplications may be responsible for a fraction of weak knock-out phenotypes³⁰, they contribute little to mutational robustness on a genomic scale.

Caveats. The data available for this analysis has shortcomings that will hopefully be overcome by future improvements of the underlying technology and analysis methods. First, in the "genetic footprinting" approach to gene disruptions taken by Smith *et al.*¹³, Ty1 element insertion far downstream from the start codon may lead to expression of a gene product with residual function. Secondary insertions of Ty1 elements at other positions may further complicate the interpretation of results. It is thus reassuring that more recent studies^{28,31} using different gene disruption techniques yield a distribution of mutant effects not inconsistent with that analyzed here. Second, various measures of sequence divergence are used here as a (less than ideal) proxy for functional divergence among duplicated genes. Some improvement can be expected once large scale structural comparisons of gene products become feasible. However, the results of this study do not rest solely on the comparison of closely related paralogous gene

pairs, but are also derived from the complementary approaches pursued here. Third, the only data sets currently available to allow comparison of gene expression patterns on a large scale are temporal mRNA expression profiles. Among the shortcomings of such data are (i) the considerable amount of noise in large scale mRNA expression data, (ii) the neglect of spatial information, as well as of translational and posttranslational regulation, and (iii) the limited ability of the microarray approach²⁶ to distinguish among mRNAs produced by closely related members of gene families.

Convergent evolution? About 40 percent of all genes with no detectable fitness effect also had no detectably related similar genes in the yeast genome. It is important to note one caveat to this finding, namely that homology search algorithms will miss a substantial number of proteins with similar tertiary structures³². However, it is not clear what fraction of proteins with similar structure but no sequence similarity share a common evolutionary history, and what fraction might be the product of convergent evolution on the structural level³³.

Could convergent evolution explain all or even most cases of genes with no detectable fitness effect and no similar genes in the yeast genome? If it can, at least one “back-up” gene with identical biochemical function but completely unrelated sequence and perhaps even unrelated structure would exist for each dispensable gene. Such back-up genes may exist but are probably rare³⁴. Extrapolating the results from chromosome V to the rest of the genome suggests the existence of approximately 1000 genes in this category (38 percent of genes with little or no fitness effect, 43 percent of which are unique). This indicates the enormous and improbable scale at which convergent evolution would have to occur. However, even if functional convergence is more frequent than currently thought, the question arises why convergence would occur on such a massive scale. The answer may well have to do with network resilience provided by functional redundancy.

Robustness, intrinsic or evolved, beneficial or detrimental? A broad distribution of mutational effects could either be an intrinsic and unchangeable feature of many genetic networks, or an evolved property. This question has been studied empirically and with mathematical models in a variety of contexts, including the evolution of dominance in enzymatic genes of *E. coli*, genetic canalization of morphological characters in *Drosophila*, and for metabolic as well as regulatory gene networks^{7-9,11,12,17,20,35}. These studies suggest unequivocally that robustness is an evolutionary response of genetic systems to stabilizing selection for either mutational stability or for stability against environmental fluctuations.

Increased mutational robustness in genetic networks would evolve via an indirect mechanism, where robust networks do not confer higher fitness on their carrier, but accumulate in a population because mutations in them are less likely to have deleterious effects. As a consequence, increased robustness increases the mean fitness of a population by an amount that may only be of the order of the mutation rate^{36,37}. Thus, evolution of robustness is not an adaptive phenomenon, even if it increases resilience to mutations of a network by orders of magnitude.

During times of prolonged environmental change, when directional selection acts, evolved robustness may even be deleterious. This is because the magnitude of a population's selection response depends on the amount of genetic variation expressed phenotypically, and robustness reduces just that amount. One would thus expect that organisms have evolved not only mechanisms to increase the production of genetic variation e.g., via elevated transposon activity^{38,39}, but also mechanisms to augment the expression of existing genetic variation during such times¹¹. These mechanisms should cause a rather unspecific increase in variation, because the exact kind of variation required can not be anticipated.

As in the case of metabolic networks, where a full mechanistic understanding of robustness and its evolution required an experimentally testable mathematical theory¹⁵, the

mechanistic causes of robustness in other genetic networks, such as signalling pathways, will remain enigmatic until similarly detailed models are available. Thus, curiously, the large scale patterns detectable and afforded by functional genomics demonstrate the pressing need to further study the smallest scale of molecular interactions.

Methods

Sources of data. I obtained amino acid sequences for all *Saccharomyces cerevisiae* open reading frames (ORFs) from the Saccharomyces genome database (ftp://genome-ftp.stanford.edu/pub/yeast/yeast_ORFs; January 1999), and publicly available micro array expression data from (<http://cmgm.stanford.edu/pbrown> and <http://genome-www.stanford.edu/cellcycle/data/rawdata/combined.txt>) for three separate experiments in which the changes in expression level of all yeast genes were assessed at multiple time points during (i) the diauxic shift (7 time points; ref. 25), (ii) sporulation (7 time points; ref. 24), and (iii) the mitotic cell cycle of cells synchronized via α -pheromone induced cell cycle arrest (18 time points; ref. 26). These three data sets represent important components of the life cycle of all yeast cells. For each gene and each time point of an experiment, the data sets give ratios of expression levels in an experimental population and a reference population of cells. To symmetrize the distribution of expression ratios r , it is expedient to transform them using the binary logarithm. Values of $\log_2(r) < 0$ and $\log_2(r) > 0$ correspond to repression and induction, respectively, of a gene. Because the cell cycle data used here was published in the form of \log_2 -transformed expression ratios, I transformed the sporulation and diauxic-shift data sets in the same way, and then pooled all three data sets for analysis.

Wolfe and Shields²⁷ provided strong evidence for a past duplication of the yeast genome. They detected 55 pairs of clusters of highly similar genes whose order between clusters was preserved. On chromosome V, 6 such blocks of genes exist, comprising a total of 45 ORFs [block

4: 12 genes with a syntenic block on chromosome II; block 13: 8 genes/chromosome IV; block 25: 5 genes/chromosome VII; block 26: 3 genes/chromosome VII; block 27: 11 genes/chromosome IX; block 28: 6 genes/chromosome X; data is available from <http://acer.gen.tcd.ie/~khwolfe/yeast>]. Genes in these 45 paralogous pairs are more closely related to each other than to any other gene in the yeast genome, with amino acid identities ranging from 21% to 100% (mean: 63.3%).

Fitness effects of knock-out mutation by Ty1 transposon insertion into each gene on chromosome V are published in ref. 13. The authors of this study measured the fitness of mutants via (i) differences in growth rates to wild-type cells in liquid culture under six different environmental conditions, and (ii) assessing the mating efficiency of mutants. Their technique of “genetic footprinting” allows a resolution of 5% fitness difference between mutant and wild-type¹³. They grouped fitness effects of knock-out mutants into four categories (WT, Q1, Q2, and Q3) corresponding to the severity of the fitness reduction. For the purpose of this study, I encoded these fitness effects numerically as follows. I assigned mutants to category zero, one, two, and three, if their fitness was indistinguishable from wild-type (WT), between 85% and 95% of wild-type (Q1), between 75% and 85% of wild-type (Q2), and less than 75% of wild-type, respectively. Only a small fraction (11.8%) of knock-out mutations showed growth defects specific to only one of the growth conditions used. This was also true for the 45 block-duplicated genes on chromosome V, i.e., only three out of 45 genes showed such specific effects: *YER139C* showed a Q1/Q2 phenotype in growth under high temperature, but a Q3 phenotype in all other selections. It was assigned to fitness category 1. *PAK1* showed a phenotype of Q1 in caffeine and Q2 in all other media (category 2). *PMD1* had a WT/Q3 phenotype for mating, but had a Q1 phenotype in all other media (category 2)¹³.

Correlation analysis of expression patterns. I used three complementary measures of statistical association to assess the degree of similarity in the expression patterns of yeast genes,

the Pearson product-moment correlation coefficient, Kendall's rank correlation coefficient, and Spearman's rank correlation coefficient⁴⁰. Pearson correlation coefficients, while easy to interpret, are sensitive to deviations from a normal distribution and are not suitable to measure nonlinear correlations.

I subjected calculated correlation coefficients r to a statistical test against the null-hypothesis that r is not significantly different from zero. Standard significance tests require that the data be sampled from a normal distribution⁴⁰. This assumption is testable in principle here, but not in practice, because measures of association among thousands of genes and three separate expression time courses were to be calculated, and each of these might have a different underlying distribution. Thus, I performed the following randomization test to assess significant deviations of calculated correlation measures from zero. Denote as $\mathbf{x}=(x_1, \dots, x_n)$ and $\mathbf{y}=(y_1, \dots, y_n)$ the expression time course for two genes x and y in this data set, represented by \log_2 -transformed expression ratios x_i and y_i . I first generated a permutation $\mathbf{x}^{(1)}$ of one of the time series by randomly shuffling all entries of the array \mathbf{x} , and then calculated a Pearson correlation coefficient $r^{(1)}$ for the shuffled time series $\mathbf{x}^{(1)}$ and \mathbf{y} , and compared its magnitude to r . I repeated this process k times, and rejected the null-hypothesis if the absolute value of $r^{(i)}$ exceeded the absolute value of r fewer than $k\mathbf{P}$ times. For the results presented here, $k=1000$ and $\mathbf{P}=0.01$.

Sequence comparisons. I used six different and complementary measures of sequence similarity in comparing paralogous genes that are part of duplicated clusters: percent amino acid identity between two genes, Z-scores obtained from a Smith-Waterman alignment⁴¹, BLAST bit scores⁴², protein distances based on PAM amino acid substitution matrices⁴³, and the estimated fraction of non-synonymous and synonymous nucleotide substitutions per site (denoted as K_a and K_s , respectively). Homology search programs such as BLAST also calculate a score estimating the statistical significance of an alignment. Because the analyzed paralogous gene pairs were

highly similar, these scores were effectively equal to zero for many gene pairs, and were thus not suitable for the purpose of this analysis. I obtained the first three of the above measures of similarity via the analysis tool provided by Wolfe and Shields (1997; <http://acer.gen.tcd.ie/~khwolfe/yeast>) which uses the packages SSEARCH 3.0⁴⁴ and BLAST⁴², both based on the BLOSUM62 similarity scoring matrix⁴⁵ and a filter to eliminate low complexity regions of a protein⁴². I calculated protein distances, which estimate the expected fraction of amino acids changed, with the phylogenetic analysis package PHYLIP⁴⁶. The method and software tool of Comeron⁴⁷ provided K_a and K_s values. All 45 duplicated genes located within conserved, syntenic blocks on chromosome V were more closely related to each other than to any other genes in the yeast genome.

I also carried out sequence searches and comparisons among two larger sets of ORFs, a search for all yeast ORFs that are similar to (i) chromosome V genes outside duplicated clusters, and (ii) all genes on chromosome V in fitness categories 0 (wild-type) and 3 (most severe). For reasons of computational feasibility I used BLAST (v2.0.1; <ftp://ncbi.nlm.nih.gov/blast/executables>), with a filter for low-complexity regions of proteins⁴². For search (i), I considered only ORFs that had at least one matching fragment of length greater than 50 amino acids, with at least 50% amino acid similarity to the query sequence on chromosome V, and for which micro array expression data for both the query sequence and for its most closely related paralogue were available. Final blast scores and amino acid identities for search (i) were based on re-aligning the previously obtained matches without a complexity filter, such that I could obtain similarity scores over the entire length of the fragments.

I restricted search (ii) to genes whose effect on growth rate was not specific to one of the six test conditions¹³, and employed a cut-off expect (E)-value⁴² of 10^{-2} , corresponding to an expected number of 0.01 false positive matches per query sequence for the queried data base. An even less conservative value of E would lead to an unacceptably high fraction of false positive

matches. A recent large scale statistical evaluation of homology search algorithms using protein structure databases³² suggests that a substantial number of proteins with similar tertiary structures will be missed by any such algorithm. However, it is not clear what fraction of proteins with similar structure but virtually no sequence similarity share a common evolutionary history, and what fraction might be the product of convergent evolution on the structural level³³.

Acknowledgements

I am indebted to E. Charnov, W. Fontana, P. d'Haeseleer, M. Lynch, D. Natvig, and M. Werner-Washburne for discussions on the subject. The financial and computational support of the Santa Fe Institute and of the Albuquerque High Performance Computing Center is gratefully acknowledged.

Table 1: Correlation of fitness effects of knock-out mutations to similarity of duplicated genes. Shown are the estimated correlation coefficients (Pearson, Kendall, or Spearman) between fitness effect and the respective measure of similarity for 45 pairs of paralogous genes. K_a is the expected fraction of non-synonymous nucleotide substitutions per site on the DNA. The hypothesis that the calculated correlation coefficient is significantly different from zero was tested by a randomization assay, and the resulting P -values are given in parentheses for each correlation coefficient. None of the calculated correlation coefficients are significantly different from zero at $P=0.05$.

	sequence similarity		
	% Identity	Z-Score	K_a
Pearson	0.24 (0.12)	0.08 (0.60)	-0.21 (0.16)
Spearman	0.23 (0.11)	0.08 (0.39)	-0.19 (0.1)
Kendall	0.19 (0.13)	0.06 (0.87)	-0.23 (0.12)

	similarity of expression patterns		
	Pearson	Spearman	Kendall
Pearson	0.32 (0.15)	0.38 (0.09)	0.41 (0.06)

Table 2: Genes highly similar in both sequence and expression pattern. For each gene pair, the gene on chromosome V is listed first. The Pearson correlation coefficient is given to indicate similarity of expression patterns.

Gene Pair	Fitness Class	% Identity	r (Expression Pattern)
<i>RPL23A - RPL23B</i>	2	100	0.96
<i>RPS8A – RPS8B</i>	3	100	0.83
<i>RPS26A - RPS26B</i>	2	98.3	0.97
<i>RPL34A – RPL34B</i>	2	97.5	0.99
<i>HOR2 – GIP2</i>	1	92.4	0.83
<i>RNR1 – RNR3</i>	3	80.3	0.89
<i>RPS24A – RPS24B</i>	3	100	0.93
<i>CYC7 – CYC1</i>	0	85.8	0.83
<i>TIF51A –TIF51B</i>	2	90.4	0.88

Figure Captions

Fig. 1 Sequence similarity vs. fitness effect of knock-out mutations in 45 duplicated yeast genes . **a**, Scatter plot of fitness effect of loss-of-function mutations as obtained by Smith *et al.*¹³ vs. amino acid identity in 45 paralogous yeast genes. A value of “3” corresponds to the most severe reduction in fitness. Shown are also a linear regression line, as well as the Pearson correlation coefficient ($r=0.24$), and its associated significance value ($P=0.12$), as determined by a randomization test. **b**, fitness effect vs. Smith-Waterman Z-score. **c**, Fitness effect vs. the expected fraction of non-synonymous nucleotide substitutions per non-synonymous site. Notice that a) and b) use measures of similarity, whereas c) uses a measure of dissimilarity between duplicates. None of the calculated correlation coefficients is significant at $P=0.05$. This also holds for three further measures of divergence, BLAST score, protein distance based on PAM substitution matrices, and Ks, the fraction of non-synonymous nucleotide substitutions per non-synonymous site (not shown).

Fig. 2 Similarity in expression pattern vs. fitness effect of knock-out mutations in 45 duplicated yeast genes. **a**, scatter plot of fitness effect of loss-of-function mutations, as determined by Smith *et al.*¹³ vs. similarity in expression pattern as measured by a Pearson correlation coefficient for each pair of duplicated genes. Open and black triangles represent correlation values that were not significantly greater than zero, and significantly greater from zero, respectively ($P=0.01$). Shown are also two regression lines, one based on all 45 points (dashed), and one based on the correlation coefficients that are significantly greater than zero (solid), as well as the Pearson correlation coefficient between the significantly similar expression patterns and fitness effect ($r=0.32$), and its associated significance value ($P=0.15$). **b**, identical to a), but for a Spearman

rank correlation coefficient to assess similarity of expression pattern. τ , identical to ρ , but for a Kendall rank correlation coefficient for similarity of expression patterns.

Fig. 3 A hypothetical distribution of fitness effects of mutations and how it changes after a genome duplication. See text for details.

Fig. 4 Similar distribution of fitness effects of block-duplicated genes and all other genes on chromosome V. Shown are percentages of genes with a given fitness effect, based on 42 genes within duplicated blocks on chromosome V (white bars), and 206 genes outside duplicated blocks (black bars). Only genes that do not have a fitness effect specific to one of the environments tested were used (ref. 13; Fig. 1 and Table 2). The two distributions are not significantly different ($\chi^2=7.1$ [3 df]; $P>0.05$).

Fig. 5 Genes with weak fitness effect do not have a greater number of related genes in the yeast genome than do genes with strong fitness effects. All genes on chromosome V were divided into two categories, those with the most severe fitness effects when mutated (category 3; white bars, 46 genes), and those with a phenotype indistinguishable from wild-type (category 0, black bars, 94 genes). Numbers shown above bars are percentages of genes on chromosome V with the number of related genes in the yeast genome shown on the x-axis. Numbers of similar genes are based on a BLAST search with a cut-off score of $E=0.01$. The distributions are significantly different at $P<0.001$ ($\chi^2=40.9$; $df=4$) but not in the direction predicted if gene duplications are responsible for mutational robustness.

Fig. 6 Genes similar to genes with weak fitness effects and to genes with strong fitness effects do not show systematic differences in their similarities. All genes on chromosome V with similar genes elsewhere in the genome were identified via a BLAST search of chromosome V genes

against the yeast genome, with a non-conservative cut-off score of $E=0.01$ ^{42,48}. Two subsets of these chromosome V genes were then analyzed separately, those with the most severe fitness effects when mutated (category 3; grey bars, 28 genes), and those with a phenotype indistinguishable from wild-type (category 0, black bars, 46 genes). **a**, and **b**, show the distribution of BLAST similarity scores and the distribution of Pearson correlation coefficients of expression patterns in a comparison of each chromosome V gene to its most similar gene elsewhere in the yeast genome. Amino acid similarity or Smith-Waterman Z-scores could not be meaningfully analyzed here, because many of the genes were similar only over very short parts of their ORFs (<20 amino acids). The distributions of sequence similarity scores are statistically indistinguishable ($\chi^2=10.3$ [5 df]; $P>0.05$), whereas the distributions of expression pattern similarity are different ($\chi^2=19.6$ [6 df]; $P<0.005$). However, these differences are not in the direction predicted if gene duplications are responsible for robustness, i.e., genes whose closest relatives have highly correlated expression patterns do not have weaker effects on fitness than do other genes.

References

1. Wang, Y. K., Schnegelsberg, P. N. J., Dausman, J. & Jaenisch, R. Functional redundancy of the muscle-specific transcription factors myf5 and myogenin. *Nature* **379**, 823-825 (1996).
2. Saga, Y., Yagi, T., Ikawa, Y., Sakakura, T. & Aizawa, S. Mice develop normally without tenascin. *Genes Dev.* **6**, 1821-1831 (1992).

3. Cadigan, K. M., Grossniklaus, U. & Gehring, W. J. Functional redundancy : the respective roles of the 2 sloppy paired genes in Drosophila segmentation. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 6324-6328 (1994).
4. Gonzalez-Gaitan, M., Rothe, M., Wimmer, E. A., Taubert, H. & Jackle, H. Redundant functions of the genes knirps and knirps-related for the establishment of anterior Drosophila head structures. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 8567-8571 (1994).
5. Hanks, M., Wurst, W., Ansoncartwright, L., Auerbach, A. B. & Joyner, A. L. Rescue of the en-1 mutant phenotype by replacement of en-1 with en-2. *Science* **269**, 679-682 (1995).
6. Hoffmann, F. M. Drosophila-abl and genetic redundancy in signal transduction. *Trends Genet.* **7**, 351-356 (1991).
7. Dun, R. B. & Fraser, A. S. Selection for an invariant character - 'vibrissa number' - in the house mouse. *Nature* **181** (1958).
8. Rendel, J. M. Canalization of the scute phenotype of Drosophila. *Evolution* **13**, 425-439 (1959).
9. Rendel, J. M. in *Quantitative genetic variation* (eds. Thompson, J. N. & Thoday, J. M.) 139-156 (Academic Press, 1979).
10. Waddington, C. H. *The strategy of the genes* (Macmillan, New York, 1959).
11. Rutherford, S. L. & Lindquist, S. Hsp90 buffers development against genetic variation and could link capacity for morphogenic change with environmental stress. *Mol. Biol. Cell* **9**, 2511-2511 (1998).
12. Kacser, H. & Burns, J. A. The molecular basis of dominance. *Genetics* **97**, 639-666 (1981).
13. Smith, V., Chou, K. N., Lashkari, D., Botstein, D. & Brown, P. O. Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **275**, 464-464 (1997).

14. Tautz, D. Redundancies, development and the flow of information. *Bioessays* **14**, 263-266 (1992).
15. Fell, D. *Understanding the control of metabolism* (Portland Press, Miami, 1997).
16. Dykhuizen, D. E., Dean, A. M. & Hartl, D. L. Metabolic flux and fitness. *Genetics* **115**, 25-31 (1987).
17. Hartl, D. L., Dykhuizen, D. E. & Dean, A. M. Limits of adaptation : the evolution of selective neutrality. *Genetics* **111**, 655-674 (1985).
18. Dykhuizen, D. & Hartl, D. L. Selective neutrality of 6pgd allozymes in Escherichia coli and the effects of genetic background. *Genetics* **96**, 801-817 (1980).
19. Edwards, S. & Palsson, B. O. Systems properties of the Haemophilus influenzae rd metabolic genotype. *J. Biol. Chem.* **274**, 17410-17416 (1999).
20. Wagner, A. Does evolutionary plasticity evolve? *Evolution* **50**, 1008-1023 (1996).
21. Nadeau, J. H. & Sankoff, D. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* **147**, 1259-1266 (1997).
22. Seoighe, C. & Wolfe, K. H. Extent of genomic rearrangement after genome duplication in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4447-4452 (1998).
23. Li, X. L. & Noll, M. Evolution of distinct developmental functions of 3 Drosophila genes by acquisition of different cis-regulatory regions. *Nature* **367**, 83-87 (1994).
24. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705 (1998).
25. DeRisi, J. L., Iyer, V. R. & Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686 (1997).
26. Spellman, P. T. *et al.* Comprehensive identification of cell-cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* **9**, 3273-3297 (1998).

27. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708-713 (1997).
28. Thatcher, J. W., Shaw, J. M. & Dickinson, W. J. Marginal fitness contributions of nonessential genes in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 253-257 (1998).
29. Kimura, M. *The neutral theory of molecular evolution* (Cambridge University Press, Cambridge, 1983).
30. Seoighe, C. & Wolfe, K. H. Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.* **2**, 548-554 (1999).
31. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901-906 (1999).
32. Hubbard, T. J. P., Ailey, B., Brenner, S. E., Murzin, A. G. & Chothia, C. Scop ; Structural classification of proteins database : applications to evaluation of the effectiveness of sequence alignment methods and statistics of protein structural data. *Acta Crystallogr. D* **54**, 1147-1154 (1998).
33. Doolittle, R. F. Convergent evolution : the need to be explicit. *Trends Biochem. Sci.* **19**, 15-18 (1994).
34. Galperin, M. Y., Walker, D. R. & Koonin, E. V. Analogous enzymes : independent inventions in enzyme evolution. *Genome Res.* **8**, 779-790 (1998).
35. Wagner, G. P., Booth, G. & Bagherichaichian, H. A population genetic theory of canalization. *Evolution* **51**, 329-347 (1997).
36. Wagner, A. Redundant gene functions and natural selection. *J. Evol. Biol.* **12**, 1-16 (1999).
37. Wagner, A. The role of pleiotropy, population size fluctuations, and fitness effects of mutations in the evolution of redundant gene functions. *Genetics (in press)* (2000).
38. Bradshaw, V. A. & McEntee, K. Dna damage activates transcription and transposition of yeast Ty retrotransposons. *Mol. Gen. Genet.* **218**, 465-474 (1989).

39. Paquin, C. E. & Williamson, V. M. Temperature effects on the rate of Ty transposition. *Science* **226**, 53-55 (1984).
40. Sokal, R. R. & Rohlf, F. J. *Biometry* (Freeman, New York, 1981).
41. Waterman, M. S. General methods of sequence comparison. *Bull. Math.Biol.* **46**, 473-500 (1984).
42. Altschul, S. F. *et al.* Gapped Blast and Psi-Blast : a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).
43. Dayhoff, M., Schwartz, R.M., Orcutt, B.C. in *Atlas of protein sequence and structure* (ed. Dayhoff, M.) 345-352 (National Biomedical Research Foundation, Silver Spring, 1978).
44. Pearson, W. R. Searching protein-sequence libraries : comparison of the sensitivity and selectivity of the Smith-Waterman and Fasta algorithms. *Genomics* **11**, 635-650 (1991).
45. Henikoff, S. & Henikoff, J. G. Amino-acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 10915-10919 (1992).
46. Felsenstein, J. PHYLIP - (Phylogeny inference package) version 3.2. *Cladistics* **5**, 164-166 (1989).
47. Comeron, J. M. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.* **41**, 1152-1159 (1995).
48. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410 (1990).