

might result in side effects. Alternatively, if the tumor suppressor activity of fumarate hydratase is indeed not due to the cellular concentration of its substrate but instead due to changes in another function, such as altered interaction with another pathway, identification of the other function or pathway might lead to a method to treat the tumors without having to replace the fumarate hydratase protein.

It is possible that many other proteins also have additional functions that have yet to be found. Characterization of a novel protein generally involves finding a function for a protein, but does not necessarily include a search for all possible additional functions of a protein. There is currently no general straightforward method to identify which proteins encoded by a genome sequence have multiple functions or for determining whether a protein of interest is a moonlighting protein. Moonlighting might be a common mechanism of communication and cooperation between the many different functions and pathways within a complex modern cell or between different cell types within an organism. When we try to determine which protein among the thousands encoded in the genome is performing a particular function in the organism, we cannot discount proteins for which an unrelated function is already known; many more proteins might be found to moonlight.

References

- Piatigorsky, J. (1992) Lens crystallins: innovation associated with changes in gene regulation. *J. Biol. Chem.* 267, 4277–4280
- Kennedy, M.C. *et al.* (1992) Purification and characterization of cytosolic aconitase from beef liver and its relationship to the iron-responsive element binding protein. *Proc. Natl. Acad. Sci. U. S. A.* 89, 11730–11734
- Chu, E. *et al.* (1991) Autoregulation of human thymidylate synthase messenger RNA translation by thymidylate synthase. *Proc. Natl. Acad. Sci. U. S. A.* 88, 8977–8981
- Barker, D.F. and Campbell, A.M. (1981) Genetic and biochemical characterization of the *birA* gene and its product: evidence for a direct role of biotin holoenzyme synthetase in repression of the biotin operon in *Escherichia coli*. *J. Mol. Biol.* 146, 469–492
- Ostrovsky de Spicer, P. *et al.* (1993) PutA protein, a membrane-associated flavin dehydrogenase, acts as a redox-dependent transcriptional regulator. *Proc. Natl. Acad. Sci. U. S. A.* 90, 4295–4298
- Jeffery, C.J. (1999) Moonlighting proteins. *Trends Biochem. Sci.* 24, 8–11
- Jeffery, C.J. (2002) Multifunctional proteins: examples of gene sharing. *Ann. Med.* 35, 28–35
- Wool, I.G. (1996) Extraribosomal functions of ribosomal proteins. *Trends Biochem. Sci.* 21, 164–165
- Chung, S.M. *et al.* (1999) Yeast ortholog of the *Drosophila* crooked neck protein promotes spliceosome assembly through stable U4/U6.U5 snRNP addition. *RNA* 5, 1042–1054
- Ben-Yehuda, S. *et al.* (2000) Genetic and physical interactions between factors involved in both cell cycle progression and pre-mRNA splicing in *Saccharomyces cerevisiae*. *Genetics* 156, 1503–1517
- Russell, C.S. *et al.* (2000) Functional analyses of interacting factors involved in both pre-mRNA splicing and cell cycle progression in *Saccharomyces cerevisiae*. *RNA* 6, 1565–1572
- Zhu, W. *et al.* (2002) Evidence that the pre-mRNA splicing factor Clf1p plays a role in DNA replication in *Saccharomyces cerevisiae*. *Genetics* 160, 1319–1333
- Cascalho, M. *et al.* (1998) Mismatch repair co-opted by hypermutation. *Science* 279, 1207–1210
- Darwiche, N. *et al.* (1999) Characterization of the components of the putative mammalian sister chromatid cohesion complex. *Gene* 233, 39–47
- Wu, R.R. and Couchman, J.R. (1997) cDNA cloning of the basement membrane chondroitin sulfate proteoglycan core protein, bamacan: a five domain structure including coiled-coil motifs. *J. Cell Biol.* 136, 433–444
- Gonzalez, F. *et al.* (2002) Recruitment of a 19S proteasome subcomplex to an activated promoter. *Science* 296, 548–550
- Citron, B.A. *et al.* (1992) Identity of 4a-carbinolamine dehydratase, a component of the phenylalanine hydroxylation system, and DCoH, a transregulator of homeodomain proteins. *Proc. Natl. Acad. Sci. U. S. A.* 89, 11891–11894
- Multiple Leiomyoma Consortium (2002), (2002) Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer. *Nat. Genet.* 30, 406–410
- Astuti, D. *et al.* (2001) Gene mutations in the succinate dehydrogenase subunit SDHB cause susceptibility to familial pheochromocytoma and to familial paraganglioma. *Am. J. Hum. Genet.* 69, 49–54
- Baysal, B.E. *et al.* (2000) Mutations in SDHD, a mitochondrial complex II gene, in hereditary paraganglioma. *Science* 287, 848–851
- Niemann, S. and Müller, U. (2000) Mutations in SDHC cause autosomal dominant paraganglioma, type 3. *Nat. Genet.* 26, 268–270

0168-9525/\$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0168-9525(03)00167-7

Genome Analysis

Evolution of *cis*-regulatory elements in duplicated genes of yeast

Balázs Papp^{1,2}, Csaba Pál^{1,2} and Laurence D. Hurst¹

¹Department of Biology and Biochemistry, University of Bath, Bath, Somerset, UK BA2 7AY

²Department of Plant Taxonomy and Ecology, Eötvös Loránd University, Pázmány Péter Sétány 1/C, Budapest, H-1117, Hungary

An increasing number of studies report that functional divergence in duplicated genes is accompanied by gene expression changes, although the evolutionary mechanism behind this process remains unclear. Our genomic analysis on the yeast *Saccharomyces cerevisiae* shows

that the number of shared regulatory motifs in the duplicates decreases with evolutionary time, whereas the total number of regulatory motifs remains unchanged. Moreover, genes with numerous paralogs in the yeast genome do not have especially low number of regulatory motifs. These findings indicate that degenerative complementation is not the sole mechanism behind

Corresponding author: Laurence D. Hurst (bssldh@bath.ac.uk).

expression divergence in yeast. Moreover, we found some evidence for the action of positive selection on *cis*-regulatory motifs after gene duplication. These results suggest that the evolution of functional novelty has a substantial role in yeast duplicate gene evolution.

Gene expression changes are recognized to be of great importance in the functional divergence of duplicate genes [1,2]. Using microarray data, a recent study revealed an unexpectedly high rate of expression divergence in duplicated genes in *Saccharomyces cerevisiae* [1]. Although it is tempting to assume that the driving force behind regulatory changes was positive selection to acquire new expression patterns at the expense of previous functions, a recent argument brought into question the generality of this idea. Force *et al.* [3] proposed that complementary degeneration of different, genetically independent transcriptional regulatory elements of duplicates might also lead to functional diversification without having to invoke positive selection in the process (see also [4] for a similar neutral scenario). Given that degenerative mutations are more common than beneficial mutations and that many eukaryotic genes have modular regulatory regions, Lynch and Force suggested that subfunctionalization might be the dominant mode of duplicate evolution [5].

In its original formulation, the subfunctionalization scenario suggests that the two daughter genes with different expression patterns partition the expression pattern of the ancestral state without gaining new function. The scenario has subsequently been extended to include several other properties of the gene (Box 1), which might be subject to subfunctionalization. Here we shall concentrate on the evolution of *cis*-regulatory sites in duplicated genes, under the assumption that these sites might be independently mutable subfunctions.

Although generally there is no simple relationship between gene expression patterns and the presence or absence of certain regulatory motifs [6], the well-studied example of Hox1b duplicates in zebrafish suggests that observed degeneration of discrete and complementary *cis*-regulatory elements might underlie the subfunctionalization of expression patterns [7]. By contrast, a recent study comparing sequences of HoxA cluster of different vertebrate lineages failed to provide evidence for the subfunctionalization model [8]. Although a couple of examples are known for the subfunctionalization process of regulatory elements in vertebrates and plants [3,7,9], these being mostly based on comparisons of duplicates with non-duplicated orthologs in outgroup species, the general implications are unclear owing to the limited number of genes investigated. Moreover, the scenario also suggests that with growing population size, the likelihood of subfunctionalization declines [5], whereas the relative importance of advantageous mutations increases. Therefore, it is worth investigating the potential role of subfunctionalization in species with relatively large population size (e.g. yeast). In this study, we attempt to shed light on the general evolutionary mechanism behind expression divergence in the unicellular yeast *S. cerevisiae* by employing a large-scale analysis to test the degenerative complementation model.

Box 1. The duplication–degeneration–complementation (or ‘subfunctionalization’) model

Broadly speaking, the model assumes that after gene duplication, subfunctions of the two copies will be subject of complementary degenerative process. At the end of this process no new functions are gained but rather both genes are required to produce the full complement of functions of the ancestral gene. For the model to work, subfunctions need to be independent, hence most mutations should affect only one. Thanks to advances in molecular genetics, it has now become clear that many genes have multiple, partly overlapping functions encoded by at least partly separate ‘unit characters’ or modules. The model has numerous predictions [3]: (1) Solo copy genes should have more molecular functions when compared to duplicated genes in related species, (2) functional specificity should increase after gene duplication and (3) subfunctionalization is more likely to work in small populations [5].

Initially, the focus was on the regulatory complexity of eukaryotic genes. It was suggested that complementary degenerative mutations in different regulatory elements of duplicated genes can facilitate the initial preservation of both duplicates, thereby increasing long-term opportunities for the evolution of new gene functions [3]. The model has subsequently modified, and now it includes several other modular features of the gene. For example, it has been suggested that subfunctionalization works on protein modules instead of *cis*-regulatory sequences [27,28]. Another possibility is that after gene duplication, alternative splicing will be lost, and both copies will retain complementary transcripts [9]. An untested prediction of the theory is that alternative splicing should be especially frequent among solo copy genes.

Last, it is worth mentioning that subfunctionalization might be a quantitative, rather than a qualitative effect [3,4]. For example, duplicates might reduce their expression level to a level when both duplicates are needed to achieve a given function. Another possibility is that enzymatic activity is reduced in the duplicates as a result of the degeneration process, and these changes are compensated by the double gene dosage provided by the two gene copies.

As genome-wide microarray data are currently not available in any close relative of *S. cerevisiae*, the ancestral expression pattern of duplicates cannot be inferred. The same problem applies to computationally or experimentally identified upstream regulatory elements. However one can circumvent this problem by comparing duplicates of different ages. To make clear the logic behind our study, consider an imaginary gene with three independent upstream regulatory motifs responsible for the whole expression profile of the gene (Fig. 1). Immediately after duplication, two genes with the same set of regulatory motifs are present in the genome. The total number of motifs is six for the two copies and the number of shared motifs is, at the outset, three. If complementary loss of regulatory motifs were the sole mechanism of divergence of expression patterns, then we would expect that both the number of shared motifs and the total number of motifs possessed by the duplicate pair should gradually decline with age (Fig. 1).

To investigate this issue, we compiled a list of independent duplicate pairs in the yeast genome by carrying out BLASTP [10] search of all available *S. cerevisiae* proteins against each other and retaining reciprocal best hits as duplicate pairs (using a cutoff expected value of $E < 10^{-20}$). Overlapping pairs and transposon-containing genes were excluded. This provides 941 pairs of duplicated genes. By this method, we ensured

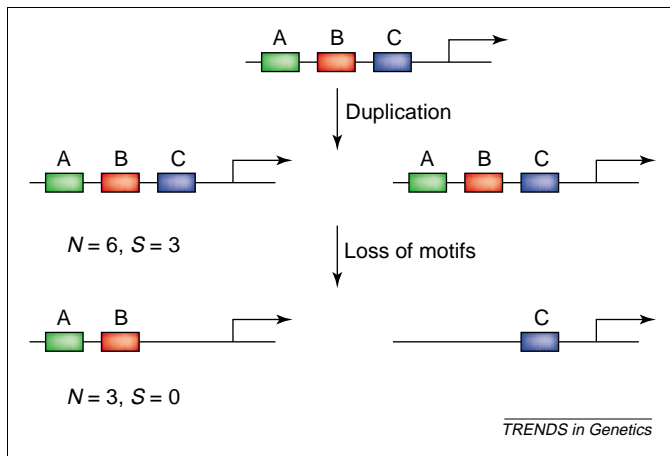


Fig. 1. The degenerative complementation model of evolution in regulatory regions after gene duplication. Abbreviations: A–C, regulatory motifs; N , total number of regulatory motifs in the duplicates; S , number of different motifs shared by the duplicates. Each daughter gene retains only a subset of promoter elements compared to the ancestral state. As a result, both S and N decline with age.

that the genes we find occur only once in our sample. We aligned the protein sequences of duplicated pairs using Clustal-W 1.81 [11] and calculated the number of synonymous (K_S) substitutions per synonymous site [12]. Only duplicates with at least 150 informative codons and $K_S > 0$ were used for the analysis. K_S can be a relatively good indicator of the age of duplicates, as long as most substitutions at synonymous sites are driven by neutral evolution. To control for selectively driven codon-usage bias [13], we considered only duplicates with low bias, measured by the effective number of codons (ENC). Based on the work of Gu and colleagues [14], only duplicates with $ENC > 35$ were considered. Next, we split duplicates into four groups according to their K_S (Fig. 2).

Information on *cis*-regulatory motifs was derived from a non-redundant set of 356 upstream motifs that were previously identified as putative upstream regulatory sites, including 37 experimentally identified motifs [6,15]. These computationally identified motifs, which are overrepresented in the upstream regions of groups of genes belonging to the same functional categories, were

derived by Hughes *et al.* [15] using Gibbs sampling algorithm. A dataset giving information on the presence of these motifs within 600 bp upstream of all known open reading frames (ORFs) in *S. cerevisiae* was downloaded from the Church laboratory website (<http://arep.med.harvard.edu/>). Genes with divergent promoters (i.e. two genes transcribed from one promoter) were excluded from the analysis. Overall, 144 duplicate gene pairs were left with appropriate information available. The average number of motifs in these pairs is 21.3. The number of shared motifs was simply defined for each pair as the number of different motifs shared by the two copies. (Only shared motifs with identical orientations were regarded as shared motifs.) The total number of motifs was calculated for each duplicate pair as the sum of the number of different motifs in each copy.

Conserved synteny of the identified shared motifs in the duplicates provides evidence that the motifs are not the result of convergent evolution, but rather they are truly paralogous. We calculated the number of shared motif pairs (e.g. A and B) with identical ($A \rightarrow B, A \rightarrow B$) and different ($A \rightarrow B, A \leftarrow B$) orientation in the duplicates. Under the assumption that the motifs are truly paralogous, the identical orientation should dominate. This is exactly what we observed (Sign test, $P < 10^{-5}$ for the 48 duplicates with at least four shared motifs).

As predicted by the subfunctionalization model, we find a highly significant negative relationship between the number of shared motifs and K_S ($F = 34.6, df = 140, P < 10^{-16}$, Spearman rank correlation for pairs with $K_S < 2$: $\rho = -0.33, P < 0.005, N = 74$; Fig. 2a). A similar gradual tendency is observed when the fraction of shared motifs is considered (data not shown).

In contrast to the decline of shared regulatory motifs, the total number of regulatory motifs possessed by the duplicates remains constant with age ($F = 0.83, df = 140, P = 0.48$, Spearman rank correlation for pairs with $K_S < 2$: $\rho = -0.037, P = 0.75, N = 74$; Fig. 2b). This last finding is not expected by complementary loss of regulatory motifs. There are some further clues pointing in this direction. The model also predicts that members of

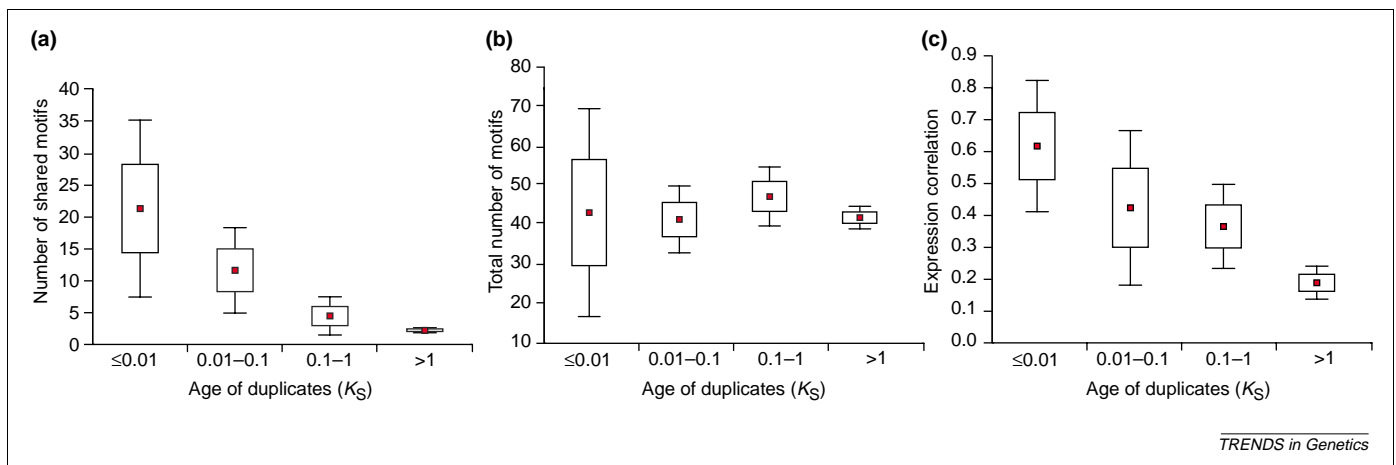


Fig. 2. (a) Negative association between duplicate age (the rate of evolution at synonymous sites, K_S) and the number of different shared motifs of duplicates. (b) No association between duplicate age and the total number of motifs possessed by the two copies. (c) Negative association between K_S and expression similarity. To estimate expression similarity of duplicates, we calculated the Pearson correlation of expression profiles compiled from public whole genome mRNA expression data by the Church laboratory [26]. Red square shows the mean; box shows ± 1 standard error; bars show ± 1.96 standard error.

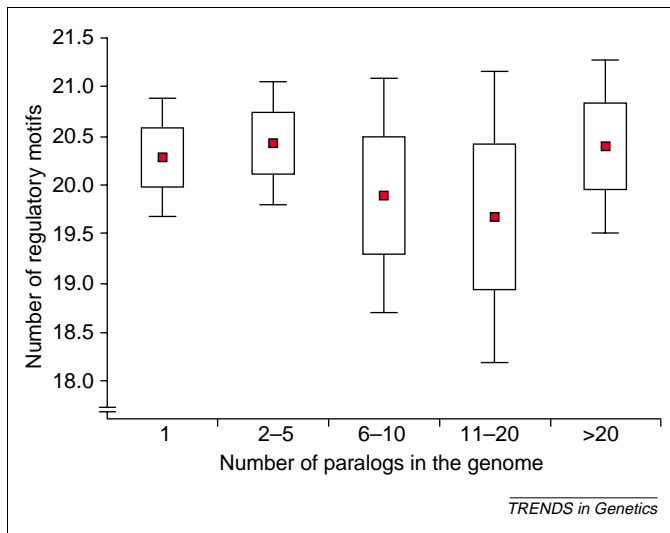


Fig. 3. No relationship between the number of regulatory motifs and the number of paralogs in the yeast genome ($N = 1601$, Spearman rank correlation $\rho = 0.0083$, $P = 0.74$). A BLASTP [10] search of all available *S. cerevisiae* proteins against each other was carried out to identify putative paralogs. The number of paralogs of a given gene was estimated by the number of its BLASTP hits with expected values 10^{-10} or less. Red square shows the mean; box shows \pm one standard error; bars show \pm two standard errors.

multigene families should have especially low number of regulatory motifs, as these genes have undergone multiple rounds of gene duplications and complementary loss of motifs [3]. In contrast to this expectation, genes with numerous paralogs in the yeast genome do not have especially low number of regulatory motifs (Fig. 3).

One potential shortcoming of our analysis is that ignores the possibility that a given transcription factor can have multiple binding sites. More precisely, multiple motifs can specify the binding of a given transcription factor, and the same motif can be involved in the binding of different transcription factors. To control for such a potential bias, we repeated all analyses using a large-scale dataset on regulatory protein–gene interactions in yeast [16]. The dataset includes a systematic, genome-wide location analysis of 106 transcription factors. From the microarray experiments, the authors provided a confidence value for each potential interactions (the cut-off value used here is the same as in [16]). We calculated the total and shared number of binding transcription factors for all pairs of duplicates investigated above. Only gene pairs with appropriate information on transcription factor binding sites were investigated (using a confidence interval as in [16]). As expected, there is a negative association between the age of duplicates and the fraction of shared regulatory proteins (Spearman rank correlation for pairs with $K_S < 2$: $\rho = -0.52$, $P < 0.05$, $N = 17$). We also found that the total number of regulatory proteins of the duplicates remain constant with age (Spearman $\rho = -0.11$, $P = 0.66$). Moreover, there is no association between the number of paralogs in the yeast genome and the total number of regulators (Spearman $\rho = -0.05$, $P = 0.19$, $N = 679$). Hence, these results are in agreement with the previous analysis on *cis*-regulatory motifs, and we can conclude that regulatory regions gradually change in the duplicates, without substantial loss of regulatory binding sites and regulators.

What others mechanisms could give rise to the patterns above? Conservation of expression patterns in the face of rapid turnover of transcription factor binding sites [17] and compensatory mutational changes in other regulatory elements [18] could result in an apparent gradual decline in the number of regulatory motifs that are shared by the two copies. This is a realistic possibility given that evolutionary analysis of transcription factor binding sites suggests that functionally conserved elements often show low sequence similarities [19]. If the divergent motifs are still recognized as motifs (just not the same motif), then we would see that the total number of motifs remains constant over time.

However, this scenario is unlikely to hold. There is positive association between the number of shared regulatory motifs and expression similarity between the duplicates (Spearman rank correlation: $\rho = 0.53$, $P < 10^{-5}$, $N = 71$, for pairs sharing at least 20% of their motifs). In a similar vein, expression similarity and the number of shared transcription factors show positive correlation ($\rho = 0.52$, $P < 10^{-6}$). Moreover, previous work [1] showed that a gradual loss of expression similarity, assayed from microarray data, takes place as duplicates diverge in sequence. We confirm this with our dataset ($F = 4.36$, $df = 133$, $P < 0.01$, Spearman rank correlation for pairs with $K_S < 2$: $\rho = -0.28$, $P < 0.02$, $N = 69$; Fig. 2c). These results suggest that loss of shared regulatory motifs is followed by expression changes.

But why does the total number of motifs remain constant with age? The most parsimonious interpretation of the data suggests that as duplicates age, either regulatory motifs with new function arise from existing ones or the loss of regulatory motifs is balanced by gain of functionally novel motifs. Either way, we need to invoke the gain of function.

As a further support of the idea above, molecular evolution of upstream DNA sequences provides some evidence for the action of positive selection on *cis*-regulatory motifs. We aligned the upstream DNA region of moderately diverged duplicates ($K_S < 0.5$) using DiAlign 2.2.1 [20] (Clustal-W 1.81 [11] gives similar results, data not shown). Next, we compared the frequency of conserved nucleotides at sites with (1) conserved *cis*-regulatory motifs, (2) no regulatory motif and (3) different regulatory motifs. We only investigated gene pairs with $ENC < 35$. As the frequency of conserved nucleotide sites with no underlying motif and K_S show strong correlation with each other (Spearman rank correlation: $\rho = 0.66$, $P = 0.005$, $N = 16$), we have good reason to suppose that these upstream regions are under relaxed purifying selection. We observed that class (1) sites show much higher sequence similarity compared with class (2) sites, (Wilcoxon median test: $P = 0.0052$, $N = 17$), suggesting that conserved *cis*-regulatory regions are under stabilizing selection.

Furthermore, we found four out of 36 duplicated pairs where nucleotide sites with no underlying motif are more conserved than sites with changed motifs (Fisher exact test: $P < 0.05$). After Bonferroni correction, one duplicate pair (*HXT6*–*HXT7*) remains (Fisher exact test: $P < 10^{-3}$).

Table 1. Expression level of duplicated and non-duplicated genes, measured by codon adaptation index (CAI), within the same functional category

Functional category	Mean CAI for genes with no paralog	Mean CAI for genes with at least one paralog	Significance level (Mann-Whitney U-test)
Cell cycle	0.142 (N = 212)	0.164 (N = 238)	$P < 10^{-6}$
Transcription	0.147 (N = 425)	0.170 (N = 345)	$P = 0.005$
DNA processing	0.151 (N = 127)	0.157 (N = 124)	$P = 0.055$
Protein synthesis	0.252 (N = 163)	0.555 (N = 194)	$P < 10^{-18}$
Protein fate	0.165 (N = 279)	0.189 (N = 313)	$P = 0.0074$
Cellular transport and transport mechanisms	0.164 (N = 210)	0.191 (N = 282)	$P = 0.0156$
Transport facilitation	0.187 (N = 70)	0.188 (N = 241)	$P = 0.3$
Cell fate	0.152 (N = 184)	0.177 (N = 243)	$P < 10^{-4}$
Control of cellular organization	0.151 (N = 111)	0.181 (N = 97)	$P = 0.018$
Stress response	0.191 (N = 45)	0.258 (N = 128)	$P = 0.007$
Cell rescue, defense and virulence	0.182 (N = 77)	0.233 (N = 199)	$P = 0.033$
Metabolism	0.194 (N = 413)	0.211 (N = 652)	$P = 0.011$
Energy	0.211 (N = 97)	0.264 (N = 155)	$P = 0.051$

Under the assumption that upstream regions with no underlying motifs are under relaxed selection pressure, the fast evolution of regions with different *cis*-regulatory motifs can best be explained by the action of positive selection on these sites.

Above we have assumed that the degenerative complementation model entails the partition of expression patterns by means of complete loss of regulatory motifs. There might also (or instead) be a quantitative effect (e.g. reduced expression levels) [3,4]. Consider, for example, a new duplicate pair, each of which produces 50 copies of the same gene product, when only a net 50 copies is necessary to fulfill a given function. Fixation of mutations reducing expression levels in both copies could lead to a situation where the two copies produce roughly half of the initial amount (e.g. 24 and 26 copies). Such a model predicts that duplicate genes should have lower expression rates than the single gene had ancestrally. This we cannot test directly, but it is noteworthy that expression levels of duplicated genes are actually higher than those of non-duplicated genes in yeast [21], even within the same functional category (see Table 1; functional categories were derived from MIPS CYGD [22], and expression level was approximated by codon adaptation index, CAI [13]). This result is compatible with the suggestion that selection for high gene dosage was a significant in determining which genes have maintained the duplicate state.

All the above findings indicate that degenerative complementation is not the sole mechanism behind expression divergence in yeast. We must emphasize however, that it is extremely hard to investigate the initial evolutionary stages in gene duplicates. First, if duplication *per se* is not highly advantageous, then the fixation time of gene duplicates is expected to be long in very large populations (such as in yeast). In this case, mutations might subsequently arise, and therefore divergence could occur before fixation of the duplication event itself. It might happen that subfunctionalization proceeds during this very early initial stage leading to the initial preservation of duplicates [23], and new or modified functions arise subsequently. Our work suggests that evolution of functional novelty most probably had a substantial role in yeast duplicate evolution. It would also be wrong to suppose that the patterns observed in

yeast need be representative of those seen elsewhere. The subfunctionalization model is perhaps more likely to work in vertebrates [5], where the population size is fairly low and more-complex regulatory regions exist. Clearly, further large-scale comparative work on gene expression patterns and regulatory sequences is needed to understand the role of drift and darwinian selection in the evolution of gene duplicates.

Last, we also wish to point out an apparent paradox of duplicate evolution. There is a gradual decline of expression [1] and regulatory motif similarity (this work) in duplicates of yeast, suggesting that these changes have an important role in the divergence of gene functions. However, even very distant duplicates can partly compensate mutations in each other [24]. Unexpectedly, we find that there is no clear association between expression similarity and the effects of gene knockouts in yeast (see also [25]) (Fig. 4). These results suggest that partial redundant functions of the duplicates could remain even after substantial sequence and expression changes. The apparent tolerance of the yeast genome to genetic

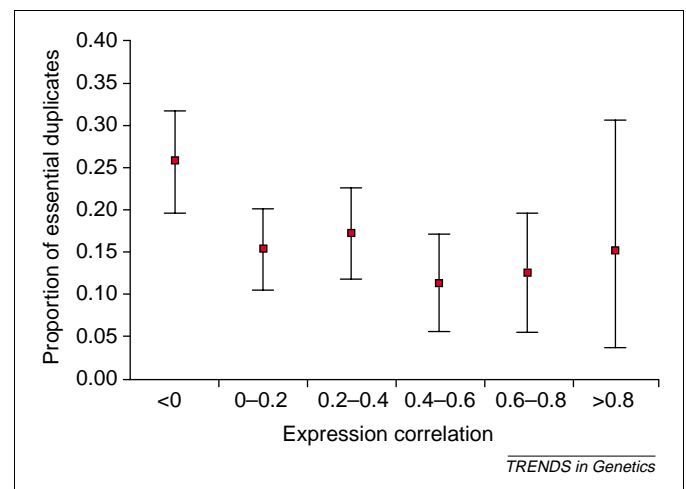


Fig. 4. No strong association between the proportion of duplicate pairs with at least one copy being essential and expression similarity (measured by Pearson correlation) of the pair ($\chi^2 = 13.68$, $df = 5$, $P = 0.018$, $N = 793$). When only gene pairs with positive correlation are considered, there is no significant association between gene dispensability and expression similarity ($\chi^2 = 2.25$, $df = 4$, $P = 0.69$, $N = 611$). Similar results were obtained when quantitative growth of knockouts were considered. Data on gene dispensability were obtained from the *Saccharomyces* Genome Deletion Project (http://www-sequence.stanford.edu/group/yeast-deletion_project/). Confidence intervals were calculated by bootstrap method.

perturbation is surprising, and it clearly deserves further investigations.

Acknowledgements

We thank Yitzhak Pilpel for providing us the upstream sequences previously used for *cis*-regulatory motif identification. B.P. was supported by Marie-Curie traveling fellowship (HPMT-CT-2001-00288), and C.P. was supported by a Nato/Royal Society fellowship and OTKA (Hungarian Scientific Research Fund). L.D.H. is supported by UK BBSRC.

References

- Gu, Z. *et al.* (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* 18, 609–613
- Ferris, S.D. and Whitt, G.S. (1979) Evolution of the differential regulation of duplicate genes after polyploidization. *J. Mol. Evol.* 12, 267–317
- Force, A. *et al.* (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545
- Stoltzfus, A. (1999) On the possibility of constructive neutral evolution. *J. Mol. Evol.* 49, 169–181
- Lynch, M. and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459–473
- Pilpel, Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29, 153–159
- Prince, V.E. and Pickett, F.B. (2002) Splitting pairs: the diverging fates of duplicated genes. *Nat. Rev. Genet.* 3, 827–837
- Chiu, C.H. *et al.* (2002) Molecular evolution of the HoxA cluster in the three major gnathostome lineages. *Proc. Natl. Acad. Sci. U. S. A.* 99, 5492–5497
- Yu, W.P. *et al.* (2003) Duplication, degeneration and subfunctionalization of the nested *synapsin-Timp* genes in Fugu. *Trends Genet.* 19, 180–183
- Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
- Thompson, J.D. *et al.* (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680
- Pamilo, P. and Bianchi, N.O. (1993) Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Mol. Biol. Evol.* 10, 271–281
- Coghlan, A. and Wolfe, K.H. (2000) Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast* 16, 1131–1145
- Gu, Z. *et al.* (2002) Extent of gene duplication in the genomes of *Drosophila*, Nematode, and Yeast. *Mol. Biol. Evol.* 19, 256–262
- Hughes, J.D. *et al.* (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296, 1205–1214
- Lee, T.I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804
- Dermitzakis, E.T. and Clark, A.G. (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* 19, 1114–1121
- Ludwig, M.Z. *et al.* (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* 403, 564–567
- Ludwig, M.Z. (2002) Functional evolution of noncoding DNA. *Curr. Opin. Genet. Dev.* 12, 634–639
- Morgenstern, B. *et al.* (1996) Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. U. S. A.* 93, 12098–12103
- Seoighe, C. and Wolfe, K.H. (1999) Yeast genome evolution in the post-genome era. *Curr. Opin. Microbiol.* 2, 548–554
- Mewes, H.W. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30, 31–34
- Lynch, M. *et al.* (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* 159, 1789–1804
- Gu, Z. *et al.* (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421, 63–66
- Wanger, A. (2000) Robustness against mutations in genetic networks of yeast. *Nat. Genet.* 24, 355–361
- Aach, J. *et al.* (2000) Systematic management and analysis of yeast gene expression data. *Genome Res.* 10, 431–445
- Dermitzakis, E.T. and Clark, A.G. (2001) Differential selection after duplication in mammalian developmental genes. *Mol. Biol. Evol.* 18, 557–562
- Gu, X. (1999) Statistical methods for testing functional divergence after gene duplication. *Mol. Biol. Evol.* 16, 1664–1674

0168-9525/\$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0168-9525(03)00174-4

Genomic analysis of gene expression relationships in transcriptional regulatory networks

Haiyuan Yu^{1,*}, Nicholas M Luscombe^{1,*}, Jiang Qian² and Mark Gerstein¹

¹Department of Molecular Biophysics and Biochemistry, Yale University, PO Box 208114, New Haven, CT 06520-8114, USA

²Wilmer Institute, Johns Hopkins School of Medicine, Baltimore, MD 21287, USA

From merging several data sources, we created an extensive map of the transcriptional regulatory network in *Saccharomyces cerevisiae*, comprising 7419 interactions connecting 180 transcription factors (TFs) with their target genes. We integrated this network with gene-expression data, relating the expression profiles of TFs and target genes. We found that genes targeted by the same TF tend to be co-expressed, with the degree of co-expression increasing if genes share more

than one TF. Moreover, shared targets of a TF tend to have similar cellular functions. By contrast, the expression relationships between the TFs and their targets are much more complicated, often exhibiting time-shifted or inverted behavior. Further information is available at <http://bioinfo.mbb.yale.edu/regulation/TIG/>

An important question in molecular biology is how gene expression is regulated in response to changes in the environment. Previous studies have explored this by making genome-wide measurements of gene expression

* These authors contributed equally to this work.

Corresponding author: Mark Gerstein (mark.gerstein@yale.edu).