Extent of Gene Duplication in the Genomes of *Drosophila*, Nematode, and Yeast

Zhenglong Gu, Andre Cavalcanti,* Feng-Chi Chen,* Peter Bouman,† and Wen-Hsiung Li**

*Department of Ecology and Evolution and †Department of Statistics, University of Chicago

We conducted a detailed analysis of duplicate genes in three complete genomes: yeast, *Drosophila*, and *Caenorhabditis elegans*. For two proteins belonging to the same family we used the criteria: (1) their similarity is $\geq I$ ($I = 30\%$ if $L \geq 150$ a.a. and $I = 0.01n + 4.8L^{-0.32(1 + \exp(-L/1000))}$ if $L < 150$ a.a., where $n = 6$ and L is the length of the alignable region), and (2) the length of the alignable region between the two sequences is $\geq$80% of the longer protein. We found it very important to delete isoforms (caused by alternative splicing), same genes with different names, and proteins derived from repetitive elements. We estimated that there were 530, 674, and 1,219 protein families in yeast, *Drosophila*, and *C. elegans*, respectively, so, as expected, yeast has the smallest number of duplicate genes. However, for the duplicate pairs with the number of substitutions per synonymous site (K_S) $<$ 0.01, *Drosophila* has only seven pairs, whereas yeast has 58 pairs and nematode has 153 pairs. After considering the possible effects of codon usage bias and gene conversion, these numbers became 6, 55, and 147, respectively. Thus, *Drosophila* appears to have much fewer young duplicate genes than do yeast and nematode. The larger numbers of duplicate pairs with $K_S < 0.01$ in yeast and *C. elegans* were probably largely caused by block duplications. At any rate, it is clear that the genome of *Drosophila melanogaster* has undergone few gene duplications in the recent past and has much fewer gene families than *C. elegans*.

Introduction

It has been proposed that gene duplication is the most important step for the origin of genetic novelties (Ohno 1970, p. 72). With the availability of complete genome sequences, it has become possible to study the extent of gene duplication on a genome-wide scale. Block duplications in *Drosophila*, yeast, and *Caenorhabditis elegans* have been studied in detail by using genomic data (Wolfe and Shields 1997; Seoighe and Wolfe 1999; Friedman and Hughes 2001). Using the BLASTP E value as the sole criterion for identifying homologous proteins, Rubin et al. (2000) studied the extents of gene duplication in yeast, *Drosophila*, and *C. elegans* genomes. However, deciding whether two proteins are homologous requires a more rigorous analysis. For example, domain shuffling or sharing is known to be a common mode for protein evolution (Doolittle 1995) and can mislead the identification of duplicate genes because a low E value between nonhomologous genes can be caused by a shared domain alone. Another difficulty in identifying homologous genes is the detection of remote homology. Deciding whether two proteins are homologous becomes difficult when their sequence identity is within the twilight zone (Doolittle 1986). Improvement in methodology often leads to the discovery of new homologous relationships and new gene family members (Krogh et al. 1994; Sonnhammer, Eddy, and Durbin 1997).

The rate of gene duplication in a genome is also of great interest. This type of study is possible only when the whole genome data is available. Lynch and Conery (2000) estimated the gene duplication rates in the yeast, *Drosophila*, and *C. elegans* genomes using the synonymous site changes (K_S) as the time scale. However, it is well known that codon usage is highly biased in some genes in these organisms (Ikemura 1982; Akashi, Kliman, and Eyre-Walker 1998). A negative correlation between synonymous rate (K_S) and strength of codon usage bias in *Drosophila* suggests that in some genes synonymous changes are not neutral (Sharp and Li 1989; Moriyama and Hartl 1993), though Dunn, Bielawski, and Yang (2001) argued against the existence of this correlation. Therefore, the K_S value might not reflect the real age of a gene duplication. A combination of K_S and the genetic distances in intron and flanking regions might be more informative.

The relatively good quality of genomic sequences and concomitant annotation for yeast, *Drosophila*, and *C. elegans* make it possible for us to investigate the above questions in these genomes. However, the presence of same genes with different names and the existence of alternative splicing forms in the database make it difficult to study the extent of gene duplication in a genome. Moreover, retrotranscriptase (RT) and protein parts derived from repetitive elements (REs) might mislead the identification of homologous proteins. For these reasons, it is important to clean the database. In this paper, after carrying out a detailed cleaning procedure for the protein databases of these three genomes, we asked two questions: How many gene families are there in each genome? How often has gene duplication occurred in the recent past in each organism? We defined two simple homology criteria by improving the criterion adopted by Rost (1999). Using the new criteria for identifying homologous genes and the single-linkage algorithm for clustering, we estimated the number of gene families in each of the three genomes. The frequency of recent gene duplication was investigated by using gene pairs with a small K_S. We excluded the gene pairs with possible gene conversion and codon usage bias by comparing K_S with the genetic distance in intron and flanking regions. The effect of codon usage on K_S in yeast was further studied.

Key words: gene families, gene duplication rate, database cleaning, codon usage bias.

Address for correspondence and reprints: Wen-Hsiung Li, Department of Ecology and Evolution, University of Chicago, 1101 East 57th street, Chicago, Illinois 60637. E-mail: whli@uchicago.edu.

Mol. Biol. Evol. 19(3):256–262. 2002
© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Materials and Methods

The protein data sets were obtained from the following websites:

Caenorhabditis elegans: http://www.sanger.ac.uk/Projects/C_elegans/wormpep/ Wormpep release 40 was used. There were 19,730 protein sequences in the database, of which 48 did not have genomic position information and 22 did not have corresponding coding sequences (cds). We used the rest of the 19,660 protein sequences in our analysis.

Yeast: ftp://ncbi.nlm.nih.gov/genbank/genomes/S_cerevisiae/ We used the NCBI October 2000 version, which was part of the Reference Sequence (RefSeq) project. The annotation for this version was based on the Saccharomyces Genome Database in the Stanford genomic resources (SGD, http://genome-www.stanford.edu/Saccharomyces/). A total of 6,297 protein sequences were in the database and used in our analysis. Information for block duplications and gene pairs within the blocks was downloaded from the website http://www.gen.tcd.ie/khwolfe/ (Wolfe and Shields 1997).

Drosophila: ftp://ncbi.nlm.nih.gov/genbank/genomes/D_melanogaster/ Release 2, October 2000 from NCBI was used. A total of 14,335 protein sequences were in the database.

The corresponding cds and genomic sequences for all three genomes were also downloaded from the above websites.

The data for each genome was processed as described below.

First Round Grouping

In each of the three genomes studied, every protein was used as the query to search against all other proteins in the database using FASTA (E = 10). Our criteria for two proteins to form a link (i.e., to be in the same family) are (1) the FASTA-alignable region between the two proteins should be longer than 80% of the longer protein, and (2) the identity between the two proteins (I) should be $I \geq 30\%$ if the alignable region is longer than 150 a.a. and $I \geq 0.01n + 4.8L^{-0.32(1 + \exp(-L/1000))}$ (Rost 1999) if otherwise, where L is the alignable length between the two proteins. Rost's formula was derived from an empirical study, which suggests that a higher I value was needed for shorter proteins. We use $n = 6$, which makes the formula continuous at $L = 150$. We call it a hit if two proteins form a link. The single-linkage algorithm was used to group proteins into clusters, i.e., if protein A hits protein B and protein B hits protein C, we group proteins A, B, and C together, regardless of whether protein A hits protein C or not.

Cleaning of Same Genes with Different Names

Occasionally, more than one name is assigned to the same gene and these names are presented as different genes in the database. Such a situation can be detected by comparing their sequence coordinates. In each of such cases, only one copy was kept in the analysis.

Isoform Cleaning

Based on gene and exon annotation, we define that two genes are isoforms if their shared coding region is longer than 20% of the entire coding region of the shorter gene. We delete one of the two isoforms from the database as follows: Delete the shorter one if both are singletons; delete the one that is a singleton if the other belongs to a multigene family; delete the shorter one if the two isoforms form a two-member cluster; delete the shorter one if both are from the same gene family with more than two members and they have the same hits; and delete the one with fewer hits if they belong to the same cluster but their hits are not all the same. We keep both proteins if they belong to different multigene families.

RE Cleaning

Each protein was used as the query to search against the RE database for the same organism using FASTAX ($E = 10^{-5}$). We delete the whole protein sequence from the database if the part of the protein hit by an RE is longer than 80% of the protein itself. We delete only the part that was hit by an RE if it is shorter than 80% of the protein.

Second Round Grouping

We repeat the steps in the first round grouping with the cleaned database. The new clusters are regarded as gene families.

Synonymous-Nonsynonymous Substitution Calculation

If two proteins are linked to each other, the FASTA-alignable regions of the two proteins are realigned using clustalW, and the corresponding coding regions of the genes are aligned based on the protein alignment. The number of substitutions per synonymous site (K_S) and nonsynonymous site (K_A) are calculated using PAML (Yang and Nielsen 2000; the default parameters are used in the calculation).

Genetic Distance in Intron and Flanking Regions

Sequences of entire introns and flanking 150 bp of cds (both upstream and downstream) are extracted using gene annotation data. ClustalW is used to do the alignment, and genetic distances are calculated using Kimura's (1980) two-parameter method when the divergence is less than 5%, but Tamura and Nei's (1993) six-parameter method is used for more divergent sequences.

Codon Usage Bias Calculation

The effective number of codons (ENC) is calculated for each gene using the CodonW package (ftp://molbiol.ox.ac.uk/cu/codonW.tar.Z) and is used as a criterion for the strength of codon usage bias: the smaller the ENC, the stronger the codon usage bias.

The results of the above analysis were stored in a MySQL database.

Table 1
Number of Cases of a Gene with Different Names, Number of Isoforms, and Number of Proteins Hit by REs that were Deleted from the Databases

	Family or Group Size	Yeast	*Drosophila*	*C. elegans*
Same gene	2	0	201	1
	>2	0	215	1
	Total	0	416	2
Isoforms	1	53	197	166
	2	0	150	165
	>2	2	108	116
	Total	55	455	447
REs	Proteins hit by REs ($E = 10^{-5}$)	110	116	506
	Hit length >80% of protein itself	101	59	255

Results

Database Cleaning

Table 1 shows the number of cases of a gene with different names and the number of isoforms that were cleaned from each database used. In the *Drosophila* database, more than 400 cases of a gene with different names were found, suggesting a relatively poor annotation. In yeast, isoforms were found mainly as singletons (i.e., they do not belong to protein families), whereas in *Drosophila* and *C. elegans* the majority of isoforms were found in protein families. The number of proteins hit by REs is also listed in table 1. In Yeast and *C. elegans*, a large number of sequences deleted belong to the RT families. On the other hand, there are much fewer RTs in the current *Drosophila* protein database. It will be interesting to see whether the low number of RTs in *Drosophila* is real or is caused by incomplete genome sequencing. We note that in many cases the part of the protein derived from a RE is less than 80% of the protein (9 in yeast, 57 in *Drosophila*, and 251 in *C. elegans*). Many of these proteins belong to non-RT protein families, although they include part(s) derived from a RE. This observation suggests that REs play an important role in protein evolution (Brosius 1999; Makalowski 2000; Nekrutenko and Li 2001).

The dramatic effect of database cleaning is shown in table 2. For example, for *Drosophila*, before the database was cleaned, there were 660 gene pairs with $K_S < 0.01$ within gene families having fewer than six members, but this number was reduced to two after database cleaning; 413 pairs were the same gene with different names and most of the rest (245 pairs) were isoforms.

Table 2
Number of Gene Pairs Before and After Database Cleaning

K_s	Group Size	Yeast Before	Yeast After	*Drosophila* Before	*Drosophila* After	*C. elegans* Before	*C. elegans* After
<0.01	<6	34	32	660	2	301	76
	All	221	58	761	7	1,700	153
<0.1	<6	68	62	703	10	438	171
	All	930	172	822	26	2,113	379
<0.25	<6	95	88	729	22	535	254
	All	1,220	262	862	52	2,426	664

Number of Gene Families

The numbers of protein families were estimated to be 530, 674, and 1,219 in yeast, *Drosophila*, and *C. elegans*, respectively (table 3). Our estimates are much smaller than those of Rubin et al. (2000) because we conducted a very detailed database cleaning, and our criteria for homology are much more rigorous than theirs (they considered two proteins homologous if the BLASTP E value between them is $\leq 10^{-6}$ but did not require any minimal length of the alignable region). If we count one gene family as one unique gene type, the numbers of unique gene types are estimated to be 5,298, 11,460, and 14,077 in yeast, *Drosophila*, and *C. elegans*, respectively. A striking point is that the number of gene families in *Drosophila* is somewhat larger than that in yeast but much smaller than that in *C. elegans*, although the genome size and total protein number in the whole genome are similar in *Drosophila* and *C. elegans* but much larger than those in yeast. The number of unique gene types in *Drosophila* is more similar to that in *C. elegans* than that in yeast. The top five gene families for each organism are listed in table 4. The largest multiple gene family in yeast is the seripauperin gene family, whereas it switches to trypsin and olfactory receptor gene families in *Drosophila* and *C. elegans*, respectively (table 4).

Table 3
Distributions of Multiple Gene Families in Yeast, *Drosophila*, and *C. elegans*

Family size	Family number: Yeast	*Drosophila*	*C. elegans*
1	4,768	10,786	12,858
2	415	404	665
3	56	113	188
4	23	46	93
5	9	21	71
6–10	19	52	104
11–20	8	26	57
21–50	0	11	33
50–80	0	0	5
>80	0	1	3
Number of gene families	530	674	1,219
Number of unique gene types[a]	5,298	11,460	14,077

[a] One gene family is counted as one unique gene type.

Table 4
Top Five Multiple Gene Families in Yeast, *Drosophila*, and *C. elegans*

Size	Representative proteins
Yeast	
20	Seripauperins
19	Hexose transporters
17	Aminoacid permeases
15	Putative helicase
12	Heat-shock proteins
Drosophila	
111	Trypsins
49	Cuticle proteins
37	GTP-binding proteins
36	P450
34	Cuticle proteins
C. elegans	
242	Ofactory receptors
181	Olfactory receptors
154	Mostly hypothetical proteins
76	Mostly hypothetical proteins
73	Mostly hypothetical proteins, containing F-box domains

Gene Duplication Rates

Table 5 lists the number of duplicate genes with $K_S < 0.01$. We compared K_S with the genetic distances in intron and flanking regions. Those pairs with both the genetic distance in intron and flanking regions (both up and down stream 150 bp) larger than 0.02 were excluded from the table. It is interesting to note that there are only six pairs of duplicate genes in *Drosophila* that have $K_S < 0.01$, whereas this number is 55 in yeast and 147 in *C. elegans* (35 in yeast if we exclude the putative helicase protein family). Assuming similar synonymous mutation rates in these three organisms and using the estimated rate in *Drosophila* (15.6 substitutions per site per 10^9 years; Li 1997, p. 191), we calculated the recent gene duplication rates in the three organisms (table 5). The recent gene duplication rate was found to be more than 10-fold lower in *Drosophila* than in yeast and *C. elegans*. We found that the number of gene pairs with a small K_S did not increase much when we reduced the alignable-length limit from 80% to 50% (six pairs for 80% coverage to seven pairs for 50% coverage). We cannot distinguish gene duplication from domain shuffling if the alignable region coverage is too small. However, even after we remove the alignable region coverage limit, there are still only 12 pairs with $K_S < 0.01$ in the whole *Drosophila* genome.

Discussion

Grouping Strategy

Gene family clustering is a difficult problem for two reasons. First, domain shuffling, which is a common mode for protein evolution, might mislead the clustering of two nonhomologous proteins into the same family because of the shared domain alone. Second, deciding whether two proteins are homologous becomes difficult when their sequence identity is low. In this paper, we improved the criterion adopted by Rost (1999), which is only based on the sequence identity, to identify the homologous relationship between two proteins. In Rost (1999), the cutoff identity is a function of the length of the alignable region. This simple criterion does not specify the proportion of alignable regions. To reduce the chance of putting two nonhomologous proteins into the same family as a result of domain sharing, we required the alignable region between two proteins to be at least 80% of the longer protein. The same criterion was used earlier in identifying homologous genes in *Arabidopsis* (The Arabidopsis Genome Initiative 2000), although most of the previous studies used only a statistic score (the E value or a related score) as the criterion. If we removed the alignable region coverage limit from our criteria and used only the identity level for inferring homologous genes, we found that more than one-third of the proteins in the *Drosophila* genome were grouped into the largest family. In table 4, the olfactory receptor protein families are the largest protein families in *C. elegans*. That the sizes of these protein families we identified are similar to those studied by other authors (Robertson 1998, 2000) suggests that our grouping strategy works reasonably well in clustering homologous proteins into the same group and preventing clustering different proteins together because of domain sharing. However, sometimes the arbitrary alignable-length limit might prevent real homologous proteins from being clustered in the same family. The effect of alignable-length limit needs to be investigated in the future. At any rate, our criteria should be regarded as operational, and our estimates of protein family numbers should be taken with these criteria in mind.

Table 5
Recent Gene Duplication Rate in Yeast, *Drosophila*, and *C. elegans*

	Yeast	*Drosophila*	*C. elegans*
Original pair number with $K_s < 0.01$	58	7	153
Pair number after correction[a]	55[b] (32)[c]	6 (10)	147 (164)
Total protein used in the analysis	6,141	13,405	18,956
Gene duplication rate (per gene per Myr)[d]	0.028[e]	0.0014	0.024

[a] The gene pairs with the genetic distances in the intron and flanking regions (both up and down stream 150 bp) larger than 0.02 were excluded from the analysis.

[b] If the Y′-helicase (putative helicase, Table 4) protein family is excluded, this number is reduced from 55 to 35.

[c] The numbers in the parenthesis are from Lynch and Conery (2000).

[d] The estimated rate of silent-site substitution in *Drosophila* of 15.6 substitutions per site per 10^9 years was used for all three species.

[e] If the Y′-helicase protein family is excluded, this number is 0.018.

Importance of Database Cleaning

We conducted an extensive database cleaning before we did protein family grouping. The presence of same genes with different names and the existence of isoforms (mainly from alternative splicing) in the database inflate the counts of protein families, whereas protein parts derived from REs may cause nonhomologous genes to be clustered into the same family. The database cleaning was found to be very necessary. For example, the number of protein families in the original *Drosophila* database was 1,094, but it was reduced to 674 after database cleaning. The selection of the database for study is also of great importance. The protein and related cds database in NCBI for *C. elegans* was used at the beginning of our study. However, it was found that the database was full of redundant protein sequences, many gene annotations were not consistent with cds sequences, and the translation of many cds were not the same with the corresponding protein sequences. We switched to Wormpep for its better quality.

Paucity of Young Duplicate Genes in *Drosophila*

It is interesting that the gene duplication rate is very low in *Drosophila*. This conclusion comes from two lines of evidence: (1) the number of gene pairs with a small K_S ($K_S < 0.01$) in *Drosophila* is much smaller than those in yeast and *C. elegans* (table 5), and (2) the number of gene families in *Drosophila* is only somewhat larger than that in yeast but much smaller than that in *C. elegans* (table 3), despite the fact that *Drosophila* has a genome size and protein number much larger than those in yeast and similar to those in *C. elegans*. However, this conclusion should be taken with caution because the sequencing strategy for *Drosophila* was different from the other two organisms. It is possible that the shotgun strategy used to sequence the *Drosophila* genome could not distinguish between very recently duplicated genes, which have K_S and K_A equal or close to 0. The fact that we still see some gene pairs with $K_S = K_A = 0$ in *Drosophila* genome and that the number of gene pairs with an intermediate K_S is still much smaller in *Drosophila* than in the other two organisms (table 2) suggests that the sequencing strategy might not be a major factor. In other words, *Drosophila* might indeed have had a much lower duplication rate than yeast and *C. elegans*, at least in the recent past.

It has been shown that block duplication occurs more frequently in yeast and *C. elegans* than in *Drosophila* (Friedman and Hughes 2001). This might largely account for the higher duplication rates in yeast and *C. elegans*. For example, we investigated the most recent gene duplications in yeast and found that among gene pairs with $K_S < 0.01$, there are at least eight duplication events involving more than one gene pair with $K_S < 0.01$. The largest block has seven embedded gene pairs with $K_S < 0.01$. There are other possible explanations. For example, a higher evolutionary rate for synonymous site, a higher deletion rate for duplicate genes, or a lower gene conversion rate in *Drosophila* than the other two organisms might lead to the observed results too. Further investigation is necessary for distinguishing among these possibilities.

Our estimates of gene pairs with $K_S < 0.01$ in the three genomes are different from those of Lynch and Conery (2000) (table 5). Lynch and Conery (2000) found 10 pairs with $K_S < 0.01$ in *Drosophila* but only six pairs are indicated in their current website. Two of these six pairs are the same as ours. In each of the remaining four pairs in our result, either one or both genes do not exist in Release 1 of the fly database from NCBI, which was used by Lynch and Conery (2000). Among the remaining four pairs in their result, there are two pairs for each of which either one or both genes do not exist in Release 2 of the fly database from NCBI, and for each of the two other pairs, the two proteins have very different lengths: 69 a.a. and 220 a.a. for one pair and 210 a.a. and 580 a.a. for the other. For yeast, if we count only the gene pairs in small gene families (size <6), as was done in Lynch and Conery (2000), there are 29 pairs with $K_S < 0.01$ in our result (32 pairs before excluding gene pairs with both genetic distances in intron and flanking regions >0.02). Lynch and Conery found 32 pairs with $K_S < 0.01$. For *C. elegans* we used WormPep40 from the Sanger Centre instead of the database from NCBI. There are 76 and 77 pairs with $K_S < 0.01$ in small (size <6) and large (size >5) gene families, respectively. After excluding gene pairs with both genetic distances in intron and flanking regions >0.02, these numbers became 73 and 74, respectively. In comparison, there were 164 pairs with $K_S < 0.01$ in small gene families (size <6) in Lynch and Conery (2000). Although we used gene pairs with $K_S < 0.01$ in all gene families to estimate the gene duplication rates, our estimates are lower than their results for *Drosophila* and *C. elegans* and similar for yeast if we exclude the putative helicase protein family.

Common Protein Families in the Three Organisms

It has been estimated that about 550 functional chemoreceptors (olfactory receptors) are scattered in the *C. elegans* genome (Robertson 1998). These proteins help nematodes detect different kinds of chemicals. Chemoreceptors can be divided into different protein families, among which *srh*, *str*, *stl*, and *srd* are large ones. The *srh* gene family was estimated to have 214 members (Robertson 2000), most of which fall into our second largest family in *C. elegans*, which has 181 members. Two closely related protein families, the *str* and *stl* (*str*-like protein) protein families, constitute our largest gene family which has 242 members (the sum of proteins in these two families was estimated to be 240, Robertson 1998). The three other most common gene families in nematode (table 4) are all hypothetical protein families. Their function and phylogenetic relationship need to be investigated in the future. It is interesting to note that among the five largest gene families in yeast there are two (seripauperins and putative helicases) located at the end of the chromosomes. Seripauperin genes encode serine-poor relatives of serine-rich proteins. Members in the putative helicase gene family are very similar to

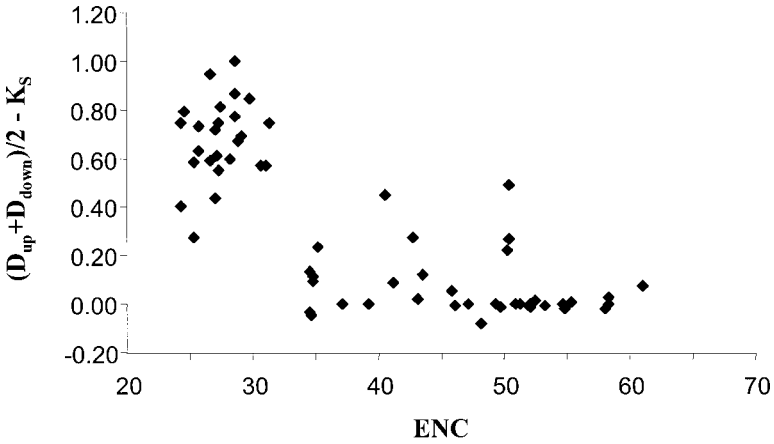

FIG. 1.—Relationship between ENC and the difference between K_S and the average number of nucleotide substitutions per site (K) in the upstream- and downstream-flanking regions for gene pairs with $K_S < 0.25$ in yeast. Only gene pairs within small families (size <6) were considered. The mean of the K values in the upstream (D_{up})- and downstream (D_{down})-flanking 150-bp regions was used to compare with K_S value.

each other. Transcriptions of these genes are not detected under normal culture conditions (Viswanathan et al. 1994; Yamada et al. 1998). However, expression of putative helicases can be detected under some stress conditions (Yamada et al. 1998). Another common gene family in yeast is also a stress response gene family: heat shock proteins. Most of the proteins within this family are cell stress chaperones. The remaining two of the five largest protein families are both membrane proteins. Although yeast has a gene duplication rate as high as that in *C. elegans* and much higher than that in *Drosophila*, there are no protein families in yeast that are as large as those in the other two organisms. Like olfactory receptors in *C. elegans*, cuticle proteins represent two out of the five largest protein families in *Drosophila*. The P450 proteins have been divided into two protein families (Tijet, Helvig, and Feyereisen 2001), one of which is among the five largest protein families in *Drosophila* (table 4), and the other is among the 10 most common protein families in flies (data not shown). The other two largest protein families in *Drosophila* are trypsin and GTP-binding proteins, which are involved in metabolic and regulatory pathways.

Nonneutral Evolution at Synonymous Sites in Yeast

As codon usage is strongly biased in many yeast genes (Ikemura 1982), we now consider whether codon usage bias can affect the estimation of K_S in yeast duplicate genes. We used gene pairs within the same duplicate block to investigate the relationship between K_S and codon usage bias, which is negatively correlated with the ENC in a gene. In principle, gene pairs within a block should have the same age, but we noted that those pairs with a small ENC value (implying strong codon usage bias) tended to have a small K_S value (data not shown). A similar pattern was found by Friedman and Hughes (2001). These results suggest that synonymous sites in genes with strong codon usage bias are under selective constraints. We now compare the K_S values with the numbers of substitutions per site (K) in both the upstream- and downstream-flanking 150-bp regions. We consider only gene pairs with $K_S < 0.25$ in small gene families (size <6) because it is easier to obtain reliable K_S and K when they are not large and because if the family size is larger than six, there may be too many nonindependent pairs. The result is shown in figure 1. For all gene pairs with average ENC ≤ 32, the K_S values are much smaller than the average K values in the flanking regions, whereas for the majority of gene pairs with average ENC ≥ 32, the K_S values are similar to the average K values in the flanking regions. This suggests that for a duplicate gene pair with strong codon usage bias, the K_S value does not increase linearly with the age of the gene duplication. In the above study of recent gene duplication rates, we considered only gene pairs with $K_S < 0.01$ and excluded those gene pairs whose genetic distances in intron and flanking regions were >0.02. This procedure reduced the effect of codon usage bias on K_S (table 5) and should give a more accurate estimate of the rate of gene duplication.

Acknowledgments

This work was supported by NIH grants GM30998, GM55759, and HD38287. A.C. was supported by CAPES-Brasilia/Brazil. We thank A. Nekrutenko, H. Wang, K. Thornton, and E. Stahl for discussion. The comments of the two reviewers and the computer support from R. Blocker are greatly appreciated.

LITERATURE CITED

AKASHI, H., R. M. KLIMAN, and A. EYRE-WALKER. 1998. Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*. Genetica **102/103**:49–60.

THE ARABIDOPSIS GENOME INITIATIVE. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408**:796–815.

BROSIUS, J. 1999. Genome were forged by massive bombardments with retroelements and retrosequences. Genetica **107**: 209–238.

DOOLITTLE, R. F. 1986. Of URFs and ORFs: a primer on how to analyze derived amino acid sequences. University Science Book, Mill Valley, Calif.

———. 1995. The multiplicity of domains in protein. Annu. Rev. Biochem. **64**:287–314.

DUNN, K. A., J. P. BIELAWSKI, and Z. YANG. 2001. Substitution rates in *Drosophila* nuclear genes: implications for translational selection. Genetics **157**:295–305.

FRIEDMAN, R., and A. L. HUGHES. 2001. Gene duplication and the structure of eukaryotic genome. Genome Res. **11**:373–381.

IKEMURA, T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J. Mol. Biol. **158**:573–597.

KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KROGH, A., M. BROWN, I. S. MIANM, K. SJOLANDER, and D. HAUSSLER. 1994. Hidden markov models in computational biology: applications to protein modeling. J. Mol. Biol. **235**: 1501–1531.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

LYNCH, M., and J. S. CONERY. 2000. The evolutionary fate and consequences of duplicate genes. Science **290**:1151–1155.

MAKALOWSKI, W. 2000. Genome scrap yard: how genomes utilize all that junk. Gene **259**:61–67.

MORIYAMA, E. N., and D. L. HARTL. 1993. Codon usage bias and base composition of nuclear genes in *Drosophila*. Genetics **134**:847–858.

NEKRUTENKO, A., and W.-H. LI. 2001. Transposable elements are found in a large number of human protein coding regions. Trends Genet. **17**:619–621.

OHNO, S. 1970. Evolution by gene duplication. Springer-Verlag, Berlin.

ROBERTSON, H. M. 1998. Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae* reveal extensive gene duplication, diversification, movement, and intron loss. Genome Res. **8**: 449–463.

———. 2000. The large *srh* family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res. **10**:192–203.

ROST, B. 1999. Twilight zone for protein sequences alignments. Protein Eng. **12**:85–94.

RUBIN, G. M., M. D. YANDELL, J. R. WORTMAN (54 co-authors). 2000. Comparative genomics of the eukaryotes. Science **287**:2204–2215.

SEOIGHE, C., and K. H. WOLFE. 1999. Updated map of duplicated regions in the yeast genome. Gene **238**:253–261.

SHARP, P. M., and W.-H. LI. 1989. On the rate of DNA sequence evolution in *Drosophila*. J. Mol. Evol. **28**:398–402.

SONNHAMMER, E. L. L., S. R. EDDY, and R. DURBIN. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins **28**:405–420.

TAMURA, K., and M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10**:512–526.

TIJET, N., C. HELVIG, and R. FEYEREISEN. 2001. The cytochrome P450 gene superfamily in *Drosophila melanogaster*: annotation, intron-exon organization and phylogeny. Gene **262**:189–198.

VISWANATHAN, M., G. MUTHUKUMAR, Y. S. CONG, and J. LENARD. 1994. Seripauperins of *Saccharomyces cerevisiae*: a new multigene family encoding serine-poor relatives of serine-rich proteins. Gene **148**:149–153.

WOLFE, K. H., and D. C. SHIELDS. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708–713.

YAMADA, M., N. HAYATSU, A. MATSUURA, and F. ISHIKAWA. 1998. Y′-Help1, a DNA helicase encoded by the yeast subtelomeric Y′ element, is induced in survivors defective for telomerase. J. Biol. Chem. **255**:335–345.

YANG, Z., and R. NIELSEN. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol. Biol. Evol. **17**:32–43.

SAITOU NARUYA, reviewing editor

Accepted October 19, 2001