

# Note

## Transcriptional Reprogramming and Backup Between Duplicate Genes: Is It a Genomewide Phenomenon?

Xionglei He and Jianzhi Zhang<sup>1</sup>

*Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan 48109*

Manuscript received August 22, 2005

Accepted for publication November 7, 2005

### ABSTRACT

Deleting a duplicate gene often results in a less severe phenotype than deleting a singleton gene, a phenomenon commonly attributed to functional compensation among duplicates. However, duplicate genes rapidly diverge in expression patterns after duplication, making functional compensation less probable for ancient duplicates. Case studies suggested that a gene may provide compensation by altering its expression upon removal of its duplicate copy. On the basis of this observation and a genomic analysis, it was recently proposed that transcriptional reprogramming and backup among duplicates is a genomewide phenomenon in the yeast *Saccharomyces cerevisiae*. Here we reanalyze the yeast data and show that the high dispensability of duplicate genes with low expression similarity is a consequence of expression similarity and gene dispensability, each being correlated with a third factor, the number of protein interactions per gene. There is little evidence supporting widespread functional compensation of divergently expressed duplicate genes by transcriptional reprogramming.

GENE duplication is the primary source of new genes (OHNO 1970; ZHANG 2003). Consequently, there are many duplicate genes in virtually every genome examined (ZHANG 2003). It is often observed that deleting a duplicate gene results in a less severe phenotype than deleting a singleton gene (GU *et al.* 2003; KAMATH *et al.* 2003; CONANT and WAGNER 2004), a phenomenon commonly attributed to functional compensation among duplicates (CONANT and WAGNER 2004; GU *et al.* 2003; KAMATH *et al.* 2003). Conceivably, this compensation relies on a similar expression pattern between duplicates. Indeed, KAFRI *et al.* (2005) recently showed in the yeast *Saccharomyces cerevisiae* that among relatively young duplicates (with synonymous nucleotide distance  $d_s < 1$ ), the probability that a gene is dispensable declines as the mean expression similarity (MES) between the duplicate pair decreases, where MES is measured under 40 different conditions. But surprisingly, among ancient duplicates ( $d_s > 1$ ), which constitute >90% of all yeast duplicates, the probability that a gene is dispensable increases when MES falls from 1 to ~0.2. KAFRI *et al.* (2005) proposed an interesting idea that these ancient duplicates have acquired an expression reprogramming ability that allows normally differen-

tially expressed duplicates to provide compensation when needed. Because gene expression diverges rapidly between duplicates (GU *et al.* 2002; MAKOVA and LI 2003), their hypothesis may explain why deleting an anciently duplicated gene still causes a smaller phenotypic effect than deleting a singleton gene (GU *et al.* 2003; CONANT and WAGNER 2004). In KAFRI *et al.*'s hypothesis, the ability of young duplicates to compensate one another is due to a shared expression inherited from their common ancestor. This ability diminishes quickly with evolutionary time, but the ability to compensate is regained long after duplication by acquisition of transcriptional reprogramming. This regaining of compensation is improbable, because to acquire the relevant expression reprogramming capability, a duplicate gene has to be able to recognize its sister copy even after substantial changes at both expression and sequence levels. Here we provide an alternative explanation to KAFRI *et al.*'s observation, which is not based on expression reprogramming.

### DATA AND METHODS

The yeast data used by KAFRI *et al.* (2005) were downloaded from <http://longitude.weizmann.ac.il/backupcircuits/>. Of 2216 duplicate gene pairs, 1551 pairs had fitness information for both copies and were used in our analyses. As in the analysis of KAFRI *et al.* (2005), a gene may be involved in more than one gene

<sup>1</sup>Corresponding author: Department of Ecology and Evolutionary Biology, University of Michigan, 1075 Natural Science Bldg., 830 North University Ave., Ann Arbor, MI 48109. E-mail: jianzhi@umich.edu

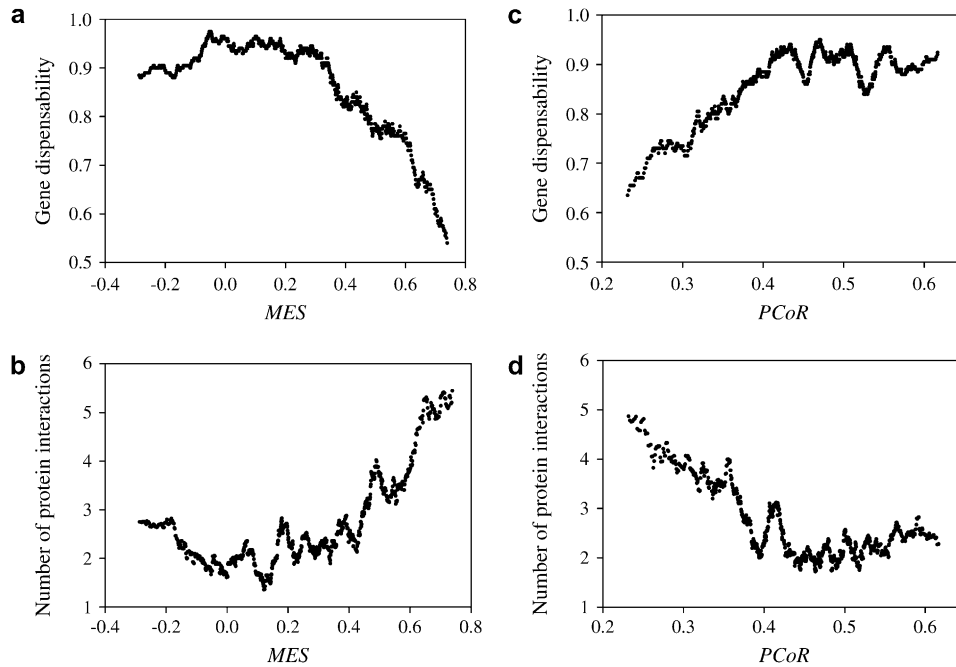


FIGURE 1.—Covariation with the number of protein interaction partners per gene explains the correlation between the MES or PCoR and gene dispensability. In a and b, duplicate gene pairs are sorted by MES, and the average MES and average gene dispensability are computed for a window size of 100 duplicate pairs, with the sliding window moving every gene pair. In c and d, duplicate genes are sorted by PCoR. A total of 1432 pairs of yeast duplicate genes with  $d_s > 1$  are analyzed here.

pair. In fact, the data set is highly redundant; on average each gene appeared  $\sim 2.3$  times. We computed  $d_s$  between duplicate genes using PAML with default parameters (YANG 1997). Those pairs with  $d_s < 1$  were regarded as recent duplicates (116 pairs), and the rest were regarded as ancient duplicates (1435 pairs). We used the yeast protein–protein interaction data compiled in our recent article (HE and ZHANG 2005). Briefly, we merged the data annotated in the Munich Information Center for Protein Sequences (<ftp://ftpmips.gsf.de/yeast/PPI>) and the high-confidence subset of the data recently compiled (VON MERING *et al.* 2002). After excluding self-interactions and interactions involving mitochondrial genes, a nonredundant protein interaction data set containing 9316 pairwise interactions among 4292 genes was derived. We obtained the yeast stable protein complex data set from Saccharomyces Genome Database ([ftp://genome-ftp.stanford.edu/pub/yeast/data\\_download/literature\\_curation/go\\_protein\\_complex\\_slim.tab](ftp://genome-ftp.stanford.edu/pub/yeast/data_download/literature_curation/go_protein_complex_slim.tab)), which contains 188 complexes comprising 1226 genes. The mean number of protein interaction partners per protein is substantially higher for proteins involved in stable protein complexes ( $6.77 \pm 0.33$ ) than for other proteins ( $3.61 \pm 0.14$ ) (Mann-Whitney  $U$ -test,  $P < 0.0001$ ). To be more conservative, we included proteins with at least one partner in the above comparison.

In the sliding-window analyses presented in Figure 1, we excluded three duplicate pairs (YER081W/YIL074C, YJR091C/YPR042C, and YKL068W/YMR047C) because of the presence of genes with extremely high numbers of protein interaction partners (94 for YER081W, 289 for YJR091C, and 128 for YMR047C) that would inappropriately influence the calculation of the mean number of partners in a window. We used the same strategy as in

KAFRI *et al.* (2005) when computing the probability that a gene is dispensable for Figure 2. Specifically, each gene was considered only once in the computation of gene dispensability in each bin.

## RESULTS AND DISCUSSION

Most proteins execute their functions through interacting with other proteins. Known as the centrality–lethality relationship, proteins involved in more protein–protein interactions tend to be indispensable (JEONG *et al.* 2001; HAHN and KERN 2005). For two reasons we suspect that, on average, ancient duplicates with high MES have more protein interaction partners than do ancient duplicates with low MES, thus generating a negative correlation between MES and gene dispensability. First, members of the same or related stable protein complexes are expected to show high MES due to transcriptional coregulation and these genes also tend to have more protein interaction partners (see DATA AND METHODS). Second, housekeeping genes are constantly expressed and therefore should have high MES among themselves. They may also engage in more protein interactions because of their involvement in many cellular processes. We thus hypothesize that the negative correlation between MSE and gene dispensability for ancient duplicates (KAFRI *et al.* 2005) is due to the covariation of these two parameters with the number of protein interactions.

By a sliding-window analysis, we first reproduced the relationship between MES and gene dispensability (Figure 1a) reported by KAFRI *et al.* (2005). Using the same window size and step length, we generated the

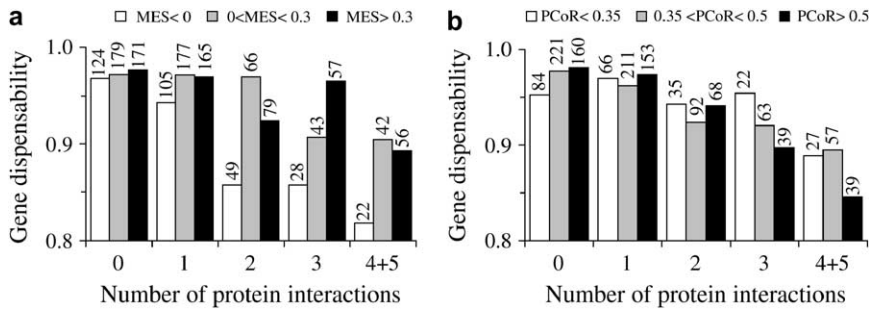


FIGURE 2.—No consistent relationships exist (a) between the MES and gene dispensability or (b) between PCoR and gene dispensability among the yeast duplicate genes with  $d_s > 1$ , when the number of protein interaction partners per gene is controlled for. Genes involved in stable protein complexes are excluded. We first separate genes according to the number of protein interactions and then compute the average gene dispensability for genes with a given number of protein interactions in three MSE or PCoR ranges. The number of genes used is indicated for each bin.

relationship between MES and the number of protein interactions (Figure 1b). In strong support of our hypothesis, the relationship between MES and the number of protein interactions almost perfectly complements that between MES and gene dispensability. KAFRI *et al.* (2005) also randomly paired genes and found that the relationship between MES and gene dispensability disappeared after random pairing. This is expected if the correlation between gene dispensability and MES is due to (i) the correlation between the number of protein interactions and gene dispensability and (ii) the correlation between the number of protein interactions and MES, because correlation ii should disappear when genes are randomly paired. As a consequence, the correlation between gene dispensability and MES disappears after random gene pairing.

It should be noted that although the numbers of protein interactions used in the above analysis may differ from the true values due to errors in high-throughput genomic studies and/or imperfect data interpretation (*e.g.*, the matrix or spoke strategy in transforming protein complex information to binary interaction information; see BADER and HOGUE 2002), the observation of a strong negative correlation between the number of protein interactions and gene dispensability (JEONG *et al.* 2001) suggests that the estimated number of protein interactions, or at least its relative value, is biologically meaningful. Note that this negative correlation between the number of protein interactions and gene dispensability exists for both singleton genes and duplicate genes (X. HE and J. ZHANG, unpublished results). Hence, the almost perfect complementation between the relationship of MES and gene dispensability and that of MES and the number of protein interactions strongly suggests the role of the number of protein interactions in causing the unexpected correlation between MES and gene dispensability among ancient duplicates. KAFRI *et al.* (2005) also examined the standard deviation of the correlation in expression response between duplicates under 40 conditions and termed it partial coregulation (PCoR). However, PCoR is highly correlated with MES (Pearson's correlation coefficient =  $-0.52$ ,  $P < 10^{-98}$ ). As expected, we found that the

relationship between PCoR and the number of protein interactions closely complements that between PCoR and gene dispensability (Figure 1, c and d).

We next examined whether the relationship between MES and gene dispensability observed by KAFRI *et al.* (2005) disappears when the number of protein interactions is controlled for. We also removed genes involved in stable protein complexes, because they tend to be similarly regulated and could have lower gene dispensability than those not involved in stable complexes even with the same number of protein interactions, due to the importance of dosage balance among components of stable complexes (PAPP *et al.* 2003; YANG *et al.* 2003). If the expression reprogramming hypothesis (KAFRI *et al.* 2005) is correct, the relationship between MES and gene dispensability should not change after the number of protein interactions is controlled for. But in fact, there are no consistent relationships between MES and gene dispensability among genes of different numbers of protein interactions (Figure 2a). Furthermore, the variation in gene dispensability at different MESs is diminutive for any given number of protein interactions (Figure 2a). A similar result is obtained between PCoR and gene dispensability (Figure 2b). Thus, MES and PCoR are uncorrelated with gene dispensability among ancient duplicates when the number of protein interactions is controlled for. Note that because the number of duplicate genes having a given number ( $N$ ) of protein interactions is too small when  $N > 5$  (*e.g.*, 30 for  $N = 6$ ), we limited our analysis to genes with  $N \leq 5$ , which constitutes  $\sim 92\%$  of all genes in our data. Lumping several bins together (*e.g.*,  $N = 6-10$ ) effectively removes the control and is not appropriate. Because the duplicate gene pairs used are not independent of each other (see DATA AND METHODS), the effect of the number of protein interactions cannot be statistically measured by a partial correlation analysis.

Our reanalysis of the yeast data suggests that the primary cause of the unexpected correlation between MES and gene dispensability among ancient duplicates is those duplicate pairs for which both proteins are involved in the same or related protein complexes. These duplicate genes tend to have similar expression (*i.e.*,

high MES) by coregulation, but low gene dispensability because of the relative importance of protein complexes. Why do such genes affect the relationship between MES and gene dispensability when  $d_s > 1$ , but not when  $d_s < 1$ ? We think that when  $d_s > 1$ , it is rare for a duplicate pair to have high MES unless they participate in the same or related protein complexes. To verify this conjecture, we separated ancient duplicate pairs into two groups on the basis of their involvement in protein complexes. Only 6.9% of complex-free duplicate pairs have MES  $> 0.6$ , while the number is 25.1% for complex-involving duplicates ( $\chi^2 = 75$ ,  $P < 10^{-17}$ ). When  $d_s < 1$ , many duplicates can have high MES without the involvement in protein complexes; their high MES is inherited from the common ancestor and has yet to be lost in evolution. In fact, 40.3% of complex-free young duplicates have MES  $> 0.6$ , in comparison to only 6.9% among complex-free old duplicates ( $\chi^2 = 89$ ,  $P < 10^{-20}$ ). In summary, our results show that the primary observation supporting the extraordinary evolutionary hypothesis of transcriptional reprogramming and backup among duplicates (KAFRI *et al.* 2005) was mainly due to a confounding factor unrelated to transcriptional reprogramming. In other words, there is no clear evidence suggesting that the reprogramming hypothesis is true at the genomewide level, although it may work in a few genes (KAFRI *et al.* 2005).

One of the interesting observations of KAFRI *et al.* (2005) is the relatively high prevalence of documented cases of synthetic lethality in paralogous pairs with low MES and high PCoR (KAFRI *et al.* 2005, Figure 5). However, this was based on a casual observation without systematic statistical comparisons. Most importantly, they failed to show that the presumable compensation among synthetic lethal genes is caused by transcriptional reprogramming. In fact, a recent study of synthetic lethal genes of *S. cerevisiae* showed that transcriptional reprogramming of a gene in responding to the deletion of its synthetic lethal partner is rare in general and virtually absent among duplicates (WONG and ROTH 2005).

One may ask if transcriptional reprogramming is not the mechanism, What would be the mechanism of functional compensation among ancient duplicates? We stress that it is important to distinguish between two distinct questions: (1) Is there prevalent functional compensation among duplicates? and (2) If there is, what is the molecular mechanism? Although the observation of a lower fitness effect caused by deleting a duplicate, rather than a singleton, gene is commonly attributed to functional compensation, there are other possible explanations. For example, intrinsically less important genes may have a higher duplicability than more important genes, a trend recently found in yeasts (HE and ZHANG 2006). If this trend is as strong as has been suggested, no functional compensation is necessary to explain the gene deletion results (HE and ZHANG 2006). From a mechanistic point of view, although it is possible that

functional compensation among young duplicates occurs frequently due to similar expressions and functions, compensation among ancient duplicates may be quite rare. Even if functional compensation does occur among ancient duplicates with low MES, it may not require transcriptional reprogramming. The reason is that low MES simply means that a duplicate pair shows different responses to given conditions, but does not mean that their expression patterns are mutually exclusive. We can imagine a scenario by which compensation could work for those low MES paralogs without transcriptional reprogramming. For example, A and B are a pair of duplicate genes. A has a high-low-high-low temporal expression pattern from phase M to phase G<sub>2</sub> in the cell cycle, while B is constitutively expressed with a constant expression level. Thus, A and B will show low MES. When A is deleted, B can most likely compensate A without reprogramming, because B is always expressed. This scenario is more parsimonious than transcriptional reprogramming in explaining functional compensation among ancient duplicates. It is possible that the overlapping expression of genes A and B is retained by certain constraints. For instance, B may perform another function that requires constitutive expression. These considerations notwithstanding, we caution that genome-wide studies of possible functional compensations among duplicate genes and the underlying molecular mechanisms have just begun (WAGNER 2000); more analyses are needed to fully address these important questions.

We thank Ran Kafri for supplying data and for discussion. Ondrej Podlaha, Wendy Grus, and two anonymous reviewers provided valuable comments. This work was supported by research grants from the National Institutes of Health and the University of Michigan to J.Z.

#### LITERATURE CITED

- BADER, G. D., and C. W. HOGUE, 2002 Analyzing yeast protein-protein interaction data obtained from different sources. *Nat. Biotechnol.* **20**: 991–997.
- CONANT, G. C., and A. WAGNER, 2004 Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. *Proc. Biol. Sci.* **271**: 89–96.
- GU, Z., D. NICOLAE, H. H. LU and W. H. LI, 2002 Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* **18**: 609–613.
- GU, Z., L. M. STEINMETZ, X. GU, C. SCHARFE, R. W. DAVIS *et al.*, 2003 Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**: 63–66.
- HAHN, M. W., and A. D. KERN, 2005 Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol. Biol. Evol.* **22**: 803–806.
- HE, X., and J. ZHANG, 2005 Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**: 1157–1164.
- HE, X., and J. ZHANG, 2006 Higher duplicability of less important genes in yeast genomes. *Mol. Biol. Evol.* **23**: 144–151.
- JEONG, H., S. P. MASON, A. L. BARABASI and Z. N. OLTVAI, 2001 Lethality and centrality in protein networks. *Nature* **411**: 41–42.
- KAFRI, R., A. BAR-EVEN and Y. PILPEL, 2005 Transcription control reprogramming in genetic backup circuits. *Nat. Genet.* **37**: 295–299.
- KAMATH, R. S., A. G. FRASER, Y. DONG, G. POULIN, R. DURBIN *et al.*, 2003 Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**: 231–237.

- MAKOVA, K. D., and W. H. LI, 2003 Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.* **13**: 1638–1645.
- OHNO, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, Berlin.
- PAPP, B., C. PAL and L. D. HURST, 2003 Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**: 194–197.
- VON MERING, C., R. KRAUSE, B. SNEL, M. CORNELL, S. G. OLIVER *et al.*, 2002 Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**: 399–403.
- WAGNER, A., 2000 Robustness against mutations in genetic networks of yeast. *Nat. Genet.* **24**: 355–361.
- WONG, S. L., and F. P. ROTH, 2005 Transcriptional compensation for gene loss plays a minor role in maintaining genetic robustness in *Saccharomyces cerevisiae*. *Genetics* **171**: 829–833.
- YANG, J., R. LUSK and W. H. LI, 2003 Organismal complexity, protein complexity, and gene duplicability. *Proc. Natl. Acad. Sci. USA* **100**: 15661–15665.
- YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**: 555–556.
- ZHANG, J., 2003 Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**: 292–298.

Communicating editor: S. YOKOYAMA