BMC Evolutionary Biology

Research article

Open Access

Duplicated genes evolve slower than singletons despite the initial rate increase

I King Jordan, Yuri I Wolf and Eugene V Koonin*

Address: National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Email: I King Jordan - jordan@ncbi.nlm.nih.gov; Yuri I Wolf - wolf@ncbi.nlm.nih.gov; Eugene V Koonin* - koonin@ncbi.nlm.nih.gov

* Corresponding author

Published: 06 July 2004

BMC Evolutionary Biology 2004, **4**:22 doi:10.1186/1471-2148-4-22

Received: 26 March 2004
Accepted: 06 July 2004

This article is available from: http://www.biomedcentral.com/1471-2148/4/22

Abstract

Background: Gene duplication is an important mechanism that can lead to the emergence of new functions during evolution. The impact of duplication on the mode of gene evolution has been the subject of several theoretical and empirical comparative-genomic studies. It has been shown that, shortly after the duplication, genes seem to experience a considerable relaxation of purifying selection.

Results: Here we demonstrate two opposite effects of gene duplication on evolutionary rates. Sequence comparisons between paralogs show that, in accord with previous observations, a substantial acceleration in the evolution of paralogs occurs after duplication, presumably due to relaxation of purifying selection. The effect of gene duplication on evolutionary rate was also assessed by sequence comparison between orthologs that have paralogs (duplicates) and those that do not (singletons). It is shown that, in eukaryotes, duplicates, on average, evolve significantly slower than singletons. Eukaryotic ortholog evolutionary rates for duplicates are also negatively correlated with the number of paralogs per gene and the strength of selection between paralogs. A tally of annotated gene functions shows that duplicates tend to be enriched for proteins with known functions, particularly those involved in signaling and related cellular processes; by contrast, singletons include an over-abundance of poorly characterized proteins.

Conclusions: These results suggest that whether or not a gene duplicate is retained by selection depends critically on the pre-existing functional utility of the protein encoded by the ancestral singleton. Duplicates of genes of a higher biological import, which are subject to strong functional constraints on the sequence, are retained relatively more often. Thus, the evolutionary trajectory of duplicated genes appears to be determined by two opposing trends, namely, the post-duplication rate acceleration and the generally slow evolutionary rate owing to the high level of functional constraints.

Background

The importance of gene duplication in the evolution of genetic novelty has long been recognized [1,2]. Because gene duplication often precedes the functional diversification between duplicates, it has been predicted that evolutionary rates should increase following duplication [3,4]. Indeed, studies on the evolutionary rates of duplicated genes showed that acceleration tends to occur immediately following duplication [5,6]. These rate accelerations may be due to either a relaxation of purifying selection on

one or both gene duplicates or to the action of positive diversifying selection between the duplicates (or some combination of both factors) [7,8]. However it is achieved, the evolutionary rate acceleration appears to be an important mechanism leading to functional diversification of duplicates [9,10]. The role of relaxed purifying selection in functional diversification has been embodied in the neofunctionalization and subfunctionalization concepts whereby duplicates accumulate mutations that either lead to the emergence of new functions or differentially inactivate subfunctions of the ancestral singleton, while the remaining subfunction is maintained or even enhanced [11-17]. Detailed studies of the effect of duplication on site-specific rates showed an increased proportion of changes in highly constrained sites, which seems to be particularly well compatible with subfunctionalization [18].

Post-duplication evolutionary rate acceleration has been revealed primarily through sequence comparisons between duplicated genes. More recently, the availability of complete genome sequences has allowed for an approach to the study of the effects of gene duplication on evolutionary rates that is qualitatively distinct from those earlier studies. The comparative-genomic approach to the study of gene duplication and evolution that is employed here relies on the distinction between genes that are related by orthology (divergence via speciation) and paralogy (divergence via duplication) [Fitch, 1970 #130;Fitch, 2000 #131;Sonnhammer, 2002 #128]. Genome-wide comparisons of proteins encoded in sequenced genomes allow for the identification of orthologs and paralogs [19,20]. Orthologous genes can then be classified into those that have paralogs (duplicates) and those that do not have any (singletons). Sequence comparisons between orthologs of these two classes can be used to assess the relationship between gene duplication and evolutionary rate [21-23]. For controlled between-species comparisons, this approach has the advantage of equalizing the time of divergence (at speciation) between the genes being compared, whereas the comparison of paralogs themselves is complicated by the fact that duplications that produced them occurred at different times. Using a combination of within and between-species sequence comparisons, we address the questions of how and to what extent gene duplication affects evolutionary rates. In particular, we address the possibility that, due to the relaxation of purifying selection after gene duplication [5,6], duplicated genes in general might evolve faster than singletons. We compare amino acid substitution levels (and nucleotide substitution levels for human-mouse) between orthologous gene pairs classified as duplicates or singletons from the following phylogenetically diverse set of species pairs: human-mouse, *Drosophila-Anopheles*, *Saccharomyces cerevisiae-Candida albicans*, *Escherichia coli-Yersinia pestis*, *Bacillus subtilis-B. halodurans* and *Pyrococcus horikoshii-P. furiosis*. We show that, in spite of the acceleration of evolution that is typically observed after duplication, duplicates tend to evolve more slowly than singletons.

Results and Discussion

Sequence substitution levels of orthologs and gene duplication

Orthologous protein sequence pairs were identified for human and mouse as described under Methods. Protein sequences were aligned, and the resulting amino acid sequence alignments were used to guide the alignment of nucleotide coding sequences (CDSs). These alignments were used to estimate sequence divergence levels (substitutions per site) for amino acids as well as for non-synonymous (dN) and synonymous (dS) CDS substitutions. For pairs of human-mouse orthologs, within-genome sequence comparisons were used to classify them as duplicates or singletons, based on whether or not they had detectable paralogs, and the average sequence divergence levels for these two classes of genes were compared. The classification of orthologous pairs as duplicates and singletons was done using three criteria: 1 – presence of paralogs in the human genome alone, 2 – presence of paralogs in the mouse genome alone, and 3 – presence of paralogs in human or mouse. For all three classification criteria, the average ortholog amino acid substitution levels for duplicates were substantially (and statistically highly significantly) lower than those of singletons (Table 1 and Figure 1a).

It is a formal, albeit unlikely, possibility that these differences in sequence diversity between the duplicate and singleton classes are due to different mutation pressures. To control for this possibility, the ratio of dN/dS was taken as an approximate measure of selective constraint and compared between the duplicate and singleton classes. As with the amino acid substitution levels, dN/dS is substantially lower for duplicates than for singletons (Table 1 and Figure 1b). The dN/dS ratio is considered to be an indicator of the mode and strength of the selection operating during evolution of a gene [9,10]. Thus, the finding that, in comparisons between human and mouse orthologs, duplicates, on average, have lower dN/dS values than singletons strongly suggests that the former are subject to stronger purifying selection than the latter.

A formal possibility exists that the observed differences between the evolutionary rates of duplicates and singletons were due solely to the presence of extremely rapidly evolving gene pairs of potentially mis-identified orthologs (Methods; Figure 5). To examine the possible contribution of this effect, all human-mouse gene pairs with dS > 2 standard deviations (sd) from the mean were removed

Table 1: Evolutionary distance (substitution level) comparisons between duplicates and singletons

Comparison[a]	% difference[b]	Duplicate – n[c]	Singleton – n[c]	*P*[d]
Human – Mouse (gamma)	47.11	10,753	2,517	4.2×10^{-50}
Human – Mouse (dN/dS)	37.79	10,753	2,517	5.1×10^{-61}
Drosophila – Anopheles (gamma)	26.37	4,902	2,233	4.8×10^{-50}
S. cerevisiae – C. albicans (gamma)	25.68	1,584	1,845	8.8×10^{-46}
E. coli – Y. pestis (gamma)	5.11	1,110	1,235	0.17
B. subtilis – B. halodurans (gamma)	4.41	1,021	1,081	0.10
P. horikoshii – P. furiosis (gamma)	13.38	574	872	0.02

[a] Species comparison used to calculate evolutionary distances. Gamma distances are the number of amino acid substitutions per site. dN/dS is the ratio of non-synonymous (dN) to synonymous (dS) nucleotide CDS substitutions. [b] Percent difference is calculated by taking the absolute value of difference between duplicate and singleton distances and dividing by the average distance for all genes [c] Numbers of orthologous gene pairs compared to calculate average distances for duplicate and singleton genes [d] *P*-value for the t-test comparing duplicate and singleton distance averages

from consideration, and the differences in average substitution levels between duplicates and singletons were re-calculated. This did not result in any appreciable differences from the original results (Figure 1) that were obtained using a cut-off of 3 sd (for the 2 sd cut-off, the amino acid gamma distance was 0.18 for duplicates and 0.29 for singletons, and the dN/dS ration was 0.14 for duplicates and 0.20 for singletons).

It is also formally possible that the differences in the rates of evolution between duplicates and singletons are due to the difficulty in the detection of paralogs of rapidly evolving genes, which would result in erroneous classification of such genes as singletons. To control for this potential bias, dS between duplicate human genes and their most closely related paralogs were determined. Duplicate genes were then re-classified as singletons if dS for a pair of human paralogs was greater than dS between each of the respective genes and its mouse ortholog. Under this procedure, only recent paralogs were classified as duplicates. The substitution rate differences between these duplicates and the resulting set of "pseudo-singletons" were re-calculated. This procedure did not result in qualitative change in the results; in fact, the magnitude of the difference between duplicate and singleton amino acid substitution rates slightly increased (from -0.11 in Figure 1a to -0.14). Thus, the differences in the evolution rates between duplicates and singletons did not seem to be due to a detection bias.

Orthologous protein sequence pairs were identified and aligned for two more pairs of eukaryotes and for three pairs of prokaryotes, all with complete genome sequences, and the pairwise sequence alignments were used to determine amino acid substitution (evolutionary) levels for duplicate versus singleton orthologs. For both additional eukaryotic species pairs (insects and yeasts), the average ortholog amino acid substitution levels for duplicates were substantially (and statistically highly significantly) lower than those for singletons (Table 1 and Figure 2). The same qualitative pattern was seen for the prokaryotic species comparisons, with the duplicate class showing consistently lower average amino acid substitution rates (Table 1 and Figure 2). However, the differences were far less pronounced than in the case of eukaryotes, and in only one case the average rate difference between duplicates and singletons was marginally statistically significant (Table 1).

A similar relationship between gene duplication and evolutionary diversity was observed when the amino acid substitution levels between orthologs were considered with respect to the number of detectable paralogs for a given ortholog. For all three eukaryotic comparisons, there are statistically significant negative correlations between the number of amino acid substitutions per site and the number of paralogs (Table 2); in other words, proteins with more paralogs tend to evolve more slowly between species than proteins with fewer paralogs. However, the magnitudes of these correlations are slight (Table 2). The effect was even less pronounced for the prokaryotic comparisons; while the correlations between the sequence diversity levels and the number of paralogs were all negative, the magnitudes of these correlations were quite small and none of them was statistically significant (Table 2).

More striking than the relationship between evolutionary sequence diversity and the number of paralogs was the correlation between the amino acid substitution levels between orthologs and those between the most closely related paralogs. For each orthologous pair with detectable paralogs, the amino acid distances between orthologs were plotted against the distances between one of the

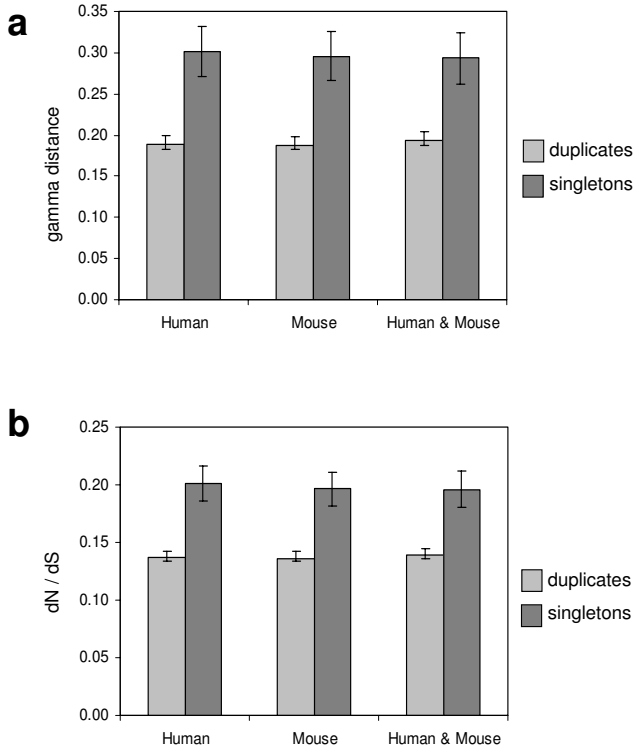

Figure 1
Average substitution levels, with 95% confidence intervals, for orthologous human-mouse sequence pairs with paralogs (duplicates – light gray bars) and with no paralogs (singletons – dark gray bars). The x-axis labels indicate comparisons where orthologous pairs were classified as duplicates or singletons based on three criteria: 1 – presence of a paralog in the human genome alone (Human), 2 – presence of a paralog in the mouse genome alone (Mouse) and 3 – presence of a paralog in the human or mouse genome (Human & Mouse). a – amino acid substitution levels calculated using the gamma correction for multiple substitutions. b – ratio of non-synonymous (dN) to synonymous (dS) nucleotide CDS substitutions.

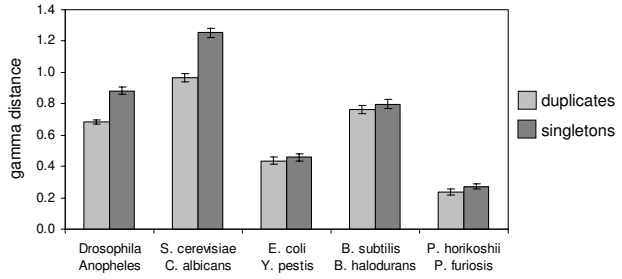

Figure 2
Average amino acid substitution levels, with 95% confidence intervals, for orthologous pairs with paralogs (duplicates – light gray bars) and with no paralogs (singletons – dark gray bars). Species comparisons are shown on the x-axis.

orthologs and its most closely related paralog. For all three eukaryotic comparisons, there was a highly significant positive correlation between the two sequence divergence levels (Table 3). Thus, orthologs that evolve relatively slowly between species tend to have more closely related paralogs within genomes, and orthologs that evolve more rapidly have less closely related paralogs. The r^2 values for these relationships were about an order of magnitude greater than those for the comparisons between sequence diversity and the number of paralogs (compare Tables 2 and 3). As in the previous cases, the relationship between ortholog and paralog amino acid substitution levels is not nearly as strong for the prokaryotes as it is for eukaryotes (Table 2). Nevertheless, the connection between gene duplication and sequence evolution of orthologs in prokaryotes is most evident in this comparison, with two out of the three correlations being statistically significant (Table 3).

Age of duplications and substitution levels

One advantage of the comparison of orthologs rather than paralogs is that the time of divergence is the same, namely, the time of speciation, for all analyzed orthologous pairs. Paralogous pairs, in contrast, will often have diverged via duplication at different times. In the case of orthologous proteins then, differences in substitution levels are primarily due to differences in the strength of purifying selection, whereas the apparent differences in substitution levels for paralogous protein pairs are additionally affected by differences in the time of duplication. This distinction is relevant to the comparison of evolutionary rates between orthologs versus evolutionary rates between closest paralogs [22]. As described above, there is a strong positive correlation between these rates. This correlation could indicate that proteins that are strongly conserved between species are also strongly conserved within genomes, or it could mean that proteins that are strongly conserved between species tend to have more recent duplicates in the genome. In an attempt to distinguish between these two explanations for the positive correlation between ortholog and paralog sequence divergence, gene duplications were partitioned into approximate isotemporal classes. For example, all-against-all sequence comparisons were performed for human, mouse and *Fugu rubripes* (Fugu), and the results

Table 2: Correlation between ortholog substitution levels and the number of paralogs

Species[a]	Slope[b]	r^{2c}	*P*[d]
Human – Mouse	-1.0×10^{-4}	3.0×10^{-3}	7.2×10^{-13}
Drosophila – Anopheles	-8.0×10^{-4}	5.2×10^{-3}	9.6×10^{-10}
S. cerevisiae – C. albicans	-4.7×10^{-3}	9.8×10^{-3}	6.0×10^{-9}
E. coli – Y. pestis	-6.0×10^{-4}	3.0×10^{-4}	4.2×10^{-1}
B. subtilis – B. halodurans	-7.0×10^{-4}	2.0×10^{-4}	4.9×10^{-1}
P. horikoshii – P. furiosis	-2.5×10^{-3}	1.5×10^{-3}	1.4×10^{-1}

[a] Species for which comparisons were performed [b] Slope of the linear regression line (i.e. y from the equation y = mx + b) [c] Square of the correlation coefficient (corresponds to the fraction of variation in ortholog evolutionary rates explained by the variation in the number of paralogs) [d] *P*-value for correlation coefficient

Table 3: Correlation between ortholog substitution levels and the substitution levels between the most closely related paralogs

Species[a]	Slope[b]	r^{2c}	*P*[d]
Human – Mouse	6.1×10^{-2}	3.3×10^{-2}	3.3×10^{-87}
Drosophila – Anopheles	6.5×10^{-2}	2.3×10^{-2}	7.6×10^{-27}
S. cerevisiae – C. albicans	2.1×10^{-1}	1.0×10^{-1}	1.1×10^{-38}
E. coli – Y. pestis	3.3×10^{-2}	3.0×10^{-3}	9.0×10^{-2}
B. subtilis – B. halodurans	1.6×10^{-1}	4.2×10^{-2}	3.1×10^{-11}
P. horikoshii – P. furiosis	6.3×10^{-2}	1.7×10^{-2}	1.0×10^{-3}

[a] Species for which comparisons were performed [b] Slope of the linear regression line (i.e. y from the equation y = mx + b) [c] Square of the correlation coefficient (corresponds to the fraction of variation in ortholog evolutionary rates explained by the variation in paralog evolutionary rate) [d] *P*-value for correlation coefficient

were used to partition duplications along three evolutionary classes (Figure 3a): 1 – duplications that occurred along the human lineage (after the human – mouse divergence), 2 – duplications that occurred along the mammalian lineage (after the divergence between Fugu and the human – mouse lineage), and 3 – relatively ancient duplications that occurred along the lineage that leads to all three species (before their divergence). The same procedure was also used to partition duplications along three yeast evolutionary lineages (Figure 3b).

Once this partitioning was complete, the ortholog versus closest paralog amino acid substitution levels were analyzed independently for each of the three classes of duplicates. This procedure has the effect of normalizing (to a degree) the time of duplication so that only proteins encoded by genes that duplicated along the same evolutionary lineage are compared. When this was done, the correlations between orthologs and paralog amino acid substitution levels within each isotemporal class of duplicates became even stronger than those seen for the pooled data (Table 4). This result strongly suggests that the same functional constraints govern a gene's evolution after speciation and after duplication.

Acceleration versus deceleration of gene duplicate's evolution

Both theoretical and empirical studies have previously pointed to an acceleration of sequence substitution following gene duplication [5,6,11,12,17].

This is thought to be due to either a relaxation of purifying selection or the action of positive, diversifying selection (or perhaps both). For instance, when pairs of paralogs were compared to pairs of orthologs that have similar levels of protein divergence, it was shown that the paralogs had higher dN/dS values [6]. This was taken as evidence for a relaxation of selection immediately after gene duplication. Consistent with this notion, two recent studies have shown that members of duplicate pairs often evolve at significantly different rates after duplication and that the more rapidly evolving duplicates have elevated dN/dS [7,8]. In light of these observations, it seems surprising that we found strong evidence here that orthologs with duplicates evolve more slowly than singletons. Indeed,

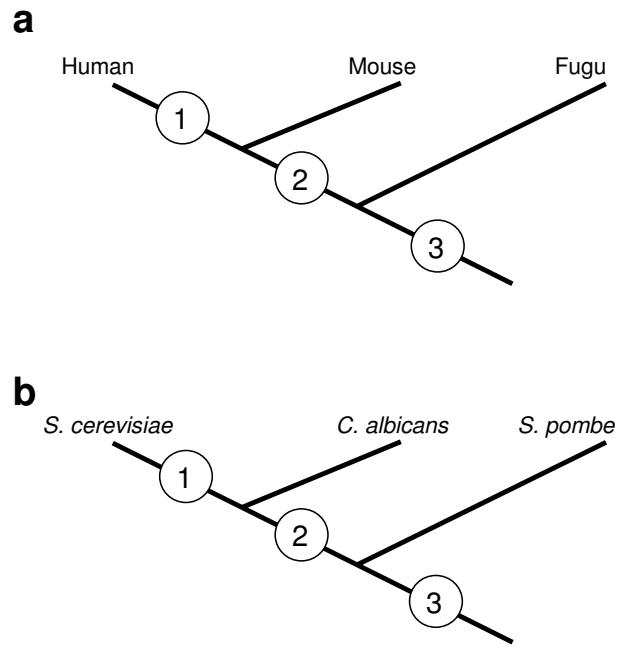

Figure 3
Mapping of lineage-specific expansions to individual branches of phylogenetic trees. Shown for vertebrates (a) and yeasts (b).

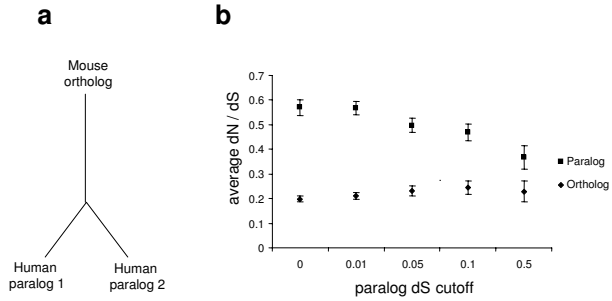

Figure 4
Post-duplication relaxation of purifying selection in paralogs. a – Schematic illustrating the rationale for the comparison of dN/dS for human-mouse orthologs versus human paralogs. dN/dS levels were averaged for sets of proteins, related as shown, where the human paralogs duplicated after the human-mouse divergence. b – Average dN/dS leves, with 95% confidence intervals (y-axis), is plotted for human-mouse orthologs (diamonds) and human paralogs (squares). A series of increasing cut-offs based on the level of dS (x-axis) between human paralogs was employed so that each set is restricted to more and more distantly related paralogs.

the evolutionary history of those orthologs that have duplicates would seem to include a period of accelerated evolution after gene duplication, whereas orthologs without duplicates are unlikely to have experienced such an acceleration. Thus, everything else being equal, duplicates would be expected to evolve faster than singletons.

To investigate this apparent contradiction, we identified triplets of genes which included a single mouse gene and a pair of human paralogs that evolved via a duplication subsequent to the human-mouse divergence (Figure 4a). The dN and dS values for the human paralogs in such gene sets were compared to the average dN and dS values for the mouse gene and each of its human co-orthologs. Comparisons between the human paralogs showed significantly higher average dN/dS ratios (t-test, $P < 5 \times 10^{-5}$) than the human-mouse ortholog comparisons (Figure 4b). This pattern holds across a series of increasingly stringent cut-offs based on the level of dS between paralogs (Figure 4b). The same pattern was also seen in a reciprocal comparison, when levels of dN and dS for human and mouse orthologs were compared to levels of dN and dS for mouse paralogs (data not shown). For the most closely related paralogs, the dN/dS ratio of paralogs was ~3-fold greater than the dN/dS ratio for the same paralogs and their single ortholog in another species (Figure 4b), which is remarkably close to the value determined previously with a different approach [6]. The magnitude of the difference declined for more distant paralogs (Figure 4b), in accord with the notion that the acceleration of evolution occurs immediately after duplication [5].

These observations suggest, consistent with previous findings, that paralogs do indeed experience a post-duplication period of accelerated evolution, which is apparently due to the relaxation of purifying selection. These results stand in stark contrast to the finding that orthologs with duplicates are more evolutionarily conserved than orthologs with no duplicates. It seems that there are two countervailing forces at work on the sequence evolution of duplicate genes: i) acceleration of substitution between paralogs caused by relaxation of purifying selection after duplication, and ii) relative reduction of substitution rate for genes with duplicates compared to singletons, which is predicated upon the stronger functional constraints affecting the former. The post-duplication acceleration has the effect of mitigating the sequence divergence differences between duplicates and singletons. This makes the differences in substitution levels that are observed between these two classes of orthologs even more notable.

Functional distribution of duplicated genes

Taken together, the measurements of sequence diversity reported here as well as previous observations and

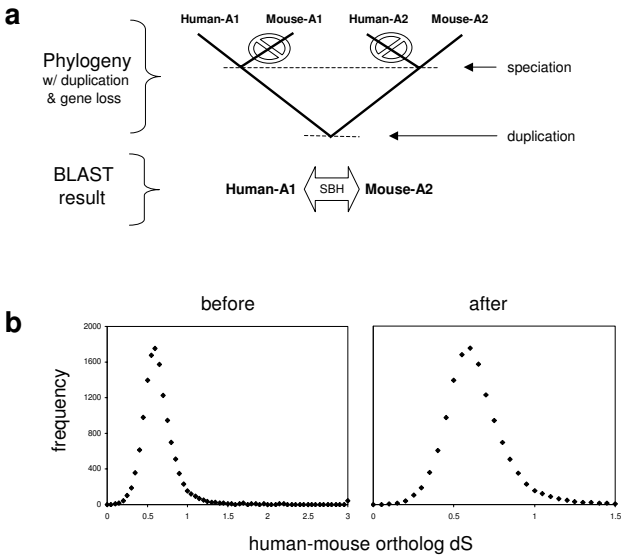

Figure 5
Ortholog identification control. a – The symmetrical best BLAST hits approach may mis-identify orthologs in rare cases where there is an ancient gene duplication followed by differential loss of paralogs. b – The dS distributions before and after removal of human-mouse orthologous pairs with dS > 3 standard deviations from the mean (see Methods).

theoretical arguments [5,6,11,12,17] suggest that the fate of duplicated genes depends greatly on their functional utility. Selection probably does continue to operate on the products of gene duplication but only in cases when the duplicates contribute substantially to organismic fitness. In order to further assess the validity of this notion, functional distributions of orthologs with and without duplicates were examined. The database of eukaryotic orthologous groups of proteins [24,25] was used to classify eukaryotic proteins into four broad functional categories: 1 – information storage and processing, 2 – signaling and other cellular processes (such as protein folding, degradation and trafficking), 3 – metabolism and 4 – poorly characterized. These distributions were then compared for the two classes of orthologs, those that possess duplicates and those that do not (Table 5). The distributions of the observed numbers of proteins in each category were compared using a χ^2 test where the expected numbers were calculated based on the functional distribution for all proteins. For all three eukaryotic comparisons, the functional distributions of the proteins in the two classes – duplicates versus singletons – were shown to be significantly different (Table 5). In almost all cases, the difference was most pronounced for the poorly characterized functional category; there are far fewer poorly characterized proteins among duplicates than expected. By contrast, the set of singletons is enriched in poorly characterized proteins. Thus, duplicates that are retained and conserved during evolution are enriched for proteins with known functions, particularly proteins that function in signaling and other cellular processes. Apparently at odds with this general pattern is the fact that duplicates have fewer information storage and processing proteins than expected, while singletons have more than expected. This could be due to the fact that many proteins involved in translation, transcription and replication function as multi-subunit complexes (e.g., the ribosome and RNA polymerase holoenzyme) such that duplication of the genes for individual subunits could lead to a dominant negative effect that would be selected against [26].

General discussion and conclusions

After the submission of the current work, we became aware of a very recent, independent study that reached the same major conclusion as we do here, namely that duplicate genes are, on average, more evolutionarily conserved than singletons in eukaryotes [27]. The analytical approach employed by Davis and Petrov was conceptually similar to ours in that it involved the comparison of ortholog substitution levels for genes designated as duplicates or singletons. However, an important difference between the two studies is that the approach of Davis and Petrov involved the characterization of genes as duplicates or singletons in one pair of species, *Caenorhabditis elegans* and *S. cerevisiae*, and the estimation of substitution levels in another pair of species, *D. melanogaster* and *A. gambiae*. This allowed for an estimate of substitution levels independent of the effects of gene duplication, whereas the substitution levels analyzed here were affected by duplication. We believe that it was important, as it is done here, to demonstrate on the same dataset that evolution of duplicated genes is shaped by the interplay of two opposing effects, the initial increase in substitution rate after gene duplication, and generally lower evolutionary rate of duplicates compared to singletons.

The results reported here point to two opposing trends in the evolution of duplicate genes. For the analyzed eukaryotic species, there is a clear relationship between gene duplication and the sequence divergence of orthologs: duplicates tend to evolve more slowly, on average, than singletons. Two recent studies reported conflicting observations on the relative rates of evolution of duplicates and singletons. Yang, Gu, and Li performed a comparison of *S. cerevisiae-C. albicans* orthologs with and without duplicates and found that the former, on average, evolved slower than the latter, in a qualitative agreement with the results described here [23]. In contrast, Nembaware and coworkers analyzed the evolutionary rates of human paralogs with varying levels of divergence and found that, in

Table 4: Correlation between ortholog substitution levels and the substitution levels between the most closely related paralogs for lineage specific expansions (Figure 3)

Lineage[a]	Slope[b]	r^2[c]	*P*[d]
Human specific (1-2a)	2.0×10^{-1}	7.8×10^{-2}	1.6×10^{-34}
Mammalian specific (2-2a)	2.1×10^{-1}	2.2×10^{-1}	2.7×10^{-114}
Vertebrate and before (3-2a)	8.2×10^{-2}	7.3×10^{-2}	9.9×10^{-132}
S. cerevisiae specific (1-2b)	8.3×10^{-1}	4.6×10^{-1}	1.7×10^{-58}
S. cerevisiae – C. albicans (2-2b)	4.6×10^{-1}	4.2×10^{-1}	3.4×10^{-18}
Yeast and before (3-2b)	3.3×10^{-1}	2.1×10^{-1}	2.8×10^{-54}

[a] Lineage specific expansions for which comparisons were performed (numbers in parentheses correspond to Figure 3a and 3b) [b] Slope of the linear regression line (i.e. y from the equation y = mx + b) [c] Square of the correlation coefficient (corresponds to the fraction of variation in ortholog evolutionary rates explained by the variation in paralog evolutionary rate) [d] *P*-value for correlation coefficient

Table 5: χ^2 test[a] of the functional distribution[b] of eukaryotic orthologs (duplicates versus singletons)

Human – Mouse	Dup Obs-Exp	Dup $(Obs-Exp)^2/Exp$	Sing Obs-Exp	Sing $(Obs-Exp)^2/Exp$
Information storage and processing	-49.5	1.46	49.5	9.76
Cellular processes and signaling	241.1	14.39	-241.1	96.04
Metabolism	-11.9	0.09	11.9	0.58
Poorly characterized	-179.6	13.04	179.6	87.07
χ^2 *P*	222.43 6.0×10^{-48}			
Drosophila – Anopheles	Dup Obs-Exp	Dup $(Obs-Exp)^2/Exp$	Sing Obs-Exp	Sing $(Obs-Exp)^2/Exp$
Information storage and processing	-93	9.75	93	26.48
Cellular processes and signaling	124.8	9.29	-124.8	25.25
Metabolism	89.1	8.78	-89.1	23.86
Poorly characterized	-120.9	12.69	120.9	34.46
χ^2 *P*	150.56 2.0×10^{-32}			
S. cerevisiae – C. albicans	Dup Obs-Exp	Dup $(Obs-Exp)^2/Exp$	Sing Obs-Exp	Sing $(Obs-Exp)^2/Exp$
Information storage and processing	-38.6	3.50	38.6	3.63
Cellular processes and signaling	55.4	5.68	-55.4	5.89
Metabolism	36.1	3.39	-36.1	3.52
Poorly characterized	-52.9	9.19	52.9	9.53
χ^2 *P*	44.33 1.3×10^{-9}			

[a] Test compares the observed (obs) numbers of proteins in each functional category for duplicates (dup) versus singletons (sing) with the expected (exp) numbers that are calculated based on the relative frequencies of all proteins in each functional category [b] Functional classifications of proteins were taken from the Clusters of Orthologous Groups database

human vs. mouse comparison, a particular class of paralogs with intermediate divergence evolved significantly faster than singletons [22]. It remains unclear what caused this difference in conclusions. However, the statistical significance and robustness of the lower level of substitutions in duplicates compared to singletons, which was observed for all compared genome pairs (albeit to a much lower extent in prokaryotes than in eukaryotes) in the present study, strongly suggests that duplicates indeed tend to evolve slower than singletons.

The finding that, on average, duplicates are more evolutionarily conserved than singletons is probably explained by the fact that the duplicates that are retained by selection are of greater functional utility than those that are lost after gene duplication. Thus, the selective pressure acting on the sequences of duplicated genes is, on average, greater than that affecting the sequences of singletons. The difference in the functional distributions between duplicated and non-duplicated genes is consistent with this notion. Apparently, genes that encode proteins with domains that are already widely employed in various cellular processes are more likely to contribute to the functional diversification of an organism via gene duplication than are genes encoding proteins with more limited functional utility.

However, in accord with the previous findings [5,6], we also demonstrate here a substantial acceleration of sequence substitution immediately after gene duplication.

Thus, the observation that, when orthologs are considered, duplicates tend to evolve more slowly than singletons is somewhat paradoxical. If one or more members of a set of paralogous genes experience a period of accelerated evolution, one might expect that, everything else being equal, this would have the effect of elevating the substitution levels between those genes and their orthologs above those characteristic of singletons. However, the results described here indicate that genes with duplicates are "more equal" than singletons in that the former, on average, are subject to more stringent purifying selection than the latter, presumably due to the relatively greater functional utility manifest in the increased likelihood of duplication fixation.

The relationship between duplication and ortholog sequence evolution also seems to be at odds with the fact that a considerable number of essential proteins, e.g., components of the core machineries of translation and transcription, do not have any paralogs but nevertheless evolve slowly. In contrast, some large multigene families, such as the immunoglobulins, encode proteins that evolve rapidly [28,29]. It appears that these two classes of proteins are exceptional: the former are subunits of stoichiometric complexes whose duplications is discourage by selection due to the deleterious effects of imbalance [26], whereas the former are adaptive linear specific expansions of paralogous families evolving under positive selection [30]. These well known exceptions to the general pattern reported here seem to render the relationship between gene duplication and ortholog substitution levels, which are averages based on comparisons of thousands of proteins, even more striking

Conclusions

It is notable that, in the prokaryotic species analyzed here, the relationship between duplication and ortholog sequence diversity is much less compelling than it is in eukaryotes. Overall, there does seem to be a similar pattern to that observed for eukaryotes, but it is far less striking and in many cases the comparisons made yield non-significant results. The causes of this difference remain uncertain. Conceivably, it could reflect the greater role that gene duplications appear to have in functional diversification of eukaryotes compared to prokaryotes and/or the preferential duplication of fast-evolving genes involved in adaptive reactions in prokaryotes [21].

Methods

Sequences

Complete sets of encoded proteins from whole genome sequences were compared for six pairs of species (three eukaryotic pairs and three prokaryotic pairs): *Homo sapiens* (human) – *Mus musculus* (mouse), *Drosophila melanogaster* – *Anopheles gambiae*, *Saccharomyces cerevisiae* – *Candida albicans*, *Escherichia coli* – *Yersinia pestis*, *Bacillus subtilis* – *Bacillus halodurans* &*Pyrococcus horikoshii* – *Pyrococcus furiosis*. All protein sequences are publicly available and were taken from ftp sites on the world wide web (Table 6). Coding nucleotide sequences that correspond to the human and mouse protein sequences were taken from the Genbank database [31,32] – using a series of Perl scripts developed specifically for the task (these are available upon request) together with programs from the SEALS software package [33].

Sequence analysis

For each pair of species considered, pairs of orthologous proteins were identified as symmetrical best hits [19,34] in all-against-all BLASTP searches [34]; between the encoded proteins of each species in the pair. All BLASTP searches were run with an expectation value (e-value) cut-off of 10^{-5}. BLASTP searches and post-processing of the

Table 6: Web sources for the protein sequences used in this study

Organism	Source	URL
H. sapiens	NCBI	ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/protein/
M. musculus	NCBI	ftp://ftp.ncbi.nlm.nih.gov/genomes/M_musculus/protein/
D. melanogaster	Ensembl	ftp://ftp.ensembl.org/pub/current_fly/data/fasta/pep/
A. gambiae	Ensembl	ftp://ftp.ensembl.org/pub/current_mosquito/data/fasta/pep/
S. cerevisiae	Saccharomyces Genome Database	ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/orf_protein/
C. albicans	Standford Genome Technology Center	ftp://cycle.stanford.edu/pub/projects/candida/
E. coli	NCBI	ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K12/
Y. pestis	NCBI	ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Yersinia_pestis_KIM/
B. subtilis	NCBI	ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Bacillus_subtilis/
B. halodurans	NCBI	ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Bacillus_halodurans/
P. horikoshii	NCBI	ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Pyrococcus_horikoshii/
P. furiosus	NCBI	ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Pyrococcus_furiosus/

results were performed using programs from the SEALS package. The symmetrical best hit procedure may misidentify orthologs in some rare cases where there is an ancient duplication (i.e. prior to the diversification of the pair of species being considered) followed by the differential loss of one of the resulting paralogs in each lineage (Figure 5a). When this occurs, the symmetrical best hits will actually represent paralogs that have artificially high levels of sequence diversity. To control for this formal possibility in the human-mouse ortholog identification, a cut-off based on the distribution of dS for human-mouse was employed (Figure 5b). Pairs of orthologs with dS values more than three standard deviations greater than the mean dS value were not considered. Implementation of this control does not result in any qualitative difference in the results obtained.

Pairs of orthologous proteins were aligned using the ClustalW program [35] and their substitution (evolutionary) levels were calculated using the gamma distance correction [36] with $\alpha = 2$. Paralogous sequences were identified by using each protein as a query in BLASTP searches, with the same settings used for ortholog identification, against the rest of the proteins encoded by its same genome. Coding nucleotide sequences were aligned to correspond (i.e. the gaps were inserted in-frame) to the amino acid sequence alignments of the proteins that they encode. The synonymous (dS) and non-syonymous substitution (dN) rates were calculated from these alignments using the Nei-Gojobori method [37] implemented in the PAML software package [38]. The lineage-specific expansions were defined as described previously using an approach that combines all-against-all BLASTP comparisons and within species clustering of the results [30]. In this report, the mammalian lineage specific expansions were defined using human versus *Fugu rupripes* (Fugu) comparisons and the human lineage specific expansions were defined using human versus mouse plus Fugu comparisons (Figure 3a). *S. cerevisiae – C. albicans* lineage specific expansions were defined using *S. cerevisiae* versus *Schizosaccharomyces pombe* comparisons and the *S. cerevisiae* lineage specific expansions were defined using *S. cerevisiae* versus *C. albicans* plus *S. pombe* comparisons (Figure 3b). The database of Clusters of Orthologous Eukaryotic proteins (KOGs) [39] was used for analysis of the functional distribution of eukaryotic proteins [25].

References

1. Haldane JBS: **The part played by recurrent mutation in evolution.** *Am Nat* 1933, **67:**5-19.
2. Fisher RA: **The possible modification of the response of the wild type to recurrent mutations.** *Am Nat* 1928, **62:**115-126.
3. Pauling L, Zuckerkandl E: **Chemical paleogenetics: molecular restoration studies of extinct forms of life.** *Acta Chem Scand* 1963, **17:**S9-S16.
4. Ohno S: **Evolution by gene duplication.** Berlin-Heidelberg-New York, Springer-Verlag; 1970.
5. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes.** *Science* 2000, **290:**1151-1155.
6. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV: **Selection in the evolution of gene duplications.** *Genome Biol* 2002, **3:**RESEARCH0008.
7. Conant GC, Wagner A: **Asymmetric sequence divergence of duplicate genes.** *Genome Res* 2003, **13:**2052-2058.
8. Zhang P, Gu Z, Li WH: **Different evolutionary patterns between young duplicate genes in the human genome.** *Genome Biol* 2003, **4:**R56.
9. Li WH: **Molecular Evolution.** Sunderland, MA, Sinauer; 1997.
10. Hughes AL: **Adaptive Evolution of Genes and Genomes.** New York - Oxford, Oxford University Press; 1999.
11. Lynch M, Force A: **The probability of duplicate gene preservation by subfunctionalization.** *Genetics* 2000, **154:**459-473.
12. Lynch M, O'Hely M, Walsh B, Force A: **The probability of preservation of a newly arisen gene duplicate.** *Genetics* 2001, **159:**1789-1804.
13. Fares MA, Wolfe KH: **Positive selection and subfunctionalization of duplicated CCT chaperonin subunits.** *Mol Biol Evol* 2003, **20:**1588-1597.
14. Braun FN, Liberles DA: **Retention of enzyme gene duplicates by subfunctionalization.** *Int J Biol Macromol* 2003, **33:**19-22.
15. Yu WP, Brenner S, Venkatesh B: **Duplication, degeneration and subfunctionalization of the nested synapsin-Timp genes in Fugu.** *Trends Genet* 2003, **19:**180-183.
16. Altschmied J, Delfgaauw J, Wilde B, Duschl J, Bouneau L, Volff JN, Schartl M: **Subfunctionalization of duplicate mitf genes associated with differential degeneration of alternative exons in fish.** *Genetics* 2002, **161:**259-267.
17. Massingham T, Davies LJ, Lio P: **Analysing gene function after duplication.** *Bioessays* 2001, **23:**873-876.
18. Seoighe C, Johnston CR, Shields DC: **Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation.** *Mol Biol Evol* 2003, **20:**484-490.
19. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278:**631-637.
20. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314:**1041-1052.
21. Jordan IK, Rogozin IB, Wolf YI, Koonin EV: **Microevolutionary genomics of bacteria.** *Theor Popul Biol* 2002, **61:**435-447.
22. Nembaware V, Crum K, Kelso J, Seoighe C: **Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs.** *Genome Res* 2002, **12:**1370-1376.
23. Yang J, Gu Z, Li WH: **Rate of protein evolution versus fitness effect of gene deletion.** *Mol Biol Evol* 2003, **20:**772-774.
24. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
25. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.** *Genome Biol* 2004, **5:**R7.
26. Papp B, Pal C, Hurst LD: **Dosage sensitivity and the evolution of gene families in yeast.** *Nature* 2003, **424:**194-197.
27. Davis JC, Petrov DA: **Preferential duplication of conserved proteins in eukaryotic genomes.** *PLoS Biol* 2004, **2:**E55.
28. Tanaka T, Nei M: **Positive darwinian selection observed at the variable-region genes of immunoglobulins.** *Mol Biol Evol* 1989, **6:**447-459.
29. Hughes AL: **Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells.** *Mol Biol Evol* 1997, **14:**1-5.
30. Lespinet O, Wolf YI, Koonin EV, Aravind L: **The role of lineage-specific gene family expansion in the evolution of eukaryotes.** *Genome Res* 2002, **12:**1048-1059.
31. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2003, **31:**23-27.
32. GenBank: **.** [http://www.ncbi.nlm.nih.gov].
33. Walker DR, Koonin EV: **SEALS: a system for easy analysis of lots of sequences.** *Ismb* 1997, **5:**333-339.

34. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
35. Higgins DG, Thompson JD, Gibson TJ: **Using CLUSTAL for multiple sequence alignments.** *Methods Enzymol* 1996, **266:**383-402.
36. Ota T, Nei M: **Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site.** *Mol Biol Evol* 1994, **11:**613-619.
37. Nei M, Gojobori T: **Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.** *Mol Biol Evol* 1986, **3:**418-426.
38. Yang Z: **PAML: a program package for phylogenetic analysis by maximum likelihood.** *Comput Appl Biosci* 1997, **13:**555-556.
39. KOGs: **.** [http://www.ncbi.nlm.nih.gov/COG].