Robustness against mutations in genetic networks of yeast

Andreas Wagner

There are two principal mechanisms that are responsible for the ability of an organism's physiological and developmental processes to compensate for mutations. In the first, genes have overlapping functions, and loss-of-function mutations in one gene will have little phenotypic effect if there are one or more additional genes with similar functions. The second mechanism has its origin in interactions between genes with unrelated functions, and has been documented in metabolic and regulatory gene networks. Here I analyse, on a genome-wide scale, which of these mechanisms of robustness against mutations is more prevalent. I used functional genomics data from the yeast *Saccharomyces cerevisiae* to test 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. My results demonstrate that interactions among unrelated genes are the major cause of robustness against mutations. This type of robustness is probably an evolved response of genetic networks to stabilizing selection.

Introduction

Both physiological and developmental processes of eukaryotes display considerable robustness against the effects of mutations. For instance, many loss-of-function mutations of developmental genes in higher organisms show no or a weak phenotypic effect[1–6]. Several morphological traits of *Drosophila melanogaster* show a lack of variation, despite a considerable amount of underlying genetic variation[7–10]. A study linking the heat-shock protein HSP90 to the buffering of developmental processes against genetic variation provides one potential molecular mechanism for this phenomenon[11]. On the cellular level, the fact that most loss-of-function mutations at enzymatic loci are recessive demonstrates the ability of biochemical pathways to compensate for changes in gene dosage[12]. Results from systematic null mutations of all genes on chromosome V of *S. cerevisiae* show that almost 40% of all yeast genes have little or no detectable effect on growth rate in 5 different environments[13].

Mechanisms for robustness. There are two principal mechanisms for such resilience against mutations. The first is derived from overlapping gene functions, in which mutations in one gene have little effect if there are one or more additional genes with similar functions. Evidence from null mutations in developmental genes supports the importance of this mechanism, stemming from considerable similarity between the gene mutated and 'back-up' genes, and many functional studies of such genes[1–6,14]. Because most eukaryotic genes have duplicates in the same genome (60% in the case of yeast), and because many of these duplicates have redundant functions, it seems plausible that redundancy among duplicates is the main mechanism for resilience against mutations. There is, however, a second possible mechanism based on interactions among genes encoding proteins with unrelated biochemical functions. It is best illustrated for the case of recessive mutations in genes encoding enzymes that are completely unrelated in function, that is enzymes that catalyse different chemical reactions, contribute to a reaction chain or network whose goal is to sustain an optimal flux of metabolites[15]. 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 routes[19]. Further potential examples include the buffering of developmental processes against genetic variation in *Drosophila* via the heat-shock protein HSP90 (ref. 11), as well as modelling work suggesting that such regulatory networks can provide physiological resistance to mutational effects without duplicate genes of similar function[20].

Which mechanism is prevalent on a genome-wide scale? Both mechanisms of robustness are well documented. How can we determine which contributes more to compensating for the phenotypic effects of mutations? The answer lies in the only aspect that distinguishes the two scenarios: the relationship between phenotypic effects of a mutation and 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 almost certainly diverge in their functions, as well as in their sequences and expression patterns. As long as their functions overlap to some extent, phenotypic effects of loss-of-function mutations will be weak, which appears to be the case for the many partially redundant genes involved in vertebrate development[1–6]. Ultimately, original and duplicate genes will diverge completely, and mutational effects will be more severe. Sequence similarity is one indicator of functional similarity among genes. Another indicator is similarity in their spatiotemporal expression patterns. Although expression

Department of Biology, University of New Mexico, and The Santa Fe Institute, Albuquerque, New Mexico, USA. Correspondence should be addressed to A.W. (e-mail: wagnera@unm.edu).

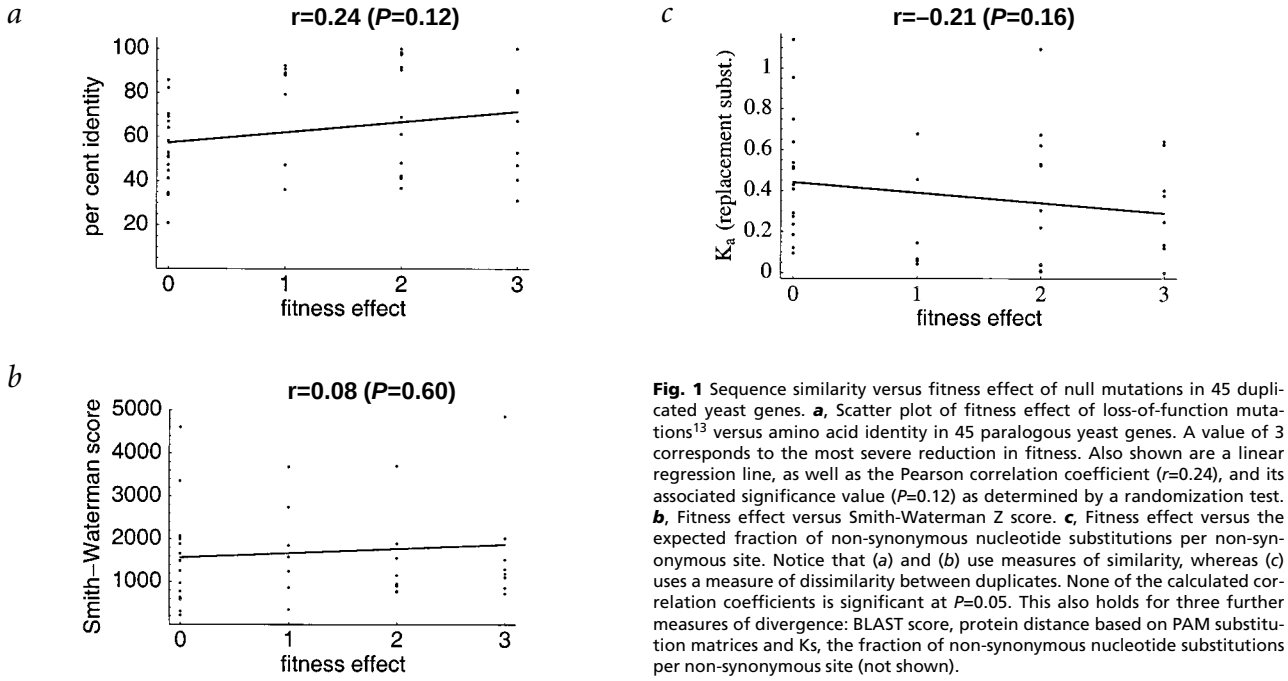

Fig. 1 Sequence similarity versus fitness effect of null mutations in 45 duplicated yeast genes. ***a***, Scatter plot of fitness effect of loss-of-function mutations[13] versus amino acid identity in 45 paralogous yeast genes. A value of 3 corresponds to the most severe reduction in fitness. Also shown are 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 versus Smith-Waterman Z score. ***c***, Fitness effect versus 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).

patterns can be similar for reasons unrelated to evolutionary history, there are several cases of overlapping gene functions in vertebrates in which expression pattern is a better indicator of functional divergence[1,23] than sequence. 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 *S. cerevisiae*, data have become available that allow a test of this prediction. First, both the complete genome sequence and microarray expression data are available for several important cellular processes[24–26]. Second, yeast underwent a genome duplication approximately 100 million years ago (Mya), the remnants of which are 55 syntenic pairs of blocks of duplicated genes[27]. Although these are only a small fraction of all duplicated yeast genes, the identical timing of duplication distinguishes these from other duplicated genes. Third, Ty1–induced loss-of-function mutations in each of 255 genes on yeast chromosome V have been generated[13]; the growth rate of the resulting 255 strains has been determined 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 event[27].

Table 1 • Correlation of fitness effects of null mutations to similarity of duplicated genes

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)

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.

Are dispensable genes really dispensable? Before testing this prediction, a brief discussion of the null-mutant data used here is necessary. These data were generated by a liquid-culture competition assay that determined growth rate (to approximately 5% resolution in 5 different environments) and mating efficiency[13]. For any of the genes whose null mutant has no detectable phenotype, the question arises whether this is due to the limited number of environments and fitness components assayed. Although one cannot 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% had growth defects specific to one environment[13]. This suggests that for most genes a mutant phenotype should manifest itself under most conditions, although a rigorous assessment of this hypothesis must await the availability of more data. More problematic is the limited resolution of growth-rate differences (5%) between null mutants and wild type. A recent study increased the experimental resolution of growth-rate differences by one order of magnitude, but it was still not possible to detect growth differences that might have resulted from 20% of the null mutations. It is possible, however, that even slight growth differences might affect the persistence of a phenotype in the population. Whether they do depends on the effective population size[29] of yeast, which is unknown.

Although there may be no truly dispensable genes, it is worth keeping in mind that synthetic null mutations are more disruptive than most point mutations, insertions and deletions to which genes are natually exposed. Results from null-mutant studies are thus best used as indicators for the effects that these more subtle mutations have on fitness. They suggest that there is a wide range of fitness effects, with many mutations having very slight effects.

Results

Phenotypic effect of mutation 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 null mutations and the similarity in sequence and expression to other yeast genes of the various genes that were mutated . First I

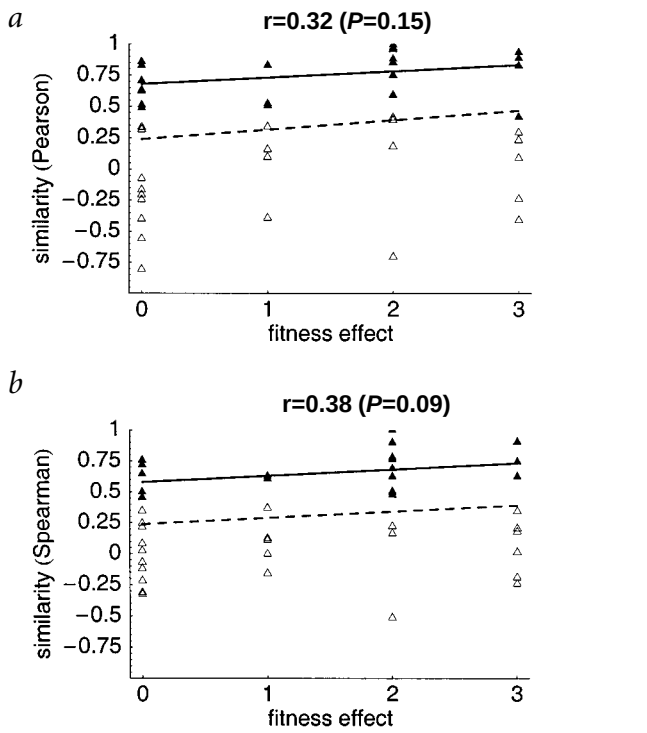

Fig. 2 Similarity in expression pattern versus fitness effect of null mutations in 45 duplicated yeast genes. ***a***, Scatter plot of fitness effect of loss-of-function mutations[13] versus similarity in expression pattern as measured by a Pearson correlation coefficient for each pair of duplicated genes. Open and filled triangles represent correlation values that were not significantly greater than zero, and significantly greater from zero, respectively (P=0.01). Also shown are two regression lines, one based on all 45 points (dashed) and one based on the correlation coefficients that are significantly greater than zero (filled), 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. ***c***, Identical to (*a*), but for a Kendall rank correlation coefficient for similarity of expression patterns.

assessed the relationship between fitness effects and sequence similarity for the 45 duplicated genes on chromosome V (ref. 27); each of the 45 genes has a paralogue in a syntenic block of genes on another yeast chromosome[27]. Data were available on the effect of null mutations for each of the 45 genes in the duplicated blocks[13]. The growth-rate effects of null mutations were determined for each of 255 genes on chromosome V, and the genes were divided into 4 categories. For the purposes 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. The results were nearly identical regardless of the similarity measure used (Fig. 1 and Table 1); genes whose null mutation had little or no fitness effect did not have paralogues with more similar sequences than did genes whose null mutation had severe fitness effects.

Table 2 • Genes highly similar in both sequence and expression pattern

Gene pair	Fitness class	% identity	r (expression pattern)
RPL23A, RPL23B	2	100	0.96
RPS8A, RPS8B	3	100	0.83
RPS26A, RPS26B	2	98.3	0.97
RPL34A, RPL34B	2	97.5	0.99
HOR2, GIP2	1	92.4	0.83
RNR1, RNR3	3	80.3	0.89
RPS24A, RPS24B	3	100	0.93
CYC7, CYC1	0	85.8	0.83
TIF51A, TIF51B	2	90.4	0.88

For each gene pair, the gene on chromosome V is listed first. The Pearson correlation coefficient is given to indicate similarity of expression patterns.

Phenotypic effect of mutation and similarity in expression pattern of duplicate genes

To assess whether differences among expression patterns might be better indicators of 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, sporulation and the diauxic shift[24–26] (the switch from anaerobic to aerobic metabolism upon depletion of glucose). 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 patterns and the fitness effects of mutations for the 45 genes in the duplicated block. The results (Fig. 2 and Table 1) showed that there was no significant association between severity of fitness effect and similarity in expression pattern.

But perhaps both similarity in sequence and expression pattern must be high 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 (Table 2). Although one cannot draw statistically sound conclusions from such a small number of genes, the percentage of genes (7/9, 77%) whose mutation leads to a detectable fitness effect is in fact greater than that for the remainder of the 45 genes (19/36, 53%). This is partly because the nine genes (Table 2) include several ribosomal genes, for which duplicates are often highly conserved in both sequence and expression pattern, yet show detectable fitness defects when mutated.

Duplicate genes outside duplicate gene blocks

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

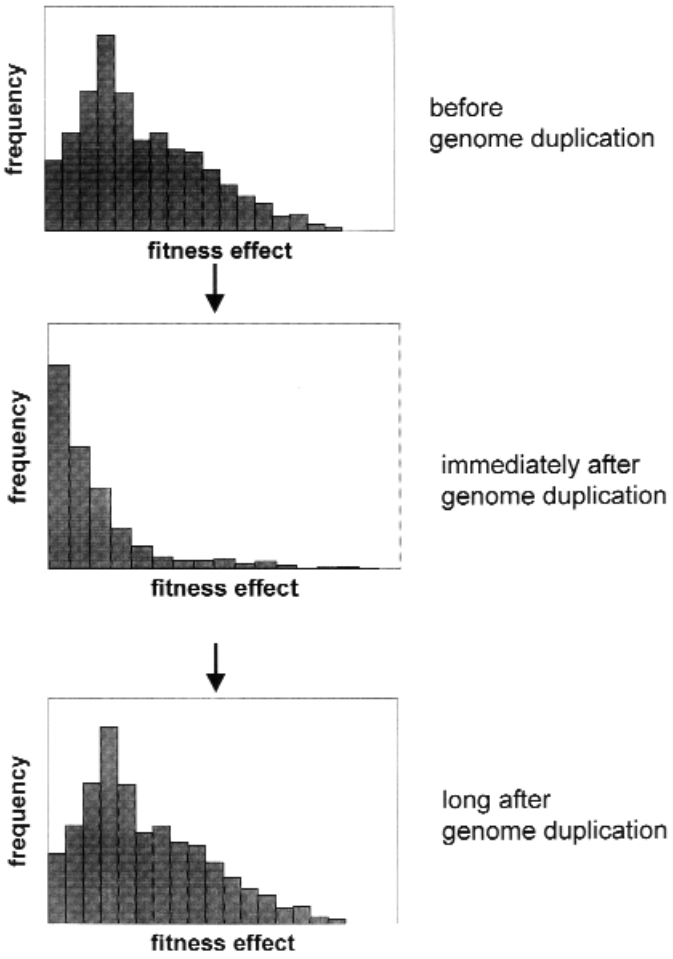

Fig. 3 Fitness effect distribution of loss-of-function mutations. The graphs show a hypothetical distribution of fitness effects of mutations and how it changes after a genome duplication.

Many mutant genes with a weak or no effect have no paralogues

In the last approach, I assigned all genes on chromosome V to one of two groups, those whose null mutation had no detectable fitness effects (94 genes) and those whose null mutation had the most severe fitness effects (46 genes). I then determined whether there were any differences between the two sets of genes with regard to the following: (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. Criterion (i) was specifically motivated by the possibility that not just one similar gene but a whole gene family might be responsible for resilience to the effects of mutations. The distribution of the number of genes with detectable similarity to chromosome V genes in the two categories is shown (Fig. 5). Of chromosome V genes with no detectable fitness effect, 41 (43.6%) also had no similar genes in the yeast genome. This result is not an artefact of too stringent a criterion to determine similarity among genes. Overall, the two distributions (Fig. 5) are significantly different (χ^2=40.9; d.f.=4; P<0.001), but the differences were not in the direction one would predict if the fitness effects of a mutation were correlated with gene-family size.

Next I determined whether paralogues of chromosome V genes with weak fitness effects displayed greater sequence similarity (Fig. 6*a*) or greater similarity in expression pattern than did paralogues of genes with strong fitness effects (Fig. 6*b*). Where there were differences in the distribution of similarity, they were not consistent with the notion that greater similarity is associated with weaker fitness effects (Fig. 6). Qualitatively identical results (data not shown) were obtained when similarity data from the five or ten most closely related genes were averaged and analysed.

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, in which mutations in some genes cause severe fitness defects, whereas other genes might be dispensable (Fig. 3). 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 Myr elapsed since the yeast genome duplication[27]. This raises the question of whether the many weak fitness effects observed in yeast are remnants of the yeast genome-duplication event.

To address this question, one must compare the distribution of mutational effects between a sample of gene pairs retained from the genome duplication and a reference set of genes not included 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 8 of them are expected to have paralogous partners that have been retained after the genome duplication[22]. I compared the distribution of fitness effects for mutations in all duplicated blocks on chromosome V with the same distribution for all genes outside the duplicated blocks (Fig. 4), and found them to be statistically indistinguishable. This suggests that the small effects of many of the mutated genes are not primarily a result of the genome duplication.

Discussion

When compared with genes whose loss of function results in severe fitness defects, genes whose loss of function results in a weak or no fitness effect are not more similar to their closest paralogues, both in sequence and temporal expression pattern. They are also not part of larger gene families whose members are, on average, more similar in sequence or expression to the gene mutated. Furthermore, they are not related to other yeast genes in about one-half of the cases studied. These cases include 45 paralogous gene pairs which also 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. Thus, although gene duplications may be responsible for a fraction of weak null-mutation phenotypes[30], they contribute little to mutational robustness on a genomic scale.

Caveats. The data available for this analysis have limitations that should be overcome by future improvements in the underlying technology and analysis methods. First, in the 'genetic footprinting' approach to gene disruptions[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]

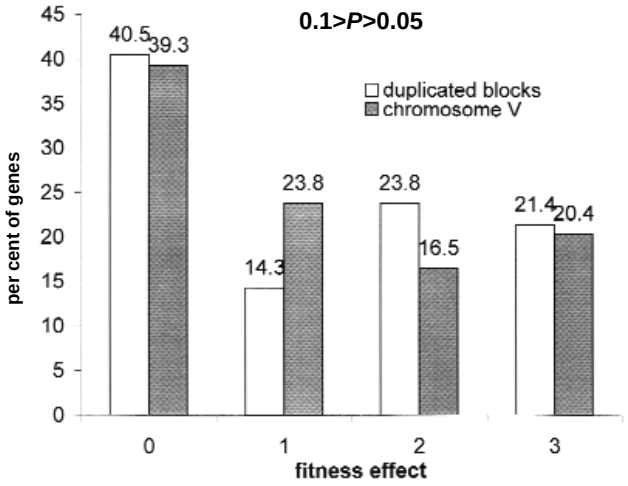

Fig. 4 Similar distribution of fitness effects of block-duplicated genes and all other genes on chromosome V. Shown are percentages of genes with a given fitness effect, based on 42 genes within duplicated blocks on chromosome V (open bars) and 206 genes outside duplicated blocks (filled bars). Only 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 (χ^2=7.1; 3 d.f.; P>0.05).

using different gene-disruption techniques have yielded a distribution of mutant effects consistent with that analysed 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. My results 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 limitations of such data are the considerable amount of noise in large-scale mRNA expression data, the neglect of spatial information, as well as of translational and post-translational regulation, and the limited ability of the microarray approach[26] to distinguish among mRNAs produced by closely related members of gene families.

Convergent evolution? Approximately 40% of all genes with no detectable fitness effect also had no related 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 structures[32]. It is not clear, however, 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 level[33].

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 so, at least one back-up gene with identical biochemical function, but completely unrelated sequence and perhaps even unrelated structure, should exist, for each dispensable gene. Such back-up genes may exist, but are probably rare[34]. Extrapolating the results from chromosome V to the rest of the genome suggests the existence of approximately 1,000 genes in this category (38% of genes with little or no fitness effect, 43% of which are unique). This indicates the enormous scale on which convergent evolution would have to occur. Even if functional convergence is more frequent than currently thought the question arises as to why convergence would occur on such a massive scale. The answer may have to do with network resilience provided by functional redundancy.

Robustness, intrinsic or evolved, beneficial or detrimental? A broad distribution of mutational effects may be either an intrinsic and unchangeable feature of many genetic networks or an evolved property. This has been studied empirically and with mathematical models in a variety of contexts, including the evolution of dominance in enzymatic genes, 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 that robustness is an evolutionary response of genetic systems to stabilizing selection for either mutational stability or stability against environmental fluctuations.

Increased robustness against mutations in genetic networks would evolve via an indirect mechanism in which robust networks do not confer higher fitness on their carrier, but accumulate in a population because mutations within networks are less likely to have deleterious effects. As a consequence, enhanced 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 within 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 the selection response of a population depends on the amount of genetic variation expressed phenotypically, and robustness reduces that amount. One would thus expect that organisms have evolved not only mechanisms to increase the production of genetic variation (for example, through elevated transposon activity[38,39]), but also mechanisms to augment the expression of existing genetic variation during such times[11]. These mechanisms should cause a rather unspecific increase in variation, because the exact kind of variation required cannot be anticipated.

As in the case of metabolic networks, where a full mechanistic understanding of robustness and its evolution required an experimentally testable mathematical theory[15], the mechanistic causes

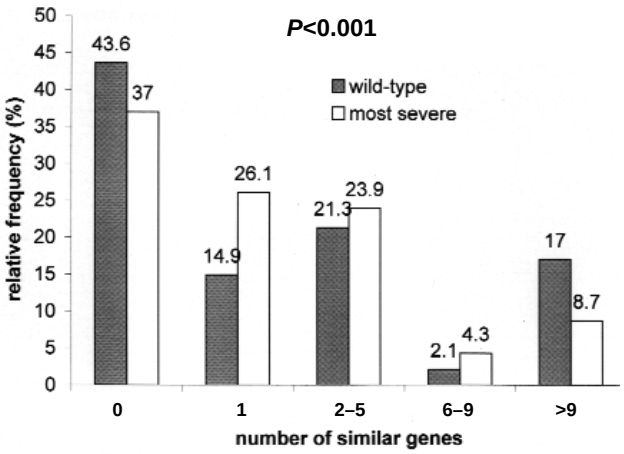

Fig. 5 Genes with weak fitness effect do not have more related genes in the yeast genome than do genes with strong fitness effects. All genes on chromosome V were divided into two categories, those with the most severe fitness effects when mutated (category 3; open bars, 46 genes), and those with a phenotype indistinguishable from wild type (category 0, filled bars, 94 genes). Numbers shown above bars are the 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 (χ^2=40.9; d.f.=4), but not in the direction predicted if gene duplications are responsible for mutational robustness.

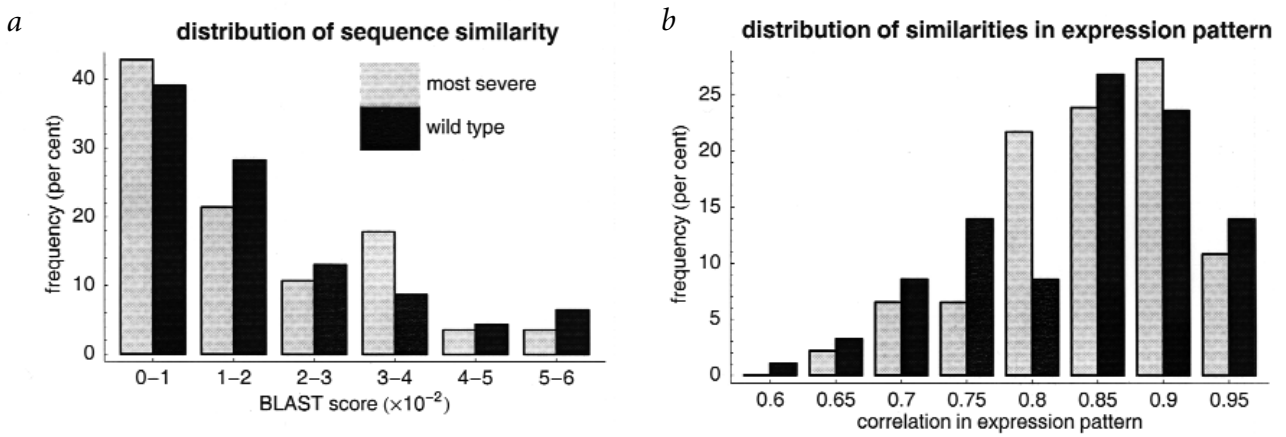

Fig. 6 Genes similar to those with weak fitness effects and to those with strong fitness effects do not show systematic differences in their similarities. All genes on chromosome V with similar genes elsewhere in the genome were identified via a BLAST search of chromosome V genes against the yeast genome, with a non-conservative cut-off score of E=0.01 (refs 42,48). Two subsets of these chromosome V genes were then analysed 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***,***b***, 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 analysed here, because many of the genes were similar over only very short regions of their ORFs (<20 aa). The distributions of sequence similarity scores are statistically indistinguishable (χ^2=10.3; 5 d.f.; $P > 0.05$), whereas the distributions of expression pattern similarity are different (χ^2=19.6; 6 d.f.; $P < 0.005$). These differences are not in the direction predicted if gene duplications are responsible for robustness; that is, genes whose closest relatives have highly correlated expression patterns do not have weaker effects on fitness than do other genes.

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

Methods

Sources of data. I obtained amino acid sequences for all *S. cerevisiae* ORFs from the *Saccharomyces* genome database (ftp://genome-ftp.stanford.edu/pub/yeast/yeast_ORFs), and publicly available microarray expression data (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 the diauxic shift (7 time points; ref. 25), sporulation (7 time points; ref. 24) and 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, then pooled all three data sets for analysis.

There is evidence for a past duplication of the yeast genome[27], including the detection of 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 are available at 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 21–100% (mean 63.3%).

Fitness effects of null mutation by Ty1-transposon insertion into each gene on chromosome V have been described[13]. The fitness of mutants was measured via differences in growth rates between wild-type cells in liquid culture under five different environmental conditions, and assessing the mating efficiency of mutants. The technique of genetic footprinting allowed a resolution of 5% fitness difference between mutant and wild type[13]. Fitness effects of null mutations were grouped 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 categories 0, 1, 2 and 3 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 null mutations showed growth defects specific to only one of the growth conditions used. This was also true for the 45 genes in the duplicated blocks on chromosome V (that is, only 3 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, and 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[13] (category 2).

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 Spearmann's rank correlation coefficient[40]. Pearson correlation coefficients, although easy to interpret, are sensitive to deviations from a normal distribution and are not suitable to measure nonlinear associations.

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 distribution[40]. 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 $x=(x_1, \ldots, x_n)$ and $y=(y_1, \ldots, y_n)$ 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 $x^{(1)}$ of one of the time series by randomly shuffling all entries of the array x, and then calculated a Pearson correlation coefficient $r^{(1)}$ for the shuffled time series $x^{(1)}$ and y, and compared its magnitude with 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 kP times. For the results presented here, k=1000 and 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: percentage amino acid identity between two genes, Z scores obtained from a Smith-Waterman alignment[41], BLAST bit scores[42],

protein distances based on PAM amino acid substitution matrices[43] 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 analysed 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 an analysis tool (1997; http://acer.gen.tcd.ie/~khwolfe/yeast) that uses the packages SSEARCH 3.0 (ref. 44) and BLAST (ref. 42), both based on the BLOSUM62 similarity scoring matrix[45], and a filter to eliminate low-complexity regions of a protein[42]. I calculated protein distances, which estimate the expected fraction of amino acids changed, with the phylogenetic analysis package PHYLIP (ref. 46). A described method and software tool[47] 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 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 proteins[42]. For search (i), I considered only ORFs that had at least one matching fragment of length greater than 50 aa, with at least 50% amino acid similarity to the query sequence on chromosome V, and for which microarray expression data for both the query sequence and 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 conditions[13], and employed a cut-off expect (E)-value[42] of 10^{-2}, corresponding to an expected number of 0.01 false-positive matches per query sequence for the queried database. 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 databases[32] suggests that a substantial number of proteins with similar tertiary structures will be missed by any such algorithm. It is not clear, however, 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 level[33].

Acknowledgements

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

Received 19 July 1999; accepted 23 February 2000.

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. USA* **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. USA* **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**, 1018–1019 (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, New York, 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, 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. USA* **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. USA* **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 Biol. Crystallogr.* **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* **154**, 1389–1401 (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. USA* **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).