Cell
PRESS

Genome Analysis

Maintenance of duplicate genes and their functional redundancy by reduced expression

Wenfeng Qian[1]*, Ben-Yang Liao[2]*, Andrew Ying-Fei Chang[2] and Jianzhi Zhang[1]

[1] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA
[2] Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Miaoli County 350, Taiwan, Republic of China

Although evolutionary theories predict functional divergence between duplicate genes, many old duplicates still maintain a high degree of functional similarity and are synthetically lethal or sick, an observation that has puzzled many geneticists. We propose that expression reduction, a special type of subfunctionalization, facilitates the retention of duplicates and the conservation of their ancestral functions. Consistent with this hypothesis, gene expression data from both yeasts and mammals show a substantial decrease in the level of gene expression after duplication. Whereas the majority of the expression reductions are likely to be neutral, some are apparently beneficial to rebalancing gene dosage after duplication.

Gene duplication without functional divergence

Gene duplication is prevalent in all three domains of life and is the major source of new genes [1–2]. Immediately after gene duplication, the two daughter genes are usually functionally redundant, especially when the entire gene together with its regulatory region is duplicated. Thus, mutations that knock out one of the duplicates are invisible to natural selection. Generally only one daughter gene is stably retained while the other degenerates into a pseudogene that is eventually lost. Therefore, with the exception of a small number of genes for which increased dosage can be beneficial (e.g. ribosomal RNA and histone genes) [2], the two daughter genes cannot be stably maintained in the same genome unless they escape from the usual fate of pseudogenization by quickly diverging in function, and this can occur by the acquisition of new functions (neofunctionalization) [1], subdivision of ancestral functions (subfunctionalization) [3], or a combination of the two [4]. Surprisingly, however, several studies in yeast and nematode have found many duplicate gene pairs with negative epistasis [5–9], meaning that deleting both gene copies produces a significantly larger defect than expected from the effects of individual deletions. Negative epistasis is caused by functional redundancy [10]. Whereas one might think that most of these negatively epistatic gene pairs are young duplicates that have not had sufficient time to diverge in function, this is not the case [7–9]. In fact, many of them are quite old [7–9] and some originated as early as a billion years ago [7]. The long-term maintenance of functional redundancy of duplicate genes is unexpected and puzzling.

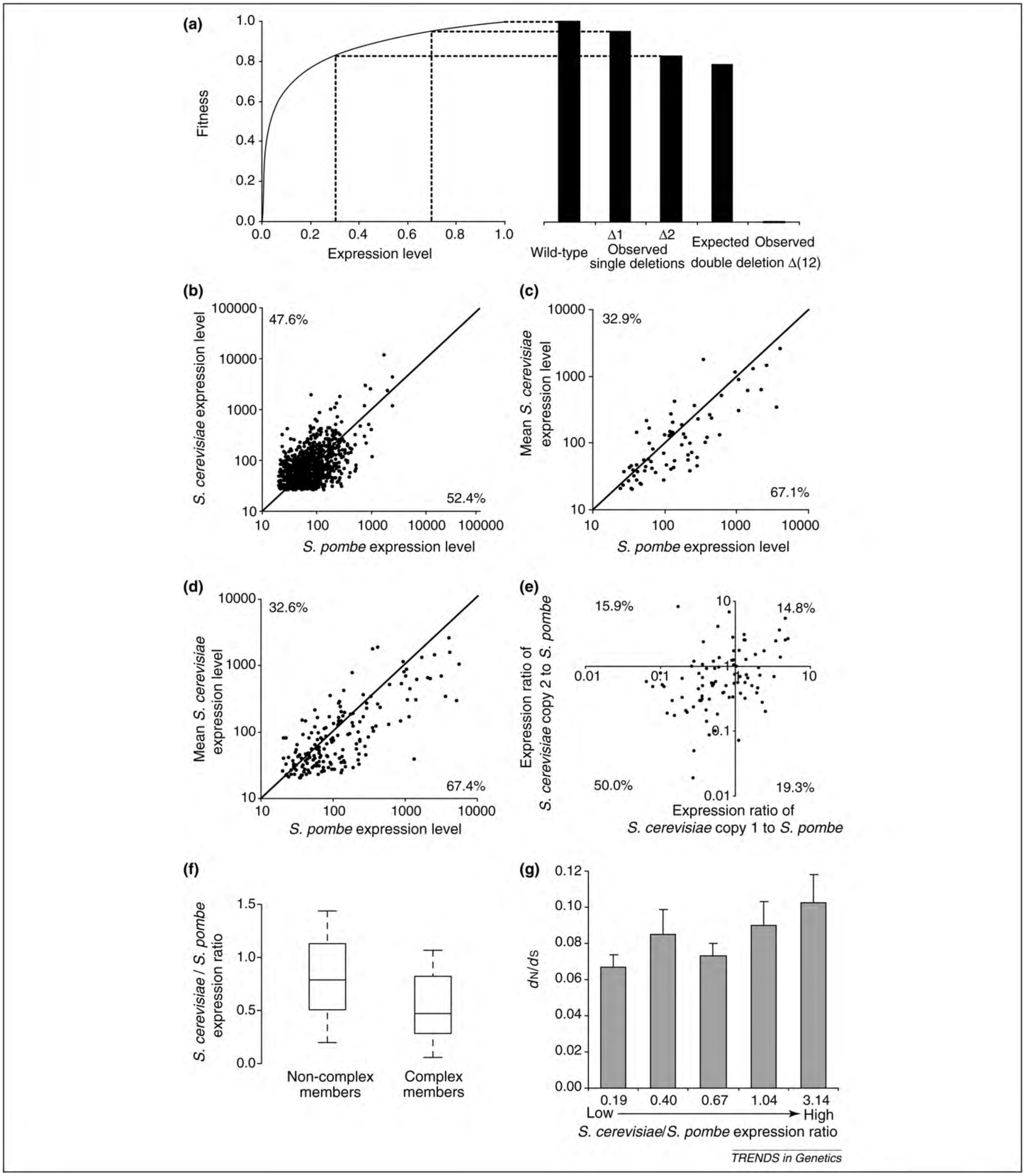
Reduced expression can lead to the maintenance of functional redundancy

Here we propose a simple mechanism for the stable maintenance of functional redundancy in duplicate genes. We propose that the amount of expression of each daughter gene is reduced compared to the expression of the progenitor gene. This expression reduction prevents the loss of either daughter gene because such loss would render the total expression level after duplication lower than that before duplication, which would be deleterious. The expression reduction, when it is sufficiently large, would require both daughter genes to retain all ancestral functions, preventing the occurrence of functional divergence. In our model, although the two daughter genes are functionally equivalent, they are not redundant in a strict sense, because the deletion of either copy is expected to cause a fitness reduction that is sufficiently large to be disfavored by natural selection. Negative epistasis between functionally equivalent duplicates, regardless of the definition of epistasis by non-multiplicative or non-additive fitness effects of individual mutations, results from the well-established nonlinear relationship between gene expression level and fitness [11] (Figure 1A). That is, the fitness effect of reducing the expression level by 50% is less than 50% [12]. This phenomenon is closely related to the observations that most genes are haplosufficient [12–13] and that most wild-type alleles are dominant to loss-of-function alleles [12–13]. Because subfunctionalization includes reductions in the joint levels or patterns of activity of the duplicate genes [3], expression reduction after gene duplication is a type of subfunctionalization in the joint levels rather than patterns of activity. So, all previous theoretical results on subfunctionalization should apply to our model. Note, however, that subfunctionalization in the joint patterns of activity cannot explain the long retention of genetic redundancy because negative epistasis is not expected if the two daughter genes have non-overlapping protein functions or tissue/condition expressions.

Substantial expression reduction after gene duplication in yeasts

To test if the expression levels of duplicate genes are indeed decreased compared to their progenitor genes we examined gene expression levels measured by the

Corresponding author: Zhang, J. (jianzhi@umich.edu)
* These authors contributed equally.



TRENDS in Genetics

Figure 1. Expression reduction after gene duplication in yeasts. **(a)** Because fitness is a concave function of expression level there is negative epistasis between duplicates with reduced expression. Synthetic lethality is observed in this hypothetical example. The fitness of the double deletion strain expected under no epistasis is calculated assuming multiplicative fitness effects of single deletions. **(b)** Expression levels of one-to-one orthologs in *S. cerevisiae* and *S. pombe*. Each dot represents a one-to-one orthologous pair. The diagonal line indicates identical expression levels in the two species. Fractions of orthologs below and above the diagonal line are indicated. **(c)** Expression levels of two-to-one orthologs that are negatively epistatic in *S. cerevisiae*. Each dot represents a two-to-one ortholog. The expression level of the single gene in *S. pombe* and the mean expression of the duplicates in *S. cerevisiae* are presented. The fraction of dots below the diagonal is significantly greater than that in **(b)** ($P = 0.006$). **(d)** Expression levels of all two-to-one orthologs. The fraction of dots below the diagonal is significantly greater than that in (B) ($P = 2 \times 10^{-5}$). **(e)** Expression ratios between *S. cerevisiae* and *S. pombe* for all two-to-one orthologs. The fraction of dots in the lower-left quadrant is significantly greater than expected ($P = 6 \times 10^{-6}$). **(f)** *S. cerevisiae* duplicates involved in the same protein complexes have lower *S. cerevisiae*/*S. pombe* expression ratios ($P < 0.05$). The values of upper quartile, median, and lower quartile are indicated in each box, whereas the bars outside the box indicate semiquartile ranges. **(g)** Two-to-one orthologs with lower mean *S. cerevisiae* expression relative to *S. pombe* expression show lower nonsynonymous/synonymous rate ratios (d_N/d_S) between *S. cerevisiae* duplicates ($P < 0.05$ for unbinned data). The five bins are of equal sample size and the mean expression ratio and mean d_N/d_S value for each bin are presented. The error bars show one standard error.

RNA-Seq method using next-generation sequencing. RNA-Seq substantively outperforms microarray-based methods in the accuracy and dynamic range of the measurement [14]. We first identified one-to-one, two-to-one, and many-to-one orthologs between the baker's yeast *Saccharomyces cerevisiae* and the fission yeast *Schizosaccharomyces pombe* (see Supplementary Methods in the supplementary material online). One-to-one orthologs are those genes with neither duplication nor gene loss in the two yeast lineages since their separation. Two-to-one (or many-to-one) orthologs have had one (or multiple) duplications in the *S. cerevisiae* lineage, but have experienced neither duplication nor gene loss in the *S. pombe* lineage. We have focused on duplicates in *S. cerevisiae* (rather than in *S. pombe*) because epistasis between duplicate genes has only been examined in *S. cerevisiae*. Because of the different sequencing depths of the RNA-Seq data of the two yeasts, we adjusted the RNA-Seq depth in *S. cerevisiae* based on the assumption that one-to-one orthologs have the same average expression levels between the two species (Supplementary Methods). Thus, one should consider our results from two-to-one and many-to-one orthologs relative to those from one-to-one orthologs. Our between-species expression comparison is meaningful because the RNA-Seq data from the two yeasts and the epistasis data from *S. cerevisiae* are all obtained under similar rich medium conditions.

There are 891 one-to-one orthologous genes with expression information from both yeasts. We found that 52.4% of these have lower expression levels in *S. cerevisiae* than in *S. pombe* (Figure 1B). By comparison, among the 70 two-to-one orthologs that are known to be negatively epistatic in *S. cerevisiae* and have expression data (see Supplementary Methods), 67.1% of the *S. cerevisiae* duplicate pairs have a lower mean expression than their single counterparts in *S. pombe* (Figure 1C). The difference between two-to-one and one-to-one orthologs is highly significant ($P = 0.006$, one-tailed Fisher's exact test). Using one-to-one orthologs as a control, we estimated that an excess of $0.671 - ((1 - 0.671) \times (0.524/0.476)) = 30.9\%$ of duplicate gene pairs with negative epistasis experienced a decrease in mean expression after gene duplication. We calculated the *S. cerevisiae*/*S. pombe* expression ratio for each two-to-one ortholog, where the *S. cerevisiae* expression is the mean expression level of the two paralogs. We found that this ratio (median = 0.74) is significantly lower than that from one-to-one orthologs (0.94; $P = 0.001$, one-tailed Mann-Whitney U test), further supporting expression reduction after gene duplication.

To examine if the above observation is more widespread than the set of negatively epistatic duplicate genes, we examined all 227 two-to-one orthologs, 69% of which either have not been tested for epistasis or have no detectable negative epistasis. Note, however, that because negative epistasis is currently detected only when the overlapping function of a duplicate pair contributes substantially to cell growth in rich medium, duplicates that do not show detectable negative epistasis could still have overlapping functions. We found that 67.4% of two-to-one orthologs have lower mean expression levels in *S. cerevisiae* than in *S. pombe* (Figure 1D), significantly greater than that in one-to-one orthologs (Figure 1B) ($P = 2 \times 10^{-5}$). The above result is supported even when only genes with significant expression reductions are considered (Supplementary Methods). We estimated that an excess of 31.5% of duplicate gene pairs have experienced mean expression reduction after gene duplication. The median expression ratio (*S. cerevisiae*/*S. pombe*) is 0.74 for all two-to-one orthologs, significantly lower than that (0.94) for one-to-one orthologs ($P = 4 \times 10^{-6}$). Similar results were obtained from 33 many-to-one orthologs between *S. cerevisiae* and *S. pombe* (Figure S1).

Because ancestral *S. cerevisiae* experienced a whole genome duplication (WGD) ~100 million years ago [15], we separated duplicates based on whether they were generated by the WGD (Figure S2A). We further separated the non-WGD group into four age groups (Figure S2A). There is no significant variation in the prevalence or degree of expression reduction between WGD and non-WGD groups or among the non-WGD age groups (Figure S2B).

In all of the above analyses we calculated the mean expression level for paralogous genes of *S. cerevisiae* and then compared it with the expression level of the single copy ortholog in *S. pombe*. However, our hypothesis predicts that both daughter genes will have lower expression levels than their progenitor gene. To verify this prediction, for each two-to-one ortholog we examined the expression ratio between each of the two *S. cerevisiae* duplicates and its *S. pombe* ortholog (Figure 1E). Because 52.4% of one-to-one orthologs have lower expression levels in *S. cerevisiae* than in *S. pombe*, we expect that by chance $(0.524)^2 = 27.46\%$ of two-to-one orthologs will have lower expression levels in *S. cerevisiae* than in *S. pombe* for both copies (i.e. in the lower-left quadrant of Figure 1E). In fact, 50.0% lie in this quadrant ($P = 6 \times 10^{-6}$, binomial test), based on which we estimated that an excess of 31.1% of duplicates experienced reduced expression levels for both copies. These results also confirmed that the expression reduction phenomenon is not an artifact of condition-specific expression levels of duplicate genes because, under the latter scenario, only one daughter gene is expected to have lowered expression.

We also examined those genes that were duplicated in the *S. pombe* lineage since its separation from the *S. cerevisiae* lineage (one-to-two orthologs), and again found the phenomenon of expression reduction after gene duplication. For example, the *S. cerevisiae*/*S. pombe* expression ratio is significantly higher for one-to-two orthologs than for one-to-one orthologs ($P = 0.03$; U test).

Evolutionary mechanisms and consequences of expression reduction

The reduction of expression after gene duplication can occur simply by random fixation of neutral regulatory mutations that decrease gene expression, as long as the total expression of the two daughter genes is not below the level required for wild-type function. Expression reduction could also be advantageous if the total gene expression upon duplication is higher than the optimal level. Excess of gene expression and protein production can be deleterious because they waste energy and raw materials [16] and result in additional misfolded protein molecules that are cytotoxic [17]. Further, the stoichiometry among different molecules in a cell can be broken by extra production of a protein. Specifically, the toxicity of dosage imbalance caused by the duplication of a gene that encodes a component of a stable protein complex is potentially high [18,19]. To explore the possibility of adaptive expression reduction after gene duplication, especially for rebalancing gene dosage, we focused on two-to-one orthologs. Because dosage balance should not be affected by WGD we excluded from our analysis all duplicates that resulted from the WGD. By contrast, individual gene duplications are unlikely to occur simultaneously to multiple components of the same protein complex and thus could cause dosage imbalance. We found that the reduction in mean expression for paralogs involved in the same protein complexes is significantly greater than that for paralogs not involved in complexes ($P < 0.05$, one-tailed U test, Figure 1F). This finding suggests that, at least in some duplicate genes, expression reduction could have been beneficial and positively selected for, owing to its role in rebalancing gene dosage after duplication.

After a substantial reduction of expression in each daughter gene, protein function is less likely to change because such a change would render the total activity of the products of the two daughter genes lower than that of the progenitor gene and be harmful. To test this prediction, for each two-to-one ortholog we estimated the ratio of mean expression in *S. cerevisiae* and expression in *S. pombe*, as well as the ratio of the nonsynonymous to synonymous nucleotide substitution rates (d_N/d_S) between the two *S. cerevisiae* duplicates. We found that the d_N/d_S ratio

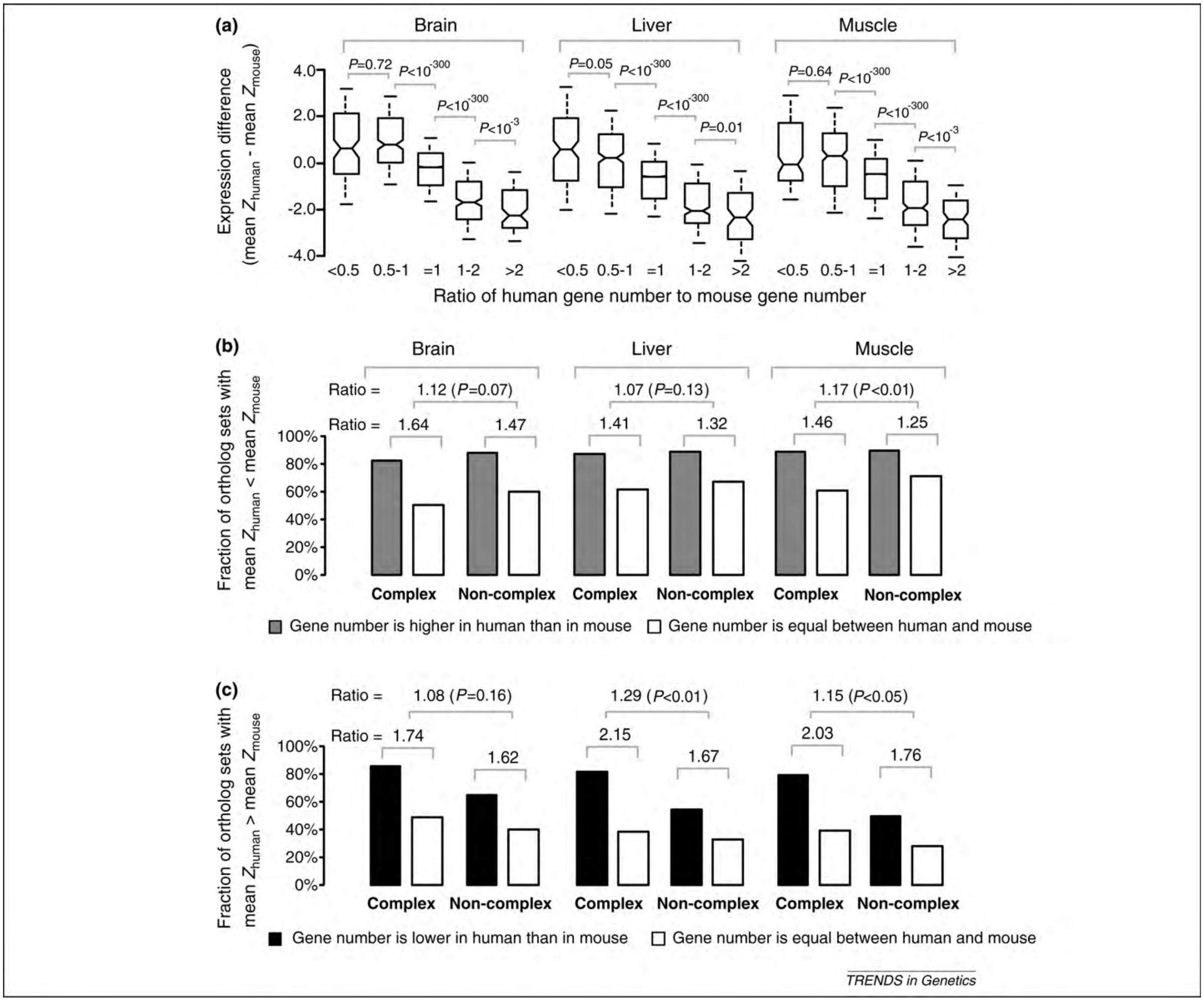


Figure 2. Expression reduction after gene duplication in mammals. **(a)** Mean expression of a human–mouse orthologous set in a species reduces as the number of paralogs in that species increases. The values of upper quartile, median, and lower quartile are indicated in each box, whereas the bars outside the box indicate semiquartile ranges. *P*-values between boxes are from one-tailed Mann-Whitney *U* tests. Spearman's rank correlation coefficients for the unbinned data of the three tissues are -0.178 ($P < 10^{-106}$), -0.159 ($P < 10^{-82}$) and −0.152 ($P < 10^{-77}$), respectively. **(b)** Proteins functioning as subunits of complexes are subject to more widespread expression reductions after gene duplication, as demonstrated by orthologous sets in which human genes outnumber mouse genes. *P* values are estimated from 10 000 parametric bootstrapping replications. **(c)** Proteins functioning as subunits of complexes are subject to more widespread expression reductions after gene duplication, evident from orthologous sets in which mouse genes outnumber human genes.

decreases as the expression ratio decreases (Spearman's correlation ρ = 0.11, $P = 0.047$, one-tailed t test; see Figure 1G for the binned results). Because lower d_N/d_S ratios indicate slower functional changes our observation suggests that the expression reduction after duplication indeed hampers functional divergence of duplicates. The above result is conservative in view of the strong negative impact of absolute expression level on the rate of protein sequence evolution of a gene [17,20]. Indeed, the partial correlation between d_N/d_S and the *S. cerevisiae*/*S. pombe* expression ratio becomes much stronger (ρ = 0.19, $P = 0.001$, one-tailed t test) when the average expression level of *S. cerevisiae* duplicates is controlled for (Supplementary Methods).

Widespread expression reduction in mammalian duplicate genes

To examine whether expression reduction after gene duplication is also found in other species, especially mammals, we analyzed RNA-Seq data from human and mouse. Because the gene expression distributions differ substantially between the two species (Figure S3), it is inappropriate to compare the expression levels of human and mouse orthologs directly. Instead, we transformed the expression levels of human and mouse genes to Z-scores after a $\log_2$ transformation (Supplementary Methods) and then compared the Z-scores of orthologous genes. For each gene that existed in the common ancestor of human and mouse we identified all of its orthologs in extant human and mouse and referred to them as an orthologous set. We then calculated the difference in mean Z score between the human genes and the mouse genes in the orthologous set, as well as the human/mouse ratio of the gene number in the set. We found that when there are more human genes than mouse genes in an orthologous set, the mean expression level per gene tends to be lower in human than in mouse, and vice versa (Figure 2A). This pattern is clear in each of the three tissues examined (brain, liver, and muscle) (Figure 2A) and therefore is unlikely to result from the unique characteristics of particular tissues. Our conclusion is also supported by analyzing gene expression ranks instead of Z scores (Supplementary Methods). Furthermore, duplicates involved in protein complexes tend to have more widespread expression reductions than those not involved in complexes (Figure 2B,C). This trend exists in all three tissues examined, although some comparisons are not statistically significant due to small sample sizes (Figure 2B,C).

Concluding remarks and implications

In this work we have proposed that expression reduction after gene duplication, a special type of subfunctionalization, facilitates the long-term maintenance of duplicate genes and their functional redundancy. We showed in both yeasts and mammals that a substantial fraction of duplicate genes experience expression reduction, and this hampers functional divergence of duplicate genes. We further showed that the expression reduction in some genes can be adaptive for dosage rebalance, although it is probably neutral in most other cases. It has been proposed that functionally redundant duplicate genes are used to backup important functions in the event of a severe mutation, much like the role of a spare tire in a car. However, theoretical population genetic analysis demonstrated that duplicates are unlikely to be maintained by the backup mechanism [21]. The present analysis further excludes the need for the backup hypothesis. Our finding is consistent with the recent discovery that only ~10% of duplicate genes are upregulated when their paralogs are deleted [22]. Even in such cases, the apparent backup phenomenon could be an evolutionary byproduct [22]. These results echo the recent finding that the abundant functional redundancies caused by alternative pathways in metabolic networks need not and cannot be explained by the backup hypothesis [23]. Together, they suggest that the genetic robustness against mutations conferred by either duplicate genes or alternative pathways is a byproduct of other evolutionary processes [24]. A few other non-backup models have been proposed to explain the evolutionary maintenance of functional redundancy between duplicates [7,25]. For instance, the piggyback hypothesis posits that two paralogs have some non-overlapping functions as well as some overlapping functions, and the latter are kept as a byproduct of the former owing to strong structural constraints [7]. Our expression reduction model requires neither the existence of non-overlapping functions nor such functional constraints, and thus could be more widely applicable. An earlier hypothesis asserts that a substantial proportion of duplicate genes are fixed and retained due to the benefit of enhanced dosage [11]. This hypothesis is not supported by either human or yeast genomic data [18]. The finding of expression reduction in approximately one third of duplicates further sets an upper limit for the fraction of duplicates whose retention could possibly be explained by this hypothesis – because under this hypothesis the expression reduction would be deleterious and hence prohibited. While one might think that, functionally speaking, there is no change if the total expression level of two daughter genes is decreased to the level of their progenitor gene, we note that the stochastic variation in the total amount of product (i.e. expression noise) is lowered after duplication, which can be beneficial or deleterious depending on the genes concerned [26,27].

Acknowledgements

We thank Meg Bakewell, Mike Lynch, Calum Maclean, Csaba Pal, Jian-Rong Yang and three anonymous reviewers for valuable comments. This work was supported by National Institutes of Health grants to J.Z. and Taiwan National Health Research Institutes intramural funding and Taiwan National Science Council grant (NSC 99-2311-B-400-003-MY2) to B.-Y.L.

Appendix A. Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2010.07.002.

References

1 Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag

2 Zhang, J. (2003) Evolution by gene duplication – an update. *Trends Ecol. Evol.* 18, 292–298

3 Lynch, M. and Force, A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 459–473

4 He, X. and Zhang, J. (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169, 1157–1164

5 DeLuna, A. *et al.* (2008) Exposing the fitness contribution of duplicated genes. *Nat. Genet.* 40, 676–681

6 Musso, G. *et al.* (2008) The extensive and condition-dependent nature of epistasis among whole-genome duplicates in yeast. *Genome Res.* 18, 1092–1099

7 Vavouri, T. *et al.* (2008) Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends Genet.* 24, 485–488

8 Dean, E.J. *et al.* (2008) Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genet.* 4, e1000113

9 Tischler, J. *et al.* (2006) Combinatorial RNA interference in *Caenorhabditis elegans* reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. *Genome Biol.* 7, R69

10 He, X. *et al.* (2010) Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks. *Nat. Genet.* 42, 272–276

11 Kondrashov, F.A. and Koonin, E.V. (2004) A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet.* 20, 287–290

12 Deutschbauer, A.M. *et al.* (2005) Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* 169, 1915–1925

13 Kim, D.U. *et al.* (2010) Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. *Nat. Biotechnol.* 28, 617–623

14 Wang, Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63

15 Wolfe, K.H. and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387, 708–713

16 Wagner, A. (2005) Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* 22, 1365–1374

17 Drummond, D.A. and Wilke, C.O. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134, 341–352

18 Qian, W. and Zhang, J. (2008) Gene dosage and gene duplicability. *Genetics* 179, 2319–2324

19 Papp, B. *et al.* (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424, 194–197

20 Pal, C. *et al.* (2001) Highly expressed genes in yeast evolve slowly. *Genetics* 158, 927–931

21 Clark, A.G. (1994) Invasion and maintenance of a gene duplication. *Proc. Natl. Acad. Sci. U. S. A.* 91, 2950–2954

22 DeLuna, A. *et al.* (2010) Need-based up-regulation of protein levels in response to deletion of their duplicate genes. *PLoS Biol.* 8, e1000347

23 Wang, Z. and Zhang, J. (2009) Abundant indispensable redundancies in cellular metabolic networks. *Genome Biol. Evol.* 1, 23–33

24 Harrison, R. *et al.* (2007) Plasticity of genetic interactions in metabolic networks of yeast. *Proc. Natl. Acad. Sci. U. S. A.* 104, 2307–2312

25 Nowak, M.A. *et al.* (1997) Evolution of genetic redundancy. *Nature* 388, 167–171

26 Zhang, Z. *et al.* (2009) Positive selection for elevated gene expression noise in yeast. *Mol. Syst. Biol.* 5, 299

27 Lehner, B. (2008) Selection to minimise noise in living systems and its implications for the evolution of gene expression. *Mol. Syst. Biol.* 4, 170

0168-9525/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2010.07.002 Trends in Genetics 26 (2010) 425–430