Journal of Structural and Functional Genomics **3:** 35–44, 2003.

The evolutionary demography of duplicate genes

Michael Lynch[1*] & John S. Conery[2]
[1]Dept. of Biology, Indiana University, Bloomington, Indiana 47405; [2]Dept. of Computer and Information Science, University of Oregon Eugene, Oregon 97403

Received 21.05.2002; accepted in final form 29.08.2002

Key words: gene duplication, genome evolution, genome size

Abstract

Although gene duplication has generally been viewed as a necessary source of material for the origin of evolutionary novelties, the rates of origin, loss, and preservation of gene duplicates are not well understood. Applying steady-state demographic techniques to the age distributions of duplicate genes censused in seven completely sequenced genomes, we estimate the average rate of duplication of a eukaryotic gene to be on the order of 0.01/gene/million years, which is of the same order of magnitude as the mutation rate per nucleotide site. However, the average half-life of duplicate genes is relatively small, on the order of 4.0 million years. Significant interspecific variation in these rates appears to be responsible for differences in species-specific genome sizes that arise as a consequence of a quasi-equilibrium birth-death process. Most duplicated genes experience a brief period of relaxed selection early in their history and a minority exhibit the signature of directional selection, but those that survive more than a few million years eventually experience strong purifying selection. Thus, although most theoretical work on the gene-duplication process has focused on issues related to adaptive evolution, the origin of a new function appears to be a very rare fate for a duplicate gene. A more significant role of the duplication process may be the generation of microchromosomal rearrangements through reciprocal silencing of alternative copies, which can lead to the passive origin of post-zygotic reproductive barriers in descendant lineages of incipient species.

For practical reasons, much of the past focus on genome evolution has been on divergence at the nucleotide level in specific genes. But with the growing proliferation of whole-genome sequences, a more global view of genomic evolution is beginning to emerge. Just as nucleotide changes continuously arise within populations via mutation, accidents at the level of chromosomal regions regularly give rise to losses and duplications of entire genes. Such genomic turnover is ultimately responsible for interspecific divergence in gene content, which may be exploited for adaptive reasons, and for modifications of gene location, which may passively give rise to post-zygotic reproductive isolating barriers (for review, see Lynch, 2002). Thus, it is of some interest to determine the rate at which new genes arise via duplication events and the frequency and mechanisms by which they are preserved.

Because of the difficulties with quantifying low probability events at the molecular level, we are almost completely lacking in direct estimates of the rate of gene duplication, although rates as high as 10^{-6} to 10^{-4} per gene per generation have been reported for *Drosophila* (Shapira and Finnerty 1986). We recently obtained indirect estimates of the rates of birth and loss of new genes through censuses of the contents of the then largely sequenced nuclear genomes of several eukaryotes (Lynch and Conery 2000), and additional estimates using somewhat different criteria have been published by Gu *et al.* (2002). Since these studies were performed, nearly complete genomic sequences have emerged for several species and all of the pre-existing databases have been refined considerably. We, therefore, take this opportunity to update and expand our previous results.

Sources of data and methods of analysis

For each of the fully sequenced eukaryotic genomes, we downloaded all coding sequences and their corresponding amino-acid sequences from the most recently curated database (as of 1 April 2001), removing all suspected pseudogenes, transposable elements, and overlapping genes prior to subsequent analyses: *Schizosaccharomyces pombe* – The Sanger Centre (ftp://ftp.sanger.ac.uk/pub/yeast/sequences/pombe); *Saccharomyces cerevisiae* – National Center for Biotechnology Information ftp://ftp.ncbi.nih.gov/genbank/genomes/S_cerevisiae; *Arabidopsis thaliana* – The Institute for Genomic Research (ftp://ftp.tigr.org/pub/data/athaliana/ath1); *Caenorhabditis elegans* – WormBase (http://www.wormbase.org); *Drosophila melanogaster* – Berkeley *Drosophila* Genome Project (http://www.fruitfly.org/sequence/download.html); and *Homo sapiens* – The Ensembl Project (ftp://ftp.ensembl.org/current/data/).

To identify duplicate genes, we used BLAST (Altschul *et al.* 1997) to compare all pairs of amino-acid sequences within each genome, retaining only those pairs for which the alignment score was below 10^{-10}. To minimize the inclusion of members of large multigene families, we excluded all genes that identified more than five matching sequences. Using each protein alignment generated by BLAST as a guide, we aligned the nucleotide sequences, and then prior to sequence analysis, we used a gap-expansion algorithm to remove ambiguous portions of the alignments (Conery and Lynch 2001).

The numbers of nucleotide substitutions per silent and replacement sites (S and R, respectively) were then estimated for each pair by using the maximum-likelihood procedure in the PAML software package (version 2.0k) (Yang 1997). Estimated rates of nucleotide substitution are sensitive to the relative rates of occurrence of transitions and transversions, which cannot be estimated accurately when the amount of sequence divergence is high. Therefore, to obtain precise estimates of the transition/transversion bias among newly arisen mutations, prior to the analyses of sequence divergences for each species, we tallied the observed substitutions at all four-fold redundant sites in all pairs of duplicate sequences that were similar enough that multiple substitutions per site were unlikely (by confining these computations to loci for which the divergence at such sites was $\leq 15\%$, after verifying that the transition/transversion ratio is essentially constant below this point). Each species-specific estimate of the transition/transversion ratio was then treated as a constant in the maximum-likelihood analyses.

In genome-wide surveys, there is a need to distinguish the number of duplication events from the number of duplicate pairs. When neither member of a duplicate pair is homologous to another gene in the data set, then the pair represents a single duplication event. However, when three or more genes are mutually related, the number of duplication events is necessarily less than the number of pairs. For example, a trio of related genes revealed as three pairs must be the result of two duplication events. With a closed loop of three similar genes, the ancestral relationships are ambiguous, so we counted each pair as two-thirds of a duplication event. For cases in which four or more genes constituted a family, we constructed a graph with nodes corresponding to genes and edges connecting two nodes whenever the PAML estimate of S was less than 5.0 for the corresponding genes. For a particular family of such genes, all possible spanning trees (excluding closed loops) were constructed (Cormen *et al.* 1990, Shioura *et al.* 1997), and the weight for each gene pair was taken to be the fraction of times the edge connecting the pair was used in the total set of trees for the family.

The age distribution of duplicate genes

Assuming the number of silent substitutions per site increases approximately linearly with time, the relative age-distribution of gene duplicates within a genome can be inferred indirectly from the distribution of S (Figure 1). As in more traditional forms of demographic analysis, the forms of such distributions are diagnostic. For all species, the highest density of duplicates tends to be contained within the youngest age classes, with the density dropping off rapidly with increasing S (Figure 1). A smooth, nearly exponential decay with age is seen for both *H. sapiens* and *C. elegans*, for which the sample sizes of duplicate genes are very large. Similar distributions are seen for the two yeasts, although these are less smooth presumably because of the lower incidence of duplicate genes in these species. On the other hand, *A. thaliana* is exceptional in showing a pronounced secondary peak in the age distribution at $S \simeq 0.75$. This bulge

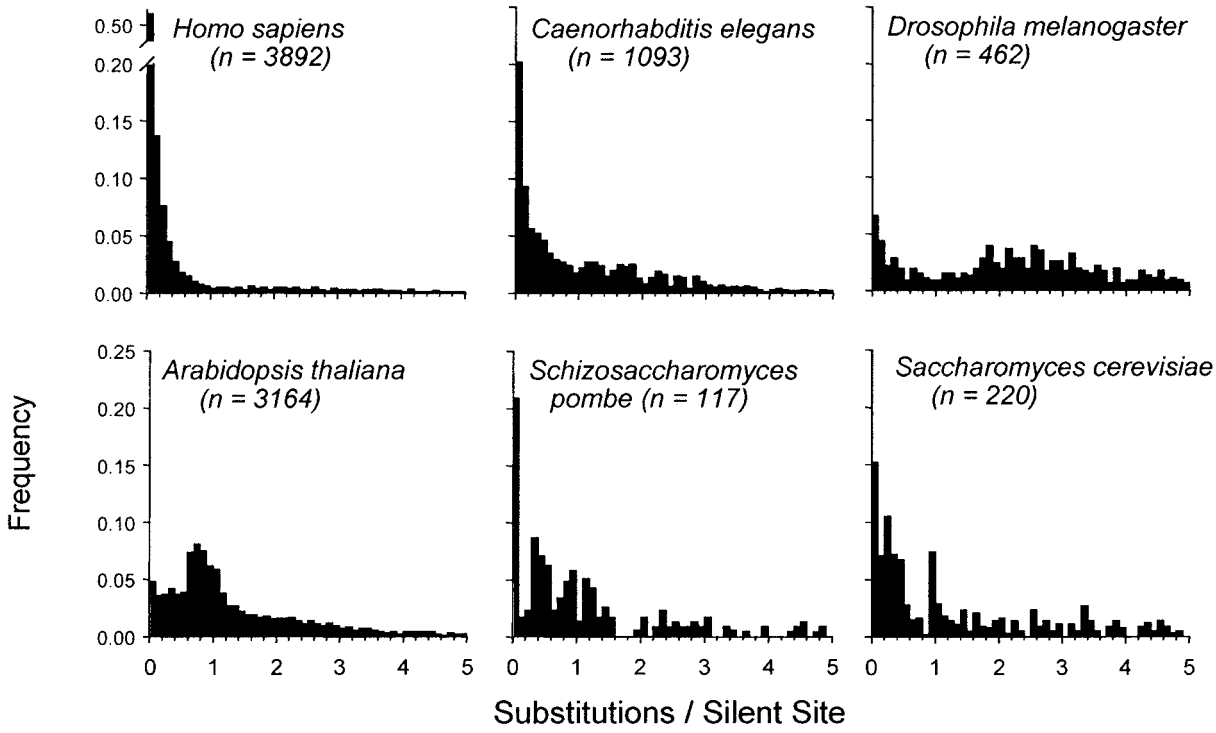

Figure 1. The age distribution of duplicate pairs of genes in six completely sequenced eukaryotic genomes. The total number of detected duplication events for each species is given in parentheses.

apparently reflects a 'baby boom' of genes that resulted from an ancient polyploidization event in the ancestor of *Arabidopsis* (Grant *et al.* 2000, Bevan *et al.* 2001). Using an estimated rate of silent-site substitution of 6.1/silent site/BY (an average of two independent estimates for vascular plants reported in Lynch (1997) and Li (1999), $S = 0.75$ is equivalent to approximately 60 million years of divergence.

Estimation of rates of birth and death

One interpretation of the taxonomically general patterns illustrated in Figure 1 is that of a birth-death process, with the very youngest age category representing newly arisen duplicates and the subsequent decline in frequency resulting from mutational processes that eliminate gene function (including large-scale deletions, frame shifts that introduce stop codons, etc.) Provided adequate numbers of young gene duplicates are available, estimates of the rates of birth and loss of such genes can be obtained directly from the observed age distribution, under the assumption that these rates have been essentially constant within the age classes employed in the analysis.

Letting n_t be the number of copies of a gene present at time t (in excess of the baseline number of one), then the dynamics of temporal change in the incidence of gene duplicates can be expressed as

$$n_t = n_{t-1} + B(1 + n_{t-1}) - Dn_{t-1}, \qquad (1)$$

where we assume that only the excess copies of a gene are subject to loss. Setting $n_t = n_{t-1}$, the equilibrium number of additional copies per gene is found to be

$$n_{tot} = \frac{B}{D - B}. \qquad (2)$$

Noting that the expected number of newborns per interval is $B(1 + n_{tot})$, the expected number of duplicates in the ith age category is

$$n_i = \frac{BD(1 - D)^i}{D - B}. \qquad (3)$$

The slope of the least-squares regression of $\ln n_i$ on S provides an estimate of the instantaneous mortality

rate $\hat{d}$, defined such that the probability of duplicate-gene loss by the time divergence at silent sites has reached S is

$$\hat{D} = 1 - e^{-\hat{d}S}. \quad (4)$$

The estimated half-life is then

$$\hat{S}_{0.5} = -\frac{\ln 0.5}{\hat{d}}. \quad (5)$$

Letting n_B be the number of duplicate pairs observed below some low level of S, then the birth rate over this span of divergence can be estimated as

$$\hat{B} = \frac{n_B \hat{d} S}{N(1 - e^{-\hat{d}S})} \quad (6)$$

where N is the total number of genes in the analysis (not including the excess duplicates), and the term $\hat{d}S/(1-e^{-\hat{d}S})$ accounts for the loss of duplicates within the range of divergence up to S.

In the analyses reported here, we let $S = 0.01$ in Equations (4) and (6), so $\hat{B}$ and $\hat{D}$ are estimated rates of birth and loss of duplicates over the time scale required for a pair of genes to diverge by 1% at silent sites. We also restricted our entire demographic analyses to duplication events with $S \leq 0.10$, so strictly speaking the assumptions regarding stationarity of rates are only relevant to this range of divergence. Consistent with this assumption, the log-arithmetic plots of n_i on S are approximately linear, although there is considerable scatter for the fungal genomes for which the numbers of duplication events are small (Figure 2).

The average half-lives of gene duplicates in the three fungal species are in the narrow range of $0.01 \leq \hat{S}_{0.05} \leq 0.02$, whereas those for the three metazoans are in the higher range of $0.04 \leq \hat{S}_{0.05} \leq 0.10$, and that for *A. thaliana* is still higher at $\hat{S}_{0.05} \simeq 0.21$ (Table 1). These half-lives in units of S can be crudely rescaled to absolute time by assuming an approximately constant rate of silent substitution. Using the rationale outlined in Lynch and Conery (2000), we assume a rate of 2.5/site/BY for human, 15.6/site/BY for invertebrates, 6.1/site/BY for *A. thaliana*, and 8.1/site/BY for fungi. Thus, the approximate half-lives for gene duplicates in humans, flies, and nematodes are 7.5, 3.2, and 1.7 MY, respectively, for an overall average of about 4 MY for animals. The average half-life for *A. thaliana* duplicates is much higher, on the order of 17.3 MY, whereas that for the three fungi is much lower, averaging to ~ 1.0 MY.

Table 1. Estimated mortality rates, half lives, and birth rates. $\hat{B}$ and $\hat{d}$ are estimated on a time scale for which silent-site divergence is 1%.

Species	$\hat{d}$ (SE)	$\hat{S}_{0.5}$	$\hat{B}$ (SE)
H. sapiens	18.4 (4.6)	0.038	0.0345 (0.0032)
C. elegans	13.0 (3.6)	0.053	0.0097 (0.0008)
D. melanogaster	6.9 (1.7)	0.100	0.0006 (0.0002)
A. thaliana	3.3 (2.0)	0.212	0.0032 (0.0005)
S. cerevisiae	30.5 (11.2)	0.023	0.0044 (0.0010)
S. pombe	42.6 (15.4)	0.016	0.0050 (0.0012)
E. cuniculi	62.3 (11.2)	0.011	0.0364 (0.0053)

Considerable interspecific variation also appears to exist for the rate of birth of duplicate genes, with the range being on the order of 0.001 to 0.04 over a time span equivalent to 1% divergence as silent sites (Table 1). At the high end, with essentially identical estimates of 0.035, are *H. sapiens* and *E. cuniculi*, whereas the remaining species fall in the range of 0.001 to 0.01. Using the molecular clocks noted above, these estimates translate into 0.009/gene/MY for humans, 0.016 for *C. elegans*, 0.001 for *D. melanogaster*, 0.002 for *A. thaliana*, 0.004 for both yeasts, and 0.030 for *E. cuniculi*. Thus, averaging over all taxa, the probability of duplication of a eukaryotic gene is at least 1% per million years. This rate is on the order of, if not greater than, most estimated rates of nucleotide substitution at silent sites (Li 1999). It is conceivable, however, that some of the duplicates that we have identified are 'dead-on-arrival.' For newborn duplicates in *C. elegans*, for example, we find that approximately one-third are complete over the entire coding region, with the remainder exhibiting one or more unique exons in one or both copies (Katju and Lynch, in prep.)

Given the potential errors in whole-genome sequences, all of the above estimates must be taken as provisional. However, there is no obvious reason to expect either the birth- or death-rate estimates to be upwardly biased, and in fact, the contrary may be true.

Whole-genome sequencing, particularly that using shot-gun approaches, is likely to lead to the exclusion of some of the youngest duplicates from the final genomic sequence, falsely interpreting them as simple redundant (or allelic) sequences. Such problems,

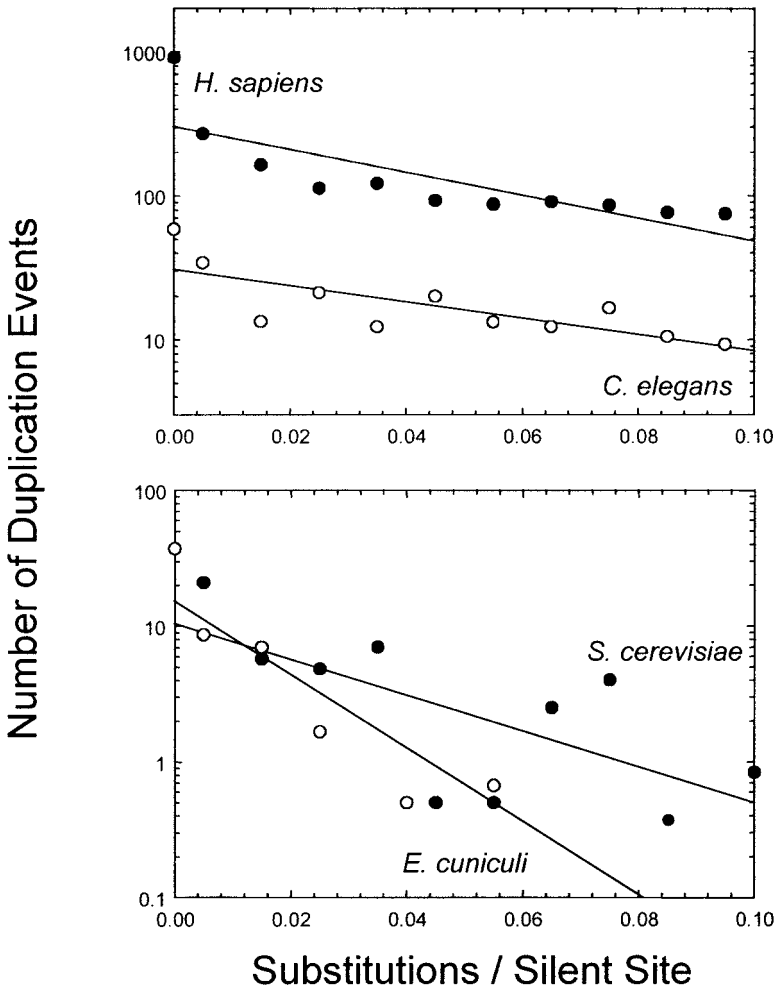

Figure 2. Survivorship curves fitted to the youngest cohorts of gene duplicates for two metazoans and two fungi.

which are most likely for the human and *Drosophila*genomes, would cause downward bias in the estimates of both the birth and death rates. In addition, our birth- and death-rate estimates may be somewhat downardly biased because we have ignored large multigene families.

Thus, the overall interpretation of these full-genome analyses, consistent with the earlier conclusions of Lynch and Conery (2000), is that the gene duplication is at least as significant as nucleotide substitution as an on-going contributor to genome evolution. Multiplying the species-specific duplication rates per gene by the genome size, the average number of newborn duplicates arising on a time scale of one million years is approximately 100 per genome. Roughly speaking, this implies that over a time span of 100 of 200 MY, nearly all of the genes within an average eukaryotic genome will have had an opportunity to duplicate. On the other hand, over 50% of such duplicates are likely to be silenced in only a few million years and most of the remainder shortly thereafter.

Evidence for genomic equilibrium

To test whether the incidences of gene duplicates in various genomes correspond to expectations under a long-term-state birth-death process, the predicted $\hat{n}_{tot}$, obtained by use of Equation (2), can be compared with the observed abundances of duplications. Such comparisons are somewhat subjective in that it is not

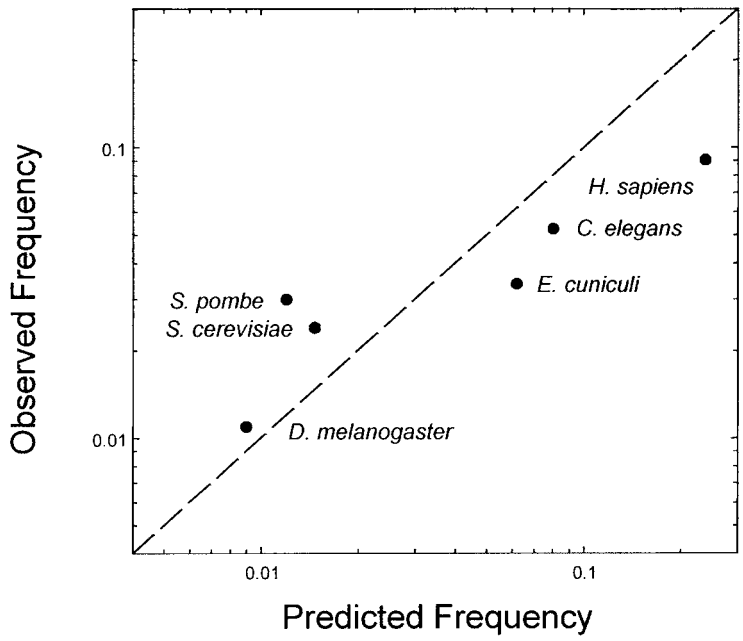

Figure 3. Comparison of the observed numbers of duplicates per gene with expectations based on a steady-state process, obtained by applying the demographic parameters in Table 1 to Equation (2). The diagonal dashed lines denotes the line of equivalence expected if the model perfectly fit the data.

entirely clear where to draw the upper boundary (with respect to *S*) for the subset of duplicates that are simply subject to the birth-death process (as opposed to the minority that are permanently preserved by positive selection). Nevertheless, there is a good correlation between observed and expected values for a broad range of cutoff values for *S*. We simply show the results for the pairs of duplicates with $S < 1.0$ in Figure 3, as estimates of *S* beyond this point are highly unreliable due to saturation effects. (We have excluded *A. thaliana* from this particular analysis because its ancestral polyploidization event clearly violates the assumption of equilibrium).

Note that because of its high rate of duplication, the microsporidian *E. cuniculi* actually has an equilibrium number of duplicates per gene approaching that of human and nematode, despite the fact that *E. cuniculi* has the smallest genome of any of the study species. On the other hand, because of its exceptionally low birth rate and moderately high loss rate, the *D. melanogaster* genome is rather depauperate with respect to gene duplicates.

Estimation of patterns of selective constraints

It is often assumed that redundancy of genes results in a relaxation of selection in one or both copies at least early in their history, and that this somehow enables one copy to take on a new function that would not otherwise be possible (e.g., Ohno 1970). From a population-genetics perspective, it is difficult to see how natural selection could isolate one gene for evolutionary exploration while keeping the other constant. This issue might be evaluated by considering the historical development of replacement- and silent-site substitutions in duplicates from the time of birth to the time of preservation or elimination, but the time scale of the mutational process necessitates an alternative approach. The problem that we are confronted with is that the observed estimates of *R* and *S* for any pair of extant duplicates are the cumulative outcomes of the joint evolutionary pressures operating on both loci since the initial duplication event. Such estimates potentially average over heterogeneous phases of molecular evolution, to an extent that increases with the age of the pair. However, some insight into the average temporal dynamics of selection can be acquired by examining the joint distribution of *R* and *S* for the

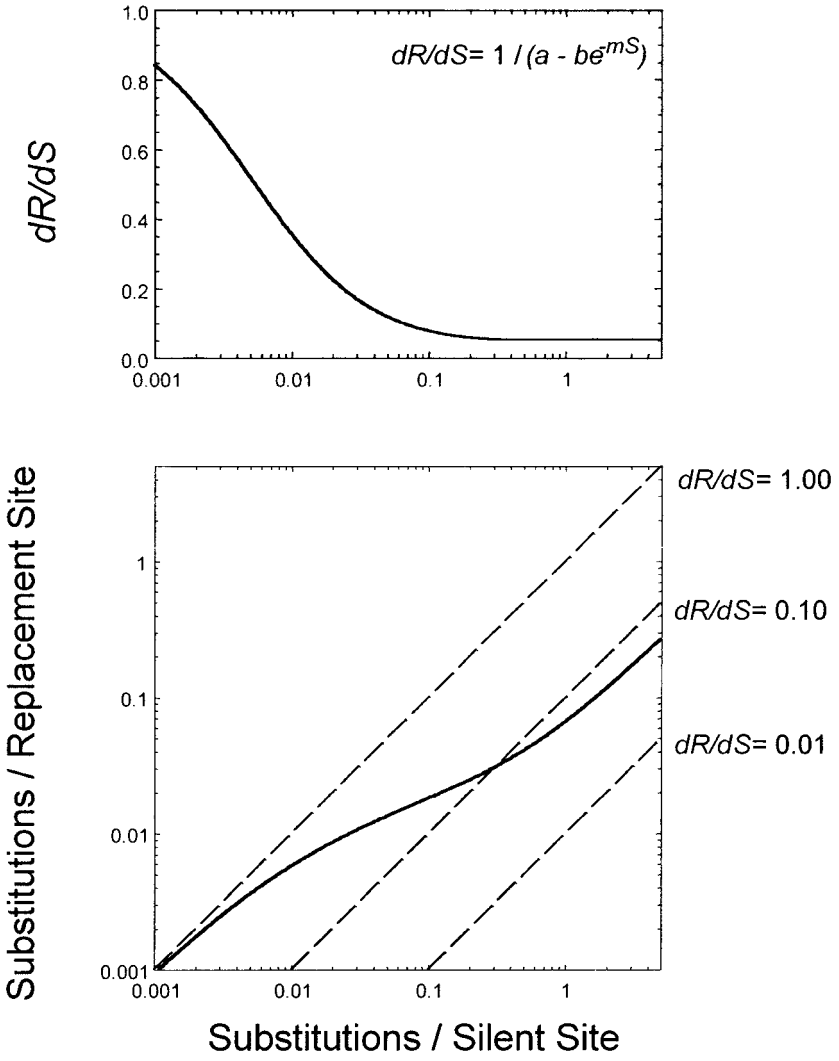

Figure 4. The change in the *R/S* ratio with increasing evolutionary time (measured in units of *S*). Upper panel: The instantaneous ratio of replacement to silent substitutions, defined by Equation (7) with $a = 20$, $b = 19$, and $m = 10$. Lower panel: The cumulative behavior of the *R* vs. *S* ratio, defined by Equation (8). Here, the dashed lines represent points of equal *R/S*. In this particular example, the ratio of replacement to silent substitutions initiates at 1.0 for newly arisen duplicates and gradually declines to a stable ratio of 0.05 as $S \rightarrow \infty$.

entire assemblage of gene duplicates within a species, under the assumption that approximately the same average temporal pattern of selection intensity operates on all cohorts of duplicates. If, for example, the intensity of purifying selection operating on duplicate genes typically increases with the age of a pair, this should be reflected in a reduction in the *R/S* ratio with increasing *S*, whereas the opposite is expected if selection is progressively relaxed.

To account for such behavior, we describe the ratio of the instantaneous rates of replacement and silent substitutions by the function.

$$\frac{dR}{dS} = \frac{1}{a - be^{-mS}}, \tag{7}$$

where a, b and m are constants (Lynch and Conery 2000). This function allows for two different phases of divergence, as well as a gradual transition between them. Assuming positive m, the ratio of rates of replacement to silent substitutions initiates with an expected value of $1/(a - b)$ at $S = 0$ (newly arisen duplicates) and declines to $1/a$ as $S \rightarrow \infty$ (ancient

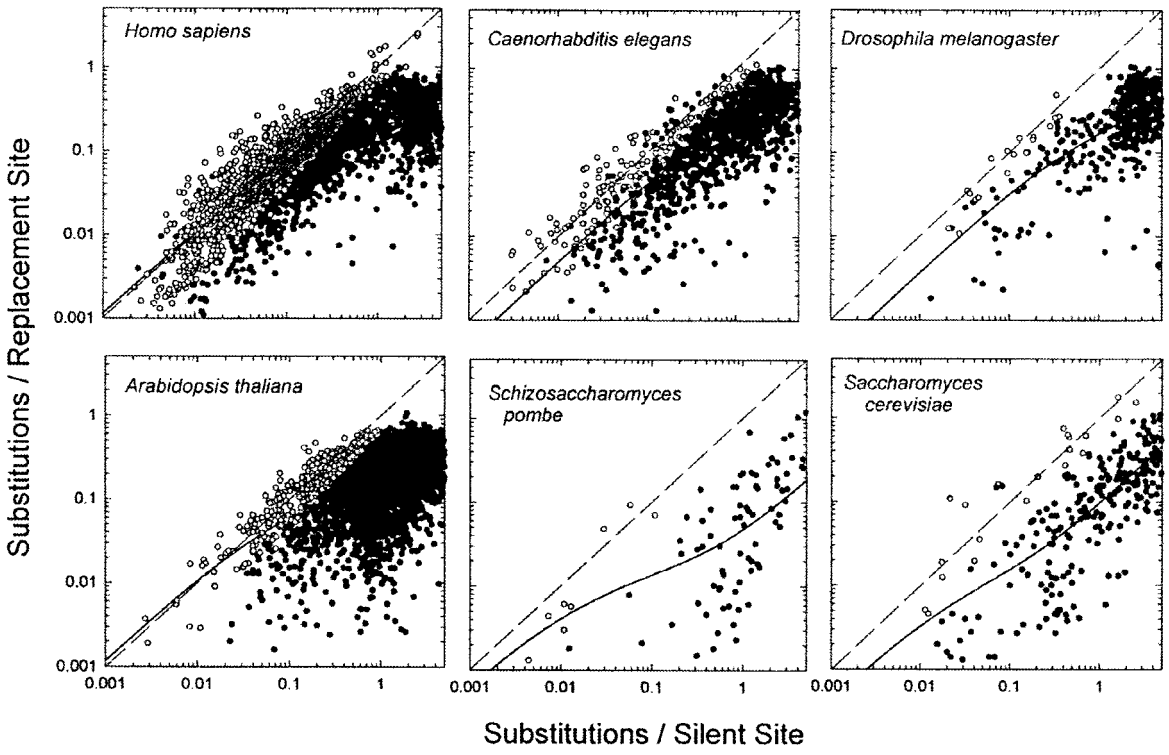

Figure 5. *R* vs. *S* plots for duplicate pairs of genes in six completely sequenced eukaryotic genomes. Open points denote pairs for which *R* is not significantly different from *S*.

duplicates). Other mathematical functions could be constructed to have qualitatively similar behavior, but Equation (7) is useful because it is readily integrated to yield a simple algebraic relationship between the cumulative number of substitutions per replacement site and the cumulative number of substitutions per silent site,

$$R = \frac{1}{am}\left[mS + \ln\left(\frac{a-b}{a-be^{-mS}}\right)\right]. \qquad (8)$$

On a log-log plot, points with equal *R/S* ratios fall on a diagonal line, with the height of the line being defined by the magnitude of *R/S* (Figure 4). Thus, with the preceding model, a linear relationship appears between log *S* and log *R* when *S* is small enough, as $dR/dS \simeq 1/(a-b)$ (phase 1). The response of log *R* to log *S* then becomes shallower as a transition is made to lower *dR/dS*, until a slope of one is again arrived at as *dR/dS* approaches $1/a$ (phase 2). The coefficient determines the rate of the transition between these two extreme phases. During a period in which genes are evolving in a neutral fashion, the response will be coincident with the diagonal describing *R/S=1.0* (the main diagonal in Figure 4).

To obtain the parameter estimates for Equation (8), we performed least-squares analyses, using logarithms of observed and expected values, so as not to give undue weight to sequence pairs with large *R*. Pairs of sequences for which *S* or *R* are equal to zero cannot be included in such an analysis, and we excluded the few data in which either *S* or *R* were <0.001, as such estimates are highly unreliable. To test for the constancy of *dR/dS*, we also obtained fits for the reduced model with a single parameter (i.e., $R/S = 1/a$, independent of *S*). Letting *n* denote the number of gene pairs, and r_f and r_r denote the correlation coefficients for the full and reduced models, the relevant test statistic is

$$F = \frac{n(r_f^2 - r_r^2)}{1 - r_f^2} \qquad (9)$$

with 1 and $n-2$ degrees of freedom (p. 633, Sokal and Rohlf 1995). (Although the models differ by two

Table 2. Estimated parameters relating replacement-site to silent-site divergence. Note that $(dR/dS)_{S\to 0} = 1/(a-b)$ and $(dR/dS)_{S\to \infty} = 1/a$. The one-parameter model was rejected for all species except *E. cunicluli*.

Species	m	$(dR/dS)_{S\to 0}$	$(dR/dS)_{S\to \}\infty}$	r^2	n
H. sapiens	0.31	0.898	0.029	0.671	3570
C. elegans	0.55	0.475	0.056	0.635	1253
D. melanogaster	0.54	0.360	0.050	0.524	445
A. thaliana	1.18	0.956	0.044	0.270	3790
S. cerevisiae	69.49	1.000	0.087	0.467	213
S. pombe	9.20	0.940	0.023	0.264	98
E. cuniculi	∞	0.296	0.296	0.919	27

parameters, m is effectively fixed at ∞ in the reduced model).

All of the eukaryotes examined except *E. cuniculi* exhibit a significant decline in *dR/dS* with increasing *S* (Table 2, Figure 5). The asymptotic values of *dR/dS* at low *S* are somewhat variable among species, with *D. melanogaster*, *C. elegans*, and *E. cuniculi* ranging between 0.3 and 0.5, the remaining species approaching 1.0, and an overall average of 0.70.

Excluding the estimate for *E. cuniculi*, the estimates of *dR/dS* at high *S* are much more homogeneous, ranging from 0.02 to 0.09, and averaging to 0.05. Thus, for most species, selection against amino-acid altering mutations is indeed relaxed early after duplication, in several cases approaching the neutral expectation. On average, the stringency of selection against replacement substitutions subsequently increases approximately 14-fold as a pair of duplicate genes ages, but the estimated values of m indicate that the transition between the two extreme phases occurs much more rapidly for the unicellular fungi than for metazoans and plants.

Discussion

These results largely corroborate the earlier conclusions in Lynch and Conery (2000). The estimated half-life of duplicate genes averaged over all species, approximately 4 MY, is identical in both studies, although it should be kept in mind that there is an order-of-magnitude range of variation among species. The average rate of origin of new duplicates in this study, 0.01/gene/MY, is about the same as that reported in our earlier study. Gu *et al.* (2002) also recently reported estimates of the rate of origin of duplicates genes in *D. melanogaster*, *C. elegans*, and *S. cerevisiae*. As in our study, these authors employed pre-screening devices to avoid pseudogenes, alternatively spliced variants, and annotation errors, but their additional methods of analysis deviated from those used herein in a number of ways. For example, Gu *et al.* (2002) did not make a distinction between the number of duplicate pairs and the actual (necessarily smaller) number of duplication events; they included very large multigene families, whereas we used an arbitrary cutoff for family size of five; they attempted to remove incomplete or chimeric duplicates; and they eliminated some duplicate pairs in which the nucleotide substitution levels in flanking regions exceeded those in silent sites. Despite these differences, it is comforting that estimates from both studies for B appear to deviate by no more than a factor of three.

Since decisions as to what constitutes a legitimate gene duplication will always involve a certain degree of subjectivity, and since the quality of current complete-genome databases is still in a state of flux, there is little question that the rates of birth and death of duplicate genes noted above will be subject to future modification. However, without a major upheaval, it appears difficult to avoid the conclusion that on-going gene turnover is a fundamental evolutionary feature of all eukaryotic genomes. Moreover, the smallest genomes may be kept small not by a low rate of origin of new duplicates but by a high rate of attrition. As a consequence of the duplication process, most species are expected to exhibit transient presence/absence polymorphisms at multiple loci (see Lynch 2003 for a review), and microchomosomal rearrangements among closely related species should commonly arise when the descendant members of duplicate pairs survive at the expense of their ancestral copy. Such reassignments of chromosomal locations provide a simple and powerful mechanism for the origin of post-zygotic isolating barriers that requires no intermediate stage of reduced fitness and

no accumulation of negative epistatic interactions between heterospecific genes (Lynch and Force 2000). Thus, although gene duplication is often regarded primarily as a mechanisms for adaptive evolution, its role in the other major engine of evolution, speciation, may be even more significant.

Finally, we note that although the majority of duplicate genes appear to be transient, contributing little to long-term phenotypic evolution, a minority of such genes become preserved for very long periods, either by subfunctionalization of neofunctionalization, as illustrated by the long, shallow shoulders on the age distributions displayed in Figure 1 and by the strong purifying selection operating on such genes (Figure 5). Although such genes may be simply selected for purely on the basis of redundancy, this appears to be unlikely on theoretical grounds (Lynch *et al.* 2001), so they almost certainly contribute to adaptive evolution. Nevertheless, unless the eukaryotic genome is undergoing a gradual and prolonged expansion, even members of these pairs must ultimately be subject to loss.

Acknowledgments

We are extremely grateful to the large number of individuals who have contributed to the genome sequencing projects from which this study draws its analyses. Our work has been supported by NIH grant GM20887 and NSF grant DEB-0003920.

References

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Bevan, M., K. Mayer, O. White, J. A. Eisen, D. Preuss, T. Bureau, S.L. Salzberg, H. W. Mewes. 2001. Sequence and analysis of the *Arabidopsis* genome. Curr. Opin. Plant Biol. 4: 105–110.

Conery, J. S., and M. Lynch. 2001. Nucleotide substitutions and the evolution of duplicate genes. Pacific Symp. Biocomput. 6: 167–178.

Cormen, T. H., C. E. Leiserson, and R. L. Rivest. 1990. Introduction to Algorithms. McGraw-Hill.

Grant, D., P. Cregan, and R. C. Shoemaker. 2000. Genome organization in dicots: genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. Proc. Natl. Sci. USA 97: 4168–4173.

Gu, Z., A. Cavalcanti, F.-C. Chen, P. Bouman, and W.-H. Li. 2002. Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. Mol. Biol. Evol. 19: 256–262.

Li, W.-H. 1999. Molecular Evolution. Sinauer Assocs., Sunderland, MA.

Lynch, M. 1997. Mutation accumulation in nuclear, organelle, and prokaryotic transfer RNA genes. Mol. Biol. Evol. 114: 914–925.

Lynch, M. 2003. Gene duplication and evolution. In A. Moya (ed.), Evolution: From Molecules to Ecosystems. Oxford University Press. (in press).

Lynch, M., and J. Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1154.

Lynch, M., and A. Force. 2000. The origin of interspecific genomic incompatibility via gene duplication. Amer. Natur. 156: 590–605.

Lynch, M., M. O'Hely, B. Walsh, and A. Force. 2001. The probability of fixation of a newly arisen gene duplicate. Genetics 159: 1789–1804.

Ohno, S. 1970. Evolution by Gene Duplication. Springer-Verlag, Berlin.

Shapira, S. K., and V. G. Finnerty. 1986. The use of genetic complementation in the study of eukaryotic macromolecular evolution: rate of spontaneous gene duplication at two loci of *Drosophila melanogaster*. Mol. Biol. Evol. 23: 159–167.

Shioura, A., A. Tamura, and T. Uno. 1997. An optimal algorithm for scanning all spanning trees of undirected graphs. SIAM J. Comput. 26: 678–692.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. 3rd Ed. Freeman, Yew York.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13: 555–556.