Review TRENDS in Genetics Vol.21 No.11 November 2005

Expression divergence between duplicate genes

Wen-Hsiung Li[1], Jing Yang[1] and Xun Gu[2]

[1]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA
[2]Department of Genetics, Development and Cell Biology, Center for Bioinformatics and Biostatistics, Iowa State University, IA 50011, USA

A general picture of the role of expression divergence in the evolution of duplicate genes is emerging, thanks to the availability of completely sequenced genomes and functional genomic data, such as microarray data. It is now clear that expression divergence, regulatory-motif divergence and coding-sequence divergence all increase with the age of duplicate genes, although their exact interrelationships remain to be determined. It is also clear that gene duplication increases expression diversity and enables tissue or developmental specialization to evolve. However, the relative roles of subfunctionalization and neofunctionalization in the retention of duplicate genes remain to be clarified, especially for higher eukaryotes. In addition, the relationship between gene duplication and evolution of transcriptional regulatory networks is largely unexplored.

Introduction

Expression divergence between duplicate genes has long been a subject of great interest to geneticists and evolutionary biologists [1–3] because it is considered an important step in the emergence of a new gene from a redundant duplicate. Investigation of divergence in tissue expression between isozymes [1], which are enzymes encoded by duplicate genes, started in the late 1950s, soon after the advent of protein electrophoresis. Such studies provided examples of how differential tissue expression of duplicate genes can contribute to functional refinement and diversification [1]. From such insights, Ohno [2] proposed that expression divergence is the first step in the functional divergence between duplicate genes and thereby increases the chance of retention of duplicate genes in a genome. Further, advances in molecular biology in the 1970s and 1980s provided a clearer understanding of the molecular basis of expression divergence between duplicate genes. However, before the advent of genomic sequencing and large-scale gene-expression technologies, studies of expression divergence were conducted on a limited number of gene families, thus providing no general picture of how fast and how often expression divergence occurs between duplicate genes in a genome. Fortunately, the availability of completely sequenced genomes and microarray gene expression and other functional genomic data have stimulated many studies on these questions (e.g. [4–6]). Here we review studies that characterized the rate or pattern of expression divergence between duplicate genes at the genomic level, which contributed to our understanding of the role of expression divergence in the evolution of duplicate genes.

Corresponding author: Li, W.-H. (whli@uchicago.edu).
Available online 2 September 2005

Coding-sequence divergence versus expression divergence

The relative roles of coding and non-coding (especially, regulatory) sequence changes in evolution have been a central but controversial issue (e.g. [7]). In this context, a more specific question has been whether coding-sequence divergence and expression divergence between duplicate genes are coupled. At the genomic level, this question was first addressed using data from the budding yeast [4]. Approximately 10% of the duplicate genes in the yeast genome were derived from a whole-genome-duplication (WGD) event that occurred ~100 million years ago (Mya) [8], whereas the remainder were the result of segmental or individual gene duplications. Wagner [4] examined microarray data from yeast genes under five physiological conditions and the sequences of 114 pairs of duplicate genes, and found no significant correlation between protein-sequence divergence and expression divergence, and concluded that protein-sequence divergence and expression divergence are decoupled. However, using synonymous divergence (*Ks*) as a proxy of divergence time between duplicate genes and more extensive data (400 non-overlapping duplicate gene pairs and microarray data of 14 physiological conditions), Gu *et al.* [5] found a positive correlation between expression divergence and *Ks* (Figure 1a). Because many gene pairs with a small *Ks* value were found to show expression divergence, the authors concluded that expression divergence between duplicate genes can occur rapidly. Furthermore, they found a positive correlation between expression divergence and non-synonymous divergence (*Ka*) for duplicates with $Ka \leq 0.30$ (Figure 1b), indicating that expression divergence and protein-sequence divergence are initially correlated. Similarly, using a phylogenetic analysis of yeast duplicate genes, Zhang *et al.* [9] found a positive correlation between expression divergence and the age of duplicate genes. In addition, Conant and Wagner [10] found a positive correlation between sequence divergence and expression divergence between duplicate genes in *Caenorhabditis elegans*.

 0168-9525/$ - see front matter doi:10.1016/j.tig.2005.08.006

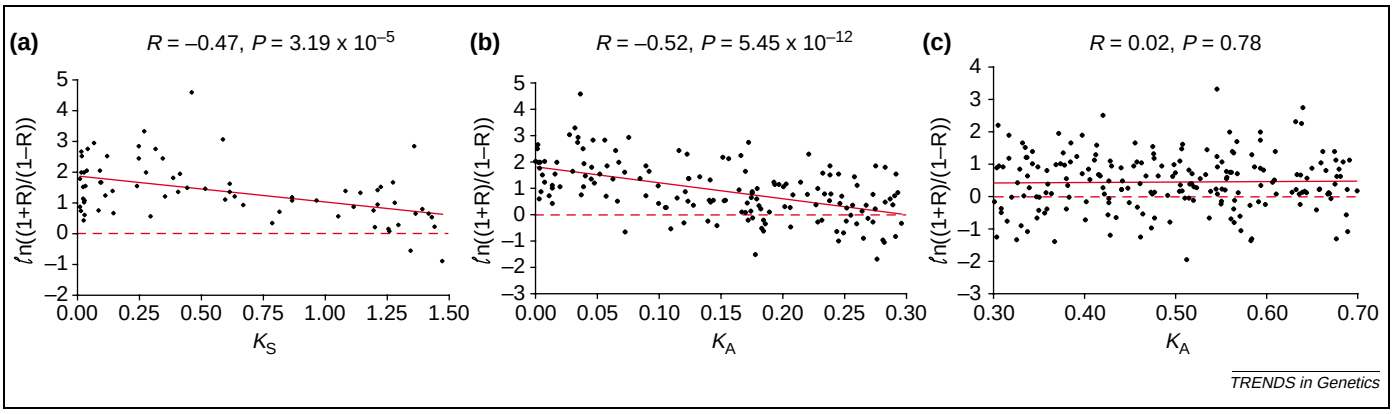

Figure 1. The relationship between the correlation coefficient (R) of gene expression (across all data points) and K_S (K_A) in duplicate genes. K_S and K_A are the numbers of nucleotide substitutions per synonymous site and per non-synonymous site, respectively, between two homogolous genes. Non-overlapping duplicate gene pairs were used in the analysis, so that all data points were independent. **(a)** A significant negative correlation between $\ell n[(1+R)/(1-R)]$ and K_S for gene pairs. This implies a positive correlation between K_S and expression divergence because $1-R$ can be regarded as expression divergence. **(b)** A significant negative correlation between $\ell n[(1+R)/(1-R)]$ and K_A for gene pairs with $K_A \leq 0.3$ (ℓ is the natural logarithm). **(c)** No correlation between $\ell n[(1+R)/(1-R)]$ and K_A for gene pairs with $K_A > 0.3$. This figure was reproduced, with permission, from Ref. [5].

A similar conclusion was reached in an analysis of human duplicate genes. Using Affymetrix expression data in 25 human tissues, Makova and Li [11] found a positive correlation between divergence in tissue expression and *Ks* (or *Ka*). Recently, Blanc and Wolfe [12] studied the expression divergence in 1137 *Arabidopsis* duplicate pairs that are believed to be the remnants of a WGD event that occurred ~24–40 Mya. They observed a weak, but significant, negative correlation between sequence similarity and expression divergence. However, in *C. elegans*, Castillo-Davis *et al.* [13] found that *Ka* and the expression difference between duplicate genes were not significantly correlated; but found a significant correlation between *Ks* and expression difference. Although they attributed the second result to a correlation between *Ks* and divergence in *cis*-regulatory sequence among duplicate genes, we think it suggests that expression divergence between duplicate genes increases with evolutionary time because *Ks* can be used as a proxy of divergence time between duplicate genes.

The *Arabidopsis thaliana* genome is rich in duplicate genes because it appears to have undergone one or more rounds of WGD [14–16]. Using whole-genome microarrays, Kim *et al.* [17] conducted an analysis of the expression of duplicate genes arising from the most recent WGD, which is estimated to have occurred ~20–40 Mya [16]. In particular, they examined the genes associated with oxidative stress response, which is an integral component of the central defense mechanism. The majority of these duplicate genes have diverged in expression, although they have retained a significant amount of coding-sequence conservation. However, in an analysis of 849 pairs of tandem duplicates and 777 pairs of segmental duplicates in *A. thaliana*, Haberer *et al.* [18] found that synonymous divergence (*Ks*) and expression divergence were not correlated; that is, although the *Ks* values for the segmental duplicates have increased by twofold compared with those for the tandem duplicates, the two sets of duplicate genes have similar proportions (19% and 26%) of pairs with a Pearson correlation coefficient of ≥ 0.8 in their expression levels. Therefore, in *A. thaliana,* the relationship between sequence divergence and expression divergence between duplicate genes is not clear. This might be because the duplicate genes in the *A. thaliana* genome were derived from a mixture of large-scale and tandem duplications, so that a definitive conclusion is more difficult to obtain.

Regulatory-motif divergence versus expression divergence between duplicates

In contrast to divergence in coding sequences, divergence in *cis*-regulatory sequences (motifs) between duplicate genes is likely to have a more direct effect on expression divergence. It is therefore interesting to study how *cis*-regulatory motifs diverge between duplicate genes. Papp *et al.* [19] showed that for yeast young duplicates, the number of shared *cis*-regulatory motifs between duplicates decreases with the age of the duplicates, as measured by *Ks* (see more details in the next section). This conclusion, which was derived largely from computationally predicted motifs, was confirmed by Zhang *et al.* [9], who used known yeast regulatory motifs.

Because duplicate genes are likely to share *cis*-regulatory motifs, one would expect a stronger co-expression pattern between duplicate genes than between two randomly selected genes. Zhang *et al.* [9] found that this is indeed the case for yeast genes. It is also reasonable to expect a negative correlation between expression divergence of duplicates and extent of their shared motifs. Zhang *et al.* [9] found a weak negative correlation, but estimated that only a limited amount (~2–3%) of expression patterns of duplicate genes can be explained by their shared *cis*-regulatory motifs. There could be several reasons for this puzzling result. First, the analysis had not included all *cis*-regulatory motifs of the duplicate genes, because *cis*-regulatory motifs are still not completely known for most yeast genes. Second, mRNA stability and chromatin structure might have contributed to the differences in expression level between duplicates. Third, *trans*-factors in the gene network might have influences on the expression divergence between duplicate genes, possibly similar to the situation of allele-specific

expression in diploid organisms. Further studies are needed to see whether any of these reasons are plausible and whether other factors are involved.

In *C. elegans*, *cis*-regulatory motifs are not well studied, so Castillo-Davis *et al.* [13] developed a measure called 'shared motif divergence' (d_{SM}), which was intended to estimate the proportion of the upstream region ($\sim$500 bp) that does not contain the *cis*-regulatory motifs that are shared between two homologous genes. They found a significant correlation (Spearman rank correlation $r=0.47$, $p<0.001$) between d_{SM} and the difference in expression level between duplicate genes.

In summary, there is a positive correlation between expression divergence and *cis*-regulatory motif divergence between duplicate genes, although differences in other factors, such mRNA stability and chromatin structure, might also contribute to expression differences between duplicate genes.

Gene duplication and expression diversification

A classical view is that gene duplication enables duplicates to become specialized in different tissues or developmental stages [1–3,20–22]. Therefore, genes that have duplicated copies should have more diversified expression profiles than single-copy genes in a group of closely related organisms. To test this idea, Gu *et al.* [23] used gene-expression data obtained from the start of metamorphosis in three species of the *Drosophila melanogaster* subgroup, to identify the effect of gene duplication on expression. They first used ANOVA to identify genes, whose expression level, at the beginning of metamorphosis, differs from that at the end of the early stage of metamorphosis within a certain lineage; next they identified the genes whose expression differed between lineages in that stage. They found that the proportion of duplicated genes that showed significant changes in expression in a lineage was significantly greater than that of single-copy genes. Furthermore, a between-genome comparison showed that duplicated genes have a much greater probability of expression divergence between species or between different strains of the same species (Table 1). The authors also examined differences in the expression of duplicated genes between two yeast strains and found a similar pattern (Table 1). Therefore, they concluded that duplicated genes are more likely to have divergent expression profiles than single-copy genes both within and between genomes.

More recently, Huminiecki and Wolfe [24] compared the expression profiles of orthologous genes in human and mouse with those genes with a lineage-specific duplication. They focused on loci where a recent lineage-specific gene duplication had created paralogous pairs in either human or mouse, and examined pairs of young duplicates that also have a comparable single-copy ortholog in the other species. The presence of lineage-specific duplicate genes increased the human–mouse divergence in the expression profiles in the homologous tissues of both species. Moreover, orthologs with multiple lineage-specific duplications showed an even greater divergence between expression profiles. They reasoned that because orthologous genes of human and mouse are of the same age (by definition they come from the last common ancestor), the increased divergence in expression is mainly due to the presence of recent lineage-specific gene duplications.

Accelerated regulatory evolution in duplicate genes has also been found in nematodes [13] and therefore appears to be a general phenomenon in eukaryotes. One can also conclude that gene duplication increases gene-expression diversity both within and between genomes.

Models of expression divergence between duplicate genes

The question of whether gene expression evolution is largely a neutral process has attracted considerable attention recently [25–28]. To this end, a statistical framework for studying expression divergence between duplicate genes is needed. Gu [25] and Oakley *et al.* [26] developed a simple model based on the brownian motion. It assumes that the expression divergence between two genes is mainly driven by small and additive genetic drifts (random effects), so it might be called a 'neutral-evolution' model of gene expression. Gu [25] also studied several other models that might involve natural selection. For example, the dramatic-shift (punctuated-equilibrium) model allows a rapid expression divergence shortly after gene duplication. These models can also be used to infer the ancestral expression profiles when the phylogeny of duplicate genes is known. To facilitate application of these models to expression and genomic data, Gu *et al.* [29] defined an additive expression distance between duplicate genes, measured by the average of squared expression differences. They analyzed yeast gene families using a multi-microarray data set and found a more than tenfold

Table 1. Distributions of singletons and duplicate genes in the comparisons between different strains or species of fly and yeast[a]

Difference in expression	Number of singletons	Number of duplicates
Comparison between *Drosophila* strains or species ($\chi^2=97.6$, d.f.$=1$, $P\approx 0$)		
Differentially expressed[b]	1201	1593
Similarly expressed[c,d]	2155	1745
Total	3356	3332
Comparison between yeast strains ($\chi^2=54.5$, d.f.$=1$. $P<10^{-12}$)		
Differentially expressed	541	392
Similarly expressed	2252	925
Total	2793	1317

[a]Data from Ref. [23].
[b]Differentially expressed between at least two strains within a species or between two species.
[c]Similarly expressed between every pair of strains within a species and between species.
[d]If only the genes that showed expression changes at the start of metamorphosis are considered, the proportion of duplicate genes with different expression patterns between species or strains is still significantly greater than that of singletons.

increase in the rate of expression evolution immediately following gene duplication.

The additive expression distance makes it possible to implement the relative-rate test for investigating whether the expression divergence has been symmetric (equal rates) between the two duplicate genes; the test requires an outgroup (reference) gene. Gu *et al.* [29] examined 111 yeast gene families and found 60 gene families (54%) that were asymmetric ($P = 0.05$), indicating a tendency for expression to evolve at different rates in the two duplicates (i.e. the expression of one copy evolves more rapidly than that of the other copy). It will be interesting to test whether the asymmetric evolution of expression divergence is related to the asymmetric evolution of coding sequences [30]. Moreover, when the phylogeny of duplicate genes is known, one can combine a well-developed ancestral-sequence-inference method and the newly developed ancestral-expression-inference method [25] to distinguish between ancestral and derived characters in expression profiles and protein functions simultaneously.

Expression divergence and retention of duplicate genes

A central issue in the evolution of duplicate genes is why so many duplicate genes have been retained in a genome even though the most likely fate for a redundant duplicate is nonfunctionalization. The neofunctionalization [2] and subfunctionalization [22,31] models are frequently invoked to explain the retention of duplicate genes. The neofunctionalization model postulates that gain of new function(s) is the major factor for the retention of both copies of duplicate genes in a genome. The subfunctionalization model is also known as the duplication-degeneration-complementation (DDC) model and it assumes that the two duplicate genes undergo complementary degeneration of their *cis*-regulatory motifs, so that both copies are required to produce the full complement of the *cis*-regulatory motifs of the ancestral gene (Figure 2). This model predicts (i) the total number of *cis*-regulatory motifs in the two genes should decrease with evolutionary time (Figure 2); and (ii) genes with several paralogs should tend to have few regulatory motifs because these genes should have undergone multiple rounds of gene duplication and complementary loss of motifs [19].

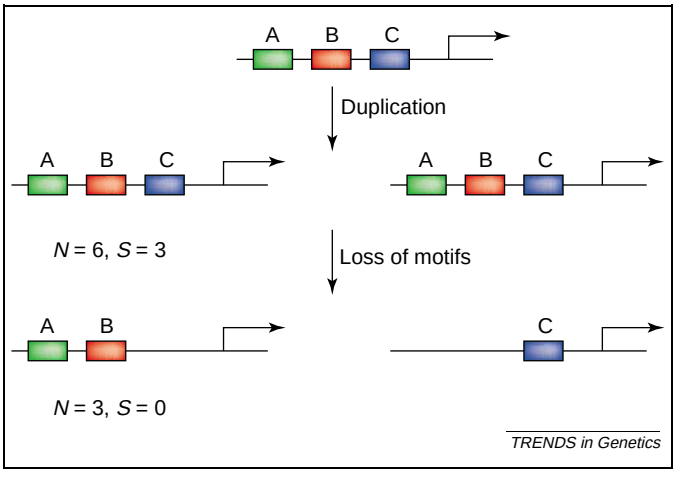

Figure 2. The degenerative complementation model of evolution in regulatory regions following gene duplication. The regulatory motifs are labeled A, B and C. Each daughter gene retains only a subset of promoter elements compared with the ancestral state. As a result, both the number of different motifs shared by the duplicates (*S*) and the total number of motifs in the duplicates (*N*) decline with age. This figure was reproduced, with permission, from Ref. [19].

Using genomic data from yeast, Papp *et al.* [19] tested the predictions of the DDC model. They found that, although the number of shared motifs decreases over time, the total number of motifs remains constant among duplicates (Figure 3) and that genes with numerous paralogs in the yeast genome do not have an especially low number of *cis*-regulatory motifs. To explain these observations, the authors suggested that either regulatory motifs with new function arise from existing motifs or the loss of regulatory motifs is balanced by the gain of new regulatory motifs, keeping the total number constant. They concluded that the DDC model alone can not fully explain duplicate gene evolution in yeast and the gain of a novel function has a substantial role in the retention of duplicate genes in the yeast genome. A similar conclusion was made following an analysis of the transcription factors of duplicate genes in the yeast genome [32].

The traditional view from the study of isozymes predicts a trend towards tissue-specific expression following gene duplication. Huminiecki and Wolfe [24] found a general trend for paralogous genes to become more specialized in their expression patterns, showing decreased breadth and increased specificity of expression as the size of gene families increases. Moreover, they also claimed that a detailed examination of some gene families revealed examples of neofunctionalization of duplicated genes, but only one case of subfunctionalization. However, this issue remains to be resolved because it is possible that neofunctionalization is a later-stage process, whereas subfunctionization is an early stage process that reduces the chance of nonfunctionalization of a duplicate [22,19,33]. Recently, He and Zhang's [33] analysis of yeast protein–protein interaction data supported the idea that a large proportion of duplicate genes undergo rapid subfunctionalization, followed by a prolonged period of neofunctionization. Their analysis of human tissue expression data also supported this idea.

Concluding remarks

Although much progress has been made in our understanding of the mode and tempo of expression divergence between duplicate genes, much remains to be learned. It is now clear that gene duplication enables tissue specialization and increases expression diversity and that expression divergence between duplicate genes increases with evolutionary time. However, the rate of expression divergence might have been overestimated for two reasons. First, the gene expression data used to date came mainly from microarray studies and the 'noisiness' of microarray data might have led to an overestimate of the rate of expression divergence. Second, most of the analyses had not taken into account the effect of gene conversion between duplicate genes, which can occur frequently, at least in yeast [34], and this negligence could have led to an underestimate of the age of duplicate genes and thus an overestimate of the rate of expression divergence. In the future, these two factors should be taken into account to obtain more reliable estimates of the rate of expression

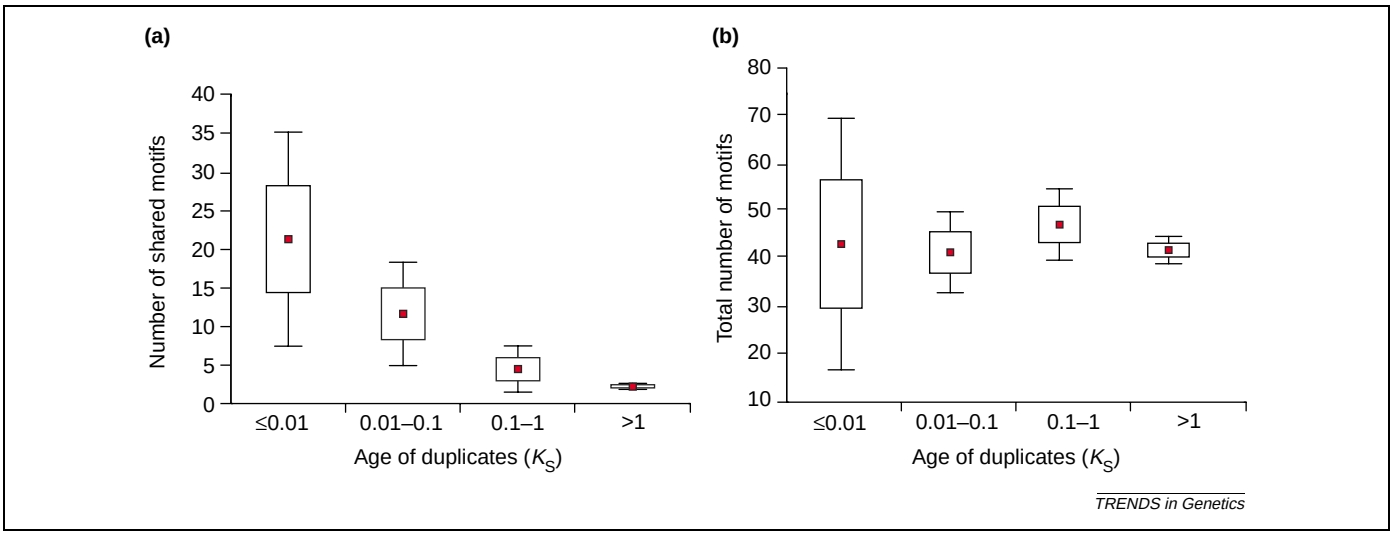

Figure 3. **(a)** Negative association between duplicate age (the rate of evolution at synonymous sites; *Ks*) and the number of different shared motifs of duplicates. **(b)** No association between duplicate age and the total number of motifs possessed by the two copies. This figure was reproduced, with permission, from Ref. [19].

divergence. We note that estimation of expression divergence using EST abundance data can provide a cross-validation of microarray data.

Another topic that needs much further investigation is the role of expression divergence in the retention of duplicate genes in a genome. Previous studies on this topic have been made mainly in yeast, and there is a need to expand this research to other eukaryotes. Yeast would have a large effective population size, which is unfavorable to the DDC model. Furthermore, yeast is a single cell organism, which does not have tissue differentiation. Therefore, the conclusions drawn from yeast data might not be applicable to other species. With the rapid accumulation of functional genomic data, scientists can soon pursue these issues in other eukaryotes.

To date, few studies have considered the relationship between gene duplication and evolution of a transcriptional regulatory network. Although available data from *Escherichia coli* and *S. cerevisiae* suggest that gene duplication plays a key role in the growth of gene networks [35], the detail of how gene duplication can affect the evolution of a network remains to be explored. Also, it is not clear how the position of a gene in a regulatory network affects the survival of duplicated genes and their subsequent expression evolution. In short, there remain many problems in the evolution of expression of duplicate genes in the context of a regulatory network.

Acknowledgements

We thank Laurence Hurst, Greg Whitt and Jianzhi Zhang for useful comments. This study was supported by NIH grants.

References

1 Markert, C.L. (1964) Cellular differentiation – an expression of differential gene function. In *Congenital Malformations*, pp. 163–174 International Medical Congress
2 Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag
3 Ferris, S.D. and Whitt, G.S. (1979) Evolution of the differential regulation of duplicate genes after polyploidization. *J. Mol. Evol.* 12, 267–317
4 Wagner, A. (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist–selectionist debate. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6579–6584
5 Gu, Z. *et al*. (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* 18, 609–613
6 Rifkin, S.A. *et al*. (2003) Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat. Genet.* 33, 138–144
7 Coyne, J.A. (2005) Switching on evolution. *Nature* 435, 1029–1030
8 Wolfe, K.H. and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713
9 Zhang, Z. *et al*. (2004) How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution? *Trends Genet.* 20, 403–407
10 Conant, G. and Wagner, A. (2004) Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. *Proc. Biol. Sci.* 271, 89–96
11 Makova, K.D. and Li, W.H. (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.* 13, 1638–1645
12 Blanc, G. and Wolfe, K.H. (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16, 1679–1691
13 Castillo-Davis, C.I. *et al*. (2004) cis-regulatory and protein evolution in orthologous and duplicate genes. *Genome Res.* 14, 1530–1536
14 The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815
15 Vision, T.J. *et al*. (2000) The origins of genomic duplications in *Arabidopsis*. *Science* 290, 2114–2117
16 Blanc, G. *et al*. (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* 13, 137–144
17 Kim, S.K. *et al*. (2005) Transcriptional divergence of the duplicated oxidative stress-responsive genes in the *Arabidopsis* genome. *Plant J.* 41, 212–220
18 Haberer, G. *et al*. (2004) Transcriptional similarities, dissimilarities, and conservation of cis-elements in duplicated genes of *Arabidopsis*. *Plant Physiol.* 136, 3009–3022
19 Papp, B. *et al*. (2003) Evolution of *cis*-regulatory elements in duplicated genes of yeast. *Trends Genet.* 19, 417–422
20 Piatigorsky, J. and Wistow, G. (1991) The recruitment of crystallins: new functions precede gene duplication. *Science* 252, 1078–1079
21 Hughes, A.L. (1994) The evolution of functionally novel proteins after gene duplication. *Proc. Biol Sci* 256, 119–124
22 Force, A. *et al*. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, 1531–1545
23 Gu, Z. *et al*. (2004) Duplicate genes increase gene expression diversity within and between species. *Nat. Genet.* 36, 577–579

24 Huminiecki, L. and Wolfe, K.H. (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res.* 14, 1870–1879
25 Gu, X. (2004) Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics* 167, 531–542
26 Oakley, T.H. *et al*. (2005) Comparative methods for the analysis of gene-expression evolution: an example using yeast functional genomic data. *Mol. Biol. Evol.* 22, 40–50
27 Yanai, I. *et al*. (2004) Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *OMICS* 8, 15–24
28 Jordan, I.K. *et al*. (2004) Evolutionary significance of gene expression divergence. *Gene* 345, 119–126
29 Gu, X. *et al*. (2005) Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc. Natl. Acad. Sci. U. S. A.* 102, 707–712
30 Kellis, M. *et al*. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* 428, 617–624
31 Lynch, M. and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459–473
32 Evangelisti, A.M. and Wagner, A. (2004) Molecular evolution in the yeast transcriptional regulation network. *J Exp. Zoolog. B. Mol. Dev. Evol.* 302, 392–411
33 He, X. and Zhang, J. (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169, 1157–1164
34 Gao, L.Z. and Innan, H. (2004) Very low gene duplication rate in the yeast genome. *Science* 306, 1367–1370
35 Teichmann, S.A. and Babu, M.M. (2004) Gene regulatory network growth by duplication. *Nat. Genet.* 36, 492–496