Gene Duplication and Evolution

Lynch and Conery (*1*) presented one of the first serious efforts to study the evolutionary fate of gene duplication using genomic sequence data. Their analysis led to several interesting observations, particularly with respect to the rate of gene duplication in eukaryotic genomes and the subsequent half-life of duplicates. These two parameters are of particular importance in studying the evolutionary processes of gene duplication and subsequent functional divergence. The most frequent class of duplications appeared to be similar in all six species, which suggests some silencing process for old duplicates. Several additional considerations in the analysis and interpretation, however, might have led to some different conclusions.

First, Lynch and Conery (*1*) used the number of substitutions per silent site, *S*, to measure the age of a duplicate-gene pair [figure 2 of (*1*)]. It is unclear, however, that silent divergence is a suitable proxy for a molecular clock involving different genes or gene duplicates. For example, Zeng *et al.* (*2*) reported 9- to 15-fold differences in *S* values and a flat distribution of *S* for 24 single-copy genes in *Drosophila*. Two points are important in this context: (i) this large variation in *S* is expected when the divergence time is low; and (ii) the divergence time for each comparison made by Zeng *et al*. (*2*) was fixed. Thus, for different genes, *S* may vary by more than an order of magnitude given a fixed divergence time. This situation differs from description of divergence time using *S* values from homologous genes across a group of organisms, in which a dependable molecular clock may exist. The same *S* values may represent duplicates of very different ages, and the different *S* values may be from duplicates of the same or similar ages. Thus, figure 2 of (1) should be viewed with caution as a description of the age distribution of gene duplications. A related issue is the reliability of estimates of *S*, because many of the values presented by Lynch and Conery (*1*) were larger than 1. Estimates larger than 1 are associated with a large variance due to saturation of substitutions and should generally be considered unreliable (*3*).

Second, the calculation of the half-life of gene duplicates was based on the untested, hidden assumption that the rate of gene duplication is constant over evolutionary time—an assumption implicit in both figure 3 and equation 3 of (*1*). Unfortunately, there are insufficient data with which to estimate the variation in the rate of gene duplication on a short time scale; nevertheless, there is some evidence that the duplication rate for some families may indeed not be stationary over a short evolutionary time. For example, in the mouse Sp100-rs family, a short lineage of *Mus musculus* has created at least 60 gene duplicates within 1.7 million years; other lineages such as the sibling taxa *Mus caroli*, a group that diverged 2.5 million years ago, contain few duplicates (*4*). If the duplication rate over the time during which divergence is observed is much lower than the recent rate of duplication, the half-life calculated by Lynch and Conery would represent a serious underestimate.

Finally, an alternative interpretation for the short half-life of duplicate genes before silencing may deserve consideration. Assuming that small values of *S* may more reliably reflect a short evolutionary time, the authors chose to estimate the half-life of duplicate genes only from gene pairs with *S* values in the range of 0 to 0.25. They estimated a mean half-life of 4 million years, concluding that "the fate awaiting most gene duplications appears to be silencing rather than preservation," and, hence, that "duplicate genes may only rarely evolve new functions." Yet their analysis appears to have ignored several important features of the data [figure 2 of (*1*)]. (i) Notwithstanding their model of "young" duplicates, the tails of the distribution are long and flat, which suggests that the data are actually heterogeneous. (ii) The proportions of the duplications that reside in the tails are high—85% for *Drosophila melanogaster*, 66% for *Caenorhabditis elegans*, and 65% for *Saccharomyces cerevisiae*. (iii) The tails include old and ancient duplications. The heterogeneity of the age distribution in figure 2 of (*1*) suggests that the short half-life calculated from young duplicate-gene pairs cannot be extended to most pairs. After all, a large proportion of these older duplicates may be much older than 4 million years, with real ages of tens or hundreds of million years. It is likely that these genes have been functional since their origin; otherwise, the duplicate sequences would have been deleted from the genome (*5*).

In addition, the absolute number of old or ancient gene duplicates is relatively large. For example, 40% of the approximately 13,600 coding sequences in the *D. melanogaster* genome appear to have arisen by gene duplication (*6*). Thus, some 34% of the fly genome, or 4624 genes [40% × 85% × 13,600, with the 85% from item (ii), above], comprise old or ancient duplicates. It is therefore misleading to assert that the vast majority of gene duplicates are quickly silenced, even if the calculation of the half-life is correct. Rather, it appears that the accumulation of "survivors" of the silencing process constitutes a large fraction of modern eukaryotic genomes.

An analogy for the application of half-lives is the mortality of newborns centuries ago: At that time the infant mortality rate was very high, because medical science was underdeveloped—but just because the "half-life" of newborns is short, it does not follow that half of all adults will die shortly. We suggest that figure 2 of (*1*) supports a conclusion opposite to the one that Lynch and Conery drew: A large proportion of duplicate genes either have evolved new functions (*7*) or have been maintained by subfunctionalization (*8*, *9*) or other mechanisms.

Manyuan Long
Department of Ecology and Evolution
University of Chicago
1101 East 57th Street
Chicago, IL 60637, USA

Kevin Thornton
Committee on Genetics
University of Chicago

References

1. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
2. L.-W. Zeng, J. M. Comeron, B. Chen, M. Kreitman, *Genetica* **102–103**, 369 (1998).
3. W.-H. Li, *Molecular Evolution* (Sinauer Associates, Sunderland, MA, 1997).
4. D. Weichenhan, B. Kunze, W. Traut, H. Winking, *Cytogenet. Cell Genet.* **80**, 226 (1998).
5. D. A. Petrov, E. R. Lozovskaya, D. L. Hartl, *Nature* **384**, 346 (1996).
6. G. M. Rubin *et al*., *Science*. **287**, 2204 (2000).
7. W. Wang, J. Zhang, C. Alvarez, A. Llopart, M. Long, *Mol. Biol. Evol.* **17**, 1294 (2000).
8. A. Force *et al*., *Genetics* **151**, 1531 (1999).
9. M. Lynch, A. Force, *Genetics* **154**, 459 (2000).

21 December 2000; accepted 21 June 2001

Lynch and Conery (*1*) have proposed a number of provocative hypotheses regarding the evolution of duplicate genes, using data from nine eukaryotic species. One hypothesis is that the ratio of replacement (*R*) to silent (*S*) nucleotide substitutions among recently duplicated genes is near 1.0, the neutral expectation. Their analysis indicates that this phase of relaxed selection is confined to recently duplicated gene pairs. Another hypothesis is that many duplicate-gene pairs are short-lived, with half-lives of 3 to 7 million years, depending on the organism.

Unfortunately, their conclusions are compromised by the fact that their data, obtained through GenBank taxon searches, included many redundant records. For example, 43.3% of the gene pairs in their *Arabidopsis* data set had no synonymous differences ($S = 0$). We randomly examined 50 of these gene pairs and found that 86% were derived from the same genomic sequence, mostly because of the presence of a single gene on two overlapping clones. These redundant sequences were used to estimate the rate at which duplicate-gene pairs reverted to single copies, a procedure that tended to overestimate the rate of gene loss. Such problems were not limited to

Arabidopsis; 58.3% of human gene pairs and 67.7% of mouse gene pairs had $R = S = 0$. Because Lynch and Conery recognized the potential problem of redundancy, human and mouse gene pairs with $S < 0.01$ were not used in their analyses. In many cases, however, both gene sequences from an $S < 0.01$ pair were compared with a more distant gene family member, which did result in the use of redundant data entries.

Also problematic are the mammalian gene pairs in the $0.01 < S < 0.05$ class, which were crucial to the conclusion by Lynch and Conery (*1*) that selective constraint is temporarily relaxed after gene duplication. We manually inspected 20 pairs from each species and found that 50% in human and 80% in mouse are actually allelic or alternatively spliced forms of the same locus. Allelism was determined primarily by GenBank annotation, which provided the same gene name for different sequence entries. These observations suggest that several data sets were problematic and cast doubt on the value of the analyses presented by Lynch and Conery.

There are additional problems with the approach used by Lynch and Conery (*1*). First, the authors applied an exponential-decay model to estimate the rate of gene turnover, assuming a steady state between the origin and loss of duplicated pairs. In yeast and *Arabidopsis*, however, this assumption has clearly been violated by episodic, large-scale genomic duplication events (*2*, *3*). Second, their model failed to account for the fact that the number of pairwise comparisons within gene families can be substantially larger than the number of actual duplication events. To estimate the rate of gene loss, one needs to know the distribution of the latter, not the former. Finally, the authors proposed a curvilinear model for the relationship of *R* to *S*, but failed to test that model against the null hypothesis that *R* is a simple linear function of *S*. In our reanalyses, we found that the curvilinear model fits significantly better than the linear model for all nine species, but we obtained substantially different parameter estimates, with smaller sums of squares than those reported. Our results, which appear to support the curvilinear model, will require independent verification in light of the problems with several data sets.

Although they have not succeeded in demonstrating empirical support for all of their hypotheses, Lynch and Conery nonetheless have offered a variety of stimulating ideas—the apparently high rate of gene duplication, the role of duplication in chromosomal repatterning, and the role of gene duplication in reproductive isolation between species—that call for further investigation.

Liqing Zhang
Brandon S. Gaut
Department of Ecology and
Evolutionary Biology
University of California, Irvine
Irvine, CA 92697, USA

Todd J. Vision
USDA-ARS Center for Agricultural
Bioinformatics
Cornell University
Ithaca, NY 14853, USA

References

1. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
2. K. H. Wolfe, D. C. Shields, *Nature* **387**, 708 (1997).
3. T. J. Vision, D. G. Brown, S. D. Tanksley, *Science* **290**, 2114 (2000).

5 March 2001; accepted 21 June 2001

Response: Our understanding of the evolutionary dynamics of duplicate genes stems largely from a handful of studies on gene families known to have functions that may enhance the likelihood of evolutionary diversification [see references in (*1*)]. These studies, though important, may lead to bias in our understanding of the usual fate of gene duplicates. To avoid this problem, our study (*1*) exploited the information inherent in fully sequenced genomes to evaluate the average evolutionary properties of the members of duplicate-gene pairs. Although this alternative approach glossed over the specific properties of individual gene pairs, the emergent patterns provided a broad description of the types of observations that a general theory for the evolution of duplicate genes must explain.

Long and Thornton have three concerns with our analyses. First, they argue that variation among genes in the rate of nucleotide substitution at silent sites reduces the reliability of the number of substitutions per silent site, *S*, as an indicator of the age of a duplicate pair. As stated in (*1*), we confined our analyses of the birth and death rates of gene duplicates to pairs with $S < 0.25$ because of the high statistical uncertainty associated with large estimates of *S*. Long and Thornton state that this problem may also be serious for pairs of duplicates with low *S*, but with only two exceptions, the study to which they refer (*2*) showed that the range of *S* for the 24 genes studied in the youngest species pair (*Drosophila subobscura* and *D. psuedoobscura*) is 0.09 to 0.50. An unknown fraction of this variation must simply be due to statistical sampling error. In any event, variation among loci in the rate of silent-site substitution would have the effect of dispersing the data points in figures 2 and 3 of (*1*) over the horizontal axis relative to the positions expected on the basis of actual ages of the pairs. This would cause our half-life estimates to be upwardly biased, but would not alter the basic conclusions reported in (*1*).

Second, Long and Thornton argue that our demographic analysis of gene duplicates made the hidden assumption of a constant rate of gene duplication over time. This assumption was actually stated explicitly in (*1*), although strictly speaking we only assumed rate constancy over the time scale for which $S < 0.25$. Long-term rate constancy was not relevant to our birthrate estimates, which were simply the average values that apply over the time scale required for *S* to reach 0.01—perhaps the past few hundred thousand to million years for the species analyzed. Rate constancy is an important assumption underlying the use of the slope of an age distribution to estimate a half-life. We note, however, that there appears to be no intrinsic reason why our half-life estimates should be biased in one direction versus the other by temporal variation in birth or mortality rates. In contrast to the scenario painted by Long and Thornton, a recent reduction in the rate of origin of duplicates would result in a flatter age distribution and an overestimate of the half-life.

Third, Long and Thornton question our assertion that the vast majority of gene duplicates enjoy a rather short half-life, arguing that many ancient pairs of duplicates can be found in most eukaryotic genomes. In principle, the probability of loss of a duplicate gene may progressively decline once preservational events such as neofunctionalization or subfunctionalization begin to take hold. However, the argument adduced by Long and Thornton themselves about inaccuracies in the estimation of *S* provides the answer for why the long, flat profile at the high end of the age distribution cannot provide quantitative insight into rates of duplicate-gene silencing—because the sampling variance of *S* becomes increasingly large with increasing *S*, the age distribution will become artificially flattened at high *S*. Ignoring the plateau on the right, as we did, will cause a slight overestimate of the initial downward slope, which will cause a slight underestimate of the average half-life of the vast majority of duplicates that appear to be quickly silenced.

Statistical problems aside, there is a fundamental problem with using the observation that "the accumulation of 'survivors' of the silencing process constitutes a large fraction of modern eukaryotic genomes" to support the claim that a large proportion of newborn gene duplicates become permanently preserved. The ancient duplicates to which Long and Thornton refer are the rare survivors of duplication events that have accumulated over vast periods of time (many tens to hundreds of millions of years). In the long run, each origin of a gene must be balanced by the loss of another to prevent indefinite genome expansion or contraction.

Many of the gene duplicates that we have identified with $S > 1$ may have arisen by processes substantially different from the incremental single-gene duplications that we focused on in (*1*). Most notable is the process of complete genome duplication, the ancient remnants of which have been implicated in yeast (*3*) and *Arabidopsis* (*4*). Although the genomic extent is not yet understood, a massive amount of gene duplication occurred early in the vertebrate lineage (*5*), and we cannot rule out the possibility of similar large-scale events prior to the radiation of the animal phyla. The probability of duplicate-gene preservation following polyploidization may be substantially elevated relative to that for single-copy duplicates for two reasons. First, as we noted previously, polyploidization maintains the dosage ratios of all pairs of genes relative to the situation in the diploid state, and selection may favor the maintenance of the ancestral stoichiometric ratios. Second, when whole chromosomes are duplicated, the constituent genes are guaranteed to initiate with all essential regulatory regions intact, and this may further reduce the likelihood of negative selection against new copies.

Zhang *et al.* argue that three of the data sets that we worked with in (*1*) contained flaws that may have influenced the outcome of our analyses. We agree that this issue merits close scrutiny, and at the close of this response, we will present some reanalyses for both the *Arabidopsis* and human genomes that take into consideration the concerns raised by Zhang *et al*. First, however, we respond to three technical issues raised by these authors:

1) As noted in (*1*), the inability to easily distinguish allelic sequences or alternative spliced forms from duplicate genes raises potential complications with some databases. This is unlikely to be a serious problem with inbred species such as *C. elegans* or haploid species such as *S. cerevisiae*, whose genomic sequences are well annotated and curated. For outbred species, it is difficult to see how one can unambiguously resolve this issue with data sets constructed from random sequences (contrary to the suggestion by Zhang *et al.*), and 5% sequence divergence seems rather high for allelic variants. Nevertheless, this problem remains a serious consideration for data sets that are not highly refined. If nonduplicate sequences are inadvertently included in a survivorship analysis for duplicate loci, the estimated half-life will be unaffected so long as the incidence of such errors is independent of the degree of divergence for the genes involved in the analysis, but the estimated rate of origin of new duplicates will be inflated.

2) As noted above, the ancient genome duplications known to have occurred in *Arabidopsis* and yeast have no bearing on our conclusions, because the duplicate pairs associated with these events were not included in our demographic analyses.

3) The distinction raised by Zhang *et al*. between numbers of extant duplicate pairs and number of actual duplication events is correct and important. However, multigene families were excluded from our analyses and the vast majority of the young gene duplicates that we identified were simple pairs (in which case, there is no ambiguity with respect to event counting), so this distinction has little effect on our estimates. Nevertheless, the reanalyses presented below are based on estimates of duplication events rather than on observed numbers of duplicate pairs.

After publication of (*1*), a well-curated version of the *Arabidopsis* genome became available (*6*) that has eliminated most of the redundancies and ambiguities noted by Zhang *et al*. A complete reanalysis of the data is beyond the scope of this response and will be reported elsewhere (*7*); to summarize, however, using our prior methods for demographic analysis, we have estimated the rate of origin of new duplicates in *Arabidopsis*, based on the new data set, to be 0.0022 per gene per million years, which is of the same order of magnitude as that observed for *D. melanogaster* (0.0023) and yeast (0.0083), but lower than that for *C. elegans* (0.0208). Because the incidence of putative duplicates in the nearly identical class is greatly reduced in this newly available data set (consistent with the arguments of Zhang *et al.*), the half-life estimate increases from our previous value of 3.2 million years to 23.4 million years, which exceeds our previous estimates for invertebrates by a factor of seven and for mammals by a factor of three.

We are also now able to provide an estimate of the rate of origin of new duplicates in the human genome, using the database of the publicly funded project (*8*, *9*). Our estimate, 0.0071 per gene per million years, falls in the middle of the range for other species. Our revised estimate of the half life for human duplicate genes, 16 million years, is about double our previous estimate. However, because assembly problems probably result in the exclusion of substantial numbers of young gene duplicates from the "complete" human genomic sequence (*9*, *10*), our estimated rate of origin of new duplicates in humans is probably downwardly biased, whereas our estimated half-life is likely upwardly biased. A recent comparison of chromosomal contents in mice and humans strongly supports our contention that a high rate of duplicate-gene turnover occurs in mammals (*11*).

One must be cautious to avoid overinterpreting the degree of precision associated with all of these estimates; most large-scale genome projects are still in a stage of maturation, with updated annotations being released regularly. At this point, however, we see no reason to alter our basic conclusions that the rate of origin of new duplicates in eukaryotes is quite high, often in the range of 0.002 to 0.020 per gene per million years, and that most gene duplicates have a relatively short life-span, the average being in the neighborhood of 1 to 10 million years (with a possible exception in *Arabidopsis*). Functional studies will be required to determine the fraction of duplicates identifiable from coding-region identity that are actually biologically active.

Michael Lynch
Department of Biology
Indiana University
Bloomington, IN 47405, USA
E-mail: mlynch@bio.indiana.edu
John C. Conery
Department of Computer and
Information Science
University of Oregon
Eugene, OR 97403, USA

References and Notes

1. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
2. L.-W. Zeng, J. M. Comeron, B. Chen, M. Kreitman, *Genetica* **102-103**, 369 (1998).
3. K. H. Wolfe, D. C. Shields, *Nature* **387**, 708 (1997).
4. T. J. Vision, D. G. Brown, S. D. Tanksley, *Science* **290**, 2114 (2000).
5. A. Sidow, *Curr. Opin. Genet. Dev.* **6**, 715 (1996).
6. www.tigr.orgtdb/e2k1/ath1/
7. A complete list of data underlying the conclusions discussed in these paragraphs appears at http://csi.uoregon.edu/projects/genetics/duplications/letters.
8. www.ensembl.org
9. International Human Genome Sequencing Consortium, *Nature* **409**, 860 (2001).
10. J. C. Venter *et al.*, *Science* **291**, 1304 (2001).
11. D. Paramvir *et al.*, *Science* **293**, 104 (2001).

12. We thank B. Haas and S. Salzberg for help in clarifying issues involving the TIGR database for *Arabidopsis* gene sequences and E. Birney for assistance in interpreting the Ensembl database resulting from the work of the International Human Genome Sequencing Consortium.

30 April 2001; accepted 31 July 2001